MAPPING GENE VARIATION IN SUB-SAHARAN AFRICAN POPULATIONS

Khanya Vokwana

A dissertation submitted to the Faculty of Health Sciences, University of Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science

Johannesburg, November 2008

DECLARATION

I, Khanya Vokwana, declare that this dissertation is my own work. It is being submitted for the degree of Science with Masters at the University of Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.

Khanya Vokwana

10th day of November, 2008

ABSTRACT

The present study examined the distribution of six genetic variants (*CYP17A1*, *CYP3A4, SRD5A2, KLK3, AR*) in the androgen biosynthesis and metabolism pathway, in 14 sub-Saharan African populations. These polymorphisms have been implicated in several complex diseases, most notably prostate cancer. In order to elucidate the frequencies of these genetic variants, PCR-RFLP and STR based methodologies were employed. Consistent with previously reported results, the frequency distribution of the gene variants in the examined populations greatly coincided with prostate cancer incidence and geographic origin. Populations of African descent had the highest frequencies of the alleles that are postulated to increase risk to prostate cancer, whilst Asian populations had the lowest. Also, there were evident differences in the frequencies of these variants between populations of different continental origin particularly between African and Eurasian populations. The distribution of these genetic variants was further used to assess the spectrum of variation within Africa. The results were greatly aligned with those previously reported, providing further support to the origin and evolution of modern humans from Africa as well as other historic events.

DEDICATION

To Phaki,

For your never-ending support and love through the many tear-filled late nights at the lab.

You are my North Star.

ACKNOWLEDGEMENTS

Words of gratitude go to the following:

Prof. Himla Soodyall for her patience, support and more importantly her lack of tolerance for mediocrity throughout the years and in the preparation of this dissertation. It undoubtedly made me a better scientist;

All my lab mates at the HGDDR unit, for their support, willingness to help and last but not least their ability to show me the lighter side of life through my many failed experiments;

My friends and family, for the love, support and encouragement when all hope seemed to be lost;

The Medical Research Council for funding my studies.

TABLE OF CONTENTS

DECLARATION i
ABSTRACT ii
DEDICATION iii
ACKNOWLEDGEMENTS iv
TABLE OF CONTENTS v
LIST OF FIGURES vi
LIST OF TABLES vii
ABBREVIATION viii
1 INTRODUCTION 1
1.1. HUMAN GENOMIC VARIATION 1
1.2. GENETIC VARIATION IN SUB-SAHARAN AFRICA 6
1.3. PROSTATE CANCER 8
1.4. AIMS 22
2 MATERIALS AND METHODS 23
2.1 SUBJECTS 23
2.2 METHODS 28
2.3 STATISTICAL ANALYSIS 39
3 RESULTS 43
3.1 SCREENING OF SAMPLES 43
3.2 ALLELE AND GENOTYPE FREQUENCY DISTRIBUTION 46
3.3 POPULATION STATISTICS 62
3.4 POPULATION AFFINITIES 71
4 DISCUSSION 73
4.1 TRENDS AND PATTERNS FROM FREQUENCY DISTRIBUTIONS 74
4.2 GENETIC VARIATION AMONG AFRICAN POPULATIONS 90
5 CONCLUSION 97
5.1 FUTURE STUDIES 98
6 APPENDICES 101
6.1 RECIPES AND SOLUTIONS 101
6.2 INTERPOPULATION COMPARISONS 103
REFERENCE 108

LIST OF FIGURES

Figure 1.1. The androgen metabolism and biosynthesis cascade. 10

Figure 1.2 A diagrammatic presentation of exon 1 of the SRD5A2 gene with the A49T and V89L polymorphisms (Adapted from Makridakis *et al.* 2001)..... 16

Figure 2.1 A map of Africa with the different sampling regions of the examined populations being highlighted in grey. 25

Figure 3.1 The banding patterns observed following digestion of the various SNP variants..... 44

Figure 3.2 An STR profile of a sample with the corresponding sequence data. ... 45

Figure 3.3 Distribution of the CYP17A1 (C) across world populations..... 47

Figure 3.4 Distribution of the CYP3A4 (G) allele amongst world populations. 49

Figure 3.5 Distribution of the SRD5A2 (V89L) allele across world populations 51

Figure 3.6 Distribution of the SRD5A2 (A49T) across world populations 53

Figure 3.7 Distribution of the KLK3 (G) across world populations 55

Figure 3.8 Allelic distribution of the CAG repeats in the present study. 60

Figure 3.9 The distribution of AR alleles (<20 and >20 repeats) in the populations examined in the current study as well as comparative data 61

Figure 3.10. Genetic diversity indices of the three continental groups for all loci.. 66

Figure 3.11 Fst estimates for different population groupings. 69

Figure 3.12 Fst estimates for different genetic variants 69

Figure 3.13 The plot of heterozygosity vs. distance from the centroid for the 14 populations. 70

Fig 3.14 A neighbour-joining tree showing the population affinities of the examined populations..... 72

LIST OF TABLES

Table 2.1 The breakdown of the examined populations according to continental descent with the corresponding code, sample sizes, groupings and sampling region. 24

Table 2.2 A detailed description of the populations and the screened markers that were used as comparative data 26

Table 2.3 PCR primers used for the amplification of regions of interest 30

Table 2.4 The PCR reaction parameters that were followed for the amplification of each SNP 31

Table 2.5 RFLP conditions for the detection of the SNP variants 34

Table 3.1 The breakdown of the distribution of the AR CAG repeat units in the examined populations. 57

Table 3.2 Hardy Weinberg Equilibrium proportions for the examined populations.63

Table 3.3. Gene diversity for the continental groups and each of the 14 populations 64

Table 3.4 Distribution of gene diversity per locus for each of the 14 populations.. 67

Table A1: *CYP17A1* pair-wise population comparisons..... 103

Table A2: *CYP3A4* pair-wise population comparisons 104

Table A3: *SRD5A2* (V89L) population pair-wise comparisons..... 105

Table A4: *SRD5A2* (A49T) pair-wise population comparisons..... 106

Table A5: *KLK3* pair-wise population comparisons..... 107

ABBREVIATION

Bp	: Base pair
BSA	: Bovine Serum Albumin
DDNTP	: Dideoxyribonucleoside triphosphate
DHEA	: Dehydroepiandrosterone
DHT	: Dihydrotestosterone
DNA	: DeoxyriboNucleic Acid
DNTP	: Deoxyribonucleoside triphosphate
Kb	: Kilobase pair
MgCL2	: Magnesium Chloride
PCR	: Polymerase Chain Reaction
PCR-RFLP	: Polymerase Chain Reaction – Restriction Fragment Length Polymorphism
PREG	: Pregnenolone
SNP	: Single Nucleotide Polymorphism
STR	: Short Tandem Repeats
UV	: Ultraviolet

1 INTRODUCTION

1.1. HUMAN GENOMIC VARIATION

The distribution of genetic variation within and amongst populations has proven to be an invaluable tool in determining the genetic structures of populations and in the inference of human evolutionary history (Zhivotovsky *et al*. 2003). Not only does analysis of human genetic variation elucidate the origins and evolution of modern humans, but it also plays a crucial part in determining which subset of variation is responsible for disease susceptibility (Jorde *et al*. 2001).

1.1.1 History of genetic variation

The curiosity of humans to know their origins (Jobling and Tyler-Smith 1995) and to understand the conspicuous differences amongst themselves dates back to antiquity (Carvalli-Sforza *et al*. 1994). The attempts to address these questions have been subject to scientific scrutiny for several centuries.

Gregor Mendel, the father of genetics, who whilst experimenting with garden peas, discovered the principles of heredity and founded the rapidly expanding discipline of genetics, provided the first window of insight into these differences (Tan and Brown 2006). However, the first studies to show clear-cut genetic variation at a molecular level were performed by Landsteiner in 1901, using ABO blood groups

(Cavalli-Sforza *et al.* 1994). For decades after the initial demonstration of genetic variation using ABO blood groups, protein polymorphisms, also known as classical polymorphism, were routinely used to study genetic variation as they were abundantly expressed in the nucleus. However, analysis using protein polymorphisms was limited due to their homogenous distribution across world populations (Harding and McVean 2004). The full extent of genetic variation amongst populations was only realized when analysis was carried out at a DNA level (Cavalli-Sforza *et al.* 1994).

Analysis at the DNA level not only led to the identification of gene variants that causes monogenic disorders but it also shattered the preconceived notion that most variation was between continental groups. Only 10% of all genetic variation could be attributed to differences between continental groups despite the clearly evident differences amongst the continental groups, whilst the bulk of genetic variation was found between individuals of the same population group (Romualdi *et al*. 2002). Although the differences in genetic variation between continental groups might seem small, it is these differences that are partially responsible for the differences in physical appearances as well as disease susceptibility (Bamshad *et al*. 2003).

With advances in DNA typing, sequencing technologies and the successful completion of the Human Genome Project (HGP) a plethora of information on genetic variation came to surface (Jorde *et al*. 1998). This information on genetic

variation is increasingly being used to understand the genetic influences in complex diseases with the intention of developing strategies for prevention, diagnosis and drug development (Cardon *et al.* 2003).

1.1.2 Analysis of genetic variation

In order to determine the extent of genetic variation amongst individuals, some form of analysis needs to be performed. Several evolutionary forces are responsible for the patterns of genetic variation that are seen within and between populations. All genetic variation is introduced into populations through mutations, whilst genetic processes like drift, selection, recombination (Cavalli-Sforza *et al*. 1994) coupled with demographic and historical processes contribute to either the maintenance or the reduction of genetic variation (Jorde *et al.* 1998).

Mutations occur randomly in the human genome at different gene loci. It is these mutations, also called genetic markers, which are used to assess the level of genetic variation amongst different individuals. These can consist of single nucleotide polymorphisms (SNPs), short tandem repeats (STRs) and insertions (Cavalli-Sforza and Feldman 2003). In the current study two types of genetic markers, SNPs and STRs, have been employed to assess genetic variation in sub-Saharan African populations.

Analysis of genetic variation using SNPs and STRs in different genetic systems such as autosomal DNA, mitochondrial DNA (mtDNA) and the non recombining portion of the Y chromosome (NRY), have been successfully exploited in order to study the full extent of human genomic variation (Ray *et al.* 2004) and to also discern the origins of man. Also more recently, there has been an increase in investigations trying to elucidate which subset of these variants is responsible for disease phenotypes (Cardon *et al.* 2003). SNPs and STRs both have attributes that make them attractive for studies on both evolution and disease susceptibility.

SINGLE NUCLEOTIDE POLYMORPHISMS (SNPs)

SNPs are single base pair differences between individuals in the same position in a DNA sequence. SNPs are the simplest and commonest type of markers (Gray *et al*. 2000). It is estimated that there are at least 10 million SNPs in the human genome (Keita *et al.* 2004). Of these, it is predicted that only 50 000- 250 000 have a small to moderate effect on disease manifestation (Rebbeck *et al*. 2004).

SNPs are sometimes referred to as unique event polymorphisms as they tend to reflect unique evolutionary events. Thus individuals with the same variant allele are most likely to share the same evolutionary history (Stoneking 2001) although the possibility of independent evolution cannot be eliminated. SNPs undergo a relatively slow rate of change, making them quite stable and thus very constructive when examining the effects of long-term population subdivision (Romualdi *et al*. 2002).

Also, studying the distribution and allele frequencies of SNPs in different populations has contributed in the understanding of human origins (Taylor *et al*. 2001) and the age of specific mutations (Erichsen and Chanock 2004).

In the post genomic era, interest in SNPs has reached burgeoning proportions. This interest is in direct correlation with the quest to find genetic variants that predispose to complex diseases (Miller *et al.* 2001) and lead to adverse drug effects (Lin and Wu 2005).

SHORT TANDEM REPEATS (STRs)

STRs are simple sequence repeats (Shastry 2002), which are broadly and evenly distributed throughout the genome. They exhibit high levels of allelic variation (Gray *et al.* 2000) due to their high mutation rate (Zhivotovsky *et al.* 2003). This attribute has made them invaluable when trying to determine patterns of recent divergence (Romualdi *et al*. 2002), elucidate the evolutionary relationships of human populations (King *et al.* 2000) as well as make inferences on population expansions and migrations (Zhivotovsky *et al*. 2003). The increased utilization of STRs in genetic variation studies has shown the higher levels of diversity that is present in African populations (Harpending and Cochran 2006).

They have also been exploited in gene mapping efforts for Mendelian diseases using linkage analysis with huge success. Although most STRs are considered neutral,

some have been reported to manifest in a disease phenotype when highly expanded (King *et al*. 2000).

1.2. GENETIC VARIATION IN SUB-SAHARAN AFRICA

Sub-Saharan Africa is not only a region that has important historical implication for all modern humans but it is also the region where the most genetic diversity is observed. However, despite all these virtues, it still remains largely understudied (Tishkoff and Williams 2002).

Following comprehensive research on the origins of mankind, the “Out of Africa” theory has become a widely accepted hypothesis. Several genetic findings have strengthened the validity of this theory (Jorde *et al*. 2001). The “Out of Africa” theory states that *Homo sapiens* initially evolved in Africa and that a group of about 1000 individuals left East Africa approximately 100 000 years ago to colonize the rest of the world. Archaeological and genetic evidence are in concordance in the view of a recent exodus of humans from Africa to colonize the rest of the world (Cavalli-Sforza and Feldman. 2003). It is thus of primary importance to study variation patterns in African populations as present day African populations harbour in their genomes information that is crucial in understanding the evolutionary processes of our species and the mechanisms that have produced the complex patterns of genomic variation found in all humans.

1.2.1 Complex diseases and Africa

There has been a noticeable increase in the prevalence of complex diseases in developed nations in the past two decades (Altmüller *et al.* 2001). Although the aetiology of complex diseases like prostate cancer is still a conundrum (Wright et al. 1999), the familial tendency of most complex diseases strongly suggests a genetic contribution to these diseases (Gray *et al.* 2002). This has prompted extensive investigations to understand the genetic basis of complex diseases.

Complex diseases result from the adverse interactions between multiple genetic variants of varying penetrance and environmental factors. However these two factors do not act solely to bring about the disease state (Wright *et al*. 1999). Association studies have been widely used to determine which gene variants predispose to complex diseases at a population level (Cardon *et al.* 2003) as linkage analysis, which had successfully mapped almost all Mendelian disease genes (Pritchard and Cox 2002), has lacked the ability to detect gene variants that have a small and modest effect on complex disease susceptibility.

The principle behind association studies is to link differences in incidence rate with differences in allele frequencies between groups. However differences in allele frequency between groups doesn't necessarily mean that the variant is associated with the disease phenotype, as demographic and historical processes are also responsible for differences in allele frequencies (Cardon *et al.* 2003). This is why it is imperative to understand the evolutionary processes that have shaped the genetic

variation in the ancestral African population as well as the populations it gave rise to, as it will aid in distinguishing disease-causing genetic variation from stochastic genetic variation that is shaped by historical and demographical processes (Reich and Lander 2001).

Individuals of African descent often have a high prevalence of complex diseases (Tishkoff and Williams 2002). Studies conducted in Jamaica and Brazil have shown that individuals of sub-Saharan African descent have a higher risk of developing prostate cancer than individuals of either European or Asian descent (Reddy *et al.* 2003). Therefore, in order to answer questions on the high prevalence of complex diseases in individuals of African descent as in the case of African Americans, studies among present day African populations would shed more light on the evolution of the gene variants associated with disease. Furthermore most of the environmental factors, which trigger disease manifestation are not highly prevalent in African populations, this will consequently allow for the discrimination of genetic influences from the environmental influences (Tishkoff and Williams 2002).

1.3. PROSTATE CANCER

Prostate cancer is one of the complex diseases that differ with prevalence among ethnic groups (Hsing *et al.* 2000). African-Americans have the highest reported incidence of prostate cancer, which is 60% greater than their European-American counterparts (Keita *et al.* 2004) and 30-50 times higher than the low risk Asian group (Hsing *et al.* 2001).

Prostatic cell growth is maintained and regulated by circulating androgens (Makridakis and Reichardt 2001). The importance of circulating androgens in prostate cells was highlighted, when the administration of testosterone to lab rats induced prostatic cell growth. Following on this observation, researchers examined the genes involved in the androgen biosynthesis and metabolism pathway as candidates for prostate cancer susceptibility. The interest in these gene markers was further elevated, when these variants were found to be differently distributed among the different populations (Habuchi *et al.* 2000). Moreover, the distribution pattern of these variants was highly similar to prostate cancer incidence (Kittles *et al.* 2002; Hsing *et al*. 2001). Currently there is paucity of data on the allele frequencies and distribution of these gene variants in African populations. Thus a robust examination of the candidate gene markers among sub-Saharan African populations would shed light in their distribution in Africa and guide future prostate cancer studies.

The current study focuses on six genetic markers (Five SNPs and one STR) that have been discovered in the androgen biosynthesis and metabolism cascade. These include the cytochrome P450c17α (*CYP17A1*, -34T>C), cytochrome P450 3A4 (*CYP3A4*, -392A>G), Steroid 5α reductase type II (*SRD5A2*, V89L and A49T), Androgen Receptor (*AR*, CAGn) and the Kallikrein Related Peptidase 3 gene (*KLK3*, -158A>G) gene variants. A diagrammatic presentation of where the different variants play a role in the cascade is shown in figure 1.1.

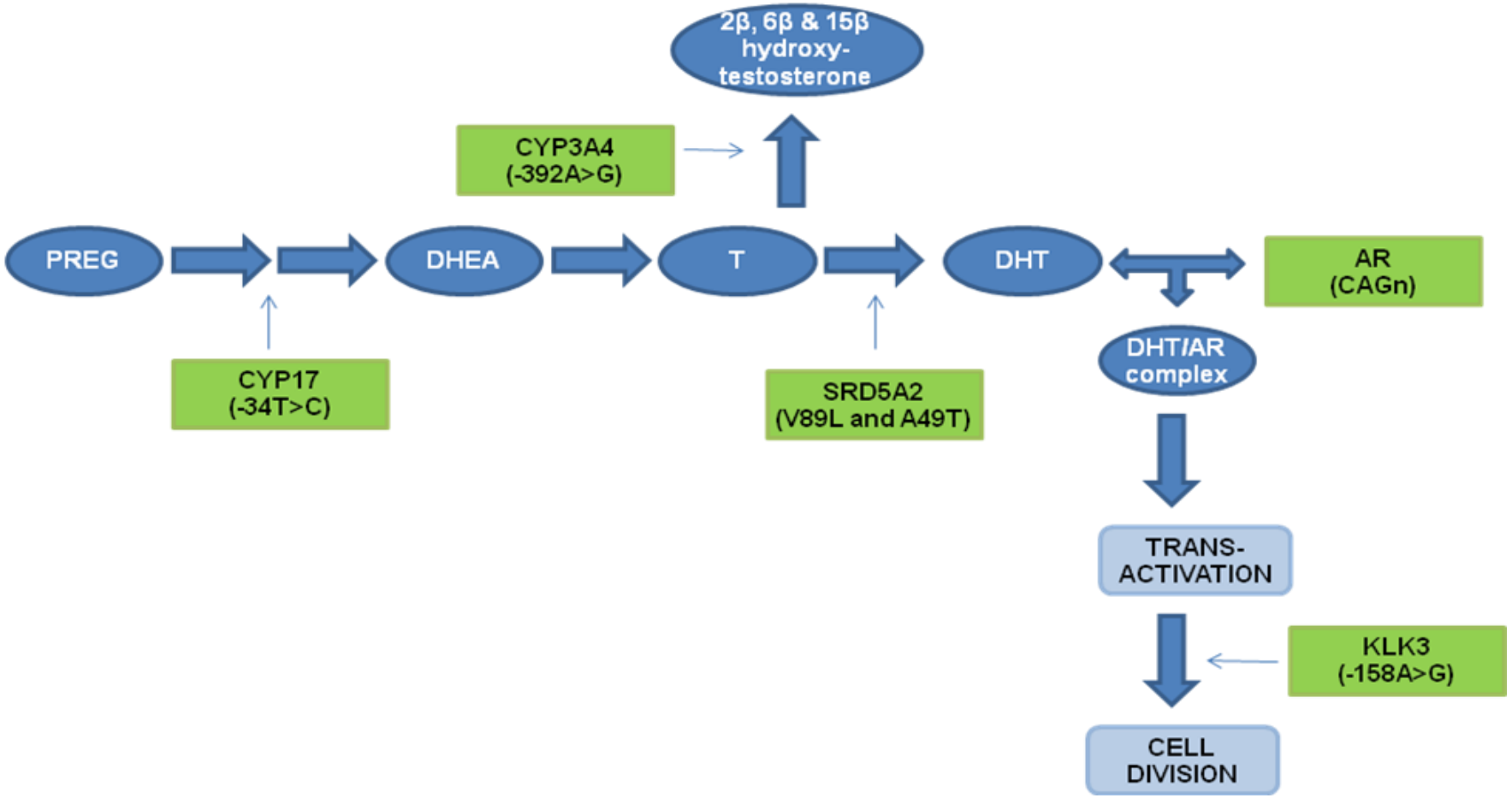


Figure 1.1. The androgen metabolism and biosynthesis cascade. The genes examined in the current study are highlighted in green text boxes. Steroid abbreviations: Preg, pregnenolone; DHEA, dehydroepiandrosterone; T, testosterone; DHT, dihydrotestosterone (Adapted from Makridakis and Reichardt 2001)

1.3.1. The cytochrome P450c17α (*CYP17A1*)

The *CYP17A1* gene maps to chromosome 10q24.3 and is made up of eight exons (Gsur *et al.* 2000). This gene encodes the cytochrome P450c17α enzyme which catalyses the activity of both the 17α-hydroxylase and the 17,20-lyase (Allen *et al.* 2001). These two enzymes play critical roles in two sequential rate-limiting steps in the biosynthesis of testosterone. Initially cholesterol enters the mitochondria and is converted to pregnenolene by *CYP11A1* (Kittles *et al.* 2001). Pregnenolene is then converted to 17αhydroxypregnenolene by 17 α-hydroxylase and finally to dehydroepiandrosterone (DHEA) by the enzyme 17,20-lyase. DHEA, the precursor of testosterone, is the first intermediate product in the testosterone biosynthetic pathway, which shows some level of androgenic activity (Gsur *et al*. 2000). The conversion of pregnenolene to DHEA takes place in the mitochondria and endoplasmic reticulum of the testis and the adrenal cortex.

A -34T>C substitution has been identified in the promoter region, 34 bases from the initiation site of translation and downstream from transcription start site (Sharp *et al.* 2004). This substitution leads to an additional Sp1-type (CCACC) promoter site. Since the number of promoter sites is directly correlated with the transcription activity of a gene, this substitution was hypothesized to lead to an increased risk to prostate cancer due to the increase in transcription activity of the *CYP17A1* gene and as a consequence increased the amount of available testosterone. It was however later shown that the variant allele did not appear to influence Sp-1 binding

(Ntais *et al.* 2003). Furthermore the variant allele did not increase levels of circulating testosterone (Allen *et al.* 2001).

Several case control studies have been conducted to try and determine the effect of the variant allele. These studies however still remain largely inconclusive. Whilst some have found a positive association of the variant C allele with risk for prostate cancer (Gsur *et al.* 2000; Lunn *et al.* 1999; Kittles *et al.* 2001), others found the more commonly distributed T allele to increase its risk (Habuchi *et al.* 2000). In a meta-analysis that was performed, it was found that the variant allele might have important implications for prostate cancer susceptibility in individuals of African descent, but more extensive research on African individuals needed to be performed in order to form conclusive results (Ntais *et al.* 2003).

The variant allele was found to be distributed with varying frequencies across the different world populations. Individuals of East Asian origin exhibit the highest frequency of homozygosity for the variant allele whilst African individuals had the lowest (Sharp *et al.* 2004).

1.3.2. The cytochrome P450 3A4 *(CYP3A4)*

Of the P450 cytochrome supergene family, the cytochrome P450 3A4 is the most abundantly expressed P450 enzyme in the human liver (Gsur *et al.* 2004). Members of the P450 cytochrome family are involved on the oxidation of a number of foreign compounds like drugs, xenobiotics and carcinogens (van Schaik *et al.* 2000). The *CYP3A4* gene, which maps to chromosome 7q21.1, is responsible for the oxidation of more than 50% of all drugs (Cavalli *et al.* 2001).

The interest in this gene in relation to prostate cancer was sparked by its role in the oxidative deactivation of testosterone to the biologically less active forms, 2β, 6β- or 15β-hydroxytestosterone. The protein product of this gene was however found to be only expressed in 61% of all prostate tumors, thus signifying tumor specific variability in the expression of this gene. An -392A>G variant was identified in the 5' regulatory region of the *CYP3A4* gene. It was suggested that this variant allele increases the risk to prostate cancer by decreasing the transcriptional activity of the gene and thus lead to elevated levels of circulating testosterone (Rebbeck *et al.* 1998).

Several studies have been performed to assess the functional implications of the variant allele with regards to prostate cancer. Positive associations for the variant allele with prostate cancer have been described for both European and African American populations (Rebbeck *et al.* 1998; Paris *et al.* 1999) whilst another study

demonstrated that population stratification can confound results of case-control studies, especially in highly admixed populations like African Americans (Kittles *et al.* 2002).

The variant allele is found to differ significantly in the different world population showing a pattern that is very similar to prostate cancer incidence (Kittles *et al.* 2002). The variant G allele was found with a frequency of 0.8 in African populations, 0.6 in African-American, 0.07 in European-American and 0 in Asian populations (Tayeb *et al.* 2000; Cavacco *et al.* 2003; Rebbeck *et al.* 1998; Zeigler-Johnson *et al.* 2002; Kittles *et al.* 2002; Paris *et al.* 1999; Van Schaik *et al.* 2000; Plummer *et al.* 2003; Cavalli *et al.* 2001).

1.3.3. Steroid 5α Reductase Type II (*SRD5A2*)

The *SRD5A2* gene, located on chromosome 2p23, covers a region of approximately 40kb and consists of 5 exons and 4 introns (Hsing *et al*. 2001). Although two isozymes of the gene exist, the *SRD5A1* and *SRD5A2*, the latter is solely expressed in prostate cells (Febbo *et al.* 1999).

The *SRD5A2* enzyme is responsible for the irreversible conversion of testosterone to its more potent form, dihydrotestosterone (DHT). DHT has a binding affinity that is 5 times higher than testosterone (Ntais *et al.* 2003). DHT subsequently binds to the androgen receptor (AR) and this DHT/AR complex transactivates a number of genes in order to bring about cellular proliferation in the prostate (Hsing *et al*.

2001). Deficiency of *SRD5A2* enzyme in males leads to a highly underdeveloped prostate (Makridakis and Reichard 2001), and also leads to presentation of pseudohermaphroditism at birth (Vilchis *et al.* 1997), thus stressing the importance of a normally functioning *SRD5A2* in prostate cells.

It was then suggested that variation in the *SRD5A2* gene could explain the differences in incidence of prostate cancer among the different ethnic groups. Several data supported this hypothesis (Hsing *et al.* 2001). Firstly, when assessing the activity of the enzyme using its known metabolites, African-Americans had the highest activity, followed by European-Americans and lastly by the low risk Japanese group (Ross *et al.* 1992). Secondly, the DHT:testosterone ratio was lowest in Asians and highest among African-Americans men (Ntais *et al.* 2003). This led to the quest to find markers in the *SRD5A2* gene that might be responsible for these discrepancies. Two missense mutations were subsequently identified in the first exon of the *SRD5A2* gene (See figure 1.4), the V89L and the A49T (Gsur *et al.* 2004).

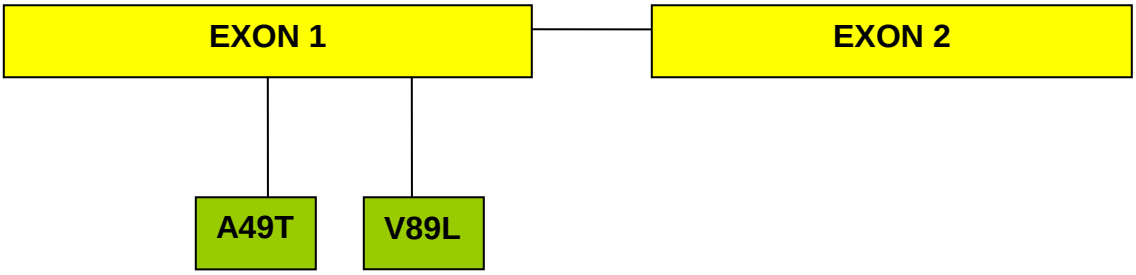


Figure 1.2 A diagrammatic presentation of exon 1 of the SRD5A2 gene with the A49T and V89L polymorphisms (Adapted from Makridakis *et al.* 2001)

THE V89L POLYMORPHISM

The V89L polymorphism is caused by a G to C transversion in codon 89, which changes the amino acid valine to leucine. The C allele was shown to decrease enzyme activity by 30% (Coughlin *et al.* 2002) and was thus assumed to confer protection to those who harboured it due to the decreased DHT production (Ribeiro *et al.* 2002). The distribution of the C allele was also highly correlated with the incidence of prostate cancer, lowest in African-American males and highest in men of Asian origin (Hsing *et al.* 2001). This variation prompted the performance of various studies to determine if the G allele was associated with risk to prostate cancer. Despite all the studies executed, association of the G allele with prostate cancer risk still remains inconsistent. Some studies found a positive association for the G allele with risk and progression to prostate cancer (Nam *et al.* 2001 and Hsing *et al.* 2001), others found the C allele to increase risk to prostate cancer (Söderström *et al.* 2002, Giwercman *et al.* 2005 and Salam *et al.* 2005), whilst

others found no association (Lunn *et al.* 1999; Jaffe *et al.* 2000; Febbo *et al.* 1999; Nam *et al.* 2003; Pearce *et al.* 2002).

THE A49T POLYMORPHISM

The A49T polymorphism is the replacement of an alanine amino acid by a threonine at codon 49. The variant allele exhibited enzyme activity that was 5 fold higher in vitro than the normal allele. Since the variant allele increases the activity of the enzyme, and consequently the production of DHT, it was postulated to increase risk to prostate cancer through increased DHT production (Gsur *et al.* 2004).

Contradictory results on the association of the variant allele with risk to prostate cancer have been observed. Some studies have noted a significant increase in risk to prostate cancer if individuals carry the variant allele (Makridakis *et al.* 1999; Jaffe *et al.* 2000), whilst some have found no association (Hsing *et al.* 2001; Söderstrom *et al.* 2002; Shibata *et al.* 2002). In a meta-analysis that was performed, it was postulated that the variant allele might have a modest effect on risk to prostate cancer but further studies were required to exclude the possibility of a bias or chance findings (Ntais *et al.* 2003).

The variant allele is found in very low frequencies across the different populations, ranging from frequencies of 0.01 in Africans and African-Americans to 0.25 in

populations of European origin. The variant allele has so far not been found in Asian populations (Allen *et al.* 2003; Söderstrom *et al.* 2002; Hsing *et al.* 2001; Makridakis *et al.* 1999; Zeigler-Johnson *et al.* 2002; Ribeiro *et al.* 2002).

1.3.4. The Kallikrein Related Peptidase 3 gene (*KLK3*)

One of the critical roles of the AR is to activate the transcription of certain genes in order to bring about cell proliferation in the prostate. It performs this task by binding to specific nucleotide sequences known as androgen response elements (AREs), which are found in the promoter region of the candidate genes. One such gene is the *KLK3* gene, also known as the Prostate Specific Antigen.

The *KLK3* gene, which maps to chromosome 19q23, was reported to have a polymorphic site (-158A>G) in one of its AREs. This polymorphism is hypothesized to cause alterations in the binding affinity of the AR with the *KLK3* gene and consequently leads to differences in *KLK3* expression (Gsur *et al.* 2004). Inconsistent results on the effect of the polymorphism on prostate cancer susceptibility have been noted. Some investigations yielded positive associations (Xue *et al.* 2000; Gsur *et al.* 2002) and even showed that interaction of the *KLK3* GG genotype with short CAG repeats (<20) increases risk to prostate cancer by five fold (Xue *et al.* 2000) whilst another study found no significant association of the polymorphism with prostate cancer risk (Sieh *et al.* 2006).

Paradoxical results on whether the polymorphism influences *KLK3* serum concentrations were obtained. The AA genotype was shown to increase the *KLK3* serum levels (Xue *et al.* 2001) but a later study refuted such claims when it found a statistically insignificant lower serum levels with the AA genotype (Xu *et al.* 2002).

The *KLK3* gene also encodes a glycoprotein that is secreted by the prostatic epithelial cells (Xu *et al.* 2002). Since it discovery in 1979, the prostate specific antigen (PSA) serum measurements has been routinely used for the early detection and in monitoring the progression of prostate cancer (Gsur *et al.* 2004), as anything that alters the size of the prostate increases PSA serum measurements. PSA serum measurements also differ according to ethnicities; African Americans have higher measurements compared to the Caucasians (Xue *et al.* 2001).

1.3.5. The Androgen Receptor gene (AR)

The AR gene, located on chromosome Xq11-12, occupies a region of approximately 90kb (Kittles *et al.* 2001). It has three domains spanned across eight exons and encodes 918 amino acids. The three domains include the DNA binding, the androgen binding and amino terminal (N-H²) domains (Hsing *et al.* 2000).

The AR gene is principally activated by DHT to form a complex that is able to activate the transcription of androgen responsive genes in order to bring about cell division in the prostate (Shibata *et al.* 2002). The amino terminal, coded by exon

one, is the domain that is responsible for the transactivation of these genes (Chen *et al.* 2002). It also contains a polymorphic trinucleotide (CAG) repeat unit that encodes a polyglutamine tract (Beilin *et al.* 2001). Normal individuals usually carry repeat units ranging from eight to 31 (Giovannucci *et al.* 1997). Highly expandable repeat units have been shown to have decreased transcriptional activities (Hsing *et al.* 2000). In vitro studies have shown that the removal of CAG repeats leads to an increase in transcriptional activity of the AR gene, thus suggesting that the CAG repeats play a part in inhibiting transcription of the AR gene (Chamberlain *et al.* 1994). Men with exceptionally long repeats (>40), suffer from a neuromuscular disorder called the spinal and bulbar muscular atrophy (Kennedy‘s syndrome) which is also characterized by testicular atrophy, androgen insensitivity and infertility to name but a few (Chen *et al.* 2002). Highly expandable alleles in other genes also lead to other well known genetic disorders like Huntington disease and spinal ataxia 1 (Chamberlain *et al.* 1994).

Since, shorter alleles lead to increased transcriptional activity of the AR gene, and consequently the transactivation of genes that bring about cell division in the prostate, it was assumed that the short allele increase risk to prostate cancer. It was even suggested that with each additional CAG repeat, there was a 3 percent reduction in prostate cancer risk (Lange *et al.* 2000). Also the shorter allele were more prevalent in the high risk African-Americans, intermediate in the Caucasian populations and lowest among the low risk Asian populations (Hsing *et al.* 2000), a pattern clearly indicative of prostate cancer incidence (Xue *et al.* 2000). This

observation launched a series of investigations to determine the association between repeat length and prostate cancer risk. Like most association studies the results from such studies proved to be contradictory. Whilst some studies found a positive association between the short alleles with the risk (Irvine *et al.* 1995; Hsing *et al.* 2000; Giovannucci *et al.* 1997; Chen *et al.* 2002), earlier onset (Beilin *et al.* 2001) and advancement of the prostate cancer (Sieh *et al.* 2006), others found no increased predisposition (Lange *et al.* 2000; Gsur *et al.* 2002), gleason grade (Shibata *et al.* 2002) or earlier onset to prostate cancer (Edwards *et al.* 2002).

1.4. AIMS

Differences in the distribution of gene variants among the different populations show that some underlying genetic structure might be present among the different populations. Whilst the current study is not a cancer genetics study, it is however a population based assessment of the frequency and distribution of the gene variants, which are hypothesised to increase risk to prostate cancer, among different world populations. This study is therefore a preliminary screen of genetic variation at these loci among sub-Saharan African populations and makes use of these data for comparative analyses with populations outside of Africa.

More, specifically this current project aims to address two key issues.

- To determine the allele frequency distribution of six gene markers (*CYP17A1*, *CYP3A4*, *SRD5A2 (*V89L and A49T), *AR* and *KLK3)* in the androgen biosynthesis and metabolism pathway in sub- Saharan African populations. These variants have been implicated in various diseases, most notably prostate cancer. Thus far the frequency and distribution of these gene variants have been poorly investigated in African populations. This study aim to address this sampling bias with the intention of adding to the growing body of knowledge on the distribution of these variants across world populations.
- To use these markers to map the spectrum of variation within Africa and to examine the genetic structure of some sub-Saharan African populations.

2 MATERIALS AND METHODS

2.1 SUBJECTS

In order to achieve the objectives outlined in section 1.4, DNA samples from 815 healthy and unrelated males from 14 ethnic groups in sub-Saharan African populations were screened for the six polymorphisms (*CYP17A1*, -34T>C; *CYP3A4*, -392A>G; *SRD5A2*, V89L and A49T; *KLK3*, -158A>G and AR, (CAG)n). The populations studied included the southeastern Bantu speakers from South Africa, southwestern Bantu speakers from Namibia, Ugandans from Uganda, Congolese from Democratic Republic of Congo, Zambians from Zambia, Ubangian speakers from Central African Republic, South African Indians, South African Coloureds, South African Whites, Nama from Namibia, !Kung from Botswana, Sekele from Angola, Kwengo from Angola and the Pygmies from Central African Republic. Table 2.1 gives a detailed description of the different ethnic groups examined with the corresponding sample sizes.

The DNA was extracted from the blood samples of the 815 individuals from the different populations. These blood samples were previously collected by Prof H. Soodyall and Prof T. Jenkins with the individuals' informed consent. The collection of samples was approved by the Ethics Committee from the University of Witwatersrand (Protocol number M050823) for research on human subjects for the use of these samples in the current study.

Table 2.1 The breakdown of the examined populations according to continental descent with the corresponding code, sample sizes, groupings and sampling region.

Name	Code	Grouping	Sample size (N)	Sampling region
AFRICAN DESCENT				
Southeastern Bantu-speakers	SEB	Bantu-speakers	91	South Africa
Southwestern Bantu-speakers	SWB	Bantu-speakers	90	Namibia
Zambians	ZAM	Bantu-speakers	55	Zambia
Congolese	DRC	Bantu-speakers	117	Democratic Republic of Congo
Ugandans	UGA	Bantu-speakers	120	Uganda
Ubangian-speakers	CAR	Bantu-speakers	32	Central African Republic
Khoisan-speaking populations	KHO	Hunter and gatherer	133	
-Nama	NAMA	Hunter and gatherer	19	Namibia
-!Kung	KUNG	Hunter and gatherer	27	Botswana
-Sekele	SEK	Hunter and gatherer	56	Angola
-Kwengo	KWE	Hunter and gatherer	31	Angola
Pygmies	PYG	Hunter and gatherer	27	Central African Republic
South African “Coloureds”	SAC	Mixed ancestry	56	South Africa
EUROPEAN DESCENT				
South African Whites	SAW	European	49	South Africa
ASIAN DESCENT				
South African Indians	SAI	Asian	45	South Africa

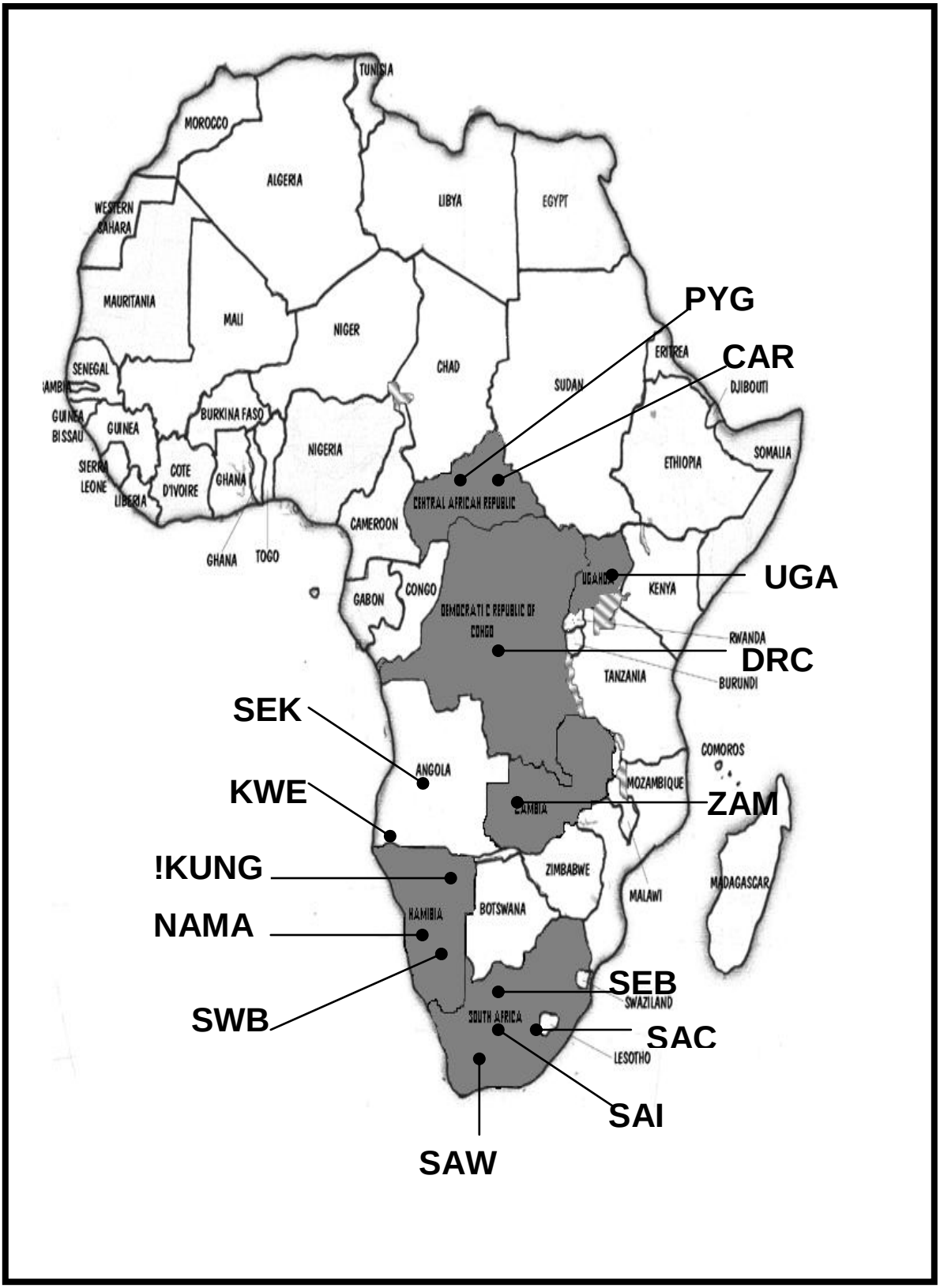


Figure 2.1 A map of Africa with the different sampling regions of the examined populations being highlighted in grey.

2.1.1 COMPARATIVE DATA

The data that was used for comparative purposes was obtained from various scientific publications. Table 2.2 gives a detailed description of the comparative data with corresponding codes, markers typed and the relevant references.

Table 2.2 A detailed description of the populations and the screened markers that were used as comparative data

NAME	CODE	MARKERS	REFERENCE
African descent			
African-Americans	AFR-AMER	CYP17A1, CYP3A4, V89L, A49T, KLK3, AR	Stanford et al. 2002; Lunn et al. 1999, Kittles et al. 2001, Plummer et al. 2003, Paris et al. 1999, Zeigler-Johnson et al. 2002, Kittles et al. 2002, Pearce et al. 2002, Makridakis et al. 1999, Xue et al. 2000 and Xue et al. 2001
Ghanaians	GHA	CYP3A4, V89L, A49T	Tayeb et al. 2000 and Zeigler-Johnson et al.2002
Nigerians	NIG	CYP17A1, CYP3A4, AR	Kittles et al. 2001 and Kittles et al. 2002
Senegal	SEN	CYP3A4, V89L, A49T	Zeigler-Johnson et al. 2002
Sierra Leone	SIL	AR	Kittles et al. 2001
European descent			
Austrians	AST	CYP17A1	Gsur et al. 2000
Britain	BRIT	CYP17A1, V89L, A49T	Allen et al. 2001
Dutch	DUTCH	CYP3A4	Van schaik et al 2000
European-Americans	EUR-AMER	CYP17A1, CYP3A4, V89L, A49T, KLK3, AR	Stanford et al. 2002, Lunn et al. 1999, Kittles et al. 2001, Haiman et al. 2001, Chang et al. 2001, Rebbeck et al. 1998, Zeigler-Johnson et al. 2002, Kittles et al. 2002, Plummer et al. 2003, Paris et al. 1999, Febbo et al. 1999, Pearce et al. 2002, Lunn et al. 1999, Nam et al. 2001, Xue et al. 2000, Xu et al. 2002, Xue et al. 2001, Gsur et al. 2002,
Portugese	POR	CYP3A4	Cavacco et al. 2003
Scottish Caucasians	SCOT	CYP3A4	Tayeb et al. 2000
Sweden	SWE	CYP17A1, V89L, A49T	Wadelius et al. 1999, Söderstrom et al. 2002, Giwercman et al. 2005*

Asian descent			
Amerindians	AMI	AR	Kittles et al. 2001
Asians	ASI	CYP3A4, KLK3, AR	Paris et al. 1999, Xue et al. 2000, Kittles et al. 2001
Chinese	CHI	V89L, AR	Hsing et al. 2001, Hsing et al. 2000
Japanese	JAP	CYP17A1, V89L, KLK3	Habuchi et al. 2000, Yamada et al. 2001, Okugi et al. 2002
Saudi Arabia	SAR	CYP3A4	Tayeb et al. 2000
U.S. Taiwanese	US. TAI	CYP17A1	Lunn et al. 1999
U.S Japanese	U.S. JAP	V89L, KLK3	Pearce et al. 2002, Xue et al. 2001
Latin-Americans			
Hispanics	HIS	CYP3A4, KLK3, V89L	Paris et al. 1999, Xue et al. 2001, Pearce et al. 2002
Mexican	MEX	V89L	Vilchis et al. 1997

2.2 METHODS

2.2.1 QUANTIFICATION OF DNA

The DNA was extracted from the blood samples using the salting out method that had initially been described by Miller *et al.* 1988. The concentrations of the DNA samples were then determined employing a spectrophotometer (Nanodrop, Thermo Fisher Scientific). Following this, the samples were diluted with TE Buffer to the final concentration of 100ng/µl.

2.2.2 PCR-RFLP

All the DNA samples were screened for the five SNPs (*CYP17A1*, T→C; *CYP3A4*, A→G; *SRD5A2*, V89L and A49T and *KLK3*, A→G) employing the Polymerase Chain Reaction- Restriction Fragment Length Polymorphism (PCR-RFLP) analysis. The specific regions containing the polymorphism of interest were amplified into multiple copies using PCR. The PCR products were subsequently digested with the appropriate restriction enzymes to screen for the presence or absence of the polymorphisms.

POLYMERASE CHAIN REACTION (PCR)

One of the inventions that revolutionized the field of molecular Biology was undoubtedly the PCR. PCR is a simple, rapid and yet robust in vitro method of rapidly amplifying target DNA regions into multiple copies. Predesigned and specific oligonucleotide primers that flank the region of interest are used to ensure maximum specificity of the amplification (Strachan and Read. 1996). For the current study, primer pairs that had been employed in previous studies were used to amplify the different loci (as shown in Table 2.3).

All reactions were carried out in 25µl volumes but the PCR reagents and thermal cycling conditions required for optimal amplification of each individual SNP varied. Table 2.4 gives a detailed description of differences in the PCR reagents, cycling conditions and product sizes for each SNP.

Table 2.3 PCR primers used for the amplification of regions of interest

GENE	POLYMORPHISM	PRIMER SEQUENCE	REFERENCE
CYP17A1	T-C	F 5' cca ttc gca ctc tgg agt cat '3 R 5' gac agg agg ctc ttg ggg ta '3	Habuchi et al. 2000
CYP3A4	A-G	F 5' gga cag cca tag aga caa ctg ca'3 R 5' ctt tcc tgc cct gca cag '3	Van Schaik et al. 2000
SRD5A2	V89L (G-C)	F 5' gca gcg gcc acc ggc gag g '3 R 5'agc agg gca gtg cgc tgc act '3	Zeigler-Johnson et al. 2002
	A49T (G-A)	F 5' gca gcg gcc acc ggc gag g '3 R 5'agc agg gca gtg cgc tgc act '3	Zeigler-Johnson et al. 2002
KLK3	G-A	F 5' ttg tat gaa gaa tcg ggg atc gt '3 R 5' tcc ccc agg agc cct ata aaa '3	Xue et al. 2000

Table 2.4 The PCR reaction parameters that were followed for the amplification of each SNP

	CYP17A1	***CYP3A4***	***SRD5A2 (V89L)***	***SRD5A2 (A49T)***	***KLK3***
PCR Reagents					
*†Buffer (10X)	1X	1X	1X	1X	1X
DNTPs (2.5mM)	0.1mM	0.1mM	0.1mM	0.1mM	0.1mM
$MgCl^2$ (25mM)	0	0	1.5mM	1.5mM	0.5mM
Primers (10µM)	0.4mM	0.2mM	0.4mM	0.4mM	0.2mM
DMSO (100%)	0	0	10%	10%	0
DNA	100ng/µl	100ng/µl	100ng/µl	100ng/µl	100ng/µl
*Taq polymerase	1 unit	1 unit	1 unit	1 unit	1 unit
Final volume	**25**	**25**	**25**	**25**	**25**
Cycling conditions	94ºC– 10min	94ºC – 10min	94ºC – 10min	94ºC – 10min	94ºC – 10min
	94 ºC - 1min	94ºC - 1min	94ºC - 1min	94ºC - 1min	94ºC - 1min
	58 ºC - 1min X35	58ºC - 1min X35	58ºC - 1min X35	58ºC - 1min X35	59ºC - 1min X35
	72 ºC - 1min	72ºC - 1min	72ºC - 1min	72ºC - 1min	72ºC - 1min
	72 ºC - 10min	72ºC - 10min	72ºC - 10min	72ºC - 10min	72ºC - 10min
PCR product sizes	421	334	330	330	300

* Roche Applied Science

† Constitutes 500mM KCl, 100mM Tris-HCl (pH 8.3) and 25mM MgCl2

VERIFICATION OF PCR PRODUCT

In order to determine the success of the PCR reactions and more importantly the absence of contamination, 5µl of the PCR product were mixed with 3µl of ficoll dye and were subsequently loaded and run on a 3% agarose gel. Ethidium Bromide was also added to the agarose gel mixture as it intercalates with the DNA fragments and forms a complex that fluoresces when exposed to Ultraviolet (UV) light. A molecular weight ladder was also run on the gel to ensure that the product visualised was of the correct fragment length. PCR products of varying length were observed for the different SNPs (shown in Table 2.4).

RESTRICTION DIGESTION

After the success of the PCR was verified, the PCR products were digested with different restriction enzymes. Restriction enzymes recognize specific DNA sequences, and each time it encounters that particular sequence it will digest the product to produce different fragment lengths. All the SNPs in the current study altered the restriction sites of specific enzymes. Thus the different banding patterns observed following restriction digestion were able to elucidate the different genotypes.

The restriction digestion mixture was made up in a 10µl volume. Included in the mixture was one unit of enzymes, 1X buffer and distilled water to make up the rest

of the volume. The restriction digestion mixture was then added to 20µl of PCR product and incubated at various temperatures for an overnight period. Table 2.5 gives a detailed description of the parameters that were employed to ensure a successful restriction digestion.

Following the overnight incubation period, the digest products were run on a 3% agarose gel. The different banding patterns were observed for each locus (as shown in Table 2.5) and the different genotypes were discerned.

Table 2.5 RFLP conditions for the detection of the SNP variants

	CYP17A1 (T-C)	***CYP3A4 (A-G)***	***SRD5A2(V89L)***	***SRD5A2(A49T)***	***KLK3 (A-G)***
RFLP					
Enzyme	†MSPA1	‡PST1	‡RSA	†MWO	†NHE1
No. of units per enzyme	1U	2U	1U	1U	1U
Buffer	†1X	‡1X	‡1X	†1X	†1X
BSA	100ηg/µl	0	0	0	100 ηg/µl
Incubation temp.	37	37	37	65	37
Banding patterns	TT- 421	AA- 220, 81, *33,	VV- 169, 105, 64, *19	AA- 90, 70, (46/47), *(17/20/21/22)	AA- 300
	TC- 421, 291, 130	AG- 220, 199, 81, *33, *21	VL- 169, 105, 83, 64	AT- 107, 90,70, (46/47),*(17/20/21/22)	AG- 300, 150
	CC- 291, 130	GG- 199, 81, *33, *21	LL- 169, 105, 83	TT- 107, 90, 70, (46/47), *(20/21/22)	GG- 150

‡ New England Biolabs and Roche Applied Science

† New England Biolabs

* The fragment was not visualized on the agarose gels due to its small size

2.2.3 SHORT TANDEM REPEATS (STR)

An STR based assay was employed to analyse the trinucleotide repeat unit, (CAG)n, in the Androgen receptor gene.

ANDROGEN RECEPTOR (CAG) REPEATS

The primer pair that was used to amplify the CAG repeats was previously described by Kittles et al. 2001. The forward primer was tagged with a fluorescent label called Hex that emits a green fluorescence whilst the reverse primer was untagged.

The reaction was carried out in a 25µl volume. The following reagents were added; 0.4µM of both primers, 0.1µM of DNTPs, 1X Taq buffer (Roche Applied Science), 1 unit of Taq (Roche Applied Science), 1mM of MgCl2 and double distilled water to make up the final volume. The Taq buffer comes standard with 500mM KCl, 100mM Tris-HCl (pH 8.3) and 25mM MgCl2.

An initial denaturing cycle of 94ºC for 10 minutes preceded 35 cycles of 94ºC for 1minute, 64ºC for 1 minute and 72 for 1 minute. This was followed by a final extension cycle of 72ºC for 10 minutes.

1µl of PCR product, 2.2 µl of Formamide Dextran and 0.3 of Rox (Applied biosystems) were mixed together. The formamide Dextran was used to keep the DNA denatured and to also prevent it from diffusing out of the gel wells. The Rox was used as an internal size marker in order to allow for the sizing of the different fragments. 1µl of the mix was loaded and run on a 4.3% polyacrylamide vertical gel in an ABI 377 Prism automated DNA sequencer. The GeneScan version 3.1.2 and Genotyper version 2.5 software packages (PE Applied Biosystems) aided in sizing the fragments whilst positive controls that had been previously sequenced were used to authenticate the results and were included in all the runs performed in the automated sequencer.

2.2.4 DNA SEQUENCING

In order to verify that the banding patterns observed after the restriction digestion were the correct ones and to have positive controls for STR based assays, a few samples were sequenced.

The samples were amplified using the previously described SNP and STR protocols. The PCR products were then mixed with ficoll dye and run on 2% Nusieve gel. Nusieve gel is the preferred gel for the excision of DNA fragments as it has a low melting temperature. The DNA fragments were then excised from the gel and purified using the Nucleospin extract column (Macherey Nagel). This purification kit works on the premise that nucleic acid in the presence of a

chaotropic salt will bind to a silica membrane. The nucleic acid can then be eluted in the presence of an elution buffer. Following the purification of the excised DNA fragments, the quality and yield of the purified DNA was visually ascertained by running it on a 3% agarose gel.

The purified DNA was then cycle sequenced. The reaction mixture consisted of 4µl of Big Dye Terminator Kit (Applied Biosystems), 6µl of purified DNA, 1µl of the forward primer (3.3mM) and double distilled water to make up a final volume of 20µl. The reverse primers of the CYP3A4 and AR (CAG) were used for sequencing as the mismatched bases and fluorescent tags in their forward primers would interfere with the sequencing analysis. Cycle sequencing is very similar to a normal PCR reaction except for the use of one primer and the inclusion of fluorescently labeled dideoxydinucleotides (ddNTPs). The ddNTPs lack a hydroxyl group at the 3' terminal of the sugar molecule and therefore cannot form a phosphodiester bond with the next base. Thus the incorporation of a ddNTP into a newly synthesized strand leads to the termination of chain elongation. The thermal cycling conditions for cycle sequencing where as follows; 25 cycles of three consecutive temperatures of 96ºC for 30 sec, 50 ºC for 15 seconds and 60ºC for 4 minute.

The products of cycle sequencing were purified using the Dyex spin column (Qiagen) in order to remove excess ddNTPs that will interfere with the sequencing analysis.

Subsequent to the purification of cycle sequencing products, the products were dried in a vacuum chamber to form a pellet. The pellet was mixed with 4µl of Formamide Dextran. The DNA was denatured at 100ºC for 2 minutes prior to loading on the sequencing machine. Iµl of denatured products were loaded on the sequencer and run for approximately seven hours. The sequences were analysed using ABI 377 prism software.

2.3 STATISTICAL ANALYSIS

2.3.1 DATA MANAGEMENT

The data was recorded and managed using a spreadsheet facilitated by the Microsoft Excel software. Genotypes at each locus were scored in all individuals and their frequencies were computed for each population using direct gene counting.

2.3.2 DATA ANALYSIS

HARDY-WEINBERG EQUILIBRIUM

In order to determine if there were any deviations from Hardy-Weinberg Equilibrium, Arlequin version 3 was used (Excoffier et al. 2005). Hardy-Weinberg is the state where the genotypic frequencies observed in the population are the result of random mating. Several factors can lead to deviations in Hardy Weinberg; these include selection and genetic drift. Individual populations were all tested at each of the loci for deviations from the state of Hardy Weinberg Equilibrium.

INTRA-POPULATION DIVERSITY

The level of intrapopulation diversity for the current study was assessed at two levels. Firstly, the overall extent of gene differentiation exhibited by each population

using all the markers was examined. This allowed for the distinction of which populations exhibited the most overall genetic diversity for the markers examined. This level of analysis was computed using the software Dispan (Dispan 1993)

The extent of gene differentiation exhibited by each population for each individual marker was also determined. This was achieved using the gene diversity (Heterozygosity) index executed by Arlequin version 3 (Excoffier *et al.* 2005). The amount of genetic diversity observed in a population and specific markers can aid to make inferences on the age of the polymorphism as well as the demographic history of a particular population.

INTER-POPULATION DIFFERENTIATION

In order to determine the statistical significance of the variation in allele frequencies between the different populations examined, two-tailed Fisher's exact test implemented by GraphPad InStat version 3.00 (for Windows XP, GraphPad Software, San Diego California USA, www.graphpad.com). The allele frequencies of various world populations were also compared to the frequencies of the currently examined populations in order to detect variations that were of statistical significance.

Furthermore, the Wrights Fst Statistic was computed using Genetic Distance Analysis programme to quantify the extent of between-population variation, among

different population groupings populations (Lewis and Zaykin 2000). The Fst statistic provides a good measure of the amount of genetic differences between populations. The higher the Fst statistic, the greater the between-population differentiation amongst that grouping.

ANALYSIS OF GENE FLOW

In order to determine the extent of gene flow within the examined populations, the model that was described by Harpending and Ward (1982) was employed. This was achieved by plotting the distance from the centroid (Ri) vs. heterozygosity. Based on this model, there is an expected simple and linear relationship between these two variants. The distance from the centroid was calculated according to this formula, Ri = (pi –P)2/(P)(1-P), where pi is the frequency of the variant alleles in a population i and P is the frequency of the variant alleles in the total populations. This equation was used for all six individual gene variants and was then averaged out across all the gene variants.

POPULATION AFFINITIES

The allele frequency data was also to generate population pair-wise genetic distances (DA distances) using the Dispan programme (Dispan 1993). Using these population pair-wise genetic distances, neighbour joining tree was constructed in

order to infer the genetic affinities of the examined populations. Tree analysis provides for a visual form of assessing population similarities and diversities.

3 RESULTS

3.1 SCREENING OF SAMPLES

Having screened 815 male samples from 14 sub-Saharan populations for six genetic markers, the individual genotypes of the various markers were ascertained by gene counting. Allele frequencies were also determined using direct counting at each locus and for each population. The genotypes of the five SNP variants were ascertained using PCR-RFLP whilst the Androgen Receptor repeats units were determined employing STR based methodologies.

Figure 3.1 (a-e) shows the various banding patterns on agarose gel of the five SNP variants that were observed following restriction digestions to elucidate the respective genotypes. Figure 3.2 is an example of an individual's STR profile for the AR (CAG) repeats with the corresponding sequence data.

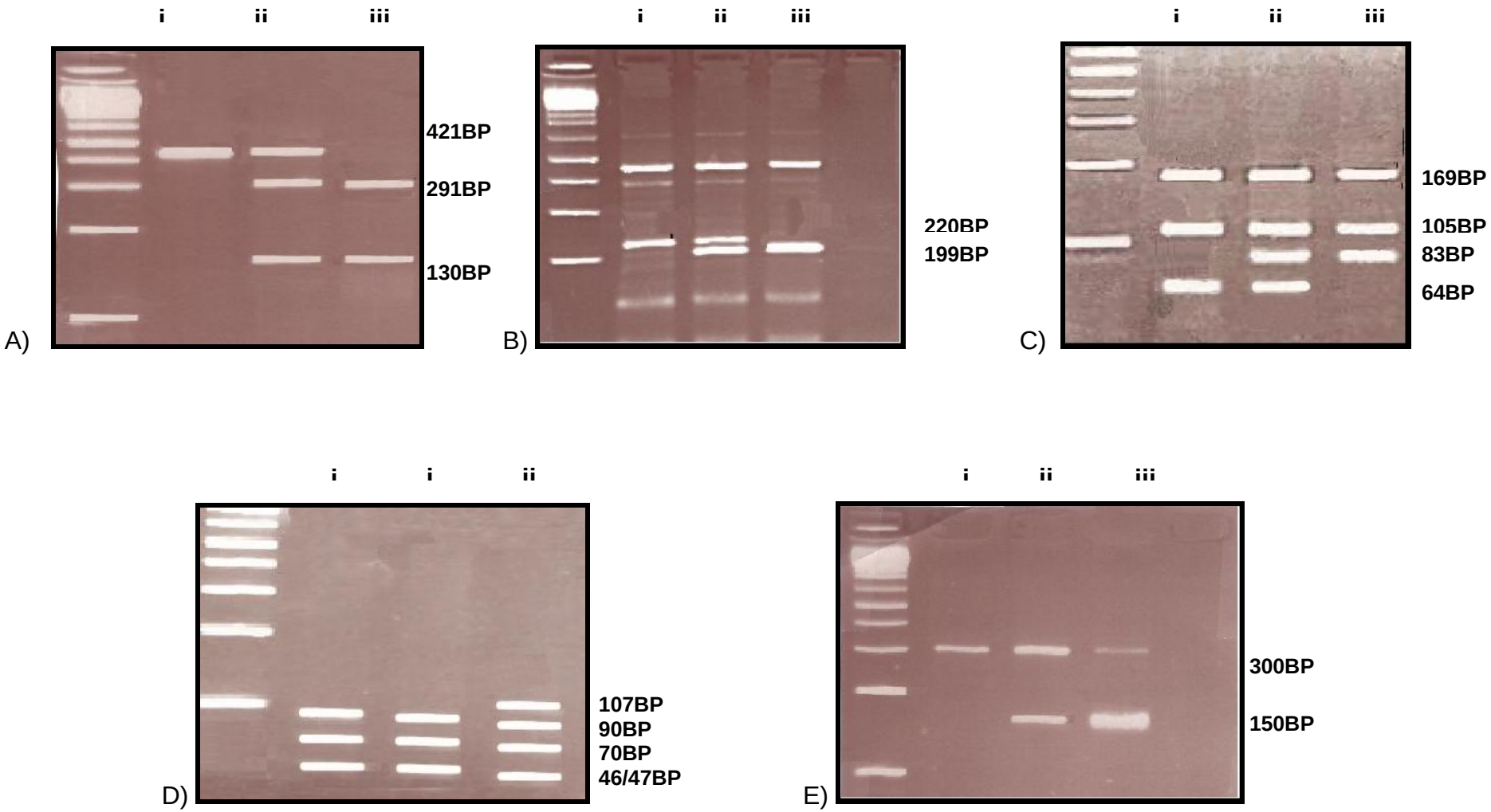


Figure 3.1 The banding patterns observed following digestion of the various SNP variants (i, Homozygous wild-type; ii, Heterozygous and iii, Homozygous for the mutant). A) CYP17A1; B) CYP3A4; C) SRD5A2 (V89L); D) SRD5A2 (A49T) and E) KLK3.

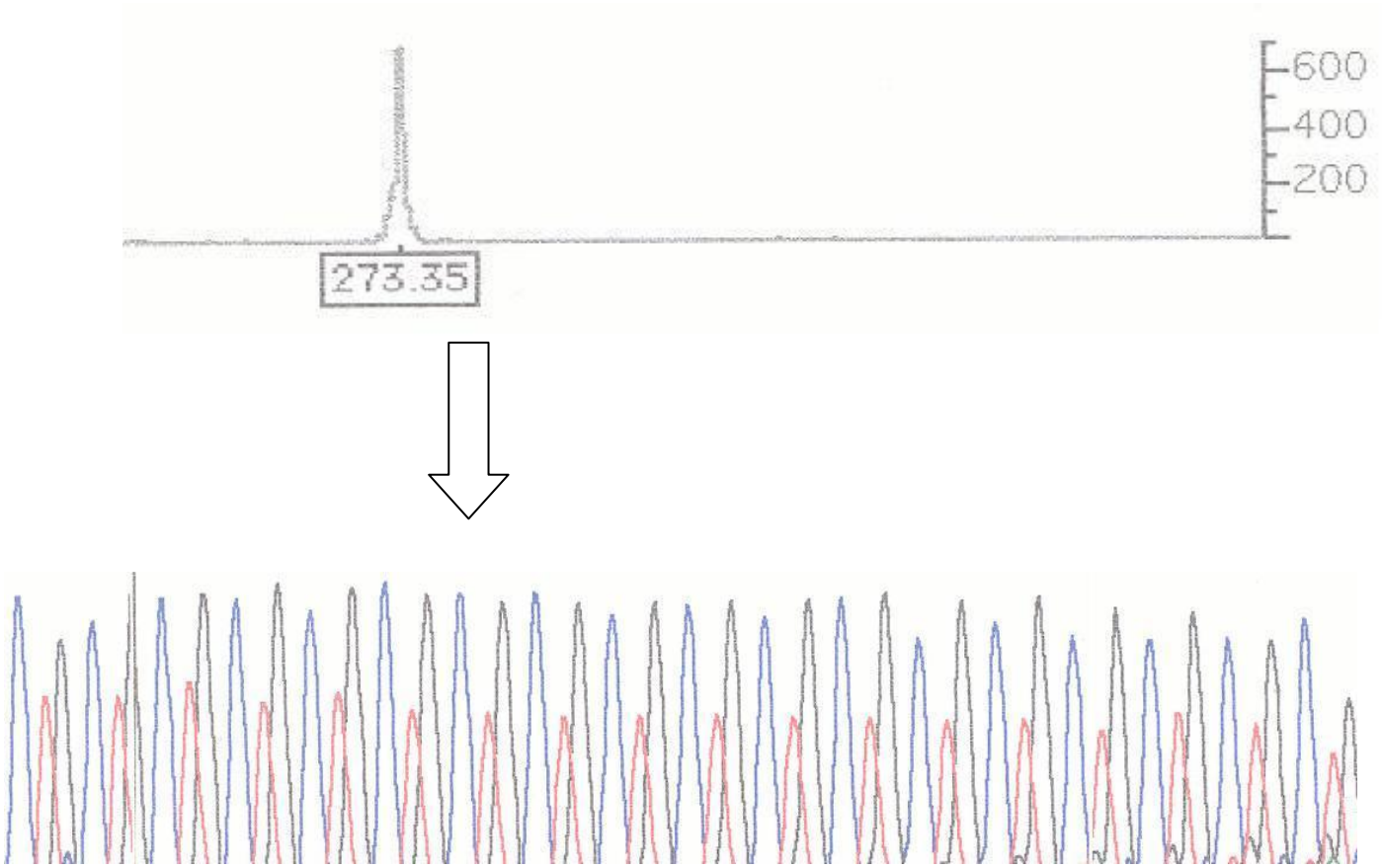


Figure 3.2 An STR profile of a sample with the corresponding sequence data.

3.2 ALLELE AND GENOTYPE FREQUENCY DISTRIBUTION

3.2.1 Cytochrome P450c17α (*CYP17A1*)

The allele frequency distribution of the C allele amongst the 14 populations as well as comparative data from other sources is given in Figure 3.3. Varying frequencies of the variant allele were noted with the highest frequencies being 0.61, 0.59, 0.59 and 0.52 in the Nama, !Kung, Pygmies and Kwengo, respectively. The SAI and Ubangian speakers from CAR had the lowest observed frequencies of 0.24 and 0.22, respectively.

Following population pair-wise comparisons, a significant difference between the Ubangian speakers from CAR and pygmies was noted ($P < 0.0001$) despite the close living proximities of these two populations. Whilst the CAR had the lowest observed frequency of the variant C allele, the Pygmies had the second highest frequency (shown in figure 3.3).

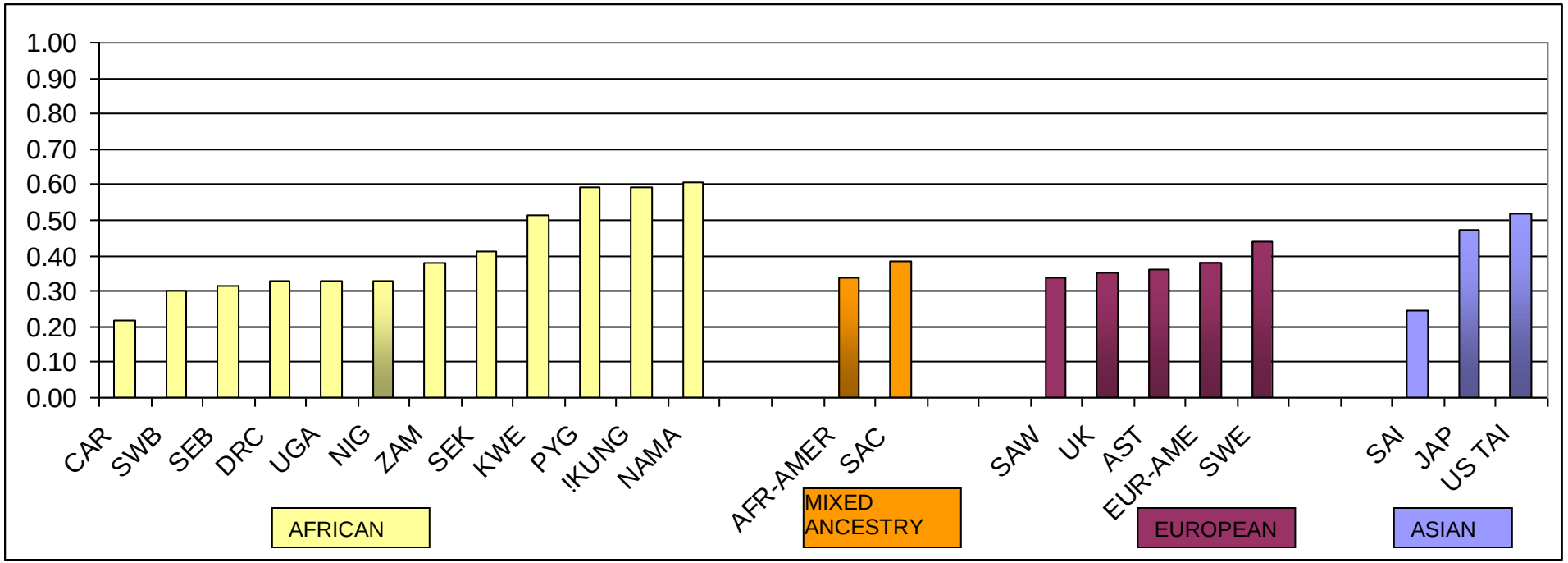


Figure 3.3 Distribution of the CYP17A1 (C) across world populations

3.2.2 The cytochrome P450 3A4 (*CYP3A4)*

There was a marked difference in the distribution of the G allele among populations from the different continents. The variant G allele was prominent in African populations with frequencies ranging from 0.67 to 0.85 (see Figure 3.4). Of the African populations, the !kung had the highest frequency (0.85) for the variant allele and was the only population that had no individuals that were homozygous for the A allele. However, frequencies of below 12% were found in the SAW and SAI populations from the current dataset. The SAI had the lowest frequency for the G allele and also did not have any individuals that were homozygous for the G allele, contrary to the !Kung population. Moreover the SAC were found to have frequencies that were intermediate between these two populations.

Consistent with patterns observed in the current dataset, data from other sources also showed that the African populations possess the highest frequencies of the variant allele. Also, the frequency distribution of the variant allele decreased considerably with increasing distance from Africa (shown in figure 3.4).

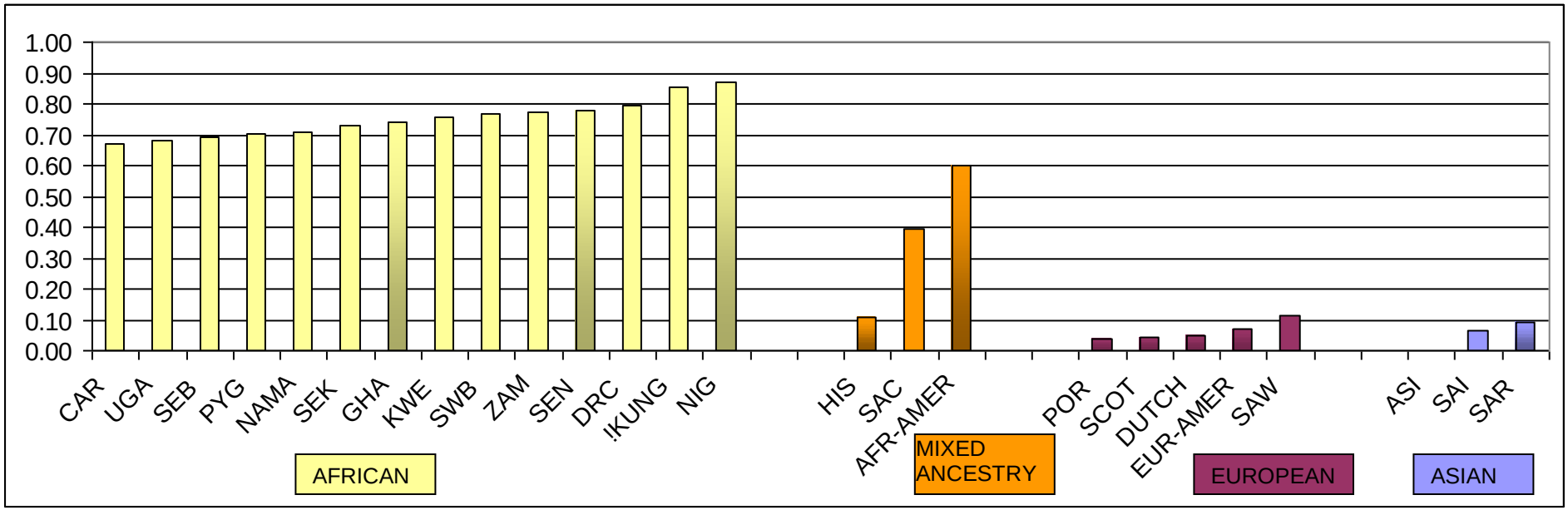


Figure 3.4 Distribution of the CYP3A4 (G) allele amongst world populations.

3.2.3 Steroid 5α Reductase Type II (*SRD5A2*) V89L

The V89L polymorphism was also found to be differently distributed among the different populations. A trend of increasing distribution with increasing distance from Africa was noted. The protective leucine amino acid was found at highest frequencies in populations of Asian (0.4 in SAI) and European (0.37 in SAW) descent in our current dataset and lowest in the hunter gatherer populations with the exception of the Kwengo.

When pair-wise population comparisons were computed, the Kwengo were found to differ significantly from the other two Khoisan populations, the Nama (P=0.0350) and the Sekele (P=0.0027). On the contrary, when the Kwengo population was compared to all the Bantu speaking populations, no significant differences were noted.

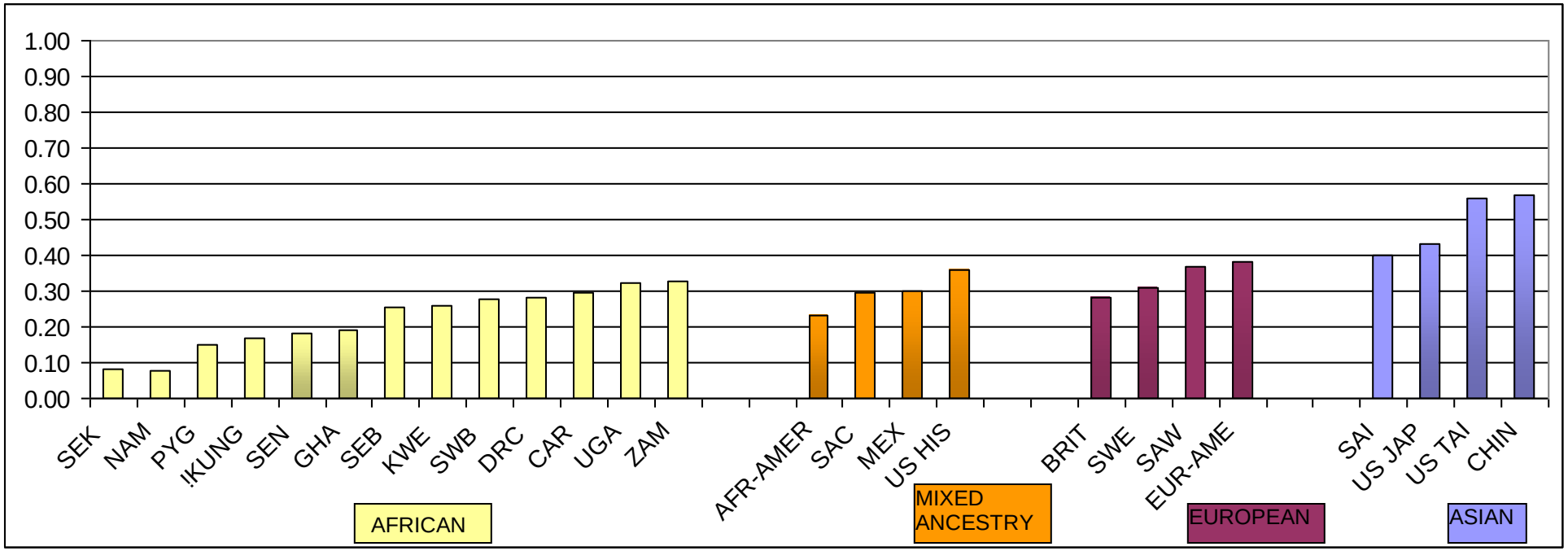


Figure 3.5 Distribution of the SRD5A2 (V89L) allele across world populations

3.2.4 Steroid 5α Reductase Type II (*SRD5A2*) A49T

Of the six loci that were examined, the A49T gene variant was the only one to be found to be highly monomorphic. The low levels of variability exhibited by this variant have been previously documented. In our current dataset, the T variant was only noted in three of the populations, the Nama, SAC and SAW with frequencies of 0.027, 0.026 and 0.02, respectively. No individuals from our current dataset were found to be homozygous for the threonine allele.

Four African individuals were found to be heterozygous for the variant allele (three SAC and one Nama). Although the threonine allele has been found in other African populations *viz.* Senegal, Ghana and African- American albeit in very low frequencies (0.01), this allele was not detected in the currently examined African populations except for the individuals in the Khoisan and SAC populations. Consistent with previous reports on the absence of the variant allele in Asian populations (Hsing et al. 2001), this allele was also not found in the SAI population. The threonine allele, in the presence of published data, was found in higher frequencies in populations of European descent.

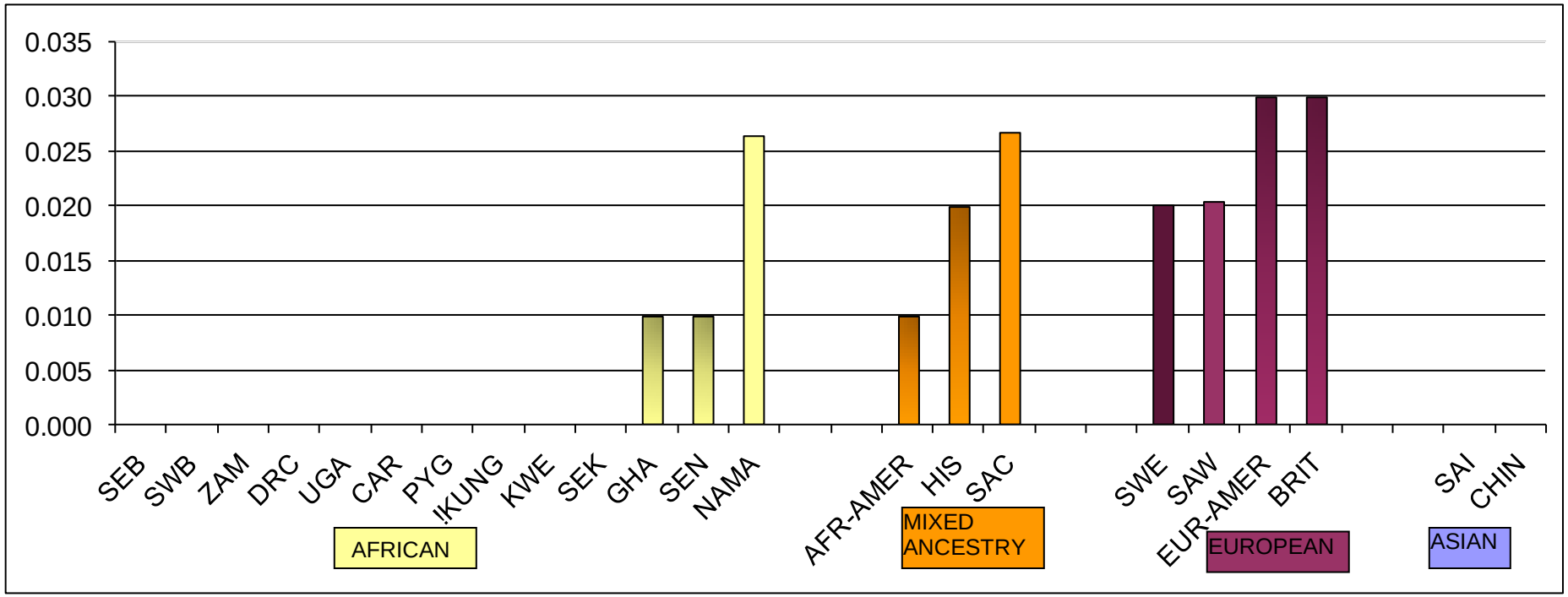


Figure 3.6 Distribution of the SRD5A2 (A49T) across world populations

3.2.5 Kallikrein Related Peptidase 3 gene (*KLK3*)

The KLK3 (G) allelic variant was found at varying frequencies in the examined populations. In contrast to the other markers, there was no distinct trend followed in the distribution of the allelic variant amongst the different populations that coincided with geographic origin. The highest frequencies were observed in the Nama and SWB with frequencies of 0.66 and 0.59, respectively whilst the lowest were however noted in the CAR and Zambia populations with frequencies of 0.36 and 0.35 respectively.

Although the variant allele has been previously reported in frequencies as high as 0.81 in populations of Asian descent, these high frequencies were not observed in the SAI populations, which was only 0.54. Also, in the presence of comparative data, Asian populations usually have the highest reported frequencies amongst the dataset, but in our current dataset, the SAI did not show this pattern, as it did not have the highest frequency of the variant G allele. When the SAI was compared to another population of Asian descent, the Japanese- Americans, a statistically significant difference was noted ($P < 0.0001$).

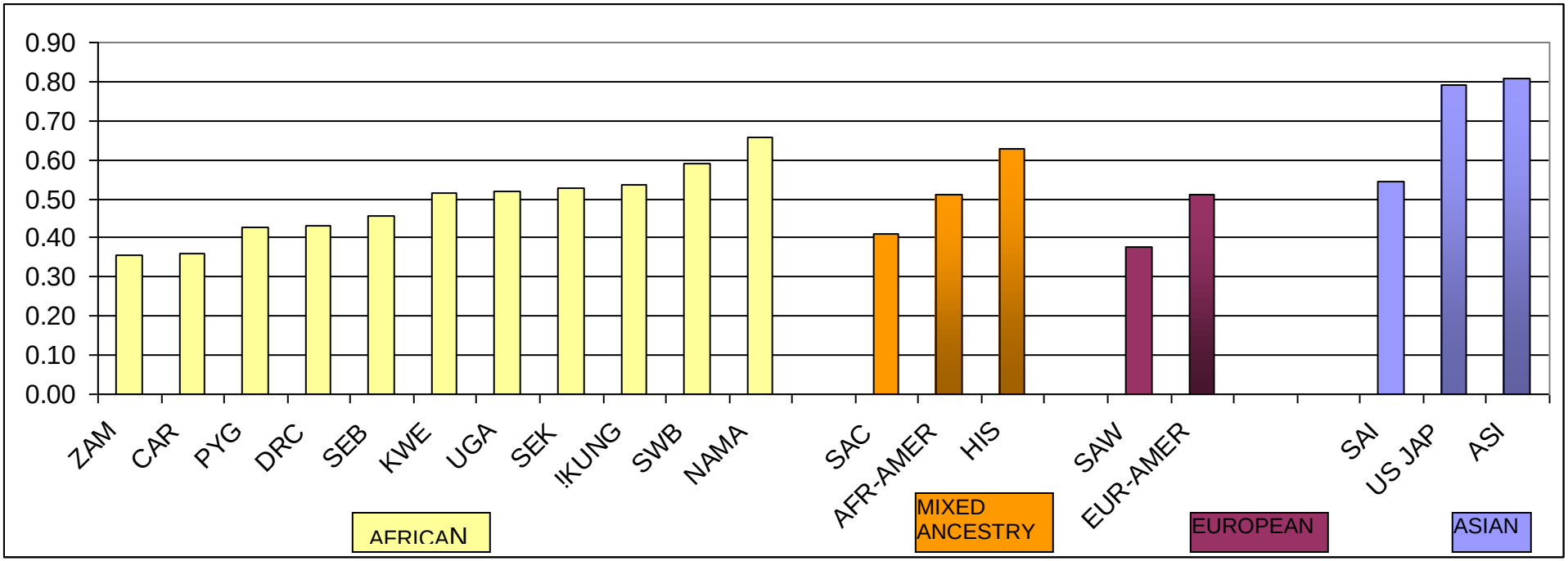


Figure 3.7 Distribution of the KLK3 (G) across world populations

3.2.6 Androgen Receptor (*AR)*

Figure 3.8- 3.9 illustrates the distribution of the *AR* CAG repeats in the examined populations. The number of repeats ranged from 7-39 repeat units. The shortest repeat units were found in two individual, one from the DRC and another from the Ugandan populations. The individual with the longest repeat unit was from the Nama population (see figure 8).

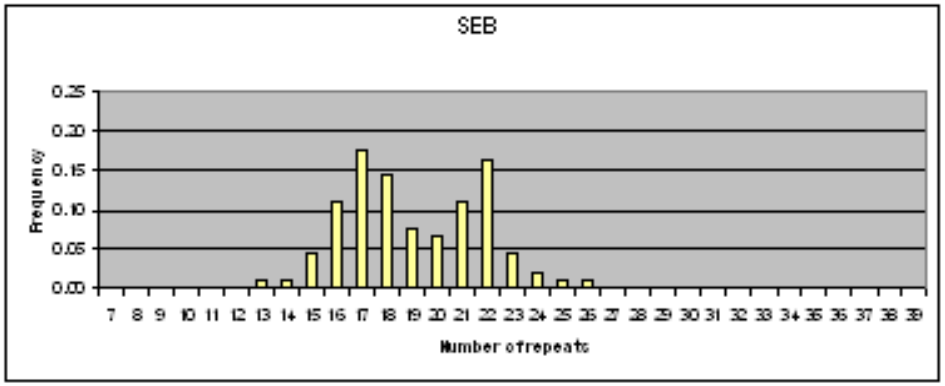
There was a great distinction in the distribution of repeat units of less than 20 amongst the populations of African descent. Whilst the Bantu-speaking populations were found to have a high prevalence of the short repeat units, the same trend was not noted in other African populations *viz.* the Hunter and Gatherer populations (Nama, !Kung, Kwengo, Sekele and Pygmies) and SAC. Moreover, all the Bantu-speaking populations with the exception of Uganda possess modal repeat units that were less than 20 (see table 3.1). The Eurasian populations (SAW and SAI), on the other hand, also exhibited very low frequencies of these repeat units.

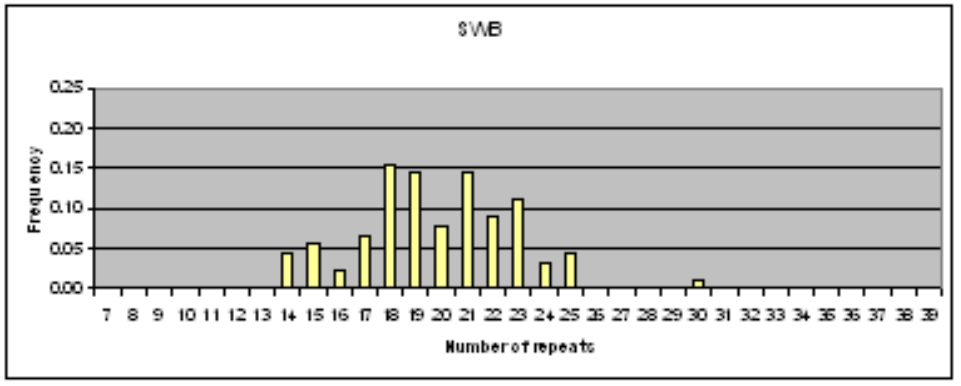
In the presence of comparative data (see Table 3.1), other populations of African descent like Nigeria, Sierra Leone and African-Americans were also found to have a high frequency of repeat units of less than 20 (Kittles *et al.* 2001).

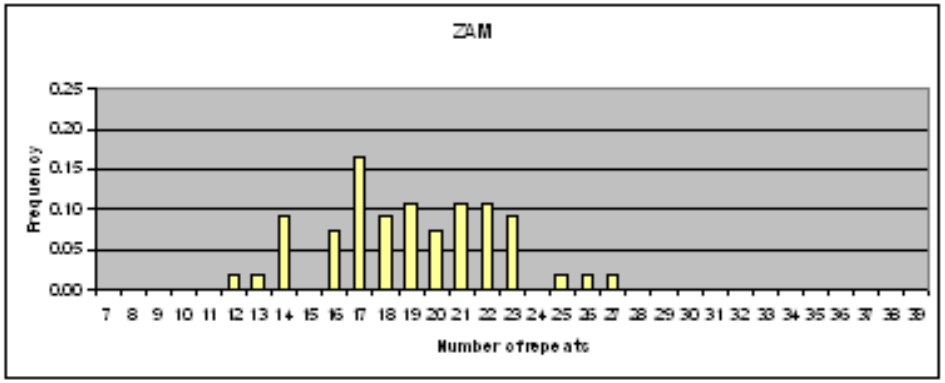
Table 3.1 The breakdown of the distribution of the AR CAG repeat units in the examined populations.

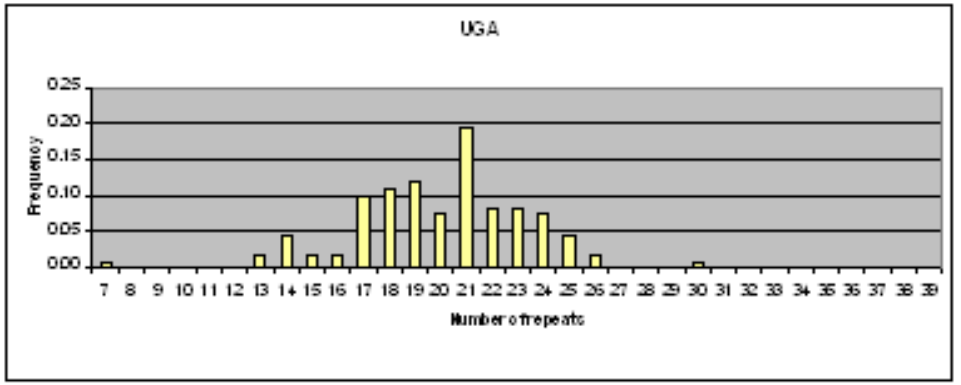
Population	N	No. of alleles	Range	Modal repeat unit
AFRICAN				
SEB	91	14	13-26	17
SWB	90	13	14-30	18
ZAM	55	14	12-27	17
DRC	117	17	7-28	18
UGA	120	16	7-30	21
CAR	32	12	14-26	16, 17, 18
PYG	27	10	12-26	22
NAMA	19	12	14-39	21
!KUNG	27	10	14-26	23
SEK	56	13	13-29	20
KWE	31	8	17-25	21,22,23,25
*SIL	230	17	10-26	16
*NIG	83	18	5-28	16
MIXED ANCESTRY				
SAC	56	14	7-32	20
*AFR-AMER	516	21	9-31	15
EUROPEAN				
SAW	49	10	18-28	23
*EUR-AMER	87	11	13-26	20
ASIAN				
SAI	45	13	17-30	24, 26
*ASI	60	12	14-26	20
*AMI	80	14	14-30	22

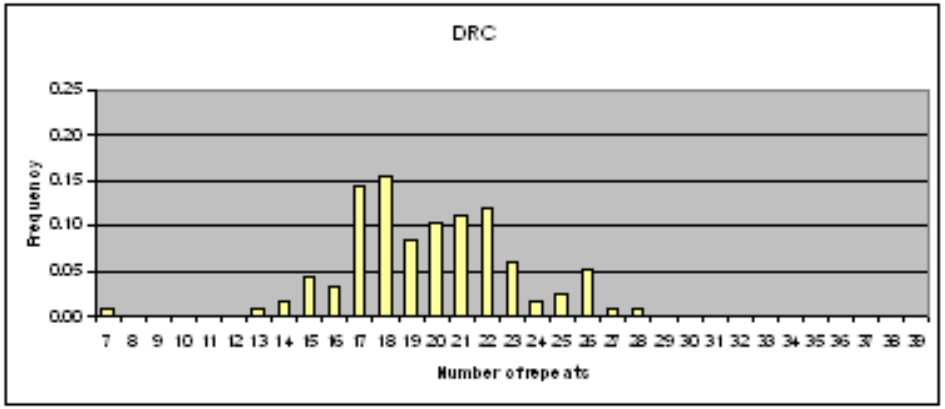
* Comparative data from Kittles et al. 2001

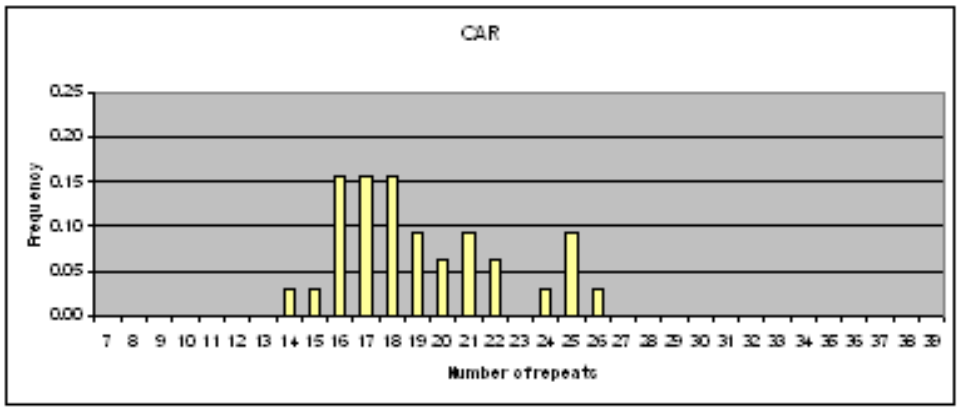

SEB
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Number of repeats

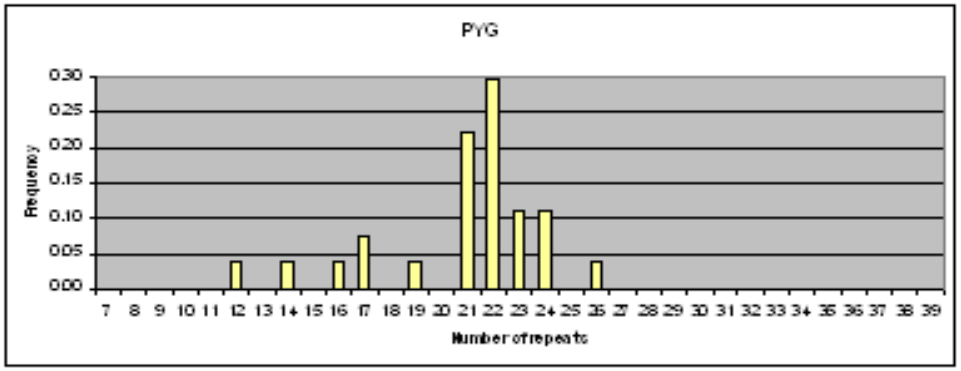

SWB
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Number of repeats

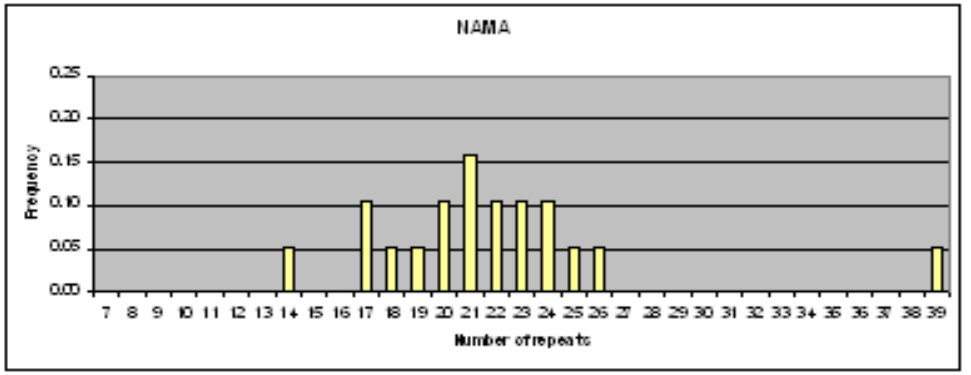

ZAM
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Number of repeats

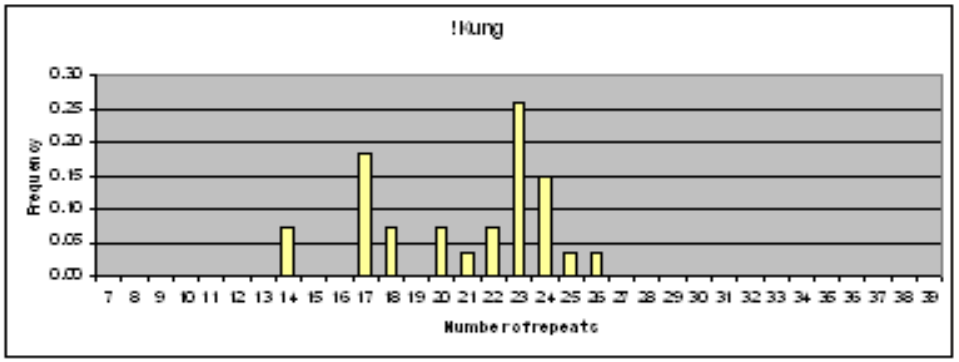

UGA
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Number of repeats

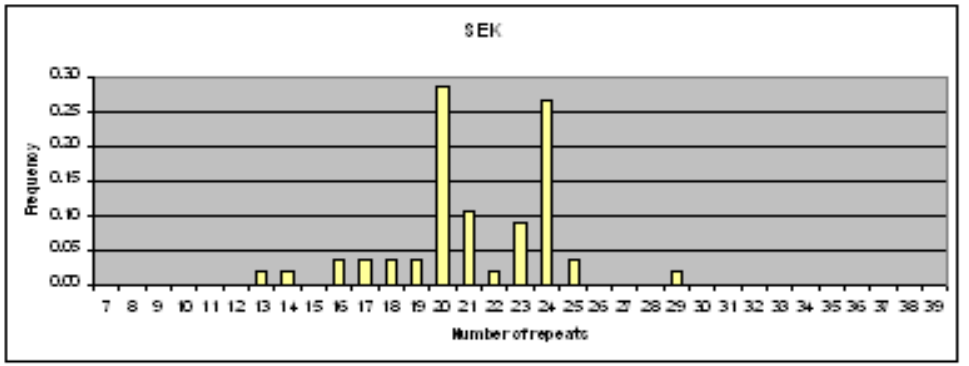

DRC
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Number of repeats

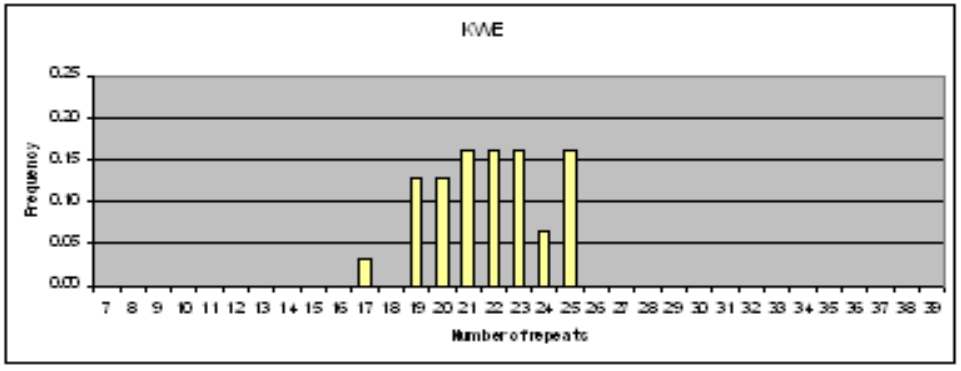

CAR
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39
Number of repeats

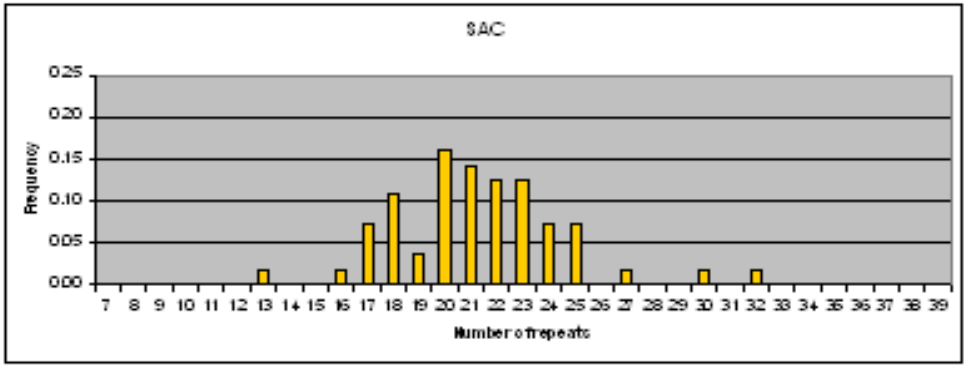

PYG
Frequency
Number of repeats


NAMA
Frequency
Number of repeats


!Kung
Frequency
Number of repeats


SEK
Frequency
Number of repeats


KWE
Frequency
Number of repeats


SAC
Frequency
Number of repeats

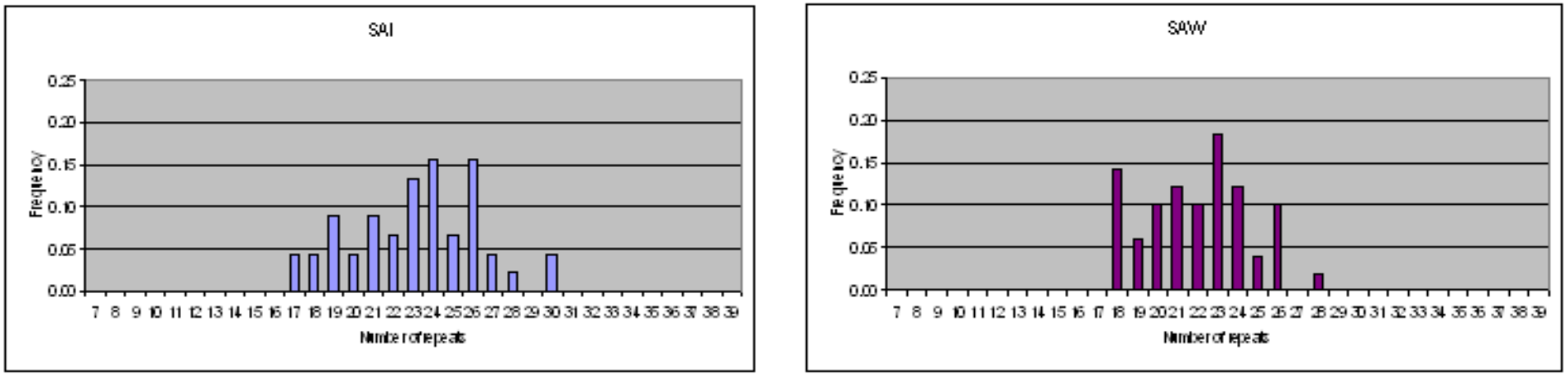


Figure 3.8 Allelic distribution of the CAG repeats in the present study (see Table 2.1).

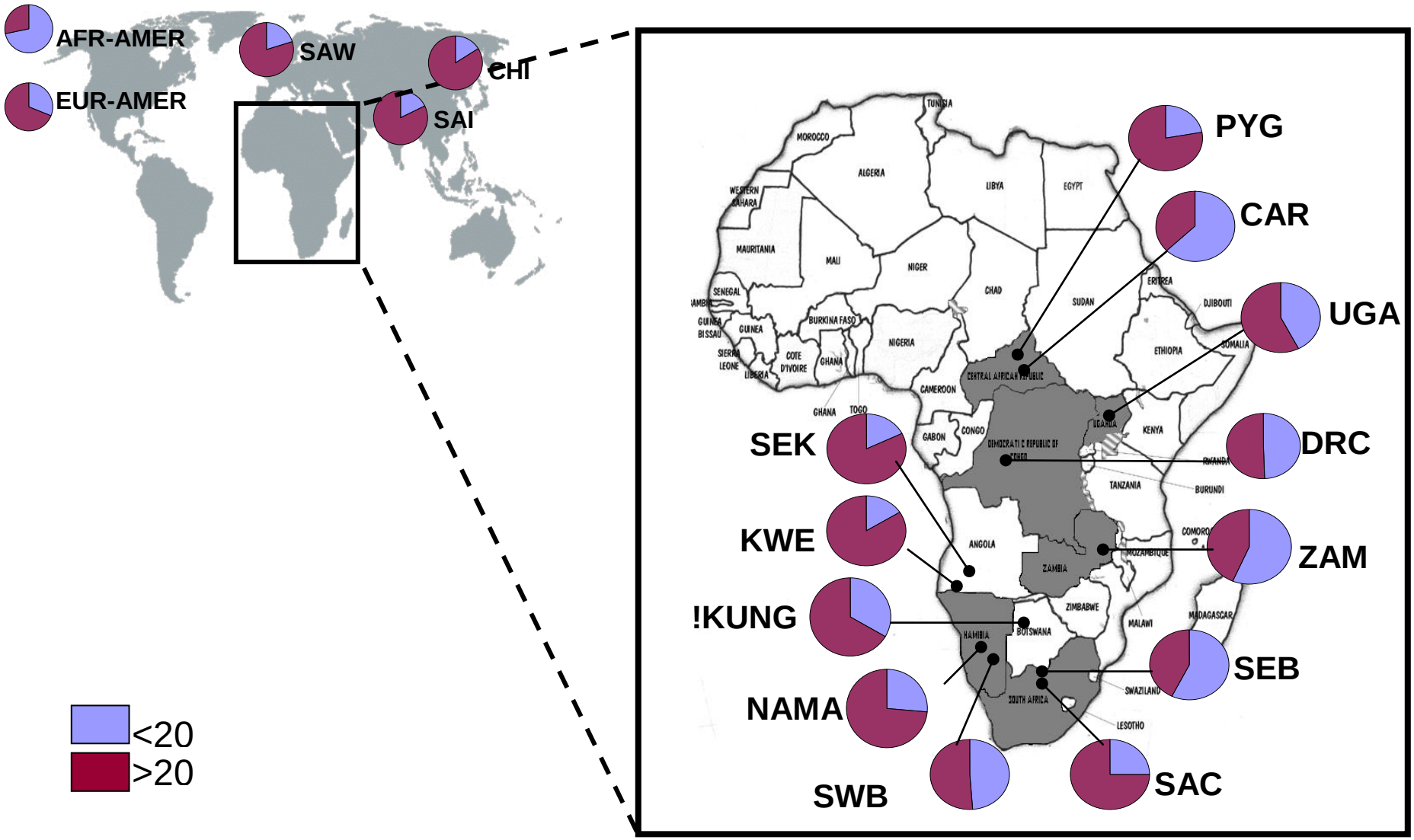


Figure 3.9 The distribution of AR alleles (<20 and >20 repeats) in the populations examined in the current study as well as comparative data (see Table 2.1)

3.3 POPULATION STATISTICS

3.3.1 Hardy Weinberg Equilibrium (HWE)

All the populations were tested to determine whether they deviated from the HWE for any of the loci tested. The CYP17A1 and the A49T loci were the only loci found to be in Hardy Weinberg Equilibrium for all the populations.

Seven out of the 70 tests performed for Hardy Weinberg Equilibrium showed significant departure from the Equilibrium state (Table 3.2). The SAW was the only population that deviated from the equilibrium state for the CYP3A4 locus showing a statistically significant P value of 0.00756. For the V89L locus, three Bantu-speaking populations, SEB, DRC and Uganda showed a statistically significant departure from Hardy Weinberg Equilibrium with P values of 0.00602, 0.00014 and 0.00213 respectively. Due to the fact that the A49T was highly monomorphic in most of the examined populations, the test could not be performed in those populations. Significant deviations from HWE were noted for the KLK3 locus in three African populations, SEB (P=0.03591), Nama (P=0.00592) and CAR (P=0.00491). However, since these departures were not restricted to specific populations or loci, they are probably the result of random statistical fluctuations.

Table 3.2 Hardy Weinberg Equilibrium proportions for the examined populations.

(Deviations from Hardy Weinberg Equilibrium are highlighted in bold)

		P-values at 95% confidence interval				
Populations	**N**	**CYP17A1**	**CYP3A4**	**V89L**	**A49T**	**KLK3**
African						
SEB	91	0.22638	0.31516	**0.00599**	-	**0.03591**
SWB	90	0.80058	0.56339	1.00000	-	0.82727
ZAM	55	1.00000	0.26192	0.76233	-	0.37786
DRC	117	0.53529	1.00000	**0.00002**	-	0.26235
UGA	120	0.14922	0.83440	**0.00118**	-	0.36585
CAR	32	1.00000	0.42754	0.68363	-	**0.00491**
PYG	27	0.69559	0.36572	0.07371	-	0.43219
NAMA	19	0.63185	0.60713	1.00000	1.00000	**0.00592**
!KUNG	27	0.42617	1.00000	0.11987	-	1.00000
SEK	56	0.78651	0.30178	1.00000	-	0.06403
KWE	31	0.48507	0.64409	0.63796	-	0.07259
SAC	56	0.77798	0.17483	0.74998	1.00000	0.40853
European						
SAW	49	0.75331	**0.00733**	1.00000	1.00000	0.12633
Asian						
SAI	45	0.09167	1.00000	0.06904	-	0.55532

3.3.2 INTRAPOPULATION DIVERSITY

The overall gene diversity for all the population was determined (shown in Table 3.3). The African populations had the highest overall diversity amongst the three continental groups, followed by the European and Asian populations with diversity indices of 0.3525, 0.3235 and 0.297278, respectively. Within the African group, the SAC had the highest overall gene diversity of 0.387291. The lowest diversity in Africa was however found in the Nama, Pygmies, Kung and Sekele (see Table 3.3).

Table 3.3. Gene diversity for the continental groups and each of the 14 populations

Population	*Genetic diversity*
AFRICAN	***0.352538***
SAC	0.387291
UGA	0.363975
KWE	0.355473
ZAM	0.347423
SEB	0.349827
CAR	0.337401
DRC	0.334089
SWB	0.334500
NAM	0.315505
PYG	0.334451
KUN	0.307757
SEK	0.307272
EUROPEAN	***0.323501***
SAW	0.323501
ASIAN	***0.297278***
SAI	0.297278

To ascertain whether African populations exhibited this high overall diversity across all the markers examined, the gene diversity per marker for each population (see Table 3.4) and per continental origin (see Figure 3.10) were determined. The highest diversity indices were found for the Androgen Receptor (CAG) repeats. However, when analysis was restricted to the SNP variants, the KLK3 gene variant exhibited the highest overall diversity whilst the A49T had the lowest (see Table 3.4).

For the *KLK3* locus, all the populations exhibited similar diversity indices, ranging from 0.4619 – 0.5016 (see Table 3.4), thus showing very little differentiation amongst the different populations. Even when the diversity was measured according to continental origin, very little variation between the continental groups was observed (see figure 3.10). Whilst the *KLK3* had the highest diversity indices, the A49T locus exhibited diversity values of 0 for all the populations with the exception of the Nama, SAC and SAW (see Table 3.4). Grouping of data according to continental descent showed that the European population had the highest gene diversity for this locus whilst the Asian population had 0% diversity for this marker (see figure 3.10).

For the *CYP3A4* locus, the lowest diversity indices were noted in populations of Asian and European descent whilst the SAC had the highest diversity for this locus (See Table 3.4). Although the Eurasian populations had the lowest diversity for the *CYP3A4*, the opposite was noted for the V89L gene variant, where they had the

highest indices. The three Khoisan-speaking populations, Nama, !Kung and the Sekele had the lowest indices for this locus whilst the Kwengo had diversity values that were similar to those of the Bantu-speaking populations.

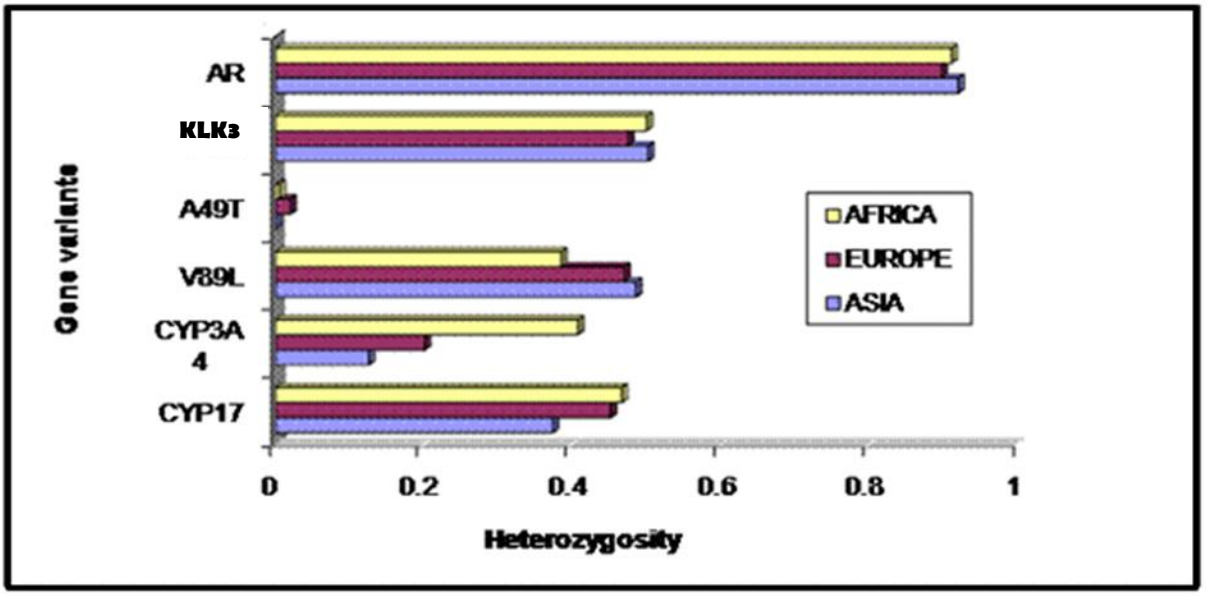


Figure 3.10. Genetic diversity indices of the three continental groups for all loci. The SAC group was included in the African sample

Table 3.4 Distribution of gene diversity per locus for each of the 14 populations.

		GENE-DIVERSITY ESTIMATES					
POPULATIONS	**N***	**CYP17A1**	**CYP3A4**	**V89L**	**A49T**	**KLK3**	**AR**
SEB	91	0.4366	0.4284	0.3852	0.0000	0.4989	0.8921
SWB	90	0.4223	0.3598	0.4035	0.0000	0.4869	0.9046
ZAM	55	0.4764	0.3545	0.4444	0.0000	0.4619	0.9172
DRC	117	0.4435	0.3275	0.4067	0.0000	0.4928	0.9078
UGA	120	0.4435	0.4346	0.4406	0.0000	0.5012	0.9056
CAR	32	0.3472	0.4479	0.4241	0.0000	0.4678	0.9173
PYG	27	0.4920	0.4249	0.2572	0.0000	0.4983	0.8575
NAM	19	0.4908	0.4225	0.1494	0.0526	0.4623	0.9532
!KUN	27	0.4920	0.2572	0.2830	0.0000	0.5066	0.8832
SEK	56	0.4884	0.3958	0.1491	0.0000	0.5031	0.8344
KWE	31	0.5077	0.3728	0.3892	0.0000	0.5077	0.8860
SAC	56	0.4773	0.4987	0.4194	0.0526	0.4884	0.9091
SAW	49	0.4513	0.2013	0.4696	0.0204	0.4749	0.8971
SAI	45	0.3735	0.1258	0.4854	0.0000	0.5016	0.9192
	815	**0.4788**	**0.4708**	**0.3799**	**0.0107**	**0.4992**	**0.9116**

3.3.3 INTERPOPULATION DIVERSITY

FST STATISTIC

To quantify the extent of between-population variation, among the populations, Wrights Fst Statistic was computed for different population groupings (see figure 3.11). The highest Fst value (0.14) was noted when the populations were grouped according to continental descent, African, European and Asian. The African populations exhibited higher between population variation (0.021) than that observed between the European and Asian populations (0.01). The lowest Fst indices were noted in the Bantu-speaking populations (0.006).

The SAC was grouped with its various parental populations. The highest between population variation was noted when the SAC was grouped with the Asian population. The lowest was noted when it was grouped with the SEB population.

The genetic variants were also assessed to determine which variant exhibit the highest level of between-population variation. The highest indices were noted for the CYP3A4 gene variant, whilst the lowest was for the AR gene variant (see figure 3.12).

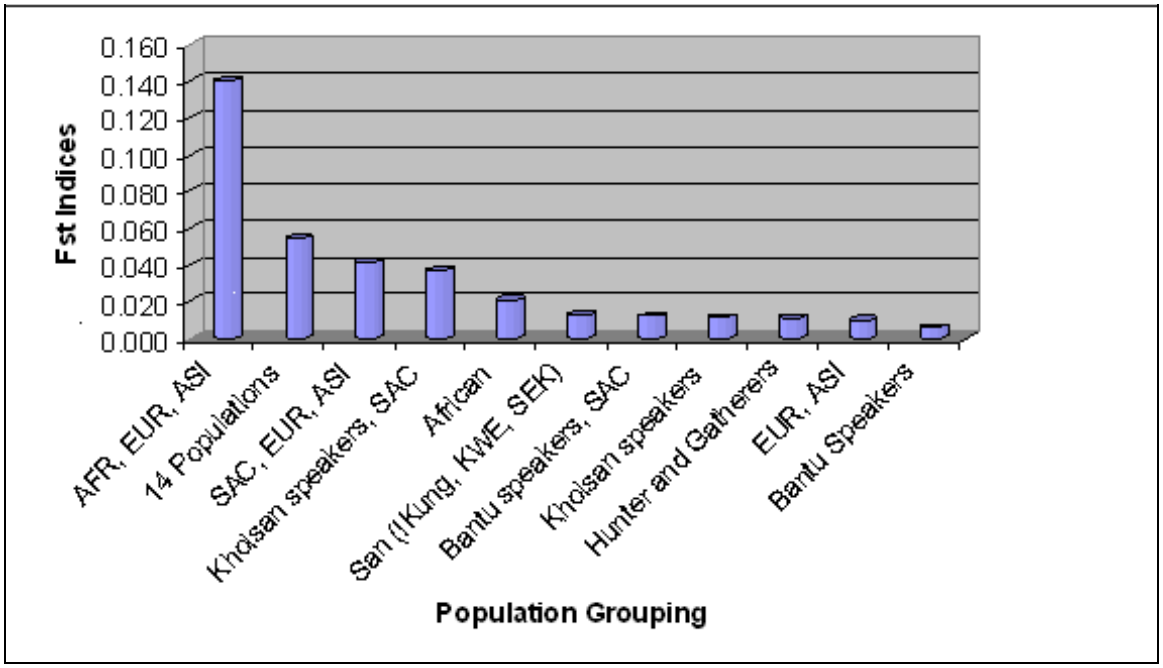


Figure 3.11 Fst estimates for different population groupings. African (AFR), Europe (EUR), Asian (ASI).

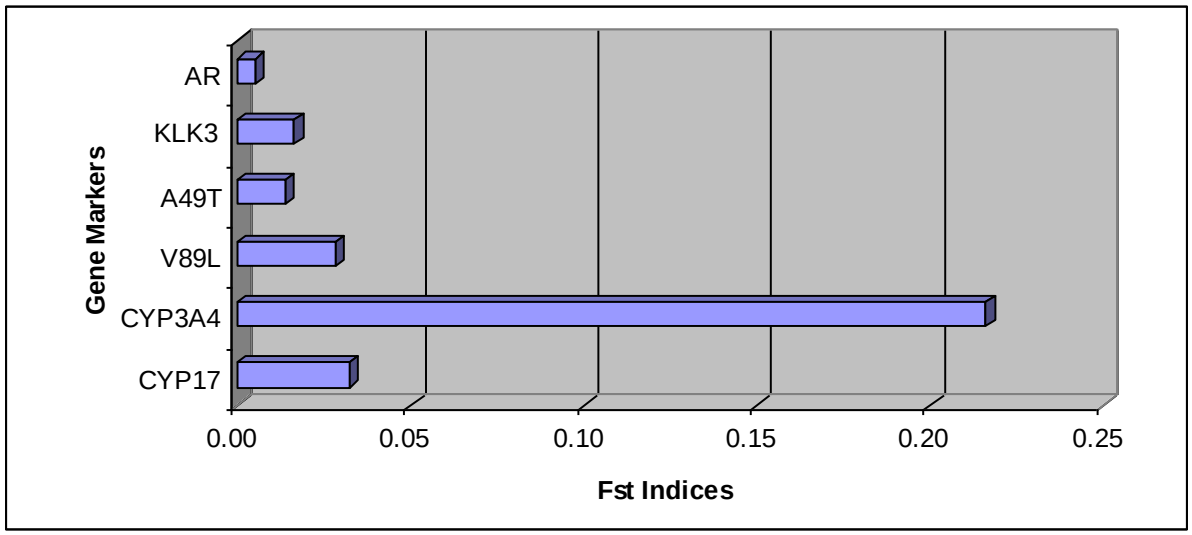


Figure 3.12 Fst estimates for different genetic variants

3.3.3 Analysis of gene flow

In order to elucidate the amount of gene flow experienced by the different 14 populations, a plot of heterozygosity vs. the distance from the centroid was drawn according to the Harpending-Ward model (Harpending and Ward 1982).

From this analysis, all the populations were found to exhibit higher genetic diversity than would be expected under this model with the exception of the Sekele. The Sekele population was the only population that exhibited less heterozygosity than would be expected under this model. The SAC was found to be one of the populations that were found to be the furthest away from the theoretical regression line, thus indicating more gene flow into the SAC population (as shown in figure 3.13).

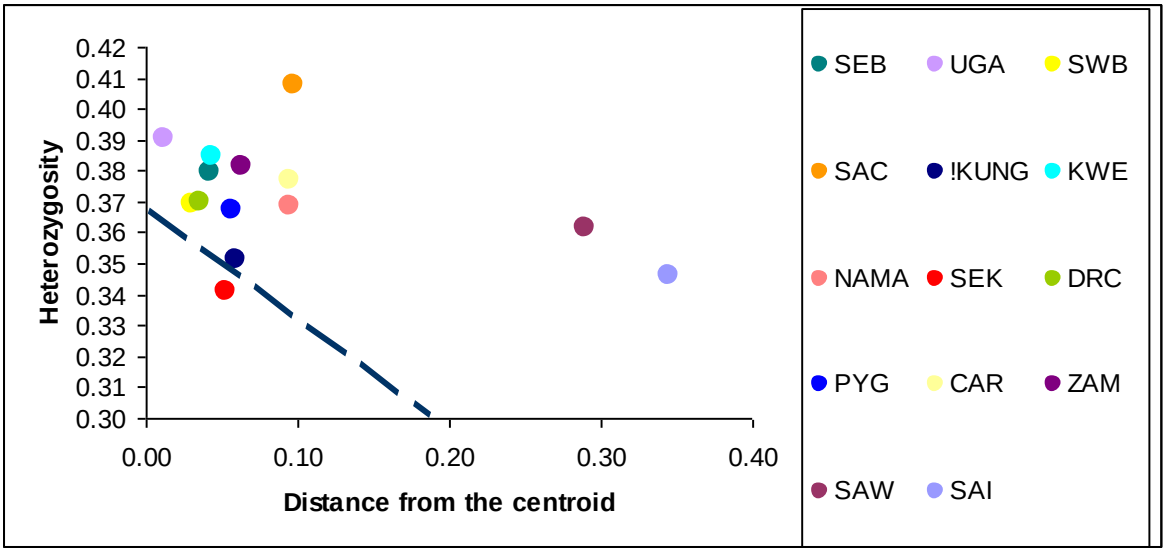


Figure 3.13 The plot of heterozygosity vs. distance from the centroid for the 14 populations.

3.4 POPULATION AFFINITIES

The genetic affinities and relationships of the examined populations were inferred using a neighbour joining tree (as shown in Figure 3.14). From this tree, three distinct clusters were noted. The first major separation is between African and non-African populations. The African populations clustered together on one branch whilst the non-African populations clustered on a separate branch. However, the SAC population was located in a branch that was between the African population and the non-African population.

The African branch was further divided into two separate clusters clearly differentiating the Bantu-speaking populations from the hunter and gatherer populations. The Khoisan-speaking populations, the Nama, !Kung, Sekele and the Kwengo formed a cluster with the Pygmies.

The Bantu–speakers branch was however not exclusively Bantu speaking as the African-Americans were found to cluster within this branch. On the contrary, the European-Americans were however found to cluster with the SAW and SAI populations on the non-African branch.

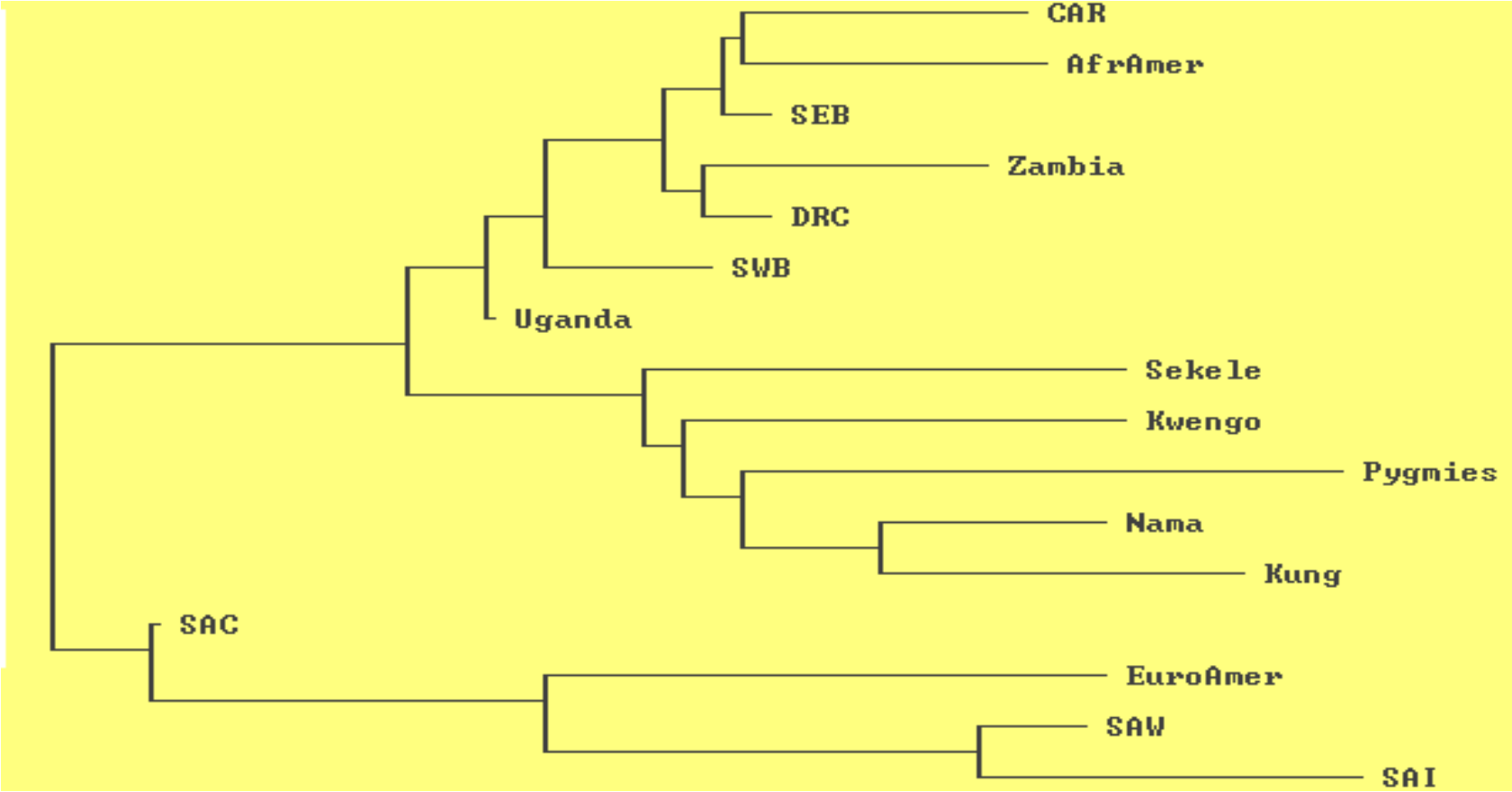


Fig 3.14 A neighbour-joining tree showing the population affinities of the examined populations

4 DISCUSSION

Prostate cancer is notorious for the extreme racial disparities observed in its incidence rate. It has been noted that individuals of African descent, especially those of the African Diaspora, have a higher risk of developing prostate cancer than individuals of either European or Asian descent (Reddy *et al.* 2003). This thus implies that some form of underlying genetic structure, which is present in both the current African populations and the exiled populations might exist, which might be responsible for the ethnic differences in incidence rates.

Several polymorphisms in the androgen biosynthesis and metabolism cascade have been identified that are hypothesised to increase the risk to prostate cancer. These gene variants have, however, only been comprehensively studied in the European, Asian and African-American populations. This bias in sampling and the scarcity of data on the frequency distribution of these gene variants in sub-Saharan African prompted the initiation of the current study. This study is the first attempt to comprehensively explore the distribution of these variants in sub-Saharan African populations and include hunter and gatherer populations in the dataset.

Furthermore, this study allows for the comprehensive analysis of genetic variation among sub-Saharan African populations, which still remain largely

understudied despite the important historical implication these populations have with regards to human evolution.

4.1 TRENDS AND PATTERNS FROM FREQUENCY DISTRIBUTIONS

Analysis of the variant alleles on an individual level, led to the identification of certain trends and patterns from the frequency distribution of these alleles. These trends and patterns were found to be in accordance with the known history of some of these populations. Thus, emphasising the critical role that population genetics should play when association studies and remedial interventions strategies are designed.

From these trends and patterns, four inferences were made. These include:

1. Historic implication of the frequency distribution of the variant alleles
2. Origins of the variant alleles
3. Association of the variants with prostate cancer
4. Interaction between genetic and environmental factors

It is however important to note that only six markers were examined in the current study and thus the inferences that were made from the data are limited.

4.1.1 Frequency distribution and Intrapopulation diversity

CYTOCHROME P450C17A (*CYP17A1*)

There was a high difference in the frequency distribution of the variant allele between the CAR and Pygmies reaching statistically significant P value of <0.0001. This in light of the close living proximities of the two populations, as well as the long history of contact between these two populations is quite surprising. Although, the Pygmies are primarily hunters and gatherers, some maintain a close relationship with the neighbouring farming populations. They not only exchange their game and labour for tobacco and alcoholic beverages, but have also adopted the language of these neighbouring populations (Cavalli-Sforza *et al.* 1994). As a result of these relations, there has been a substantial amount of gene flow from the Pygmies into their neighbouring Bantu-speaking populations and vice versa. Thus when the extent of the genetic contributions made by Bantu-speaking males into the Pygmy population was calculated, it was found to be more than 50% (Cruciani *et al.* 2002).

The distribution patterns of the CYP17A1 variant allele, however does not illustrate this extensive contact and significant gene exchange between these two populations. It seems that the Pygmies have retained their ancestral frequencies of the variant allele. This is also illustrated by frequency distribution patterns and

gene diversity indices that are similar to those of groups that were priorily known as hunter and gatherer populations (Khoisan-speaking populations).

THE CYTOCHROME P450 3A4 (*CYP3A4)*

Extreme ethnic variation in the frequency distribution of the CYP3A4 (G) allele was observed in the dataset. This allele was observed in high frequencies in African populations but was found in frequencies of less than 12% in non-African populations (European and Asian descent). Furthermore, the highest between-population diversity in this data was observed for this marker. Thus, this variant in conjunction with other ancestry informative markers would be very useful in making inferences on the geographic origins of individuals. This continentally determined distribution was also noted in the presence of comparative data, further accentuating the African specificity of this marker.

The SAC had intermediary frequencies between the African populations and European populations as well as the highest gene diversity for this polymorphism. These are signatures that indicate a state of admixture. The precise biological history of the South African Coloured population is not clear, as there was no single source but rather through a number of successive events. However, the emergence of the SAC population can be traced back to the shores of Table Bay a few years after the Dutch settlement. This population formed as a result of relations between the Dutch Settlers and the Khoisan women. Later,

various genetic contributions from the Bantu-speaking, Asian and other European populations (Nurse *et al.* 1985) entered the SAC genetic pool. Thus the SAC population is a highly admixed population having received contributions from multiple parental populations. Therefore the presences of intermediary allele frequencies between the parental populations as well as the high gene diversity and gene flow are testimony to this state of admixture.

STEROID 5α REDUCTASE TYPE II (*SRD5A2*) V89L

The frequency distribution of the leucine variant increased with a pattern of increasing distance from Africa. As a result, populations of Asian origin harboured the highest frequencies for this allele. Thus, signifying that this allele is likely to have gained importance with the Out of Africa migration.

The lowest gene diversity was observed in the hunter and gatherer populations with the exception of the Kwengo. The low diversity indices in these populations could be a consequence of the demographic histories of these populations. The hunter and gatherer populations maintain small population size and live in isolation, making the effects of genetic drift more pronounced in these populations (Oota *et al.* 2005). The Khoisan-speaking populations (the Nama and the !Kung), have in the recent past undergone two major successive bottleneck events. The wars with the settlers over Khoisan speakers' cattle stealing tendencies and the small pox epidemic greatly reduced the effective population

size of these populations in Southern Africa. One of the great consequences of major bottleneck events is the reduction in diversity of the affected populations.

The Kwengo population on the other hand, was found to statistically differ to the other Hunter and gatherer populations. However, when it was compared to the Bantu-speaking populations, no significant differences were noted. The Kwengo have often been called the Black Bushmen, due to their striking similarities in physical appearance to the Bantu-speaking populations. It was postulated that the Kwengo were initially Bantu speaking populations that adopted the Khoisan language. Studies based on mitochondrial DNA have shown that the Kwengo are genetically more similar to the Bantu-speaking populations than to the Khoisan-speaking populations (Chen *et al*. 2000).

STEROID 5α REDUCTASE TYPE II (*SRD5A2*) A49T

Of the loci examined, the A49T locus was the least variable and had frequencies of zero in 11 out of the 14 populations examined. The variant allele was absent in all the population in our current dataset with the exception of the Nama, SAC and SAW where it was found with frequencies of less than 3%. Even in the presence of comparative data, this allele is found in minute frequencies in world populations. However, the European populations seem to generally have higher frequencies of this polymorphism as compared to other populations.

The fact that this allele is found in such small frequencies across the globe strongly indicates that this allele is a newly arisen allele. Up to date, there has been no report of this allele in Asian populations.

KALLIKREIN RELATED PEPTIDASE 3 GENE (*KLK3*)

The KLK3 locus exhibited the highest diversity values in all the SNP variants that were examined. This suggests that this polymorphism might be the oldest of the SNP variants that was examined, as there is a direct correlation between gene diversity and the age of the mutation (Jorde *et al.* 2001). Also there was no evident clustering pattern of frequency distribution amongst the examined population. This observation suggests that this locus is highly unlikely to have been under any selective pressures or gained importance during the Out of Africa migrations

The SAI were noted to have frequencies that were significantly different to those of other Asian populations. This pattern was also noted in both the *CYP17A1, V89L* and *AR* variants. It was unfortunate that in the case of SAI, we were restricted in comparative analysis to populations that did not directly contribute to the gene pool of SAI. This could be the reason for the significant differences that were noted between the SAI and populations of Asian descent.

ANDROGEN RECEPTOR (*AR)*

The highest gene diversity of all the examined loci was observed in the AR locus. This high diversity is attributable to the fact the STR markers have a high mutation rate and are therefore extremely fast evolving (Zhivotovsky *et al.* 2003). Thus making inferences on relatively recent population divergence possible (Romualdi *et al.* 2002). This is evident from the ability of this marker to decipher variation patterns within Africa.

There was a great distinction in the distribution of repeat units of less than 20 between the Hunter and gatherer populations and the Bantu speaking populations. Furthermore all the Bantu-speaking populations with the exception of Uganda had modal repeat units that were less than 20 whilst the hunter and gatherer populations had modal repeat units that were either above or equal to 20.

The close genetic similarities of Bantu speaking populations have been previously documented. The Bantu speaking populations currently occupy more than one third of the African continent and had come to occupy this vast land through one of the most significant population expansions in Africa, the Bantu expansion. This expansion was due to the introduction of plants from South East Asia and was the resultant migration in search for more land to accommodate the increase in population size (Murdock *et al.* 1959). It took place approximately

5000 years ago and led to the spread of agriculture to the southern parts of the continent (Pereira *et al.* 2001), which were originally inhabited by the Khoisan speaking populations (Soodyall *et al.* 1992).

The Bantu expansion took place in two routes, the south-westerly and the south-easterly routes. The south-westerly route spread to the south along the shores of the Atlantic Ocean, whilst the south-easterly route passed through the rainforest and settled briefly in the Great lakes and later advanced to the southern tip of the continent. Both routes are postulated to have originated from the ancestral homeland that was in the border of Cameroon and Nigeria. It is however contemplated that intermingling between the two routes had occurred at different times during the expansion although the specifics are still unclear (Plaza *et al*. 2004).

This repeat unit, in combination with other ancestry informative markers would be useful in the elucidation of variation patterns within Africa especially because of the differences in frequency distribution between Bantu speaking and hunter and gatherer populations. The presence of repeat units of less than 20 appears to be highly prevalent in Bantu speaking population and in conjuction with other markers, can thus be used to trace and form conclusive results on the spread of Bantu speaking populations across the African continent.

4.1.2 ORIGINS OF THE VARIANT ALLELES

The presence of four gene variants (CYP17A1, V89L, KLK3 and AR) in frequencies of more than 10% across the global populations suggests that these variants predate the human divergence period. The distribution of these four gene variants in the African as well as non African populations indicates that these polymorphisms were present in the ancestral African populations and are therefore present in the non African populations due to the contribution from the ancestral population.

The A49T polymorphism was however not observed in African and Asian populations with the exception of one Nama and three SAC individuals. Higher frequencies of this polymorphism, although in small frequencies, are observed in populations of European descent as compared those of either African or Asian descent, thus suggesting that the roots of this polymorphism are highly likely to be traced back to Europe. However, this observation does not explain the presence of this variant in the two African populations. When the Y chromosome data of these African individuals carrying the variant allele was ascertained, all four were found to belong to the M207 haplogroup, a haplogroup that is found predominantly in the European populations and also known as a European specific Y chromosome marker. Due to the known history of contact between the SAC and Nama with European populations, the presence of this polymorphism in the African populations is likely to be as a result of this extensive contact.

Furthermore, the low frequency with which this variant is found even in the European populations in which it originated from indicates that it is newly arisen polymorphism. This variant allele must have already been present in the European populations prior to the year 1652 when the Dutch settler arrived in South Africa and were thus able to introduce it to the Khoisan speaking populations and in the newly formed SAC population.

There is very evident frequency differences in the distribution of the CYP3A4 (G) allele between African and non-African populations. The G allele is found in frequencies ranging from 67% to 90% in African populations whilst the A allele is found in frequencies of 89 to 93%. This then begs the question of which of these two alleles is the ancestral one. Since the African populations are regarded as the ancestral population that gave rise to all other existing populations, there is a high probability that the G allele might trace its origins back to Africa and be the ancestral allele. But due to the ascertainment bias of this allele in European populations, the CYP3A4 (A) allele was thought to be the ancestral allele (Rebbeck *et al.* 1998). Furthermore, the presence of the A allele in frequencies of more than 10% in the ancestral African populations strongly suggests that this allele arose before the Out of Africa expansion. Selective pressures coupled with the bottleneck events and rapid expansion of the European and Asian populations with the Out of Africa expansion must have led to the high frequencies with which it is seen in these populations. However, comparing the

human CYP3A4 sequence to that of the Pan Tryglodes might help to form conclusive results on which of these allele is the ancestral one.

Many might argue that we should move towards classifying DNA changes by state rather than calling them ancestral or derived, particularly since both these alleles are found in frequencies of more than 80% in the African and Eurasian populations. Whilst this argument is true and valid, it is critical for geneticists to understand which alleles are ancestral, as this allows them to fully understand the selective factors that drove the variant allele to such high proportions in either the African or Eurasian populations. This will consequently add to the growing body of knowledge on the drivers of human genomic variation.

4.1.3 ASSOCIATION OF VARIANTS WITH PROSTATE CANCER

Four gene variants out of the six that were studied in the current dataset showed frequency distribution patterns that greatly mimicked prostate cancer incidence. These include the *CYP17A1*, *CYP3A4*, V89L and AR gene variants. Populations of African descent, particularly African-Americans, have a extremely high incidence of prostate cancer whilst the Asian populations have the lowest reported incidence, which is 50 fold lower than the high risk African American populations (Keita *et al.* 2004; Hsing *et al.* 2001).

When analysis was restricted to the current data set, the *CYP17A1* (C) allele that is said to increase risk to prostate cancer was found in higher frequencies in the high risk African population as compared to the SAI, which are considered the low risk group. The SAW population was however noted to exhibit intermediary frequencies. However, when comparative data was added, the reverse was seen. The Asian populations (Japanese and Taiwanese) populations had the highest frequencies of the variant allele, followed by the African populations. It is highly probable that the small sample size of SAI population (n=45) in comparison to the 538 Asian individuals in comparative data, is the reason for this deviation. A larger sample size of the SAI might show a different picture, one that is in concordance with other Asian populations.

The *CYP3A4* showed the most distinctive pattern that greatly resembled that of prostate cancer incidence. Extremely high frequencies of the G allele, which is said to increase prostate cancer risk, were found in African populations. Frequencies of below 10% were found in Asian and European populations. The Asian populations had the lowest frequencies for this allele.

The leucine allele is said to confer a protective effect against prostate cancer to those carrying this allele. This allele was found in high frequencies in Asian populations, followed by European populations. The African populations as well as the African- American population had the least frequency of the protective leucine allele.

The effect of short repeat units (<20) has long been researched as a potential risk factor for prostate cancer. These short repeat units were found to be extremely common in Bantu speaking populations as well as the extremely high risk African-Americans. The short repeat units were however found in low frequencies in the both the Asian and European populations. This frequency distribution pattern further increases the likelihood of this polymorphisms being associated with prostate cancer.

The distribution of *KLK3* and A49T polymorphisms did not however illustrate a pattern that coincided with prostate cancer incidence. It is highly probable that the KLK3 acts in conjunction with other markers, the most likely candidate being the AR (CAG) repeat unit, to exert an effect. The combination of the KLK3 (G) allele and the short AR repeat units (<20) have been previously shown to increase risk to prostate cancer (Xue *et al.* 2000).

The A49T gene variant, because of its likely origins in Europe and the low frequencies with which it is found in Africa, is highly unlikely to greatly increase risk to prostate cancer in African populations. This polymorphism might however be a genetic predisposing factor in European populations and populations with European contributions. The distribution of this variant allele highlights some of the challenges faced by disease studies on the risk classification of subjects in to groups that may not be very robust. Due to the high level of admixture in the SAC

and Nama populations as well as the low frequencies with which this allele is found in these populations, spurious associations for this allele might be noted. This emphasises the importance of understanding the genetic structure and history of populations prior to conducting such studies.

4.1.4 INTERACTION OF GENETIC AND ENVIROMENTAL FACTORS

Most complex diseases, like prostate cancer, manifest as a result of complex interactions between environmental and genetic predisposing factors. These two factors cannot act solely to bring about the disease state (Wright *et al.* 1999). Therefore in order to develop the disease state, a combination of the genetic predisposing makeup coupled with exposure to the environmental triggers is required.

Due to the unique history of African-Americans, the genetic affinities of this population were inferred using a neighbour joining tree. From this, it was observed that African-Americans and African populations have similar genetic makeup for the predisposing gene variants. This observation is consistent with their known history of a sub-Saharan African origin. The African-Americans are descendants of African slaves that were taken from sub-Saharan Africa to work at sugar plantations in various American colonies. From 1518 up until 1870, when the practise of slavery was abolished, a total of 10 million slaves had been sent to the Americas. The great source of slaves during this period was the West and Central African states. However by the late 18th century, the source had

extended to regions in Southern and East Africa. Additional genetic contributions from European American and American Indians entered the genetic pool of the African-Americans with the settlement in the American colonies, thus making this population highly admixed (www.history.com).

However, despite these genetic similarities, there is an evident difference in the incidence rate of prostate cancer in the continents where these two populations currently reside. American populations have an incidence rate that is almost three times as high as the incidence rate observed within African continent (Parkin *et al.* 2005). Since both these populations have the same genetic makeup, the only varying factor that could possibly explain the discrepancies in the incidence rate is the exposure to the environmental stimuli.

One of the environmental factors that have long been researched for its contribution to prostate cancer risk is the intake of a diet high in fat content. It has been noted that the incidence of prostate cancer increases when individuals from the low risk population groups move to the high risk population regions for example the Japanese living in Brazil have a higher incidence of prostate cancer than the Japanese living in Japan (Iwasaki *et al.* 2004).

Furthermore, in other cancers particularly breast cancer where a diet high in fat intake is also considered a risk factor, although inconclusive results have been noted in this regard (Fung et al.), the highest incidence rates are observed in

North America (Parkin *et al.* 2005). These observations accentuate the fact that a diet high in fat intake is very likely to represent one of several candidates that increase risk to prostate cancer and might be the one of the reasons for the differences in incidence rate between African-American and their founding African populations. As African populations adopt a more westernised diet, an increase in the prostate cancer incidence might be seen. It is however important to note that there are other potential environmental factors, such as lifestyle and smoking, that are likely to increase risk to prostate cancer. The current study supports the view on the requirement of both the genetic and environmental factors concurrently in order to bring about the disease state. It also further emphasises the point made by Tishkoff and Williams 2002, which stressed how studying genetic variation in Africa might consequently lead to the discrimination of environmental factors from genetic contributions in complex diseases (Tishkoff and Williams 2002).

4.2 GENETIC VARIATION AMONG AFRICAN POPULATIONS

The analysis of genetic variation and structure in African and non-African populations using all the six polymorphisms highlights and cements genetic findings that have been previously and extensively documented even though only a limited number of markers were used.

1. Evolution of modern human
2. Prehistory of the Khoisan speaking populations
3. The genetic structure of the SAC
4. The Bantu Expansion

4.2.1 Evolution of modern humans

The current data supports the theory that has been previously postulated on the evolution of modern humans within Africa. A number of key findings from the current study show support for the “Out of Africa” theory and the rapid expansion of the Eurasian populations.

Firstly, African populations had the highest overall diversity for the markers examined, despite the fact that most of these markers had been ascertained in European, Asian and American populations. Also, a trend of decreasing diversity with increasing distance from Africa was noted, with Asian populations having the lowest diversity indices. Higher diversity among African population as compared to non-African populations has been previously documented for various genetic

systems, Y chromosome, Mitochondrial, Autosomal SNPs and STRs as well as for Alu polymorphisms (Watkins *et al.* 2001)

Secondly, the highest measure of between-population variation (Fst) was noted when the populations were grouped according to their continental origin. The removal of African populations reduced this statistic from 14% to 1%, illustrating that African populations contribute the majority of intercontinental variation and that more variation occurs within African populations than between the two non-African (European and Asian) populations.

Lastly, the clustering patterns of the neighbour-joining tree are in concordance with those observed with Y chromosome and mtDNA data (Jobling and Tyler-Smith 1995, Jorde *et al.* 1998 and Zhitovosky *et al.* 2003). The first distinct separation is between the African and non-African populations, signifying one of the prominent characteristics of the Out of Africa migration.

All the above-mentioned observations are aligned with the origin of modern humans from Africa. The high genetic diversity observed in Africa might be indicative of the fact that the African populations had more time to diversify as they were the founding populations. Furthermore, the reduction of diversity in non-African populations could be the consequence of the exodus of a limited subset of variation from Africa, which was further reduced by the series of bottleneck events followed by rapid expansion.

4.2.2 Prehistory of Khoisan-speaking populations

The term Khoisan was invented by Schulze to collectively group the Khoi-Khoi populations with the San populations. However, there has been ongoing debate on the collective grouping of the Khoisan-speaking populations due the cultural and linguistic differences observed between these populations, which might be indicative of separate origins for these populations (Soodyall *et al*. 1992).

When the genetic affinities of the Khoisan speaking population were assessed using a neighbour joining tree, all the Khoisan speaking populations were found to form a cluster within the African branch. There were no distinctions in the clustering of the the Khoi-Khoi (Nama) and the San populations (!Kung, Sekele and Kwengo) that would suggest separate origins of these two populations. Infact the Khoi-Khoi population was found to cluster in a branch with two of the San population (Kwengo and !Kung) further showing lack of support for the separate origins of these populations.

Moreover, when the between-population variation of the San populations was assessed, an Fst value of 1.3% was noted. This value was reduced to 1.1% when the Khoi-Khoi population was included in the analysis. This further emphasises the lack of population differentiation between the Khoi-Khoi and the San populations.

Even though the separate origins of the Khoi-Khoi (Nama) and San (Sekele, Kwengo and! kung) populations have been previously postulated, the current data does not support this view and suggest that these two populations are mostly likely to share a common origin and ancestry. Despite the fact that the Kwengo have been previously suggested to have a Bantu-speaking origin and mitochondrial DNA studies have shown the Sekele to be genetically similar to the Bantu speaking populations (Soodyall *et al*. 1992), these observations were not noted in the current dataset when all the markers were assessed collectively. These populations do not exhibit large genetic differences that would render them as populations of separate origins.

4.2.3 The genetic structure of the SAC

Consistent with the observation noted for the CYP3A4 genetic variants, the analysis of the genetic structure of the SAC, yielded results that were in accordance with the known history of the South African Coloured.

Of the populations examined, the SAC had the highest overall genetic diversity. Furthermore when the analysis of gene flow was performed, the SAC were found to be the furthest away from the theoretical regression line, indicating more gene flow has occurred into this population. These observations are consistent with the contribution of multiple parental populations into the SAC genetic pool.

However, when the SAC was grouped with the different parental populations, the lowest Fst value was observed when it was grouped with the Bantu-speaking populations whilst the highest value was noted when grouped with the Eurasian populations. This observation suggests that the sample of SAC that were used in the current study is genetically more similar to the Bantu-speaking populations as opposed to the other SAC parental populations.

The genetic structure of the SAC poses a variety of challenges for disease studies using these gene variants in the SAC population. Prostate cancer association studies have been plagued with inconsistencies on the link between these gene variants and prostate cancer susceptibility, particularly in the African-

American populations. Population structure has often been cited as the reason behind the lack of replicability observed in association studies. Many maintain that populations that are extremely heterogeneous and exhibit large genetic diversities are highly likely to be affected by population stratification (Kittles *et al.* 2002). Due to the highly admixed nature of the SAC, spurious associations in disease studies might be noted, particularly due to the large variation in the frequency distribution of these gene variants in the parental populations of the SAC. Thus the underlying genetic structure of such a population is an important consideration to be taken into account when designing association studies using these gene variants.

4.2.4 The genetic structure of the Bantu-speaking populations

The patterns of genetic variation that were observed in the AR repeat unit for the Bantu-speaking populations were also noted when analysis was done for all the gene variants. The results of the overall analysis further highlighted the relatively recent divergence of the Bantu speaking population from the same ancestral population on the border of Cameroon and Nigeria. The recent divergence of these populations was not only emphasised by their clustering pattern on the neighbour joining tree but also by their levels of between-population variation.

When the population affinities of the Bantu-speaking populations were assessed using the neighbour joining tree, these populations were found to cluster on the African branch, which was then further divided into the Bantu-speaking and the

Hunter and gatherer branches. Also, the clustering pattern in the Bantu-speaking branch did not coincide with the two different paths (south-easterly and south-westerly) that were taken by these populations from their ancestral homeland to the southern tip of the African continent. Furthermore, the lowest between-population variance (Fst) for the different groupings was observed among the Bantu-speaking populations.

These observations provide further testimony to the rapid expansion and the relatively recent dispersion of these populations within the southern parts of the African continent and support the linguistic evidence on this expansion (Murdock *et al.* 1959 and Plaza *et al*. 2004).

5 CONCLUSION

This study is the first to comprehensively examine the frequency distribution of the gene variants in the androgen biosynthesis and metabolism cascade in sub-Saharan Africa and include the hunter and gatherer populations in the sample. The frequency distribution patterns of these genetic variants in the current study are in accordance with the other published data and highlighted key facts.

- The distribution patterns of these genetic variants are aligned to the demographic histories of these populations. Thus signalling the importance of fully understanding the history of these genetic variants as well as that of the populations where they will be examined.
- Also, the distribution patterns of the genetic variants amongst the populations showed great alignment to prostate cancer susceptibility. Thereby providing support to the likelihood that these variants increase risk to prostate cancer susceptibility.
- Furthermore, the current study highlights the importance of the link between both the genetic and environmental factors to disease susceptibility, particularly prostate cancer. The African-American population, although they share great genetic similarity for these variants to their founding African populations, have a greater incidence for prostate cancer.
- This study also provides further testimony to the “Out of Africa” theory and other important historical events in Africa.

5.1 FUTURE STUDIES

Future studies should focus on the link between these gene variants and prostate cancer risk. In order to get more comprehensive results, these markers should not be examined in isolation, as they are most likely to be low penetrance genes and exert small effects on their own, an effect that might go unnoticeable in association studies. It is however imperative that knowledge of genetic structure be discerned before performing large-scale association studies especially in admixed populations. The following are great considerations for future studies.

5.1.1 POPULATION GENETICS

The knowledge of population history is essential when designing association studies. The frequency distribution of alleles and genomic diversity observed within and between populations is heavily influence by the history of a particular populations. Since association studies are based on the premise that differences in allele frequencies between groups (cases and control) is due to association with the disease, knowledge and analysis of factors that can potentially skew the results is of paramount importance. Knowledge of population history is one of the elements that are critical in the design of these studies. The genetic structure in the SAC and African-American populations needs to be comprehensively examined when designing association studies using these gene variants;

especially since these gene variants are distributed differently in the parental populations that gave rise to these two populations.

5.1.2 STUDIES IN AFRICA

As previously stated by Tishkoff and Williams 2002, this study illustrates that studying African populations will allow for the discrimination of environmental contributions from genetic ones. This was highlighted by the fact that even though African-American populations have similar genetic makeup for the gene susceptibility variants to the sub-Saharan African population, differences in incidence rate are still evident. This highlights the role of environmental factors in increasing risk to prostate cancer such as a high fat diet, lifestyle and smoking.

Furthermore, this study has added to the growing body of knowledge on genetic variation in African populations and the evolution of modern humans. Analysis of genetic variation in the current study, showed the same patterns of African genetic variation that have been noted in other studies and further supported the “Out of Africa” theory. More studies on the analysis of genetic variation in African population need to be performed as this could help form a clearer picture on the evolution of modern humans.

5.1.3 Epidemiological studies

Another worthwhile experiment would be the elucidation of the relationship between a diet high in fat intake and prostate cancer incidence. Large-scale epidemiological studies would need to be performed to determine if there is a correlation between fat intake and the presence of genetic predisposing gene variants to developing prostate cancer.

6 APPENDICES

6.1 RECIPES AND SOLUTIONS

TE Buffer

Dissolve 1ml of Tris HCl(IM) in 200μl of EDTA (ph 8.0) . Add 100ml of dH2O. The ph of the solution should be 8.0

10X TBE

Dissolve 108 g of Tris base and 55 g boric acid in 800 ml of dH2O. Add 40 ml of 0.5 M EDTA (pH 8.0) and make up to 1 litre. The pH of this stock should be approximately 8.3

1X TBE

Mix 100 ml of 10X TBE stock and 900 ml of dH2O to make 1 litre of 1X TBE solution

3% AGAROSE GEL

Add 3g of agarose to 100ml of 1X TBE, and heat to dissolve. Allow to cool slightly and then add 3μl of ethidium bromide before setting

3% NUSIEVE GEL

Add 3g of Nusieve to 100ml of 1X TBE, and heat to dissolve. Allow to cool slightly and then add 3μl of ethidium bromide before setting.

2% NUSIEVE GEL

Add 2g of Nusieve to 100ml of 1X TBE, and heat to dissolve. Allow to cool slightly and then add 3µl of ethidium bromide before setting.

4.3% POLYACRYLAMIDE GEL

Dissolve 36g urea in 10.6ml of deionised 40% bis-acrylamide, 10 ml of 10X TBE and approximately 50 ml of ddH2O. Stir over heat till all contents are dissolved; then bring volume to 100ml. Filter through Nalgene filtration system under vacuum pressure. Store covered in foil at 4°C.

Sequencing gel preparation: Mix 50ml of 4.3% polyacrylamide gel with 30µl TEMED (as a catalyst) and 250µl 10% APS (to facilitate cross-binding) immediately before pouring.

BROMOPHENOL BLUE FICOLL DYE

Dissolve 50g sucrose, 1.86g EDTA, 0.1g bromophenol blue and 10g Ficoll in 50ml dH2O, stir overnight, pH to 8.0 and filter through Whatmann filter paper.

DEXTRAN/ FORMAMIDE DYE

Dissolve 20 mg Dextran (Sigma) in 1 ml Formamide

2.5mM DNTPs

Use 100mM premade stocks (GibcoBRL) of dATP, dGTP, dCTP and dTTP. Add 10µl of each stock dNTP to 360µl sterile ddH2O to make 400µl of 2.5mM dNTPs.

6.2 INTERPOPULATION COMPARISONS USING THE FISHERS EXACT TEST

Table A1: *CYP17A1* pair-wise population comparisons

	SWB	ZAMBIA	DRC	UGANDA	CAR	PYGMIES	KHOISAN	SAW	SAI	SAC
SEB	0.7337	0.3090	0.8336	0.8345	0.1519	**0.0004**	**0.0001**	0.7898	0.2579	0.2582
SWB		0.1592	0.5942	0.5963	0.2568	**0.0002**	**< 0.0001**	0.5885	0.3901	0.1602
ZAMBIA			0.3950	0.3362	**0.0294**	**0.0126**	**0.0410**	0.5635	**0.0476**	1.0000
DRC				1.0000	0.0946	**0.0005**	**0.0001**	0.8989	0.1779	0.3355
UGANDA					0.0951	**0.0005**	**0.0001**	0.8992	0.1430	0.3372
CAR						**< 0.0001**	**< 0.0001**	0.1148	0.8471	**0.0295**
PYGMIES							0.2346	**0.0034**	**< 0.0001**	**0.0130**
KHOISAN								**0.0063**	**< 0.0001**	**0.0425**
SAW									0.1997	0.5650,
SAI										**0.0484**
SAC										

Table A2: *CYP3A4* pair-wise population comparisons

	SWB	ZAMBIA	DRC	UGANDA	CAR	PYGMIES	KHOISAN	SAW	SAI	SAC
SEB	0.1246	0.1772	**0.0222**	0.9156	0.7564	1.0000	0.1287	**< 0.0001**	**< 0.0001**	**< 0.0001**
SWB		1.0000	0.5482	0.0631	0.1390	0.3712	0.9099	**< 0.0001**	**< 0.0001**	**< 0.0001**
ZAMBIA			0.6724	0.0989	0.1575	0.3441	0.8940	**< 0.0001**	**< 0.0001**	**< 0.0001**
DRC				**0.0048**	**0.0453**	0.1503	0.3900	**< 0.0001**	**< 0.0001**	**< 0.0001**
UGANDA					1.0000	0.8715	0.0593	**< 0.0001**	**< 0.0001**	**< 0.0001**
CAR						0.8424	0.1550	**< 0.0001**	**< 0.0001**	**0.0005**
PYGMIES							0.3924	**< 0.0001**	**< 0.0001**	**0.0002**
KHOISAN								**< 0.0001**	**< 0.0001**	**< 0.0001**
SAW									0.3173	**< 0.0001**
SAI										**< 0.0001**

Table A3: *SRD5A2* (V89L) population pair-wise comparisons

	SWB	ZAM	DRC	UGA	CAR	PYG	KHOI	SAW	SAI	SAC
SEB	0.7223	0.2290	0.6569	0.1617	0.6230	0.1020	**0.0020**	0.0741	**0.0247**	0.5029
SWB		0.4268	1.0000	0.3354	0.7499	0.0713	**0.0004**	0.1362	0.0523	0.7903
ZAMBIA			0.4477	1.0000	0.7369	**0.0154**	**< 0.0001**	0.5623	0.3032	0.6642
DRC				0.5392	0.8763	**0.0381**	**< 0.0001**	0.2270	0.1003	1.0000
UGANDA					0.7609	**0.0046**	**< 0.0001**	0.6567	0.3782	0.6467
CAR						**0.0171**	**0.0048**	0.3688	0.1819	1.0000
PYGMIES							0.8317	**0.0006**	**0.0001**	**0.0258**
KHOISAN								**< 0.0001**	**< 0.0001**	**0.0007**
SAW									0.7714	0.2925,
SAI										0.1366

Table A4: *SRD5A2* (A49T) pair-wise population comparisons

	SWB	ZAMBIA	DRC	UGANDA	CAR	PYGMIES	KHOISAN	SAW	SAI	SAC
SEB							1.0000	0.3500		**0.0198**
SWB							1.0000	0.3525		**0.0209**
ZAMBIA							1.0000	0.4712		0.1217
DRC							1.0000	0.2952		**0.0106**
UGANDA							1.0000	0.2899		**0.0099**
CAR							1.0000	1.0000		0.2980
PYGMIES							1.0000	1.0000		0.3050
KHOISAN								0.4665	1.0000	**0.0285**
SAW									1.0000	1.0000
SAI										0.1302

Table A5: *KLK3* pair-wise population comparisons

	SWB	ZAM	DRC	UGA	CAR	PYGMIES	KHOISAN	SAW	SAI	SAC
SEB	**0.0119**	0.1113	0.6206	0.2020	0.1900	0.7564	0.0681	0.2545	0.1977	0.4697
SWB		**0.0002**	**0.0021**	0.1976	**0.0021**	**0.0425**	0.3822	**0.0010**	0.5155	**0.0038**
ZAMBIA			0.1962	**0.0040**	1.0000	0.3958	**0.0010**	0.7739	**0.0098**	0.4101
DRC				0.0541	0.3197	1.0000	**0.0122**	0.3941	0.0816	0.7288
UGANDA					0.0245	0.2301	0.5935	**0.0224**	0.7123	0.0668
CAR						0.5701	**0.0083**	1.0000	**0.0328**	0.5251
PYGMIES							0.1349	0.6050	0.2281	0.8679
KHOISAN								**0.0064**	1.0000	**0.0183**
SAW									**0.0279**	0.6723
SAI										0.0661

REFERENCE

1) Allen NE, Forrest MS, Key TJ (2001) The Association between Polymorphisms in the CYP17A1 and 5α Reductase (SRD5A2) Genes and Serum Androgen Concentration in Men. Cancer Epidemiology & Prevention 10: 185-189
2) Allen NE, Reichardt JKV, Nguyen H, Key TJ (2003) Association between two polymorphisms in the SRD5A2 Gene and Serum Androgen Concentrations in Brittish Men. Cancer Epidemiology, Biomarkers and Prevention 12: 578-581
3) Atmulller J, Palmer LT, Fischer G, Scherb H, Wjst M (2001) Genomewide Scans of Complex human diseases: True Linkage Is Hard to Find. American Journal of Human Genetics 69: 936-950
4) Bamshad MJ, Wooding S, Watkins WS, Ostler T, Batzer MA, Jorde LB (2003) Human population Genetic Structure and inference of group membership. American Journal of Human Genetics 72:578-589
5) Beilin J, Harewood L, Frydenberg M, Mameghan H, Martyres RF, Farish SJ, Yue C, Deam DR, Byron KA, Zajac JD (2001) A Case-Control Study of the Androgen Receptor Gene CAG Repeat Polymorphism in Australian Prostate Carcinoma Subjects. Cancer 92(4): 941-949
6) Cardon LR, Palmer LJ (2003) Population stratification and spurious allelic association. Lancet 361: 598-604
7) Carvalli-Sforza LL, Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. Nature genetics 33: 266-275
8) Carvalli-Sforza LL, Menozzi P, Piazza A (1994) The history and geography of human genes. Princeton university Press

9) Cavaco I, Gil JP, Gil-Berglund E, Ribeiro V (2003) CYP3A4 and MDR1 Alleles in a Portuguese Population. Clinical Chemical Laboratory Medicine 41(10): 1345-1350

10) Cavalli SA, Hirata MH, Hirata DC (2001) Detection of MboII Polymorphism at the 5' Promoter Region of CYP3A4. Clinical Chemistry 47(2): 348-351

11) Chamberlain NL, Driver ED, Miesfeld RL (1994) The length and location of CAG trinucleotide repeats in the androgen receptor N-terminal domain affect transactivation function. Nucleic Acid Research 22(15): 3181-3186

12) Chang BI, Zheng S, Issacs SD, Wiley KE, Carpten JD, Hawkins GA, Bleecker ER, Walsh PC, Trent JM, Meyers DA, Isaacs WB, Xu J (2001) Linkage and association of CYP17A1 gene in hereditary and sporadic prostate cancer. Int, J. Cancer 95: 354-359

13) Chen C, Lamharzi N, Weiss NS, Etzioni R, Dightman DA, Barnett M, DiTommaso D, Goodman G (2002) Androgen Receptor Polymorphism and the Incidence of Prostate Cancer. Cancer Epidemiology, Biomarkers and Prevention 11: 1033-1040

14) Coughlin SS, Hall IJ (2002) A review of genetic polymorphisms and Prostate Cancer Risk. Annals of Epidemiology 12(3): 182-196

15) Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coja V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA (2002) A back migration from Asia to sub-Saharan Africa is supported by high resolution analysis of human Y-

chromosome Haplotypes. American Journal of Human Genetics 70(5) 1197-214

16) Dispan (1993):Genetic distance and phylogenetic analysis proram, ver 1.1 Ota T and Pennyslvania State university, USA

17) Edwards KMJ, Paterson PJ, Hehir M, Underwood MA, Bartlett JMS (2002) The CAG trinucleotide repeat length in the Androgen Receptor does not predict the early onset of prostate Cancer. British Journal of Urology International 90: 573-578

18) Erichsen HC, Chanock SJ (2004) SNPs in cancer research and treatment. British Journal of Cancer 90: 747-751

19) Excoffier, Laval LG, Schneider S (2005) Arlequin ver. 3.00: An integrated software package for population genetics data analysis. Evolutionary Bioinformatics Online 1: 47-50

20) Febbo PG, Kantoff PW, Platz EA, Casey D, Batter S, Giovannucci E, Hennekens CH, Stampfer MJ (1999) The V89L Polymorphism in the 5α-Reductase Type 2 Gene and Risk of Prostate Cancer. Cancer Research 59: 5878-581

21) Fung TT, Hu FB, Holmes MD, Rosner BA, Hunter DJ, Colditz GA, Willet WC (2005) Dietary patterns and the risk of postmenopausal breast cancer. International journal of cancer 116(1):116-21

22) Giovannucci E, Stampfer MJ, Krithivas K, Brown M, Brufsky A, Talcott J, Hennekens CH, Kantoff PW (1997) The CAG repeat within the androgen

receptor gene and its relationship to prostate cancer. Proceedings of the National Academy of Science 94: 3320-3323

23) Giwercman YL, Abrahamsson P, Giwercman A, Gadaleanu V, Ahlgren G (2005) The 5α-Reductase Type II A49T and V89L High-Activity Allelic variants are More Common in Men with Prostate Cancer Compared with the general population. European Urology 48: 679-685

24) Gray IC, Campbell DA, Spurr NK (2000) Single nucleotide polymorphisms as tools in human genetics. Human Molecular Genetics 9(16): 2403-2408

25) Gsur A, Bernhofer G, Hinteregger S, Haidinger G, Schaltzl G, Madersbacher S, Marberger M, Vutuc C, Micksche M (2000) A Polymorphism in the CYP17A1 gene is associated with prostate cancer risk. International Journal of Cancer 87: 434-437

26) Gsur A, Feik E, Madersbacher S (2004) Genetic polymorphisms and prostate cancer risk. World Journal of Urology 21: 414-423

27) Gsur A, Preyer M, Haidinger G, Zidek T, Madersbacher S, Schatzl G, Marberger M, Vutus C, Micksche M (2002) Polymorphic CAG repeats in the Androgen Receptor gene, Kallikrein Related Peptidase 3 gene polymorphism and prostate cancer risk. Carcinogenesis 23(10): 1647- 1651

28) Habuchi T, Liqing Z, Suzuki T, Sasaki R, Tsuchiya N, Tachiki H, Shimoda N, Satoh S, Sato K, Kakehi Y, Kamoto T, Ogawa O, Kato T (2000) Increased risk of prostate cancer and benign hyperplasia associated with a CYP17A1 gene polymorphism with a gene dosage effect. Cancer Research 60: 5710-5713

29) Haiman CA, Stampfer MJ., Giovannucci E, Ma J, Decalo NE,. Kantoff PW, Hunter DJ. The Relationship between a Polymorphism in *CYP17A1* with Plasma Hormone Levels and Prostate Cancer. Cancer Epidemiology Biomarker & Prevention 10: 743-748

30) Harding RM, Mcvean G (2004) A structured ancestral population for the evolution of modern humans. Current Opinion in Genetics & Development 14: 667-674

31) Harpending H, Cochran G (2006) Genetic diversity and genetic burden in humans. Infect Genet Evol 6(2): 154 -162

32) Harpending HC, Ward RH (1982) Chemical systematics and human population Biochemical aspects of evolutionary biology. University of Chicago Press, Chicago: 213-256

33) Hsing A-W, Chen C, Chollakingam AP, Gao Y-T, Dightman DA, Nguyen HT, Deng J, Cheng J, Sesterhenn IA, Mostofi FK, Stanczyk FZ, Reichardt JKV (2001) Polymorphic markers in the SRD5A2 gene and prostate cancer risk: A population based case-control study. Cancer Epidemiology, biomarkers and Prevention 10:1077-1082

34) Hsing AW, Gao Y, Wu G, Wang X, Deng J, Chen Y, Sesterhenn IA, Mostoffi FK, Benichou J, Chang C (2000) Polymorphic CAG and GGN Repeat lengths in the Androgen Receptor Gene and Prostate Cancer Risk: a population-based Case-Control Study in China. Cancer Research 60: 5111-5116

35) Irvine RA, Yu MC, Ross RK, Coetzee GA (1995) The CAG and GGC microsatellites of the androgen receptor gene are in linkage disequilibrium in men with prostate cancer. Cancer Research 55: 1937-1940

36) Iwasaki M, Mameri CP, Hamada GS, Tsugane S (2004) Cancer mortality among Japanese immigrants and their descendants in the state of São Paulo, Brazil, 1999-2001. Japan Journal of Clinical Oncology 34(11):673-80

37) Jaffe JM, Malkowicz SB, Walker AH, MacBride S, Peschel R, Tomaszewski J, Van Arsdalen K, Wein AJ, Rebbeck TR (2000) Association of SRD5A2 Genotype and Patohological Characteristics of Prostate Tumors. Cancer Research 60: 1626- 1630

38) Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. Trends in Genetics 11(11): 449-456

39) Jorde LB, Bamshad M. Rogers AR (1998) Using mitochondrial and nuclear DNA markers to reconstruct human evolution. Bioessays 20: 126-136

40) Jorde LB, Watkins WS, Bamshad MJ (2001) Population genomics: a bridge from evolutionary history to genetic medicine. Human Molecular Genetics 10(20): 2199-2207

41) Keita SOY, Kittles RA, Royal CDM, Bonney GE, Furbert-Harris P, Dunston GM, Rotimi CM (2004) Conceptualising human variation. Nature Genetics Supplement 36(11): S17-S20

42) King JP, Kimmel M, Chakraborty R (2000) A Power Analysis of Microsatellite-Based Statistics for inferring Past Population Growth. Molecular Biology and Evolution 17(12): 1859-1868

43) Kittles R.A., Young D., Weinrich S., Hudson J., Argyropoulos G., Ukoli F., Adams-Campbell L. and Dunston G.M. 2001. Extent of linkage disequilibrium between the androgen receptor gene CAG and GGC arepeats in Human populations: implications for prostate cancer risk. Human Genetics 109: 253-261

44) Kittles RA, Chen W, Panguluri RK, Ahaghotu C, Jackson A, Adebamowo CA, Griffin R, Williams T, Ukoli F, Adams-Campbell L, Kwagyan J, Isaacs W, Freeman V, Dunston GM (2002) CYP3A4 and prostate cancer in African-American: causal and confounding association because of population stratification? Human Genetics 110: 553-560

45) Kittles RA, Panguluri RK, Chen W, Massac A, Ahaghotu C, Jackson A, Ukoli F, Adams-Campbell L, Isaacs W, Dunston GM (2001) CYP17A1 Promoter Variant Associated with Prostate Cancer Aggressiveness in African-American. Cancer Epidemiology, Biomarkers & Prevention 10: 943-947

46) Lange EM, Chen H, Brierley K, Livermore H, Wojno KJ, Langefeld CD, Lange K, Cooney KA (2000) The Polymorphic Exon 1 Androgen Receptor CAG Repeat in Men with a Potential Inherited Predisposition to Prostate Cancer. Cancer Epidemiology, Biomarkers and Prevention 9: 439-442

47) Lewis PO, Zaykin D (2000) Genetic Data Analysis: computer program for the analysis of allelic data. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs

48) Lin M, Wu R (2005) Theoretical basis for the identification of allelic variants that encode drug efficacy and toxicity. Genetics 170: 919-928

49) Lunn RM, Bell DA, Mohler JL, Taylor JA (1999) Prostate cancer risk and polymorphism in 17 hydroxylase (CYP17A1) and steroid reductase (SRD5A2). Carcinogenesis 20(9): 1727-1731

50) Makridakis NM, Reichardt JKV (2001) Molecular epidemiology of hormone-Metabolic Loci in Prostate Cancer. Epidemiologic Reviews 23(1): 24-29

51) Makridakis NM, Ross RK, Pike MC, Crocitto LE, Kolonel LN, Pearce CL, Henderson BE and Reichardt JK (1999) Association of mis-sense substitution in SRD5A2 gene with prostate cancer in African-American and Hispanic men in Los Angeles, USA. Lancet 354(9183): 975-8

52) Miller RD, Kwok P (2001) The birth and death of human single-nucleotide polymorphisms: new experimental evidence and implication for human history and medicine. Human Molecular Genetics 10(20): 2195-2198

53) Miller SA, Dykes DD, Polesky HF (1988) A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic Acid Research 16:1215

54) Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51:263-273

55) Murdock GP (1959) Africa Its peoples and their Culture History. McGraw-Hill Book Company

56) Nam RK, Toi A, Vesprini D, Ho M, Chu W, Harvie S, Sweet J, Trachtenberg J, Jewet MA, Narod SA (2001) V89L polymorphism of type-2, 5-alpha reductase enzyme predicts prostate cancer presence and progression. Urology 57: 199-204

57) Nam RK, Zhang WW, Trachtenberg J, Jewett MAS, Emami M, Vesprini D, Chu W, Ho M, Sweet J, Evans A, Toi A, Pollak M and Narod SA (2003) Comprehensive assessment of candidate genes and serological markers ofr the detection of prostate cancer. Cancer Epidemiology, Biomarkers and Prevention 12: 1429-1437

58) Ntais C, Polycarpon A, Ioannidis JPA (2003) Association of the CYP17A1 Gene Polymorphism with the Risk of Prostate Cancer: A Meta-Analysis. Cancer Epidemiology, Biomarkers & Prevention 12: 120-126

59) Nurse GT, Weiner JS, Jenkins T (1985) The peoples of Southern Africa and their affinities. Clarendon Press : Oxford

60) Oota H, Pakendorf B, Weiss G, von Haeseler A, Pookajorn S, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M (2005) Recent Origin and Cultural Reversion of a Hunter–Gatherer Group. PLoS Biol 3(3): e71

61) Paris PL, Kupelian PA, Hall JM, Williams TL, Levin H, Klein EA, Casey G, Witte JS (1999) Association between a CYP3A4 Genetic Variant and Clinical Presentation in African-American Prostate Cancer Patients. Cancer Epidemiology, Biomarkers & Prevention 8: 901-905

62) Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. CA Cancer J. Clin 55 (2):74-108

63) Pearce CL, Makridakis NM, Ross RK, Pike MC, Kolonel LN, Henderson BE, Reichardt JKV (2002) Steroid 5-α reductase type II V89L Substitution Is Not Associated with Risk of Prostate Cancer in a Multiethnic Population Study. Cancer Epidemiology, Biomarkers and Prevention 11: 417-418

64) Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A (2001) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansion and the slave trade. Annals of Human genetics 65: 439-458

65) Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, Carracedo Á, Comas D (2004) Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Human Genetics 115(5)439-447

66) Plummer SJ, Conti DV, Paris PL, Curran AP, Cassey G, Witte JS (2003) CYP3A4 and CYP3A5 Genotypes, Haplotypes and Risk of Prostate Cancer. Cancer Epidemiology, Biomarkers and Prevention 12: 928-932

67) Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease- common variant…or not? Human Molecular Genetics 11(20): 2417-2423

68) Ray DA, Walker JA, Hall A, Llewellyn B, Ballantyne J, Christian AT, Turteltaub K, Batzer MA (2004) Inference of human geographic origins using Alu insertion polymorphism. Forensic Science International 153(2-3):117-124

69) Rebbeck TR, Ambrosone CR, Bell DA, Chanock SJ, Hayes RB, Kadbular FF, Thomas DC (2004) SNPs, Haplotypes, and Cancer: Applications in Molecular Epidemiology. Cancer Epidemiology Biomarkers 13(5): 681-7

70) Rebbeck TR, Jaffe JM, Walker AM, Wein AJ, Malkowics SB (1998) Modification of Clinical Presentation of Prostate Tumors by a Novel Genetic Variant in CYP3A4. Journal of the National Cancer Institute 90(16): 1225-1229S

71) Reddy S, Shapiro M, Morton R, Brawley OW (2003) Prostate cancer in Black and White Americans. Cancer and metastasis Reviews 22:83-86

72) Reich DE and Lander ES (2001) On the allelic spectrum of human disease. Trends in Genetics 17(9):502-10

73) Ribeiro ML, Santos A, Carvallo-Salles AB, Hackel C (2002) Allelic frequencies of six polymorphic markers for risk of prostate cancer. Brazillian Journal of Medical and Biological Research 35: 205-213

74) Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, Batzer MA, Barbujani G (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Research 12:602-612

75) Ross RK, Bernstein L, Lobo RA, Shimizu H, Stanczyk FZ, Pike MC, Henderson BE (1992) 5-alpha-reductase activity and risk of prostate cancer among Japanese and US white and black males. Lancet 339: 887-89

76) Salam MT, Ursin G, Skinner EC, Dessissa T, Reichardt JKV (2005) Association between polymophisms in the steroid 5-α reductase type II (SRD5A2) gene and benign prostatic hyperplasia and prostate cancer. Urologic Oncology 23: 246-253

77) Sharp L, Cardy AH, Cotton SC, Little J (2004) CYP17A1 Gene Polymorphisms: Prevalence and Association with Hormone Levels and Related Factors. A Huge Review. American Journal of Epidemiology 160(8): 729-739

78) Shastry BS (2002) SNP alleles in human disease and evolution. Journal of Human Genetics 47: 561-566

79) Shibata A, Garcia MI, Cheng I, Stamey TA, McNeal JE, Brooks JD, Henderson S, Yemoto CE, Peehl DM (2002) Polymorphisms in the Androgen Receptor and

Type II 5α-Reductase Genes and Prostate Cancer Prognosis. The Prostate 52: 269-278

80) Sieh W, Edwards KL, Fitzpatrick AL, Srinouanprachanh SL, Farin FM, Monks SA, Kronmal RA, Eaton DL (2006) Genetic susceptibility to Prostate Cancer: prostate-specific antigen and its interaction with the androgen Receptor (United States). Cancer Causes and Control 17: 187-197

81) Söderström T, Wadelius M, Andersson S, Johansson J, Johansson S, Granath F, Rane A (2002) 5α-Reductase 2 polymorphism as risk factors in prostate cancer. Pharmacogenetics 12(4): 307-312

82) Soodyall H, Jenkins T (1992) Mitochondrial DNA polymorphisms in Khoisan populations from Southern Africa. Annals of Human Genetics 56: 315-324

83) Stanford JL, Noonan El, Iwasaki L, Kolb S, Chadwick RB, Feng Z, Ostrander EA (2002) A Polymorphism in the *CYP17A1* Gene and Risk of Prostate Cancer.

84) Stoneking M (2001) Single nucleotide polymorphisms. From the evolutionary past. Nature 409(6822): 821-2

85) Tan SY, Brown J (2006) Medicine in stamps: Gregor Mendel (1822- 1884) Man of God and Science. Singapore Med. J 47(11) 922

86) Tayeb MT, Clark C, Ameyaw MM, Haites NE, Evans DAP, Tariq M, Mobarek A, Ofori-Adjei D, Mcleod HL (2000) CYP3A4 promoter variant in Saudi, Ghanaian and Scottish Caucasian populations. Pharmacogenetics 10: 753-756

87) Taylor JG, Choi E, Forster CB, Chanock SJ (2001) Using genetic variation to study human disease. Trends in Molecular Medicine 7(11): 507-11

88) Tishkoff SA, Williams SM (2002) Genetic analysis of African populations: Human evolution and Complex diseases. Nature Reviews 3: 611-620

89) Van Schaik RHN, de Wildt SN, van Iperen NM, Uitterlinden AG, van den Anker JN, Lindermans J (2000) CYP3A4-V Polymorphisms Detection by PCR-Restriction Fragment Length Polymorphisms Analysis and Its Allelic Frequency among 199 Dutch Caucasians. Clinical Chemistry 46(11): 1834-1836

90) Vilchis F, Hernández D, Canto P, Mendez JP, Chaves B (1997) Codon 89 polymorphism of the human 5α-steroid reductase type 2 gene. Clinical Genetics 51: 399-402

91) Wadelius M, Andersson AO, Johansson JE, Wadelius C, Rane E (1999) Prostate cancer associated with CYP17A1 genotype. Pharmacogenics 9: 635-639

92) Watkins WS, Ricker CE, Bamshad MJ, Carrol ML, Nguyen SV, Batzer MA, Harpending HC, Rogers AR, Jorde LB (2001) Patterns of ancestral human diversity: An analysis of Alu-insertions and Restriction-site Polymorphisms. American Journal of Human Genetics 68:738-752

93) Wright AF, Carothers AD and Pirastu M (1999) Population choice in mapping genes for complex diseases. Nature Genetics 23: 397-404

94) www.history.com

95) Xu J, Meyers DA, Sterling DA, Zheng SL, Catalona WJ, Cramer SD, Bleecker ER and Xu JO (2002) Association Studies of Serum Prostate-specific Antigen Levels and the Genetic Polymorphisms at the Androgen Receptor and Prostate-specific Antigen Genes. Cancer Epidemiology Biomarker & Prevention 11: 664-669

96) Xue W, Coetzee GA, Ross RK, Irvine R, Kolonel L, Henderson BE and Ingles SA (2001) Genetic Determinants of Serum Prostate-specific Antigen levels in Healthy Men from a Multiethnic Cohort. Cancer Epidemiology Biomarkers & Prevention 10: 575-579

97) Xue W, Irvine RA, Yu MC, Ross RK, Coetzee GA, Ingles SA (2000) Susceptibility to Prostate Cancer: Interaction between Genotypes at the Androgen Receptor and Prostate-specific Antigen Loci. Cancer Research 60: 839-841

98) Yamada Y, Watanabe M, Murata M, Yamanaka M, Kubota Y, Ito H, Katoh T, Kawamura J, Yatani R, Shiraishi T (2001) Impact of genetic polymorphisms of 17-hyfroxylase cytochrome P-450 (CYP-17) and steroid 5α-reductase type II (SRD5A2) genes on prostate-cancer risk among Japanese population. Int. J. Cancer 92:683-686

99) Zeigler-Johnson CM, Walker AH, Mancke B, Spangler E, Jalloh M, McBride S, Malowicz SB, Ofori-Adjei D, Gueye SM, Rebbeck TR (2002) Ethnic Differences in the Frequency of Prostate Cancer Susceptibility Alleles at SRD5A2 and CYP3A4. Human Heredity 54: 13-21

100) Zhivotovsky LA, Rosenberg NA, Feldman MW (2003) Features of evolution and expansion of modern humans, inferred for genome wide microsatellite markers. American Journal of Human Genetics 72: 1171-1186