SNP and Haplotype Characterisation of Apobec 3G, a Protein Involved in Retroviral Defence, in Black South Africans

Roshilla Ramdin

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, August 2012

DECLARATION

I declare that this is my own original, unaided work being submitted to the University of the Witwatersrand in fulfillment of the degree of MSc (Genetics and Developmental Biology)

Roshilla Ramdin

28 August 2012

DEDICATION

I would like to thank my dad Vishum, mum Urmila, my sister Raksha and my brothers and for their encouragement, patience, love and help throughout my academic career. I could not have achieved all that I have if it were not for their support. Thank you and I love you. To my daughter, Sanusha thank you for your understanding and love.

ACKNOWLEDGEMENTS

I would like to acknowledge the NRF for providing the funding for my studies,

Table of Contents

CHAPTER THREE – RESULTS

CHAPTER FOUR – DISCUSSION

List of Figures

List of Tables

List of Abbreviations

AA	African American
AMP	adenosine monophosphate
ARG	AIDS restricting gene
CCR5	chemokine receptor 5
GF	genotype frequency
GP	general population
GWS	genome wide screen
HIVNET	HIVNET 028
JHB	Johannesburg General Hospital
LD	linkage disequilibrium
MAF	minor allele frequency
MRCA	most recent common ancestor
NWM	New World monkeys
OWM	Old World monkeys
PPi	pyrophosphate
RAO	recent out of Africa
RT	reverse transcriptase
SNP	single nucleotide polymorphism
T_m	melting temperature
vif	viral infectivity factor
UNG	uracil DNA glycosylase

ABSTRACT

It is known that infectious agents elicit different responses in different individuals which strengthens the view that susceptibility and resistance to infectious diseases has a genetic component. These differences in susceptibility to disease can be observed in populations. *APOBEC3G* is a member of the cytidine deaminase gene family located on chromosome 22. It is crucial in non-permissive cells as it functions as part of the innate immunity system and is an inhibitor of the HIV-1 accessory protein vif.

The goal of the study was to develop genotyping assays and estimate allele frequencies. Thus, genetic variation within *APOBEC3G* was identified and characterized in black South Africans. Indirect genotyping assays were designed to amplify regions within the upstream non-coding region, and in exon 4 of the coding region of the gene. Selected polymorphisms were then genotyped using allele-specific PCR, RFLP-PCR and Pyrosequencing™ assays.

Reanalysis of sequence data from 2003 showed numerous SNPs were well represented. Comparison of sequence data at various SNPs showed that allele frequencies were similar to frequencies in other African populations. The only sequenced SNP that deviated from the frequencies in Ensembl was -590. Thus the sequencing was a useful tool for detection of variation. ASA proved to be the least reliable genotyping technique as the minor allele frequency of -571 (0.59) deviated from the published frequency of 0.894 in Africans. RFLP analysis

proved more reliable for genotyping -571 and H186R. The minor allele frequency was estimated to be 0.84 and 0.32 for -571 and H186R respectively. The frequency of H186R is similar to published data from An et al (2004) and Reddy et al (2010). If SNPs are in LD they occur together on the same haplotype more often than by chance. Usually SNPs that are in LD are in close proximity. However our data suggests -571 and H186R SNPs which are 5kb apart are not in LD. A LD map of chromosome 22 shows highly variable pattern of LD (Dawson et al, 2002). Widespread regions of nearly complete LD up to 804 kb in length are intermingled with regions of little or uundetectable LD. Haplotype analysis showed the most frequent haplotype was GA. This was the most frequent haplotype when the sample types were subdivided according to spoken language. in comparison to studies from An et al, (2004) *D'* of the two SNPs was estimated at 0.967. The linkage disequilibrium (LD) revealed a non-independence of allele segregation because the loci analyzed were strongly linked in the Apobec 3 G gene. The data are consistent with greater genetic diversity of African populations and can form the basis for further evaluation of the role of variation in this gene in response to HIV.

Introduction

1.1 Human Immunodeficiency Virus (HIV) and Acquired Immune Deficiency Syndrome (AIDS) in Sub-Saharan Africa

Human Immunodeficiency Virus (HIV) is transmitted by an exchange of body fluids (O'Brien & Nelson, 2004). The virus affects the immune system. It kills the $CD4^+$ cells, which are critical for the prevention of infections and diseases (Barre-Sinoussi *et al.*, 1983). Consequently, a loss of immunity results, allowing organisms that would normally be warded off to cause infections (O'Brien & Nelson, 2004). This is life threatening as it increases susceptibility to infectious diseases. This condition caused by HIV is called Acquired Immune Deficiency Syndrome (AIDS) (Barre-Sinoussi *et al.*, 1983). AIDS is clinically defined when the $CD4^+$ T cell count is less than 200cells/mm^3 or by being HIV positive and having and AIDS-associated illness (UNAIDS/WHO, 2007).

Statistics reveal that Sub-Saharan Africa has the highest prevalence of HIV infections. In this region there were 22.5 million adults living with HIV and approximately 1.7 million adults and children had become infected with the virus in one year (UNAIDS/WHO, 2007). These alarming statistics are cause for concern especially since Southern Africa accounts for one third of all new HIV infections and AIDS deaths globally. South Africa is the hardest hit country with the largest number of people infected and has the highest HIV prevalence as compared to any country (UNAIDS/WHO, 2007). The virus is spreading throughout the population and is not limited to high risk groups such as sex-workers (UNAIDS/WHO, 2005). Women and girls have become especially

vulnerable to HIV/AIDS and in 2007 it was estimated that 61 % of adults in South Africa living with HIV were women (UNAIDS/WHO, 2007). The key to curbing the spread of the virus is to prevent transmission and to implement efficient and practical strategies to help those infected with and affected by HIV/AIDS.

1.2 HIV diversity in Africa

There are two types of HIV; HIV-1 and HIV-2. HIV-1 is further divided into groups called Groups M, N, O and P. These groups are further subdivided subtypes A-K. HIV-1 group M accounts for more than 95 % of all HIV infection around the world apart from HIV-2 infections in certain parts of Africa (Lemey *et al*, 2003). All subtypes of HIV-1 are found in Africa. Subtype B is found mainly in Europe, the Americas, Japan and Australia (Heeney *et al*, 2006).

The origins of HIV-1 have been very contentious. The first HIV-1 like virus called Simian Immunodeficiency Virus (SIVcpz) was characterised in captive chimpanzees in 1989. Initially the SIVcpz virus was thought not to be responsible for the disease in humans as it could not be determined if wild chimpanzees are naturally infected by the virus. Chimpanzees are divided into four subspecies; the western, Nigerian, central and eastern chimpanzees (Gagneux et al, 1999). Each subspecies occupy a different geographical niche. Sequence analysis of all the SIVcpz strains from captive chimpanzees described by Peeters et al (1989) and Peeters et al (1992) showed that all these strains clustered together within the central chimpanzee subspecies and also formed one cluster within all HIV-1

strains. This is indicative of the central subspecies, which encompasses Cameroon and the Congo, being the causative agent of HIV-1 virus in humans.

DNA sequencing of archival samples from Zaire (now the Democratic Republic of the Congo (DRC)) was used to date the origin of HIV-1 and confirmed the evolutionary history (Worobey *et al*, 2008) as this is the epicentre of diversity. The archival samples are designated ZR 59 (a blood plasma sample from 1959) and DRC60 (biopsy specimen from female patient in the DRC). The phylogenetic analysis demonstrated a short nodal distance between the two. Particularly the DRC60 sequence was found to cluster close to the A subtype ancestral node in the phylogenetic tree while the ZR59 sequence clustered closer to the subtype D (Worobey *et al*, 2008). This indicates that even 50 years ago group M of HIV-1 strains had evolved into distinct subtypes that were circulating within the populations of this region. The phylogenetic analysis indicates too that there is substantial genetic diversity between the two ancestral sequences. This further confirms that the virus was present in humans long before the epidemic was characterized.

An interesting point has been raised that urbanization has played a vital role in the rise of the disease in Africa because the two ancestral sequences cluster with other strains from the same region rather than the same subtype, giving rise to viral lineages which are more diverse within viral subtypes. Thus it was concluded that

the diversification of HIV-1 group M viruses began in Kinshasa (Keele *et al*, 2006).

The causative agent of HIV-2 is known to be SIVsm (SIV from Sooty Mangabeys). The Sooty Mangabeys naturally are found in the forest of Senegal east to Ghana. The origin of HIV-2 has not been so contumacious. A natural reservoir of the virus was detected in these monkeys as early as 1989 (Hirsch *et al*, 1989). Unlike HIV-1, each HIV-2 subtype was apparently the result of independent cross-species transmission. The most recent common ancestor of HIV-2 subtype A was dated to be ~1940 and subtype B 1945 in Guinea-Bissau (Lemey *et al*, 2003). Subtypes A and B are linked to the epidemic in this region while other subtypes have been identified in people infected with a single subtype. Like HIV-1 transmission of the virus has been marked by a period of untraceability followed by an exponential rise in infections. This rise of infections is estimated to occur around the same time as the War of Independence between 1963 and 1974. Once again it is evident that viral epidemics are reliant on socio-economic conditions.

HIV -1 subtype C is well documented as being the predominant strain in South Africa (Rodenberg *et al*, 2001 & Papathanasopoulos *et al*, 2002). However there is evidence of minor strains such as non-C subtypes and various recombinant subtype viruses also being present in South Africa. In addition HIV-1 subtype B is

present in the homosexual population (van Harmelen *et al*, 1997 & Williamson *et al*, 1995).

1.3 HIV Life Cycle

The HIV infection begins with HIV adsorption to the CD4 cell via the gp120 molecules exposed on the surface of the virion. This binding induces a conformational change in gp120, another protein of the viral envelope, allowing it to bind CCR5 (Doms and Trono, 2000). The conformational change allows gp41 which becomes exposed to initiate fusion of the viral and host membranes. The virus then releases its contents (containing the two RNA strands) into the cytoplasmic compartment of the host cell (Figure 1.1). Once in the host cell there is a partial uncoating of the capsid to expose the RNA strands. Viral reverse transcriptase converts the RNA to DNA, which is a much more stable molecule in the cell, and can then recombine with the host's DNA. The DNA then enters the host nucleus where together with viral integrase it is integrated into the host genome (Doms and Trono, 2000). Once it is integrated into the host genome it is called a provirus. This provirus can remain dormant but become active at any stage. If the provirus is active it will express viral proteins for new virions. It uses the host cell's own enzymes to transcribe the double stranded DNA to mRNA. Once the viral mRNA is processed in the nucleus, it is transported to the cytoplasm where the viruses diverts the host protein synthesis machinery to produce the viral proteins Env(gp160), Gag, Gag-pol, Vif, Vpr, Vpu, Rev, Tat and Nef. The initial env protein is processed in the ER and Golgi into the gp120 and gp41. The gp120 is glycosylated. The gag and gag-pol polyproteins aggregate

near the membrane and interact with plasma membrane and the gp41 present in the membrane (Doms and Trono, 2000). As the gag and gag-pol aggregate at the plasma membrane the virion begins to form and will be extruded from the host cell membrane. As budding occurs the virion takes the host cell lipid layer in which the env protein is bound. The virion then undergoes maturation. A virally-encoded proteinase cleaves the precursor gag, gag-pol into functional proteins.

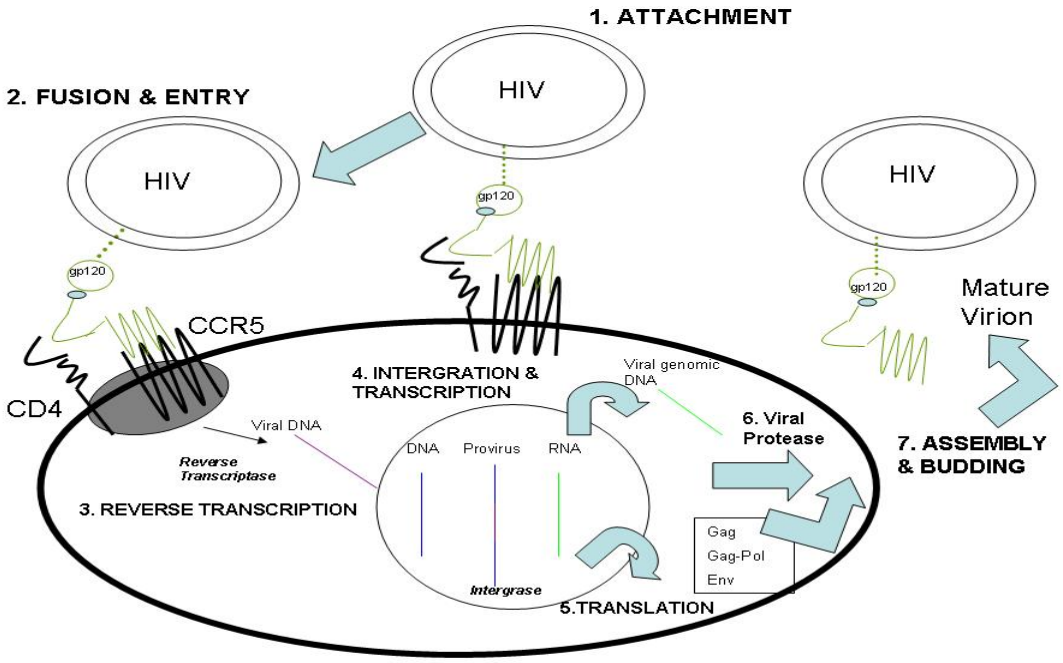

Figure 1.1 Schematic overview of HIV life cycle.

1.4 Host proteins involved in HIV

Many approaches have been used to identify host proteins important in HIV infection. Classical analysis of host genetics and its involvement in viral genetics made use of candidate gene studies, leading to the discovery of numerous AIDS

restriction genes that effect susceptibility to viral infection (Hutcheson *et al*, 2007). Candidate gene approaches are hypothesis-based. A candidate gene is studied for association in order to ascertain any frequency differences between the case and control. The advantage of this method is that the study population need not be large, though it is often better to have a larger study population to ensure accuracy of predictions. The disadvantage of this method lies in the fact that the investigator must have some knowledge about the candidate gene or the study could be uninformative. Most candidate gene studies have focused on European populations (or populations of European origin); little information is available for the effects of variation in these genes in African populations where the epidemic is expanding at an alarming rate. Nonetheless, a large number of population studies have focussed on African Americans. This does not give us a complete picture of the disease in Africans due to admixture of genes, but is a good platform from which tobase additional studies.

Candidate gene products include chemokine receptors and their variants (CCR5), chemokine receptor ligands (SDF), cytokines (IL), the HLA system and various factors involved in cellular immunity such as TRIM5α, APOBEC3G and 3F(Sheehy *et al*, 2000).

One must study the effects of important genes in Africans to gain the correct and comprehensive picture of disease pathogenesis. APOBEC3G is a good candidate

for HIV restriction because it allows the expression of an antiviral phenotype in non-permissive cells; consequently this innate immune defence is an alternative in the design of new therapies. In addition, polymorphisms in the host factors characterized in genetic studies are found predominantly in the promoter region or regulatory regions. Hence, this region is a good starting point to investigate the variation in *APOBEC3G* and its contribution to HIV/AIDS pathogenesis. In addition, numerous variants which modulate HIV pathogenesis have been discovered by sequencing the upstream non coding region such as the Duffy Antigen Receptor for Chemokines (DARC) (Winkler *et al*, 2004).

The innate cellular defence system is crucial in detecting and limiting infection caused by viruses such as HIV. Apolipoprotein B mRNA editing enzyme, catalytic polypeptide – like 3G (APOBEC3G*)* is part of this system and was discovered via the candidate gene approach. Non-permissive cells which possess APOBEC3G allowed for an antiviral phenotype that is overcome by viral protein Vif (Sheehy *et al*, 2000).

1.5 *APOBEC 3G*

APOBEC3G is a member of the cytidine deaminase gene family located on chromosome 22. It is positioned at 22q13.1- q13.2. It is one of seven genes clustered on chromosome 22. The *APOBEC3G* gene is 10 613 bases in length and is transcribed into a protein 384 amino acids long. The protein is homodimeric

and is mainly localised within the cytoplasm but small amounts are found within the nucleus. As mentioned earlier this protein is an inhibitor of HIV-1 accessory protein Vif which will be discussed later.

APOBEC3G has eight exons arranged in tandem (Jarmuz *et al*, 2002) (Figure 1.2). Exons 2, 3, 4 are duplicated within exons 5, 6, 7, respectively. The duplication results in the presence of two active sites (exons 2, 5), two linker regions (exons 3, 6) and two pseudo-catalytic domains (exons 4, 7) (Jarmuz *et al*, 2002). The active site is responsible for target binding and specificity. It contains a zinc-finger motif where zinc-binding ligands such as histidine, glutamic acid, proline and cysteine are critical for the functioning of the putative deaminase. The two aromatic amino acids phenylalanine and tyrosine are responsible for the binding of the target to the active site (Jarmuz *et al*, 2002). Point mutations in the zinc-finger motif diminished the activity of APOBEC3G, once again illustrating the importance of this domain in establishing an antiviral phenotype in the absence of Vif (Mangeat *et al*, 2003). The pseudo-catalytic domain however lacks the zinc-binding ligands and thus has no target binding abilities. It is hypothesized that this domain may stabilize the hydrophobic core of the active site in addition it may bind auxiliary factors essential for deamination (Jarmuz *et al*, 2002).

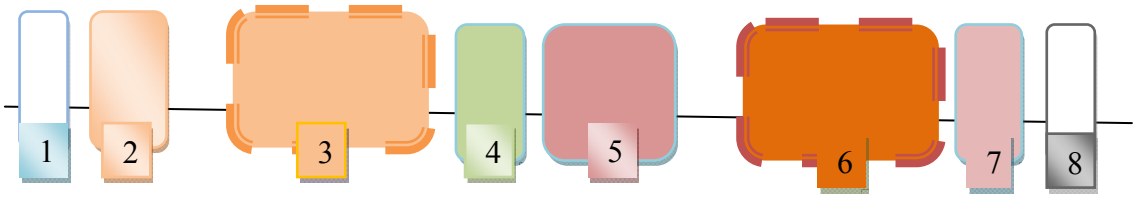

Figure 1.2 Schematic representation of exons within *APOBEC 3G*. The translated and untranslated portions of the exons are indicated with filled and unfilled boxes respectively. The exons encoding the zinc-coordinating domain are shown as weighted dashed boxes.

1.6 Origin and Evolution of APOBEC Deaminases

APOBEC1 was the first member of this family to be described. Its crystal structure and gene organization were based on *E. coli* cytidine deaminase. Homology modelling revealed that by removing the sequences of nucleotides termed the gaps from *E. coli* cytidine deaminase (ECCDA) the signature sequence of APOBEC 1 was derived (Chester *et al*, 2000). Homology modelling works by alignment of the amino acid sequence of the catalytic domains of the ECCDA and then APOBEC protein's amino acid sequence is fitted to these domains (Huthoff and Malim, 2005). Other approaches have been used to elucidate the structure and evolution of APOBEC proteins. In one instance, DNA and protein sequences of all deaminases were pooled from BLAST searches (Conticello *et al*, 2005). The BLAST searches were focused on the deaminases that had the first cluster of the active site containing a single zinc ligand. These sequences were then used to construct a phylogenetic tree. From these trees the AID/APOBEC family was very distinct from other cytidine deaminases as they contained the characteristic zinc

coordinating domain. *AID/APOBEC1/APOBEC3* was clustered together and has diverged from APOBEC 2. *APOBEC2* and *AID* each had homologs that could be traced back to bony fish. However *APOBEC1* did not have any non-mammalian homologs, suggesting that it was actually derived from AID. In addition, *APOBEC2* was an ancestral sequence from which the other members of the family have diverged (Conticello *et al*, 2005).

1.6.1 Evolution of the APOBEC3 family

It has been proposed that APOBEC3 zinc domains were the result of the diversification of two ancestral domains that either constituted a double-domain protein or a single domain protein (Conticello *et al*, 2005). This duplication gave rise to seven *APOBEC3* genes in humans designated *APOBEC3* A, B, C, D, F, G, H (Figure 1.3).

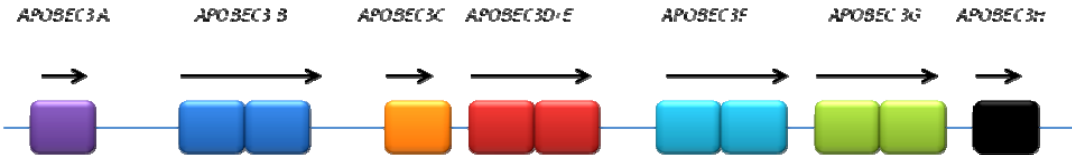

Figure 1.3 APOBEC 3 locus position on chromosome 22.

The sequence which was initially assigned to *APOBEC3E* was later found to encode a second domain of *APOBEC3D* (Wedekind *et al*, 2003 & Conticello *et al*, 2005). This 130kb *APOBEC3* cluster is syntenic to the mouse APOBEC 3 gene found on chromosome 15. However the human APOBEC 3 cluster is much

expanded. *APOBEC3* homologs are present in primates but absent from bony fish and chicken genomes. In addition the sequences flanking the *APOBEC3* are highly conserved in bony fish and chicken suggesting that the human *APOBEC3* locus may have evolved much later in mammals. This confirms that like *APOBEC1*, *APOBEC3* orthologs are restricted to mammals and evolved from AID. The duplication and divergence of *APOBEC3* locus is thought to occur from two ancestral domains termed Z1 and Z2. It is hypothesized that Z1 proteins in primates diverged to form single and double domain proteins within the *APOBEC3* locus.

1.7 HIV-1 viral infectivity factor (vif)

HIV-1 encodes a basic 23 kDA protein designated Vif. The protein is translated from a singly spliced mRNA and is located in both the cytosol and the nucleus (Turner & Summers, 1999). The C-terminal domain is functionally important and binds many membranes and is associated with Gag precursors (Khan *et al*, 2001). Deletion of this domain abolishes Vif activity. Comparison of the Vif codon of lentiviruses and HIV-1 infected individuals revealed that the *Vif* is consistently conserved (Sova *et al*, 1995). It is postulated that this limited variability in the gene is important in natural HIV-1 infection.

1.8 The Interaction of APOBEC3G and Vif

In permissive cells infected with Vif defective HIV, APOBEC3G is packaged into the virions. The encapsidated APOBEC3G then deaminates the dC of the minus

strand of the cDNA. This converts the cytosine to uracil and the uracil DNA glycosylase (UNG) then removes this base (Gu & Sundquist, 2003 & Vartanian *et al*, 2003). The unstable DNA fails to integrate into the host genome to form a provirus and it is consequently degraded. If the dU escapes degradation by UNG and the mutation becomes incorporated to the HIV genome it will lead to large proportion of G to A mutations, which may disrupt the open reading frame leading to non-infectious or weakly infectious virions (Gu & Sandquist, 2003 & Vartanian *et al*, 2003). Wild type virus that infects non permissive cells ensures that Vif binds to APOBEC3G, and this blocks viral uncoating and thus preventing deamination.

In non-permissive cells, Vif has a three-fold action on APOBEC3G. It can bind to it and target it for degradation by proteasomes, it can mediate its destruction via poly-ubiquitination and it can stop encapsidation of the APOBEC3G into newly produced virions in the absence of degradation. The region of APOBEC3G implicated in this binding of Vif is the amino terminal (Zhang *et al*, 2008). The process is detailed below.

The most complex function of Vif is its inhibition of APOBEC3G via poly-ubiquitination which is a proteolytic modification of proteins. The enzymes E1, E2, E3 are involved in the covalent conjugation of ubiquitin to a substrate consequently ensuring that the protein is degraded by the proteosome. Classically

this process is divided into three stages: Activation, Conjugation, and Ligation. Initially there is an ATP-dependent activation of ubiquitin by E1. Following this, the activated substrate is conjugated to E2 via a thioester bond. E2 acts with E3 to transfer the substrate to the target. An iso-peptide bond is formed between the terminal lysine of the substrate and the C-terminal glycine of the ubiquitin. The target is poly-ubiquinated and is tagged to be degraded by the 26S proteosome. This method of ubiquitination is relatively conserved even though all the components of the process may not be characterized.

Vif interacts with cellular proteins Cul 5, Elongin B & C and Rbx1 (He *et al*, 2008). Vif is probably the F-box protein, which determines target specificity. These proteins form SCF-like complex, more specifically, it bears resemblance to the VCB-like complex (Figure 1.4). Both of these complexes belong to an E3 ubiquitin ligase super-family. These complexes are biologically important as they selectively bind and ubiquinate specific proteins, targeting them for destruction (Yu *et al*, 2003). They are also vital in regulating and stabilizing signal transduction pathways and maintaining the cell cycle. Over-expression of Rbx1 in the presence of APOBEC3G decreases infectivity of the virus and Cul 5 inhibits ubiquitination of APOBEC3G. Vif interacts with this complex via a conserved motif SLQXLA because mutations in this motif drastically reduces the interaction between Vif and SCF-like complex (Yu *et al*, 2003, Mehle *et al*, 2006). Although mutations in this motif reduce interaction with the complex, it does not affect the interaction of Vif with APOBEC3G. This is direct evidence that Vif activity alone

is not sufficient to overcome the antiviral action of APOBEC3G. However, the cellular target of the SCF-like complexes remains elusive. Proteasomes have also been implicated in Vif functioning. When a proteosome inhibitor is added to non-permissive cells Vif fail to exclude APOBEC3G from the resulting progeny and the viruses are weakly infectious (Yu *et al*, 2003). Thus, Vif in conjunction with SCF-like complexes and proteosomes are essential in suppressing host antiviral phenotype.

Vif is established to interact closely with the components, which are collectively termed an E3 ligase. They are said to be an E3 ligase as they have a similar structure to the SCF complexes. Rbx1 is an important component of the complex and interacted intimately with Cul 5 as seen in mutagenesis studies. Mutants in Cul 5 with no Rbx1 binding affect the production of infectious virions in non-permissive cells (Yu *et al*, 2003). Vif associates with the E3 ligase by means of a conserved SLQ motif (Yu *et al*, 2003). Mutation in this motif decreases the association of Vif with E3 ligase but not with APOBEC3G. Consequently it is speculated that Vif interacts with the ligase and simultaneously with APOBEC3G acting as a bridge between the two facilitating ubiquitination. The N terminus of Vif binds APOBEC3G via its N terminal residues 54 -124. The zinc binding domain of Vif has two conserved cysteine that bind Cul5 (Mehle *et al*, 2006).

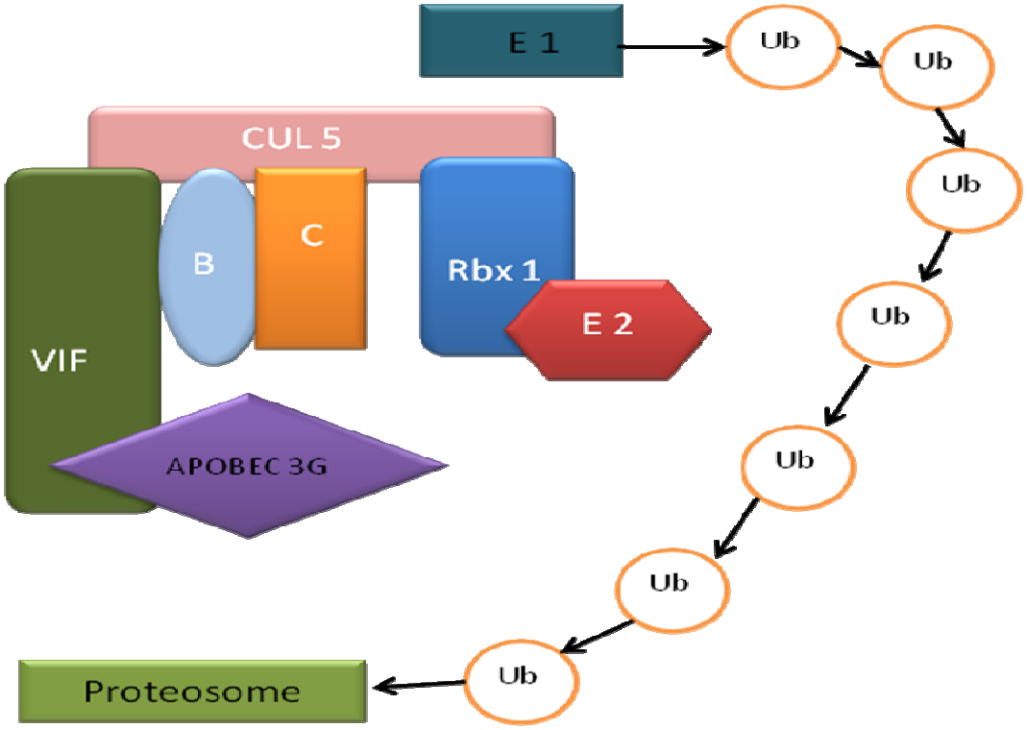

Figure 1.4 The interaction of Vif and APOBEC3G in non-permissive cells.

The dimerization domain of Vif is functionally important in blocking the incorporation of APOBEC3G into virions (Miller *et al*, 2007). This "proof of concept" was proved by using antagonists against Vif dimerization. This process is localized to the SQL motif. The agonists disrupted the expression of Vif in addition to the dimerization of Vif (Miller *et al*, 2007). This disruption facilitated the incorporation of APOBEC3G into subsequently produced virions and subsequently decreased the HIV infectivity. The regulation of the APOBEC3G by Vif is also mediated by functionally important residues such as amino acid 124 to 127. These are integral to packaging the APOBEC3G into virions while the amino acid 128 determines species specificity.

1.9 Selection of APOBEC3G and Vif

The interaction between APOBEC3G and Vif is antagonistic. APOBEC3G is under a positive selection pressure to decrease the activity of the Vif, while Vif is adapted to enhance the relationship between the two. This antagonistic relationship drives the process of evolutionary change. Comparative sequence data from non human primates such as Old World monkeys (OWM) New World monkeys (NWM) and hominids indicate that APOBEC3G has been under positive selection pressure in primates for at least 33 million years (J, Lui et al, 2010 & L, Sawyer, 2004).

A characteristic of positive selection is the excess of non-synonymous changes (changes which alter the amino acid sequence) over synonymous changes. Typically the non-synonymous changes are harmful to the organism and are removed from the lineage. However examination of *APOBEC3G* reveals that non-synonymous changes are far greater that synonymous changes which is the driving force for the fixation of variants with altered proteins (Meyerson and Sawyer, 2011). This ultimately affects how the gene variants interact with one another.

Interestingly the selection pressure acting on APOBEC3G appears to be ancient (Sawyer *et al*, 2004). APOBEC3G in Old World monkeys and hominids appear to have diverged from each other 23 million years ago. The New World monkeys,

Old World monkeys and hominids shared a common ancestor 33 million years ago. This selection is not limited to a specific domain. It appears that different domains within the APOBEC3G proteins have been selected in different primates through the years (Sawyer *et al*, 2004). It is thought that the positive selection of APOBEC3G is favored by the changing the charge of the amino acid. This is supported by the fact that the Asp 128 is conserved in hominids but Lys 128 in conserved in OWM because when the aspartic acid in the human APOBEC3G is replaced with the lysine that is found naturally in African Green Monkey protein, it becomes resistant to HIV-1 Vif but not SIV Vif (Zhang & Webb, 2004). One can infer that the duplication of the primate APOBEC3 locus is in direct response to viral infectivity. It may serve as a means for the host to further protect itself from infection.

1.10 DNA Polymorphisms

Innate cellular genes such as *APOBEC3G* are under tremendous selection pressure as indicated previously. The evidence of selection is the variation or polymorphisms found within this gene. There are different polymorphisms that distinguish individuals within a species (Syvanen, 2001). Insertions, deletions and simple sequence repeats fall into the category of length polymorphisms the second category of polymorphism consists of single base changes. This category comprises point mutations and single nucleotide polymorphisms (SNP's). The most common type of variation within the human genome is the SNP. They occur at one out of every thousand base pairs (Syvanen, 2001) but recent evidence

shows that this frequency is much higher (Sabeti *et al*, 2007). The consequence of detecting and characterizing SNP's are varied. SNP is in the regulatory or coding region result in synonymous or non-synonymous changes. Synonymous changes so characterized because there is change of nucleotide but not amino acid. Non-synonymous changes are classed as being missense, conservative or non-conservative mutations. They may also cause the formation of a premature stop codon.

1.11 Linkage disequilibrium and haplotypes

Mendel's postulates state that there is an independent assortment of genes so that each individual/offspring has an equal chance of inheriting either gene of a pair. The conclusion is that these genes are in linkage equilibrium. Though, when variation at two sites is inherited together more often than by chance then these genes are said to be in linkage disequilibrium (LD) (VanLiere & Rosenberg, 2008). Often genes and SNPs are inherited as part of a unit with other genes which lie close to it. This unit is referred to as a haplotype. LD has an important part to play in genetic studies because it can be used to infer the role of SNPs in disease.

1.12 Clinical significance of variation within *APOBEC3G*

APOBEC3G variants have a profound influence on the progression to AIDS in a study of a cohort of seroconvert, seropositive, seronegative participants of European and African American descent (Winkler *et al*, 2004). Seven SNP were identified within the gene. Three were in the putative regulatory region (-571, -199, -90), one the codon 3 (F119F), one in exon 4 (H186R) and two within the introns. This study revealed through haplotype analysis that there are six frequent haplotypes which cause these SNPs to be inherited together. Of particular interest is the frequency of substitution of histidine to arginine at amino acid position 186 was higher in African Americans (AA) than European Americans (EA). In particular the 186R allele was associated with faster progression to AIDS in the African Americans. The frequency was 37 % in AA as compared to 29 % in EA (Winkler *et al*, 2004). However, these observations have not been confirmed in other studies. A recent study investigated this polymorphism in Indians. The 186R allelic variant could not be found in the study sample and thus no conclusions could be drawn about its influence on disease progression in this study population (Rathore *et al*, 2008). There is further evidence that APOBEC3G does not have an association with AIDS progression (Do *et al*, 2005). Genotyping of a French cohort which contained subpopulations of seropositive individuals with slow progression and fast progression and healthy control subjects failed to show significant association of variation in *APOBEC3G* with disease progression. Twenty-nine polymorphisms were identified; these included 14 novel polymorphisms all with the exception of one found within the coding region of *APOBEC3G*. Particularly the P-value calculated at non-synonymous change from histidine to arginine at position 186 in exon 4 was calculated at 0.69 in pair wise

linkage disequilibrium analyses (Do *et al*, 2005). If there was some significant association between the polymorphism and estimated haplotypes then P value would be between 0.05 and 0.10. In contrast the P values for the estimated haplotypes and the arginine to histidine substitution within the AA seropositive sub populations studied was less than 0.024 indicating statistical significance (Winkler *et al*, 2004). However there is some consistency amongst the different studies; variation in *APOBEC3G* showed no association with AIDS progression in Europeans.

The distribution of allele frequencies and haplotypes amongst various ethnic groups gives an indication as to the importance of population demographics in the study of disease progression. It bears testament that ancestry has a strong influence on the interaction of populations with viruses.

1.13 Origin of Modern Humans

The patterns of LD and haplotypes have been used as tools for elucidating the genealogical and demographic history of populations. There are two schools of thought. The least popular idea is that race or ethnicity can be used to predict genetic classification (Tishkoff and Kidd, 2004, Jorde and Wooding, 2004, Mountain and Risch; 2004). Historically race has been classified according to biological factors such as skin colour, morphology. This in itself is complicated and not always correct; traits that produce phenotypic differences are a result of

genetic component adapting to the environment in which an individual lives. In contrast ethnic races are clustered in groups but this is a consequence of the geographic expansion of population out of Africa (Tishkoff & Kidd, 2004). This expansion has not given rise to any race specific genes.

However the most popular ideology is that LD patterns and population haplotypes are useful in determining the geographical distribution and lineage of populations (Rosenberg et al, 2002, Lane et al, 2002). These studies show that populations that arise from the same geographic region have similar LD patterns suggesting that the origins of human populations is important in assessing the genetic diversity

The accepted theory for the origin of modern humans is the Out of Africa theory popularly called the Recent Out of Africa theory (ROA) (Cavalli-Sforza *et al*, 1992 & Cavalli-Sforza *et al*, 1997). The theory stipulates that *Homo sapiens* evolved in Africa around 1 million to 2 million years ago (Cavalli-Sforza *et al*, 1988 & Cavalli-Sforza *et al*, 1992 & Cavalli-Sforza, 2006). The Old World (OW) was inhabited by numerous morphologically varied groups of hominids. Evidence suggests that in Africa and the Middle East *Homo sapiens* was present, in Europe *Homo erectus* and *Homo neanderthalensis* were present. It is argued that by roughly 30 000 yrs ago *Homo sapiens* was dominant in Europe.

Early human migration can be deciphered using mitochondrial DNA and the Y-chromosome (Ingman *et al*, 2003 & Cavalli-Sforza, 2000). Human mitochondrial DNA haplotyping shows that all present-day women have inherited mitochondria from one woman called Mitochondrial Eve, who lived in Africa 160 000 years

ago (Cann *et al*, 1987 & Vigilant *et al*, 1991). Similarly Y-chromosomal haplotyping confirms that present day men have inherited their Y chromosome from a common male ancestor present in Africa 60 00 years ago.

The first exodus out of Africa to Near East occurred some 60 000 years ago (Cavalli-Sforza *et al*, 1988 & Cavalli-Sforza *et al*, 1992 & Cavalli-Sforza, 2006). It is believed that the subsequent migrations to East Asia, Europe & Australia, and South Asia occurred 30 000, 40 000 and 50 000 years ago respectively (Wilson & Cann, 1992 & Bakewell, Oliver and Haas, 2007). These migrations are confirmed by tracing the mitochondrial and Y-chromosome lineages and numerous nuclear markers. When genetic and archeologically data are matched with linguistic data it seems to support the earlier evidence that the oldest language families are in Africa (Cavalli-Sforza, 2006). This mirrors what Charles Darwin postulated over a hundred years ago that a genetic tree of evolution will encapsulate the tree of linguistics evolution.

Africans have the highest genetic diversity owing to the largest number of variable genes and alleles (Jorde *et al*, 2000 & Tishkoff & Williams, 2002 & Cavalli-Sforza *et al*, 1997 & Jakobson *et al*, 2008). In comparison, non-African, populations do not have high genetic diversity predominantly due to genetic drift, which occurred during the migrations of modern humans out of Africa and the resulting small populations (Tishkoff & Williams, 2002). In addition, Africans

have a lower LD than other populations. The lower LD is indicative of a decrease in genetic distance in 'haplotype blocks' (Shifman et al, 2003).

1.13.1 Bantu Expansion

It has been established that the genetic history of Africa is important as it is the centre of the evolution of anatomically modern humans. The genetic footprint of Africa lies deeply rooted in the migrations of early humans. Human population adaptation to viruses shapes the evolution of the genome. That is they are provided traces of adaptive evolution in human populations. Therefore detecting the association of SNPs in HIV restricting genes can be very valuable in Africans because local patterns of LD have been characterized across the genome for Sub-Saharan South Africans (Donfack *et al* 2006).

The Bantu are classified as a group of related individuals who originated in West Africa. Bantu, a word meaning "people", is also used for a collection of related languages spoken by these people. The great Bantu Migration started with the migration of one group to Cameroon, coastal Congo and inland Kinshasa (Diamond *et al*, 2003, Holden *et al*, 2001). Later a second wave of migration occurred to the east (Diamond *et al*, 2003, Holden *et al*, 2001). These Bantu occupied regions now known as Uganda. Then groups from Uganda spread farther east into Kenya and Tanzania while others continued southward to colonize areas that are now the countries Zimbabwe, Botswana, Mozambique, and South Africa (Diamond *et al*, 2003, Holden *et al*, 2001). As the Bantu migrated so too did their

languages. In some regions the language reflects a mixture of Bantu influences (Vansina, 1995). These migrations were accompanied by spread of agricultural practices. Specifically in South Africa several ethnic groups descended from the Nguni and the Tswana-Sotho families within the Bantu Nation (Herbert, 1990 & Mitchell & Whitelaw, 2005). These ethnic groups are today known as Zulus, Xhosa, Swazi, Ndebele, Basuto, Tswana and Sepedi peoples (Mitchell & Whitelaw, 2005).

1.13.2 Genetic substructure of South African populations

The genetic substructure of seven South African populations, Zulu, Xhosa, Tsonga, Sotho, Pedi, Tswana, and Venda, was assessed by studying the Y-chromosome and the autosomal DNA of these populations (Lane *et al*, 2002 & Mitchell P, 2010). The Y-chromosome is a good indicator of inherited variation within and across populations because the Y-chromosome is inherited only by males from their fathers. In essence the Y-chromosome gives an indication of how males influence the gene pool. The contribution of females and males is assessed by the autosomal DNA.

Black South Africans compromise approximately 77 % of the country's population. In addition, South Africa has 11 official languages. 9 or the 11 official languages are Bantu speaking languages. These linguistic groups all belong to the Eastern Bantu-speaking group. From Y chromosome data it was shown that the languages cluster into 3 specific groups; Tswana/Sotho, Nguni and Venda

language groups. The Nguni group comprises the Zulu, Xhosa, Tsonga/Shagaan linguistic groups. The Sotho/Tswana group comprises the North Sotho, South Sotho and Tswana language groups.

Measurements of F_{st}, was very low, indicating these population groups although linguistically diverse share more than 98% of their genetic variation, suggesting that they all share a common ancestor

These linguistic groups also show genetic differences between these populations. The Nguni linguistic groups split in two with the Zulu and Xhosa forming a cluster while the Tsonga and Shangaan form a separate cluster with the Venda. The Sotho /Tswana group clusters midway between the two. This is indicative that migration events also influence underlying genetic diversity.

1.14 Aim

The aim of the project was to study variation in *APOBEC3G* in Black South Africans in more detail. In 2003 an initial study showed that there is detectable variation in the upstream non-coding region of *APOBEC3G* (Ramdin, 2003). The definitive role of *APOBEC3G* in HIV/AIDS pathogenesis could not be clarified as all sequenced study samples were HIV positive and there was no control group against which to make a comparison. Further genotyping and sequencing were needed to clarify this issue. In addition there was no detection of heterozygotes at various SNPs locations within any of the sequenced samples, as sequencing was not very reliable. Thus an assay was needed to facilitate rapid heterozygote detection and it is vital that this gene be re-examined in detail.

In addition H186R polymorphism was characterised within the study samples. As the origin of most South Africans stems from the Bantu expansion 2000 years ago, the Bantu-speaking sub-populations found today are still genetically similar (Lane *et al*, 2002). The objectives included the characterization of variation of *APOBEC3G* in black South Africans with particular emphasis on the non-coding region. The non-coding region is known to be important in determining the functional variant produced as it controls transcription, translation and processing events of a gene. Particular attention was paid to SNP -571 as this was found within the non-coding region of sequences characterised within Ensembl database as well as in the initial sequencing of the samples collected and characterised in 2003. It was well represented in the 2003 sequences and thus it was decided to

further characterise this SNP in a bigger sample size. In addition, the Bantu-speaking Johannesburg population is representative of the major Bantu-speaking ethnic groups in South Africa.

1.15. Objectives

1. Re-analyse my Honours sequence data obtained from direct sequencing using Sequencher 4.0.

2. Develop genotyping assays for -571 SNP in the upstream non-coding region and H186R in the coding region of *APOBEC 3G*

3. Determine allele and genotype frequencies of the two SNPs.

4. Determine linkage disequilibrium of the SNP data and infer haplotypes.

Chapter 2

Materials and Methods

2.1 Sample Description

Seventy-one samples were collected from HIV positive patients at Johannesburg General Hospital (JHB). These samples were used in the initial investigations to detect variation within the upstream non coding region of APOBEC3G. In the present study 69 of the 71 samples were used for follow-up investigations as material was limited. In addition 45 samples were collected from staff and students at the University of the Witwatersrand from the Bantu speaking population and termed the General Population (GP) (Table 2.1). The HIV status of these individuals was unknown. This set was not used as a control but to ensure that a representative panel of variation within the Bantu population was detected and not only variation pertaining to HIV positive individuals. Short self-reported patient histories of date of first infection, recent CD 4 count, and secondary illnesses were obtained from the JHB sample set.

Adding to these 56 samples were received from the HIVNET 028 Study (HIVNET). The patients were obtained from five clinic sites in four Southern African countries; Malawi, Zimbabwe, Zambia and South Africa. These samples included CD4 counts and viral loads; however information on ancestry of individuals was not available. In addition, a further set of 91 samples was

collected from Helen Joseph Hospital (HJ) (Table 2.1). Short self reported patient histories date of first infection, recent CD 4 count, and secondary illnesses were also obtained from patients. All samples were collected under informed consent (Appendix III).

Table 2.1 Samples collected for the study of variation within APOBEC3G, their HIV status and the genotyping methods used to detect variation.

	Samples			
HIV Status	JHB	HIVNET	HJ	GP
HIV positive	71	56	91	
Unknown				45
Genotyping method	ASA-571 (69), RFLP-571(13), PYRO-H186R (13), RFLP-H186R (32)	ASA-571 (56) RFLP-H186R (38)	RFLP-571 (91) PYRO-H186R (91)	ASA-571 (40), RFLP-571 (30), PYRO-H186R (30), RFLP-H186R (41)

Data was also collected on geographic origin of participants from JHB, GP and HJ study individuals. The participants all filled out a short questionnaire detailing information about themselves, their parents and grandparents. This was used to ascertain the language affiliation of all participants. Language is a better indication of subpopulation affiliation than geographic origin for some parts of South Africa, particularly Limpopo Province. Ethics clearance for this study has been obtained from the Human Research Ethics Committee by Prof Tracy

McLellan. The clearance number is M040221 (Appendix II).

Within the South African samples it was found that in 57 % there was one language present over three generations, 32 % of the recent ancestors were mixed with respect to language. In 8.2 % of the sample population the language across all three generations was incomplete. 2.9 % of the sample was mixed with respect to ethnicity. Within the study populations where there was a single language across all three generations, Zulu was the most frequently sampled at 45 %. At 17 % the Xhosa ethnicity was second in frequency within this grouping, followed closely by Tswana, Sotho, Pedi languages occurring at a frequency of 11 %. The Sotho and Pedi languages were present at frequency of 9.2 % and 7.5 % respectively within the study sample. The Tsonga, Venda, Ndebele and languages represent 4.2 %, 2.5 % and 1.7 % respectively.

2.2 DNA extraction

Genomic DNA extraction was performed using the Qiagen DNA Blood Mini kit as per manufacturer's instructions. This DNA was used to detect variation in *APOBEC3G*. DNA was run on a 1.0 % agarose gel in 1 x TBE at 7V/cm for 75 min. Gels were stained with 10 µg/ml ethidium bromide (EtBr). Gels were visualized under UV light. The DNA concentration in the samples was quantified using the Nanodrop ND-1000. DNAs were aliquotted and stored at -20°C.

2.3 Re-analysis of sequence data

Fifteen samples, sequenced in forward and reverse directions in 2003, were analysed with Sequencher 4.0 (Gene Codes). Sequences were edited and then aligned with each other and a reference sequence from the Human Genome Database (NT011520). The sequence alignments appeared as sequence chromatograms and allowed detection of discrepancies between and within samples representing either SNPs or sequencing errors. Sequences were visually inspected in conjunction with chromatograms and discrepancies were classified as SNP according to the following:

•The peaks were of good height and have little baseline noise.

•Discrepancies were located in good quality sequence.

▪Discrepancies were not located within homopolymeric stretches as these cause many peaks where one should be.

2.4 Detection of Variation in *APOBEC3G*

2.4.1 The genotyping of position -571 using allele specific amplification

Table 2.4.1 Primer set sequences used in ASA, annealing temperatures, and size of product sequence.

Primer	Primer Sequence	SNP	Annealing temperature	Product size
-571 G PCR Reverse	5 'CGCCATGGGAACACGCTACCAG 3' 5' TGAAGCCTCACTTCAGGTACC GCTGC 3'	-571G	63 C	850 bp
-571 C PCR Forward	5'GCGCGTCTCACAGCTCCCTTCCCG 3' 5' AGTTCACAGGGGTCACAATGGCT 3'	-571C	63.5 C	450 bp

Polymerase Chain Reaction (PCR) is a relatively easy method used for amplifying DNA. The method is so sensitive that a single DNA molecule can be amplified and consequently visualized on an agarose gel as bands.

The *APOBEC 3G* was accessed using NCBI and the accession number (NT011520) (http://www.ncbi.nlm.nih.gov/gene/60489). To detect variation of the -571 SNP Allele Specific Amplification (ASA) was used. ASA was used as an alternative to direct sequencing because it is applicable to a large number of samples and is therefore inexpensive. In ASA both primers have the identical sequence except the base at the 3' end is different for each primer (Okayama et al, 1989). Each primer will have one alternative for the SNP at the 3' end. Homozygotes will only yield PCR product with either primer. Heterozygotes will yield a product with both primers (Okayama et al, 1989).

The 50 µl reaction mixture for the PCR consisted of 2X Fermentas Master Mix (0.05 u/µl *Taq* DNA polymerase (recombinant), reaction buffer, 4 mM $MgCl_2$, 0.4 mM of each dNTP (dATP, dCTP, dGTP, and dTTP). Primers (Table 2.4.1) were diluted to give a stock concentration of 2 µg/µl (mass (µg) + 0.5 (mass (µg) TE) and a working solution of 20 ng/µl (2 µl stock soln + 198 µl water) and 20 – 100ng of DNA.

The cyclic conditions for the -571G PCR are denaturation at 94 °C for 2 min, followed by 35 cycles of denaturation at 94 °C for 30s, annealing temperature at 63 C for 25 s, extension at 72 °C for 28 s, the final extension is at 72 °C for 5

minutes. The cyclic conditions for the -571C PCR are denaturation at 94 °C for 2 min, followed by 35 cycles of denaturation at 94 °C for 30s, annealing temperature at 63.5 C for 25 s, extension at 72 °C for 20 s, the final extension is at 72 °C for 5 min. PCR products were electrophoresed on a 1% agarose gel stained with 10µg/ml EtBr in 1X TBE buffer at 7V/cm for 45 min.

2.4.2 Detection of SNP -571 using Restriction Fragment Length Polymorphism (RFLP)

A gradient PCR facilitated the selection of the optimal annealing temperature for the amplification of product. Once the gradient PCR established the optimal annealing temperature, this temperature was used to amplify the subsequent samples using a 50 µl reaction mixture for the PCR consisted of 2X Fermentas Master Mix (0.05 u/µl *Taq* DNA polymerase (recombinant), reaction buffer, 4 mM $MgCl_2$, 0.4 mM of each dNTP (dATP, dCTP, dGTP, and dTTP). Primers were diluted to give a stock concentration of 2 µg/µl (mass (µg) + 0.5 (mass (µg) TE) and a working solution of 20 ng/µl (2 µl stock soln + 198 µl water) and 20 – 100ng of DNA.

The amplicons were then visualized on 1% agarose gel stained with 10µg/ml EtBr in 1X TBE buffer at 7V/cm for 1 hour. The amplified samples were then subjected to restriction digestion for RFLPs. The restriction digestion was optimized by using sequence samples with known nucleotide sequence for GG and GC genotypes. The optimization did not include a control for the CC

genotype as only one was found by direct sequencing and DNA pertaining to this sample was not available.

The RFLP technique was also used to detect variants of SNP -571. The 50µl reaction mixture consisted of 10 X Buffer R with BSA, 2.5 U/ µl *Mva*I (Fermentas), 6 µl amplified PCR product and nuclease free water. The digestion was carried out for 4 hours at 37 °C. The product was run on a 4% agarose gel stained with 10µg/ml EtBr in 1X TBE buffer at 7V/cm for 3 hours at 100V. The gels were visualized with the UV transilluminator. The individuals homozygous for the C allele will have three bands present at 215 bp, 114bp and 49 bp. Those homozygous for the G allele will have 215bp, 163 bp bands present. The heterozygotes will have four bands present (215bp, 163bp, 114bp and 49 bp) (Figure 2.1).

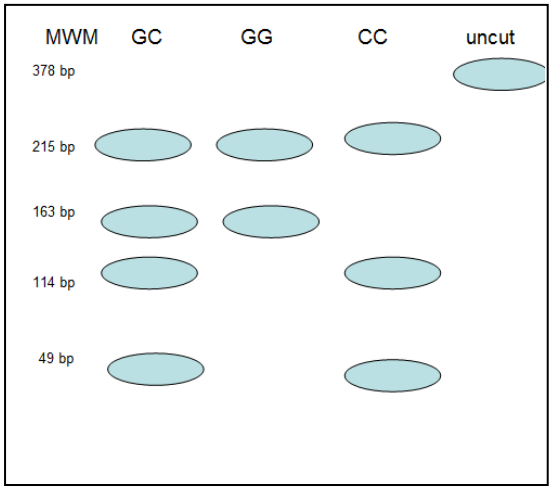

Figure 2.1 Schematic of three genotypes of position -571 of the upstream non-coding region after digestion by restriction enzyme *Mva*I.

2.4.3 Genotyping of position 186 in exon 4 of APOBEC3G

The Ensembl database showed two SNPs in this exon next to each other (Figure 2.2); hence conventional genotyping assays such as allele specific PCR or RLFP proved to be problematic because of their close proximity. The solution to this was to use pyrosequencing to discern the genotypes in the population and to see if the SNP upstream of H186R codon changing variant is also found in the Bantu-speaking South African population (Figure 2.7).

41101 GCAGCCTGTGTCAGAAAAGAGACGGTCCGCGTGCCACCATGAAGATCATGAATTATGACG 41160
41161 GTGAGAAGTGGGAGGTTCAGGGGTGTGGGAGAGACTGCTTAAGTGTMTGTGATGGGTCCT 41220
41221 TCCCACACATACCTGTGGGTCTGCTCTGATGCCTGCAAAGGCCAAGTGTCCCAGGGGAGC 41280
41281 CTGTGGGGTTGGGTCTGGCGCTGASTGTAACTAGTATCYAGAATATGTCTGGGAGGGGAG 41340
41341 GGTCCCGAGGTCACAGAAGAGAGGCCAGCTGGGCTTGACTGCKTTCTCTCTTCTTTTTCT 41400
41401 TAGAATTTCAGCACTGTTGGAGCAAGTTCGTGTACAGCCAAAGAGAGCTATTTGAGCCTT 41460
41461 GGAATAATCTGCCTAAATATTATATATTACTRCRCATCATGCTGGGGGAGATTCTCAGGT 41520
41521 GAGGGTCTCCCTCCAGGCTCATCGCCTCGCTCCTCTCACCTCCTGCTCATCCTCTTGAGG 41580
41581 CCTCCYCTCTGTTCCAGACCAGGTCCTCTCCTGGCCAGGCCCTCCTGCCTTCCCTCCTGC 41640
41641 CCCCTGCCTGCCCTCGTGGTTACACTCCCTCACCCACACTCCTCGTGCTCCCTCCACCTC 41700

Figure 2.2 Excerpt from Ensembl database: ENST0000026324. Exon 4 is highlighted in Blue. The codon changing variant of interest is highlighted in red and yellow and another SNP is found in the adjacent codon.

2.4.3.1 Sequencing of exon 4

The region in exon 4 containing SNP 186 was sequenced by conventional sequencing to confirm the exon 4 nucleotide sequence obtained from the Ensembl database was correct. A 600bp fragment was amplified by PCR with the following primers: Codon 4 fw (AAGCTGCATCGTGACCAGGAGTAT) and Aporev (AGAGGAGCGAGGCGATGA). The 40µl reaction mixture for the PCR consisted of 2X Fermentas Master Mix (0.05 u/µl *Taq* DNA polymerase (recombinant), reaction buffer, 4 mM $MgCl_2$, 0.4 mM of each dNTP (dATP, dCTP, dGTP, and dTTP). Primers were diluted to give a stock concentration of 2

µg/µl (mass (µg) + 0.5 (mass (µg) TE) and a working solution of 20 ng/µl (2 µl stock soln + 198 µl water) and 20-100ng of DNA.

The cycling conditions for the PCR were denaturation at 94 °C for 2 min, followed by 40 cycles of denaturation at 94 °C for 30s, annealing temperature of 59.6 °C for 30 s, extension at 72 °C for 30 s, the final extension was at 72 °C for 5 min. PCR products were electrophoresed on 1% agarose gel stained with 10µg/ml EtBr in 1X TBE buffer at 7V/cm for 1 hour. Two samples were sequenced in both the forward and reverse directions using sequencing primer ApoSeq1.

2.4.3.2 Detection of variation in exon 4 using pyrosequencing

Pyrosequencing technology is based on a 4-enzyme real-time monitoring of DNA by bioluminescence. Essentially there are four reactions that ultimately result in the quantitation of a light signal and a sequence of synthesized strand of DNA (Ahmadian et al, 2005). For pyrosequencing the sequence surrounding SNPs is known and only the dNTP that matches the sequence is added to the reaction and it pairs with the sequencing template and an inorganic pyrophosphate is released. The tagged amplified fragments serve as the template for the universal primer. The released phosphate serves as a template for ATP sulfurylase to produce ATP (Figure 2.3). ATP is converted to light by luciferase (Ronaghi, 2008). The unincorporated nucleotides and ATP are removed from the reaction by Apyrase (Ronaghi *et al*, 1998). This is crucial in ensuring that the light signal is only the result of the correct base being added. The sequence is consolidated as a pyrogram

where the peaks give the approximate light signal intensity. The light intensity is directly proportional to the sequence of the synthesized DNA (Ronaghi, 2003). The software produces a theoretical output of all possible genotypes and the pyrograms need to be compared to these to verify the genotypes.

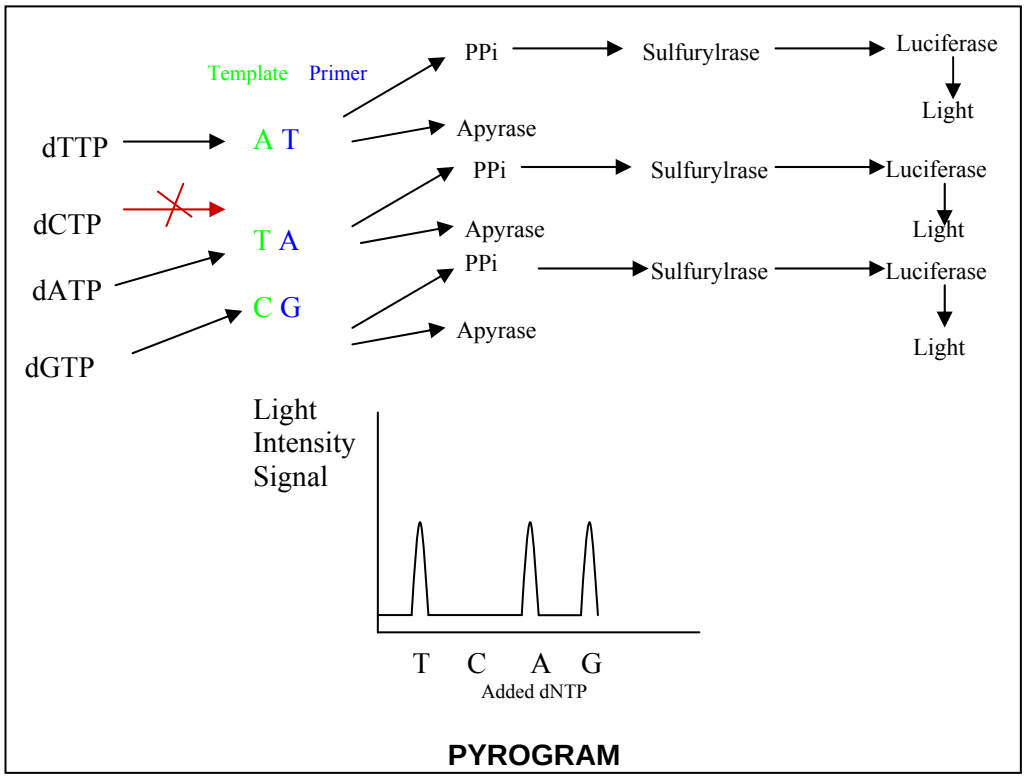

Figure 2.3 Schematic overview of pyrosequencing system. When the correct dNTP is added it pairs with the template and light is produced. When the dNTP that is added is not complementary with the template, no light is produced.

The simplicity of the technique allows for high throughput DNA analysis. There are many applications of pyrosequencing. This technique was used for genotyping SNPs within the coding region of APOBEC3G (Figure 2.7). The primer design is very important in the SNP genotyping assay. There are sequence specific primers that used to amplify the region of interest. The primers are APOfor

(5'GACGGGGACACCGCTGATCGTTTAGCAAGTTCGTGTACAGCCAAAG A 3') and APOrev (AGAGGAGCGAGGCGATGA) and were designed with the pyrosequencing software by Dr Zane Lombard from the NHLS. The forward primer is tagged at the 5' end with a sequence:

(GACGGGGACACCGCTGATCGTTTA) that matches the biotin labelled universal primer (GACGGGGACACCGCTGATCGTTTA). Essentially the process is that the forward and reverse primers will amplify the region of 159 bp (Figure 2.4).

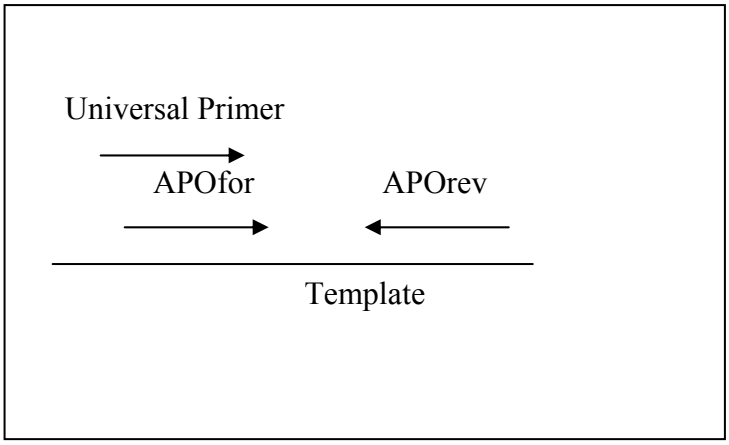

Figure 2.4 Position of the primers used for pyrosequencing.

The 50 µl reaction mixture for the PCR consisted of 2X Fermentas Master Mix (0.05 u/µl *Taq* DNA polymerase (recombinant), reaction buffer, 4 mM $MgCl_2$, 0.4 mM of each dNTP (dATP, dCTP, dGTP, and dTTP). Primers were diluted to give a stock concentration of 2 µg/µl (mass (µg) + 0.5 (mass (µg) TE) and a working solution of 20 ng/µl (2 µl stock soln + 198 µl water) and 20 – 100ng of DNA.

The cycling conditions for the PCR was denaturation at 94 °C for two min, followed by 40 cycles of denaturation at 94 °C for 30s, annealing temperature of 58C for 30 s, extension at 72 °C for 30 s, the final extension was at 72 °C for five min. After the successful optimization and amplification of the desired region the samples were viewed on a four % agarose gel stained with 10µg/ml EtBr in 1X TBE buffer at 7V/cm for three hours at 100V to ensure the correct size and minimal primer dimers.

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 three minutes.

Because the forward primer was tagged the software assigns the genotype of the anti-sense strand.

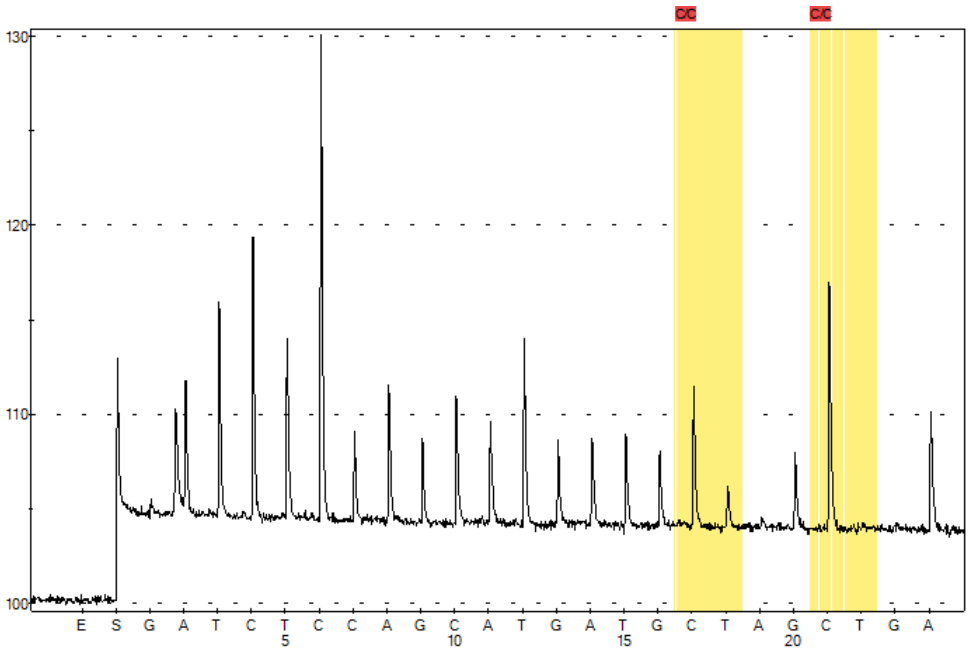

Figure 2.5 Pyrogram of variation at codon positions 185 and 186 in exon 4 of Apobec 3G

2.5 Data Analysis

2.5.1 Allele and Genotype frequencies

To determine the allele and genotype frequencies of the various SNPs the genotypes were counted. The genotype of an individual is defined as its genetic makeup with reference to a specific locus. Consequently the genotype frequency is the frequency or proportion individuals carrying a certain genotype (Hartl and Clark, 1989). The allele frequency on the other hand is the measure of the frequency in a population of a specific allele at a given locus for that trait.

If a locus is bi-allelic (i.e. it has two forms at the locus) then the frequency of the three possible genotypes can be represented by f (AA), f (AA), f (AA). If the

numbers of individuals (obtained by direct counting of the three genotypes) carrying those genotypes is represented by x, y, z respectively.

Genotype frequency of the three genotypes is calculated as follows:

F (AA) = x/n where n is the number of individuals present

f (Aa) = y/n

f (aa) = z/n

The allele frequency is determined from the genotype frequency as follows:

Let f (A), f (a) represent the alleles frequencies of the A allele and the a allele respectively. Then,

f (A) = (2x+ y)/2n where 2x + y is the number of A alleles

f (a) = (2z + y)/2n where 2z + y is the number of a alleles

2.5.2 Hardy-Weinberg Equilibrium

The Hardy-Weinberg law predicts genotype frequencies from allele frequencies under certain conditions. For a population to be in Hardy-Weinberg equilibrium certain conditions have to hold true; there must be random mating, no gene flow in or out of the population, the population must be infinitely large and there must be equal fertility of all genotypes (Hartl and Clark, 1989). Thus a consequence of this model is that allele frequencies will not change from one generation to the next. Genotype frequencies can be predicted from allele frequencies.

The model

If f (A) = p,

and f (a) = q,

then the expected genotypes will be

$f(AA) = p^2$

f (Aa) = 2 pq

$f(aa) = q^2$

If the population is in equilibrium then,

$p^2 + 2pq + q^2 = 1$

The statistical Chi-squared test ($\chi 2$) is then used to determine if the frequency of the expected genotypes is much different from the observed genotypes.

$X^2 = \sum (O-E)^2 / E$

Where O = observed number of genotypes

E = expected number of genotypes

A P value of 0.05 indicates a lack of significant deviation from the Hardy-Weinberg Model. In this study the estimation of gene frequencies was further used to look for differences

2.5.3 Linkage disequilibrium and haplotype analysis

Linkage disequilibrium describes the non-random assortment of alleles at two or more loci. It essentially states that some genotypes may occur more or less frequently than would be expected if the loci were not linked. Often genes and

SNPs are inherited as part of a unit with genes that lie in a close physical proximity and this is termed a haplotype. Thus measurements of LD are based on comparisons of genotype frequencies of haplotypes.

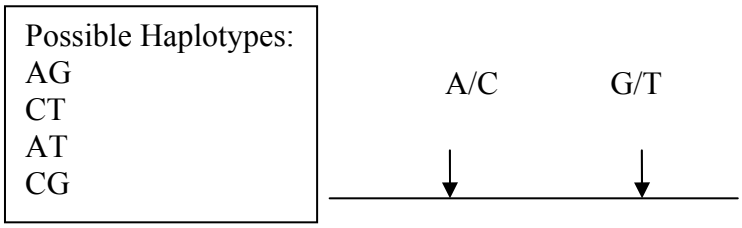

Figure 2.6 Diagrammatic representation of two loci on a gene. There are two alternative bases for each loci giving rise to four possible haplotypes during independent assortment of genes.

Linkage disequilibrium was analysed using Linkage Disequilibrium Analyzer 1.0. This program implements the EM algorithm (Keyue et al, 2001). To explain LD, consider an example.

Consider two loci each with two alternative alleles on one chromosome

Haplotype Frequency:

$A_1 B_1 x_{11}$

$A_1B_2 x_{12}$

$A_2B_1 x_{21}$

$A_2B_2 x_{22}$

Then the allele frequency will be:

Allele Frequency:

$A_1 p_1 = x_{11} + x_{12}$

$A_2 p_2 = x_{21} + x_{22}$

$B_1\ q_1 = x_{11} + x_{21}$

$B_2\ q_2 = x_{12} + x_{22}$

If these alleles are independent then

$x_{11} = p_1 q_1$

But if the alleles are not independent and there is a deviation from the observed frequencies compared to the expected then it is measured by a parameter D (Devlin and Risch, 1995).

$D = x_{11} - p_1 q_1$

Allele frequencies can only be between 0 and 1. When either applies then there can no D. Therefore the D is normalized by dividing it with the theoretical maximum (0.5) of observed allele frequencies (Lewotin, 1964).

$D' = D / Dmax$

Where $D >= 0$ or $D < 0$

Another important LD measure is r^2. This measure informs if alleles at two loci are related and is often used to detect loci that influence disease susceptibility (VanLiere and Rosenberg, 2008).

Haplotypes analysis was determined using PHASE v 2.1 (Stephens et al, 2001). This software implements Bayesian algorithms. One hundred and thirty six samples with known genotypes at loci -571 and H186R were used for the haplotype analysis. To obtain reliable results the developers suggested that at least 10 runs were done and inter run variability checked to ensure correct analysis. Specifically it was suggested that the Freq output file is checked between runs as this will provide the most reliable look at run performance.

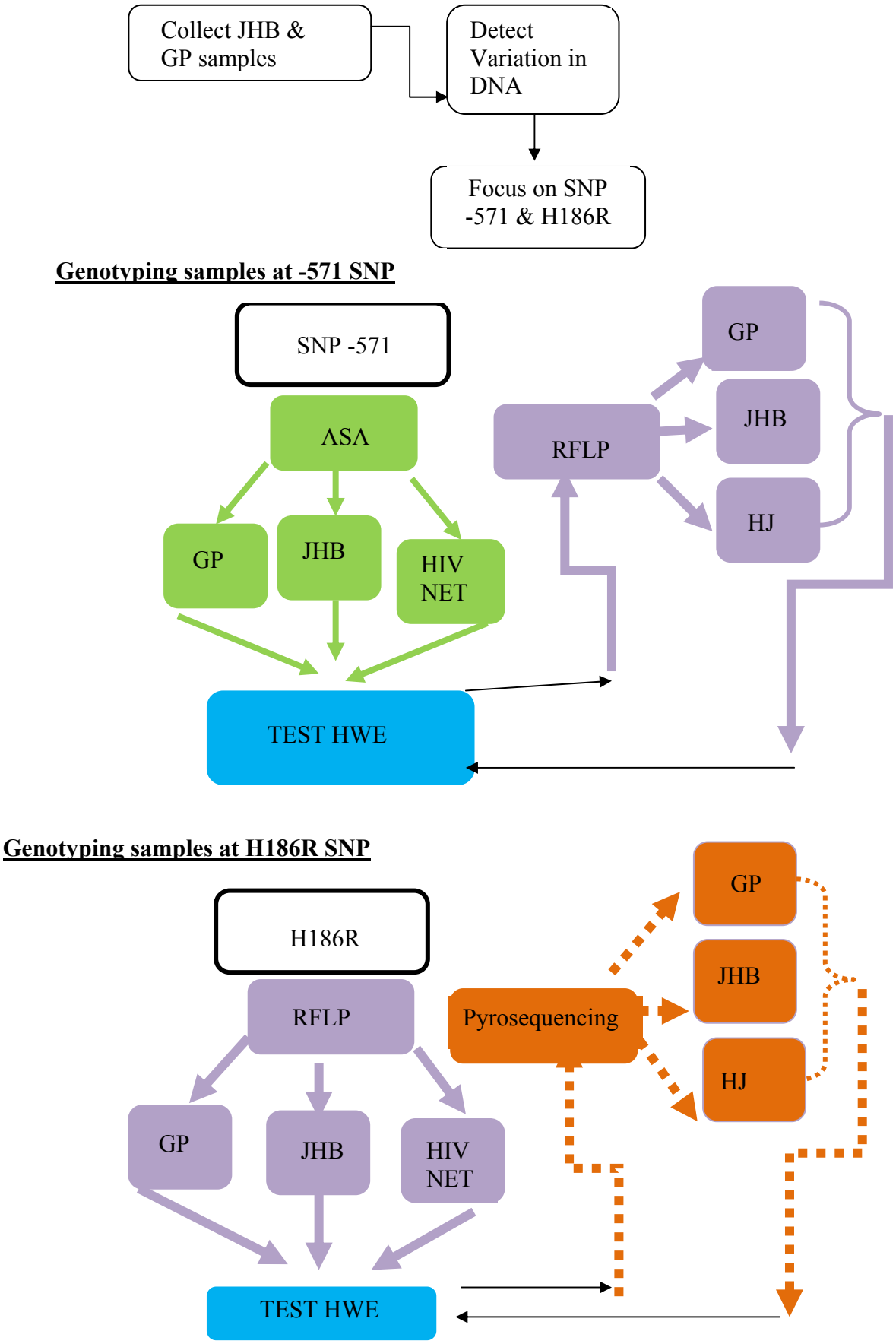

Figure 2.7 Schematic overview of all methods used on each sample type.

Chapter 3

Results

3.1 Analysis of upstream non-coding region sequences

3.1.1 Reanalysis of previously sequenced samples

Fifteen samples were sequenced by Inqaba Biotech in 2003. The samples were sequenced in the forward and reverse directions and the average length of the sequences was 900bp. In 2003 the forward and reverse sequences of each sample were automatically aligned with the putative upstream non-coding region of APOBEC 3G (NCBI reference sequence NT_011520.11) to form a contig. Each contig was then edited according to the chromatograms. Editing involved examining the forward and reverse chromatograms while simultaneously looking at the sequences in relation to the reference sequence. Thereafter the reference sequence was removed from the contig and a new consensus sequence was created from the forward and reverse sequences. The consensus sequence of each sample was then aligned automatically to form a new contigs. The relative position of the mutations was then noted.

In 2003 sequencing showed that there were numerous point mutations present (Figure 3.1). Different combinations of transitions and transversions were present in each sample. Transversions were the most common point mutations. Six of the eight possible transversions were present in the sequences, T-A, C-A, C-G, T-G, G-C, and A-C. Only three transitions were detected, C-T, A-G, G-A, with C-T

being the most frequent transition. Numerous insertions and deletions were also found. SNPs at -972, -963, -960, -881, were not observed in these samples but other new SNP were characterized. During consensus alignment of the samples, two sites were of particular interest. At position -590 all sequences were different from the reference sequence, there being an A-G transition. Furthermore, at position -571 11 of the 15 sequences were different from the reference sequence there being a G-C transversion (Figure 3.1.1.3). Another SNP (G-C) was found at -571 in most of the samples. One of the previously characterized SNP at -90 was also present in some samples. -286 A-G transition was represented in 5 of 10 sequences.

In 2003 -90 SNP was found at a position of 91 bases upstream of transcription initiation. Reanalysis demonstrated after additional editing only 6 from initial 15 sequences provided informative data. Four heterozygotes (GC) and 2 CC homozygotes at position -90 were detected. The allele frequencies could not be estimated accurately for this SNP because of the very small sample size. However comparative data from Ensembl shows that the allele frequency is relatively the same in the Yoruba from Nigeria (G and C allele frequency is 0.562 and 0.438 respectively) while the frequency of the ancestral allele (C) in Europeans is much less than the frequency of the G allele (Table 3.1)

SNPs at position -163, -166 and -199 were not characterized in 2003 but re-analysis showed them to be well represented in 15 reanalysed samples. Five heterozygotes were observed at each position -163, -166 and -199. Heterozygote genotypes observed for these positions were found in the same samples.

Homozygotes for the ancestral alleles according to the dbSNP database were observed at these positions in the remainder of the samples. No homozygote genotypes were observed for the minor alleles at each SNP position. The remaining 10 samples were homozygous for the major allele at all three loci. There is no frequency data available for SNP -163 and -166 for Africans or Europeans. However the data available for SNP -199 shows that the frequency of the ancestral allele (A) in Africans is similar to the estimated values within the study (Table 3.1).

Table 3.1 Sequenced data from 2003 and comparative data of other populations (http://www.ensembl.org/Homo_sapiens/Variation/Population)

	NCBI rrs ID	rs5750743						rs34550797		rs5757463		rs17496004		rs8142124	
	SNP	-90		-163		-166		-199		-571		-590		-881	
	Alleles	C	G	C	A	T	A	A	G	C	G	A	G	C	T
Allele Frequency	Study	0.667	0.333	0.167	0.833	0.167	0.833	0.167	0.833	0.733	0.267	0	1	0.611	0.389
	African (Yoruba in Ibadan, Nigeria,)	0.438	0.562					0.062	0.938	0.894	0.106	0.994	0.006	0713	0.287
	European (Utah Residents (CEPH) with Northern and Western European ancestry)	0.318	0.682					0	1	0.900	0.100	0.934	0.066	0.999	0.001
	American (Mexican Ancestry from Los Angeles USA)									0.97	0.03	0.970	0.030	0.983	0.017
	Asian (Han Chinese in Bejing, China,)									0.913	0.087	1	0	0.997	0.003

In subsequent reanalysis, I found that the SNP at -286 was the result of a sequencing artifact as may happen when G is preceded by T in direct sequencing.

The SNP at -571 observed during analysis in 2003 was still polymorphic after re-analysis. Of 15 samples six were heterozygous, eight homozygous for the C allele and one was homozygous for the G allele. The frequency of the C allele is 0.733 and the frequency of the G allele is 0.267 in this group. The population diversity on the dbSNP shows that in African populations the allele frequency of the C and G alleles to be 0.894 and 0.106 respectively (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=5757463.) This difference will be discussed later.

Polymorphism at position -881 were not characterized in the 2003 analysis but were seen upon re-analysis. Seven heterozygotes and two homozygotes for the allele C were observed. However, frequency data could not be ascertained from the sequences as the sample size is too small. Frequency data from Ensembl was also not available for this SNP.

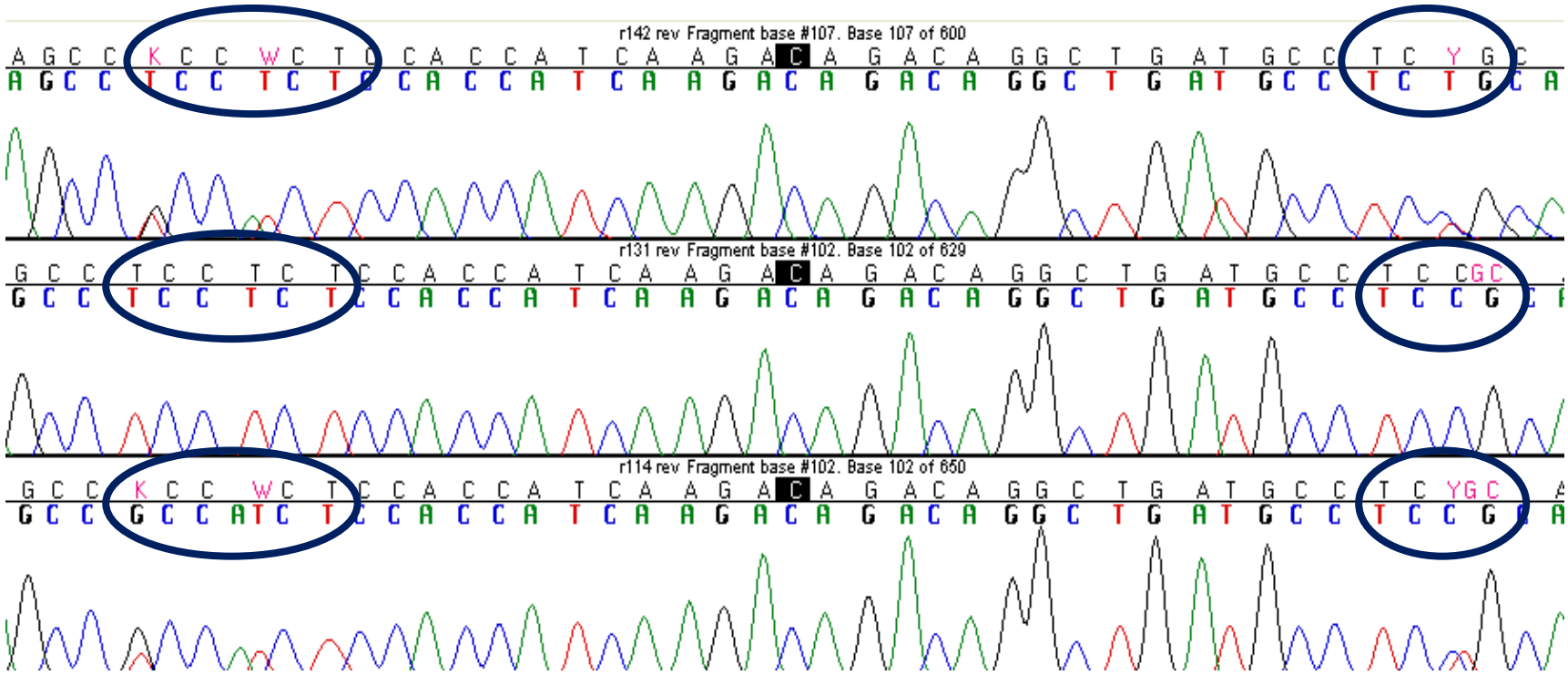

Figure 3.1. The chromatograms show samples 114, 142 and 131 at SNP positions -163, -166 and -199. These SNPs appear to be in linkage disequilibrium as they are all either heterozygous at all positions in sample 114, 142 or homozygous at all position as in sample 131.

Position -571 -590

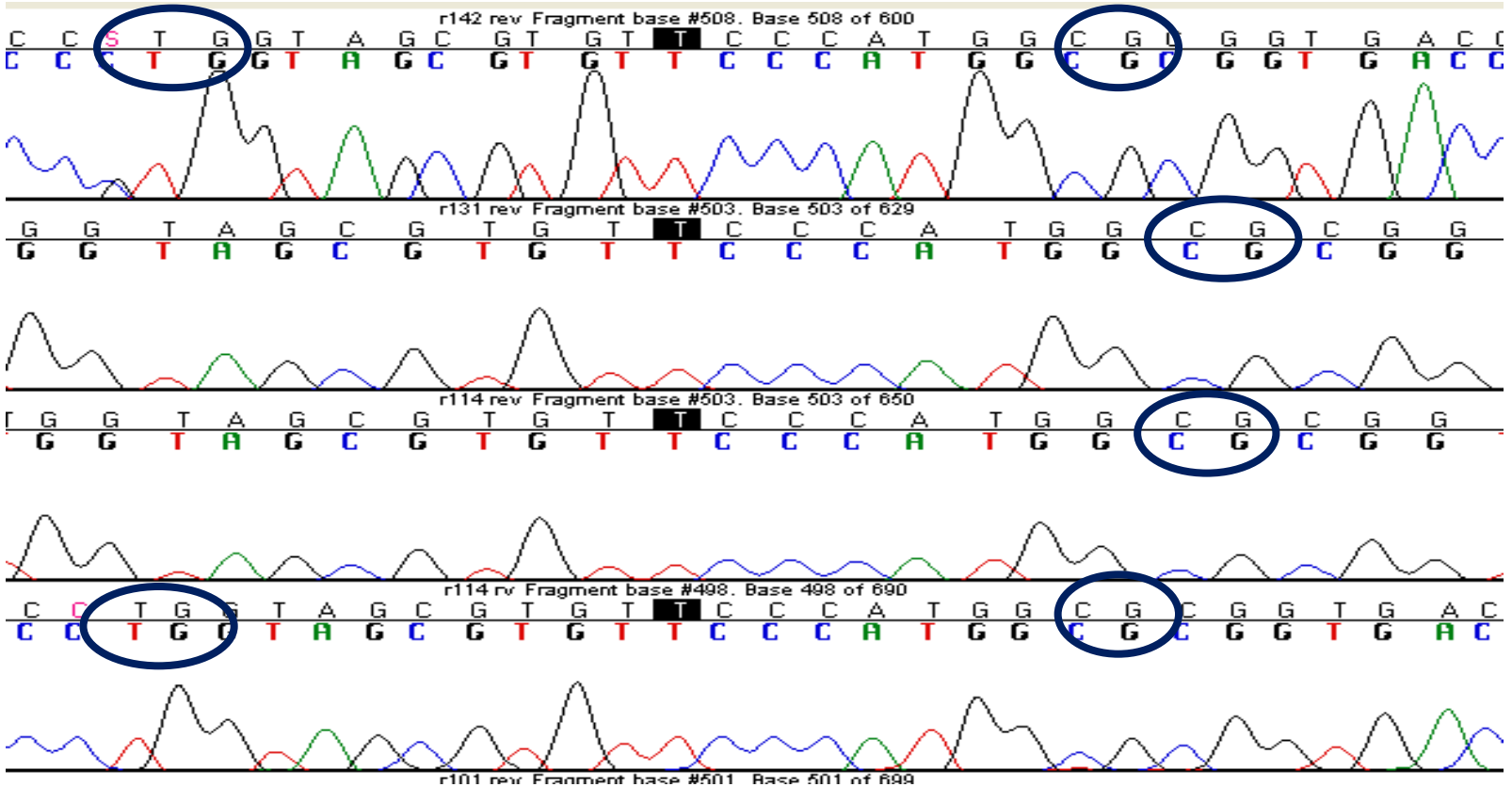

Figure 3.2 Chromatograms showing -571 and -590 loci (circles). The -571 SNP is well represented in the samples. The chromatograms show heterozygotes and homozygotes for the alleles. All samples at -590 deviated from the major allele which is a G. Thus this is an example of a fixed polymorphism.

3.2 Detection of Variation in *APOBEC 3G* using Genotyping Assays

Genotyping was not possible in all of the samples of each sample set as the DNA was not available for all samples at the time of testing. Thus only subset of each group was genotyped as indicated in Appendix I.

3.2.1 The genotyping of position -571 using Allele Specific Amplification

Allele-specific PCR was used to genotype the insertion at position -571 in 165 samples (69 JHB, 56 HIVNET and 40 GP) (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 control with no sample DNA was included to preclude any false positives as a result of DNA contamination. The PCR allowed the SNP to be detected in a large sample. The PCR products for each allele of each SNP were run together on the same agarose gel to facilitate their correct genotyping (Figure 3.3).

Individuals homozygous for a SNP yield a product in one reaction and not the other. While heterozygotes yield a product of the same intensity in both reactions. The sizes of the products for the different alleles differ allowing accurate genotyping. GG homozygotes were more frequent than CC homozygotes. The sample population deviated from Hardy-Weinberg equilibrium at a $p < 0.05$.

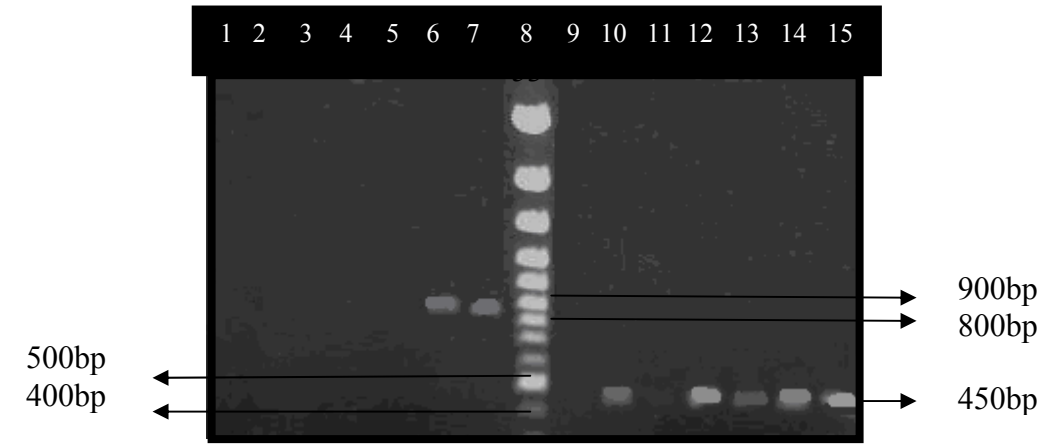

Figure 3.3 ASA products from -571 SNP were run on a 1 % agarose gel. Two separate reactions are performed, each resulting in the amplification of one of the two allelic variants. Genotypes were resolved based on the presence or absence of a PCR product in each or both of the reactions. The products of both reactions were visualized on a 1% agarose gel. The -571 G reactions produces a band of size 850bp and the -571 C reaction produces a band of size 450 bp. Lanes 1-7 are the -571 G reaction of samples, lane 8 is a GeneRuler™ 100bp Plus DNA Ladder (Fermentas) and lanes 9-15 are the -571 C reactions of the same samples. Sample 1 and 3 were not amplified in either reaction. Samples 2, 4, and 5 are genotyped CC while samples 6 and 7 are genotyped GC.

3.2.2 Detection of SNP -571 using Restriction Fragment Length Polymorphism (RFLP)

A RFLP-PCR assay was designed to genotype the -571 SNP at position within APOBEC 3G in 132 samples (13 JHB, 88 HIVNET and 32 GP). The PCR-RFLP allowed the SNP to be detected in a large sample size with relative ease. The

gradient PCR allowed for the detection of the optimal cyclic conditions to amplify the ~400bp region. An annealing temperature of 58°C was chosen as the most favourable temperature because two of the three samples amplified to a better degree than at annealing temperatures of 59.3°C and 60°C.

The amplified samples were digested with *Mva*I and fragments run on a 4 % agarose gel. Genotypes were assigned based on the restriction profile obtained (Figure 3.4).

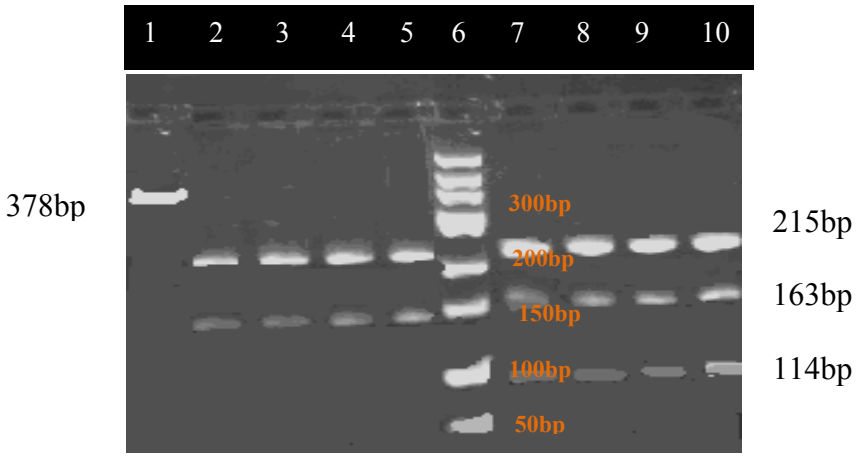

Figure 3.4 The restriction fragments generated when a 378bp fragment of the APOBEC 3G is digested with *Mva*I. The genotype at this position can then be determined by the different restriction profiles produced by digestion. The restriction digest was resolved on a 4% agarose gel. Lane 1 shows the undigested "no enzyme" control. Lanes 2-5 represent homozygotes for the G allele with fragment sizes at 215bp and 163bp. Lane 6 is a GeneRuler^{TM} 50bp DNA ladder (Fermentas). Lane 7-10 represents heterozygotes with fragment sizes at 215bp, 163bp, 114bp and 49bp. The 49bp fragments are not clearly visible on the gel.

The assay identified 88 GG homozygotes, 35 heterozygotes and 1 CC homozygote. The remaining 12 genotypes were discerned from re-analysis of sequence data. Therefore there were in total 95 GG homozygotes, 39 heterozygotes and 2 CC homozygotes. The minor allele frequency when employing the RFLP-PCR assay was low (0.16) in comparison to the frequency found when using the ASA genotyping method. The sample population did not deviate from Hardy-Weinberg equilibrium at $p<0.5$ at RFLP however there is deviation from the equilibrium with ASA.

3.2.3 Detection of H186R using pyrosequencing

Variation in codon 186 within exon 4 was detected in 132 samples (13 JHB, 88 HJ, and 32 GP). These were sequenced using pyrosequencing (Figure 3.5).

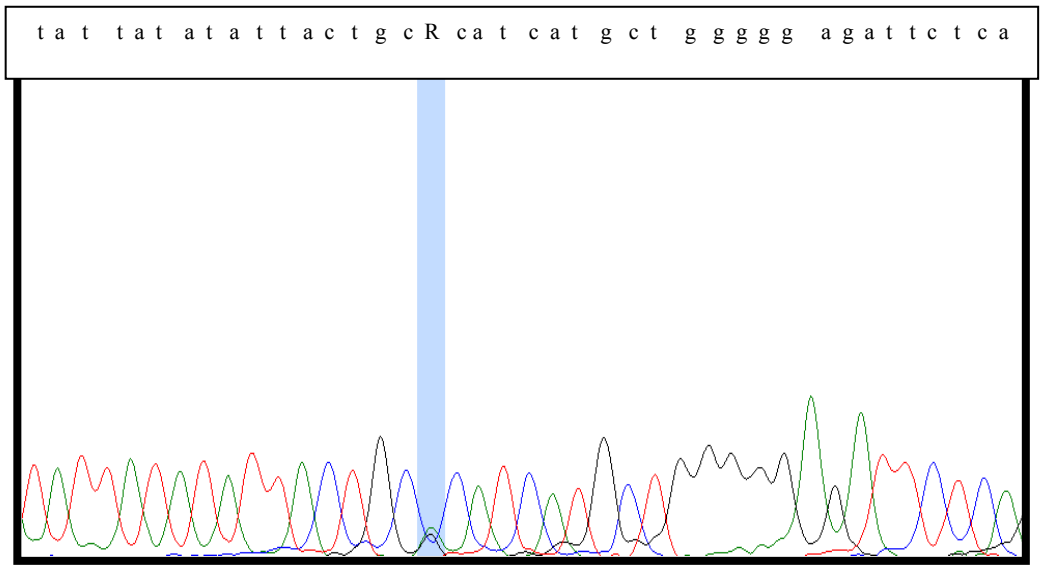

Figure 3.5 Chromatogram of sample 310 sequence. The highlighted strip shows a heterozygote for this sample at position 186. Codon 185 represented by nucleotides 530, 531 and 532 in the chromatogram is not heterozygous in this sample.

The sequencing of a 600bp region of exon 4 confirmed that the sequence from Ensembl database was indeed correct (Figure 3.6).

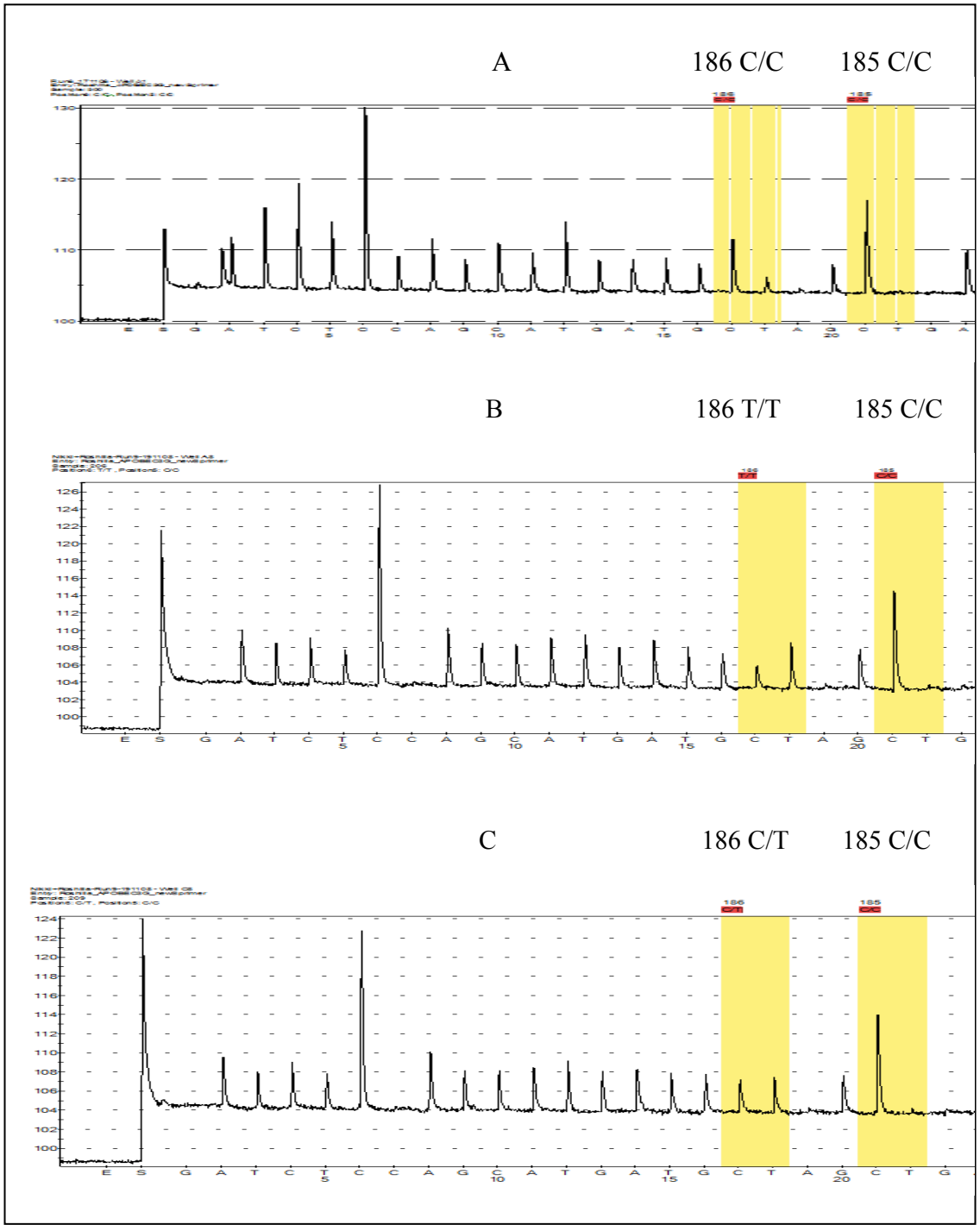

Figure 3.6 Pyrogram output files for codons 186 and 185. Pyrogram A indicates the CC genotype at position 186 and 185.Pyrogram B shows TT genotype was present at 186 in this sample and position 185 was homozygous for C allele. Pyrogram C shows this sample to be heterozygous at position 186 and homozygous for the C allele at position 185.

3.3 Estimation of Gene Frequencies

3.3.1 Differences at -571 and H186R using various genotyping methods

Table 3.3.1 A summary of the genotyping data collected at all 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.

Genotyping Method	SNP	GENOTYPE	N	ALLELE	MAF	GF	χ^2 Value	P-Value
ASA (69 JHB, 56 HIVNET and 40 GP sample sets)	-571	CC CG GG	46 44 75 165	C G	0.41	0.279 0.267 0.455	33.708	0.0001
RFLP (13 JHB, 88 HJ and 32 GP sample sets)	-571	CC CG GG	2 38 92 132	C G	0.16	0.015 0.288 0.697	0.762	0.6833
RFLP (32 JHB,38 HIVNET and 41 GP sample sets)	H186R	CC CT TT	3 66 42 111	C T	0.32	0.027 0.594 0.378	14.326	0.0008
Pyro-sequencing (13 JHB, 88 HJ and 32 GP sample sets)	H186R	CC CT TT	24 84 24 132	C T	0.50	0.182 0.636 0.182	9.818	0.0074

The sample population deviated significantly from Hardy-Weinberg equilibrium at – 571 using ASA at a $p < 0.05$. The GG homozygotes were more frequent than the CC homozygote (Table 3.3.1). The sample population did not deviate from Hardy-Weinberg equilibrium at -571 using RFLP at a $p > 0.05$ even though the GG homozygotes were more frequent than the CC homozygote. The H186R SNP

does deviate from the Hardy-Weinberg Equilibrium using RFLP at $p < 0.05$. There is a large heterozygote excess and almost equal proportion of either homozygote. Conversely there is no deviation at this SNP when pyrosequencing is used for genotyping.

3.3.2 Estimation of Pair-wise Allelic Linkage Disequilibrium

Table 3.3.2 Linkage disequilibrium in APOBEC3G gene

Site	D	D'	r^2
-571/ H186R (Present Data)	-0.127	0.216	0.0087
-571/H186R EA (An et al, 2004)		1.000	
-571/H186R -AA (An et al, 2004)		0.967	

Pair-wise linkage disequilibrium analysis across the two SNPs which are 5030 base pairs apart on chromosome 22 showed that they are not in linkage disequilibrium because the D' value for this association was less than 0.46 (Kidd et al, 1988). There is no correlation between the alleles at the two loci because the r^2 is very low (Table 3.3.2).

Ensembl does not give LD values because allele frequencies are low in European populations. However the D' value for African Americans was 0.967 and 1.000 in European Americans (An et al, 2004).

3.3.3 Differences between the Bantu language groups

A comparison of the allele frequencies (Table 3.3.3) was made between the different ethnic groups represented in this study. The classification was based on the home language spoken by the individual in question and their immediate family. The analysis was based on the data from RFLP and pyrosequencing at -571 and H186R as these were the most reliable genotyping methods tested for each SNP. There were a total of 132 samples (Appendix I, Table A 4, A 5 and A 6) which represent 135 individuals within each genotyping method. However the final comparison was conducted using 46 individuals who reported a single language spoken by their relatives for three generations, in both their maternal and paternal lineages and 40 individuals with uncertain lineages. Zulu speakers comprised 40 % of the sample, while five of the ethnic groups (Venda, Tsonga, Swazi, Pedi and Ndebele) were represented by fewer than ten individuals and as recommended by Lane et al (2008) these groups were pooled for subsequent analysis.

The minor allele frequency of the Zulu group at position -571 was lower than that of the Xhosa. At position H186R the Xhosa and mixed lineage group showed similar frequencies, while the Zulu group had a frequency lower than those of the other two groups (Table 3.3.3). The group comprising individuals of mixed or unknown lineage had a similar frequency distribution to the Zulu group at -571. The Sotho/Tswana group had similar frequencies to that of the Xhosa language group.

Table 3.3.3 The minor allele frequencies at each of the four polymorphic sites. Only groups comprising >10 individuals are included.

Language Group	n	Minor Allele Frequency	
		-571	H186R
Zulu	34	0.21	0.37
Xhosa	12	0.04	0.46
Sotho/ Tswana	17	0.06	0.47
Other[1]	40	0.20	0.45

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

In Table 3.3.3 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. These groups are classified as macrogroups according to Lane et al (2002).

A comparison of the allele frequencies between the macrogroups (Table 3.3.3) 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.

Table 3.3.3.1 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 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	***Genotype***	***Group 1***					***Group 2***					***Group 3***					***Group 4***				
		n	**GF**	**MAF**	**χ^2 Value**	**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**
-571	CC	2	0.27				2	0.06				2	0.04				0	0			
	CG	19	0.25	0.15	0.01[1]	0.97	10	0.29	0.21	0.06[5]	0.81	10	0.24	0.16	0.02	0.89	8	0.28	0.14	0.01	0.92
	GG	54	0.72				22	0.65				33	0.72				21	0.72			
+H186R	CC	10	0.13				3	0.09				5	0.11				5	0.17			
	CT	44	0.59	0.43	0.54	0.46	19	0.56	0.37	0.19	0.66	26	0.57	0.39	0.25	0.62	18	0.62	0.48	0.39	0.53
	TT	21	0.28				12	0.35				15	0.33				6	0.21			

3.4 Haplotype Analysis

PHASE 2.1 (Stephens *et al.*, 2001) was used to construct the haplotype structure surrounding the two polymorphic sites genotyped and to determine the frequencies of these haplotypes in the sample sets JHB, GP and HJ examined by RFLP and Pyrosequencing (Table 3.4.1) as well in the four ethnic groups under investigation (Table 3.4.2). A total of four possible haplotypes was 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 three possible haplotypes were identified in the 22 individuals comprising group 4.

Analysis of haplotype frequencies (Table 3.4.1) in the pooled data from samples sets JHB, GP and HJ there was a similar frequency between the GG and GA haplotypes of 0.419 and 0.413 respectively. The CG and CA haplotypes also exhibited a similar frequency.

Analysis of the haplotype frequencies (Table 3.4.2) in each of the groups revealed the frequency distributions were similar between groups 1, 2, 3 and 4 of haplotype GA and CG. Groups 1, 2 and 3 had similar frequencies of the CA haplotype in comparison to group 4. Additionally, while GA was still the most common haplotype in this population group, the CG haplotype was the most infrequent haplotype amongst all groups within this population.

Table 3.4.1 The estimated haplotype frequencies in JHB, GP and HJ sample sets generated by genotyping data from RFLP and Pyrosequencing genotyping assays in this study, as calculated using PHASE 2.1. (Stephens *et al.*, 2001).

Haplotype	**Haplotype Frequency**
GG	0.42
GA	0.41
CG	0.08
CA	0.09

Table 3.4.2 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	**Haplotype Frequency**			
	Group 1	**Group 2**	**Group 3**	**Group 4**
GG	0.36	0.29	0.33	0.42
GA	0.49	0.50	0.50	0.51
CG	0.07	0.08	0.06	0.06
CA	0.09	0.13	0.10	<0.01

Chapter 4

4.0. Discussion

4.1. Direct Sequencing

The initial goal of the Honours project (Ramdin, 2003) was to find variation within the Apobec 3G locus. Once variation was characterised by sequencing analysis it was compared to variation within the Ensembl database. Thereafter the focus shifted to develop genotyping assays and estimate allele frequencies in the SA population as a basis for further work. These were then used to estimate haplotypes within all sample sets used. Thereafter as an addition the frequencies were estimated based on differences between ethnic groups based on spoken language.

In this work the re-analysis of the direct sequencing of the upstream non-coding region showed six polymorphisms (--90, 163, -166, -199,-571,-881) in the 15 samples each sequenced in the forward and reverse directions. Three of these sites had previously been identified, while the remaining three were not identified by other studies (http://www.ensembl.org/Homo_sapiens/Variation/Population). These novel polymorphisms may thus be unique to African populations. This supports the observation that Africans have the largest number of population-specific alleles and that the variation present in non-African populations is largely 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).

Six SNPs were identified within the 1000bp region upstream of the non-coding region via direct sequencing. SNP -90 was characterised only in six samples. SNPs -163, -166 and -199 appeared to be in linkage disequilibrium and well represented in most samples. All three sites were either heterozygous or homozygous for the major alleles at each position. Both alleles at SNP -571 were well represented in the sequenced samples. SNP -881 was only well represented in two samples as the sequences were edited on reanalysis and this region at the end of the sequence was often eliminated.

The sequenced polymorphisms were compared to data from Ensembl. The allele frequency of -90 SNP in this study is 0.667 of the ancestral allele and 0.333 of the G allele. In comparison the frequencies of the African (Yoruba in Ibadan, Nigeria) population from Ensembl showed the ancestral allele frequency to be 0.438. The G allele is 0.562 (http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=22:39472546-39473546;v=rs5750743;vdb=variation;vf=32060683). The discrepancy between these two groups is attributed to the detection of specific genotypes. With the direct sequencing from 6 sequences 2 CC and 4 CG and no GG genotypes were detected. Examining the African population data genotypes CG and GG are present in the population with equal frequency of 0.375 the CC genotype occurs at a slightly lower frequency of 0.250. In the European (Utah residents with Northern and Western European ancestry) population the CG and GG genotypes also occur at the same genotype frequency however the numbers of these genotypes detected is higher than in the African population

(http://www.ensembl.org/Homo_sapiens/Variation/Individual?db=core;r=22:39472546-39473546;v=rs5750743;vdb=variation;vf=32060683). Therefore the CC genotype is detected at a lower rate in this group. The frequency of minor allele in African Americans was 0.319, 0.340 in European Americans in work of An et al (2004). Furthermore a very similar minor allele frequency of 0.327 was observed in Africans in work of Reddy et al, (2010). These frequencies are very similar to the study frequency of 0.333.

Only SNP -199 has been characterised before (rs 34550797). The sequencing data indicate very similar frequency of genotypes as in dbSNP. The GG occurred 42 out of 48 samples in Africans (Yoruba in Ibadan, Nigeria) (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=34550797). In addition the GG genotype also occurred in 22 out of 24 samples (http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=22:39472437-39473437;v=rs34550797;vdb=variation;vf=32183674). The allele frequency of the G allele is 0.938 and 1.000 in Yoruba and Utah residents respectively. In our study the allele frequency is comparable at 0.833 with the Yoruba population. The allele frequency of the minor allele is lower in the Yoruba population (0.062) than in the study (0.167) (http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=22:39472437-39473437;v=rs34550797;vdb=variation;vf=32183674). Sequencing data shows SNP -199 is tightly linked to SNPs -163, -166. They always occur in the same samples and if one SNP is heterozygous or homozygous for ancestral allele

the other two follow the same pattern. An et al (2004) makes reference to this SNP and terms it rare. It was not rare in the sequenced samples though.

SNP -571 sequencing data frequencies are different from the Ensembl population data. The frequency of the minor allele is 0.267 as estimated from the sequencing data. In contrast the frequency within Africans is 0.106, 0.100 in Europeans, and 0.03 and 0.087 in Americans and Asians respectively. The frequencies from the study and the Ensembl are vastly different from those of An et al (2004) and Reddy et al (2010). The frequencies in these studies are very similar. For Africans the frequency is 0.0.91 and 0.089 for African Americans and Africans respectively. The European Americans exhibit a frequency of 0.063.

SNPs -590 is a fixed polymorphism in our samples. All the sequenced samples deviated from the reference sample at this position. The sequencing data is dissimilar from the Ensembl data (http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=22:39471748-39472748;v=rs17496004;vdb=variation;vf=32167980). In fact the frequency of the "minor" allele was 1 and the ancestral allele was not present in any of the sequenced samples. However the minor allele frequencies of Americans (Mexican Ancestry from Los Angeles USA) (0.03) and Europeans (Utah Residents (CEPH) with Northern and Western European ancestry) (0.07) are comparable to each other. The frequencies of the Yoruba (0.006) and Asians (Han Chinese in Bejing, China) (0.000) were also comparable.

The minor allele frequency at SNP -881 for Europeans (Iberian population in Spain, British in England and Scotland, Finnish in Finland, Toscani in Italia, Utah Residents (CEPH) with Northern and Western European ancestry) and Asians (Han Chinese in Bejing, China, Japanese in Tokyo, Japan, Southern Han Chinese, Chinese Dai in Xishuangbanna, China, Kinh in Ho Chi Minh City, Vietnam) is 0.001 and 0.003 respectively. In comparison the study data the minor allele frequency is 0.389 versus 0.287 for Africans (Yoruba in Ibadan, Nigeria, Luhya in Webuye, Kenya, Gambian in Western Divisons in The Gambia, Mende in Sierra Leone, Esan in Nigeria, Americans of African Ancestry in SW USA, and African Carribbeans in Barbados).

Comparing the sequencing data to Ensembl it is clear that direct sequencing does remain the method of choice for detecting variation. The frequency values differed amongst various populations. It is relatively expensive so in the context of this study it was used an exploratory tool. Thereafter, indirect genotyping assays were designed to genotype the study population.

4.2. Genotyping of -571 and H186R SNP

Allele specific amplification did allow the genotyping of some samples of study population at -571.

However, in this study the results of ASA differed from the RFLP analysis, as the population deviated from Hardy Weinberg equilibrium. In addition the minor

allele frequency is 0.59 and it deviates from the sequencing (0.267) and Ensembl data in Africans (0.106), Europeans (0.100), Americans (0.03), and Asians (0.087). In ASA homozygotes were represented by a single band in either reaction while heterozygotes are represented by two bands of equal intensity. Certain samples amplified with two bands but of very unequal intensity, a very bright band and a very dim band. Thus one has to use discretion when genotyping these samples by this method. The deviation from Hardy Weinberg is the result of the incorrect genotyping, which is the result of the assay not being specific enough to detect the variants correctly.

A great deal of time and effort has been devoted to the characterization and genotyping of SNPs but genotyping errors are not uncommon in population studies (Hosking *et al*, 2004). These can occur at any step of the genotyping process from sampling through to the actual procedure. While all efforts are made to eliminate errors it is important to characterise and quantify this error. This estimation become increasing important if populations of unrelated individuals (such as here) are genotyped because one cannot use Mendelian inheritance to check inconsistencies as in family studies. It has been shown that the source of the genotyping error can also be dependent on the method used to evaluate the SNP as is the case in this study where ASA revealed incorrect genotypes while RFLP analysis was reproducible and more reliable in genotyping (Hosking *et al*, 2004). While HWE is a good measure to identify genotyping errors it is important to note that this too is also dependant on the minor allele frequency of the SNP under investigation and the number of samples analysed. It has been documented that in

SNPs heterozygotes may be problematic as they may lead to a phenomenon of allelic dropout (Cutler, 2001). Commonly two reactions are used to genotype each locus and allelic dropout occurs when one of the reactions fails; this is interpreted as a homozygote even though in reality the individual is heterozygous at that locus. This can account for the increase in GG genotype of -571 when using ASA and subsequent deviation from HWE. In ASA the success of this technique is also 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). Thus it seems that the initial starting point of any study should be to determine the real error by an initial pilot study on a subset of samples which was not done in this study.

The RFLP analysis was a more reliable technique. The discriminatory power is great because it allows one to design the assay at a population level as in our study. The uncut control in the RFLP analysis ensured the assay was working

correctly. In addition, the sequence data allowed the assay to be tested with control samples where the genotype was known. The drawback of using this technique is that a large amount DNA is required. This can be overcome by an initial PCR amplification as in this study. Genotyping by RFLP analysis the study population did not deviate from the HWE. However SNP -571 minor allele frequency (0.84) is dissimilar from sequencing data (0.267) and Africans (Yoruba, Nigeria) (0.106). In this instance it seems that RFLP technique was not reliable in detecting the variation in-571. However it proved very reliable when detecting variation in H186R. The minor allele frequency of the study population was 0.32. This is similar to the frequency of African Americans from the work An et al and Africans in Reddy et al (2010) where the frequencies were 0.368 and 0.307 respectively.

Similarly, pyrosequencing is a reliable technique because it is based on sequencing. However, it does require a lot of technical work to get the assay working. Once the assay is optimized a large number of samples can be genotyped at once. On the other hand genotyping as the frequency in H186R of the minor allele (0.50) is different from published literature.

4.3. Genetic variation in South Africans

It is necessary to have an understanding of the local demographic history which has helped to outline 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 false associations between random markers with no physical linkage to susceptibility loci and a disease phenotype (Pritchard and Rosenberg, 1999).

The Zulu and Xhosa ethnic groups showed differences in allele frequency to each other and the group comprising individuals of mixed or unknown ancestry at polymorphic positions. The ethnic groups showed differences in allele frequency to each other and the group comprising individuals of mixed or unknown ancestry at all polymorphic positions. This supports the findings of Lane *et al.* (2002), that the black South African population does show distinct differences among the representative ethnic groups. 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 polymorphic positions. This inconsistency may however, simply is a consequence of the relatively small sample sizes available for study in this investigation.

The samples were representative of the population (Zulu speakers, Xhosa, Tswana, Sotho, Pedi, Tsonga, Venda, and Ndebele); some samples had to be pooled as suggested by Lane *et al* (2002) for subsequent analysis as they would be uninformative if not grouped.

Macrogroup 1 comprised speakers from all 9 ethnic groups, macrogroup 2 constituted Zulu speakers, macrogroup 3 consisted of Zulu and Xhosa Speakers

while macrogroup 4 encompassed Sotho, Tswana and Pedi speakers. A comparison of allele frequencies showed that group 1 had similar frequencies to that of macrogroups 2 and 3. This is not surprising that considering that Zulu and Xhosa speakers comprised 45 % of the population. Macrogroup 2 had similar frequencies to group 3, consistent with the findings by Lane *et al* (2002) that Zulu and Xhosa speakers have similar patterns of variations as they are linguistically similar. However group 4 had different allele frequencies in comparison to macrogroups 2 and 3. This is once again consistent with the findings of Lane *et al* (2002) that Sotho/Tswana speakers have different patterns of variation from Zulu and Xhosa speakers and are in fact linguistically diverse.

There is strong evidence for population substructure at the polymorphic loci across the macrogroups. If one compares the allele frequency of groups, group 1 has the same frequency as the average of macrogroups 3 and 4. This confirms that group 3 and 4 are subpopulations of group 1.

I did not find evidence of the Wahlund effect, which is the reduction of heterozygosity in pooled populations by comparison of genotype frequencies (Hartl and Clark, 1989). This effect is also indicative of population substructure and may also account for the pooled sample deviating from HWE, which is not the case in this study. However it is important to note that if there is no noteworthy difference of allele frequencies between the different groups, the heterozygosity reduction will not be sufficient to cause a deviation from HWE. This is consistent with the findings in this study that shows that the macrogroups

did deviate from HWE even though there is some evidence of population substructure at the loci explained earlier.

In addition LD also plays a pivotal role in population substructure. Africans have the highest genetic diversity owing to the largest number of variable genes and alleles (Jorder *et al*, 2000 & Tishkoff & Williams, 2002 & Cavalli-Sforza *et al*, 1997 & Jakobson *et al*, 2008). In comparison non-African populations do not have high genetic diversity predominantly due to genetic drift, which occurred during the migration of modern humans out of Africa and resulted in a small population (Tishkoff & Williams, 2002). Thus LD extends over shorter distances in Africans versus Non-Africans and that distance is estimated to be approximately 3-10kb (Reich *et al.*, 2001). However studies within the population's genetics laboratory reveal that this LD acts over an even shorter distance in African populations (Heitkamp *et al*, personal communication).

The pairwise allelic Linkage Disequilibrium of -571 and H186R showed that they are not in Linkage Disequilibrium, |D'| is 0.216. In contrast they appear to be linked in AA and the |D'| is 0.967 and 1.000 in EA (An *et al*, 2004). This lack of correlation is surprising as the SNPs are just over 5000bp apart. Thus should be in strong linkage equilibrium as suggested by the literature and our own laboratory estimates. However this shows that LD is strongly affect by population history, selection, and sample size. The sample size may not have been sufficient to detect the LD. In addition another compounding reason may be the selection of study participants. These participants were all matched for ethnicity but not other parameters which may have underlying effects on the LD. As mentioned

previously both D' and r^2 are measures used to quantify LD. However their interpretation is different. It is thought that while both ranges from 0 to 1, r^2 are the more accurate measure. This is so because there is an inverse relationship between r^2 and sample size. While the calculated r^2 value is small it is not dependant on the sample size. Importantly because genes are linked on the same chromosome does not mean they will be in linkage disequilibrium. It is known that intermediate values of D' from ~0.3 to 0.7 is difficult to interpret as the D' value can be highly variable in pairs of sites that are separated by large distance (Wall & Pritchard, 2003).

Haplotype analysis revealed four haplotypes between the SNPs within the whole population and when the population was grouped according to different languages. GA was the most common while CG was the most infrequent amongst all macrogroups. These differences in haplotype frequency, however small, 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.

APOBEC3G remains a promising target in the study of HIV/AIDS pathogenesis. Functional studies need to be conducted to ascertain the exact contribution of this gene product in HIV/AIDS. Future studies should include expression analysis on well characterised samples. A good example of expression analysis to use in

future studies is siRNA analysis directed against *APOBEC3G*. This will show the efficiency of *APOBEC3G* proteins on retroviral infection in a South African context where HIV-1 type C is the most prevalent subtype. To determine with accuracy if this gene is involved in susceptibility to HIV and progression to AIDS well characterised subjects need to be followed over time.

References

Ahmadian, A., Ehn, M., Hober, S. 2006. Pyrosequencing: History, biochemistry, future. *Clinica Chimica Acta* **363**: 83-94.

Alkhatib, G., Combadiere, C., Broder, C. C., Feng, Y., Kennedy, P. E., Murphy, P. M., Berger, E. A. 1996. CC CKR5: A RANTES, MIP- 1aplha, MIP- 1beta β receptor as a fusion co factor for the macrophage-trophic HIV. *Science* **272:** 1955-58.

An, P., Bleiber, G., Duggal, P., Nelson, G., May, M. Mangeat, B., Alobwede, I., Trono, D., Vlahov, D., Donfield, S., Goedert, J. J., Phair, J., Buchbinder, S., O'Brien, S. J., Telenti, A., Winkler, C. A. 2004. *APOBEC3G* Genetic Variants and Their Influence on the Progression to AIDS. *Journal of Virology* **78:** 11070-11076.

Armour, J. A. L., Anttinen, T., May, C. A., Vega, E. E., Sajantila, A., Kidd, J. R., Kidd, K. K., Bertranpetit, J., Pääbo, S., Jeffreys, A. J. 1996. Minisatellite diversity supports a recent African origin for modern humans. *Nature Genetics* **13**: 154-160.

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

Bakewell, Oliver and Hein de Haas (2007) African Migrations: continuities, discontinuities and recent transformations. in Patrick Chabal, Ulf Engel and Leo de Haan (eds.) *African Alternatives*. Leiden: Brill: 95-118.

Bailey, J. R., Zhang, H., Wegweiser, B. W., Yang, H., Herrera, L., Ahonkhai, A., Williams, T. M., Siliciano, R. F., Blankson, J. N. 2007. Evolution of HIV-1 in a HLA-B*57-positive patient during virologic escape. *Journal of Infectious Disease* **196**: 50-55.

Barre-Sinoussi, F., Chemann, J.C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest , J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., Montagnier, L. 1983. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science* **220**: 868-871.

Berglund, J., Pollard, K. S., Webster, M. T. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biology* **7**: e1000026.

Brumme, Z. A., Brumme, C. J., Heckerman, D., Korber, B. T., Daniels, M., Carlson, J., Kadie, C., Bhattacharya, T., Chui, C., Szinger, J., Mo, T., Hogg, R. S., Montaner, J. S., Frahm, N., Brander, C., Walker, B. D., Harrigan, P. R. 2007. Evidence of Differential HLA Class I-Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1. *PLoS Pathogens* **3**: e94.

Cann, R. L., Stoneking, M., Wilson, A. C. 1987. Mitochondrial DNA and human evolution. *Nature* **325**: 31-36.

Carrington, M., Nelson, G., O'Brien, S. J., 2001.Considering genetic profiles in functional studies of immune responsiveness to HIV-1. *Immunology Letters* **79**: 131-140.

Cavalli-Sforza, L. L. 1997. Genes, people and languages. *Proceedings of the National Academy of Sciences* **94**: 7719-7724.

Cavalli-Sforza, L. L., Minch, E., Mountain, J.L. 1992. Co evolution of genes and languages revisited. *Proceedings of the National Academy of Sciences* **89***:* 5260-5264.

Charron, D. 2005. Immunogenetics today: HLA, MHC and much more. *Current Opinion in Immunology* **17**: 493-497.

Chester, A., Scott, J., Anant, S., Navaratnam, N. 2000. RNA editing: cytidine to uridine conversion in apolipoprotein B mRNA. *Biochimica et Biophysica Acta* **1494:** 1-13.

Clapham, P.R, McKnight, A. 2001. HIV-1 receptors and cell tropism. *British Medical Bulletin* **58**: 43-59.

Conticello, S. G., Thomas, C. J. F., Petersen-Mahrt, S. K., Neuberger, M. S. 2005. Evolution of the AID/APOBEC family of polynucleotide (deoxy) cytidine deaminases. *Molecular Biology and Evolution* **22**: 367-377.

Devlin, B., Risch, N.1995. A comparison of Linkage Disequilibrium measure for fine scale mapping. *Genomics* **29**: 311-322.

Diamond, J., and Bellawood, P. 2003. Farmers and Their Languages: The First Expansions. *Science* **300**: 597-603.

Do, H., Vasilescu, A., Diop, G., Hirtzig, T., Heath, S., Coulonges, C., Rappaport, J., Therwath, A., Lathrop, M., Matsuda, F., Zagury, J. F. 2005. Exhaustive genotyping of the CEM15 (APOBEC3G) gene and the absence of association with AIDS progression in a French cohort. *The Journal of Infectious Disease* **191**: 159-163.

Doms, R. W., Trono, D. 2000. The plasma membrane as a combat zone in the HIV battlefield. *Genes and Development* **14:** 2677-2688.

Donfack, J., Buchinsky, F. J., Post C., Ehrlich, G. D. 2006. Human susceptibility to viral infection: The search for HIV protective alleles among Africans by means of genome-wide studies. *AIDS Research and Human Retroviruses* **22**: 925-930.

Fellay, J., Shianna, K. V., Ge, D., Colombo, S., Ledergerber, B., Weale M, Zhang, K., Gumbs, C., Castagna, A., Cossarizza, A., Cozzi-Lepri, A., De Luca, A., Easterbrook, P., Francioli, P., Mallal, S., Martinez-Picado, J., Miro, J. M., Obel, N., Smith, J. P., Wyniger, J., Descombes, P., Antonarakis, S. E., Letvin, N. L., McMichael, A. J., Haynes, B. F., Telenti, A., Goldstein, D. B. 2007. A whole genome association of major determinants for host control of HIV-1. *Science* **317**: 944-947.

Galtier, N., Duret, L., Glemin, S., Ranwez, V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. *Trends in Genetics* **25:** e1-5.

Goldstein, D. B., Weale, M. E. 2001. Population genomics: linkage disequilibrium holds the key. *Current Biology* **24**: R576-R579.

Gu, Y., Sundquist, W. L. 2003. Good to CU. *Nature* **424**: 21-22.

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

He, Z., Zhang, W., Chen, G., Xu, R., Yu, X. F. 2008. Characterization of conserved motifs in HIV-1 Vif required for APOBEC3G and APOBEC3F interaction. *Journal of Molecular Biology* **381**: 1000-1011.

Hedrick, P. W., Verrelli, B. C. 2006. 'Ground truth' for the selection on CCR5-Δ32. *Trends in Genetics* **22:** 293-296.

Heeney, J. L., Dalgeish, A. G., Weiss, R. A. 2006. Origins of HIV and the evolution of resistance to AIDS. *Science* **313**: 462-466.

Herbert, R. K. 1990. The Sociohistory of clicks in Southern Bantu. *Anthropological Linguistics* **32**: 295-303.

Hirsch, V., Olmsted, R. A., Murphey-Corb, M., Purcell, R. H., Johnson, P., R. 1989. An African primate SIV_{sm} closely related to HIV-2. *Nature* **339**: 389-392.

Holden, C. J. 2001. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *The Royal Society* **269**: 793-799.

Hosking, L., Lumsden, S., Lewis, K., Yeo, A., McCarthy, L., Bansal, A., Riley, J., Purvis, I., Xu, C. F. 2004. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics* **12**: 395-399.

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

Hudson, R. R. The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631 (1985).

Hutcheson, H. B., Lautenberger, J. A., Nelson, G. W., Pontius,J. U., Kessing, B. D., Winkler, C. A., Smith, M. W., Johnson, R., Stephens, R., Phair, J., Goedert, J. J., Donfield, S., O'Brien, S. J. 2008. Detecting AIDS restriction genes: From candidate genes to genome-wide association discovery. *Vaccine* **26:** 2951-2965.

Huthoff, H., Malim, M. H. 2005. Cytidine deamination and resistance to retroviral infection: towards a structural understanding of the APOBEC proteins. *Virology* **334**: 147-153.

Ingman, M., and Gyllensten, U. 2003. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Res.* **13:** 1600-6.

Jakobsson, M., Scholz, S. W., Scheet, P., Raphael Gibbs, J., VanLiere, J. M., Fung, H.C., Szpiech, Z. A., Degnan, J. H., Wang, K., Guerreiro, R., Bras, J. M., Schymick, J. C., Hernandez, D. G., Traynor, B. J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H. M., Hardy, J. A., Rosenberg, N. A., Singleton, A. B. 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998-1003.

Jarmuz, A., Chester, A., Bayliss, J., Gisbourne, J., Dunham, I., Scott, J., Navaratnam, N. 2002. An anthropoid specific locus of orphan C to U RNA editing enzymes on Chromosome 22. *Genomics* **79**: 285-296.

Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E. 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y chromosome data, *American Journal of Human Genetics* **66**: 979–988.

Keele, B. F., Van Herverswyn, F., Li, Y**.,** Bailes, E., Takehisa, J., Santiago, M. L., Bibollet-Ruche, F., Chen, Y., Wain, L. V., Liegeois, F., Loul, S., Ngole, E. M., Bienvenue, Y., Delaporte, E., Brookfield, J. F., Sharp, P. M., Shaw, G. M., Peeters, M., Hahn, B. H. 2006. Chimpanzees reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**: 523-526.

Keyue, D., Zhou, K., He, F., Shen, Y. 2003. LDA – A java-based linkage. *Bioinformatics* **19**: 2147- 2148.

Khan, M., Garcia-Barro, M., Powell, M. 2001. Restoration of wild-type infectivity to human immunodeficiency virus type 1 strains lacking nef by intravirion reverse transcription. *Journal of Virology* **24**: 12081-12087.

Kidd, K. K., Morar, B., Castiglione, C. M., Zhao, H., Pakistis, A. J., Speed, W. C., Bonne-Tamir, B., Lu, R. B., Goldman, D., Lee, C., Nam, Y. S., Grandy, D. K., Jenkins, T., Kidd, J. R. 1998. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Human Genetics* **103**: 211-227.

Klenerman, P., Wu, Y., Phillips, R. 2002. HIV: current opinion in escapology. *Current Opinion in Microbiology* **5:** 408-413.

Klitz, W., Brautbar, C., Schito, A. M., Barcellos, L. F., Oksenberg, J. R. 2001. Evolution of the CCR5 Delta32 mutation based on haplotype variation in Jewish and Northern European population samples. *Human Immunology* **62**: 530-538.

Kwok, S., Kellogg, D. E., McKinney, N., Spasic, D., Goda, L., Levenson, C., Sninsky, J. J. 1990. Effects of primer-template mismatches on the polymerase chain reaction: human immunodeficiency virus type 1 model studies. *Nucleic Acids Research* **18**: 999-1005.

Lane, A. B., Soodyall, H., Arndt, M. E., Ratshikhopha, E., Jonker, C., Freeman, C., Young, L., Morar, B., Toffie, L. 2002. Genetic substructure in South African Bantu-speakers: evidence from autosomal DNA and Y-chromosome studies. *American Journal of Physical Anthropology* **119**: 175-185.

Lemey, P., Pybus, O. G, Wang, B., Saksena, N. K., Salemi, M., Vandamme, A. M. 2003. Tracing the origin and history of the HIV-2 epidemic. *Proceedings of National Academy of Sciences* **100**: 6588-6592.

Letvin, N., Barouch, D., Montefiori, D. 2002. Prospects for vaccine protection against HIV-1 infection and AIDS. *Annual Review of Immunology* **20**: 73-99.

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

Libert, F., Cochaux, P., Beckman, G., Samson, M., Aksenova, M., Cao, A., Czeizel, A., Claustres, M., de la Rúa, C., Ferrari, M., Ferrec, C., Glover, G., Grinde, B., Güran, S., Kucinskas, V., Lavinha, J., Mercier, B., Ogur, G., Peltonen, L., Rosatelli, C., Schwartz, M., Spitsyn, V., Timar, L., Beckman, L., Parmentier, M., Vassart, G. 1998. The ΔCCR5 mutation conferring protection against HIv-1 in Caucasian population has a single and recent origin in northeastern Europe. *Human Molecular Genetics* **7**: 3399- 3406.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., Trono, D. 2003. Broad antiretroviral defense by human APOBEC3G through lethal editing of nascent reverse transcripts. *Nature* **424**: 99-103.

Mehle, A. J., Goncalves, M., Santa-Maria, M., McPike, M., Gabudza, D. 2004. Phosphorylation of a novel SOCS-box regulates the assembly of the HIV-1 Vif-Cul5 complex that promotes APOBEC3G. *Genes & Development* **18**: 2861-2866.

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

Meyerson, N., and Sawyer, S., (2011).Two stepping through time: mammals and viruses. *Trends in Microbiology* **19**: 286-294.

Miller, J. H., Presnyak, V., Smith, H. C. 2007. The dimerization domain of HIV-1 viral infectivity factor Vif is required to block virion incorporation of APOBEC3G. *Retrovirology* **4**: 81-92.

Mitchell, P. 2010. Genetics and southern African prehistory: an archaeological view. *Journal of Anthropological Sciences*, **88**:73–92.

Mitchell, P., Whitelaw, G. 2005. The archaeology of southernmost Africa from c. 2000 BP to early 1800s: A review of recent research. *The Journal of African History* **46:** 209-241.

Mountain, J. L., Risch, N. 2004. Assessing genetic contributions to phenotypic differences among "racial" and "ethnic" groups. *Nature Genetics* **36**: S48-S53.

Newton, C. R., Graham, A., Heptinstall, L., Powell, S. J., Summers, C., Kalsheker, N., Smith, J. C., Markham, A. F. 1989. Analysis of any point mutation in DNA. The amplification refractory mutation system (ARMS). *Nucleic Acids Research* **17**: 2503-2515.

Novembre, J., Galvani, A. P., Slatkin, M., 2005. The geographic spread of the CCR5-Δ32 HIV-Resistance allele. *PLoS Biology* **3**: e339.

O'Brien, S.J., Nelson, G. W. 2004. Human genes that limit AIDS. *Nature Genetics* **36**: 565-574.

Okayama, H., Curiel, D. T., Brantly, M. L., Holmes, M. D., Crystal, R. G. 1989. Rapid, nonradioactive detection of mutations in the human genome by allele-specific amplification. *Journal of Laboratory and Clinical Medicine* **114**: 105-113.

Papathanasopoulos, M., Cilliers, T., Morris, L., Mokii, J., Dowling, W., Birx, D. L., McCutchan, F. E. 2002. Full length genome analysis of HIV-1 subtype c utilizing CXCR4 and intersubtype recombinants isolated in South Africa. *AIDS Research and Human Retroviruses* **18**: 879-886.

Pritchard, J.K., Rosenberg, N.A. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *American Journal of Human Genetics* **65**: 220-228.

Ramdin, R., 2003. Population variation in CEM 15, a gene involved in HIV replication. BSc (Hons). University of Witwatersrand.

Reddy K., Winkler C.A, Werner L., Mlisana K., Abdool Karim SS, Ndung'u T. 2010. APOBEC3G expression is dysregulated in primary HIV-1 infection and

polymorphic variants influence CD4+ T-cell counts and plasma viral load. *AIDS* **24**: 195-204.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R., Lander, E. S. 2001. Linkage disequilibrium in the human genome. *Nature* **411**: 199-204.

Ronaghi, M. 2003. Pyrosequencing for SNP genotyping. *Methods in Molecular Biology* **212**: 189-195.

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

Sabeti, P. C., Walsh, E., Schaffner, S. F., Varilly, P., Fry, B., Hutcheson, H. B., Cullen, M., Mikkelsen, T. S., Roy, J., Patterson, N., Cooper, R., Reich, D., Altshuler, D., O'Brien, S., Lander, E. S. 2005. The case for selection at CCR5-Delta32. *PLoS Biology* **3**: e378.

Samson, M., Libert, F., Doranz, B. J., Rucker, J., Liesnard, C., Farber, C. M., Saragosti, S., Lapoumeroulie, C., Cognaux, J., Forceille, C., Muyldermans, G., Verhofstede, C., Burtonboy, G., Georges, M., Imai, T., Rana, S., Yi, Y., Smyth, R. J., Collman, R. G., Doms, R. W., Vassart, G., Parmentier, M. 1996. Resistance

to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **383**: 722-725.

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

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

Sawyer, S. L., Emerman, M., Malik, H. S. 2004. Ancient adaptive evolution of the primate antiviral DNA-Editing enzyme APOBEC3G. *PLoS Biology* **2**: 1278-1285.

Sheehy, A. M., Gaddis, N. C., Choi, J. D., Malim, M. H. 2002. Isolation of human gene that inhibits HIV-1 infection and is suppressed by the viral Vif protein. *Nature* **939**: e1-5.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., Darvasi, A. 2003. Linkage disequilibrium patterns of the human genome across populations. *Hum. Mol. Genet*. **12**: 771-776.

Sova, P., van Ranst, M., Gupta, P., Balachandran, R., Chao, W., Itescu, S., McKinley, G., Volsky, D. J. 1995. Conservation of an intact human

immunodeficiency virus type 1 vif gene in vitro and in vivo. *Journal of Virology* **69**: 2557-2564.

Stephens, J. C., Reich, D. E., Goldstein, D. B. 1998. Dating the origin of the CCR5-Δ32 AIDS-Resistance Allele by the coalescence of Haplotypes. *American Journal of Human Genetics* **62:** 1507-1515.

Stephens, M., Smith, N., Donnelly, P. 2001. A new statistical method for haplotype reconstruction based from population data. *American Journal of Human Genetics* **68**: 978-989.

Syvanen, A. 2001.Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Genetics* **2**: 930-940.

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

Tishkoff, S. A, Williams, S. 2002. Genetic analysis of African populations: human evolution and complex disease. *Nature Review Genetics* **3**: 611- 621.

Tishkoff, S. A., Kidd, K. K. 2004. Implications of biogeography of human populations for "race" and medicine. *Nature Genetics* **36**: S21-S27.

Tishkoff, S. A., Verrelli, B.C. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics* **4**: 293-340.

Tishkoff, S. A., Dietzsch, E., Speed, W., Pakstis, A. J., Kidd, J. R., Cheung, K., Bonné-Tamir, B., Santachiara-Benerecetti, A. S., Moral, P., Krings, M. 1996. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**: 1380-1387.

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. *American Journal of Human Genetics* **67**: 901-925.

Turner, B., Summers, M. 1999. Structural biology of HIV. *Journal of Molecular Biology* **285**: 1-32.

Van Harmelen, J., Wood, R., Lambrick, M., Rybicki, E. P., Williamson, A. L, Williamson, C. 1997. An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. *AIDS* **11**: 81-87.

Wall, J.D., Pritchard, J.K. 2003. Haplotype Blocks and Linkage disequilibrium in the Human Genome. *Nature Review Genetics* **4**: 587-597

VanLiere, J. M., Rosenberg, N. A. 2008. Mathematical properties of the r2 measure of linkage disequilibrium. *Theoretical Population Biology* **74**: 130-137.

Vartanian, J. P., Sommer, P., Wain-Hobson, S. 2003. Death and the retrovirus. *Trends in Molecular Medicine* **9**: 409-413.

Vasina, J. 1995. An association between HIV-1 subtypes and mode of transmission in Cape Town, South Africa. *Journal of African History* 36: 173-195.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., Wilson, A. C. 1991 African populations and the evolution of human mitochondrial DNA. *Science* **253**: 1503-1507.

Watson, E., Bauer, K., Aman, R., Weiss, G., von Haeseler, A., Pääbo, S. 1996. mtDNA sequence diversity in Africa. *American Journal of Human Genetics* 59: 437-444 .

Wedekind, J., Dance, G., Sowden, M., Smith, H. 2003. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends in Genetics* **19**: 207-216.

Williamson, C., Engelbrecht, S., Lambrick, M., van Rensburg, E., Wood, R., Bredell, W., Williamson, A. L. 1995. HIV-1 subtypes in different risk groups in South Africa. *Lancet* **346**: 782.

Wilson, A. C., Cann, R. L. 1992. The recent African genesis of humans. *Scientific American* **266:** 68-73.

Winkler, C. A., An, P., O'Brien, S., 2004. Patterns of ethnic diversity among the genes that influence AIDS. *Human Molecular Genetics*: R9-R19.

Worobey, M., Gemmel, M., Teuwen, D. E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J. J., Kabongo, J. M., Kalengayi, R. M., Van Marck, E., Gilbert, M. T., Wolinsky, S. M. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* **455**: 661-665.

Wu, D. Y., Ugozzoli, L., Pal, B. K., Wallace, R. B. 1989. Allele-specific enzymatic amplification of β-globin genomic DNA for diagnosis of sickle cell anemia. *Proceedings of National Academy of Sciences* **86**: 2757-2760.

Yu, X., Yu, Y., Liu, B., Luo, K., Kong, K., Mao, P., Yu, X. F. 2003. Induction of APOBEC3G ubiquitination and degradation by an HIV-1Vif-Cul5-SCF complex. *Science* **302**: 1056-1060.

Zhang, J., Webb, D. M. 2004. Rapid evolution of primate antiviral enzyme APOBEC3G. *Human Molecular Genetics* **13**: 1785-17.

Zhang, L., Saadatmand, J., Li, X., Guo, F., Niu, M., Jiang, J. 2008. Function analysis of sequences in human APOBEC3G involved in Vif-mediated degradation. *Virology* **370**: 113-112.

The Joint United Nations Programme on HIV/AIDS / World Health Organization**.** 2007. AIDS Epidemic Update.

The Joint United Nations Programme on HIV/AIDS / World Health Organization. 2005. AIDS Epidemic Update

Appendix I

Table A.1: Raw genotyping data obtained using the allele-specific PCR at SNP -571 for the JHB samples, RFLP-PCR and Pyrosequencing™ assays

Sample Description	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA
JHB	104	CC	118	CC	147	CC	160	CC	177	GC	191	GC
	105	CC	120	CC	148	CC	161	CC	179	GC	192	GC
	106	CC	123	CC	149	CC	163	CC	180	GC	193	GC
	109	CC	124	CC	150	CC	164	CC	181	GC	194	GC
	110	CC	126	CC	151	CC	167	CC	182	GC	195	GC
	111	CC	127	CC	152	CC	168	CC	183	GC	197	GC
	112	CC	131	CC	153	CC	170	CC	184	GC	198	GC
	113	CC	138	CC	154	CC	171	CC	185	GC	199	GC
	114	CC	140	CC	155	CC	172	CC	187	GC	200	GC
	115	CC	141	CC	156	CC	173	CC	188	GC		
	116	CC	143	CC	158	CC	175	GC	189	GC		
	117	CC	146	CC	159	CC	176	GC	190	GC		

Table A.2: Raw genotyping data obtained using the allele-specific PCR at SNP -571 for the GP samples.

Sample Description	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA
GP	201	GC	208	GC	214	GC	221	GC	228	GG	235	GG
	203	GC	209	GC	216	GC	222	GC	229	GG	236	GG
	204	GC	210	GC	217	GC	223	GG	230	GG	237	GG
	205	GC	211	GC	218	GC	224	GG	231	GG	239	GG
	206	GC	212	GC	219	GC	225	GG	232	GG	241	GG
	207	GC	213	GC	220	GC	227	GG	234	GG	243	GG
											244	GG
											245	GG

Table A.3: Raw genotyping data obtained using the allele-specific PCR at SNP -571 for the HIVNET samples.

Sample Description	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA	Sample #	571-ASA
HIVNET	521-160	GG	536-149	GG	541-193	GG	541-73	GG	615-31	GG	616-445	GG
	521-171	GG	536-173	GG	541-228	GG	541-98	GG	615-325	GG	616-453	GG
	521-298	GG	536-31	GG	541-234	GG	614-121	GG	615-332	GG	616-457	GG
	521-301	GG	541-115	GG	541-256	GG	615-107	GG	615-67	GG	616-472	GG
	521-327	GG	541-131	GG	541-353	GG	615-11	GG	615-78	GG	616-486	GG
	521-343	GG	541-144	GG	541-36	GG	615-136	GG	615-80	GG	616-499	GG
	536-015	GG	541-178	GG	541-49	GG	615-15	GG	615-93	GG	616-503	GG
	536-107	GG	541-180	GG	541-62	GG	615-26	GG	616-17	GG	616-90	GG
											615-366	GG

Table A.4: Raw genotyping data obtained using the RFLP-PCR at SNP -571 and Pyrosequencing at SNP H186R for the JHB samples.

Sample Description	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571-RFLP
JHB	101	CT	GC	113	TT	GG	126	CC	GG	142	CT	GG
	105	CT	GG	114	TT	CC	127	TT	GC			
	106	CT	GC	118	CC	GC	131	CT	GG			
	111	CC	GG	119	CC	GG	141	CT	GC			

Table A.5: Raw genotyping data obtained using the RFLP-PCR at SNP -571 and Pyrosequencing at SNP H186R for the GP samples.

Sample Description	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571-RFLP
GP	206	TT	GG	217	CT	GG	230	CT	GG	240	TT	GG
	208	TT	GG	221	CT	GG	231	CT	GG	241	TT	GG
	209	CT	GG	222	CT	GG	232	CT	GG	242	TT	GC
	210	CT	GG	223	TT	GG	233	CC	GG	243	CC	GC
	212	TT	GG	224	TT	GG	234	CT	GG	245	CT	GG
	213	CT	GG	226	CT	GG	235	TT	GG			
	214	CT	GG	227	CT	GC	237	CT	GG			
	215	CT	GG	228	CT	GG	238	TT	GG			
	216	CT	GG	229	CT	GG	239	CC	GG			

Table A.6: Raw genotyping data obtained using the RFLP-PCR at SNP -571 and Pyrosequencing at SNP H186R for the HJ samples.

Sample Description	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571-RFLP	Sample	H186R-PYRO	571- RFLP	Sample	H186R-PYRO	571-RFLP
HJ	300	CT	GG	322	CT	GG	344	CT	GC	375	CT	GC
	301	CC	GG	323	CT	GG	346	CT	GG	376	CT	GC
	302	CC	GG	324	CT	GC	347	TT	GC	377	CT	GC
	303	CC	GC	325	CT	GC	348	CT	GC	378	CT	GG
	304	CC	GC	326	CT	GC	349	TT	CC	381	CT	GG
	305	CC	GC	327	CC	GC	357	CT	GG	383	CT	GC
	306	CT	GG	328	CT	GC	358	CT	GG	384	CT	GG
	307	CC	GG	329	CT	GC	356	CT	GC	385	CT	GG
	308	CC	GG	330	CT	GC	379	CT	GG	386	CT	GG
	309	CT	GG	331	CT	GC	361	TT	GG	387	TT	GG
	310	CT	GG	332	CC	GC	362	CT	GG	390	CT	GG
	311	CT	GG	333	CT	GC	363	CT	GG	391	CC	GG
	312	CT	GC	334	CT	GC	365	CT	GG	392	CT	GG
	313	CT	GG	335	CC	GC	366	CT	GG	393	TT	GG
	314	TT	GC	336	CT	GC	367	CT	GG	394	CT	GG
	315	TT	GG	337	CT	GC	368	CT	GG	395	CT	GG
	316	CT	GG	338	CT	GC	369	CC	GG	396	CC	GG
	317	CC	GG	339	CT	GG	370	CC	GG	397	CC	GG
	318	CT	GG	340	TT	GG	371	TT	GG	398	CT	GG
	319	CT	GG	341	CT	GC	372	CT	GC	399	TT	GG
	320	CT	GG	342	CC	GG	373	CT	GC	400	CT	GG
	321	CT	GG	343	TT	GG	374	TT	GC	401	CT	GG
										402	CT	GG

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