The Influence of Genetic Variation in *PSIP1* on HIV-1 Infectivity in Black South Africans

Nikki Gentle

A dissertation submitted to the Faculty of Science,
University of the Witwatersrand in fulfillment of the
requirements for the degree of Master of Science

Johannesburg, 2009

Declaration

I, Nikki Gentle, declare that this dissertation is my own work. It is being submitted in fulfillment of the requirements for the degree of Master of Science at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university.

……………………………………..

(Signature of candidate)

On this the………………..day of the...................month in the

year………………..

Dedication

To my family

Table of Contents

CHAPTER ONE - INTRODUCTION

CHAPTER TWO – MATERIALS AND METHODS

CHAPTER THREE – RESULTS

CHAPTER FOUR – DISCUSSION

List of Figures

List of Tables

List of Abbreviations

AIDS	acquired immune deficiency syndrome
AMP	adenosine monophosphate
AP-1	activating protein complex
APS	ammonium persulphate
ARG	AIDS restricting gene
ATP	adenosine triphosphate
bp	base pair
CCR5	chemokine receptor 5
CD/CV	common disease/common variant
Cl	chlorine
cm	centimeter
CR	charged region
DNA	deoxyribonucleic acid
dNTP	dinucleotide triphosphate
EDTA	ethylene diamine tetra-acetic acid
E2F	E2F transcription factor 2
g	gram
G6PD	Glucose-6-phosphate dehydrogenase
GF	genotype frequency
GRE	glucocorticoid response element
HCl	hydrogen chloride
HDGF	hepatoma-derived growth factor
HIV	human immunodeficiency virus
HSE	heat shock element
IBD	integrase binding domain
IN	integrase
Indel	insertion/deletion polymorphism
IRF	interferon regulatory factor
kb	kilobase

kDa	kilodalton
LD	linkage disequilibrium
LEDGF	lens epithelial-derived growth factor
LTR	long terminal repeat
MAF	minor allele frequency
Mg	magnesium
MgCl	magnesium chloride
mg	milligram
ml	millilitre
mM	millimolar
MRCA	most recent common ancestor
mRNA	messenger RNA
mtDNA	mitochondrial DNA
NADPH	nicotinamide adenine dinucleotide phoshosphate
NaCl	sodium chloride
NaOH	sodium hydroxide
NF-κB	nuclear factor of kappa light chain enhancer of activated B-cells
ng	nanogram
NLS	nuclear localization signal
Oct-1	octamer transcription factor 1
PCR	polymerase chain reaction
PIC	pre-integration complex
PPi	pyrophosphate
PSIP1	PC4 and SFRS1 interacting protein
PWWP	proline-tryptophan-tryptophan-proline
RAO	recent out of Africa
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
RT	reverse transcriptase
SDS	sodium dodecyl sulphate
SNP	single nucleotide polymorphism
SP-1	transcription factor SP-1

SRE	serum response element
STAT	signal transducer and activator of transcription
STRE	stress-related regulatory element
TBE	Tris-boric acid-EDTA buffer
TBS	transcription factor binding site
TFIID	transcription factor II D
TGF-β	transforming growth factor beta
TIE	TGF-β inhibitory element
T_m	melting temperature
Tris	tris(hydroxymethyl)aminomethane
µg	microgram
µl	microlitre
µM	micromolar
UTR	untranslated region
UV	ultra violet
V	volt
VDR/RXR	vitamin D receptor/retinoic acid X receptor

Abstract

Genetic variation plays an important role in determining an individual's susceptibility to infectious disease. *PSIP1* encodes LEDGF/p75, which stably associates with the core domain of HIV-1 integrase via a highly-conserved integrase binding domain (IBD) located in its C-terminal. Through this interaction, the protein tethers HIV-1 IN to chromosomes at sites corresponding to regions of high LEDGF/p75-mediated transcription. Genetic variation within *PSIP1* was identified and characterized in black South Africans to establish whether variation in this influences an individual's susceptibility to HIV infection. PCR assays were designed to amplify regions within the upstream non-coding region, IBD and DNA-binding domains of the gene and selected polymorphisms were then genotyped using allele-specific PCR, RFLP-PCR and Pyrosequencing™ assays. Three insertion-deletion (indel) and eight single nucleotide polymorphisms (SNP) where identified through sequencing. Four of the SNPs had been recorded previously, while the seven other polymorphisms had not and appear to be unique to our population. Differences in allelic and genotypic frequencies where found between the various ethnic groups represented in this study, which were reflected in the underlying haplotype structure within this gene, suggesting that genetic substructure exists within the black South African population. Differences in allele and genotype frequencies were also seen between HIV$^+$ individuals and the general population. Thus variation within *PSIP1* may influence an individual's susceptibility to HIV-1 infectivity and/or rate of disease progression.

Chapter 1

Introduction

Human immunodeficiency virus-1 (HIV-1), a member of the lentiviridae family of retroviruses, infects CD4$^+$ cells, causing acquired immune deficiency syndrome (AIDS) (Barre-Sinoussi *et al.*, 1983; Popovic *et al.*, 1984; Levy *et al.*, 1984). Thus, like other lentiviruses, HIV is able to infect non-dividing, terminally differentiated cells, a property that distinguishes them from oncoretroviruses. Effective infection of these cells by the virus requires that the virus' RNA genome (Figure 1.1) be reverse transcribed into cDNA that can be stably incorporated into the DNA of the host cell (Bushman *et al.*, 1990).

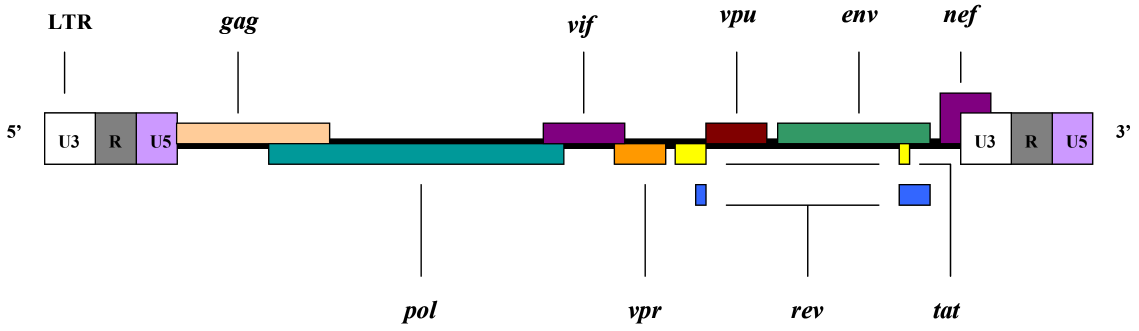


Figure 1: A schematic representation of the genome of the HIV-1 provirus. The ~9kb genome contains nine genes that encode fifteen proteins.

Reverse transcription of the viral RNA into cDNA is performed by the viral reverse transcriptase (RT), within a complex known as the reverse transcription complex (Fassati and Goff, 2001). The synthesis of a minus-strand DNA from the virus' plus-strand RNA is followed by synthesis of a plus-strand DNA complementary to the minus-strand, resulting in the formation of a blunt-ended, double-stranded DNA molecule**.** This product of reverse transcription then remains in the cytoplasm as part of the viral pre-initiation complex (PIC) (Miller *et al.*, 1997).

The HIV-1 PIC is a nucleoprotein complex containing the newly synthesized double-stranded viral cDNA, several viral proteins (including reverse transcriptase, integrase, matrix and nucleocapsid) and a number of the host's cellular proteins. The PIC is essential for integration of the viral cDNA into the host chromosome and is initially assembled in the cytoplasm of the host cell, but later imported into the nucleus (Miller *et al.*, 1997).

1.1 HIV-1 Integration

The process of viral integration takes place within the PIC and involves two enzymatic reactions, both catalyzed by the viral integrase. The first is referred to as 3' processing and is characterized by the removal of a GT dinucleotide from the 3' end of each of the viral long terminal repeats (LTRs). These GT dinucleotides are adjacent to a highly conserved CA dinucleotide. This reaction takes place while the PIC is still in the cytoplasm (Pauza, 1990; Engelman *et al.*; 1991).

The second reaction, strand transfer, takes place following nuclear import of the PIC and is initiated following nucleophilic attack of the 5' end of the host's DNA by the 3' hydroxyl groups of the viral DNA. This attack is directed at two phosphodiester bonds located on either side of the major groove of the host's DNA. Nucleophilic attack is followed by a transesterification reaction, in which the 3' ends of the viral DNA (carrying the conserved CA dinucleotides) are ligated to the 5' ends of the host's DNA (Engelman *et al.*, 1991; Bushman and Craigie, 1991).

Following strand transfer, the 5' ends of the viral DNA remain unattached to the 3' ends of the host's DNA, and as a result are flanked by two 5-nucleotide gaps and two unmatched AC dinucleotides (Bushman and Craigie, 1991). Cleavage of the unpaired dinucleotides from the 5' ends of the viral DNA and repair of the gaps by the host's DNA repair enzymes complete the integration process (Yoder and Bushman, 2000).

1.2 HIV-1 Integrase

HIV-1 integrase (IN) is a 32-kDa protein, encoded by the viral *pol* gene, and is initially present as part of a large Gag-Pol polyprotein. Maturation of the IN protein is achieved by cleavage of the precursor by the viral protease. The mature protein consists of three domains (Figure 1.2.1), namely an N-terminal domain, a core catalytic domain (containing the enzyme's catalytic site) and an α-helical, arginine/lysine-rich C-terminal domain (Engelman *et al.*, 1993). IN predominantly accumulates in the nucleus of human cells and stably associates with condensed chromosomes during mitosis (Maertens *et al.,* 2003).

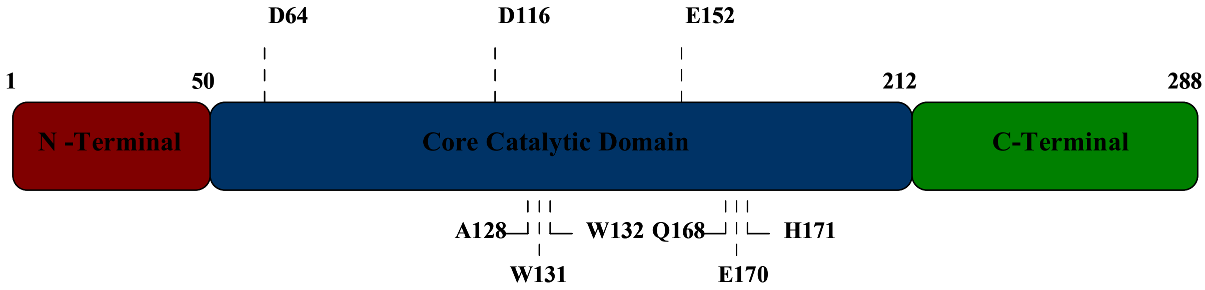
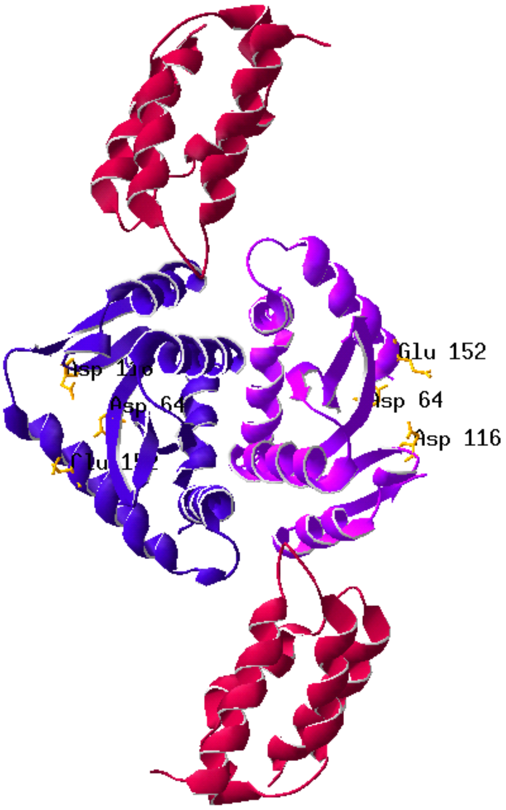


Figure 1.2.1: The domain structure of HIV-1 integrase, highlighting the key residues of the catalytic site and those involved in binding LEDGF/p75. Residues D64, D166 and E152 are responsible for coordinating magnesium ions in the catalytic centre. The remaining residues shown are located in the pocket at the p75/IN dimer interface.

It has been demonstrated that in the nucleus, HIV-1 IN is present in complex with the host-encoded protein identified as lens epithelium-derived growth factor/transcriptional co-activator 75 (LEDGF/p75), which forms a component of the PIC. This interaction between HIV-1 IN and LEDGF/p75 (Figure 1.2.2) involves the formation of a symmetrical complex consisting of a pair of IN tetramers in association with two subunits of LEDGF/p75 (Cherepanov *et al.*, 2003; 2005a; 2005b).

A.


Glu 152
Asp 64
Asp 116
Asp 64

B.

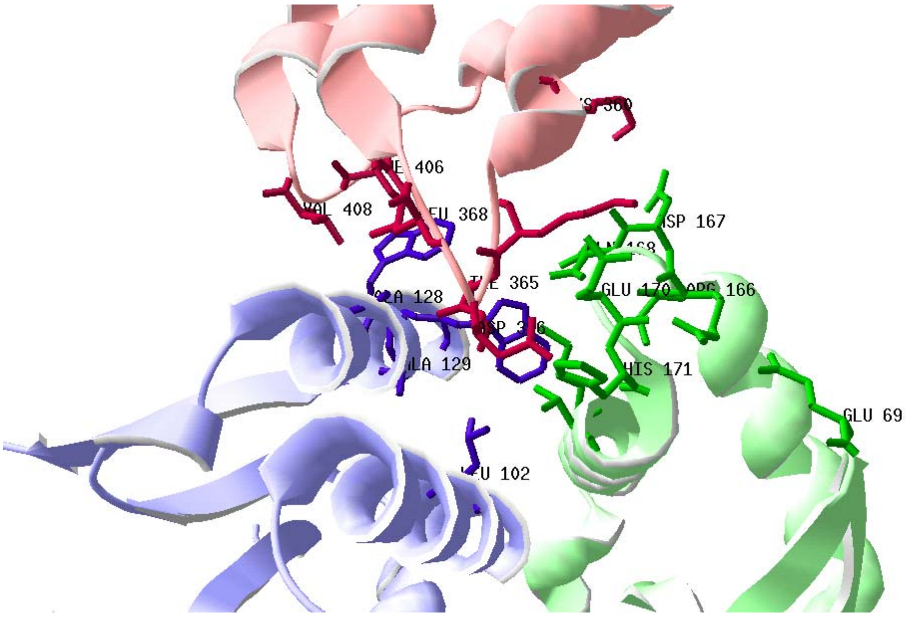

E 406
VAL 408
EU 368
SP 167
GLU 170
ILE 365
LA 128
LA 129
HIS 171
GLU 69
EU 102

Figure 1.2.2: The interaction between HIV-1 IN and LEDGF/p75. (A) The overall structure of the tetrameric HIV-1 IN-LEDGF/p75 complex. IN chains A and B are shown in blue and purple, respectively, and the LEDGF/p75 subunits are shown in red. The side chains of the DDE catalytic centre residues are shown in orange. (B) The key residues involved in the interaction between HIV-1 IN and LEDGF/p75. IN chains A and B are shown in blue and green, respectively, and a single LEDGF/p75 subunit is shown in pink. Both figures were made with Swiss PDB Viewer, using Protein Data Base crystal structure file 2B4J (www.pdb.org).

1.3 LEDGF/p75

LEDGF/p75 is a ubiquitously expressed 60 kDa protein, consisting of 530 amino acids arranged in a number of functional domains (Figure 1.3) (Ge *et al.*, 1998a; Nishizawa *et al.*, 2001). The protein is a member of the hepatoma-derived growth factor (HDGF) family and like all HDGF-related proteins, contains a PWWP (Pro-Trp-Trp-Pro) motif in its N-terminal domain (Figure 1.3) that can mediate both protein-protein and DNA-binding interactions (Nakamura *et al.*, 1994; Ishimoto *et al.*, 1997). Proteins containing PWWP domains are capable of mediating the protein-protein interactions involved in the regulation of chromatin structure, suggesting a role for these proteins in the regulation of transcriptional processes (Stec *et al.*, 2000; Ge *et al.*, 2004).

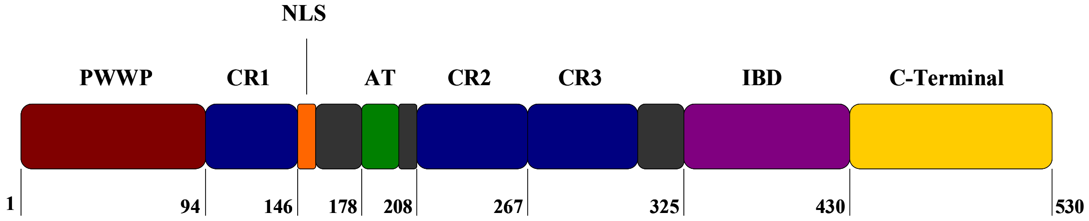


Figure 1.3: The domain structure of LEDGF/p75. The amino acid residues at the boundaries of each domain are also shown.

A functional nuclear localization signal (NLS) and a pair of AT hooks have also been identified in this protein (Figure 1.3) (Cherepanov *et al.*, 2004; Maertens *et al.*, 2004; Vanegas *et al.*, 2005). AT hooks mediate the binding of proteins to AT-rich DNA regions (Aravind and Landsman, 1998). These, together with the PWWP domain and NLS (Figure 1.3), form a tripartite DNA-binding motif. This conserved tripartite element is necessary and sufficient to mediate the binding of this protein to DNA *in vitro*. Additional charged regions in the N-terminus (CR1, CR2 and CR3) (Figure 1.3) further enhance the binding activity of the DNA binding motif (Llano *et al.*, 2006a).

The region of the LEDGF/p75 protein that has excited the most interest, however, is its C-terminal domain (Figure 1.3), which contains a highly-conserved integrase-binding domain (IBD) (Cherepanov *et al.*, 2004) that binds JPO2 (Maertens *et al.*, 2006; Bartholomeeusen *et al.*, 2007) and lentiviral integrases, including HIV-1 IN (Cherepanov 2007).

1.4 The Role of LEDGF/p75 in HIV-1 Integration

The association between LEDGF/p75 and HIV-1 IN is formed between the IBD within the C-terminal of LEDGF/p75 and the catalytic core domain of IN (Cherepanov *et al.*, 2004; 2005a; 2005b). This association prevents proteosomal degradation of HIV-1 IN (Llano *et al.*, 2004b) and provides a means whereby IN may be tethered to host chromosomes (Maertens *et al.*, 2003; Emiliani *et al.*, 2005) at AT-rich regions within LEDGF/p75-regulated genes (Ciuffi *et al.*, 2005; Hombrouck *et al.*, 2007; Shun *et al.*, 2007; Marshall *et al.*, 2007).

LEDGF/p75 is able to perform this function because while it binds to IN via its C-terminal IBD (Cherepanov *et al.*, 2003), it binds DNA though a tripartite DNA binding domain located in its N terminal (Llano *et al.*, 2006a). As a result, both ends of LEDGF/p75 must be present and functional for HIV-1 IN to associate with chromosomes (Llano *et al.*, 2006b; Shun *et al.*, 2007).

LEDGF/p75 has also been shown to interact with the INs of a wide range of other lentiviruses (Llano *et al.*, 2004a; Busshots *et al.*, 2005; MacNeil *et al.*, 2006; Cherepanov, 2007; Marshall *et al.*, 2007) and like HIV-1, these viruses have all shown a preference for integrating into transcriptional units (Schröder *et al.*, 2002; Ciuffi *et al.*, 2005; Crise *et al.*, 2005; MacNeil *et al.*, 2006; Marshall *et al.*, 2007). Conversely, LEDGF/p75 does not interact with non-lentiviral INs (Llano *et al.*, 2004a; Busshots *et al.*, 2005; Cherepanov, 2007), which display very different integration preferences (Mitchell *et al.*, 2004; Narezkina *et al.*, 2004; Barr *et al.*, 2005; Lewinski *et al.*, 2006). Collectively, these data suggest that the LEDGF/p75-IN interaction is the primary cellular determinant of lentiviral integration site selection.

1.5 *PSIP1*

LEDGF/p75 is encoded by *PSIP1*, which maps to chromosome 9p22.2 (Singh *et al.*, 2000). This 46 923bp gene consists of sixteen exons and encodes five transcripts (www.ensembl.org), of which only two (LEDGF/p75 and an alternatively splice product, p52) have been identified as functional proteins (Ge *et al.*, 1998a; Nishizawa *et al.*, 2001). p52 shares 325 N-terminal residues with LEDGF/p75, but has a unique 8bp C terminus that does not contain an IBD (Ge *et al.*, 1998a; 1998b) and as a result, does not associate with HIV-1 IN.

Analysis of the gene's promoter region has revealed it is a TATA-less promoter with three transcriptional start sites - a major site (an A at position +1) and two minor sites (a G at +35 and a C at +55) (Singh *et al.*, 2002). TATA-less promoters display different mechanisms of transcription initiation than other promoter elements such as initiator sequences (Inrs) (Martinez *et al.*, 1994) and Sp-1 binding sites (Smale *et al.*, 1990) are required to recruit TFIID to the promoter (Crawford *et al.*, 1999). To this end, a wide variety of putative regulatory elements have also been identified within the promoter, including AP1, HSE, STRE, SRE, E2F1, IRF-2, IRF-1, GRE, VDR/RXR, NF-κB, SP1, TGF-ß inhibitory (TIE), STAT and Oct1 elements (Singh *et al.*, 2002; Magana-Arachchi *et al.*, 2003). Because LEDGF/p75 binds to HSE and STRE elements in the promoters of several stress-related genes, the presence of these elements in the promoter of *PSIP1* suggests the gene may be self-regulating (Sharma *et al.*, 2000; Shinotara *et al.*, 2002; Singh *et al.*, 2002)

1.6 DNA Polymorphisms

DNA sequence polymorphisms are defined as genetic variations for which the most common allele occurs at a frequency of less than 99% in a given population (Knight, 2005) Single nucleotide polymorphisms (SNPs) are by far the most common form of human genetic variation, occurring on average every 1 000bp across the genome, and constituting an estimated 90% of all genetic variation (Collins *et al.*, 1998; The International SNP Map Working Group, 2001). Much time has been devoted to the discovery and study of these polymorphisms, with the result that a dense SNP map of the human genome is now available (The International SNP Map Working Group, 2001)

Insertion-deletion polymorphisms (indels) are a second, less common type of variation, which together with SNPs constitute the major mutational processes driving gene evolution (Taylor *et al.*, 2004). While a great deal of time and effort has been devoted to the characterization and genotyping of SNPs, little is known about the frequency of indels and the mechanisms whereby they arise. This can largely be attributed to the fact that they are predominantly deleterious mutations and often produce unstable phenotypic effects that prevent them from becoming fixed within a population, making them difficult to detect and characterize. However, recent estimates suggest these polymorphisms constitute as much as 15-18% of the genetic variation within the human genome and could be useful as genetic markers (Dawson *et al.*, 2002; Weber *et al.*, 2002; Bhangale *et al.*, 2005; Mills *et al.*, 2006).

1.6.1 Regulatory Polymorphisms

Traditionally, studies have focused on identifying polymorphisms within the coding regions of genes. This strategy is based on the premise that these polymorphisms were most likely to have a functional effect on the given protein. However recently, in light of the discovery that allelic differences in gene expression exist within and between individuals, interest has been renewed in identifying and characterizing regulatory polymorphisms (Yan *et al.*, 2002; Cheung *et al.*, 2003; Morley *et al.*, 2004; Pastinen *et al.*, 2004). These polymorphisms generally alter gene expression at the level of transcription, mRNA stability or translation and are usually located within the 5' and 3' untranslated regions (UTRs), enhancer and repressor elements, intronic and intergenic regions, at splice junctions and within promoters (Buckland, 2006).

Of particular interest are polymorphisms located in the 5' region immediately upstream of genes. This region is often the site of the gene's proximal promoter and as such, is usually a rich source of elements involved in the initiation and regulation of transcription – including RNA polymerase II binding sites, binding sites for a number of elements involved in the formation of the transcription initiation complex and transcription binding sites (TBSs) (Buckland *et al.*, 2005; Buckland, 2006). Polymorphisms within any of these *cis*-acting elements (e.g. a TBS) which alter their sequence are of particular interest, as these are considered most likely to produce an alteration in gene expression by preventing the binding of the relevant *trans*-acting element (e.g. the transcription factor).

However, the mechanisms involved in gene expression and regulation are vast, complex and often poorly understood; with the expression of any given gene being under the control of a variety of *cis*- and *trans*-acting elements, which often act in concert to exert their effects. Thus, the influence of a polymorphic variant within a *cis*-acting element may be masked by that of another variant in a related *trans*-acting element (or vice versa), often making it difficult to positively identify a single allelic variant as being the causative agent of a difference in gene expression. As a result, focus has shifted towards characterizing linkage disequilibrium (LD) in the regions surrounding possible regulatory polymorphisms and identifying haplotypes that may be associated with changes in gene expression (Knight, 2005; Buckland, 2006).

1.7 Linkage Disequilibrium and Haplotype Structure

Individuals who carry a particular allele at one polymorphic locus have generally been found to carry specific alleles at other nearby variable sites on the same chromosome. This non-random correlation between neighboring polymorphisms is referred to as linkage disequilibrium (LD) (Lewontin, 1964) and the combinations of alleles arising as a result of these non-random associations are referred to as haplotypes (Gabriel *et al.*, 2002). LD is created when a new mutation arises on a chromosome carrying a particular allele at a nearby polymorphic locus. This new mutation disrupts the existing haplotype structure of the chromosome in question and creates a new haplotype along this chromosome. This haplotype in turn, is transmitted to subsequent generations until it too is disrupted by mutation or recombination (Ardlie *et al.*, 2002; Gabriel *et al.*, 2002).

1.7.1 The Importance of Identifying LD Patterns and Haplotype Structure

Because LD between neighboring polymorphisms reflects haplotypes descended from ancestral chromosomes (Reich *et al.*, 2001), variations in the extent and distribution of LD across the human genome offer invaluable insight into both the genealogical and demographic history of the human population (Ardlie *et al.*, 2002). This is highlighted by a number of studies which have detected LD patterns consistent with migration, population expansion and admixture and have used these findings to trace the path of human evolution and draw conclusions about the major migratory events which have helped to shape modern patterns of genetic variation, both within and between populations (Quintana-Murci *et al.*, 1999; Ingman *et al.*, 2000; Alonso and Armour, 2001; Salas *et al.*, 2004).

The elucidation of LD patterns and haplotype structure across the genome is also an important component to identifying loci involved in the development of disease (Tishkoff and Verrelli, 2003). Because a dense genome-wide map of SNPs and other variation is now available (The International SNP Map Working Group, 2001; Mills *et al.*, 2006), differences in LD patterns and haplotype structure can now also be used to identify loci involved in the development of disease indirectly by detecting LD between these disease-associated loci and nearby genetic markers, for which genotyping data are available (Reich *et al.*, 2001; Gabriel *et al.*, 2002).

1.7.2 Using LD and Haplotype Structure to Define Populations

While it is well established that comparison of LD patterns and haplotype structure within and between populations provides a useful tool for the elucidation of human genealogical and demographic history and identifying loci involved in the development of disease (Reich *et al.*, 2001; Gabriel *et al.*, 2002), the criteria selected to define population boundaries remains a contentious issue. Perhaps the most hotly contested of these is the issue surrounding the use of race or ethnicity as a basis for genetic classification. Evidence has been presented for both sides of this argument (Tishkoff and Kidd, 2004; Jorde and Wooding, 2004; Mountain and Risch; 2004) and while some studies reveal that certain loci exhibit signatures of selection that can be correlated with traditional concepts of race, the majority of the evidence based on LD patterns and haplotype structure suggests that geographical distribution and ancestry may be more useful parameters for defining population boundaries (Rosenberg *et al.*, 2002; Lane *et al.*, 2002; Ramachandran *et al.*, 2005; Li *et al.*, 2008).

These studies have shown that individuals (regardless of race or ethnic background) who share geographical ancestry exhibit similar patterns of LD and haplotype structure - more so than individuals of the same race or ethnic group from geographically distinct locations (Rosenberg *et al.*, 2002; Ramachandran *et al.*, 2005; Li *et al.*, 2008). This would then imply that some knowledge of both global and regional demographic history is necessary to interpret LD patterns and haplotype structure detected within specific populations when in search of disease susceptibility loci (Tishkoff and Verrelli, 2003).

1.8 Genetic Variation and Human Demographic History

While several models of human evolution have been proposed (Tishkoff and Williams, 2002; Excoffier, 2002), the most widely accepted of these, the recent out of Africa (RAO) model, proposes that all modern humans are descended from a common *Homo sapiens* ancestor that evolved in Africa 100 000-200 000 years ago (Stringer and Andrews, 1988; Armour *et al.*, 1996; Ingman *et al.*, 2000). All non-African populations then arose as a result of the migration of one or more groups of these ancestral humans through East Africa into Asia 44 000-200 000 years ago (Quintana-Murci *et al.*, 1999, Macualay *et al.*, 2005). Following the migration out of Africa, the non-African population expanded rapidly and spread (Alonso and Armour, 2001; Zhivotovsky *et al.*, 2003), supplanting any and all archaic *Homo* populations still present outside of Africa (such as the Neanderthals), with little or no admixture (Nordborg, 1998; Serre *et al.*, 2004; Plagnol and Wall, 2006).

This model has received overwhelming support from a number of studies based on autosomal (Tishkoff *et al.*, 1996; Zietkiewicz *et al.*, 1997; Tishkoff *et al.*, 1998; Tishkoff *et al.*, 2000) mitochondrial (mtDNA) (Chen *et al.*, 1995; Ingman *et al.*, 2000; Macaulay *et al.*, 2005) Y-chromosomal (Seielstad *et al.*, 1999; Underhill *et al.*, 2000; Hammer *et al.*, 2001) and X-chromosomal (Armour *et al.*, 1996; Hammer *et al.*, 1997; Kaessermann *et al.*, 1999) DNA variation, which have highlighted the different patterns of variation present in non-African populations relative to African populations.

Africans have the largest number of population-specific alleles and the variation present in non-African populations is only a subset of the variation present in African populations (Armour *et al.*, 1996; Watson *et al.*, 1997; Zietkiewicz *et al.*, 1997; Kidd *et al.*, 1998), which is consistent with the hypothesis that all non-African populations are derived from a small, ancestral group of individuals who migrated to Asia through East Africa 44 000-200 000 years ago (Quintana-Murci *et al.*, 1999; Macaulay *et al.*, 2005).

Africans also typically display lower levels of LD than non-African populations, with the result that African populations carry a greater number of haplotypes (Reich *et al.*, 2001; Gabriel *et al.*, 2002). This has largely been ascribed to the fact that African populations have always maintained a larger effective population size than non-African populations, which have been subject to extensive genetic drift as a consequence of having undergone bottleneck and founder effects following the migration out of Africa. This has allowed more time (i.e. more generations) for LD to decay as a result of mutation and recombination (Stoneking *et al.*, 1997; Kidd *et al.*, 1998; Reich and Goldstein, 1998; Relethford and Jorde, 1999; Scozzari *et al.*, 1999). This again can be attributed to the initial migration out of Africa (Quintana-Murci *et al.*, 1999; Macaulay *et al.*, 2005).

1.9 Genetic Variation in Africans

The differences in genetic variation seen between African- and non-African populations can thus be seen as a testament to how migratory events can have far-reaching consequences with regard to the shaping of variation patterns both within and between populations. Other migratory events within Africa, prior to and following the exit of non-African populations out of northeast Africa, have also had a significant impact on shaping African genetic diversity patterns (Tishkoff and Williams, 2002; Zhivotovsky *et al.*, 2003; Henn *et al.*, 2008). Perhaps the most significant of these events was the migration of Bantu-speaking West Africans into eastern and southern African 1 500-3 000 years ago (Scozzari *et al.*, 1999; Salas *et al.*, 2002; Tishkoff and Williams, 2002).

1.9.1 The Influence of the Bantu Expansion

Based on archeological evidence, the Bantu expansion is proposed to have originated in the Cross River Valley of Central Africa, moving first into the Great Lakes region of eastern Africa and subsequently into southern Africa (Salas *et al.*, 2002; Tishkoff and Williams, 2002; Tishkoff and Verrelli; 2003). The archeological evidence is well supported by patterns of mtDNA (Soodyall *et al.*, 1996; Chen *et al.*, 2000; Salas *et al.*, 2002; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007) and Y-chromosomal (Passarino *et al.*, 1998; Scozzari *et al.*, 1999; Hammer *et al.*, 2001) variation. However, slight discrepancies exist between the findings of these studies, suggesting that males and females may have contributed differently to shaping the patterns of variation present in modern Africans.

Mitochondrial DNA variation can be classified into several (often geographically distinct) haplogroups (Gonder *et al.*, 2007). While all of these haplogroups are present in Africans populations, 3 main mitochondrial lineages (namely L1, L2 and L3) can be clearly distinguished (Salas *et al.*, 2002; Tishkoff and Williams, 2002). The L1 lineage is the most ancient lineage of these and includes the most recent common ancestor (MRCA) of human mtDNA (Ingman *et al.*, 2000; Salas *et al.*, 2002). L1 is thought to then have given rise to the L2 and L3 lineages 60 000-103 000 years ago (Chen *et al.*, 1995; Watson *et al.*, 1997; Chen *et al.*, 2000). This is supported by the observation that both the L2 and L3 lineages carry only a subset of the variation found in the L1 lineage. All non-African populations are descended from a subgroup of the L3 lineage (and thus carry only a subset of the variation found in this lineage) (Chen *et al.*, 1995; Gonder *et al.*, 2007).

L1 can be further subdivided into several subclades, which show distinct geographical distributions. The most common of these subclades, L1a, is common in East, Central and southeast Africa but is virtually absent in North, West and southern Africa. L1b on the other hand is most frequent in West Africa (and regions of North and Central Africa), but is rarely found in East, southeast or southern Africa. L1c occurs primarily in Central Africa, while L1d is common in the Khoisan people of southern Africa and is found at much lower frequencies in southeast and East Africa (Chen *et al.*, 2000; Salas *et al.*, 2002; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). These observations are consistent with the view that the L1 lineage arose in East Africa and spread into other parts of Africa as a result of a series of migratory events.

L2 is commonly subdivided into four main subclades, namely L2a-d. L2 is proposed to have originated in West Africa and then spread to southeast Africa during the Bantu expansion. This view is supported by the observation that L2a is common in southeast Africa, while L2b, -c and –d are most common in West and Central Africa (Chen *et al.*, 2000; Salas *et al.*, 2002). The subclades of the L3 lineage also show distinct geographical distributions that suggest this lineage arose in East Africa (where it is most frequent) and then spread into West, Central and southern Africa (Watson *et al.*, 1997; Salas *et al.*, 2002).

Y-chromosomal data also supports the view of a common ancestry between East Africans and southern African Khoisan populations (Seminò *et al.*, 2002), despite the vast distances separating these populations. When combined with observations that specific Y chromosomal haplotypes are common in both East Africans and West Africans, while other haplotypes are common to both southern Africans and West Africans (Passarino *et al.*, 19998; Scozzari *et al.*, 1999); it becomes clear that the Bantu expansion was a defining event in the demographic history of the African continent, which helped to shape present day patterns of genetic variation in sub-Saharan Africans.

1.9.2 Genetic Variation in South Africans

The South African population is composed of a wide range of linguistic groups. Bantu languages (Zulu, Xhosa, Pedi, Tswana, Southern Sotho, Tsonga, Swazi, Venda and Ndebele) comprise eight of the eleven official languages of South Africa and the country is also home to the Kung and Khwe Khoisan-speaking populations. Studies aiming to characterize mtDNA and Y-chromosomal variation in the Kung and Khwe (Passarino *et al.*, 1998; Scozzari *et al.*, 1999; Chen *et al.*, 2000; Salas *et al.*, 2002; Seminò *et al.*, 2002; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007) have revealed distinct genetic differences between these populations, despite the similarities in their linguistic patterns. While the Kung share notable similarities with other Khoisan-speaking populations in East Africa (Chen *et al.*, 2000; Salas *et al.*, 2002; Knight *et al.*, 2003; Tishkoff *et al.*, 2007), the Khwe show closer affinity with West African Bantu-speaking populations (Chen *et al.*, 2000; Salas *et al.*, 2002; Gonder *et al.*, 2007).

The genetic affinities of the Bantu-speaking populations, however, are less clear. A study by Lane *et al.* (2002) based on both autosomal and Y-chromosomal data revealed substructure exists within South African Bantu-speaking populations and that commonalities in linguistic patterns are not necessarily reflected at the genetic level. The Bantu languages spoken in South Africa all belong to the southern branch of the eastern Bantu-speaking linguistic group, but these languages can be subdivided into a further three language groups, namely Sotho/Tswana, Nguni and Venda (Lane *et al.*, 2002).

The Nguni language group comprises Zulu, Xhosa and Tsonga speakers and members of this group constitute more than 40% of the South African population. The Sotho/Tswana language group is comprised of Southern Sotho, Tswana and Pedi speakers and its members represent approximately 25% of the population. While members of both these groups understand the languages spoken by other members within the group, the differences between the two groups prevent them from understanding each other. Venda speakers represent a very small percentage of the population and this language is distinct from both Sotho/Tswana and Nguni languages (Lane *et al.*, 2002).

The study by Lane *et al.* (2002) found that while Sotho/Tswana speakers show similar patterns of variation that correlate with their linguistic patterns, the same could not be said for speakers of Nguni languages. Zulu and Xhosa speakers showed very similar patterns of variation to each other; but were distinctly different from Tsonga speakers, despite the linguistic similarities displayed by the three groups. The Tsonga were rather found to resemble the Venda genetically, a pattern attributable to shared demographic history between these two groups (Lane *et al.*, 2002). These findings thus provide further evidence to support the idea that demographic history can have an important influence on the shaping of patterns of genetic variation.

1.10 Genetic Variation and Disease

It is now well established that genetic variation plays an important role in determining an individual’s susceptibility to disease (Tishkoff and Verrelli, 2003). Because a dense genome-wide map of SNPs and other variation is now available (The International SNP Map Working Group, 2001; Mills *et al.*, 2006), differences in LD patterns and haplotype structure can now also be used to identify loci involved in the development of disease by detecting LD between these loci and nearby genetic markers (Reich *et al.*, 2001; Gabriel *et al.*, 2002).

While the Common Disease/Common Variant (CD/CV) hypothesis proposes that the genetic factors underlying common diseases will be reflected by a few common alleles that are present in high frequency across all populations (Chakravarti, 1999); recent findings suggest complex diseases may be influenced by susceptibility alleles at many loci, present at different frequencies in geographically distinct populations (Pritchard, 2001; Pritchard and Cox, 2002). These geographical restrictions in frequency may be as result of mutation, recombination, migratory events, genetic drift, population expansion or differential exposure to selective pressures (Tishkoff and Verrelli, 2003). Regulatory polymorphisms have been identified as being particularly important in determining susceptibility to complex disease, as the presence of these polymorphisms within a gene results in differences in gene expression within and between populations at the affected locus (Knight, 2005; Buckland, 2006).

A classic example of how differences in allele frequencies that arise as a result of geographically restricted selective pressures can influence susceptibility to complex disease is that of the influence of variation at the *Glucose-6-phosphate dehydrogenase* (*G6PD*) locus on an individual's susceptibility to malaria. The enzyme encoded by this gene is involved in glucose metabolism and is responsible for generating nicotinamide adenine dinucleotide phosphate (NADPH) in red blood cells (Verrelli *et al.*, 2002).

A variety of G6PD variants, with varying levels of enzyme activity, have been identified and classified on the basis of their electrophoretic mobility (Verrelli *et al.*, 2002). The B variant, which has normal enzyme activity, has been identified as the ancestral allele and has a worldwide distribution. However, variants A, A- (both restricted to sub-Saharan Africa) and Med (found in North African, Middle Eastern and Mediterranean populations) result in enzyme deficiencies and the distributions of these deficiency variants are restricted to regions with past and present histories of high malaria incidence (Vulliamy *et al.*, 1992; Beutler, 1994; Ruwende *et al.*, 1995, 1998; Tishkoff *et al.*, 2001; Verrelli *et al.*, 2002).

Only the A- variant, which has only 12% enzyme activity, is thought to offer a protective effect against malaria caused by *Plasmodium falciparum* (Ruwende *et al.*, 1995; Tishkoff *et al.*, 2001). This variant is the result of a single amino acid substitution in exon 4 of *G6PD* and is always associated with the amino acid change that gives rise to the A variant, which has 85% enzyme activity (Vulliamy *et al.*, 1992; Verrelli *et al.*, 2002).

While pairwise LD is low between these and neighbouring polymorphisms (Verrelli *et al.*, 2002), analysis of the entire gene has revealed distinct haplotypes that characterize each variant and result in the varying levels of enzyme deficiency (Tishkoff *et al.*, 2001). This has led to the suggestion that selection favours the resulting enzyme deficiency, rather than the specific allelic variants responsible (Verrelli *et al.*, 2002). As a result, these alleles have been maintained at relatively high frequencies by balancing selection, despite in some cases being associated with haemopathologies (Vulliamy *et al.*, 1992; Beutler, 1994; Tishkoff *et al.*, 2001; Verrelli *et al.*, 2002).

1.10.1 Genetic Variation and HIV-1

The influence of genetic variation on an individual's susceptibility to HIV infection and rate of disease progression has been clearly highlighted by the identification and characterization of several AIDS restriction genes (ARGs) (O'Brien and Nelson, 2004). Many of these genes encode products that are involved in viral entry into the cell, immune recognition and antigen presentation, and as a result, polymorphic variations in these genes can have profound effects on host-pathogen interactions (Winkler *et al.*, 2004).

Perhaps the best characterized of these ARGs is the gene encoding CCR5, a major chemokine co-receptor for HIV-1 strain R5. A rare 32bp deletion within the open reading frame of this gene (Δ32) which results in a non-functional protein offers a protective effect against HIV-1 infection. Individuals homozygous for the Δ32 allele are highly resistant to HIV-1 infection and individuals who are heterozygous at this position show delayed disease progression (Dean *et al.*, 1996). The distribution of this allele is restricted to European populations, with its highest frequencies seen in Scandinavian populations (Gonzalez *et al.*, 2001).

Several regulatory polymorphisms, some of which increase an individual's susceptibility to HIV-1 or rate of disease progression, while others confer a protective effect, have subsequently also been identified within this gene as its promoter has been well characterized both in terms of the variation present and underlying haplotype structure (Martin *et al.*, 1998; Carrington *et al.*, 1999; O'Brien and Nelson, 2004). These findings further emphasize the importance of LD patterns and haplotype structure in determining an individual's susceptibility to complex disease.

1.11 Problem Identification and Objectives

African populations, and sub-Saharan populations in particular, are the most genetically diverse in the world and variation within all non-African populations represents only a subset of that present in Africans (Armour *et al.*, 1996; Watson *et al.*, 1997; Zietkiewicz *et al.*, 1997). Furthermore, while distinct differences in variation patterns exist between African and non-African populations (Reich *et al.*, 2001; Gabriel *et al.*, 2002), differences in variation also exist between geographically distinct African populations (Watson *et al.*, 1996; Scozzari *et al.*, 1999; Chen *et al.*, 2000, Salas *et al.*, 2004). Because genetic variation has been shown to influence an individual's susceptibility to disease, an understanding of these inter-population differences could be useful in determining which loci are responsible for the development of disease (Tishkoff and Williams, 2002; Tishkoff and Verrelli, 2003).

Much has been done to identify and characterize variation in non-African populations, such that LD patterns and haplotype structure in these populations are well understood (Reich *et al.*, 2001; Gabriel *et al.*, 2002). However, little work has been done to characterize African specific variation, both within and between distinct populations (Tishkoff and Williams, 2002; Tishkoff and Verrelli, 2003). So much so, that the extent to which African populations differ from each other genetically remains unclear (Lane *et al.*, 2002).

The aim of this investigation was to identify and characterize genetic variation within *PSIP1* in black South Africans. This gene encodes LEDGF/p75, a protein that interacts with HIV-1 integrase (Cherepanov *et al.*, 2003; 2005a; 2005b) and thereby determines the integration site selection for this virus (Llano *et al.*, 2004a; Busshots *et al.*, 2005; MacNeil *et al.*, 2006; Cherepanov, 2007; Marshall *et al.*, 2007). Because of the direct interaction between these two proteins, it has been suggested that alterations in the expression of this protein could influence an individual's susceptibility to HIV infection and/or rate of disease progression (Llano *et al.*, 2004b; Vandegraaff *et al.*, 2006; Zielske and Stevenson, 2006; Vandekerckhove *et al.*, 2006; Llano *et al.*, 2006b).

Genetic variation within this gene was identified and characterized in both HIV$^+$ individuals and individuals whose HIV status was unknown. Genotyping was performed at four polymorphic sites, using one of three genotyping techniques, and LD and haplotype analysis was performed using the genotyping data generated to try and correlate variation patterns with disease susceptibility. Because of the known substructure present in the black South African population (Lane *et al.*, 2002), efforts were also made to identify signatures of population substructure at this locus.

Chapter 2

Materials and Methods

2.1 Samples

Genomic DNA from whole blood samples was used during the course of this investigation to detect and characterize genetic variation within *PSIP1*. All samples were aseptically collected in ethyldiaminetetraacetic acid (EDTA)-containing tubes, under an ethics clearance certificate obtained from the Human Research Ethics Committee at the University of the Witwatersrand (Appendix II). The sample set was comprised of 97 samples collected from HIV^{+} individuals at the Themba Lethu Clinic at Helen Joseph Hospital under informed consent and 39 samples (hereafter referred to as the general population samples), which had previously been collected from staff and students of Wits University whose HIV status is unknown.

For each sample group, affiliation with language groups was determined by asking participants to complete a short questionnaire (Appendix III) that requested details of their place of birth and the home language spoken by the subject, their parents and grandparents. All nine South African Bantu language groups were represented, with 63 % of the sample reporting a single language in three generations. More than one language group occurred in 31% of the sample and 6 % did not know the languages spoken by their relatives. The most common languages spoken were Zulu, followed by Xhosa, Southern Sotho, Tswana, Pedi, Tsonga and the others.

The HIV^+ positive samples were also accompanied by a short patient history that included details about viral load and $CD4^+$ counts, date of diagnosis, history of HIV infection and history of any secondary infections (e.g. tuberculosis or pneumonia).

2.2 DNA Extractions

The HIV^+ whole blood samples used for all genotyping procedures were centrifuged at 2 500 x *g* for ten minutes, separating them into three distinct fractions. Genomic DNA was then extracted from the leukocytes (buffy coat layer) using the QIAmp® DNA Blood Mini Kit, according to the manufacturer's instructions (Qiagen). This kit uses DNA-adsorbing silica-gel membrane spin column technology. The buffy coat layer was treated with proteinase K (20mg/ml) and RNase A (100mg/ml) to remove any protein and RNA contaminants. The leukocytes were lysed with a SDS-containing lysis buffer and the DNA was then precipitated in ethanol (96-100%), eluted in elution buffer (10mM Tris-Cl; 0.5mM EDTA; pH 9.0) and stored at -20°C.

To confirm the success of the extraction procedure, agarose gel electrophoresis was performed using a 1.0% agarose gel, containing 0.3g of agarose in 30ml of 1 x TBE (89mM Tris, 89mM Boric acid and 2mM EDTA) and stained using ethidium bromide (10µg/ml). Gels were electrophoresed in 1 x TBE at 7V/cm for approximately 90 minutes and visualized under UV light. The purity and concentration of each sample was then determined using the NanoDrop ND-1000 Spectrophotometer (ISOGEN).

2.3 Detection of Variation within *PSIP1* by Direct Sequencing

Selected non-coding regions within *PSIP1* were amplified using the polymerase chain reaction (PCR) (Saiki *et al.*, 1988). Portions of the upstream non-coding region and intronic regions of the DNA-binding domain and IBD were selected for sequencing because variations within these regions may disrupt putative regulatory elements and could thus potentially influence gene expression. Additionally, non-coding regions are more informative for the identifying signatures of population substructure and admixture, as these regions are less likely to be subject to selection (Tishkoff and Verrelli, 2003).

For each region, primers were designed (Table 2.3) and reaction conditions were optimized to obtain a single fragment of the desired size, at yields sufficient to perform sequencing reactions. Conditions including the annealing temperature, the annealing- and extension times, DNA- and primer concentrations and the number of cycles used were varied in an attempt to achieve the desired product. PCR products from 20 of the general population samples were then sent to Inqaba BioTec for automated Sanger sequencing (Sanger *et al.*, 1977). The chromatograms obtained were edited and analyzed using Sequencher® version 4.5 (Gene Codes Corporation) and the sequence data obtained for each region was compared with a reference sequence obtained from the Ensembl database (www.ensembl.org) to identify any polymorphisms within these regions. The sequences were also aligned with a corresponding reference sequence from *Pan troglodytes*, also obtained from the Ensembl database (www.ensembl.org), in order to infer the ancestral allele.

2.3.1 Detection of Variation within the Upstream Non-coding Region

A 721bp fragment of the upstream non-coding region of *PSIP1* was amplified in preparation for sequencing using the primers LEDF and LEDGuR (Table 2.3). PCR was performed in a reaction volume of 50µl, containing 2X PyroStart^{TM} Fast PCR Master Mix (0.05u/µl hot start *Taq* DNA polymerase, PCR buffer, 4mM MgCl and 0.4mM of each of the four dNTPs) (Fermentas), 1µM of each of the primers and 20-100ng of template DNA. The amplification reaction consisted of initial denaturation at 95.0°C for 60 seconds, followed by 38 cycles of denaturation at 95.0°C for 1 second, annealing at 55.0°C for 8 seconds and extension at 72.0°C for 50 seconds, with a final extension at 72.0°C for 10 seconds.

To ensure a fragment of the correct size was obtained in quantities sufficient for sequencing, the PCR product was electrophoresed on a 1% agarose gel in 1X TBE buffer at 7V/cm for 45 minutes. The remainder of the PCR product was then purified and sequenced in both directions by Inqaba Biotec, using the sequencing primers LEDGF-INTL ($^{5'}$CCCTTCGCATTTTGCATT$^{3'}$) and LEDGF-INTR ($^{5'}$TCCCCAAGTTCGCTTTA$^{3'}$).

2.3.2 Detection of Variation within the DNA-Binding Domain

A 566 bp fragment of intron 5 of *PSIP1* was amplified in preparation for sequencing using the primers SeqIF5-6 and SeqIR5-6 (Table 2.3). PCR was performed in a reaction volume of 50µl, containing 2X PCR Master Mix (0.05u/µl *Taq* DNA polymerase, PCR buffer, 4mM MgCl and 0.4mM of each of the four dNTPs) (Fermentas), 1µM of each of the primers and 20-100ng of template DNA. The amplification reaction consisted of initial denaturation at 94.0°C for 5 minutes, followed by 35 cycles of denaturation at 94.0°C for 30 seconds, annealing at 61.0°C for 45 seconds and extension at 72.0°C for 90 seconds, with a final extension at 72.0°C for 5 minutes.

To ensure only a fragment of the correct size was obtained, the PCR product was electrophoresed on a 1% agarose gel in 1X TBE buffer at 7V/cm for 40 minutes. The remainder of the PCR product was then purified and sequenced in the forward direction by Inqaba Biotec, using the primer SeqIF5-6 (Table 2.3).

2.3.3 Detection of Variation within the Intronic Regions of the IBD

A 970bp fragment of intron 12 of *PSIP1* had previously been amplified by Miss Daniella Grantcharov using the primers INTL and INTR (Table 2.3) and sequenced in both directions by Inqaba Biotec. A total of 52 chromatograms were available for analysis. These were edited, analyzed and aligned with a reference sequence obtained from the Ensembl Genome Browser (www.ensembl.org) using Sequencher® version 4.5 (Gene Codes Corporation), in order to identify any possible polymorphisms.

Table 2.3: Primers used to amplify the three regions of *PSIP1* selected for sequencing. All primers were designed using the web-based tool, Primer3 (Rozen *et al.*, 2000) and subjected to a BLAST search (Altschul *et al.*, 1990) to ensure their target specificity.

Primer Name:	**Sequence:**	**T_m[1] (°C):**	**GC Content (%):**	**Fragment Size (bp):**
LEDF	GCCCAAACTCACATCCTATCTAAA	61.15	41.67	
LEDGuR	CGACCAACTGTTTACCGAGAGA	62.67	50.00	721
SeqIF5-6	CAGTACCAACTGCTGCCTCA	62.45	55.00	
SeqIR5-6	GCACTCAAAGTTTAATTCGATGG	59.20	39.13	566
INTL[2]	CACTGCATGTTGCTTTTCTCA	58.66	42.86	
INTR[2]	CAGTCCTGGCAAATGGTTTA	58.35	45.00	970

[1] T_m represents the melting temperature of the primer.

[2] Primers which had previously been available.

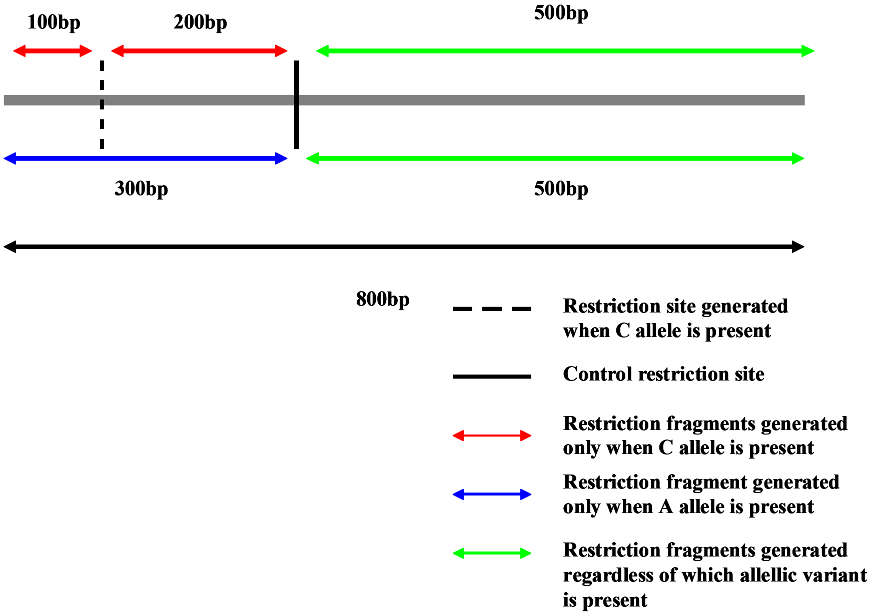
2.4 Genotyping

Depending on the nature of the polymorphism in question, one of three genotyping methods (RFLP-PCR, allele-specific PCR and Pyrosequencing™) was employed to establish the allele and genotype frequencies of these polymorphisms within the sample set.

2.4.1 PCR-RFLP

Polymorphisms which result in either the introduction or abolition of the recognition sequence of a type II restriction enzyme provide a means whereby these polymorphisms may be genotyped, as only one of the possible allelic variants will render the sequence resistant to digestion by the enzyme. Following PCR amplification of the region of interest, subsequent digestion of the PCR product with the given enzyme will result in fragments of different sizes, depending on which of the allelic variants are present. In this way, a distinction can then be made between homozygous and heterozygous individuals based on the results of agarose gel electrophoresis (Figure 2.4.1) (Deng, 1989). Another restriction site, which is present regardless of the genotype at the polymorphic site in question, is usually included in the assay design. This serves as a control for the assay, confirming the efficacy of the restriction enzyme.

A.


100bp
200bp
500bp
300bp
500bp
800bp
Restriction site generated when C allele is present
Control restriction site
Restriction fragments generated only when C allele is present
Restriction fragment generated only when A allele is present
Restriction fragments generated regardless of which allellic variant is present

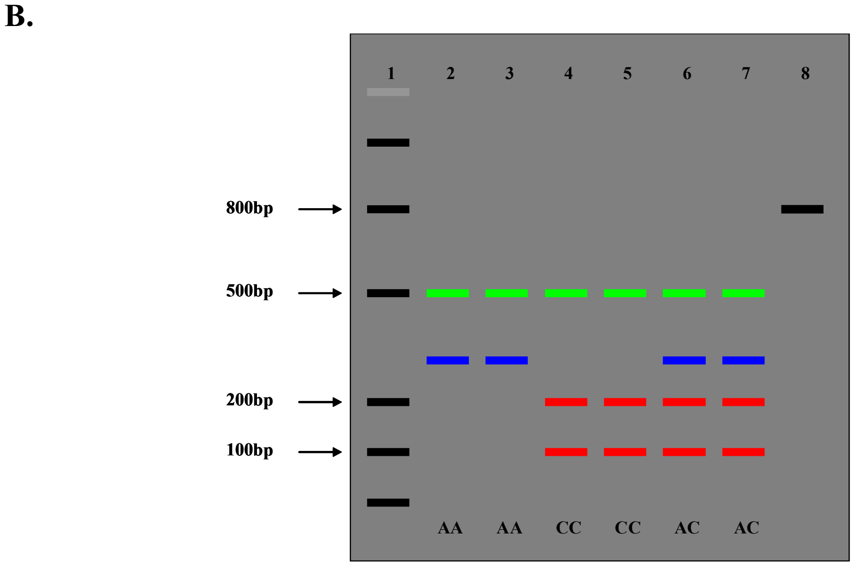


Figure 2.4.1: A schematic representation of a RFLP-PCR assay. (A) Different restriction fragments will be generated based on which of the two polymorphic variants is present. (B) The patterns that should be detected by gel electrophoresis after digestion with the restriction enzyme of choice. In this way, the different genotypes can be identified by the different electrophoretic patterns they produce.

2.4.1.1 Genotyping of the 5bp Deletion within Intron 13 of the IBD

A 621 bp fragment of intron 13 was amplified in preparation for restriction digestion using the primers IBDL ($^{5'}$GCATGTTGCTTTTCTCAACCAC$^{3'}$) and IBDR ($^{5'}$ACAAAATTCAAAGAATCCACATGAC$^{3'}$). PCR was performed in a reaction volume of 20µl, containing 2X PCR Master Mix (0.05u/µl *Taq* DNA polymerase, PCR buffer, 4mM MgCl and 0.4mM of each of the four dNTPs) (Fermentas), 1µM of each of the primers and 20-100ng of template DNA. The amplification reaction consisted of initial denaturation at 94.0°C for 5 minutes, followed by 35 cycles of denaturation at 94.0°C for 30 seconds, annealing at 55.0°C for 45 seconds and extension at 72.0°C for 90 seconds, with a final extension at 72.0°C for 5 minutes.

A restriction digest was then performed, in order to determine the genotype at this position (Figure 2.4.1.1). Digestion was performed in a 30µl reaction volume; consisting of 15µl of PCR product, 10x Buffer R (Fermentas) and 5U *Mbo*I, incubated for 14 hours at 35°C. The digestion product was then electrophoresed on a 4% agarose gel in 1X TBE buffer at 7V/cm for 3 hours and genotypes were recorded based on the electrophoresis patterns observed.

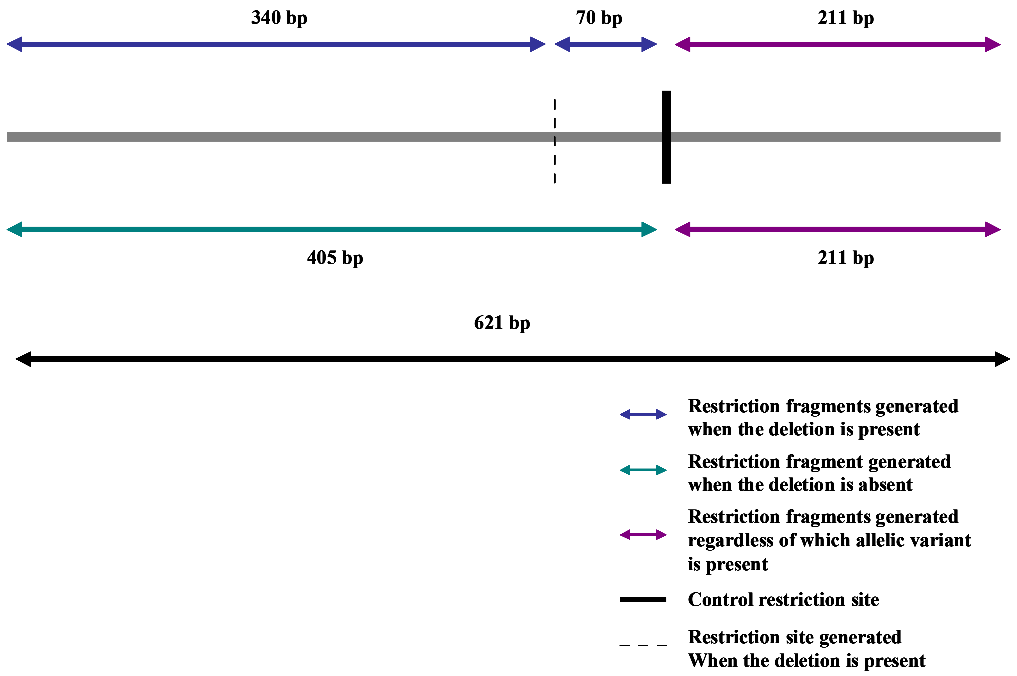
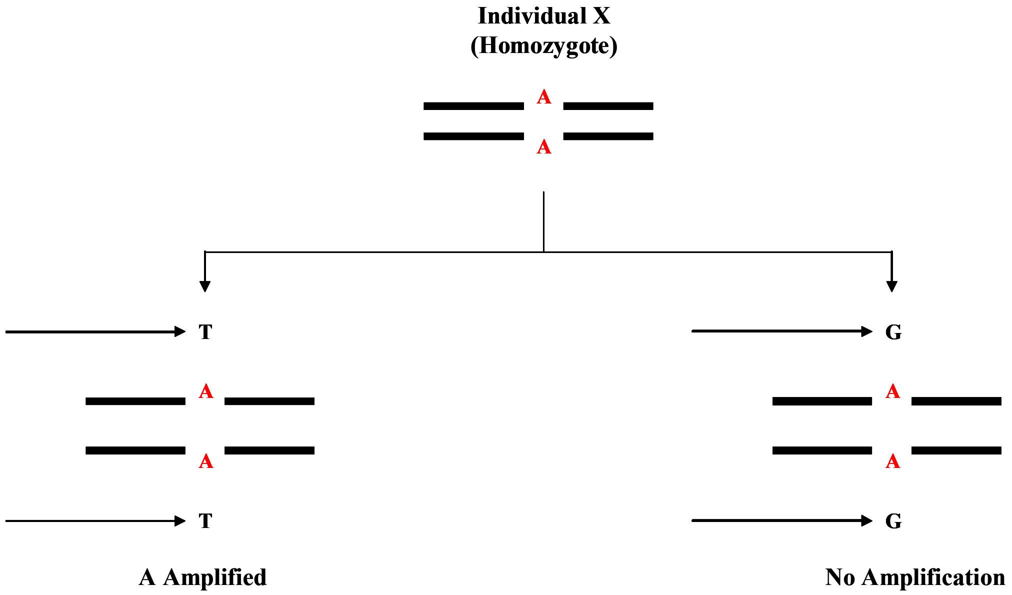

340 bp
70 bp
211 bp
405 bp
211 bp
621 bp
Restriction fragments generated when the deletion is present
Restriction fragment generated when the deletion is absent
Restriction fragments generated regardless of which allelic variant is present
Control restriction site
Restriction site generated When the deletion is present

Figure 2.4.1.1: The restriction map generated when a 5bp deletion at position + 41 796 is both present and absent in a 621bp fragment of the IBD. When the deletion is present, an additional *Mbo*I restriction site is introduced and a different restriction map is generated.

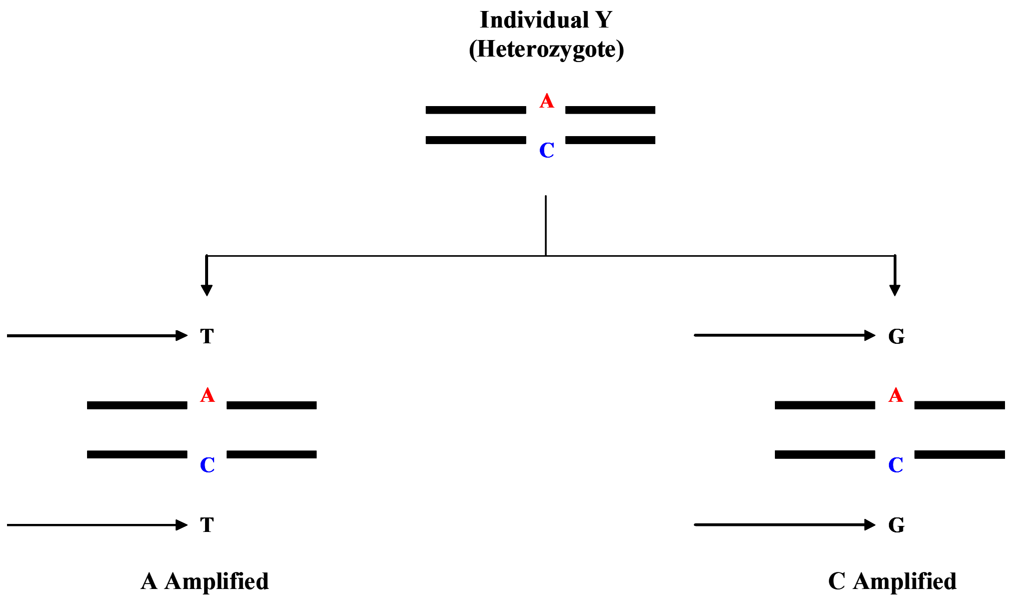
2.4.2 Allele-specific PCR

This technique involves the use of three primers (two forward primers and a common reverse primer, or vice versa); rather than two, as is necessary for conventional PCR. The two forward (or reverse) primers are designed to differ by a single base at the final or penultimate positions of their 3' ends, so that each primer can detect and amplify only one of two possible allelic variants. In each case, if the allele the primer is designed to detect is present, amplification of the desired product will occur; whereas if the allele the primer is designed to detect is not present in a given sequence, a mismatch will result and little or no amplification will be observed. Genotypes may then be scored based on the presence or absence of a PCR product following electrophoresis (Figure 2.4.1.2) (Newton *et al.*, 1989; Wu *et al.*, 1989).

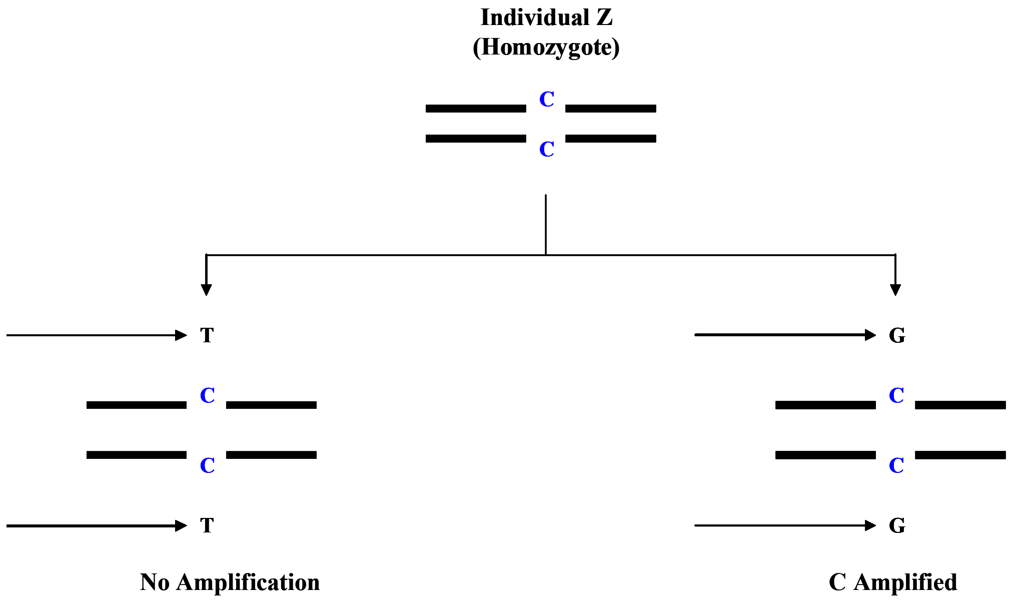
A.



B.


Individual Y
(Heterozygote)
A
C
T
A
C
T
A Amplified
G
A
C
G
C Amplified

C.


Individual Z
(Homozygote)
C
C
T
C
C
T
No Amplification
G
C
C
G
C Amplified

D.

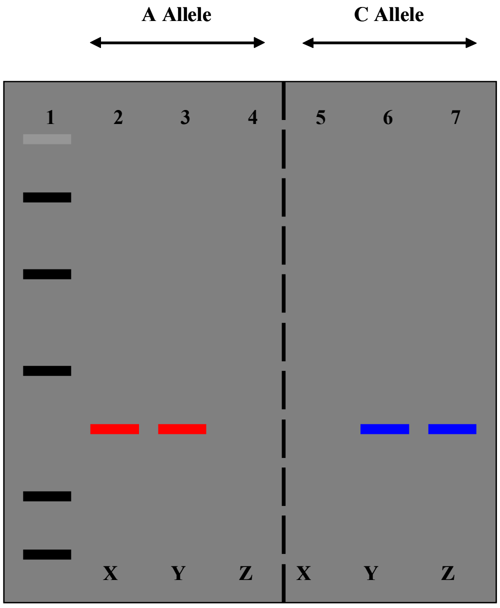


Figure 2.4.2: A schematic representation of the principles involved in the design of an allele-specific PCR assay. (A-C) In each case, two separate reactions are performed – each with a different allele-specific primer. In this way, only the allelic variant corresponding to the primer used will be amplified. (D) The genotyped may then be identified based on the presence or absence of a PCR product.

2.4.2.1 Genotyping of the Insertion within the Upstream Non-coding Region

Initially, an assay using two allele-specific forward primers (UpInW and UpInM) and a common reverse primer (LEDGuR) (Table 2.4.2) was designed to genotype an insertion within the promoter region of *PSIP1* at position -417. In attempting to optimize the two reactions, parameters such as the annealing temperature, the annealing- and extension times, primer concentrations and the number of cycles used were repeatedly varied. All optimization procedures were performed using samples of known genotype as a control. However, because optimal specificity of these reactions could not be achieved, it was later deemed necessary to redesign the assay.

Rather, a new assay involving two allele-specific reverse primers (UpstrmWT and UpstrmMT) and a common forward primer (LEDGuF) (Table 2.4.2) was designed. Genotyping reactions were performed in a reaction volume of 10µl, containing 2X PCR Master Mix (0.05u/µl *Taq* DNA polymerase, PCR buffer, 4mM MgCl and 0.4mM of each of the four dNTPs) (Fermentas), 1µM of each of the primers and 20-100ng of template DNA. Both reactions were performed under identical cycling conditions - namely initial denaturation at 94.0°C for 5 minutes, followed by 30 cycles of denaturation at 94.0°C for 30 seconds, annealing at 64.5°C for 45 seconds and extension at 72.0°C for 90 seconds, with a final extension at 72.0°C for 5 minutes and both resulted in the formation of a 415bp fragment. The PCR product was then electrophoresed on a 1% agarose gel in 1X TBE buffer at 7V/cm for 40 minutes.

2.4.2.2 Genotyping of SNPs within Intron 5 of the DNA-Binding Domain

Initially an attempt was made to genotype two adjacent SNPs within intron 5 by means of two separate allele-specific assays. In the case of the SNP at position +31 040, two allele-specific forward primers (31 040T and 31 040C) were designed in conjunction with a common reverse primer (31 040R) (Table 2.4.2); while in the case of the SNP at +31 041, two allele-specific reverse primers (31 041T and 31 041C) were designed with a common forward primer (31 041F) (Table 2.4.2). The assay for genotyping the SNP at position +31 041 was later redesigned to include a new allele-specific primer (041T – $^{5'}$TAAAAATAAAGCTAATATTCTTGATGCA$^{3'}$), designed to replace 31 041T.

Parameters including the annealing temperature, the annealing and extension times, primer concentrations and the number of cycles used were repeatedly varied in an effort to optimize the conditions for both these reactions. Similarly, all optimization procedures were performed using samples of known genotypes. However, once again, it was determined that this technique was unsuitable for genotyping these particular SNPs and that an alternative technique, namely Pyrosequencing™, would have to be explored.

2.4.2.3 Genotyping of a 2bp Deletion within Intron 12 of the IBD

An allele-specific assay was also designed to genotype a 2bp deletion at position +41 780 within intron 12 of the IBD. Because of the presence of another 5bp deletion within 20bp of this polymorphism, the only option available was to design the reverse primers (IBDd and IBDtt) to be allele-specific, with a common forward primer (INTR) (Table 2.4.2). The aforementioned PCR parameters were varied and optimization procedures were performed using samples of known genotype; but again, this technique was found to be unsuitable for genotyping this particular polymorphism. Unfortunately, because of the presence of other polymorphisms in the region surrounding this deletion, alternative primers could not be designed and genotyping of this site could ultimately not be performed.

Table 2.4.2: Primers used for allele-specific genotyping. All primers were designed using the web-based tool, Primer3 (Rozen *et al.*, 2000) and subjected to a BLAST search (Altschul *et al.*, 1990) to ensure their target specificity.

Primer Name:	Sequence:	T_m[1] (°C):	GC Content (%):	Fragment Size (bp):
UpInW	AAACCTCCCCACCCTGGA	62.68	61.11	414
UpInM	AACCTCCCCACCCTGGG	63.17	70.59	414
LEDGuR	CGACCAACTGTTTACCGAGAGA	62.67	50.00	
UpstmWT	GATTCATGTTCTTGTATCGTTTCCA	59.66	36.00	415
UpstmMT	ATTCATGTTCTTGTATCGTTTCCC	59.44	37.50	415
LEDGuF	ACATTGTACCACCTACCAGCTCCT	64.57	50.00	
31040T	GTGTAATCACATACTTTGTTCTCCATAT	60.22	32.14	387
31040C	GTGTAATCACATACTTTGTTCTCCATAC	61.69	35.71	387
31040R	GCAGGTCGTCCTCTTTTAGG	62.45	55.00	
31041T	AAATAAAGCTAATATTCTTGATGCA	54.74	24.00	417
31041C	ATAAAGCTAATATTCTTGATGCG	55.64	30.43	415
31041F	CCCATCTCCTCCTTTGTCT	62.45	55.00	
IBDtt	CAGTACTGCATTTATAGCTTCATCTTTT	60.22	32.14	666
IBDd	CAGTACTGCATTTATAGCTTCATCTTA	60.22	25.93	665
INTR	CAGTCCTGGCAAATGGTTTA	58.35	45.00	

[1] T_m represents the melting temperature of the primer.

2.4.3 Pyrosequencing^{TM}

Pyrosequencing^{TM} (Ronaghi *et al.*, 1996) is a sequencing method based on the real-time detection of the pyrophosphate (PPi) released during the synthesis reaction catalyzed by DNA polymerase. Automated solid-phase Pyrosequencing™ (Ronaghi *et al.* , 1998) employs a four-enzyme system that involves Klenow DNA polymerase from *Escherichia coli*, ATP sulfurylase from *Saccharomyces cerevisiae* (yeast), luciferase from *Photinus pyralis* (the North American firefly) and apyrase from *Solanum tuberosum* (potato tubers).

An pre-amplified PCR product is generally first rendered single-stranded (although double-stranded PCR product may also serve as a template for Pyrosequencing™) and incubated in a microtiter plate (which is under constant agitation) with these four enzymes, APS (the substrate for ATP sulfurylase), D-luciferin (the substrate for luciferase), as well as a short sequencing primer. Each of the four deoxynucleotides is then sequentially added to the reaction mixture in a predetermined order by an inkjet cartridge. When the correct nucleotide is added to reaction mixture the primer is extended by DNA polymerase and PPi (of equal molarity to the incorporated deoxynucleotide) is released (Ronaghi *et al.*, 1998).

$$\text{(a)}\qquad (\text{DNA})_n + \text{dXTP} \xrightarrow{\text{DNA polymerase}} (\text{DNA})_{n+1} + \text{PPi}$$

This PPi is then converted to ATP by the sulfurylase (Ronaghi *et al.*, 1998).

$$\text{(b)}\qquad \text{PPi} + \text{APS} \xrightarrow{\text{ATP sulfurylase}} \text{ATP} + \text{SO}^{2-}_4$$

The ATP generated by this reaction is then utilized by luciferase to produce light (Ronaghi *et al.*, 1998),

$$\text{(c)} \qquad \text{ATP + luciferin + O}_2 \xrightarrow{\text{luciferase}} \text{AMP + PPi}$$

which is detected by a CCD camera and used by computer software to generate a pyrogram. ATP is also utilized by apyrase to degrade any non-incorporated deoxynucleotides (Ronaghi *et al.*, 1998).

$$\text{(d)} \qquad \text{ATP + dXTP} \xrightarrow{\text{apyrase}} \text{AMP and dXMP + 4Pi}$$

2.4.3.1 Genotyping of the SNPs within Intron 5 of the DNA-Binding Domain

Pyrosequencing™ was used to genotype the adjacent SNPs at positions +31 040 and +31 041. A 227bp fragment was amplified by PCR, using PSIPfor ($^{5'}$GACGGGGACACCGCTGCTCGTTTATGTGTTAGTTGCAGTGTAATCACA$^{3'}$) and PSIPrev ($^{5'}$GTCTATGGTAACGTTGAGTTCAAG$^{3'}$). Both primers were designed by Dr. Zane Lombard using PSQ™ Assay Design Software. The forward primer was designed to include an additional 23bp oligonucleotide tag, which is complementary to the sequence of a universal biotin-labeled primer (Aydin *et al.*, 2005). A sequencing primer (PSIPseq – $^{5'}$AAAGCTAATATTCTTGATGC$^{3'}$), positioned immediately adjacent to the SNPs of interest, was also designed by Dr. Lombard.

PCR was performed in a reaction volume of 50µl, with a reaction mixture consisting of 2X PCR Master Mix (0.05u/µl *Taq* DNA polymerase, PCR buffer, 4mM MgCl and 0.4mM of each of the four dNTPs) (Fermentas), 0.2µM of both the reverse and universal primers, 0.02µM of the tagged forward primer and

20-100ng of template DNA. Optimized cycling conditions involved initial denaturation at 94.0°C for 5 minutes, followed by 30 cycles of denaturation at 94.0°C for 30 seconds, annealing at 60.0°C for 45 seconds and extension at 72.0°C for 90 seconds, with a final extension at 72.0°C for 5 minutes.

A total volume of 40µl of PCR product was required for each Pyrosequencing™ reaction. The PCR products were immobilized to streptavidin sepharose beads in the presence of binding buffer (10mM Tris-HCl, 2M NaCl, 1mM EDTA, 1% Tween 20), before strand separation was performed by transferring the templates between 70% ethanol, denaturation solution (0.2M NaOH) and washing solution (10mM Tris-Acetate, pH 7.6). Sequencing primer annealing was then performed by heating the templates and primer at 80°C in the presence of annealing buffer (20mM Tris-Acetate, 2mM Mg-Acetate) for 3 minutes.

Sequencing was performed with the PSQ^{TM} 96MA Instrument (Pyrosequencing AB), using the PSQ™ 96 SNP Reagent Kit (Biotage). Sequencing was performed at 28°C in a volume of 50µl. Because the sequence surrounding the SNPs was known, deoxynucleotides were added sequentially and the sequencing primer was extended for only 5 bases. Pyrograms were generated and genotypes were detected using PSQ™ 96MA SNP v.2.1 software (Pyrosequencing AB). All computationally derived and any ambiguous genotypes were confirmed by manual base-calling.

2.5 Data Analysis

2.5.1 Estimation of Gene Frequencies by Gene Counting

For each biallelic, polymorphic locus under investigation, the frequencies of each of the three possible genotypes were determined as a proportion of the total sample size. These genotype frequencies were then used to estimate the allele frequencies at each of the polymorphic loci (Ceppellini *et al.*, 1955).

In this way, assuming x, y and z represent the number of individuals carrying the genotypes A_1A_1, A_2A_2 and A_1A_2 in a population of size of n; the frequencies of each of the genotypes may then be calculated as:

$$\text{Frequency of genotype } A_1A_1 = \frac{x}{n}$$

$$\text{Frequency of genotype } A_2A_2 = \frac{y}{n}$$

$$\text{Frequency of genotype } A_1A_2 = \frac{z}{n}$$

Thus in a population of size n, where each individual carries two alleles; the frequencies of alleles A_1 and A_2 may be calculated from the genotype frequencies using the formulae:

$$\text{Frequency of allele } A_1 = \frac{2x + z}{2n}$$

and

$$\text{Frequency of allele } A_2 = \frac{2y + z}{2n}$$

2.5.2 Test for Hardy-Weinberg Equilibrium

Independently published calculations by Hardy and Weinberg demonstrated that within populations of sexually reproducing diploid species, genotype frequencies reach equilibrium after one generation of random mating and fertilization - provided selection and migration have no effect on the genotypes in question; and these equilibrium frequencies persist throughout several generations unless a force powerful enough to alter the allele frequencies arises (Crow, 1986).

Thus in a population in Hardy-Weinberg equilibrium, at a given biallelic, polymorphic locus the frequencies of alleles A_1 and A_2 (hereafter denoted as p and q, respectively) can be used to estimate the expected frequencies of the three possible genotypes A_1A_1, A_2A_2 and A_1A_2; which will then be given by p^2, q^2 and 2pq, respectively (Crow, 1986), from

$$p^2 + 2pq + q^2$$

In order to establish if the population deviated significantly from Hardy-Weinberg equilibrium at the polymorphic positions under investigation, a χ^2 test for goodness-of-fit was performed to compare the genotype frequencies observed within the population with those expected under Hardy-Weinberg conditions:

$$\chi^2 = \frac{\Sigma\,(o - e)^2}{e}$$

where *o* represents the observed genotype number and *e* represents the expected genotype number.

The tests were performed with 1 degree of freedom and at a significance level of $\alpha = 0.05$. Thus for values of $p<0.05$, the null hypothesis is rejected and the data is not seen to deviate significantly from Hardy-Weinberg equilibrium (Crow, 1986).

When multiple individual χ^2 tests were performed, a sequential Bonferroni test (Rice, 1988) was subsequently performed. This technique makes allowances for the fact that when multiple tests are performed, the significance level selected must be adjusted accordingly to control the overall type-1 error rate.

2.5.3 Estimation of linkage disequilibrium

Pairwise LD refers to the non-random association of alleles at two separate loci on the same chromosome (Wall and Pritchard, 2003). Consider two loci - each with two alleles (A_1, A_2, B_1 and B_2), with frequencies p_1, q_1, p_2 and q_2, respectively. If these result in four gametic types, namely A_1B_1, A_1B_2, A_2B_1 and A_2B_2, with frequencies g_1, g_2, g_3 and g_4, respectively then the linkage disequilibrium parameter D may be calculated using the formula (Lewontin, 1988):

$$D = g_1 - p_1p_2$$

If $D = 0$ then the alleles at the two loci are randomly associated; but if $D > 0$ or $D < 0$, the association between the alleles at the two loci is non-random and they are said to be in linkage disequilibrium (Lewontin, 1988; Weiss and Clark, 2002).

However, D is very heavily dependent on allele frequency. So much so that the largest value D can take (D_{max}) is the smaller of p_1q_2 or p_2q_1 if D is positive or the smaller of p_1p_2 or q_1q_2 if D is negative (Lewontin, 1988; Weiss and Clark, 2002). Thus Lewontin's co-efficient (D') (Lewontin, 1964), given by:

$$D' = \frac{D}{D_{max}}$$

is considered to be a more robust measure for quantifying linkage disequilibrium, as this parameter is less dependent on allele frequency (Lewontin, 1988).

An additional measure of LD is Pearson's correlation (r^2), which is calculated from the formula:

$$r^2 = \frac{D^2}{p_1q_1p_2q_2}$$

Under a standard model of selectively neutral evolution, the expected value of r_2 is $1/(4Nc + 1)$, where N is the effective population size and *c* is the recombination rate between the two loci (Pritchard and Przeworski, 2001). Additionally, the sample size required to detect statistically significant LD is inversely proportional to r^2 (Pritchard and Przeworski, 2001). Thus this parameter not only provides a means whereby LD may be quantified, but also supplies useful information about population history and the significance of the data itself.

Calculation of D, D' and r^2 values was performed using Linkage Disequilibrium Analyzer (LDA) version 1.0 (Ding *et al.*, 2003).

2.5.4 Haplotype Analysis

While LD analysis is a useful tool for studying patterns of genetic variation, the picture presented by the data can often be noisy and erratic. A better impression can be obtained by identifying the underlying haplotype structure of the chromosomal region of interest (Daly *et al.*, 2001). Haplotypes can be defined as consecutive sites between which there is little or no evidence of historical recombination, as determined by calculation of pairwise LD (Gabriel *et al.*, 2002; Wall and Pritchard, 2003).

Haplotype analysis was performed using the software package, PHASE 2.1 (Stephens *et al.*, 2001), which uses a coalescence-based Markov-chain Monte Carlo approach, based on a pseudo-Gibbs sampler, to statistically infer phase and reconstruct haplotypes from genotyping data (Niu, 2004). Only samples with complete genotyping data at all four polymorphic positions were used for haplotype analysis. In cases where individuals were heterozygous at more than one position, the most common haplotype pairing was selected as the assigned haplotype pair. Haplotype phase could not be determined for one of the HIV$^+$ samples, which was excluded from further analysis.

Chapter 3

Results

3.1 Detection of Variation in *PSIP1* by Direct Sequencing

3.1.1 Detection of Variation within the Upstream Non-coding Region

Direct sequencing of a 721bp fragment (Figure 3.1.1) of the upstream non-coding region of *PSIP1* in twenty of the general population samples in both the forward and reverse directions led to the identification of a single variable site, the insertion of a G at position -417. Only one individual was found to be homozygous for the insertion, while ten individuals were identified as being heterozygous at this position. The minor allele was fairly frequent. The G allele was present in the *Pan troglodytes* sequence and is therefore likely to be the ancestral allele.

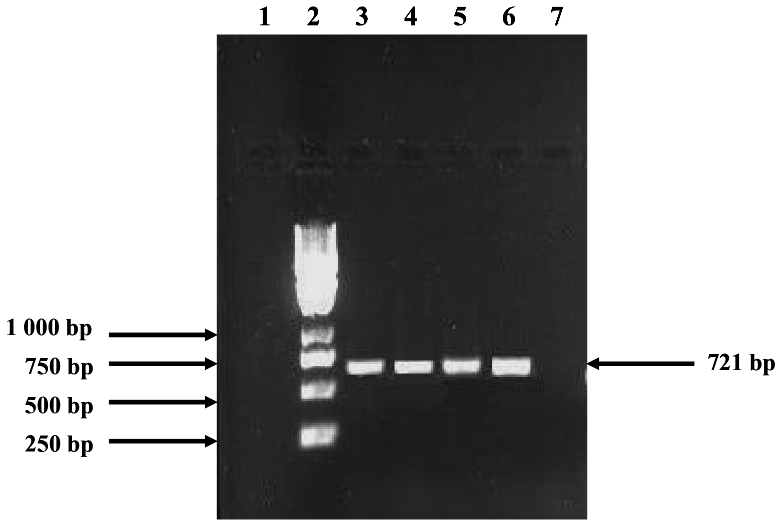


Figure 3.1.1: PCR amplification of the upstream non-coding region in preparation for sequencing. The resulting 721bp fragment was visualized on a 1% agarose gel. Lanes 3-6 show the amplified PCR product, while lane 2 is a GeneRuler^{TM} 1kb DNA ladder (Fermentas) and lane 7 shows the negative "no DNA" control.

3.1.2 Detection of Variation within the DNA-Binding Domain

Direct sequencing of a 566bp fragment of the DNA-binding domain (Figure 3.1.2.1) revealed the presence of seven variable sites, with minor allele frequencies ranging between 0.03 and 0.48 (Table 3.1). Two of the sites identified (at positions +31 040 and +31 041) were located immediately adjacent to each other (Figure 3.1.2.2). While both were present as T/C variants, C was the minor allele at position +31 040, while T was found to be the minor allele at position +31 041. No individuals were homozygous for the minor allele at positions +30 816, +30 846, +30 903 and +31 041, while only one individual was homozygous for the minor allele at position +31 040. The ancestral allele corresponded to the major allele at positions +30 816, +30 846, +30 903 and +31 040; and corresponded to the minor allele at positions +30 830, +31 041 and +31 097.

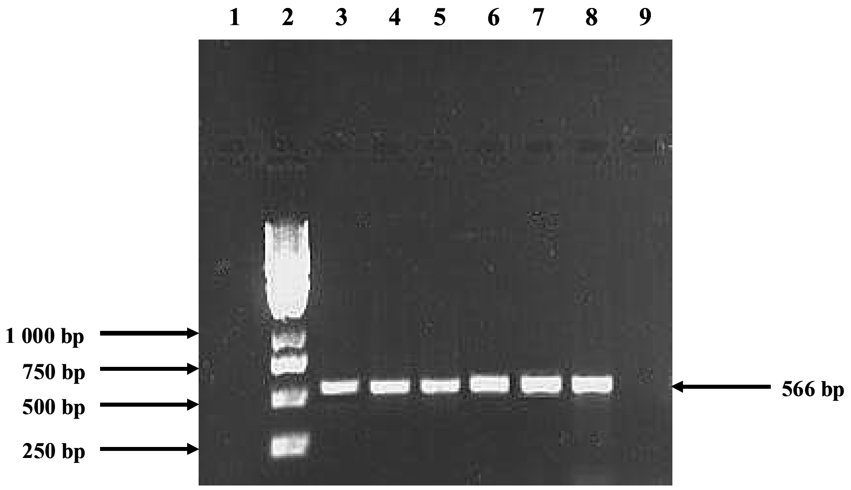


Figure 3.1.2.1: PCR amplification of the DNA-binding domain in preparation for sequencing. The resulting 566bp fragment was visualized on a 1% agarose gel. Lanes 3-8 show the amplified PCR product, while lane 2 is a GeneRuler^{TM} 1kb DNA ladder (Fermentas) and lane 9 shows the negative “no DNA” control.

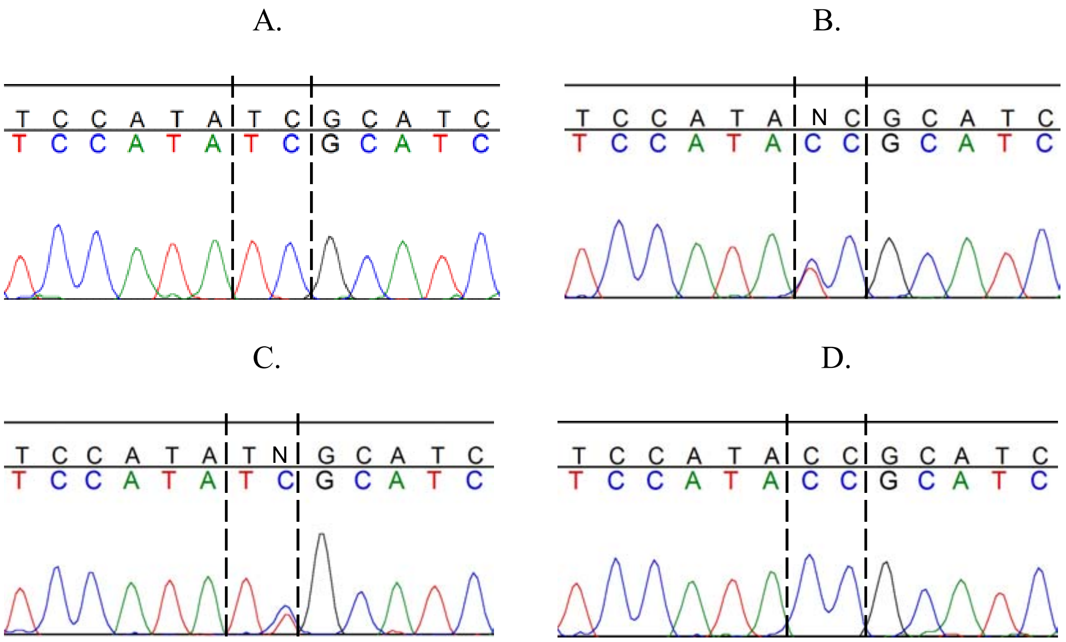


Figure 3.1.2.2: Chromatograms showing two adjacent SNPs present within the DNA-binding domain. The different genotypes can be identified based on the different fluorescence patterns generated by the sequencing software. (A) An individual homozygous for the major alleles at both loci. (B) An individual heterozygous at the first locus and homozygous for the major allele at the second. (C) An individual homozygous for the major allele at the first locus and heterozygous at the second. (D) An individual homozygous for the minor allele at the first locus and homozygous for the major allele at the second.

3.1.3 Detection of Variation within the IBD

Analysis of sequencing data obtained for a 970bp fragment of the IBD revealed the presence of a single SNP (+42 357 (C/T) and two deletions (at positions +41 780 and +41 796), located within 20bp of each other (Table 3.1). Both deletions were absent in the *Pan troglodytes* sequence, while the C was found to be the ancestral allele at the SNP position. No individuals were homozygous for the minor allele at any of the three positions. In the case of the SNP, the minor allele was present in four heterozygous individuals, while the minor allele was only present in one heterozygous individual for both of the deletions.

Table 3.1: The genetic variation detected within *PSIP1* through direct sequencing. Genotype and minor allele frequencies for each of the sites are given, as well as the χ^2- and P values for the goodness of fit of the data to Hardy-Weinberg equilibrium.

Polymorphism Position [1]	Genotype	n	Genotype Frequency	Minor Allele Frequency
- 417	-/- -/G G/G	9 10 1	0.45 0.50 0.05	0.30
+ 30 816	G/G G/C C/C	12 8 0	0.60 0.40 0.00	0.20
+ 30 830	G/G G/A A/A	5 12 3	0.25 0.60 0.15	0.45
+30 846	A/A A/C C/C	19 1 0	0.95 0.05 0.00	0.03
+30 903	G/G G/C C/C	18 2 0	0.90 0.10 0.00	0.05
+31 040	T/T T/C C/C	15 4 1	0.75 0.20 0.05	0.15
+31 041	C/C C/T T/T	16 4 0	0.80 0.20 0.00	0.10
+31 097	A/A A/T T/T	4 13 3	0.20 0.65 0.15	0.48
+41 780	TT/TT TT/- -/-	24 3 0	0.89 0.11 0.00	0.06
+41 796	TCTTA/TCTTA TCTTA/- -/-	26 1 0	0.96 0.04 0.00	0.02
+42 357	C/C C/T T/T	18 4 0	0.82 0.18 0.00	0.09

[1] Positions are given relative to the start of transcription.

3.2 Genotyping

3.2.1 Genotyping of the Insertion within the Upstream Non-coding Region

Allele-specific PCR was used to genotype the insertion at position -417 (Figure 3.2.1) in 136 samples (Appendix I). Two separate reactions were performed, each designed to amplify only one of the possible allelic variants at this position. Genotypes were then assigned based on the presence or absence of a PCR product following each reaction. Reaction conditions were optimized using samples of known genotype (based on the results of direct sequencing) and in each case, a "no DNA" negative control was included to preclude any false positives as a result of DNA contamination. As an additional control, the procedure was repeated using thirty randomly-selected samples (approximately 20% of the total sample size), to confirm the accuracy and specificity of both genotyping reactions. The results obtained confirmed that the previously recorded genotypes were accurate.

A total of sixteen individuals were found to be homozygous for the minor allele, while 52 individuals were heterozygous at this position. The minor allele frequency at this position was 0.31 (Table 3.2), a value similar to that obtained from analysis of the direct sequencing data. The population sample did not deviate significantly from Hardy-Weinberg equilibrium at this position (Table 3.2).

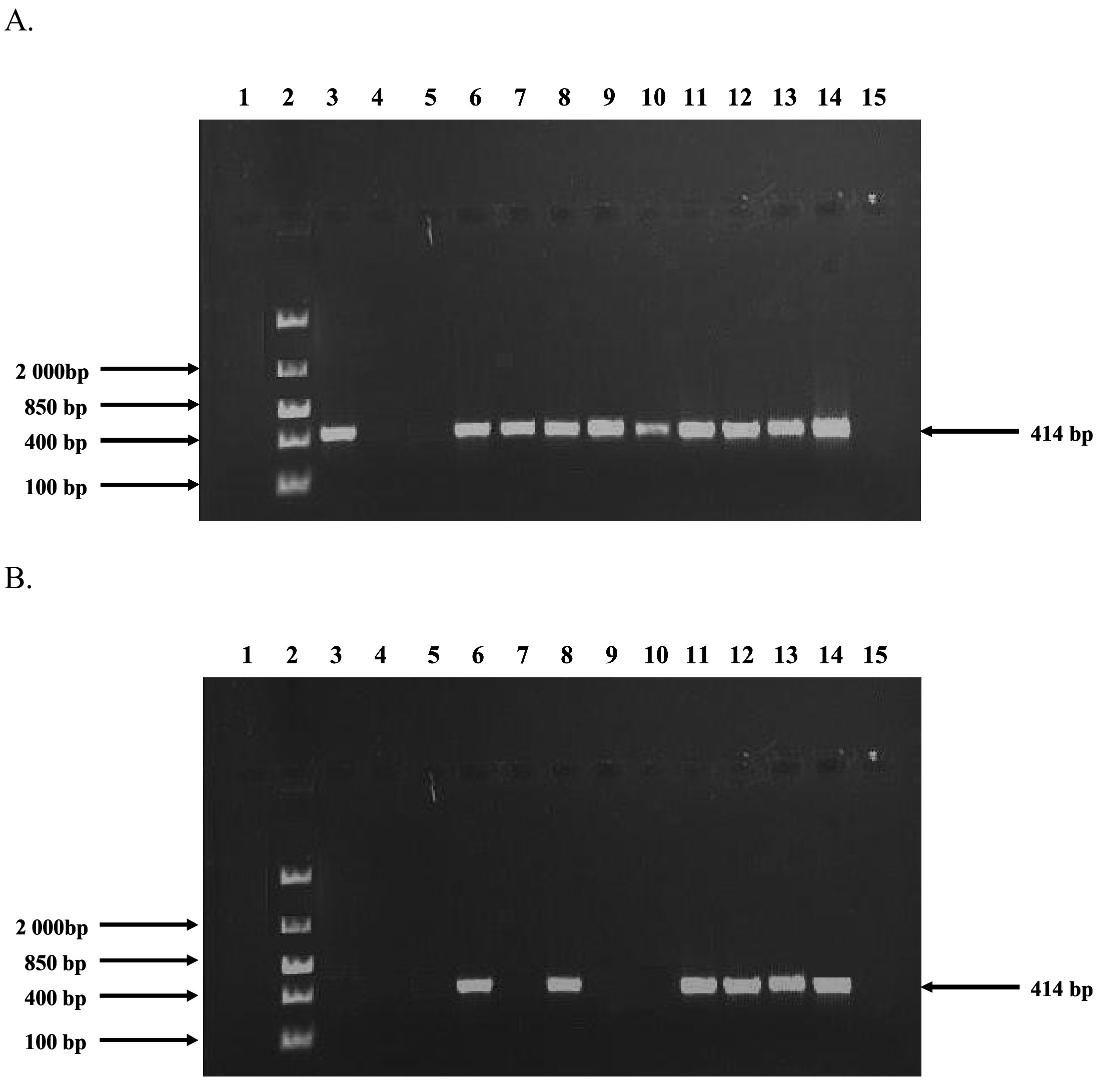


Figure 3.2.1: Allele-specific PCR amplification of an insertion within the upstream non-coding region. Two separate reactions are performed, each resulting in the amplification of one of the two allelic variants. Genotypes were resolved based on the presence or absence of a 414bp PCR product in each or both of the reactions. The products of both reactions were visualized on a 1% agarose gel. (A) The PCR product produced when the insertion is absent. (B) The PCR product produced when the insertion is present. (A and B) lanes 3-14 show the results of the respective allele-specific PCR reactions, while lane 15 shows the "no-DNA" negative control and lane 2 is a FastRuler™ Middle Range DNA Ladder (Fermentas).

3.2.2 Genotyping of the 5bp Deletion within the IBD

A RFLP-PCR assay was designed to genotype the 5bp deletion at position +41 796 within the IBD. First, conventional PCR was used to amplify a 621bp fragment of the IBD (Figure 3.2.2.1). Then, a restriction digest was performed to confirm whether or not the deletion was present, based on the presence or absence of an additional *Mbo*I restriction site. Genotypes were assigned based on the restriction profile obtained (Figure 3.2.2.2).

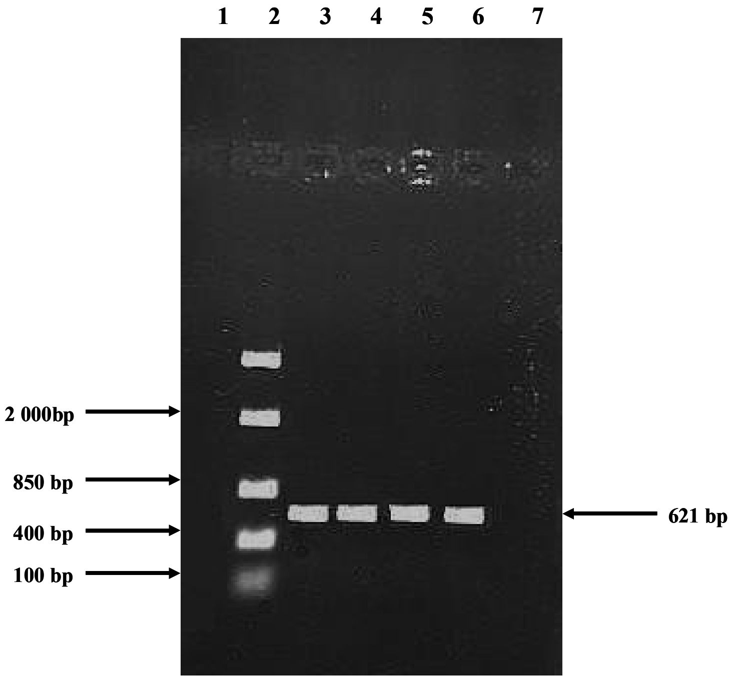


Figure 3.2.2.1: PCR amplification of a 621bp fragment of the IBD in preparation for restriction digestion. The PCR product was visualized on a 1% agarose gel. Lanes 3 -6 show the amplified PCR product, while lane 2 is a FastRuler^{TM} Middle Range DNA ladder (Fermentas) and lane 7 shows the "no-DNA" negative control.

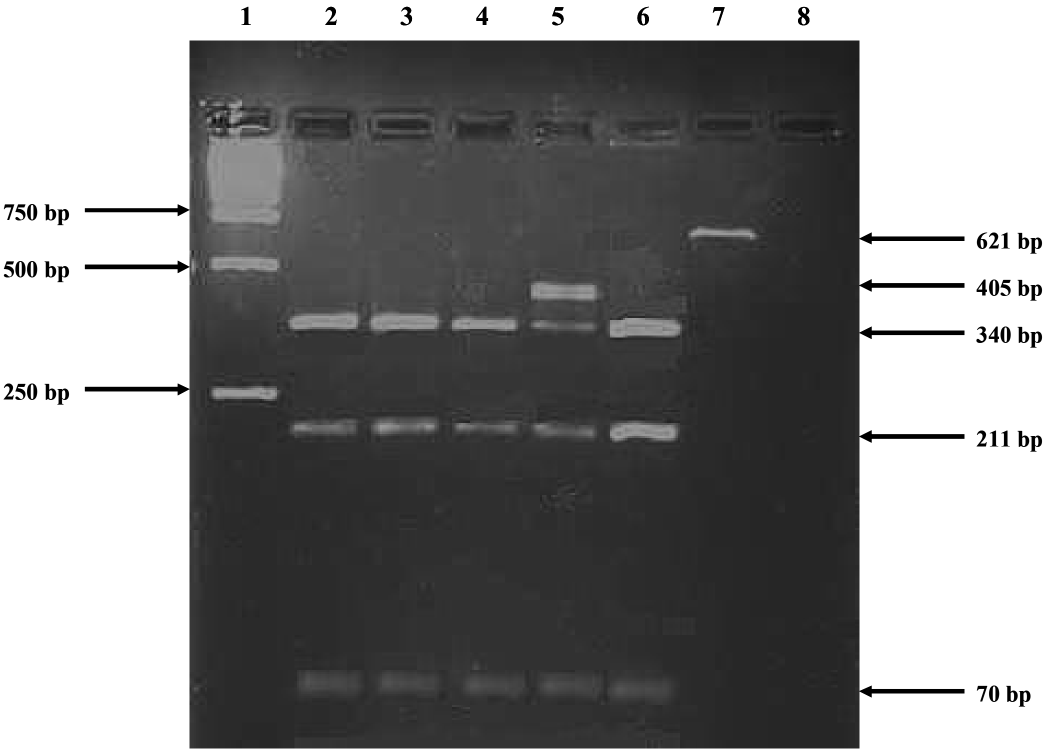


Figure 3.2.2.2: The restriction fragments generated when a 621bp fragment of the IBD is digested with *Mbo*I. When 5bp are deleted at position +41 796 an additional restriction site is introduced. The genotype at this position can then be determined by the different restriction profiles produced by digestion. The restriction digest was resolved on a 4% agarose gel. Lanes 2, 3, 4 and 6 show individuals in which the deletion is absent, while lane 5 shows an individual heterozygous for the deletion. Lane 7 shows the undigested "no enzyme" control and lane 1 is a GeneRuler^{TM} 1kb DNA ladder (Fermentas).

Genotyping of the deletion at position +41 796 was performed in a total of 122 samples (Appendix I). No individuals homozygous for the deletion were found within this population, but fifteen heterozygotes were identified. The minor allele was found to be quite rare, with a frequency of only 0.06. The population sample did not deviate significantly from Hardy-Weinberg equilibrium at this position (Table 3.2).

3.2.3 Genotyping of the Adjacent SNPs within the DNA-Binding Domain

Pyrosequencing™ was used to genotype two adjacent SNPs at positions +31 040 and +31 041 within the DNA-binding domain. Initially, a 227bp PCR fragment was amplified (Figure 3.2.3.1). Genotypes were then assigned based on the differential fluorescence patterns generated by the sequencing reaction (Figure 3.2.3.2). Genotyping of both SNPs was performed in 126 samples (Appendix I). While only six individuals were homozygous for the minor allele at both positions, almost twice as many individuals were found to be heterozygous at position +31 040 than at position +31 041. Consequently, the minor allele was more frequent at position +31 040 than +31 041. The population sample did not deviate from Hardy-Weinberg equilibrium at position +31 040, but did deviate significantly at position +31 041 with P = 0.03 (Table 3.2). However, correction for multiple tests revealed this deviation was not significant at the table-wide level (P<0.01).

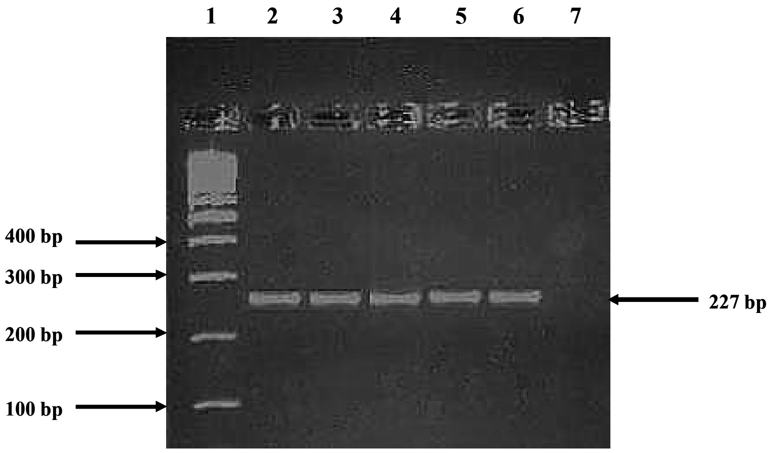


Figure 3.2.3.1: PCR amplification of a 227bp fragment of the DNA-binding domain in preparation for Pyrosequencing™. The PCR product was visualized on a 3% agarose gel. Lanes 2-6 show the amplified PCR product, while lane 7 shows the "no DNA" control and lane 1 is a Quick Load 100bp DNA Ladder (New England Biolabs).

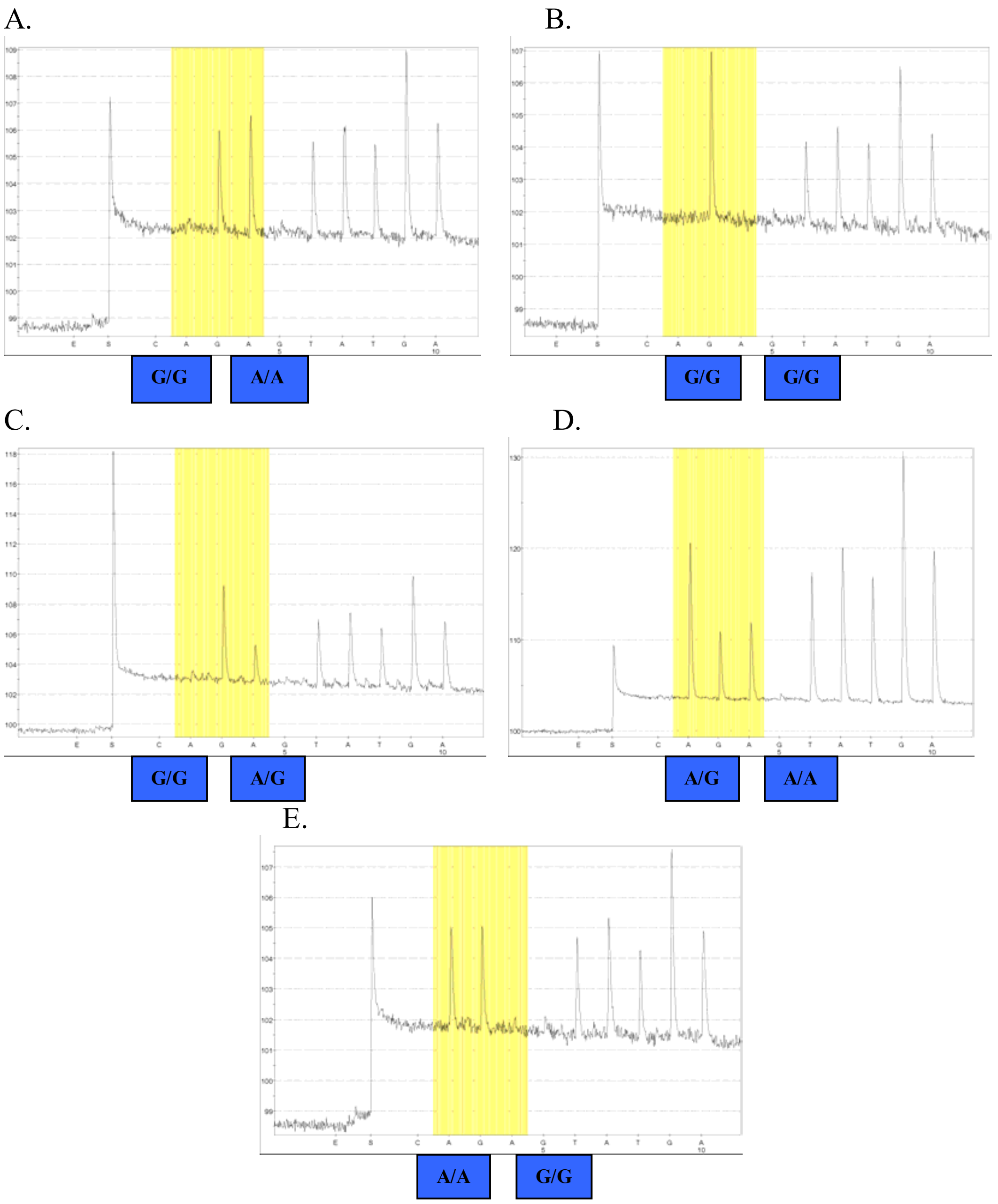


Figure 3.2.3.2: The pyrograms generated during Pyrosequencing™. (A-E) Each of the genotype profiles obtained is shown. Sequencing was performed in the reverse direction, thus all genotype are given in the reverse complement with the genotype at position +31 041 given first and then that at position +31 040.

Table 3.2: A summary of the genotyping data collected at all four polymorphic positions using allele-specific PCR, RFLP-PCR and Pyrosequencing™ assays. The genotypes, numbers of individuals genotyped, genotype and allele frequencies, χ^2 and P values are given for each position.

Polymorphism Position [1]	**Genotype**	**n**	**Genotype Frequency**	**Minor Allele Frequency**	**χ^2 Value:** [2]	**P Value**
- 417	-/- -/G G/G	68 52 16	0.50 0.38 0.12	0.31	1.48	0.22
+31 040	T/T T/C C/C	76 44 6	0.60 0.35 0.05	0.22	0.01	0.92
+31 041	C/C C/T T/T	94 26 6	0.75 0.21 0.05	0.15	4.74	0.03
+41 796	TCTTA/TCTTA TCTTA/- -/-	107 15 0	0.88 0.12 0.00	0.06	0.52	0.47

[1] Positions are given relative to the start of transcription.

[2] χ^2 values were calculated with 1 degree of freedom and at a significance level of $\alpha = 0.05$.

3.3 Estimation of Linkage Disequilibrium

While all the alleles were found to be in linkage disequilibrium, the degree of non-random association was not very strong between any of the four polymorphic sites. D and |D’| values indicated that LD was strongest between the SNP at position + 31 040 and the 5bp deletion at + 41 796 and weakest between +31 040T/C and the insertion in the upstream non-coding region, while r^2 values indicated that the strongest association was between the adjacent SNPs at +31 040 and 31 041 and the weakest association was between the SNP at +31 041 and the 5bp deletion at + 41 796. Interestingly, all three of these measures indicated that very low LD between the adjacent SNPs at +31 040 and +31 041.

Table 3.3: Results of linkage analysis showing the D, |D’| and r^2 values for pairwise LD between each of the four genotyped sites, as calculated using LDA v. 1.0 (Ding *et al.*, 2003).

+31 040			***+31 041***			***+41 796***			***SNP Position*** [1]
D	**|D’|**	**r^2**	**D**	**|D’|**	**r^2**	**D**	**|D’|**	**r^2**	
-0.03	0.08	0.0009	0.30	0.38	0.06	0.14	0.20	0.01	-417
			0.24	0.26	0.05	0.34	0.41	0.04	+31 040
						-0.01	0.07	0.0001	+31 041

[1] Positions are given relative to the start of transcription.

3.4 Estimation of Gene Frequencies

3.4.1 Differences between the General Population and HIV$^+$ Groups

In an effort to establish if a possible association exists between variation in *PSIP1* and HIV-1 infectivity, the genotype frequencies at each of the four polymorphic positions were compared between the general population and HIV$^+$ sample groups (Table 3.4.1). The two groups did not significantly differ from each other at positions -417, +31 041 and +41 796. There was, however, a significant difference between the two groups at position +31 040, where the heterozygote frequency in the HIV$^+$ group was more than twice that observed in the general population.

Table 3.4.1: The genotype frequencies at each of the four polymorphic positions, in both the general population and HIV$^+$ samples. The number of individuals genotyped and P values for Fisher's exact test are also given for each group.

SNP Position[1]	Genotype	General Population		HIV$^+$ Positive		P Value
		n	Genotype Frequency	n	Genotype Frequency	
-417	-/- -/G G/G	16 20 3	0.41 0.51 0.08	52 32 13	0.54 0.33 0.13	0.15
+31 040	T/T T/C C/C	29 7 3	0.74 0.18 0.08	47 37 3	0.54 0.43 0.03	0.01
+31 041	C/C C/T T/T	30 7 2	0.77 0.18 0.05	64 19 4	0.74 0.22 0.05	0.94
+41 796	TCTTA/TCTTA TCTTA/- -/-	36 3 0	0.92 0.08 0.00	71 12 0	0.86 0.14 0.00	0.38

[1] Positions are given relative to the start of transcription.

3.4.2 Differences between the Ethnic Groups

A comparison of the allele frequencies (Table 3.4.2) was made between the different ethnic groups represented in this study. Ethnic classification was based on the home language spoken by the individual in question and their immediate family. The comparison was conducted using 86 individuals who reported a single language spoken by their relatives for three generations, in both their maternal and paternal lineages and 50 individuals with uncertain lineages. Zulu speakers comprised 41% of the sample, while five of the ethnic groups (Venda, Tsonga, Swazi, Pedi and Ndebele) were represented by fewer than ten individuals (Table 3.4.2).

The minor allele frequency of the Zulu group at position -417 was lower than that of the Xhosa, Tswana and Sotho, who showed similar distributions. At position +31 040 the Zulu and Tswana showed similar frequencies, while the Xhosa and Sotho had similar frequencies that were lower than those of the other two groups. At position +31 041, the frequency distribution was similar for the Zulu, Tswana and Sotho, but was lower in the Xhosa. The Zulu and Sotho had identical frequencies at position +41 796, while the Xhosa and Tswana showed higher frequencies at this position. The group comprising individuals of mixed or unknown lineage had a virtually identical frequency distribution to the Zulu group at all four polymorphic positions.

Table 3.4.2: The minor allele frequencies at each of the four polymorphic sites, in five of the ethnic groups represented in this study. Only groups comprising more than ten individuals are given.

Language Group	n	Minor Allele Frequency			
		-417	+31 040	+31 041	+41 796
Zulu	35	0.29	0.27	0.15	0.05
Xhosa	13	0.42	0.14	0.05	0.13
Tswana	10	0.50	0.28	0.17	0.20
Sotho	13	0.46	0.17	0.17	0.05
Other[1]	50	0.25	0.25	0.15	0.05

[1] Group comprising individuals who had parents or grandparents who spoke different languages or who did not know the languages spoken by their relatives.

While the sampling distribution in this study was fairly representative of the population from which the sample was drawn, the sample sizes of several of the groups were too small to be sufficiently informative. The samples were thus pooled into four macrogroups according to the findings of Lane *et al.* (2002). Group 1 comprised all individuals from all nine ethnic groups with complete genotyping data at all four of the polymorphic positions. Group 2 comprised only individuals who reported Zulu as their home language in three generations. Group 3 comprised all individuals who reported Zulu or Xhosa as their home language and that of their relatives and group 4 comprised all individuals who reported Tswana, Pedi or Sotho as the home language of them and their relatives.

A comparison of the allele frequencies between the macrogroups (Table 3.4.4) revealed minor allele frequencies were similar between groups 2 and 3 at all four polymorphic positions. Allele frequencies in group 1 were similar to those in groups 2 and 3. However, allele frequencies in group 4 differed from those of the other three groups, at all four polymorphic positions. Interestingly, the allele frequencies in group 1 were equal to the average of those in groups 3 and 4 at positions +31 040 and +31 041.

A similar trend was observed when genotype frequencies were compared (Table 3.4.4). The frequencies in group 1 were similar to those in groups 2 and 3. However, frequencies in group 4 differed from those of the other three groups, at all four polymorphic positions. This pattern was particularly evident at positions +31 040 and +31 041. The four macrogroups did not deviate significantly from Hardy-Weinberg equilibrium at three of the polymorphic positions, but did deviate significantly from Hardy-Weinberg equilibrium at position +31 041. However, correction for multiple tests revealed this deviation was not significant at the table-wide level (P<0.01).

Table 3.4.4: The genotype (GF) and minor allele frequencies (MAF) at all four polymorphic positions, in the four groups generated by pooling genotyping data from the nine ethnic groups represented in this study. The numbers of individuals genotyped, as well as the χ^2- and P-values for the χ^2 test for goodness-of-fit to Hardy-Weinberg equilibrium are also given.

SNP Position[1]	Genotype	*Group 1*					*Group 2*					*Group 3*					*Group 4*				
		n	GF	MAF	χ^2 Value[2]	P Value	n	GF	MAF	χ^2 Value	P Value	n	GF	MAF	χ^2 Value	P Value	n	GF	MAF	χ^2 Value	P Value
-417	-/- -/G G/G	55 49 14	0.47 0.42 0.12	0.33	0.36	0.55	16 14 3	0.48 0.42 0.09	0.30	0.06[5]	0.98	20 19 5	0.45 0.43 0.11	0.33	0.02	0.89	10 10 6	0.38 0.38 0.23	0.42	1.17	0.28
+31 040	T/T T/C C/C	71 41 6	0.60 0.35 0.05	0.22	0.06[3]	0.98	17 14 2	0.52 0.42 0.06	0.27	0.16	0.69	25 17 2	0.57 0.39 0.05	0.24	0.18	0.67	18 7 1	0.69 0.27 0.04	0.17	0.09	0.76
+31 041	C/C C/T T/T	86 26 6	0.73 0.22 0.05	0.16	4.01	0.05	25 6 2	0.76 0.18 0.06	0.15	2.81	0.09	35 7 2	0.80 0.16 0.05	0.13	3.25	0.07	18 7 1	0.69 0.27 0.04	0.17	0.09	0.76
+41 796	TCTTA/TCTTA TCTTA/- -/-	102 13 0	0.89 0.11 0.00	0.06	0.41	0.52	30 2 0	0.94 0.06 0.00	0.03	0.06	0.81	38 5 0	0.88 0.12 0.00	0.06	0.17	0.68	19 5 0	0.79 0.21 0.00	0.10	0.32	0.57

[1] Positions are given relative to the start of transcription.

[2] χ^2 values were calculated with 1 degree of freedom and at a significance level of $\alpha = 0.05$.

[3] Value is given as x 10^2

3.5 Haplotype Analysis

PHASE 2.1 (Stephens *et al.*, 2001) was used to construct the haplotype structure surrounding the four polymorphic sites genotyped and to determine the frequencies of these haplotypes in the four population groups under investigation (Table 3.5). A total of sixteen possible haplotypes were identified in group 1 (which comprised of individuals from all of the nine ethnic groups represented) and all these haplotypes were also present in groups 2 and 3. However, only thirteen possible haplotypes were identified in the 26 individuals comprising group 4.

Analysis of the haplotype frequencies (Table 3.5) in each of the groups revealed the frequency distributions were virtually identical between groups 1, 2 and 3, with the ATCB haplotype being the most common and the ATTD haplotype being the most infrequent in all three groups. However, group 4 showed slightly different frequencies of the ACCB, ITCD and ITTB haplotypes. Additionally, while ATCB was still the most common haplotype in this population group, the ITTD haplotype (rather than ATTD) was the most infrequent haplotype in this group. A median joining network (Figure 3.5.1) was constructed using the haplotypes identified in group 1. This network had a cuboidal shape, rather than a tree-like structure, indicating there have been high levels of recombination and/or recurrent mutation between the polymorphisms.

Table 3.5: The estimated haplotype frequencies in each of the four macrogroups generated by pooling genotyping data from the nine ethnic groups represented in this study, as calculated using PHASE 2.1. (Stephens *et al.*, 2001).

Haplotype[1]	**Haplotype Frequency**			
	Group 1	**Group 2**	**Group 3**	**Group 4**
ATCB	0.45	0.43	0.43	0.40
ATCD	0.02	0.02	0.03	0.02
ATTB	0.05	0.04	0.03	0.04
ATTD	<0.01	<0.01	<0.01	0.00
ACCB	0.11	0.16	0.13	0.07
ACCD	0.01	<0.01	0.01	0.02
ACTB	0.03	0.04	0.03	0.03
ACTD	<0.01	<0.01	<0.01	0.00
ITCB	0.20	0.20	0.23	0.23
ITCD	0.01	<0.01	<0.01	0.05
ITTB	0.05	0.03	0.03	0.08
ITTD	<0.01	<0.01	<0.01	<0.01
ICCB	0.04	0.03	0.03	0.03
ICCD	<0.01	<0.01	<0.01	<0.01
ICTB	0.02	0.03	0.02	0.02
ICTD	<0.01	<0.01	<0.01	0.00

[1] I and D represent the presence of an insertion and deletion, respectively. A and B correspond to their absence.

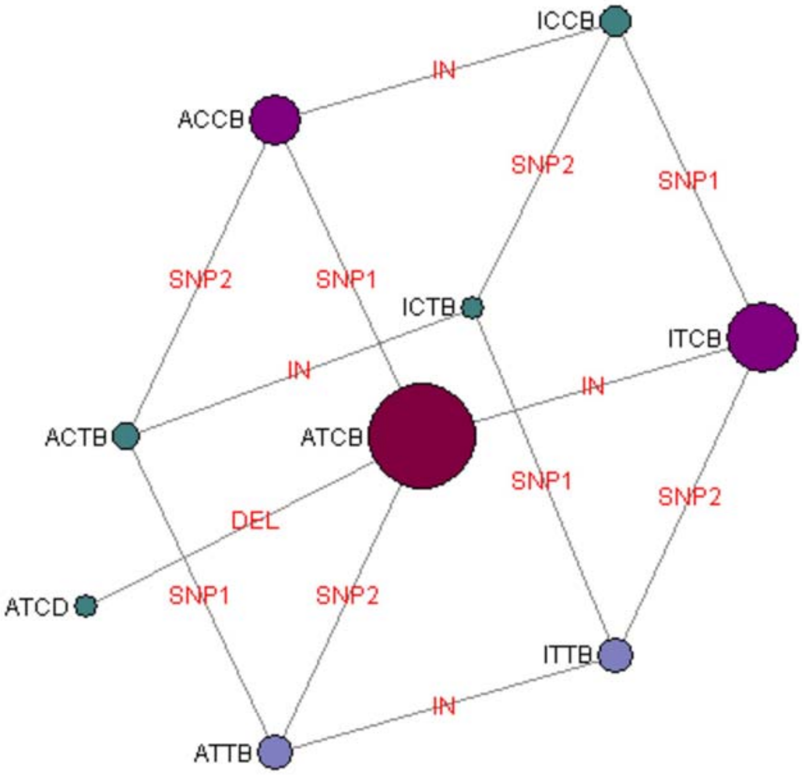


Figure 3.5.1 Median-joining network of the haplotypes present in the black South African population, constructed using Network 4.5.1.0. I and D represent the presence of an insertion and deletion, respectively. A and B correspond to their absence. Only haplotypes present at frequencies greater than 0.01 were used to construct the network. Each circle represents a haplotype and the mutations differentiating between them are shown in red. Circles are coloured according to haplotype frequency, with frequencies >0.20 in red, <0.20 in purple, <0.10 in blue and ≤0.05 in green.

Chapter 4

Discussion

Direct sequencing of three regions of *PSIP1*, spanning a total of 2 257bp, in twenty individuals revealed the presence of eleven variable sites. Five of these sites had previously been identified, while the remaining six had not (www.ensembl.org). These novel polymorphisms may thus be unique to African populations. This further illustrates the observation that Africans have the largest number of population-specific alleles and that the variation present in non-African populations is only a subset of the variation present in African populations (Armour *et al.*, 1996; Tishkoff *et al.*, 1996; Watson *et al.*, 1997; Kidd *et al.*, 1998; Tishkoff *et al.*, 2000).

A 721bp fragment of the upstream non-coding region contained one polymorphism, an indel at position -417. No variation had previously been detected in this region. This polymorphism was originally classified as an insertion based on comparison with the human reference sequence used in this study (www.ensembl.org). However, when this region was aligned with the reference sequence from *Pan troglodytes* the site was found to be monomorphic for the G allele in chimpanzees. Thus, assuming the G allele corresponds to the ancestral allele, this indel can be seen as representative of a deletion event rather than an insertion. It should be noted, however, that only one chimpanzee sequence was used to infer the state of the ancestral allele and this therefore may not be reflective of the variation within this gene in chimpanzees.

Seven variable sites were identified within a 566bp fragment of the DNA-binding domain, with minor allele frequencies ranging between 0.03 and 0.48. Four of these SNPs (30 830G/A, 30 846A/C, 31 040T/C and 31 097A/T) had previously been reported, while the remaining three (30 816G/C, 30 903G/C and 31 041C/T) had not. Conversely, an additional SNP (31 115T/C) which had previously been identified within this region was not found in the population under investigation. When aligned with the reference sequence from *Pan troglodytes*, the chimpanzee site was monomorphic and corresponded to the major allele at three of the polymorphic positions (+30 816G/C, +30 846A/C and +30 903G/C) and the minor allele at three others (+30 830G/A, +31 041C/T and +31 097A/T). However, the chimpanzee sequence was also polymorphic at position +31 040. Therefore, the state of the ancestral allele could not be inferred for this polymorphism. The fact that this position is polymorphic in both humans and chimpanzees could suggest that this polymorphism has been maintained in both species as a result of selection acting on this locus or another site that is in linkage disequilibrium with this one.

A further two deletions (2bp at position +41 780 and 5bp at position +41 796) and another SNP (+42 357T/C) were identified within a 970bp fragment of the IBD. The SNP had previously been reported, while the two deletions had not. No individuals were found to be homozygous for the minor allele at any of the three sites. Aligning this region with the corresponding region in *Pan troglodytes* revealed that at the SNP position the *P. troglodytis* sequence was monomorphic for the C allele, while both deletions are not present in chimpanzees.

Unlike the other polymorphisms identified in this study, the SNP within the IBD occurs within an exonic region of the gene. As a result, this polymorphism is subject to different selective and evolutionary forces. The SNP is found within a codon encoding an asparagine residue, but because it occurs at the third "wobble" position of the codon it does not alter the coding sequence. Given that this synonymous SNP has previously been identified in non-African populations, it is probably the result of a point mutation that occurred prior to the migration of non-Africans out of Africa. The deletions on the other hand appear to be unique to African populations and, as such, are probably much more recent products of deletion events in African populations. Alternatively these deletions may have occurred prior to the migration of non-African populations out of Africa and may later have been lost due to random genetic drift or the effects of natural selection. While it is impossible to positively rule out either explanation, given the very low frequencies of the minor alleles observed at both these positions, the former scenario seems more likely.

While the majority of the polymorphisms detected in this study were SNPs, three deletions were also identified within the three regions sequenced. This is consistent with recent findings which suggest that indels may represent as much as 15-18% of the genetic variation within the human genome (Dawson *et al.*, 2002; Weber *et al.*, 2002; Bhangale *et al.*, 2005; Mills *et al.*, 2006). However, these studies are based on whole-genome analyses which estimate that indels occur every 7-10kb across the genome. Given that only 2 257bp were sequenced during this study this number then seems unusually high. When one considers that a number of other indels have been identified in other non-coding regions of the gene (www.ensembl.org), it would seem that this gene appears to have a propensity for insertion-deletion polymorphisms.

While a great deal of time and effort has been devoted to the characterization and genotyping of SNPs, little is known about the frequency of indels and the mechanisms whereby they arise. Most often the presence of these mutations can be explained by Streisinger's "strand-slippage" hypothesis (Streisinger *et al.*, 1966). This theory proposes that indel mutations arise as a result of a strand slippage event during DNA replication that generates a misaligned intermediate with one or more unpaired nucleotides. Provided the unpaired nucleotide(s) avoid post-replicative repair mechanisms, an insertion or deletion is generated depending on whether the unpaired nucleotide was located in the primer or template strand, respectively. These replicative errors tend to occur at a higher frequency in sequence regions characterized by homopolymeric repeats (Streisinger and Owen, 1985). More recent NMR and crystallographic studies have provided structural support for this theory by demonstrating that conformational changes in DNA that orient the template strand at right angles to the primer-template junction during the interaction with DNA polymerase disrupt the van der Waals contacts and hydrogen bonds that ensure the correct base is added to the growing daughter strand (Garcia-Diaz and Kunkel, 2006).

In light of findings which show that DNA sequence complexity has an important role to play in influencing local DNA conformation, much has been done to try and characterize the sequences surrounding indels in an effort to identify possible "hotspots" of indel formation (Krawczak and Cooper, 1991; Cooper and Krawczak, 1991; Krawczak *et al.*, 2000; Chuzhanova *et al.*, 2002; Kondrashov and Rogozin, 2004; Ball *et al.*, 2005). As a result, a variety of sequence elements can be implicated in the formation of short insertions and deletions, several of which include direct repeats, inverted repeats and palindromic elements.

The influence of direct repeats on indels can be explained by a modified version of Streisinger's "strand-slippage" hypothesis (Krawczak and Cooper, 1991). This theory may provide a possible explanation for the formation of the deletions at positions -417 and +41 780, which both represent deletions of repetitive bases, but does not account for the deletion at position +41 796. Inverted repeats and palindromic elements on the other hand influence indel formation in a slightly different manner. These sequences often produce secondary structures like hairpin loops during replication which may be excised by the DNA replication repair enzymes, resulting in the formation of either a deletion or an insertion depending on which DNA strand is involved (Krawczak *et al.*, 2000; Chuzhanova *et al.*, 2002; Kondrashov and Rogozin, 2004; Ball *et al.*, 2005). This mechanism of indel formation may account for the 5bp deletion at position +41 796, which is flanked by inverted TGA repeats.

Four of the polymorphic sites were selected and genotyped using three different genotyping methods. While ultimately Pyrosequencing™ was used to genotype the adjacent SNPs within the DNA-binding domain, initially an attempt was made to genotype these SNPs using two allele-specific PCR reactions. This technique involves the use of different primers, which differ by a single base at the final or penultimate positions of their 3' ends, to detect the different allelic variants present at a given polymorphic position. If the allele the primer is designed to detect is present, amplification of the desired product will occur, whereas if the allele the primer is designed to detect is not present at a given position, a mismatch will result and little or no amplification will be observed (Newton *et al.*, 1989; Wu *et al.*, 1989).

The success of this technique is thus heavily dependent on the decreased ability of *Taq* DNA polymerase to extend mismatched bases at the 3' end of an oligonucleotide primer (Newton *et al.*, 1989; Wu *et al.*, 1989; Sarkar *et al.*, 1990; Huang *et al.*, 1992; Ayyadevara *et al.*, 2000) due to its lack of 3' to 5' exonuclease activity (Tindall and Kunkel, 1988). However, the ability of *Taq* polymerase to extend mismatches is not decreased to the same extent for all base pairs, as the resulting changes in the thermodynamic parameters that govern these reactions are different for each of the mismatched pairs (Newton *et al.*, 1989; Kwok *et al.*, 1990; Huang *et al.*, 1992; Ayyadevara *et al.*, 2000). The complexity of this situation is further compounded by the influence of the base immediately 5' to the mismatch on these same thermodynamic parameters (Breslauer *et al.*, 1986; Mendelman *et al.*, 1989; SantaLucia *et al.*, 1996).

While little consensus has been reached as to precisely what extent each mismatch pair reduces the efficiency of extension by *Taq* polymerase, several patterns have emerged. Firstly, the efficiency of mismatch extension is significantly increased when the 3' terminal base of the primer is an A or T (Kwok *et al.*, 1990; Ayyadevara *et al.*, 2000). Secondly, purine-pyrimidine and pyrimidine-purine mismatches are extended with greater efficiency that purine-purine or pyrimidine-pyrimidine mismatches (Newton *et al.*, 1989; Huang *et al.*, 1992). Thus, given that the mismatches involved at position +31 040 were T (primer)·G(template) and C·A, while those involved at position +31 041 were A·C and G·T, it becomes clear why allele-specific PCR was not a plausible option for genotyping these SNPs.

Allele-specific PCR was also initially used in an attempt to genotype the indel at position +41 780 within the IBD. At this position, a TT pair was deleted from a homopolymeric stretch of four Ts, resulting in the presence of either two or four Ts. As a result, allelic discrimination at this position was impossible as local misalignment of the primer and template during the polymerase reaction (Kunkel, 1990) resulted in extension of both primers, regardless of which allelic variant (either two Ts or four) was present. Thus this technique was again deemed to be inappropriate to genotype the polymorphism at this position.

However, despite these two instances where this technique was not the appropriate choice, allele-specific PCR was successfully used to genotype the indel at position -417 in the upstream non-coding region. The accuracy of the genotyping data obtained at this position was confirmed by comparing it with the sequencing data and by repeating several of the samples at random to confirm the two reactions concur. Thus while there are limitations to the use of allele-specific PCR for genotyping purposes, when all factors are carefully reviewed and considered during the design of an assay, this technique can provide accurate and reliable genotyping data.

The genotype and allele frequencies observed following genotyping corresponded well with those seen in the sequencing data and the population sample did not deviate significantly from Hardy-Weinberg equilibrium at three of the four positions. According to the standard chi-squared test for goodness-of-fit the population sample did deviate significantly from Hardy-Weinberg equilibrium at position +31 041 ($P<0.05$). However, this observation reflects the results of an individual significance test and makes no allowances for the fact that when multiple tests are performed, the significance level selected must be adjusted accordingly to control the overall type-1 error rate (Rice, 1988). Subsequent correction for multiple tests revealed this deviation was not significant at the table-wide level ($P<0.01$).

It is now well established that genetic variation plays an important role in determining an individual's susceptibility to complex disease (Tishkoff and Verrelli, 2003). The influence of genetic variation on an individual's susceptibility to HIV infection and rate of disease progression has been clearly highlighted by the identification and characterization of several AIDS restriction genes (ARGs) (O'Brien and Nelson, 2004). Because LEDGF/p75 interacts directly with HIV-1, tethering it to chromosomes (Maertens *et al.*, 2003; Emiliani *et al.*, 2005) at AT-rich regions within LEDGF/p75-regulated genes (Ciuffi *et al.*, 2005; Hombrouck *et al.*, 2007; Shun *et al.*, 2007; Marshall *et al.*, 2007), the genotype frequencies at each of the four polymorphic positions were compared between the general population and HIV$^+$ sample groups in an effort to establish whether a possible association exists between variation in *PSIP1* and HIV-1 infectivity. The two groups did not significantly differ from each other at three of the four polymorphic sites. There was however a significant difference between the two groups at position +31 040, where the frequency of heterozygotes observed in the HIV$^+$ group was twice that in the general population. This could indicate that this polymorphism may influence an individual's susceptibility to HIV-1 infection or rate of disease progression or may be linked to another as yet unidentified susceptibility allele. In the absence of another control group of individuals known to be HIV$^-$, one can only speculate that the minor allele at this position may provide a protective effect against HIV-1 infection and that in the heterozygous state this allele may perhaps slow the rate of HIV-1 disease progression.

Several studies have demonstrated that changes in LEDGF/p75 expression levels do result in differential patterns of HIV-1 integration (Llano *et al.*, 2004b; Vandegraaff *et al.*, 2006; Zielske and Stevenson, 2006; Vandekerckhove *et al.*, 2006; Llano *et al.*, 2006b), but consensus has not been reached as to whether or not this has a corresponding effect on HIV-1 susceptibility. One such study has shown that HIV-1 infectivity is decreased in cells expressing LEDGF/p75 mutants defective for DNA-binding (Ciuffi *et al.*, 2006). Given that the SNP at +31 040 is located within an intronic region of the DNA-binding domain, if this allele is in fact linked to HIV-1 susceptibility, it would most likely exert its influence at the regulatory level. In which case, this polymorphism may affect mRNA expression levels by affecting splicing (possibly resulting in truncated protein defective for DNA-binding) or at the level of mRNA stability.

However before conclusions can be drawn about disease association, it is necessary to have at least some understanding of the regional demographic history which has helped to shape patterns of genetic variation at this locus (Tishkoff and Verrelli, 2003). This is necessary because if population substructure is present, it is possible to detect spurious associations between arbitrary markers with no physical linkage to susceptibility loci and a disease phenotype (Pritchard and Rosenberg, 1999). Intronic polymorphisms are particularly useful for detecting population substructure, as they are subject to less stringent functional constraints and are thus more selectively neutral (Tishkoff and Verrelli, 2003).

As a result, allele frequencies were compared between the different ethnic groups represented in this study in order to establish if population substructure has had an influence on shaping patterns of genetic variation in the black South African population. Zulu speakers comprised 41% of the sample, Xhosa- and Sotho speakers each comprised 15%, Tswana speakers comprised 9% and the remaining five ethnic groups (Venda, Tsonga, Swazi, Pedi and Ndebele) each comprised less than 6% of the sample. While the sampling distribution in this study was fairly representative of the population from which the sample was drawn, the sample sizes of several of the groups were too small to be sufficiently informative. Groups comprising fewer than ten individuals were thus excluded from comparative analysis.

The four remaining ethnic groups showed differences in allele frequency to each other and the group comprising individuals of mixed or unknown ancestry at all four polymorphic positions. This supports the findings of Lane *et al.* (2002), that the black South African population does show distinct differences between the representative ethnic groups. Consistent with their findings, the Tswana and Sotho groups showed similar allele frequencies at two of the four positions, suggesting these population groups share a common ancestry that corresponds to their shared linguistic patterns. However while Lane *et al.* (2002) found the Zulu and Xhosa shared similar patterns of variation, the allele frequency distribution between these two groups in this study differed at all four polymorphic positions. This inconsistency may however, simply be a consequence of the relatively small sample sizes available for study in this investigation.

In order to reduce errors based on small sample sizes the samples were pooled into four macrogroups according to the findings of Lane *et al.* (2002). A comparison of the allele frequencies between the macrogroups revealed minor allele frequencies were similar between groups 2 (Zulu speakers) and 3 (Zulu and Xhosa speakers). This is consistent with the findings of Lane *et al.* (2002) that Zulu and Xhosa speakers show similar patterns of variation to each other. Allele frequencies in group 1 (comprising individuals from all nine ethnic groups) were also similar to those in groups 2 and 3. This unsurprising as Zulu and Xhosa speakers collectively comprised 37% of group 1. However, allele frequencies in group 4 (Tswana, Sotho and Pedi speakers) differed from those of the other three groups at all four polymorphic positions, again consistent with the findings of Lane *et al.* (2002) that Sotho/Tswana speakers have different patterns of variation to those of Zulu and Xhosa speakers. Another interesting trend was observed by comparing allele frequencies between the four macrogroups at positions +31 040 and +31 041. At these positions, the allele frequencies in group 1 were equal to the average of those in groups 3 and 4. This pattern is generally considered to be indicative of population substructure (Hartl and Clark, 1989) and thus suggests that group 1 (which is representative of the black South African population) is comprised of several subpopulations, which include groups 3 and 4, in accordance with the findings of Lane *et al.* (2002).

Further evidence for population substructure was provided by comparison of the genotype frequencies. When structured populations with different allele frequencies and that do not deviate from Hardy-Weinberg equilibrium are pooled, the resulting pooled population sample will display a reduction in heterozygosity relative to the subpopulations from which it is derived. This phenomenon is referred to as the Wahlund effect and often results in the pooled population sample deviating significantly from Hardy-Weinberg equilibrium (Hartl and Clark, 1989). The Wahlund effect can thus account for the observation that group 1 did deviate significantly from Hardy-Weinberg equilibrium at position +31 041, while the other three groups did not. However, this reduction in heterozygosity is dependent on the variance in allele frequencies between the subpopulations comprising the pooled population sample. Thus, if the allele frequencies in the two population samples do not differ dramatically, the reduction in heterozygosity will be insufficient to significantly alter the Hardy-Weinberg genotypic ratios (Hartl and Clark, 1989). This then accounts for why the four macrogroups did not deviate significantly from Hardy-Weinberg equilibrium at position +31 040, despite the clear evidence for population substructure at this position. The four macrogroups also did not deviate significantly from Hardy-Weinberg equilibrium at positions -417 and +41 796, but there was no evidence for the population substructure at either of these positions.

Signatures of population substructure can also be observed in LD patterns, which can also indicate genetic drift, population growth, admixture, migration, natural selection, variations in recombination and mutation rates and gene conversion (Ardlie *et al.*, 2002). While population substructure increases LD within each of the population groups comprising a larger subdivided population (Pritchard, 2001), LD is decreased within this larger population as its quantification is confounded by the divergent patterns of LD present in each of the smaller subgroups (Tishkoff and Verrelli, 2003).

Numerous studies have shown that LD in Africans extends over shorter distances than it does in non-African populations (Kidd *et al.*, 1998; Tishkoff *et al.*, 1998; Tishkoff *et al.*, 2000; Reich *et al.*, 2001; Gabriel *et al.*, 2002). This has largely been attributed to the fact that African populations have always maintained a larger effective population size than non-African populations (Tishkoff *et al.*, 1996; 1998; 2000). Nonetheless, these studies propose that LD in Africans can extend over 3-10kb (Collins *et al.*, 1999; Reich *et al.*, 2001; Gabriel *et al.*, 2002). However, these studies are based on variation within Yoruban populations that have been shown to be genetically distinct from South African populations (Soodyall *et al.*, 1996; Chen *et al.*, 2000; Salas *et al.*, 2002; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). Studies within our own population have revealed that LD in South African Bantu speakers extends over even shorter distances (Heitkamp *et al.*, personal communication).

In this instance, linkage analysis revealed that while all the alleles were found to be in linkage disequilibrium, the degree of non-random association was not very strong between any pair of the four polymorphic sites. Therefore since the polymorphisms at positions -417 and +41 796 are further than 40kb apart it is unsurprising that LD is not very strong between these sites. However, the fact that similarly low values of LD were observed between the SNPs at positions +31 040 and +31 041 is unexpected. These SNPs are adjacent to each other and as such it is unlikely that LD between these sites has been heavily influenced by recombination. Rather, given that the SNP at +31 041 appears to have very different allelic distributions within the different ethnic groups represented in this study, the low levels of LD between these two SNPs may reflect population substructure within the black South African population.

Elucidation of the haplotype structure underlying the variation present in the four macrogroups identified sixteen haplotypes in groups 1, 2 and 3 and these occurred at roughly the same frequency across all three groups. However only thirteen of these haplotypes were present in group 4 and three of these had frequencies that differed from those in the other groups. These differences in haplotype frequency reflect both the low levels of LD between the different pairs of polymorphisms and the differences in the allele frequencies observed between the groups. These results thus concur with those based on the analysis of genotype and allele frequencies and support the view that population substructure exists within the black South African population.

The median-joining network constructed based on the most frequent haplotypes observed assumed a cuboidal shape, which is often indicative of high levels of recombination and/or recurrent mutation. As seen from the LD analysis, it is likely that recombination has played a role in breaking down ancestral haplotypes in this region. This is especially likely between the indel at -417 and the other polymorphisms (which are separated by more than 30kb). However, the SNPs at +31 040 and +31 041 are immediately adjacent to each other, yet the ancestral haplotype structure surrounding these polymorphisms has been broken down dramatically. In this instance recurrent mutation seems a more likely explanation.

Considering the nature of the polymorphisms in question and the mechanisms involved in their formation, it seems unlikely that the two indels could have occurred more than once and thus they are unlikely to have been subject to back mutation (Tishkoff *et al.*, 1996). However, nucleotide substitutions can occur more than once and have been known to back mutate. Since the SNP at position +31 040 has previously been identified in both non-African populations and in *Pan troglodytes*, suggesting that this polymorphism may have been maintained within the population for some length of time. Thus it seems that if back mutation has indeed occurred within this gene, the SNP at position +31 041 provides the most likely candidate. Given the strong evidence for population substructure at this position, the back mutation event probably occurred within one of the subpopulations comprising the black South African population. This would also account for the very different allele frequencies seen between the various subpopulations at this position.

So what effect does the presence of population substructure present in the black South African population have on the power of this study to detect an association between genetic variation patterns and an individual's susceptibility to HIV-1 infection? The primary danger involved in looking for disease associations in substructured populations is that substructure can confound LD analysis by inflating D' and r^2 values. This can result in spurious associations between a disease phenotype and arbitrary markers that have no real physical link to causative loci. This is particularly a problem when sampling of case and control samples is done without any regard to ethnicity, so that these groups then contain different frequencies of each ethnic group and subsequent pooling of samples results in different allele and genotype frequency distributions (Pritchard and Rosenberg, 1999).

However, in this study no such inflation of LD was evident. Rather, the opposite effect was observed. Also the HIV^+ (case) and general population samples (control) used were very similar in terms of their ethnic composition, ensuring the effects produced by pooling the data in these groups were similar for both groups. Thus with these measures in place to reduce the type-1 error rate, the association detected at +31 040 may indeed reflect more than just a spurious association. And when coupled with the knowledge that this SNP has previously been identified in both non-African populations and in *Pan troglodytes* and thus probably represents an ancestral mutation that has been maintained within the human population since before the divergence between humans and chimpanzees, it appears plausible that genetic variation at this position may be linked to HIV-1 infectivity.

References

Alonso, S. and Armour, J.A.L. 2001. A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. *PNAS* **98**: 864-869

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lippman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403-410

Aravind, L. and Landsman, D. 1998. AT-hook motifs identified in a wide variety of DNA-binding proteins. *Nucl. Acids Res.* **26**: 4413-4421

Ardlie, K.G., Kruglyak, L. and Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **3**: 299-309

Armour, J.A.L., Antttinen, T., May, C.A., Vega, E.E., Sajantila, A., *et al.*. 1996. Minisatellite diversity supports a recent African origin for modern humans. *Nat. Genet.* **13**: 154-160

Arndt, P.F., Hwa, T. and Petrov, D.A. 2005. Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density and telomere-specific effects. *J. Mol. Evol.* **60**: 748-763

Aydin, A., Tolait, M.R., Bähring, S., Becker, C. and Nürnberg, P. 2006. New universal primer facilitates Pyrosequencing™. *Electrophoresis* **27**: 394-397

Ayyadevara, S., Thaden, J.J. and Reis, R.J. 2000. Discrimination of primer 3'-nucleotide mismatch by *Taq* DNA polymerase during polymerase chain reaction. *Anal. Biochem.* **284**: 11-18

Ball, E.V., Stenson, P.D., Abeysinghe, S.S., Krawczak, M., Cooper, D.N. *et al.*. 2005. Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mut.* **26**: 205-213

Barr, S.D., Leipzig, J., Shinn, P., Ecker, J.R. and Bushman, F.D. 2005. Integration targeting by avian sarcoma–leukosis virus and human immunodefieciency virus in the chicken genome. *J. Virol.* **79**: 12035-12044

Barre-Sinoussi, F., Chemann, J.C., Rey, F., Nugeyre, M.T., Chamaret, S., *et al.*. 1983. Isolation of a T-lymphotropicretrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**: 868-871

Bartholomeeusen, K., De Rijck, J., Busshots, K., Desender, L., Gijsbers, R., *et al.*. 2007. Differential interaction of HIV-1 integrase and JPO2 with the C terminus of LEDGF/p75. *J. Mol. Biol.* **372**: 407-421

Beutler, E. 1994. G6PD deficiency. *Blood* **84**: 3613-3636

Bhangale, T.R., Rieder, M.J., Livingston, R.J. and Nickerson, D.A. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum. Mol. Genet.* **14**: 59-69

Breslauer, K.J., Frank, R., Blöcker, H. and Marky, L.A. 1986. Predicting DNA duplex stability from the base sequence. *PNAS* **83**: 3746-3749

Buckland, P.R. 2006. The importance and identification of regulatory polymorphisms and their mechanisms of action. *Biochim. Biophys. Acta* **1762**: 17-28

Buckland, P.R., Hoogendoorn, B., Coleman, S.L., Guy, C.A., Smith, S.K., *et al.*. 2005. Strong bias in the location of functional promoter polymorphisms. *Hum. Mut.* **26**: 214-223

Bushman, F.D. and Craigie, R. 1991. Activities of human immunodeficiency virus (HIV) integration protein *in vitro*: specific cleavage and integration of HIV DNA. *PNAS* **88**: 1339-1343

Bushman, F., Fujiwara, T. and Craigie, R. 1990. Retroviral DNA integration directed by HIV integration protein *in vitro*. *Science* **249**: 1555-1558

Busshots, K., Vercammen, J., Emiliani, S., Benarous, R., Engelborghs, Y., *et al.*. 2005. The interaction of LEDGF/p75 with integrase in lentivirus-specific and promotes DNA binding. *J. Biol. Chem.* **280**: 17841-17847

Bustemante, C.D., Fledel-Alon, A., Williamson, S., Nielsen, R., Hubisz, M.T., *et al.*. 2005. Natural selection on protein-coding genes in the human genome. *Nature* **437**: 1153-1157

Carrington, M., Dean, M., Martin, M.P., and O'Brien S.J. 1999. Genetics of HIV-1 infection: chemokine receptor CCR5 polymorphism and its consequences. *Hum. Mol. Genet.* **8**: 1939-1945

Ceppellini, R., Siniscalco, M. and Smith, C.A.B. 1955. The estimation of gene frequencies in a random-mating population. *Ann. Hum. Genet.* **20**: 97-115

Chakravarti, A. 1999. Population genetics – making sense out of sequence. *Nat. Genet. Supp.* **21**: 56-60

Chen, Y., Olckers, A., Schurr, T., Kogelnik, A.M., Huoponen, K. *et al.*. 2000. mtDNA variation in the South African Kung and Khwe-and their genetic relationship to other African populations. *Am. J. Hum. Genet.* **66**: 1362-1383

Chen, Y., Torroni, A., Excoffier, L., Santachiara-Benerecetti, A.and Wallace, D.C. 1995. Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am. J. Hum. Genet.* **57**: 133-149

Cherepanov, P., Maertens, G., Proost, P., Devreese, P., Van Beeumen, J., *et al.*. 2003. HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J. Biol. Chem.* **278**: 372-381

Cherepanov, P., Devroe, E., Silver, P.A. and Engelman, A. 2004. Identification of an evolutionarily conserved domain in human lens epithelium-derived growth factor/ transcriptional co-activator 75 (LEDGF/p75) that binds HIV-1 integrase. *J. Biol. Chem.* **279**: 48883-48892

Cherepanov, P., Sun, Z.J., Rahman, S., Maertens, G., Wagner, G., *et al.*. 2005a. Solution structure of the HIV-1 integrase-binding domain in LEDGF/p75. *Nat. Struct. Mol. Biol.* **12**: 526-532

Cherepanov, P., Ambrosio, A.L.B., Rahman, S., Ellenberger, T. and Engelman, A. 2005b. Structural basis for the recognition between HIV-1 integrase and transcriptional coactivator p75. *Proc. Nat. Acad. Sci.* **102**: 17308-17313

Cherepanov, P. 2007. LEDGF/p75 interacts with divergent lentiviral integrases and modulates the enzymatic activity *in vitro*. *Nucl. Acids Res.* **35**: 113-124

Cheung, V.G., Conlin, L.K., Weber, T.M., Arcado, M., Jen, K.Y. *et al.*. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* **33**: 422-425

Chuzhanova, N.A., Anassis, E.J., Ball, E.V., Krawczak, M. and Cooper, D.N. 2002. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. *Hum. Mut.* **21**: 28-44

Ciuffi, A., Llano, M., Poeschla, E., Hoffmann, C., Leipzig, J., *et al.*. 2005. A role for LEDGF/p75 in targeting HIV DNA integration. *Nature Med.* **11**: 1287-1289

Ciuffi, A., Diamond, T., Hwang, Y., Marshall, H. and Bushman, F. 2006. Modulating target site selection during human immunodeficiency virus DNA integration *in vitro* with an engineered tethering factor. *Hum. Gene Ther.* **17**: 1-8

Collins, A., Lonjou, C. and Morton, N.E. 1999. Genetic epidemiology of single-nucleotide polymorphisms. *PNAS* **96**: 15173-15177

Cooper, D.N. and Krawczak, M. 1991. Mechanisms of insertional mutagenesis in human genes causing genetic disease. *Hum. Genet.* **87**: 409-415

Crawford D.L., Segal, J.A. and Barnett, J.L. 1999. Evolutionary analysis of TATA-less proximal promoter function. *Mol. Biol. Evol.* **16**: 194-207

Crise, B., Li, Y., Yuan, C., Morcock, D.R., Whitby, D., *et al.*. 2005. Simian immunodeficiency virus integration preference is similar to that of human immunodeficiency virus type 1. *J. Virol.* **79**: 12199-12204

Crow, J.F. 1986. Basic Concepts in Population, Quantitative and Evolutionary Genetics. W.H. Freeman and Co. New York, USA. pp 6-10, 239

Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T.J. and Lander, E.S. 2001. High-resolution haplotype structure in the human genome. *Nat. Genet.* **29**: 229-232

Dawson, E., Abecasis, G.R., Bumpstead, S., Chen, Y., Hunt, S., *et al.*. 2002. A first generation linkage disequilibrium map of human chromosome 22. *Nature* **418**: 544-548

Dean, M., Carrington, M., Winkler, C., Huttley, G.A., Smith, M.W., *et al.*. 1996. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CCR5 structural gene. *Science* **273**: 1856-1862

Deng, G. 1989. A sensitive non-radioactive PCR-RFLP analysis for detecting point mutations at 12th codon oncogene c-Ha-*ras* in DNAs of gastric cancer. *Nucl. Acids Res.* **16**: 1988

Ding, K., Zhou, K., He, F. and Shen, Y. 2003. LDA – a java-based linkage disequilibrium analyzer. *Bioinformatics* **19**: 2147-2148

Emiliani, S., Mousnier, A., Busshots, K., Maroun, M., van Maele, B., *et al.*. 2005: Integrase mutants defective for interaction with LEDGF/p75 are impaired in chromosome tethering and HIV-1 replication. *J. Biol. Chem.* **280**: 25517-25523

Engelman, A., Bushman, F.D. and Craigie, R. 1993. Identification of the discrete functional domains of HIV-1 integrase and their organization within an active multimeric complex. *EMBO J.* **12**: 3269-3275

Engelman, A., Mizuuchi, K. and Craigie, R. 1991. HIV-1 DNA integration: mechanism of viral DNA cleavage and DNA strand transfer. *Cell* **67**: 1211-1221

Excoffier, L. Human demographic history: refining the recent African origin model. *Curr. Opin. Genet. Dev.* **12**: 675-682

Fassati, A. and Goff, S. 2001. Characterization of intracellular reverse transcription complexes of human immunodeficiency virus type 1. *J. Virol.* **75**: 3626-3635

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., *et al.*. 2002. The structure of haplotype blocks in the human genome. *Science* **296**: 2225-2229

Garcia-Diaz, M. and Kunkel, T.A. 2006. Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem Sci* **31**: 206-214

Ge, H., Si, Y. and Roeder, R.G. 1998a. Isolation of cDNAs encoding novel transcription coactivators p52 and p75 reveals an alternate regulatory mechanism of transcriptional activation. *EMBO* **17**: 6723-6729

Ge, H., Si, Y. and Wolfe, A. 1998b. A novel transcriptional coactivator, p52, functionally interacts with the essential splicing factor ASF/SF2. *Mol. Cell.* **2**: 751-759

Ge, Y.Z., Pu, M.T., Gowher, H., Wu, H.P., Ding, J.P., *et al.*. 2004. Chromatin targeting of de novo DNA methyltransferases by the PWWP domain. *J. Biol. Chem.* **279**: 25447-25454

Gojobori, T., Li, W. and Graur, D. 1982. Patterns of nucleotide substitution in pseudogenes and functional genes. *J. Mol. Evol.* **18**: 360-369

Gonder, M.K., Mortensen, H.M., Reed, F.A., de Sousa, A. and Tishkoff, S.A. 2007. Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol. Biol. Evol.* **24**: 757-768

Gonzalez, E., Dhanda, R., Bamshad, M., Mummidi, S, Geevarghese, R., *et al.*. 2001. Global survey of genetic variation in CCR5, RANTES and MIP-1alpha: impact on the epidemiology of the HIV-1 pandemic. *PNAS* **98**: 5199-5204

Hammer, M.F., Karafet, T.M., Redd, A.J., Jarjanazi, H., Santachiara-Benerecetti, A., *et al.*. 2001. Hierarchical patterns of global human Y-chromosome diversity. *Mol. Biol. Evol.* **18**: 1189-1203

Hammer, M.F., Spurdle, A.B., Karafet, T., Bonner, M.R., Wood, E.T., *et al.*. 1997. The geographic distribution of human Y chromosome variation. *Genetics* **145**: 787-805

Hartl, D.L. and Clark, A.G. 1989. Principles of population genetics, 2nd edition. Sinauer Associates Inc., Sunderland, Massachusetts. pp 282-288

Henn, B.M., Gignoux, C., Lin, A.A., Oefner, P.J., Shen, P., *et al.*. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania into southern Africa. *PNAS* **105**: 10693-10698

Hombrouck, A., De Rijck, J., Hendrix, J., Vandekerckhove, L., Voet, A., *et al.*. 2007. Virus evolution reveals an exclusive role for LEDGF/p75 in chrosomal tethering of HIV. *PLoS Path.* **3**: e47

Huang, M., Arnheim, N. and Goodman, M.F. 1992. Extension of base mispairs by *Taq* DNA polymerase: implications for single nucleotide discrimination in PCR. *Nucl. Acids Res.* **20**: 4567-4573

Ingman, M., Kaessermann, H., Pääbo, S. and Gytiensten, U. 2000. Mitochondrial genome variation and the origin of modern humans. *Nature* **408**: 708-713

Ishimoto, Y., Kuroda, T., Harada, H., Kishimoto, T. and Nakamura, H. 1997. Hepatoma-derived growth factor belongs to a gene family in mice showing significant homology in the amino terminus. *Biophys. Res. Commun.* **238**: 26-32

Jorde, L.B. and Wooding, S.P. 2004. Genetic variation, classification and "race". *Nat. Genet.* **36**: S28-S33

Kaessermann, H., Heibig, F., von Haeseler, A. and Pääbo, S. 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.* **22**: 78-81

Kidd, J.R., Pakstis, A.J., Zhao, H., Lu, R., Okonofua, F.E., *et al.*. 2000. Haplotypes and linkage disequilibrium at the phenylalanine hydroylase locus, *PAH*, in a global representation of populations. *Am. J. Hum. Genet.* **66**: 1882-1899

Kidd, K.K., Morar, B., Castiglione, C.M., Zhao, H., Pakistis, A.J., *et al.*. 1998. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum. Genet.* **103**: 211-227

Knight, A., Underhill, P.A., Mortensen, H.M., Zhivotovsky, L.A., Lin, A.A., *et al.*. 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr. Biol.* **13**: 464-473

Knight, J.C. 2005. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* **83**: 97-109

Kondrashov, A.S. and Rogozin, I.B. 2004. Context of deletions and insertions in human coding sequences. *Hum. Mut.* **23**: 177-185

Krawczak, M., Chuzhanova, N.A., Stenson, P.D., Johansen, B. N., Ball, E.V., *et al.*. 2000. Changes in primary DNA sequence complexity influence the phenotypic consequences of mutations in human gene regulatory regions. *Hum. Genet.* **107**: 362-365

Krawczak, M. and Cooper, D.N. 1991. Gene deletions causing human genetic disease: mechanisms of mutagenesis and the role of the local DNA sequence environment. *Hum. Genet.* **86**: 425-441

Kunkel, T.A. 1990. Misalignment-mediated DNA synthesis errors. *Biochemistry* **29**: 8003-8011

Kwok, S., Kellogg, D.E., McKinney, N., Spasic, D., Goda, L., *et al.*. 1990. Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nuc. Acids Res.* **18**: 999-1005

Lane, A.B., Soodyall, H., Arndt, M.E., Ratshikhopha, E., Jonker, C., *et al.*. 2002. Genetic substructure in South African Bantu-speakers: evidence from autosomal DNA and Y-chromosome studies. *Am. J. Phy. Anthro.* **119**: 175-185

Levy, J.A., Hoffman, A.D., Kramer, S.M., Landis, J.A., Shimabukuro, J.M., *et al.*. 1984. Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS. *Science* **225**: 840-842

Lewinski, M.K., Yamashita, M., Emerman, M., Ciuffi, A., Marshall, H., *et al.*. 2006. Retroviral DNA integration: viral and cellular determinants of target site selection. *PLoS Path.* **2**: e60

Lewontin, R.C. 1964. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**: 49-67

Lewontin, R.C. 1988. On measures of gametic disequilibrium. *Genetics* **120**: 849-852

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., *et al.*. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100-1104

Li, W., Wu, C. and Luo, C. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J. Mol. Evol.* **21**: 58-71

Llano, M., Vanegas, M., Fregoso, Saenz, D., Chung, S., *et al.*. 2004a. LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. *J. Virol.* **78**: 9524-9537

Llano, M., Delgado, S., Vanegas, M. and Poeschla, E.M. 2004b. Lens epithelium-derived growth factor/p75 prevents proteasomal degradation of HIV-1 integrase. *J. Biol. Chem.* **279**: 55570-55577

Llano, M., Vanegas, M., Hutchins, N., Thompson, D., Delgado, S., *et al.*. 2006a. Identification and characterization of the chromatin-binding domains of the HIV-1 integrase interactor LEDGF/p75. *J. Mol. Biol.* **360**: 760-773

Llano, M., Saenz, D.T., Meehan, A., Wongthida, P., Peretz, P., *et al.*. 2006b. An essential role for LEDGF/p75 in HIV integration. *Science* **314**: 461-464

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., *et al.*. 2005. Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* **308**: 1034-1036

MacNeil, A., Sankale, J.L., Meloni, S.T., Sarr, A.D., Mboup, S., *et al.*. 2006. Genomic sites of human immunodeficiency virus type 2 (HIV-2) integration: similarities to HIV-1 *in vitro* and possible differences *in vivo*. *J. Virol.* **80**: 7316-7321

Maertens, G., Cherepanov, P. and Engelman, A. 2006. Transcriptional co-activator p75 binds and tethers the Myc-interacting protein JPO2 to chromatin. *J. Cell Sci.* **119**: 2563-2571

Maertens, G., Cherepanov, P., Pluymers, W., Busschots, K. and De Clercq, E., *et al.*. 2003. LEDGF/p75 is essential for nuclear and chromosomal targeting of HIV-1 integrase in human cells. *J. Biol. Chem.* **278**: 33528-33539

Maertens, G., Cherepanov, P., Debyser, Z., Engelborghs, Y. and Engelman, A. 2004. Identification and characterization of a functional nuclear localization signal in the HIV-1 integrase interactor LEDGF/p75. *J. Biol. Chem*. **279**: 33421-33429

Magana-Arachchi, D.N., Sharma, P., Fatma, N., Singh, D.P. and Shinohara, T. 2003. Identification of regulatory elements in the promoter of LEDGF. *Invest. Ophthalmol. Vis. Sci.* **44**: E-Abstract

Marshall, H.M., Ronen, K., Berry, C., Llano, M., Sutherland, H., *et al.*. 2007. Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting. *PLoS One* **12**: e1340

Martin, M.P., Dean, M., Smith, M.W., Winkler, C., Gerrard, B., *et al.*. 1998. Genetic acceleration of AIDS progression by a promoter variant of CCR5. *Science* **282**: 1907-1911

Martinez, E., Chiang, C.M., Ge, H. and Roeder, G. 1994. TATA-binding protein-associated factor(s) in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. *EMBO J.* **13**: 3115-3126

Mendelman, L.V., Boosalis, M.S., Petruska, J. and Goodman, M.F. 1989. Nearest neighbor influences on DNA polymerase insertion fidelity. *J. Biol. Chem.* **264**: 14415-14423

Miller, M.D., Farnet, F.M. and Bushman, F.D. 1997. Human immunodeficiency virus type 1 preintegration complexes: studies of organization and composition. *J. Virol.* **71**: 5382-5390

Mills, R.E., Luttig, C.T., Larkins, C.E., Beauchamp, A., Tsui, C., *et al.*. 2006. An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res.* **16**: 1182-1190

Mitchell, R.S., Beitzel, B.F., Schroder, A.R.W., Shinn, P., Chen, H., *et al.*. 2004. Retroviral DNA integration: ASLV, HIV and MLV show distinct target site preferences. *PLoS Biol.* **2**: e234

Morley, M., Molony, C.M., Weber, T.M., Devlin, J.L., Ewens, K.G., *et al.*. 2004.Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743-747

Mountain, J.L. and Risch, N. 2004. Assessing genetic contributions to phenotypic differences among “racial” and “ethnic” groups. *Nat. Genet.* **36**: S48-S53

Nakamura, H., Izumoto, Y. Kambe, H., Kuroda, T., Mori, T., *et al.*. 1994. Molecular cloning of complementary DNA for a novel hepatoma-derived growth factor. *J. Biol. Chem.* **269**: 25143-25149

Narezkina, A., Taganov, K.D., Litwin, S., Stoyanova, R., Hayashi, J., *et al.*. 2004. Genome-wide analyses of avian sarcoma virus integration sites. *J. Virol.* **78**: 11656-11663

Newton, C., Graham, A., Heptinstall, L., Powell, S.J., Summers, C. *et al.*. 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucl. Acids Res.* **17**: 2503-2515

Nishizawa, Y., Usukura, J., Singh, D.P., Chylack, T. and Shinohara, T. 2001. Spatial and temporal dynamics of two alternatively spliced regulatory factors, lens epithelium-derived growth factor (LEDGF/p75) and p52, in the nucleus. *Cell Tissue Res.* **305**: 107-114

Nordborg, M. 1998. On the probability of Neanderthal ancestry. *Am. J. Hum. Genet.* **63**: 1237-1240

O'Brien, S.J. and Nelson, G.W. 2004. Human genes that limit AIDS. *Nat. Genet.* **36**: 565-57462

Passarino, G., Seminò, O., Quintana-Murci, L., Excoffier, L., Hammer, M., *et al.*. 1998. Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am. J. Hum. Genet.* **62**: 420-434

Pastinen, T., Sladek, R., Gurd, S., Sammak, A., Ge, B., *et al.*. 2004. A survey of genetic and epigenetic variation affecting human gene expression. *Physiol. Genomics* **16**: 184-193

Pauza, C.D. 1990. Two bases are deleted from the termini of HIV-1 linear DNA during integrative recombination. *Virol.* **179**: 886-889

Plagnol, V. and Wall, J.D. 2006. Possible ancestral structure in human populations. *PLoS Genet.* **2**: e105

Popovic, M., Samgadharan, M.G., Read, E. and Gallo, R.C. 1984. Detection, isolation and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. *Science* **224**: 497-500

Pritchard, J.K. 2001. Are rare variants responsible for susceptibility to complex disease? *Am. J. Hum. Genet.* **69**: 124-137

Pritchard, J.K. and Cox, N.J. 2002. The allelic architecture of human disease genes: common disease-common variant…or not? *Hum. Mol. Genet.* **11**: 2417-2423

Pritchard, J.K. and Przeworski, M. 2001. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**: 1-14

Pritchard, J.K. and Rosenberg, N.A. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**: 220-228

Quintana-Murci, L., Seminò, O., Bandelt, H., Passarino, G., McElreavey, K., *et al.*. 1999. Genetic evidence of an early exit of *Homo sapiens* from Africa through eastern Africa. *Nat. Genet.* **23**: 437-441

Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenbeg, N.A., Feldman, M.W. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *PNAS* **102**: 15942-15947

Reich, D.E. and Goldstein, D.B. 1998. Genetic evidence for a Paleolithic human population expansion. *Proc. Natl. Acad. Sci.* **95**: 8119-8123

Reich, D.E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P.C., *et al.*. 2001. Linkage disequilibrium in the human genome. *Nature* **411**: 199-204

Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., *et al.*. 2002. Human genome sequence variation and the influence of gene history, mutation and recombination. *Nat. Genet.* **32**: 135-142

Relethford, J.H. and Jorde, L.B. 1999. Genetic evidence for larger African population size during recent human evolution. *Am. J. Phy. Anthro.* **108**: 251-260

Rice, W.R. 1988. Analyzing tables of statistical tests. *Evolution* **43**: 223-225

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlen, M. and Nyren, P. 1996. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**: 84-89

Ronaghi, M, Uhlen, M. and Nyren, P. 1998. A sequencing method based on real-time pyrophosphate. *Science*.**281**: 363-365

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., *et al.*. 2002. Genetic structure of human populations. *Science* **298**: 2381-2384

Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**: 365-386

Ruwende, C. and Hill, A. 1998. Glucose-6-phosphate deficiency and malaria. *J. Mol. Med.* **76**: 581-588

Ruwende, C., Khoo, S.C., Snow, R.W., Yates, S.N.R., Kwlatkowski, D., *et al.* 1995. Natural selection for hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* **376**: 246-249

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., *et al.*. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* **239**: 487-491

Salas, A., Richards, M., De la Fe, T., Lareu, M., Sobrino, B., *et al.*. 2002. The making of the African mtDNA landscape. *Am. J. Hum. Genet.* **71**: 1082-1111

Salas, A., Richards, M., Lareu, M., Scozzari, R., Coppa, A., *et al.*. 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am. J. Hum. Genet.* **74**: 454-465

Sanger, F., Nicklen, S. and Coulson, A.R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74**: 5463-5467

SantaLucia, J., Allawi, H.T. and Seneviratne, P.A. 1996. Improved nearest-neighbor parameters for predicting DNA duplex stability. *Biochemistry* **35**: 3555-3562

Sarkar, G., Cassady, J., Bottema, C. and Sommer, S. 1990. Characterization of polymerase chain reaction amplification of specific alleles. *Anal. Biochem.* **186**: 64-68

Schröder, A.R.W., Shinn, P., Chen, H., Berry, C., Ecker, J.R. and Bushman, F. 2002. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110:** 521-529

Scozzari, R., Cruciani, F., Santolamazza, P., Malaspina, P., Torroni, A. *et al.*. 1999. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am. J. Hum. Genet.* **65**: 829-846

Seielstad, M., Bekele, E., Ibrahim, M., Tourè, A. and Traorè, M. 1999. A view of modern human origins from Y chromosome microsatellite variation. *Genome Res.* **9**: 558-567

Seminò, O., Santachiara-Benerecetti, A., Falaschi, F., Cavalli-Sforza and Underhill, P.A. 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am. J. Hum. Genet.* **70**: 265-268

Serre, D., Langaney, A., Chech, M., Teschler-Nicola, M., Paunovic, M., *et al.*. 2004. No evidence of Neanderthal mtDNA contribution to early modern humans. *PLoS Biol.* **2**: 0313-0317

Sharma, P., Singh, D.P., Fatma, N., Chylack, L.T. and Shinohara, T. 2000. Activation of LEDGF gene by thermal- and oxidative-stresses. *Biochem. Biophy. Res. Comm.* **276**: 1320-1324

Shinohara, T., Singh, D.P. and Fatma, N. 2002. LEDGF, a survival factor, activates stress-related genes. *Prog. Ret. Eye Res.* **21**. 341-358

Shun, M., Raghavendra, N., Vandegraaff, N., Daigle, J.E., Hughes, S., *et al.*. 2007. LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration. *Genes Dev.* **21**:1767-1778

Singh, D.P., Kimura, A., Chylack, L.T. and Shinohara, T. 2000. Lens epithelium-derived growth factor (LEDGF/p75) and p52 are derived from a single gene by alternative splicing. *Gene* **242**: 265-273

Singh, D.P., Fatma, N.P., Sharma, P., Hayakawa, K., Chylack, L.T., *et al.*. 2002. Structural and functional organization of the human lens epithelium-derived growth factor (LEDGF) gene promoter. *Invest. Ophthalmol. Vis. Sci.* **43**: E-Abstract 2345

Smale, S.T., Schmidt, M.C., Berk, A.J. and Baltimore, D. 1990. Transcriptional activation by SP1 as directed through TATA or initiator: specific requirement for mammalian transcription factor IID. *PNAS* **87**: 4509-4513

Soodyall, H., Vigilant, L., Hill, A.V., Stoneking, M. and Jenkins, T. 1999. mtDNA control-region sequence variation suggests multiple independent origins of an “Asian-specific” 9-kb deletion in sub-Saharan Africans. *Am. J. Hum. Genet.* **58**: 595-608

Spielman, R.S., Bastone, L.A., Burdick, J.T., Morley, M., Ewens, W.J., *et al.*. 2007. Common genetic variants account for differences in gene expression among ethnic groups. *Nat. Genet.* **39**: 226-231

Stec, I., Nagl, S.B., van Ommen, G.B. and den Dunnen, J.T. 2000. The PWWP domain: a potential protein-protein interaction domain in nuclear proteins influencing differentiation? *FEBS Letters*: **473**: 1-5

Stephens, M, Smith, N. and Donnelly, P. 2001. A new statistical method for haplotype reconstruction based from population data. *Am. J. Hum. Genet.* **68**: 978-989

Stoneking, M., Fontius, J.J., Clifford, S.L., Soodyall, H., Arcot, S.S., *et al.*. 1997. *Alu* insertion polymorphisms and human evolution: evidence for a larger population size in Africa. *Genome Res.* **7**: 1061-1071

Streisinger. G. 1966. Frameshift mutations and the genetic code. *Cold Spring Harb. Symp. Quant. Biol.* **31** (Abstract)

Streisinger, G. and Owen, J. 1985. Mechanism of spontaneous and induced frameshift mutation in bacteriophage T4. *Genetics* **109**: 633-659

Stringer, C.B. and Andrews, P. 1988. Genetic and fossil evidence for the origin of modern humans. *Science* **239**: 1263-1268

Taylor, M.S., Ponting, C.P. and Copley, R.R. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* **14**: 555-566

The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**: 928-933

Tindall, K.R. and Kunkel, T.A. 1988. Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Biochemistry* **27**: 6008-6013

Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R. *et al.*. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**: 1380-1387

Tishkoff, S.A., Goldman, A., Calafell, F., Speed, W.C., Deinard, A.S., *et al.*. 1998. A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am. J. Hum. Genet.* **62**: 1389-1402

Tishkoff, S.A., Gonder, M.K., Henn, B. H., Mortensen, H., Knight, A., *et al.*. 2007. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol. Biol. Evol.* **24**: 2180-2195

Tishkoff, S.A. and Kidd, K.K. 2004. Implications of biogeography of human populations for “race” and medicine. *Nat. Genet.* **36**: S21-S27

Tishkoff, S.A., Pakstis, A.J., Stoneking, M., Kidd, J.R., Destro-Bisol, G. 2000. Short tandem-repeat polymorphism/*Alu* haplotype variation at the PLAT locus: implications for modern human origins. *Am. J. Hum. Genet.* **67**: 901-925

Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., *et al.*. 2001. Haplotype diversity and linkage disequilibrium at human *G6PD*: recent origin of alleles that confer malarial resistance. *Science* **293** 455-462

Tishkoff, S.A. and Verrelli, B.C. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu. Rev. Genomics Hum. Genet.* **4**: 293-340

Tishkoff, S. and Williams, S. 2002. Genetic analysis of African populations: human evolution and complex disease. *Nat. Rev. Genet.* **3**: 611-621

Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., *et al.*. 2000. Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26**: 358-361

Vandekerckhove, L., Christ, F., van Maele, B., De Rijck, D., Gijsbers, R., *et al.*. 2006. Transient and stable knockdown of the integrase cofactor LEDGF/p75 reveals its role in the replication cycle of human immunodeficiency virus. *J. Virol.* **80**: 1886-1896

Vandergraaff, N., Devroe, E., Turlure, F., Silver, P.A. and Engelman, A. 2006. Biochemical and genetic analyses of integrase-interacting proteins lens epithelium-derived growth factor (LEDGF)/p75 and hepatoma-derived growth factor related protein 2 (HRP2) in preintegration complex function and HIV-1 replication; *Virology* **346**: 415-426

Vanegas, M., Llano, M., Delgado, S., Thompson, D., Peretz, M. and Poeschla, E. 2005. Identification of the LEDGF/p75 HIV-1 integrase-interaction domain and NLS reveals NLS-dependent chromatin tethering. *J. Cell Sci.* **118**: 1733-1743

Verrelli, B.C., McDonald, J.H., Argyropoulos, G., Destros-Bisol, G., Froment, A., *et al.*. 2002. Evidence for balancing selection from nucleotide sequence analysis of human *G6PD*. *Am. J. Hum. Genet.* **71**: 1112-1128

Voight, B.F, Kudaravalli, S., Wen, X. and Pritchard, J.K. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* **4**: e72

Vulliamy, T., Mason, P. and Luzzatto, L. 1992. The molecular basis of glucose-6-phosphate dehydrogenase deficiency. *Trends Genet.* **8**: 138-143

Wall, J.D. and Pritchard, J.K. 2003. Haplotype blocks and linkage disequilibrium in the human genome. *Nat. Rev. Genet.* **4**: 587-597

Watson, E., Bauer, K., Aman, R., Weiss, G., von Haeseler, A., *et al.*. 1996. mtDNA sequence diversity in Africa. *Am. J. Hum. Genet.* **59**: 437-444

Watson, E., Forster, P., Richards, M. and Bandelt, H. 1997. Mitochondrial footprints of human expansions in Africa. *Am. J. Hum. Genet.* **61**: 691-704

Weber, J.L., David, D., Heil, J., Fan, Y., Zhao, C., *et al.*. 2002. Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* **71**: 854-862

Weiss, K.M and Clark, A.G. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**: 19-24

Winkler, C., An, P. and O'Brien, S.J. 2004. Patterns of ethnic diversity among the genes that influence AIDS. *Hum. Mol. Genet.* **13**: R9-R19

Wu, D.Y., Ugozzoli, L., Pal, B.K. and Wallace, R.B. 1989. Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. *Proc. Natl. Acad. Sci.* **86**: 2757-2760

Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B., Kinzler, K.W., *et al.*. 2002. Allelic variation in human gene expression. *Science* **297**: 1143

Yoder, K.E. and Bushman, F.D. 2000. Repair of gaps in retroviral DNA integration intermediates. *J. Virol*. **74**: 11191-11200

Zhang, Z. and Gerstein, M. 2003. Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. *Nucl. Acids Res.* **31**: 5338-5348

Zhivotovsky, L.A., Rosenberg, N.A. and Feldman, M.W. 2003. Features of evolution and expansion of modern humans, inferred from genome-wide microsatellite markers. *Am. J. Hum. Genet.* **72**: 1171-1186

Zielske, S.P. and Stevenson, M. 2006. Modest but reproducible inhibition of human immunodeficiency virus type 1 infection in macrophages following LEDGF/p75 silencing. *J. Virol.* **80**: 7275-7280

Zietkiewicz, E., Yotova, V., Jarnik, M., Laskowska, M., Kidd, K.K., *et al.*. 1997. Nuclear DNA diversity in worldwide distributed human populations. *Gene* **205**: 161-171

Appendix 1

Figure A.1: Raw genotyping data obtained using the allele-specific PCR, RFLP-PCR and Pyrosequencing™ assays

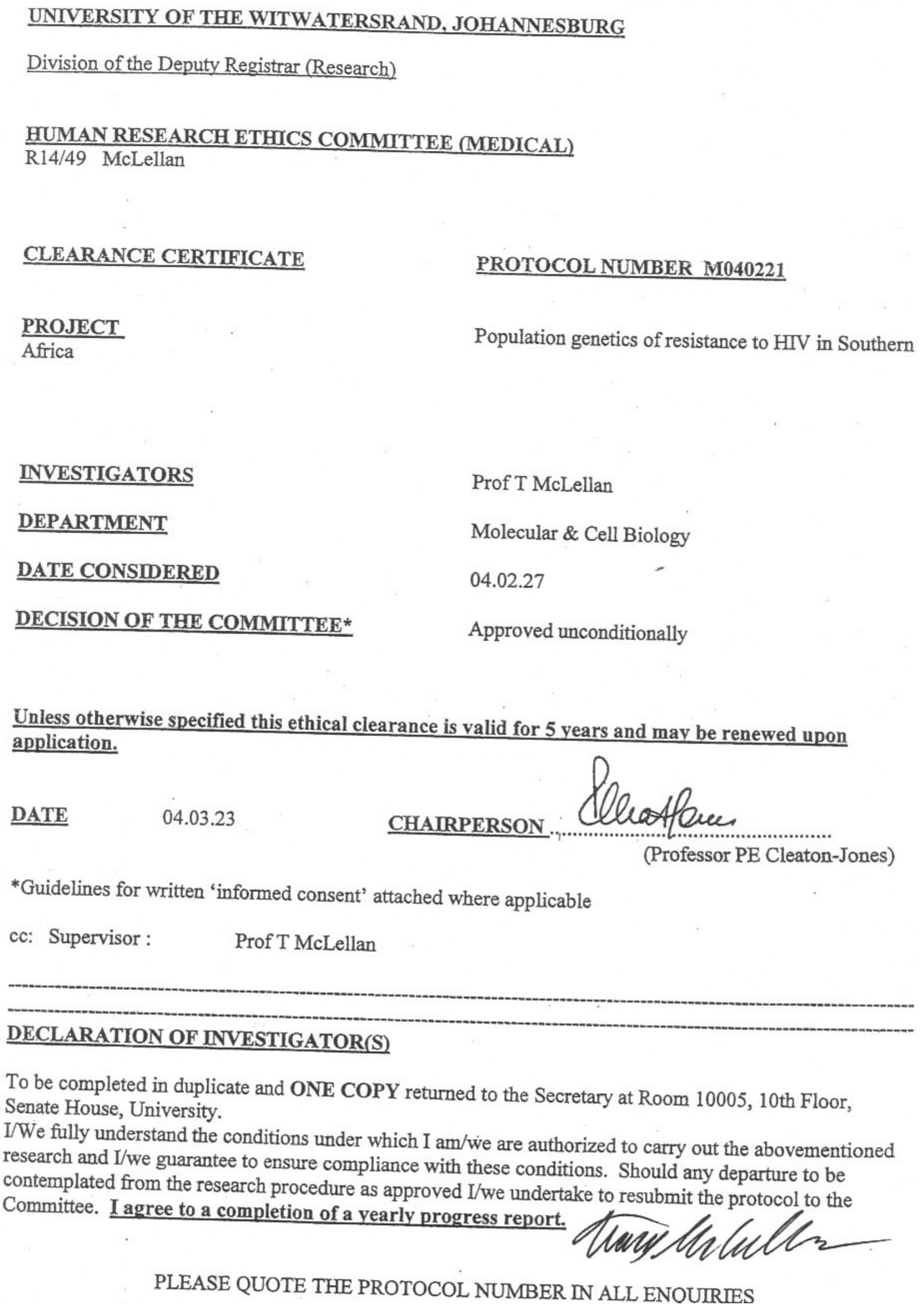
Sample No.	Genotype at -417	Genotype at + 31 040	Genotype at + 31 041	Genotype at +41 796	HIV-1 Status	Assigned Haplotype Pairs[1]	
206	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
207	GG	TT	CT	TCTTA/_	Unknown	ITCD	ITTB
208	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
209	_G	TC	CC	TCTTA/TCTTA	Unknown	ATCB	ICCB
210	_G	TT	CT	TCTTA/TCTTA	Unknown	ATCB	ITTB
211	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
212	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
213	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
214	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
215	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
216	_G	CC	TT	TCTTA/TCTTA	Unknown	ATCB	ITCB
217	_G	TT	CT	TCTTA/TCTTA	Unknown	ATCB	ITTB
218	_G	TT	CT	TCTTA/TCTTA	Unknown	ATCB	ITTB
219	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
221	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
222	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
223	_/_	TC	CC	TCTTA/TCTTA	Unknown	ATCB	ACCB
224	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
225	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
226	_G	TT	CC	TCTTA/_	Unknown	ATCB	ITCD
227	GG	TT	CT	TCTTA/TCTTA	Unknown	ITCB	ITTB
228	GG	CC	TT	TCTTA/TCTTA	Unknown	ICTB	ICTB
229	_G	CC	CC	TCTTA/TCTTA	Unknown	ACCB	ICCB
230	_G	TT	CT	TCTTA/TCTTA	Unknown	ATCB	ITCD
231	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
232	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
233	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
234	_G	TT	CT	TCTTA/TCTTA	Unknown	ATCB	ITTB
235	_/_	TC	CC	TCTTA/TCTTA	Unknown	ATCB	ACCB
236	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
237	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
238	_G	TC	CC	TCTTA/TCTTA	Unknown	ACCB	ITCB
239	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
240	_/_	TC	CC	TCTTA/_	Unknown	ATCD	ACCB
241	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
242	_/_	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ATCB
243	_G	TC	CC	TCTTA/TCTTA	Unknown	ATCB	ICCB
244	_G	TT	CC	TCTTA/TCTTA	Unknown	ATCB	ITCB
245	_/_	TC	CC	TCTTA/TCTTA	Unknown	ATCB	ACCB
300	_G	TT	CC		HIV$^+$		
301	_/_	TT	CC	TCTTA/TCTTA	HIV$^+$	ATCB	ATCB
302	_/_	TC	CC		HIV$^+$		
304	_/_	TC	CC	TCTTA/TCTTA	HIV$^+$	ATCB	ACCB

305	**_/_**	**TC**	**CC**	**TCTTA/_**	**HIV$^+$**	**ATCD**	**ACCB**
307	**_/_**	**CC**	**TT**	**TCTTA/TCTTA**	**HIV$^+$**	**ACTB**	**ACTB**
308	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
309	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
310	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
311	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
312	**_/_**	**TC**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACTB**
313	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
314	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
316	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
317	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
318	**GG**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ITCB**
319	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
320	**_/_**	**TC**	**CC**	**TCTTA/_**	**HIV$^+$**	**ATCB**	**ACCD**
321	**_G**	**CC**	**TT**	**TCTTA/TCTTA**	**HIV$^+$**	**ACTB**	**ICTB**
322	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
323	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
324	**_/_**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATTB**
325	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
326	**GG**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ITCB**
327	**_/_**	**TT**	**CC**	**TCTTA/_**	**HIV$^+$**	**ATCB**	**ATCD**
328	**_/_**	**TT**	**CC**		**HIV$^+$**		
329	**_G**	**TT**	**CC**		**HIV$^+$**		
330	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
331	**_/_**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATTB**
332	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
333	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
334	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
335	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
336	**_G**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ACCB**	**ITCB**
337	**_/_**	**TT**	**CC**		**HIV$^+$**		
338	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
339	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
340	**_G**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ACCB**	**ITCB**
341	**_/_**	**TC**	**CC**	**TCTTA/_**	**HIV$^+$**	**ATCD**	**ACCB**
342	**_G**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ICCB**
343	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
344	**_/_**	**TC**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACTB**
345	**_G**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ICCB**
346	**_G**	**TC**	**CC**	**TCTTA/_**	**HIV$^+$**	**ATCB**	**ITTB**
347	**_G**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITTB**
348	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
349	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
350	**GG**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ITTB**
352	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
353	**_/_**	**TC**	**CC**		**HIV$^+$**		
354	**_/_**	**TC**	**CC**		**HIV$^+$**		
355	**_G**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITTB**
356	**GG**	**CC**	**TT**	**TCTTA/_**	**HIV$^+$**	**ICTB**	**ICTD**
357	**_G**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITTB**
358	**_G**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ICCB**
359	**GG**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ITCB**

360	**GG**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ITCB**
361	**GG**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ICCB**
362	**GG**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ICCB**
263	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
364	**GG**	**TT**	**CC**	**TCTTA/_**	**HIV$^+$**	**ITCB**	**ITCD**
366	**_/_**	**TT**	**CC**		**HIV$^+$**		
367	**_/_**	**TC**	**CC**		**HIV$^+$**		
368	**_/_**	**TC**	**CC**		**HIV$^+$**		
369	**_G**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITTB**
370	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
371	**GG**			**TCTTA/TCTTA**	**HIV$^+$**		
372	**_/_**			**TCTTA/TCTTA**	**HIV$^+$**		
373	**_G**	**TT**	**TT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATTB**	**ITTB**
374	**_/_**	**TC**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACTB**
375	**_G**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITTB**
377	**_G**	**TC**	**CC**	**TCTTA/_**	**HIV$^+$**		
378	**_G**	**TT**	**CC**	**TCTTA/_**	**HIV$^+$**	**ATCB**	**ITCD**
379	**_G**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITTB**
380	**GG**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ITCB**	**ITTB**
381	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
382	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
383	**_/_**	**TT**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATTB**
384	**_/_**	**TC**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACTB**
385	**_G**	**TC**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATTB**	**ICCB**
386	**_/_**	**TC**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACTB**
387	**_G**	**TC**	**CT**	**TCTTA/_**	**HIV$^+$**		
388	**_G**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ACCB**	**ITCB**
389	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
390	**_/_**	**TC**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ACCB**
391	**_/_**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ATCB**
392	**_/_**			**TCTTA/TCTTA**	**HIV$^+$**		
393	**_G**	**TT**	**CC**	**TCTTA/TCTTA**	**HIV$^+$**	**ATCB**	**ITCB**
394	**GG**	**TC**	**CC**		**HIV$^+$**		
395	**_/_**			**TCTTA/TCTTA**	**HIV$^+$**		
396	**GG**				**HIV$^+$**		
397	**_/_**			**TCTTA/_**	**HIV$^+$**		
398	**_G**			**TCTTA/TCTTA**	**HIV$^+$**		
399	**_/_**				**HIV$^+$**		
400	**_/_**			**TCTTA/_**	**HIV$^+$**		
402	**_G**	**TC**	**CT**	**TCTTA/TCTTA**	**HIV$^+$**	**ATTB**	**ICCB**

[1] I and D represent the presence of an insertion and deletion, respectively. A and B correspond to their absence.

Appendix II



UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 McLellan

CLEARANCE CERTIFICATE PROTOCOL NUMBER M040221

PROJECT Population genetics of resistance to HIV in Southern
Africa

INVESTIGATORS	Prof T McLellan
DEPARTMENT	Molecular & Cell Biology
DATE CONSIDERED	04.02.27
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 04.03.23 CHAIRPERSON ....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof T McLellan

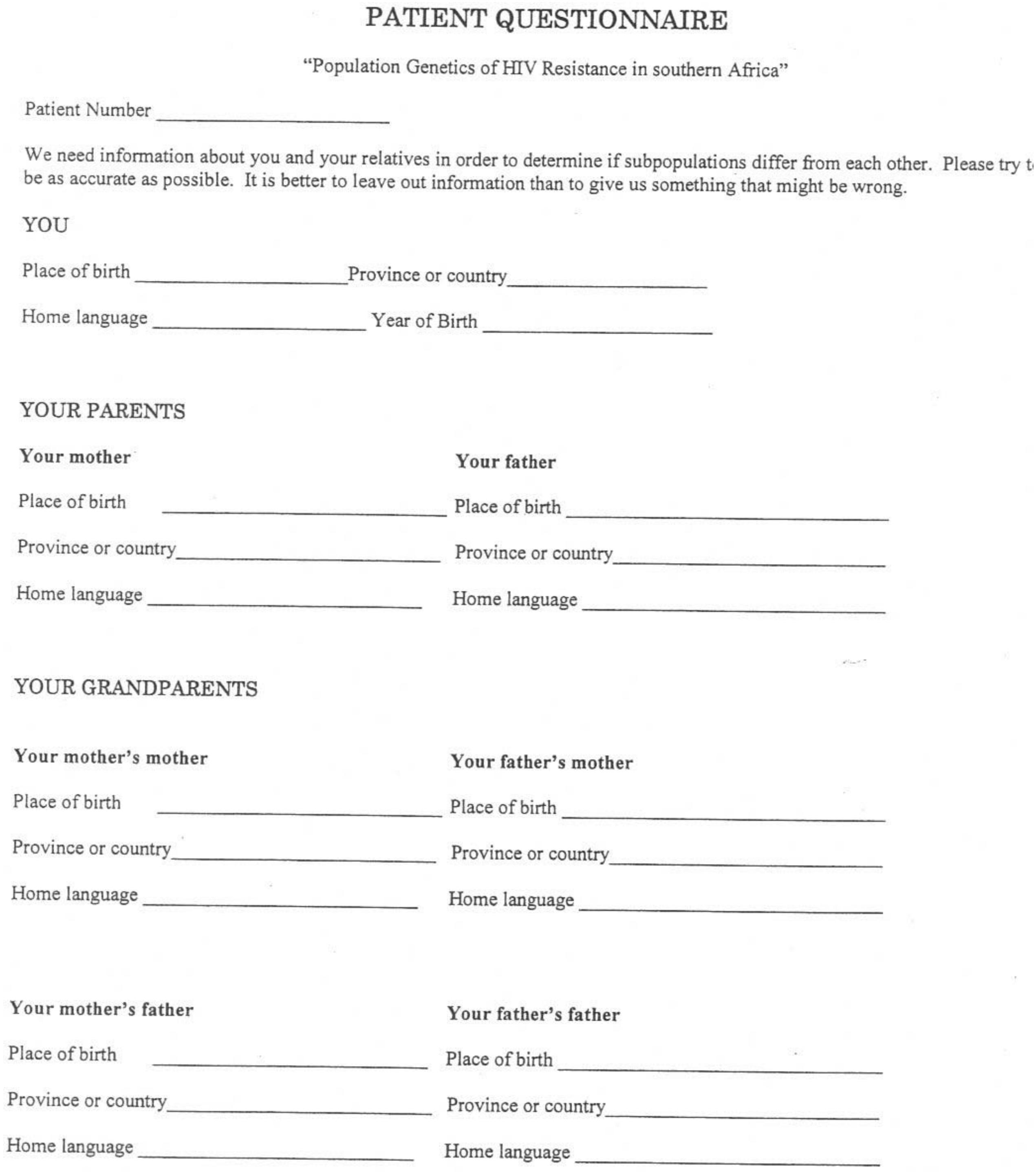
DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10005, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Figure A.1: Ethics clearance certificate obtained from the Human Research Ethics Committee at the University of the Witwatersrand

Appendix III



PATIENT QUESTIONNAIRE

"Population Genetics of HIV Resistance in southern Africa"

Patient Number ____________________

We need information about you and your relatives in order to determine if subpopulations differ from each other. Please try t be as accurate as possible. It is better to leave out information than to give us something that might be wrong.

YOU

Place of birth ____________________ Province or country____________________

Home language ____________________ Year of Birth ____________________

YOUR PARENTS

Your mother

Place of birth ____________________

Province or country____________________

Home language ____________________

Your father

Place of birth ____________________

Province or country____________________

Home language ____________________

YOUR GRANDPARENTS

Your mother's mother

Place of birth ____________________

Province or country____________________

Home language ____________________

Your father's mother

Place of birth ____________________

Province or country____________________

Home language ____________________

Your mother's father

Place of birth ____________________

Province or country____________________

Home language ____________________

Your father's father

Place of birth ____________________

Province or country____________________

Home language ____________________

Figure A.2: Questionnaire completed by blood donation volunteers, detailing their linguistic history and, where applicable, history of HIV-1 infection.

Acknowledgements

- My supervisor, Prof. McLellan for her support and guidance.
- The National Research Foundation (NRF) and the University of the Witwatersrand for providing the funding for this project.
- Dr. Patrick McPhale and Sister Dora for their assistance with sample collection.