SNP and Haplotype characterisation of *APOBEC3G*, 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, March 2009

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). It has not been submitted before for any degree or examination in any other University.

_____

Roshilla Ramdin

_____

Date

Abstract

Heritable variation is important in disease progression, therefore its association with HIV/AIDS was analyzed. *APOBEC3G* is a unique cellular gene that influences HIV infectivity. It belongs to family of cytidine deaminases and is both an RNA and DNA editing enzyme. *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 defense may provide the basis for the design of new therapies for HIV. Variation in the upstream non-coding region of *APOBEC3G* was studied. Six base variants were found at positions -90, -163, -166, -571, -590 and -821. In addition, promoter analysis identified promoters in the upstream non-coding region. Indirect genotyping assays were developed to genotype the participants at -571 and H186R. The frequency of -571 GG was 70 %. The frequency of the TT genotype of H186R was 20 %. The GG genotype was selected against in the HIV + group of the study participants. This is indicative that this SNP has disease modifying effects. The TT genotype was related to increased progression to AIDS confirming the results of previous studies.

ACKNOWLEDGEMENTS

Firstly, I would like to extend my gratitude to my supervisor Prof T McLellan for her constant help and support. Without her encouragement, I would not have succeeded in completing this MSc.

Secondly, 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.

TABLE OF CONTENTS

Methods

Results

List of Tables

List of Figures

Figure 10. Chromatograms of re-analyzed sequences from Honours project in 2003. A) 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. B) These chromatograms show the artifact of sequencing which was erroneously interpreted as a SNP in 2003. If look at sample 114 then one can see that at the circled position it seems as if this sample is heterozygous for both alleles. C) Chromatograms showing -571 SNP and -590 loci (purple 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 allele. Thus this is an example of a fixed polymorphism.

Figure 11. PCR products from -571 SNP were run on a 0.7 % agarose gel. The reactions for each allele were done separately but run on the same gel. This gel sows the -571 G and -571 C reactions respectively. 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 7 samples 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, 5, and 7 are genotyped CC while sample 6 is genotyped GC.

Figure 12. PCR cycling conditions for genotyping -571 SNP.

Figure 13. The gradient PCR products (400bp) for three samples 216, 340, 365 that were run on a 1 % agarose gel. Lanes 1, 2, 3 represent sample 216,340, & 365 at 60°C. Lanes 5, 6, & 7 represent samples 216, 340 & 365 at 59.3°C. Lanes 9, 10 & 11 represent samples 216, 340, 365 at 58°C. The annealing temperature of 58°C was chosen as it gave the best amplification in two of the three samples.

Figure 14. DNA fragments after restriction digestion with MvaI for 4 hours at 37 ℃. Lanes 1 represents the uncut control DNA of size ~378 bp. Lanes 2-5 represent homozygotes for the G allele with fragment sizes at 215bp and 163bp. Lane 7-10 represent heterozygotes with fragment sizes at 215bp, 163bp, 114bp and 49bp. The 49bp fragments are not clearly visible on the gel. No homozygotes for the C allele were present.

Figure 15. Chromatogram of sample 310 sequence. The highlighted strip shows a heterozygote for this sample at position 186. Codon 185 represented by nucleotides 530, 531, 532 in the chromatogram are not polymorphic in this sample.

Figure 16. 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.

1. HIV/AIDS

Universal statistics reveal that Sub-Saharan Africa has the highest HIV/AIDS infections. In Sub-Saharan Africa there are 22.5 million adults living with HIV and approximately 1.7 million adults and children had become infected with the virus in 2007.These alarming statistics are cause for concern especially since Southern Africa accounts for one third of all new HIV infection and AIDS deaths globally. South Africa is the hardest hit country with largest number of people infected and has the highest HIV prevalence of any country. The virus is spreading throughout the population and is not limited to high risk populations such as sex-workers (UNAIDS/WHO, 2005). The key to curbing the spread of the virus is to implement efficient and practical strategies to help those infected with and affected by HIV/AIDS.

2. History of HIV

2.1. From SIV to HIV

The explosion of HIV-1 infection in humans can be traced to the origins of the ancestral virus. HIV originated from the zoonotic transfer of SIV from non human primates. Molecular evidence is consistent with this transfer having occurred in Cameroon. SIV was previously only identified in captive chimpanzees *(Pan troglodytes troglodytes)* (Keele et al, 2006). The ability to detect this virus in wild living chimps and use these viral sequences in phylogenetic analysis, allows for the expectation of whether SIV is indeed the origin of HIV-1 infection.

Chimpanzees are native to Cameroon, which is located north of the DRC (Keele et al, 2006). Two subspecies of chimps inhabit this region and are geographically isolated by the Sanaga River. *P.t vellerosus* inhabits north of the river while the *P.t.traglodytes* is found south of the river much closer to the DRC (Keele et al,

2006). Different regional sites were sampled in the southern region of Cameroon. HIV-1 proteins such as gag, pol, and env were detected in the samples. The Pol and gp41 sequences from wild living chimps showed that HIV-1 group M subtype clusters identified new strains that were sequenced (Keele et al, 2006). Interestingly the HIV-1 M/A is most closely related to the samples from the region bordering the DRC. HIV-1 N clusters more close to the samples from further north of the border, between Cameroon and the DRC.

This clustering is indicative of an ancestral virus, most likely SIV, which infected an individual/s who then migrated into the DRC, specifically Leopoldville (Kinshasa), as it was the metropolis of the early twentieth century where the HIV-1 epidemic has its origins. However, the origin of SIVcpz itself remains elusive. SIVs occur in 26 nonhuman primate species (Keele et al, 2006). They do not, however, cause disease in the natural hosts, but form a natural reservoir of the virus (Hahn et al, 2000). This is confirmed by the taking a closer look at the viral lineages. The divergence of several clusters was most likely caused by the evolution of the host lineage because SIV form distinct host specific clusters within the phylogenetic tree (Hahn et al, 2000). SIV cpz as mentioned earlier is found in wild *P.t.troglydytes* endemic in central Africa and these viruses form a monophyletic clade distinct from any other lentiviruses. Thus again confirming another example of host-dependant viral evolution (Hahn et al, 2000). Sequence analysis has helped clarify zoonotic transmission of SIVs to humans. It is estimated that there have been no fewer than 7 independent cross species transmission to humans giving rise to HIV-1 groups (Hahn et al 2000).The characterization of SIVcpz across the gag, pol and env regions has shown that this virus is clearly a recombinant virus. This is so because the SIV cpz is not particularly closely related to any one human or chimp virus. The phylogenetic analysis revealed a single breakpoint in the tree when looking at Env and pol proteins. SIVcpz Pol sequences clustered more closely with SIVrcm (SIV from Red Capped Mangabeys) during the phylogeny. SIV cpz Env sequences clustered

more closely with SIVgsn (Greater Spot-nosed monkeys) sequences than any other SIV lineages. This is direct evidence that the hybrid SIVcpz has a more recent transmission to humans. Therefore by extrapolation it can be argued that SIVs from monkeys can also have been transmitted to humans and may also allude to the origins of HIV-1. Human exposure to SIVs is mostly likely through contact with animal blood either through consumption of bush meat or through biting and of predation of wild living species (Worobey et al, 2008). The only plausible explanation for the spread of SIVs to humans and subsequent evolution of HIV is that SIVs must have had to undergo some changes to adapt to the host in addition to taking advantage of the socio-behavioural factors that were rife in Africa at the time (Gao et al, 1999).

2.2 HIV in humans

HIV-1 group M accounts for more than 95 % of all HIV infection around the world save for HIV-2 infections in certain parts of Africa (Lemey et al, 2003). Group M is composed of numerous different subtypes, each endemic in certain parts of the world, e.g. subtype B is found mainly in Europe, the Americas, Japan and Australia (Heeney et al, 2006). In southern Africa, subtype C is the most common. (Lemey et al, 2003).

DNA sequencing of archival samples from Zaire (now the DRC) was used to date the origin of HIV -1 and to discern its evolutionary history (Worobey et al, 2008). The archival samples are designated ZR 59 (a blood plasma sample from 1959) and DRC60 (biopsy specimen from female patient). The phylogenetic analysis showed that these sequences shared a common ancestor at least 50 years ago because of the short nodal distance between the two (Worobey et al, 2008). 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 50yrs ago group M strains had evolved into distinct subtypes that were circulating within the

populations of this region (Worobey et al, 2008). The phylogenetic analysis indicates too that there is substantial genetic diversity between the two ancestral sequences. This further confirms that the virus was present long before the epidemic was characterized.

Urbanization seems to have played a vital role in the exponential rise of the disease in Africa because the two ancestral sequences cluster with other strains from the same region rather then 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).

Like HIV-1 the causative agent of HIV-2 is known to be SIVsm (SIV from Sooty Mangabey). The Sooty Mangabeys naturally inhabit the forest of Senegal east to Ghana. The origin of HIV-2 has not been so contenscious. 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 a result of 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 singly infected people. Like HIV-1 transmission of the virus as been marked by a period of untracability 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 (Lemey et al, 2003). Once again it is evident that viral epidemics are reliant on socio-economic conditions.

3. HIV Life Cycle

HIV-1 has a high mutation rate of 3×10^{-5} per nucleotide base per replication cycle (Sharp, 2002). Because of this high mutation rate the virus evolves rapidly. This is challenging because this generates masses of sequence change per decade

of the pandemic. Nonetheless, the viral sequence analysis is what is producing the most important information on the origins of HIV-1 and its co-evolution with humans.

The HIV infection begins with HIV attaching to the CD4 cell via the gp120 molecules exposed on the surface of the virion. This binding induces a conformational change in gp 120 allowing it to bind to 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 capsid (which contains the two RNA strands) into the cellular compartment of the host cell. Once in the host cell there is a partial uncoating of the capsid to expose the RNA strands (Doms and Trono, 2000). The viral reverse transcriptase converts the RNA to DNA as this is a much more stable molecule in the cell. The DNA then enter in host nucleus where together with viral intergrase is spliced 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 and can become active at any stage. If the provirus is active it will generate viral proteins for new virions. It uses the host cells own enzymes to transcribe the double stranded DNA to mRNA. Once the mRNA is processed in the nucleus, it is transported to the cytoplasm where the viruses once again hi-jacks the host protein making processes to produce the viral proteins such as env(gp 160), gag, gag-pol, vif vpr, vpu, rev, tat and nef (Doms and Trono, 2000). The initial env protein is processed in the ER and Golgi into the gp120 and gp41. The gp 120 is glycosylated .The gag and gag-pol polyproteins aggregate near the membrane and interact with plasma membrane and the gp 41 present in the membrane (Doms and Trono, 2000). As the gag and gag-pol aggregate at the plasma membrane the virion begins to assemble. Subsequently the new virions are 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 enzyme cleaves the precursor gag, gag-pol into functional proteins.

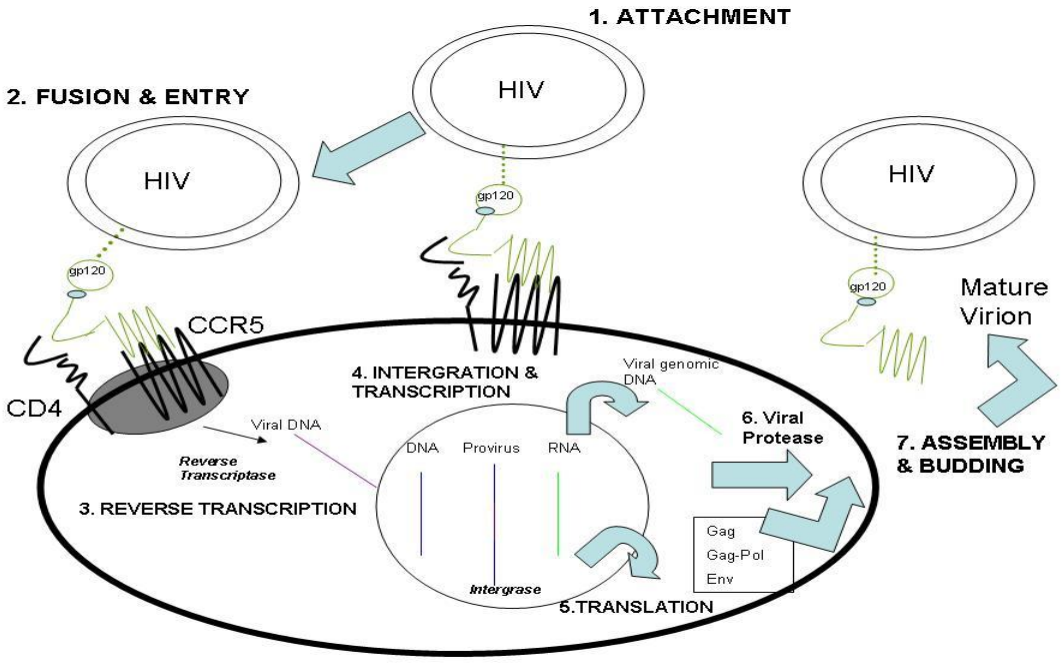


Figure 1. Overview of HIV replication detailing the steps of HIV infection.

4. Identification of host genes important in HIV

The HIV genome is composed of nine genes which encode 3 structural proteins (gag, pol, env), 2 regulatory proteins (tat, rev) and four accessory proteins (nef, vif, vpu, vpr) (Doms and Trono, 2000). It is important to understand how natural selection has had an impact on the HIV and SIV genomes because genome variation undoubtedly has profound effects on the virus as well as the host. SIVs do not cause disease in their natural hosts probably because this virus and the host have reached equilibrium because of the long occurrence of the virus in chimps and monkeys (Bailes et al, 2008). It is reasonable to assume that the SIV genes would have to adapt to their host environment in order for the virus to become transmissible. Examination of the HIV-1 and HIV-2 groups reveals subtypes with differing rates of transmission in different populations. In addition some subtypes are mosaics of two distinct subgroups e.g. Subtype A + CRF02_AG accounts for a large proportion of new infections worldwide. This viral heterogeneity is also

common in SIVs indicated by SIVcpz which is a hybrid of SIVrcm and SIVgsn (Bailes et al, 2008). This enforces that the viruses have had to adapt to their changing environments in order to ensure transmissibility.

There are many approaches that have been used to identify host proteins important in HIV. Classical analysis of host genetics and its involvement in viral genetics was studied with candidate gene studies. Numerous AIDS restriction genes that effect susceptibility to viral infection have been discovered via this approach (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. It is often better to have a larger study population to ensure accuracy of predictions. The disadvantage of this method lies with the fact that the investigator must have some knowledge about the candidate gene or the study could be uninformative and a waste. 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 Sub-Saharan blacks where the epidemic is flourishing at an alarming rate. Nonetheless, a lot of population studies have focussed on African Americans. This does not give us a complete picture of the disease in Sub-Saharans Africa but is a good platform from which to grow.

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 have all been discovered by candidate gene approaches.

5. Candidate genes for HIV

5.1. CCR5

The most widely characterized cell surface molecule is CCR5, a chemokine receptor (Carrington et al, 2001). Chemokines are secreted from cells in response to an inflammatory reaction. This attracts white blood cells to the site of inflammation. Chemokines CCL 5, CCL3, CCL4, formerly called RANTES, MIP 1α and MIP 1β, bind to chemokine receptors.

Co-receptor selectivity in HIV-1 is dependant on the V3 loop of HIV. A portion of the V3 loop has a remarkably similar structure to the CXC and CC chemokines (natural ligands for CXCR4 and CCR5, respectively) β2-β3 hairpin loop (Cardozo et al, 2007). Functional role of V1/V2 region of HIV-1 in infection showed that the V3 loop is necessary for viral tropism, although the exact mechanism of action was not known (Koito et al, 1994). Modelling of the V3 loop showed the importance of three amino acids; 11, 24, 25 in determining this viral tropism. When these three amino acid residues have a positive charge then the virus will bind using CXCR4. The converse is true for negatively charged amino acid residues (Cardozo et al, 2007). In addition V3 loop like V1/V2 region protects the env from being targeted by neutralizing antibodies present in the host sera. Truncation of this loop confirms that HIV will adapt their use of CCR5 for their propagation (Laakso et al, 2007). This interesting feature gives insight into the mechanisms of drug resistance that HIV has and the novel manner it will interact with the cell surface to ensure its propagation. The mutations discussed above are in specific reference to HIV and how its mutations can alter the use of chemokines receptors which undoubtedly has an effect on viral propagation.

The CCR5 co-receptor is a seven-transmembrane–spanning G-protein coupled receptor. The deletion allele of CCR5 has a 32 base pair deletion within the coding region. This leads to a frame shift mutation that introduces a stop codon in

the reading frame. As a result, a non-functional, truncated protein is produced (Alkhatib et al, 1996 & Carrington et al, 2001). Thus, there is no CCR5 on the cell surface. People homozygous for the mutation are resistant to initial infection by viruses utilizing CCR5 (Samson et al, 1996). However, CCR5 Δ32 homozygotes the viruses use CXCR4 co-receptor (Clapham & McKnight, 2001). Heterozygotes, however, are not resistant to infection but rate of progression to AIDS is delayed.

The CCR5 deletion is found at a high frequency in northern Europe, a lower frequency in southern Europe and is non-existent in Africans. The CCR5 deletion is remarkable in that it may be a good example of a locus under adaptive evolution. The locus may have become preferentially selected to protect individuals where there are other illnesses endemic in that region (Stephens et al, 1998). There is conflicting evidence for this positive selection. Some evidence is consistent with CCR5 del32 having been selected in human populations. Other evidence indicates that it could have become common by random population processes. Examination of frequencies of the CCR5 deletion in samples from the Bronze Age as compared with modern samples and 14^{th} century samples confirmed that this deletion was present long before it was selected as a preventative haplotype for plague or smallpox (Hedrick & Verrelli, 2006). But if the CCR 5 deletion is under positive selection then the linked variants around the CCR5 locus will have a reduced heterozygousity and high LD as the variant becomes common in the population. However, this is not the case for the CCR5 locus. The LD for the SNPs around this locus was the same across the European, Asian and African populations (Sabeti et al, 2005). Extensive genetic analysis confirms that indeed this locus is under neutral evolution because deletion was not a significant in terms of the F_{st} among all the study populations. This confusion arose because there has been a steady increase in the frequency of the deletion in European populations across Europe (Novembre et al, 2005).

Indeed recent studies have shown the earlier methods used for detection of hotspots for positive selection may be flawed. Hotspots which have been evolutionary accelerated may be a result of GC-biased gene conversion (gBGC) and not signatures of positive selection after all (Berglund et al, 2009). Hotspots in the human genome are detected as candidates for positive selection but a bias is observed in the genes within the hotspots. The bias is such that there is consistent change of AT to GC (Galtier et al, 2009). This bias was observed in all regions of the genes and in non human primates too. Thus positive selection cannot be inferred unless gBGC can also be rejected as the cause of these signatures in human genes.

5.2. The HLA system

The HLA gene system is another gene discovered via the candidate gene approach. The HSA locus on chromosome 8 was important in susceptibility to HIV. The allele A of this marker SNP had a higher association to susceptibility of HIV in addition in HIV positive people the allele is associated with higher viral load and faster progression to AIDS (Loeuillet et al, 2008). The HIV viral load of an individual determines how long individuals will progress to AIDS (Fellay et al, 2007). A GWS identified variants accountable for the difference in viral load between individuals (Fellay et al, 2007). The association study showed that polymorphisms in HLA loci were the major determinant for viral load variation (Fellay et al, 2007). The polymorphism of in the HLA complex P5 accounts for 9.6 % of viral load variation in the GWS. This polymorphism is in high Linkage Disequilibrium with HLA-B * 57 polymorphism which to be overly expressed in infected individuals with low viral loads. Other studies show HIV mutations that allow the virus to escapes cytotoxic T lymphocyte killing are linked to certain HLA alleles in the population. This immune escape shapes viral evolution through imprinting of a specific HLA genotype. The presence of other escape mutants in other studies too, such as HLA-B *57 and HLA-B*51 (in the RT codon) confirms

that the active CTL epitopes will not disappear completely from a population (Brumme et al, 2007). Virologic escape is possible from HIV-1 infected individuals who possess HLA-B * 57 mutation (Bailey et al, 2007). Nonetheless, this may not explain virologic suppression in these individuals because escape mutants are known to have low viral fitness and thus may also lead to decreased viral load. In this specific case this patient had a mutation in gag epitopes of the CD +8 cells, drug resistant mutations (M184V & T215Y) in reverse transcriptase and polymorphism in VPU gene which is responsible for the release of virions from an infected cell (Bailey et al,2007). When these mutations were present the patient had a higher viral load. Reversion of the mutations to the consensus sequence did not increase viral fitness either. Although it has not been shown implicitly, the reversion is more likely the cause of the virologic escape of the patient and thus the consequence of the patient's long-term survival. HLA-B 27 is another escape mutant that has been associated with long term non-progression in infected individuals. It is hypothesized that its activity is due to the presentation of a conserved Gag epitope on its surface (Brumme et al, 2007). These active CTL epitopes intensify the diversity of the circulating viral strains in a population. HLA alleles associated with delayed progression to Aids shows that they have a preference for HIV-1 p24 Gag protein.

6. siRNA technology

Technology is rapidly advancing thus providing much stronger and novel ways to detect the involvement of genes in disease. One such technique is the use of gene knockdowns to discover the gene function in an attempt to change disease genotypes and ultimately control disease phenotypes. Viral proteins are dependant on the use of the host machinery to ensure fully functional virions are produced and go on to continue the infection. Host proteins that play a role in HIV infection were detected by using small interfering RNA (siRNA) molecules (Brass et al, 2008). RNA molecules able to regulate gene expression are known as RNA

interfering molecules. RNA molecules which regulate gene expression fall into two categories; long ds RNA and small hairpins (Fire et al, 1998). The siRNA are generated from dsRNA which may be viral in origin or can be artificially introduced into the cell (Kanzaki et al, 2008). Micro RNA (miRNA) on the other hand is generated from within the cell from the cleaving of short hairpin structures. siRNA will induce the specific knockdown of the gene of interest and to facilitate this, the sequence of the gene of interest must be known for the siRNA to pair with the gene and induce the knockdown (Kanzaki et al, 2008). The earliest known involvement of RNAi in HIV pathogenesis described latency which is a very important contributing factor to the success of antiviral therapy. HIV -1 encodes its own siRNA. In addition to this it also encodes a suppressor protein tat which blocks the action of Dicer (RNA processing enzyme) (Bennasser et al, 2005). siRNA act on reverse transcription by degrading RNA transcribed from the proviral DNA (Gao et al, 2008). This degradation is only effective before the newly synthesized RNA is encapsidated. This encapsidation has evolved to protect degradation of the RNA. In addition to this HIV-1 may have evolved an added escape mechanism during its evolution to the host. The complementarity between the siRNA and its target must be perfect. HIV-1 averts the complementarity by the introduction of mutations, which may alter target sequence thereby reducing the efficacy of this approach (Kanzaki et al, 2008). Despite this RNAi therapy provides a novel way of reducing infection. Targeting highly conserved HIV-1 genes may be an interesting target for siRNA because the mutation rates may be lowered.

7. Genome Wide Association studies for the detection of HIV genes

Genome Wide Screens are increasing in popularity for the detection of genes involved in complex diseases such as HIV. Genome wide screens or association studies are particularly useful because they permit the examination of the entire genome across many unrelated individuals without any knowledge of any gene

functions or associations (Loeuillet et al, 2008). Direct genome wide association studies detect an increased number of particular functional variant in affected and unaffected individuals (Pearson & Manolio, 2008). Indirect genome wide association relies on Linkage Disequilibrium. Genome Wide studies are important in determining disease pathogenesis, which will ultimately lead to the production of effective therapies. They are advantageous for the identification of genes that exert small effects. These small cumulative effects are thought to underlie complex diseases (Lazarus et al, 2002). This approach has significant advantages over other approaches such as candidate gene studies. No prior knowledge regarding the gene function of the variants is required thus no assumptions are necessary regarding the type of variant involved (Hutcheson et al, 2007). In addition to this the GWS are particularly useful as the variants do not need to be localized within a genomic region. Advances such as low cost of genotyping a large variety of SNPs in a large population makes this ideal for high throughput genotyping. However there are disadvantages. There is a bias based on the selection of participants for the association study. The relative risk is also an estimate based on the population stratification which can lower detection power of genotyping (Hirschhorn & Daly, 2005). A major shortcoming of this type of study is its potential for false-positives and negatives because of the large number of statistical tests that are performed (Hirschhorn & Daly, 2005). In any event this type of analysis is critical to initial discovery of genetic variants related to common diseases as well as Quantitative Trait Loci (QTL) (Hirschhorn & Daly, 2005). Numerous steps such as careful selection of study participants, inclusion of certain statistical tests have been implemented in these studies to ensure true association of disease and reduce rate of falsehood.

8. Human Population Diversity

GWS have been used to reconstruct the evolutionary history of human populations. This has led to the completion of sequencing of the human genome in

2003 which provides a diverse map of human population evolution.

The GWS have numerous applications but the most versatile is the use of SNPs to study diseases. Single Nucleotide Polymorphisms (SNPs) within human genome (Hutchenson et al, 2007). SNPs occur at one out of every thousand base pairs (Syvanen, 2001) but recent evidence shows that this frequency is much higher, between 0.03 % to 0.01 % in certain blocks of genes (Sabeti et al, 2007). The consequence of detecting and characterizing SNPs are varied. The variations can result in a change of phenotype; it can also change the susceptibility of certain individuals to other diseases. SNP is in the regulatory or coding region result in synonymous or non-synonymous changes (Syvanen, 2001). Synonymous changes result in a 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.

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. Bearing this in mind it leads to the conclusion that these genes are in linkage equilibrium. Though, when genes are inherited more often than chance then these genes are said to be in linkage disequilibrium (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. This Linkage Disequilibrium (LD) has an important part to play in genetic studies because it can be used to infer the role of SNPs in disease.

That being said these GWS had revealed European populations had linkage disequilibrium extending 60kb from common alleles (Reich et al, 2001). This, however, is the subject of much debate. The LD in Europeans extends over a longer distance compared to Africans (Reich et al, 2001, Jakobson et al, 2008).

Variation and demography play a vital role in linkage disequilibrium. SNPS within 10Kb of each other are in strong linkage disequilibrium (Goldstein & Weale, 2001). A study of the linkage disequilibrium within these two populations showed that in a European population the linkage disequilibrium extends over a greater distance. In the Nigerian population the linkage disequilibrium is much less. This is directly related to the events of population structure. This conclusion supports the idea that ancestral European populations stem from Africa and that. The populations have undergone genetic drift when they migrated out of Africa. In so doing the ancestral haplotypes were limited and have consequently given rise to all the present day haplotypes seen in European populations (Jakobson et al, 2008).

In contrast to the rest of the world, Africa has the largest genetic diversity both within and among populations due to the historically high population size. The analysis of autosomal markers, the Y-chromosome and mitochondrial DNA among Africans, European and Asian populations showed the Fst (measure of population co-ancestry) is the highest in Africans (Jorder et al, 2000). Consequently, there are a greater number of variable genes and alleles in Africans (Tishkoff & Williams, 2002 and 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). It is estimated that Africa has approximately 2000 ethnically diverse groups. Thus one can see the shortfall of candidate gene studies is that not all these populations have been characterized therefore the genetic diversity may be under represented.

Therefore applying GWS to detect the association of SNPs in HIV 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 2005). 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, 2001). 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. The Zulu, Xhosa, Pedi, Sotho, Tswana and Venda groups are linguistically similar in origin and are seen to cluster together in the constructed phylogeny. The measure of population structure, 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. Thus because of the large genetic diversity there may be a greater number of restrictive HIV alleles and these will be able to be located with precision because of the low LD in African populations. As mentioned before the selection of the study participants thus the population structure is critical in determining the success of the association study. Population history is very important in dissecting disease causing variants and the issue of racially biased genes in disease has compounded the quest for answers. There has been some favour given to the theory that certain traits are racially biased (Mountain & 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. Nevertheless, the manifestation of certain degrees of the same condition in different ethnic populations does make this assumption attractive.

The human population adaptation to the virus shapes the evolution of the genome therefore it is important to dissect the structure and evolution of specific

populations. This has become increasing important when looking into the evolutionary forces that HIV exerts in populations. Infectious agents like HIV elicit different responses in different individuals leading to a difference in response to pathogenic agent. (Cooke & Hill, 2001). In the study of malarial infection in different West African populations, one group had increased immunity to the disease. Scientific evidence reveals a similar situation with individuals infected with HIV. A relatively small percentage of individuals are resistant to initial infection from HIV even though they form part of a high-risk group. Other individuals are not resistant to initial infection but the rate of progression to AIDS is delayed. These long-term non-progressors remain asymptomatic for 15 years or longer. Other HIV positive individuals develop AIDS in as little as 3 years (Carrington et al, 2001). Thus, the life expectancy of patients is dependant on numerous factors such as genetic make-up, immune response to HIV, CD4 + cells in the blood and the pathogenicity of the infecting virus.

The innate cellular defence system is crucial in detecting and limiting infection. *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 was overcome by viral protein Vif (Sheehy et al, 2000). *APOBEC3G* can function independently of its enzymatic activity. In essence the enzymatic activity can be dissociated from the other cellular functions of *APOBEC3G*.

9. Genetic organization of *APOBEC3G*

This gene has eight exons, all of which are transcribed, arranged in tandem (Jarmuz et al, 2002). 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 psuedocatalytic 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 psuedocatalytic 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).

Promoter elements serve as recognition sites for the binding of DNA proteins which are known to control gene expression mostly at the level of transcription but also through mRNA processing and translation. Typically, there is a 5'-TATA-3' box at -10 base pairs (bp) and a further promoter element at – 25 bp. In addition to these elements, a third element 5'-CAAT-3' region is located at -80bp (Klugg & Cummings, 1997). *APOBEC3G* does not contain the elements at -10 bp and -25 bp but the 5'-CAAT-3' region is present at -75 bp. Although *APOBEC3G* does not have the typical promoter elements it has other distinct elements located upstream of the open reading frame directing the transcription and translation of the exons. A 1025bp sequence upstream of the transcription initiation site displays constitutive promoter activity (Muckenfuss et al, 2007). The activity of this promoter was dependant on a sequence about 78-87 upstream of transcription initiation (Muckenfuss et al, 2007). It is at this site that transcription factors are bind and induce transcription. This box is termed the GC box.

10. Evolution of the gene family

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 some sequences of nucleotides termed the gaps from *E.coli* cytidine deaminase (ECCDA) the signature sequence of ***A****POBEC1* 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 proteins' 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 focused on the deaminases that had the first cluster of the active site which contained 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 (Conticello et al, 2005). *AID/ APOBEC1/APOBEC3* were clustered together and have diverged from *APOBEC2. APOBEC2* and *AID* 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). There are nine paralogs of *APOBEC3G* in the humans.

It has been proposed that *APOBEC3* zinc domains were the result of the diversification of two ancestral domains that either constituted a double-domained protein or a single domained protein such as *APOBEC3 A* and *H* (Conticello et al, 2005).

11. Deaminase independent activity of *APOBEC3G*

APOBEC3G antiviral activity is not correlated with its deamination ability. *APOBEC3G* has an effect at every stage of viral replication. On entry into the cell the *APOBEC3G* physically blocks the reverse transcriptase from moving along the template (Chiu & Greene, 2008). This inhibition is not always complete and rather decreases the production of reverse transcripts. *APOBEC3G* may be expressed in two forms (Goila-Gaur and Strebel, 2008). It is packaged into high molecular weight ribonucleotide complexes (RNP) in resting CD4 cells in the lymphoid tissues by cytokines (Gallios-Montbrun et al, 2006), localized within the cytosol of the cells. When *APOBEC3G* is associated in this form it cannot block integration and reverse transcription of the virus. In contrast, resting CD4 cells in the peripheral blood are not permissive to HIV infection because *APOBEC3G* is packaged into low molecular weight ribonucleotide complexes (Goila-Gaur and Strebel, 2008) which block integration and reverse transcription of the virus, independent of the editing mechanism (Gallios-Montbrun et al, 2006). In essence the enzymatic activity can be dissociated from the other cellular functions of *APOBEC3G*

12. Deaminase dependant activity of *APOBEC3G*

Inhibition of viral reverse transcription is not completely negated via the independent pathway and viral DNA does enter the cell and is transcribed (Chiu & Greene, 2008). In so doing a viral minus strand is generated and deamination occurs. Deamination is the conversion of cytidine in RNA to a uridine of the first strand reverse transcripts in target cells. This results in G-to-A hypermutation in the coding strand and is associated with premature DNA degradation. The deamination of the cytosine will result in one normal C: G pair and the mutated pair (U: A) (Sousa et al, 2007). The mutated pair will be fixed in subsequent generations as T: A. Viruses are sensitive to the incorporation of uracil and have

subsequently developed a mechanism to ensure the removal of the uracil and the propagation of the virus. After the incorporation of the uracil, it may also be removed by uracil DNA glycosylases, leading to the cleavage of the sites by an endonuclease and subsequent viral degradation as a result of DNA fragmentation. cDNA complementation does occur after deamination but that the synthesized DNA is degraded before integration into the host genome (Mangeat et al, 2003). Thus the G to A mutation functions to curb the spread of the virus (Harris et al, 2003). Mutations are produced at dC dinucleotides or pools (Harris et al, 2003). Thus they occur at hotspots which are preceded by 5'pyrimidine-dC consensus. Hypermutation may function to diversify expressed sequences. Hypermutation is common in mammals and plays a vital role in normal processes within the body such as antibody diversification in B lymphocytes. This mechanism of C-to-U substitution editing mechanism is similar to the editing of apo mRNA editing and the Apobec-like protein editing.

13. 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.

In permissive cells infected with vif defective HIV, *APOBEC3G* is packaged into the virions. The virions on contact with a new cell infect the cell and the minus strand cDNA is used to synthesize DNA. The encapsidated *APOBEC3G* then deaminates the newly synthesized DNA at specific cytosine residues to uracil. The unstable DNA fails to integrate into the host genome to form a provirus and it is consequently degraded. Wild type virus that infects non permissive cells ensures that vif binds to *APOBEC3G,* blocking uncoating of the virus and thus preventing deamination.

The structure of human *APOBEC3G* was discerned by modelling it on the outer monomer of Apobec 2. This model fitted the predicted secondary structure. In addition, this model allowed confirmation of the residues important in the functioning of the *APOBEC3G*. Specifically, the D128 residue, previously found to be important in determining species specificity in vif mediated inhibition of *APOBEC3G*. Comparative sequence data from non human primates such as Old world monkeys (OWM) New world monkeys (NWM) and hominids indicate that *APOBEC3G* was under positive selection pressure in primates for at least 33 million years. This positive selection pressure is the driving force for the fixation of variants with altered proteins. This ultimately affects how the gene variants interact with one another. Interestingly the selection pressure appears to be ancient (Sawyer et al, 2004). *APOBEC3G* in Old world monkeys and hominids appears to have diverged from each other 23 million years ago (Sawyer et al, 2004). The new world monkeys, old world monkeys and hominids shared a common ancestor 33 million years ago (Sawyer et al, 2004). 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. Residue D128 of *APOBEC3G* contributes a negative charge to the cluster under positive selection. The 3D modelling confirmed what Huthoff and Malim (2005) postulated that the position of this residue allows the direct interaction of *APOBEC3G* with vif. Two functionally important residues emerge R122 and W127. Mutations in these residues show that these mutants failed to inhibit HIV-1 in vif defective virions as *APOBEC3G* is not packaged into these virions (Zhang et al, 2007). Residues 124-127 are aromatic and have been implicated in having an important role in the incorporation of *APOBEC3G* into the virion (Huthoff and Malim, 2007). It is thought that the positive selection of APOBEC3G is favored by the changes in charge of the amino acid. This is supported by the fact that the Asp 128 is conserved in hominids and Lys 128 is conserved in OWM. When the aspartic acid in the human *APOBEC3G* is replaced with the lysine that is found naturally in African Green Monkey *APOBEC3G*, it becomes resistant to HIV-1 vif but not

SIV vif (Zhang& Webb, 2004). Having said this it is very hard to ascertain the exact driving evolutionary force acting on this gene. The gBGC is unlikely to be the explanation for the high number of non-synonymous polymorphisms (the hallmark of positive selection) in *APOBEC3G* because it favours the fixation of advantageous GC to AT mutations. In contrast *APOBEC3G* functions by converting uridines to cytosines ultimately leading to the production of G to A mutations in the plus strand of the viral DNA. It is possible to conclude that the interaction between *APOBEC3G* and vif is the driving force for this selection. They act in an antagonistic manner with each other and so it is reasonable to assume that *APOBEC3G* evolution should be considered in conjunction with that of vif.

14. Interaction of *APOBEC3G* and Vif

Vif has a three fold action on *APOBEC3G*. It can bind to it and target it for degradation by proteosomes. It can mediate its destruction via polyubiquitination and 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 polyubiquitination. Ubiquitination is proteolytic modification of proteins with ubiquitin. E1, E2, E3 enzymes are involved in the covalent conjugation of ubiquitin to a substrate consequently ensuring that the protein is degraded by the proteosome. Classically ubiquitination is divided into three stages: Activation, Conjugation, and Ligation. Initially there is activation of ubiquitin with ATP by ubiquitin-activating enzyme E1. Following this, the activated ubiquitin is conjugated to E2 via a thioester bond. E2 acts with E3 to transfer the ubiquitin to the target. An isopeptide 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 are of ubiquitination 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. 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* decreased infectivity of the virus and Cul 5 inhibited ubiquitination of *APOBEC3G*. Vif interacts with this complex via a conserved motif SLQXLA because mutations in this motif drastically reduced the interaction between Vif and SCF-like complex (Yu et al, 2003, Mehle et al, 2006). Although mutations in this motif reduced interaction with the complex, it did not effect the interaction of Vif with *APOBEC3G*. This is direct evidence that Vif activity alone is not enough to overcome the antiviral action of *APOBEC3G*. However, the cellular target of the SCF-like complexes remains elusive. Proteosomes have also been implicated in Vif functioning. When a proteosome inhibitor is added to non-permissive cells Vif failed to exclude APOBEC3G from the resulting progeny and the viruses were 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 as an E3 ligase. They are said to be an E3 ligase as they have a similar structure to the SCF complexes. Rbx1 was an important component of the complex and interacted intimately with Cul 5 as seen in mutagenesis studies. Cul 5 mutants with no Rbx1 binding affected 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 decreased 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, 2004).

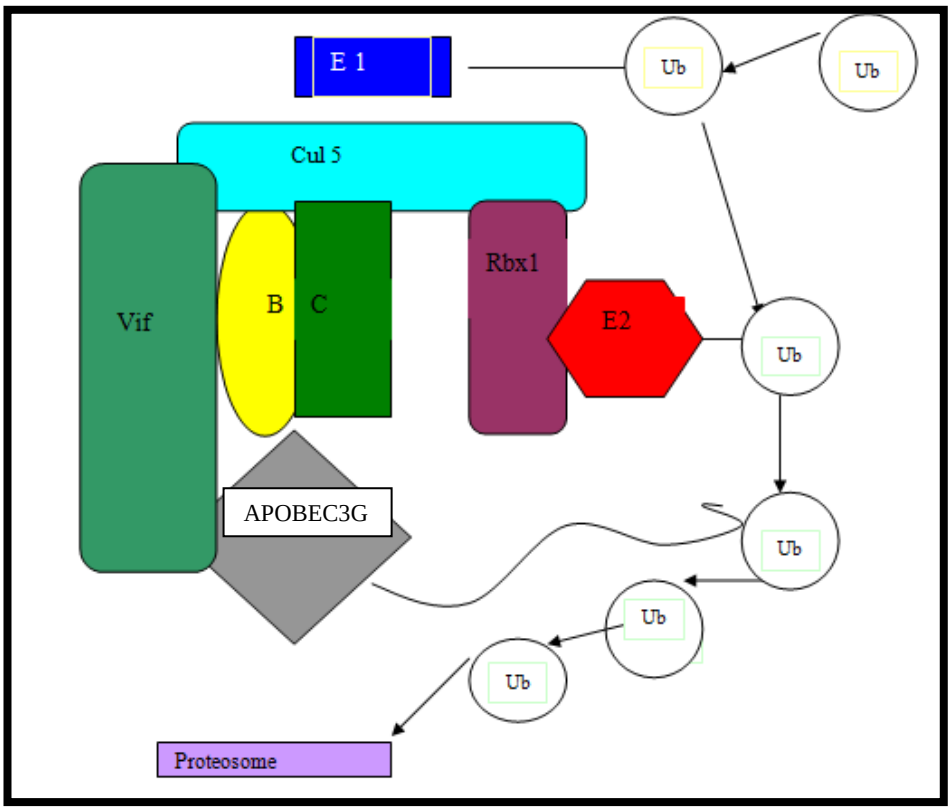


Figure 2. The interaction of Vif and APOBEC3G in non-permissive cells.

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 the dimerization of vif is localized to the SQL motif with Vif. 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 the newly produced virions and subsequently HIV infectivity was decreased. 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. *APOBEC3G* variants have a profound influence on the progression to AIDS in a cohort of seroconverter, seropositive, seronegative participants of European and African American descent. 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 haplotypes 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 in amino acid position 186 was higher African Americans than European Americans. In particular the 186R allele was associated with faster progression to AIDS in the African Americans (An et al, 2004).The frequency was 37 % in African Americans as compared to 2, 9 % in European Americans (An et al, 2004). However, these observations have not being independently confirmed in other studies. A recent study investigated this polymorphism in Indians. The 186 R allelic variant could not be found in the study sample and thus not conclusions could be drawn about its influence on disease progression in this study population (Rathore et al, 2008).

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).

15. Aim

The aim of the project was to describe variation in *APOBEC3G*. The objectives included the characterization of variation of *APOBEC3G* in Bantu-speaking South Africans with particular emphasis on the non-coding region of *APOBEC3G*. 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.

16. Objectives

1) Analyze sequence data from sequences received in Honours.
2) Develop genotyping assays for -571 SNP in the upstream non-coding region and H186R in the coding region of *APOBEC 3G.*
3) To determine allele and genotype frequencies of the two SNPs.
4) To determine linkage disequilibrium of the SNP data and to infer haplotypes.

Materials and Methods

1. Sampling

1.1.Sample collection

The search for variation in *APOBEC3G* was carried out using blood samples from Bantu-speaking South Africans. Samples were collected from patients at The Infectious Disease Clinic at Johannesburg General Hospital and from the Themba Lethu Clinic at Helen Joseph Hospital. A subset of 45 samples was collected from a general population of Bantu-speaking South Africans surrounding University of the Witwatersrand. Data was collected on geographic origin of participants, their parents and grandparents and in addition, specifics such as age, recent $CD4^+$ T cell count, date of first infection and secondary illnesses was also gathered from the participants. Ethics clearance has been obtained from the Human Research Ethics Committee. The clearance number is M040221 (Appendix).

As the origin of most South Africans stems from the Bantu expansion 2000 years ago, the Bantu-speaking subpopulations found today are still genetically similar (Lane et al, 2002). Within the samples it was found that in 57 % there was one language present over three generations, 31 % of the recent ancestors were mixed with respect to language. In 7.4 % of the sample population the language across all three generations was incomplete. 3.7 % of the sample was mixed with respect to ethnicity. Within the sample group that has a single language across all three generations, Zulu is the most frequently sampled at 45.4 %. At 15.6 % the Xhosa ethnicity is the second frequent within this grouping, this is followed closely by Tswana and Sotho languages both occurring at a frequency of 11.7 %. The Pedi language group accounts for 6.5 % of the sample in this subgroup, with Venda, Ndebele and Tsonga representing 3.9 %, 2.6 % and 2.6 % respectively. Therefore the Bantu-speaking Johannesburg population is representative of the major Bantu-speaking ethnic groups in South Africa. Thus, the sample is adequate as it provides an estimation of variation present in HIV/AIDS patients in the province.

1.2. Privacy and confidentiality

Privacy and confidentiality was observed throughout the sample collection process. Discussions with patients took place in a vacant doctor's room in the clinic. To ensure confidentiality numerous steps will be taken:

*Codes were used instead of participants' names

*Information shared with the investigator was not disclosed with a third party.

*Information was sealed at all times in a locked cabinet.

1.3. Informed consent

Participants were required to fill in consent forms (Appendix). It was instrumental that the participants gave voluntary informed consent. To ensure this numerous steps were taken:

*An explanation of the project and the consent form was undertaken in a language preferred by the participant.

*Written and oral explanations were provided in a manner that was easily understandable to the participant.

*Participants were given the opportunity to ask questions and only then did they sign the consent forms.

1.4. DNA extraction

Genomic DNA extraction was performed at Medical School and at the P2 facility in Molecular and Cell Biology using the Qiagen DNA Blood Mini kit. The Vacutainer® tubes used to collect and store the blood had EDTA, which prevents the blood from clotting. Genomic DNA was extracted from the white blood cells in the buffy coat, which is the middle layer when blood that has been centrifuged for 10 minutes at 13.4 rpm. Initially the blood was lysed with an ethanol-containing buffer. This was followed by the addition of protease and RNase. The

protease degraded any proteins and the RNase eradicated any traces of HIV virus. The solution was then added to spin columns. This facilitated the binding and elution of the DNA from silica particles. Washing with an ethanol-containing buffer ensured that any impurities such as divalent ions and proteins were removed so that pure DNA could be eluted from the membrane. DNA was eluted with another buffer and was concentrated by using a smaller volume of buffer. This DNA was used to detect variation in the *APOBEC3G* gene.

2. Analysis of upstream non-coding region sequences

2.1 Analysis of sequence data

In my BSc (Honours) project of 2003 I sequenced the upstream non coding region of *APOBEC3G*. This variation was detected by direct sequencing and analysed using Sequencher 4.0. The analysis was very rudimentary. Hence as more insight was gained regarding the gene and software it became evident that these sequences need to be re-analyzed now.

2.2 Promoter analysis

All eukaryotes contain promoter elements upstream of the first open reading frame. Promoter elements serve as recognition sites for the binding of DNA proteins which are known to control gene expression mostly at the level of transcription but also through mRNA processing and translation (Muckenfuss et al, 2007).

Software such as FPROM, TSSG, TSSW were employed to search for promoter elements using the sequenced data from 2003. FPROM, TSSG, TSSW allows for the recognition of human POLII promoter region and the start of transcription. The algorithm is based on the recognition of functional motifs as well as nucleotide sequence of genes (Solovyev and Salamov, 1997). The interesting

point is these programs were modeled using microarrays to search for functional enrichment of transcription binding sites (Solovyev and Salamov, 1997). The output file included the number and position of the transcription start sites, their relative weights and the TATA box position (if any). In addition TSSG and TSSW predicted the transcription factor binding sites for the Transcription Start Sites (TSS).

3. The genotyping of position -571

3.1 Detection of SNP -571 using Allele Specific Amplification

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 use of thermo stable DNA polymerases and automation of the method increases its efficiency. As a result many PCR applications have been developed. These include screening and sequencing of inserts from phages and bacteria and the visualization of single-copy genes. There are three main steps in PCR, denaturation, annealing and extension. Initially double stranded template DNA is denatured. This is termed a hot start. This is then followed by a ramp down to the annealing temperature where primers anneal to their target sites. Lastly extension of the primers, with nucleotides, by DNA polymerase is performed. Primer design is crucial in PCR and numerous things have to be taken into consideration in order to yield successful results. Primers should have GC content between 40 % - 60 %. The higher the GC content the more stable the primers. The primers should have a length of between 18 nucleotides and 30 nucleotides. The primers should not be self complementary or complementary to each other. If primers are self-complementary they will fold back and bind on themselves. If complementary to each other they will bind to each other and not to the annealing sites on the target molecules resulting in inefficient or no amplification.

The *APOBEC3G* gene was accessed using Pubmed and the accession number (NT011520). To detect variation of the -571 SNP Allele Specific Amplification was used. Allele Specific Amplification was used as an alternative to direct sequencing because it is applicable to a large number of samples and is therefore inexpensive. In Allele Specific Amplification (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' prime end. Homozygotes will only yield PCR product with either primer. Heterozygotes will yield a product with both primers (Okayama et al, 1989).

The 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).

The cyclic conditions for -571G PCR is denaturation at 94°C for 2 min, this is followed by 35 cycles of denaturation at 94°C for 30s, annealing temperature at 63°C for 25 s, extension at 72 °C for 18 s, the final extension is at 72°C for 5 minutes. The cyclic conditions for -571C PCR is denaturation at 94°C for 2 min, this is 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 minutes. The PCR products were then visualized on a 1% agarose gel stained with ethidium bromide. The gel was run at a constant voltage of 70V for 1 hour. Thereafter the agarose gel was visualized on the UV transilluminator at wavelength 254 nm.

Table 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’ TGAAGCCTCACTTCAGGTACC GCTGC	-571G	63 C	850 bp
-571 C PCR Forward	5’GCGCGTCTCACAGCTCCCTTCCCG 3’ AGTTCACAGGGGTCACAATGGCT	-571C	63.5 C	450 bp

3.2 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. The PCR master mix reaction is the same as for conventional PCR. The difference is that the PCR heat block where the PCR tubes are placed is divided into different temperature zones which are selected by the user. The PCR is setup for one sample for each temperature zone tested. Essentially the same sample can be amplified using different annealing temperatures at once there by reducing time.

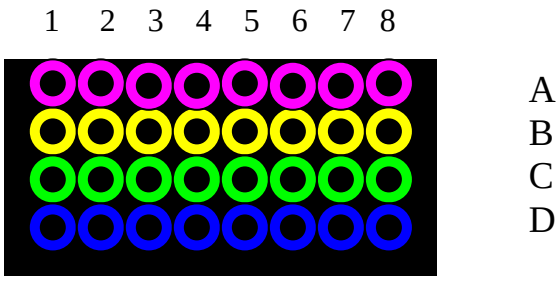


Figure 3: PCR Heat Block. Letters A, B, C, D represent the different temperature zones. Row A would have one temperature; row B another temperature and so on. Eight samples can be loaded in each row.

The 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).

The cyclic conditions for the gradient PCR is denaturation at 94°C for 2 min, this is followed by 35 cycles of denaturation at 94°C for 30s, annealing temperature of between 60°C-58°C for 30 s, extension at 72°C for 30 s, the final extension is at 72°C for 5 min. The PCR products were then visualized on a 1% agarose gel stained with ethidium bromide. The gel was run at a constant voltage of 70V for 45 minutes. Thereafter the agarose gel was visualized on the UV transilluminator at 254 nm.

Once the gradient PCR established the optimal annealing temperature, this temperature was used to amplify the subsequent samples using the PCR reaction mixture as above and then visualize after being run on a 1% agarose gel at 70 V for 1 hour. Successful samples are then subjected to restriction digestion to 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 for this sample was finished.

The RFLP technique was also used to detect variants of SNP -571. RFLP is a difference in the DNA sequence of a genome which when cut with a restriction enzyme will yield fragments of differing size which are separated and visualized on an agarose gel. This method can be used to determine genotypes because each combination of alleles will produce a predicted pattern of fragments. RFLP is especially helpful in detecting known SNPs in the genome in large sample set and thus also inexpensive when compared to direct sequencing.

The reaction mixture consisted of 10 X Buffer R with BSA, 2.5 µ/ µl MvaI, 6 µl 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 ethidium bromide for 2 hours at 70V. 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 4).

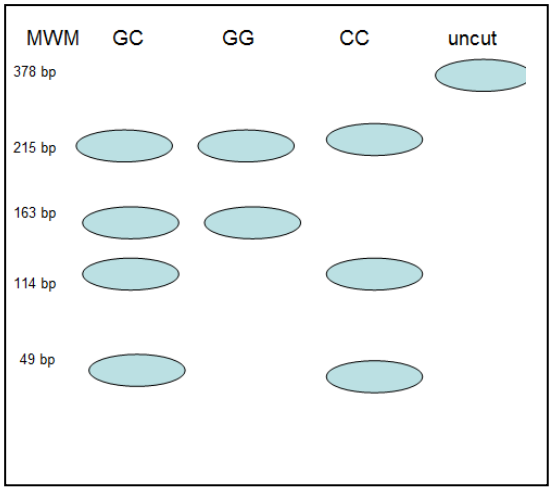


Figure 4. Schematic of three genotypes of position -571 of the upstream non-coding region after digestion by restriction enzyme MvaI.

4. Detection of variation in exon 4 using Pyrosequencing

Pyrosequencing technology is based on sequencing. It is based on a 4-enzyme real-time monitoring of DNA by bioluminescence. Essentially there are 4 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. ATP is converted to light by luciferase (Ronaghi, 2001). The unincorporated nucleotides and ATP are removed from the reaction by Apyrase (Ronaghi, 2001). 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, 2001). The software produces a theoretical output of all possible genotypes and the pyrograms need to be compared to these to verify the genotypes.

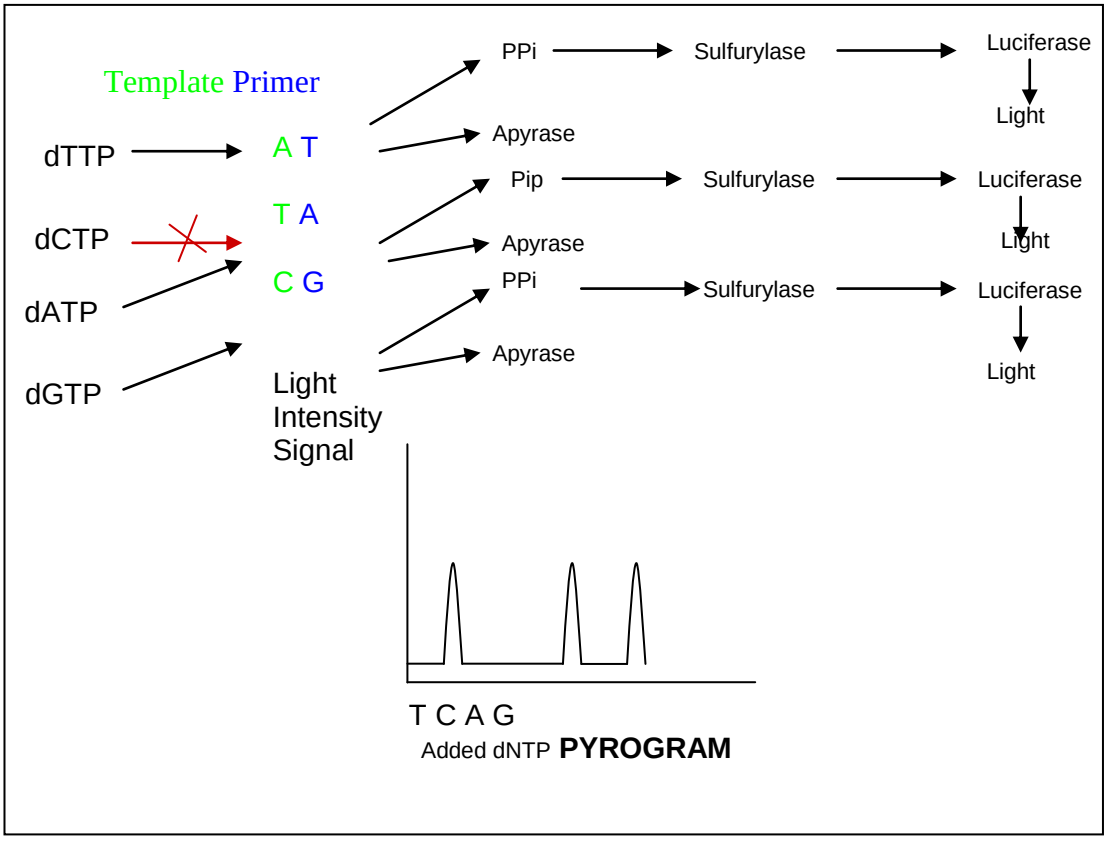


Figure 5. Schematic overview of pyrosequencing system. When the correct dNTP is added it pairs with the template and light is produced. When the dNTP that is added is not complementary with the template, no light is produced.

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*. 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 that were used are APOfor (GACGGGGACACCGCTGATCGTTTAGCAAGTTCGTGTACAGCCAAAGA) 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 labeled

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

After the successful optimization and amplification of the desired region the samples were viewed on a 4 % agarose gel to ensure the correct size and minimal primer dimmers. Because the forward primer was tagged the software assigns the genotype of the anti-sense strand.

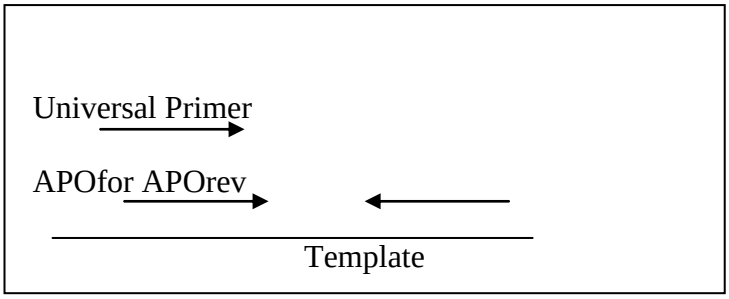


Figure 6. Position of the primers used for pyrosequencing.

The 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).

The cyclic conditions for the PCR was denaturation at 94°C for 2 min, followed by 40 cycles of denaturation at 94°C for 30s, annealing temperature of 58°C for 30 s, extension at 72°C for 30 s, the final extension was at 72°C for 5 min.

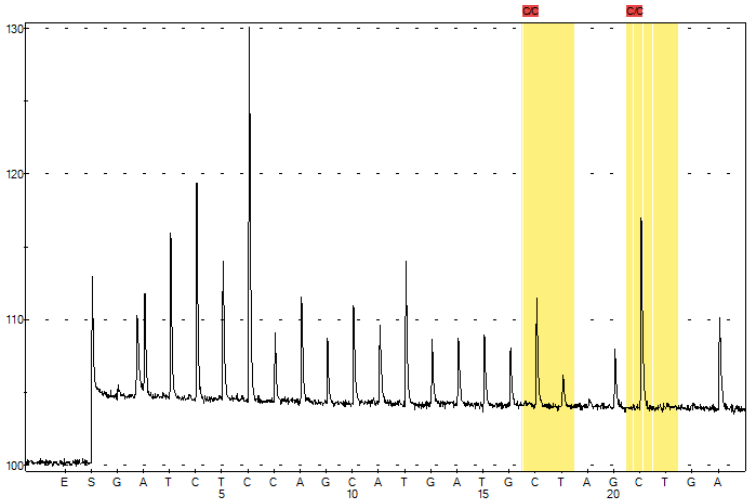


Figure 7. Pyrogram of variation at two positions

The Ensembl database showed that there are two SNPs in this exon which reside next to each other; hence conventional genotyping assays such as allele specific PCR or RLFP may be problematic because of the proximity of the SNPs. 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.

```
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
41701 CCTGCCTCCCACCTGCTTTCCTGGGCCCTTCCTGTGAGTGAGAGGCCCCTTCTGCCTCCA 41760
41761 GAGCAACCTCCATCCACCCCCACAGCCTGGGAGCCCCAACCTGGCCCCTTCCATCTCCCT
```

Figure 8. 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.

5. 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. PCR was used to amplify a 600bp fragment was amplified using the following primers: Codon 4 fw (AAGCTGCATCGTGACCAGGAGTAT) and Aporev (AGAGGAGCGAGGCGATGA). The 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).

The cyclic 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. The PCR products were then visualized on a 1% agarose gel stained with ethidium bromide. The gel was run at a constant voltage of 70V for 1 hour. Thereafter the agarose gel was visualized on the UV transilluminator. Two samples were sequenced in both the forward and reverse directions using sequencing primer ApoSeq1 (AGACCCTCACCTGAGA).

6. Data Analysis

6.1 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

6.2 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 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

$f(AA) = p^2$

$f(Aa) = 2pq$

$f(aa) = q^2$

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.

6.3 Linkage Disequilibrium and Haplotype Analysis

Linkage Disequilibrium describes the non-random assortment of alleles at two or more loci (Devlin & Risch, 1995). 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.

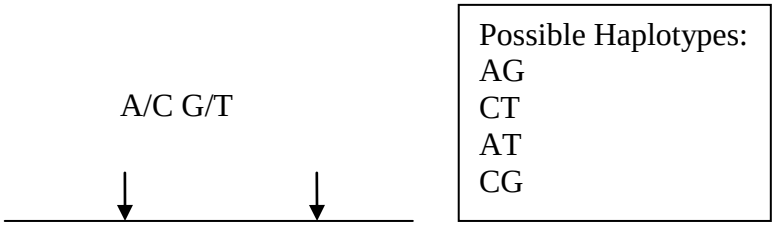


Figure 9. Diagrammatic representation of Haplotypes at two loci

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

Consider two loci each with two alternative alleles on one chromosome

Haplotype Frequency:

$A_1 B_1 \; x_{11}$

$A_1 B_2 \; x_{12}$

$A_2 B_1 \; x_{21}$

$A_2 B_2 \; x_{22}$

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. 136 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.

6.4. Disease Status Association

To determine whether *APOBEC3G* is associated to susceptibility to HIV the sample population was divided into HIV positive and general population. The genotype and allele frequencies were calculated in the sub groups and tested for with the Hardy Weinberg chi squared test.

6.5. Independent Chi-squared test

A co-dominant model was used in this test to ascertain the frequency distribution within the HIV + and general population. The Fishers probability will determine if the distribution is significant.

RESULTS

1. Analysis of upstream non-coding region sequences

1.1 Reanalysis of sequencing data

Reanalysis of sequencing data using Sequencher 4.0 yielded numerous insights into the upstream non-coding region.

In 2003 -91 SNP was found at a position of 91 bases upstream of transcription initiation. After trimming the sequences near the end of the sequences to remove all nonsensical sequence only 6 sequences provided informative data. Four heterozygotes (GC) and 2 CC homozygotes were detected. The allele frequencies could not be calculated for this SNP in this sample group because of the very small sample size.

SNPs at position -163, -166 and -199 were not characterized in 2003 but re-analysis showed it to be well represented in 18 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 (T) 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 (Figure 10). The remaining 13 samples were homozygous for the major allele at all three loci.

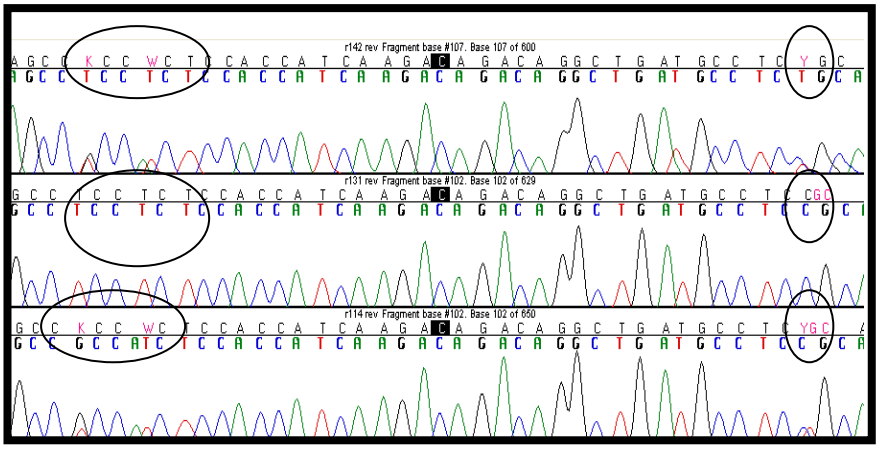
SNP at -286 was not actually a SNP. It is the result of a sequencing artifact which happens when G dNTP is preceded with a T dNTP in direct sequencing.

The SNP at -571 was observed during analysis in 2003 and was still polymorphic. In 20 samples 8 were heterozygous, 11 homozygous for the C allele and 1 was homozygous for the G allele. The frequency of the C allele is 0.75 and the frequency of the G allele is 0.25 in these samples. The population diversity on the dbSNP shows that in Sub-Saharan population the allele frequency of the C and G

alleles to be 0.917 and 0.083 respectively. The difference is the result of the sample size and will be resolved in later discussions.

At position -881 7 heterozygotes and 2 homozygotes for the allele C were observed. Frequency data can not be ascertained from the sequences as the sample size is too small

A



B C

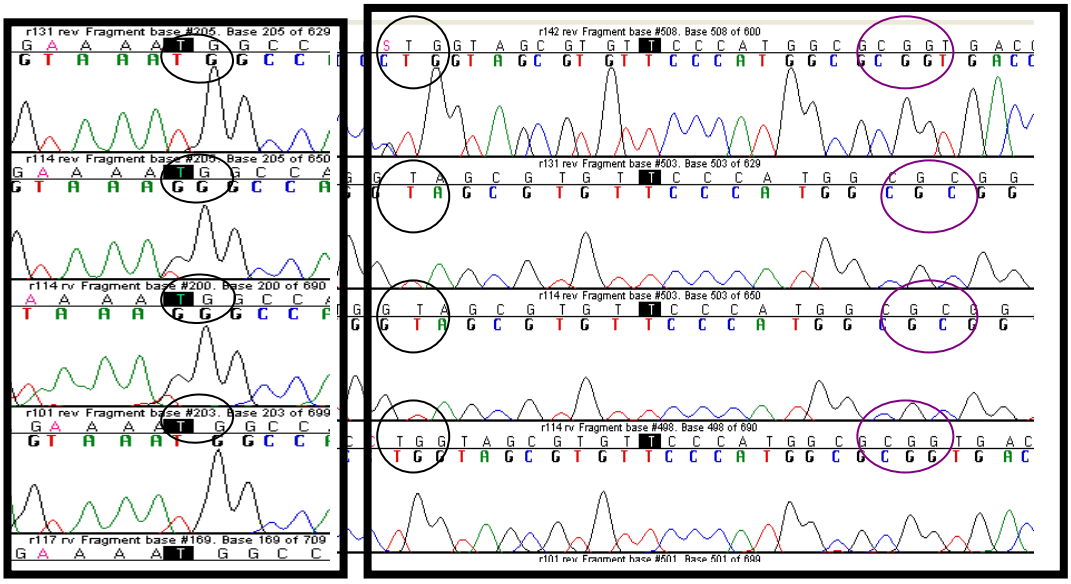


Figure 10. Chromatograms of re-analyzed sequences from Honours project in 2003. A) The chromatograms show samples 114, 142 and 131 at SNP positions -163, -166 and -199. These SNPs appear to be in Linkage Disequilibrium as they are all either heterozygous at all positions in sample 114, 142 or homozygous at all position as in sample 131. B) These chromatograms show the artifact of sequencing which was erroneously interpreted as a SNP in 2003. If look at sample 114 then one can see that at the circled position it seems as if this sample is heterozygous for both alleles. C) Chromatograms showing -571 and -590 loci (purple 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 allele. Thus this is an example of a fixed polymorphism.

1.2 Promoter analysis

Table 1. Genomic locations of predicted Transcription Start Sites and TATA box positions on *APOBEC3G* on chromosome 22 using three different software programs.

	Promoter analysis software		
Transcription Start Site (TSS) on chr 22	**FPROM**	**TSSG**	**TSSW**
	37783673	-	37783676
	-	37790636	37790636
	37791680	-	-
	-	37791823	-
	37792953	37792952	37792948
	37793945	-	-
	37794350	-	-
	37796881	37796880	37796877
	37797902	37797907	37797899
	37800332	-	-
	Promoter analysis software		
TATA box position on chr 22	**FPROM**	**TSSG**	**TSSW**
	37783643	-	37783644
	-	37791323	37791323
	37791637	-	-
	-	37791783	-
	37792822	37792821	37792821
	37793902	-	-
	37794319	-	-
	37796841	37796840	37796840
	37797870	37797869	37797869
	37800289	-	-

The FPROM program predicts more TSS and TATA box positions on Chr 22. The FPROM identified eight TSSs and TATA box motifs. TSSG and TSSW identified five TSSs and TATA box motifs. However the programs predict three of the same TSSs (highlighted in green) and the corresponding TATA box positions- highlighted in blue on chromosome 22.

2) The genotyping of position -571

2.1 Detection of SNP -571 using Allele Specific Amplification

The PCR allowed the SNPS to be detected in a large sample size. The PCR products for each allele of each SNP were run together on the same agarose gel to facilitate the correct genotyping of samples. Individuals homozygous for a SNP will yield a product in one reaction and not the other. While heterozygotes will yield a product of the same intensity in both reactions. The sizes of the products for the different alleles differ allowing accurate genotyping. A total of 165 samples were genotyped for this SNP using ASA.

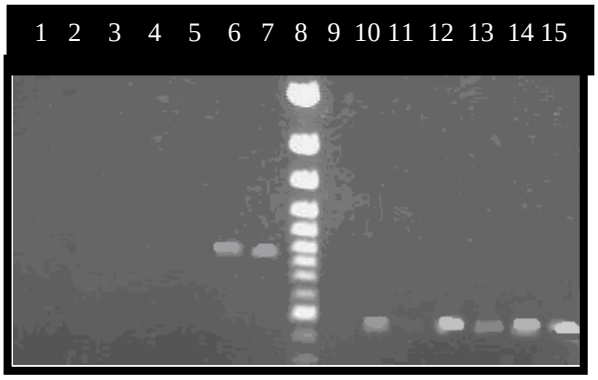


Figure 11. PCR products from -571 SNP were run on a 0.7 % agarose gel. The reactions for each allele were done separately but run on the same gel. This gel sows the -521 G and -571 C reactions respectively. The -571 G reactions produces a band of size 850bp and the -571 C reaction produces a band of size 450 bp. Lanes 1-7 are the -571 G reaction of 7 samples 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, 5, and 7 are genotyped CC while sample 6 is genotyped GC.

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

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. The 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. However, this is very subjective (Figure 13).

The cyclic conditions are described below.

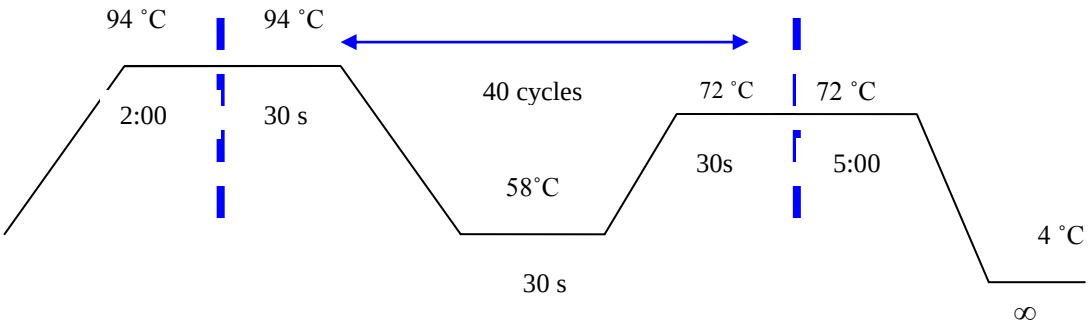


Figure 12. PCR cycling conditions for genotyping -571 SNP.

The amplified samples were digested with MvaI and fragments run on a 4 % agarose gel. The assay identified 88 GG homozygotes, 35 heterozygotes and 1 CC homozygote. The remaining 12 genotypes were discerned from sequence data from 2003. Therefore there were in total 95 GG homozygotes, 39 heterozygotes and 2 CC homozygotes.

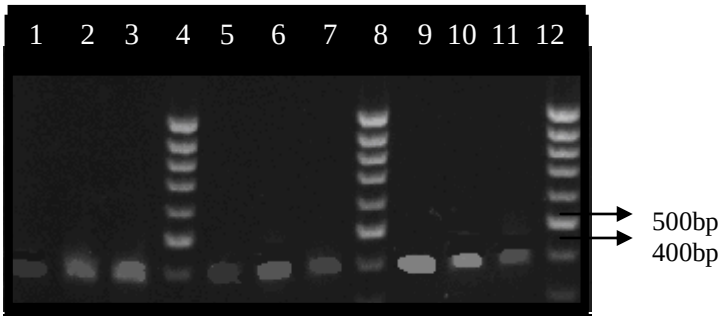


Figure 13. The gradient PCR products (400bp) for three samples 216, 340, 365 that were run on a 1 % agarose gel. Lanes 1, 2, 3 represent sample 216,340, & 365 at 60°C. Lanes 5, 6, & 7 represent samples 216, 340 & 365 at 59.3°C. Lanes 9, 10 & 11 represent samples 216, 340, 365 at 58°C. The annealing temperature of 58°C was chosen as it gave the best amplification in the three samples.

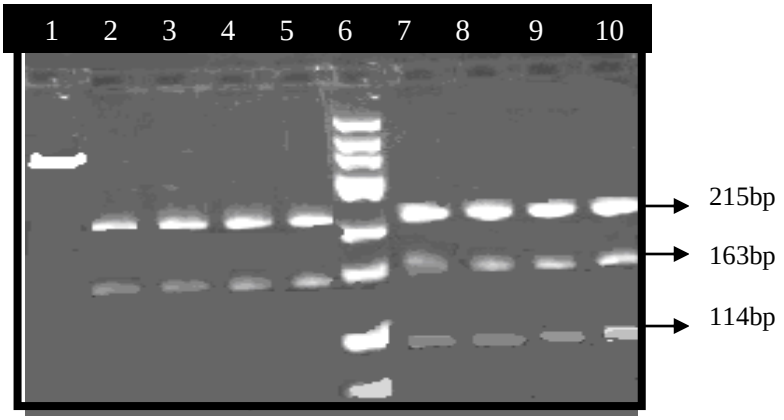


Figure 14 DNA fragments after restriction digestion with MvaI for 4 hours at 37 °C. Lanes 1 represents the uncut control DNA of size ~378 bp. Lanes 2-5 represent homozygotes for the G allele with fragment sizes at 215bp and 163bp. Lane 7-10 represent heterozygotes with fragment sizes at 215bp, 163bp, 114bp and 49bp. The 49bp fragments are not clearly visible on the gel. No homozygotes for the C allele were present.

3. Genotype and Allele Frequencies

Table 2: Genotype and allele frequencies for -571 SNPs detected by Allele specific amplification in subpopulation A of the whole study population and the χ^2 test

	GENOTYPE	N	ALLELE	ALLELE FREQ	GENOTYPE FREQ	χ^2
571	CC	46	C	0.41	0.279	33.363
	CG	44	G	0.59	0.267	
	GG	75			0.455	
		165				

The sample population deviated from Hardy-Weinberg equilibrium at a $p < 0.05$. The GG homozygotes were more frequent than the CC homozygote as in below. The CC and CG genotypes are similar.

Table 3: Genotype and allele frequencies for -571 SNPs detected by RFLP in subpopulation B of the whole study population and the χ^2 test

	GENOTYPE	N	ALLELE	ALLELE FREQ	GENOTYPE FREQ	χ^2
571	CC	2	C	0.16	0.014	1.330
	CG	39	G	0.84	0.287	
	GG	95			0.699	
		136				

The sample population did not deviate from Hardy-Weinberg equilibrium at $p<0.05$ even though the GG homozygotes were more frequent than the CC homozygote. The allele frequencies indicate that the G allele is more frequent in this subpopulation.

4. Detection of H186R using pyrosequencing

The sequencing of a 600bp region of exon 4 showed that the sequence from Ensembl database was indeed correct (Figure 15).Variation was detected in 136 samples in codon 186 within exon 4. 136 samples were sequenced using pyrosequencing (Figures 16).

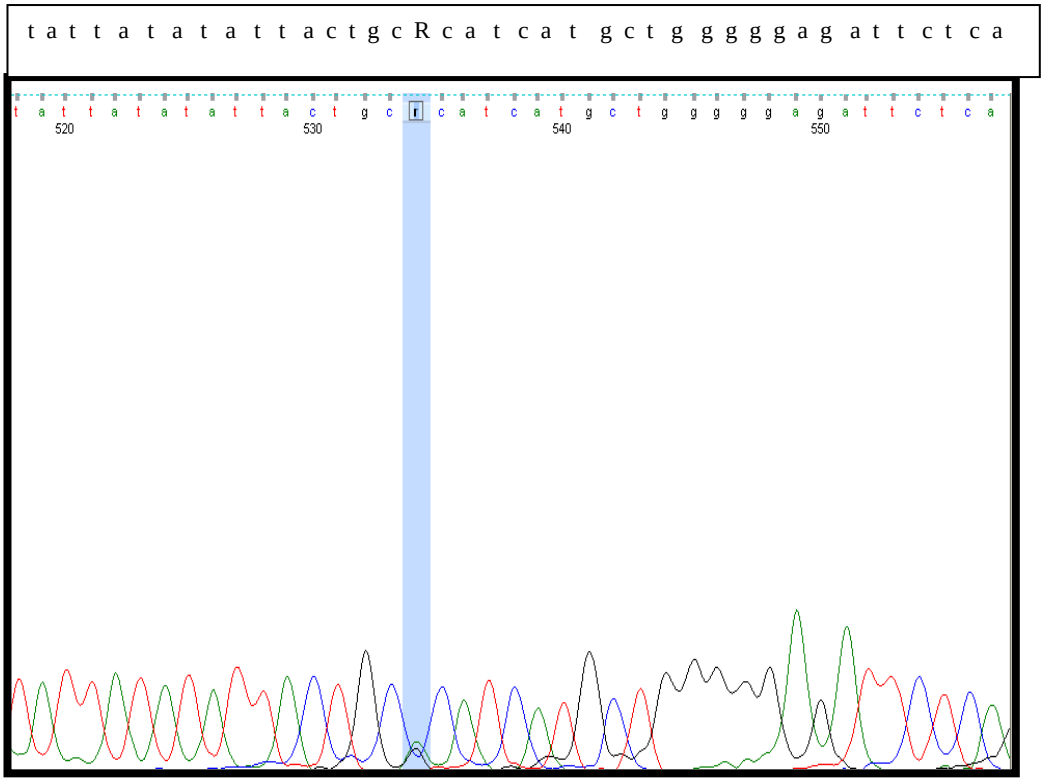


Figure 15. Chromatogram of sample 310 sequence. The highlighted strip shows a heterozygote for this sample at position 186. Codon 185 represented by nucleotides 530, 531, 532 in the chromatogram are not polymorphic in this sample.

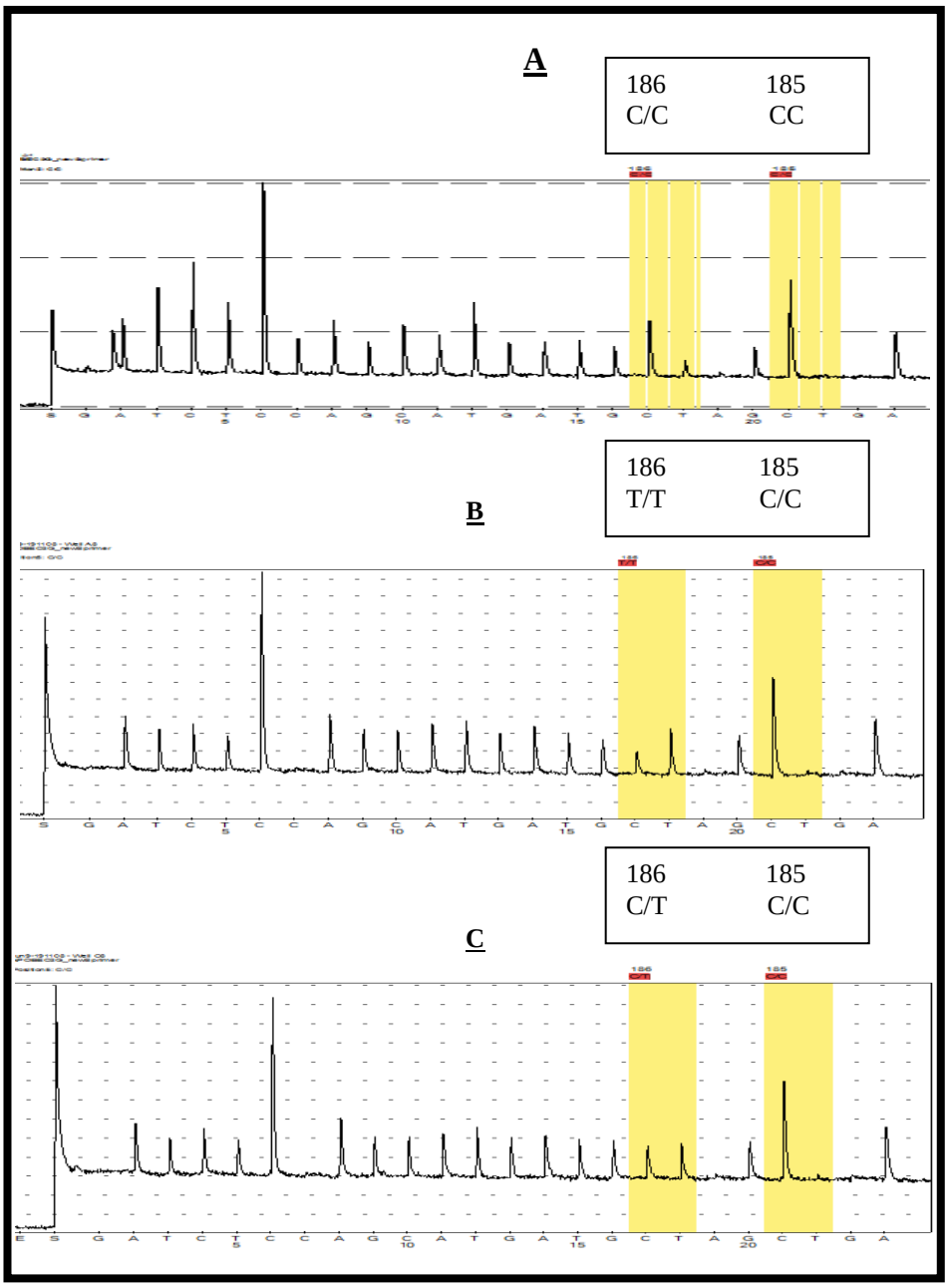


Figure 16. Pyrogram output files for codons 186 and 185. Pyrogram A indicates the CC genotype at position 186 and 185.Pyrogram B shows TT genotype was present at 186 in this sample and position 185 was homozygous for C allele. Pyrogram C shows this sample to be heterozygous at position 186 and homozygous for the C allele at position 185.

Table 4: Genotype and allele frequencies for H186R SNP in a subpopulation C of the whole study population and the χ^2 test (RFLP)

	GENOTYPE	N	ALLELE	ALLELE FREQ	GENOTYPE FREQ	χ^2
H186R	CC	4	C	0.34	0.040	11.007
	CT	60	T	0.66	0.594	
	TT	37			0.366	
		101				

Table 5: Genotype and allele frequencies for H186R SNP in a different subpopulation B of the whole study population and the χ^2 test (Pyrosequencing)

	GENOTYPE	N	ALLELE	ALLELE FREQ	GENOTYPE FREQ	χ^2
H186R	CC	24	C	0.49	0.176	7.102
	CT	86	T	0.51	0.632	
	TT	26			0.191	
		136				

The H186R SNP does deviate from the Hardy-Weinberg Equilibrium at $p < 0.05$. The genotypes are representative of those on the antisense strand. There is a large heterozygote excess and almost equal proportion of either homozygote.

5. Association of *APOBEC3G* genotypes with HIV/AIDS disease status

Table 6. Genotypes in the General population and HIV positive sub grouping

		General population				HIV +			
Upstream non coding region	Genotype	n	Freq	χ^2	P	n	Freq	χ^2	P
- 571	CC CG GG	0 3 29 32	0 0.094 0.906	0.369	0.5	2 37 65 104	0.019 0.356 0.625	3.603	0.05
H186R	CC CT TT	3 19 10 32	0.094 0.594 0.313	11.224	0.001	21 67 16 104	0.202 0.644 0.154	8.477	0.01

The HIV positive and the General population did not deviate from Hardy Weinberg Equilibrium at $P<0.05$ at -571 locus. The GG genotype is frequent in both sub groups. There is an absence of the CC genotypes in the general population and a very low frequency in the HIV positive group. There is a significant difference at H186R locus. There is a large heterozygous excess in general population and HIV positive groups. In the general population the TT genotype is more frequent than the CC genotype. However, in the HIV positive group the CC genotype has a higher frequency than the TT genotype.

6. Chi-squared Test: 2x 2 table

Table 7. Independent Chi-squared test for frequency distributions between the General population and HIV positive group for SNPs -571 and H186R

SNP	Chi squared value	P-value
-571 GP/HIV pos	9.068	0.0026
H186R GP/HIV pos	5.059	0.0245

Chi-squared test using the co-dominant model was conducted to compare frequencies between the General population and HIV + populations at H186R. There was a significant difference in frequencies between the two groups. This is suggestive of them playing a role in HIV/AIDS pathogenesis.

7. Pair-wise Allelic Linkage Disequilibrium

Table 8 Linkage disequilibrium in *APOBEC3G* gene

Site	D	|D'|	r^2
-571/ H186R	-0.127	0.216	0.0087

Pair-wise Linkage disequilibrium analysis across the two SNPs 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.

8. Haplotype analysis

The haplotypes across upstream non coding SNP and coding region SNP were analyzed using PHASE 2.0.

Table 9. Haplotype frequencies for the sample population.

Haplotype Name	Haplotype	Frequency
1	GC	0.418
2	GT	0.419
3	CC	0.074
4	CT	0.087

There are four haplotypes present in this study population. Haplotype 1 and 2 have equal frequencies while Haplotype 3 and 4 and comparable at 0.074 and 0.084 respectively.

Table 10. Haplotype analysis in General population and HIV positive subgroups

Haplotype	Haplotype	Frequency in HIV positive pop. (104)	Frequency in General pop. (32)
1	GC	0.431	0.368
2	GT	0.371	0.585
3	CT	0.104	0.024
4	CC	0.092	0.022

The GC haplotype was more common in the HIV positive population while the GT haplotype had a higher frequency in the general population (Table 7). The least common haplotype in both groups was haplotype 4. It occurred at a low frequency of 0.092 and 0.022 in the HIV positive and the general population respectively. The frequency of haplotype 1 was comparable between the two groups. However the remaining haplotypes were dissimilar between the groups.

DISCUSSION

Analysis of the direct sequencing of the upstream non-coding region showed six polymorphisms in 20 samples. SNP -90 was characterised only in six samples. SNPs -163, -166 and -199 appear to be in linkage disequilibrium. All three sites are either heterozygous or homozygous for the major alleles at each position. SNP -571 was well represented in the sequenced samples. Promoter analysis identified numerous transcription start sites (TSS) as well as several TATA box motifs (TBM). The FPROM programs identified eight TSSs and eight TBMs. The remaining two identified five promoter elements and five TBMs, four of which are identical. FPROM only identified three of those detected by TSSG and TSSW. FPROM seemed less stringent in promoter identification than TSSG and TSSW. The -571 SNP was genotyped with Allele Specific amplification (ASA) and RFLP analysis. While the H186R was genotyped with RFLP analysis and pyrosequencing. Genotyping of the study population at -571 and H186R showed that these SNPs deviated from Hardy Weinberg equilibrium (HWE). However, when the -571 study population was genotyped with RFLP analysis the population did not show any significant deviation from HWE. A large heterozygotes excess was found at the H186R locus. The difference in genotype frequency distributions between the HIV + and general populations at -571 and H186R was significant. The GT haplotype was more prevalent in the whole study population. However it was not the most frequent in the HIV+ group.

Direct sequencing remains the unsurpassed method for detecting variation. However it is relatively expensive so in the context of this study it was used an exploratory tool. There after indirect genotyping assays were designed to genotype the study population. Allele specific amplification did allow the genotyping of the study population at -571. However, in this study the ASA erroneously genotyped some samples when compared to the RFLP analysis. 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. The RFLP analysis was a more reliable technique. The discriminatory power is great because it allows one to design the assay at a species level or 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. Thus, this method was more reliable in genotyping. Consequently, genotyping by RFLP analysis the study population did not deviate from the HWE. 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.

The promoter analysis identified promoters. However, they were not the same as those identified in a previous study (Muckenfuss et al, 2007). The first TSS located with the FPROM program was 2748 bp upstream of the transcription initiation start site. All other promoters identified were between 2798 bp and 17208 bp upstream of the start site of the APOBEC3G mRNA. In contrast a TSS was located 66 nucleotides upstream of the start of the mRNA sequence (Muckenfuss et al, 2007).The core promoter of APOBEC3G was located within the region -114 and +66 (Muckenfuss et al, 2007). These positions correspond to 37802858 and 37803081 on chromosome 22 respectively, and were not identified with the promoter software programs. In addition the GC box was mapped to -87/ -78 (37802929/ 37802938) of the promoter region. This box was not identified by the software analysis. No independent studies have confirmed the results obtained by Muckenfuss et al (2007) but these results seem promising when compared

against what has been written about the features of promoters in the literature. Promoters in general have a TATA box located 10 nucleotides upstream of the TSS, followed by an element at 25 nucleotides upstream of the TSS and finally a CAAT element at 80 nucleotides upstream of TSS. Although APOBEC3G does not the typical elements at -10 and -25 it does appear to have the CAAT element present encompassing the same region (Klugg & Cummings, 1997).

Comparison of the genotyping results with the literature shows some marked differences. This study's frequencies were compared to the Yoruba population from Nigeria and African Americans. The Yoruban and African American data was obtained from dbSNP and Ensembl.

Table. 11. Comparison of population frequencies across three populations groups

	-571			H186R		
	Genotype	Genotype Frequency	Allele Frequency	Genotype	Genotype Frequency	Allele Frequency
This study	CC	0.279	C= 0.I6	CC	0.176	C=0.41
	CG	0.267		CT	0.632	
	GG	0.455	G= 0.84	TT	0.191	T=0.59
Yoruban	CC	0.850	C=0.917	CC	0.217	C=0.467
	CG	0.133		CT	0.533	
	GG	0.017	G=0.083	TT	0.283	T=0.550
African American	X	X	X	CC	X	C=0.390
	X	X	X	CT	X	
	X	X	X	TT	X	T=0.610

*The frequency data for the Yoruban and African Americans was obtained from dbSNP and Ensembl respectively.

The -571 allele frequencies are similar between the Yoruban and this study however, the genotype frequencies between the two are very different. The GG genotype has a higher frequency in this study but has a very low frequency in the Yoruba. The CC genotype was considerably higher in the Yoruban than in this study. There was no data available for African Americans thus no conclusions can be drawn regarding the similarities and/or differences in frequencies between AA and the Bantu population. Nonetheless, African has 2000 ethnically diverse populations. This difference in frequencies shows that the Yoruban population, from Nigeria, is different from the Bantu speaking study population. These ethnic differences do affect the genotype frequencies. The genotype frequencies for H186R are different between this study and the Yoruban. However, there remains a heterozygote excess in both populations. The TT genotype has a frequency of 33 % in African Americans (An et al, 2004). In contract TT genotype within this study population was 20 %. CC and TT genotype frequencies are analogous within both populations. In contrast the allele frequencies in the Yoruban population are equivalent to those in this study. However they appear to differ from the allele frequencies of African Americans (AA). No genotype data was obtained for AA at H186R and it would be interesting to note if there is still a heterozygote excess in this population group. This heterozygotes excess is not present in the CEPH (Utah residents with ancestry, from Northern and Western Europe), the Han Chinese and the Japanese populations studied in the HapMap. The frequency of the TT genotype for the CEPH, Han Chinese, and Japanese are 0.933, 0.911, and 0.822 respectively. This indicates that selection must be operating with this study population and the Yoruban allowing fro the heterozygote excess.

The heterozygote excess at H186R could be indicative of incorrect genotyping. However, two reliable methods employed showed similar results, suggesting that that this is not a result of incorrect genotyping. The TT genotype was more frequent in the general population than in the HIV+ group. This is expected

because more people would be advancing to AIDS and death hence less of this genotype would be found within an HIV + population. This consistent with a previous study where the TT genotype is associated with faster progression to progression AIDS (An et al, 2004). The substitution of histidine (H186) to arginine (186R) results in a change of charge from negative to positive. This change in polarity may influence the structure of *APOBEC3G* binding to vif thus leading to faster progression to AIDS. There also seems to be selection against the GG genotype at -571 in the HIV + group and thereby increasing the frequency of the CC and CG genotypes. The GG genotype of -571 is more frequent in the general population than in the HIV + group. Likewise, as more individuals progress to AIDS and death the GG genotype is reduced. Therefore, there is selection against this genotype in HIV + group. There is no reported clinical implication associated with the GG genotype at -571. The p-values indicate that the frequency distributions at -571 (0.0026) and H186R (0.0245) are statistically significant. To know for certain the impact of these SNPs on Bantu speaking population a well characterised sample needs to be studied. This means that a cohort of HIV -, HIV + and high risk exposed uninfected (HREU) individuals need to be followed over a period of time. In addition reported clinical details of the HIV+ group needs to be noted such as date of initial infection, CD 4 counts, and other illnesses such as TB. HIV status for the general population sampled in this study is unknown. This population was initially used to detect variation and may not be suitable to study disease susceptibility and progression. In addition, no reported clinical details were obtained for the HIV + group. Thus, the samples to be investigated need to be well characterised.

The pair wise 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 (An et al, 2004). This lack of correlation is interesting because these SNPs are relatively close together on the chromosome and the literature suggests that SNPs within 10kb are always in strong Linkage

Disequilibrium (Goldstein & Weale, 2001). In addition, linkage studies on Yoruban population indicates the |D'| drops below 0.5 on average every ~5kb. The -571 and H186R SNPs are just over 5000bp apart. However it is known that Africans have weak Linkage Disequilibrium based on studies of the Yoruban population (Reich et al, 2001). So the lack of linkage in this Bantu-speaking population is not surprising considering that this population may have a lower LD and therefore much more variant alleles.

An alternative explanation for the heterozygotes excess is copy number variations (CNVs). *APOBEC3G* gene has 8 exons. Exons 2, 3, 4 are duplicated within 5, 6, 7 resulting in two active site (2, 5), two linker regions (3, 6) and two psuedocatalytic domains (4, 7). This architecture has the potential for CNVs. CNVs have been reported to occur in regions within large homologous regions such as these duplicated exons (Freeman et al, 2009). In addition these duplications can be maintained by non homology based mechanisms leading to grater chromosomal re-arrangements within the gene. A consequence of CNVs is are allele frequencies that deviated form the expected Hardy-Weinberg equilibrium such as the heterozygotes excess of H186R SNP obtained through RFLP analysis and confirmed via pyrosequencing. However the Bantu speakers from South Africa have comparable number of CNV per loci pee individual when compared to the rest of the world (Jakobsson et al, 2008). This approach needs further investigation with well characterised individuals to ascertain the contribution of CNVs to disease especially HIV.

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 in HIV/AIDS. Future studies 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. These samples were very hard to acquire in Johannesburg. Nonetheless, this study proves that APOBEC3G plays a pivotal role in disease.

CONCLUSION

The study succeeded in genotyping the study participants at -571 and H186R. The frequency distributions of the SNPs between the HIV+ and the general population were statistically significant indicating that -571 and H186R variants do have a modifying role on HIV/AIDS pathogenesis. In particular these variants seem to be under selective pressure. Interestingly these variants though in close proximity are not linked and this warrants further investigation with better characterised samples. The differences in allele, genotype and haplotype frequencies seem to be in part population-specific. This is consistent with literature which indicates that Bantu speaking populations have been shaped by different selection pressures compared to Europeans and African Americans.

The only way to facilitate the proper Linkage Disequilibrium and Haplotype analysis is to have a more extensive reference panel. This panel must encompass populations from Southern Africa including the Bantu-speaking population. This will ensure better a more relevant predication of LD and consequently the haplotype inference will be more meaningful.

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 et al (1996). CCR5: A RANTES, MIPα, MIP 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., (2004).APOBEC3G genetic variants and their influence on the progression to AIDS. *Journal of Virology* **78**: 11070-11076.

Bailes, E., Gao, F., Bibollet-Ruche, F., Courgnaud, V., Peeters, M**.,** et al (2003). Hybrid origin of SIV in chimpanzees. *Science* **300**: 1713.

Bailey, J. R., Zhang, H., Wegweisr, B. W., Yang, H., Herrera, L., et al (2007). Evolution of HIV-1 in a HLA-B*57-positive patient during virologic escape. Journal of Infectious Disease 196: 50-55.

Bennasser, Y., Le, S., Benkirane, M., Jeang, K., (2005). Evidence that HIV-1 encodes a siRNA and a suppressor of RNA silencing. *Immunity* **22**: 607-619.

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

Brass, A., Dykxhoorn, D. M., Benita, **Y.,** Yan, N., Engelman, A., et al (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science* **319**: 921-926.

Brumme, Z. A., Brumme, C. J., Heckerman, D., Korber, B. T., Daniels, M., et al (2007). Evidence of Differential HLA Class I-Mediated Viral Evolution in Functional and Accessory/Regulatory Genes of HIV-1. PLos Pathogens 3: e94.

Cardozo, T., Kimura, T., Philpott, S., Weiser, B., Burger, H., et al (2007). Structural basis for coreceptor selectivity by the HIV Type 1 V3 loop. *AIDS Research and Human Retroviruses* **23:** 415-426.

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.

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

Chiu, Y. L., Greene, W. C., (2009). APOBEC3G: an intracellular centurion. *Philosophical Transactions: Biological Sciences* **364**: 689-703.

Clapham, P. R., McKnight, A., (2001). HIV-1 receptors and cell tropism. *British Medical Bulletin* **58**: p43.

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.

Cooke, G.S., & Hill, A.V., (2001).Genetics of susceptibility to human infectious disease. *Nat.Gen* **2**: 967-77.

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

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., et al (2007). A whole genome association of major determinants for host control of HIV-1. *Science* **317**: 944-947.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S., Driver, S., et al (1998). Potent and specific genetic interference by double stranded RNA in *Caenorhabditis elegans. Nature* **391**: 806-811.

Kidd KK, Jenkins T, Morar B, *et al* (1998), A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Human Genetics* **103**: 211-227.

Klugg, W. S., Cummings, M. R., (1997). Regulation of gene expression in Eukaryotes. Concepts of Genetics: 435.

Gallios-Montbrun, S., Kramer, B., Swanson, C. M., Byers, H., Lynham, S., et al (2007). Antiviral Protein APOBEC3G Localizes to Ribonucleoprotein Complexes Found in P Bodies and Stress Granules. *Journal of Virology* **81**: 2165-2178.

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.

Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., et al (1999). Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes. Nature* **397**: 436-441.

Gao, X., Nelson, G. W., Karacki, P., Martin, M. P., Phair, J., et al (2001).Effects of single amino acid change in MHC class1 molecules on the rate of progression to AIDS. *The New England Journal of Medicine* **344**: 1668-75.

Gao, Y., Lobritz, M. A., Roth, J., Abreha, M., Nelson, K. N., et al (2008). Targets of small interfering RNA restriction during Human Immunodeficiency Virus Type 1 replication. *Journal of Virology* **82**: 2938-2951.

Goila-Gaur, R., Strebel, K., (2008). HIV-1 Vif, APOBEC, and intrinsic immunity. Retrovirology 5: 51.

Goldstein, D. B., Weale, M. E., Linkage disequilibrium holds the key. *Current Biology* **11**: R576-R579.

Hahn, B. H., Shaw, G. M, De Cock, K. M., Sharp, P. S., (2000). AIDS as zoonosis: scientific and public health implications. *Science* **287**: 607-614.

Harris, R.S., Bishop ,K. N., Sheehy, A. M, Craig, H. M., Petersen-Mahrt, S. K., et al (2003).DNA deamination mediates Innate Immunity to Retroviral Infection. *Cell* **113**: 803-9.

Hartl DL and Clark AG (1989), Principles of population genetics, 2nd ed, Sinauer Associates, Sunderland, Massachusetts, pp 682.

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. *J. Mol. Biol.* **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.

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.

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

Hutcheson, H. B., Lautenberger, J. A., Nelson, G. W., Pontius,J. U., Kessing, B. D., et al (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.

Jakobsson, M., Scholz, S.W., Scheet, P., Raphael Gibbs, J., VanLiere, J. M., et al (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., et al (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, *Am J Hum Genet* **66** : 979–988.

Kanzaki, L. I. B., Ornelas, S. S., Arganaraz, E. R., (2008). RNA interference and HIV-1 infection. *Rev Med Virol* **18**: 5-18.

Keele, B. F., Van Herverswyn, F., Li, Y**.,** Bailes, E., Takehisa, J., et al (2006). Chimpanzees reservoirs of pandemic and nonpandemic HIV-1. *Science* **313**: 523-526.

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

Koito, A., Harrowe, G., Levy, J. A., Cheng-Mayer, C., (1994). Functional role of the V1/V2 region of human immunodeficiency virus type 1 envelope glycoprotein gp120 in infection of primary macrophages and soluble CD4 neutralization. *J Virol*. **68**: 2253-2259.

Laakso, M. M., Lee, F. H., Haggarty, **B**., Agrawal, C., Nolan, K. M., (2007). V3 Loop truncations in HIV-1 envelope impart resistance to coreceptor inhibitors and enhanced sensitivity to neutralizing antibodies. *PLos Pathogens* **3**: e117.

Lane, A.B., Soodyall, H., Arndt, S., Ratshikhopha, M. E., Jonker, E., et al (2002). Genetic substructure in South-African Bantu-speakers: evidence from autosomal DNA and Y-chromosome studies. *American Journal of Physical Anthropology* **119**: 175-185.

Lazarus, R., Vercelli, D., Palmer, L.J., Klimecki, W. J., Silverman, E. K., et al (2002). Single nucleotide polymorphisms in innate immunity genes: abundant variation and potential role in complex human disease. *Immunological Review* **190**: 9-23.

Lemey, P., Pybus, O. G, Wang, B., Saksena, N. K., Salemi, M., et al (2003). Tracing the origin and history of the HIV-2 epidemic. *PNAS* **100**: 6588-6592.

Lewotin RC (1964). The interaction of selection and linkage considerations; heterotic models, *Genetics* **49:** 49-67.

Loeuillet, C., Deutsch, S., Ciuffi, A., Robyr, D., Taffé, P., et al (2008). In vitro whole genome analysis identifies a susceptibility locus for HIV-1. *PLos Biology* **6:** e32.

Mangeat, B., Turelli, P., Caron, G., Friedli, M., Perrin, L., et al (2003).Broad antiretroviral defense by human APOBEC3G through 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 Dev.* **18**: 2861-2866.

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.

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

Muckenfuss, H., Kaiser, J. K., Krebit, E., Schwer, M., Cichuteket, C., et al (2007). Sp1 and Sp3 regulate basal transcription of the human APOBEC3G gene. *Nucleic Acid Research***:** e1-13.

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

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. *J Lab Clin Med* **114**: 105-113.

Pearson, T.A., and Manolio, T.A., (2008). How to interpret a genome wide association study. *JAMA* **299**: 1335-1344.

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

Rathore, A., Chatterjee, A., Yamamoto, N., Dhole, T. N., (2008). Absence of H186R polymorphism in exon 4 of the APOBEC3G gene among North Indians individuals. *Genetic Testing* **12**: 453-456.

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

Ronaghi, M., (2001). Pyrosequencing sheds light on DNA sequencing. Genome Research 11: 3-11

Sabeti, P. C., Walsh, E., Schaffner, S. F., Varilly, P., Fry, B., et al (2005). The case for selection at CCR5-Delta32. *PLos Biology* 3: e378.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., et al (2007). Genome wide detection and characterization of positive selection in human populations. *Nature* **449**: 913-918.

Salamov, A. A., Solovyev, V. V., (1997). Protein secondary structure prediction

using local alignments. *Journal of Molecular Biology* **268**: 31-36.

Samson, M., Libert, F., Doranz, B. J., Rucker, J., Liesnard, C., et al (1996). Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature* **383**: 722-725.

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.

Sharp, P., (2002). Origins of human virus diversity. *Cell* 108: 305-312.

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

Sousa, M. M. L., Krokan, H. E., Slupphaug, G., (2007). DNA-uracil and human pathology. Molecular aspects of Medicine 28: 276-306.

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. *Am. J. Hum. Gen.* **62:** 1507-1515.

Stephens, M., Smith, N. J., Donnelly, P., (2001). A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal for Human Genetics* **68**: 978-989.

Syvanen, A., (2001).Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat.Genet.* **2**: 930-40.

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

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

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

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., et al (2008). Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. *Nature* 455: 661-665.

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

Zhang, J., and Webb, D. M., (2004). Rapid evolution of primate antiviral enzyme APOBEC3G. *Human Molecular Genetics* **13**: 1785-1791.

Zhang, K. L., Mangeat, B., Ortiz, M., Zoete, V., Trono, D., (2007). Model structure of APOBEC3G. *PLos one* **2:** e378.

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**: 113-121.

APPENDIX

Raw Data Subpopulation A							
Sample	571-ASA	Sample	571-ASA	Sample	571-ASA	Sample	571-ASA
104	CC	171	CC	219	GC	541-353	GG
105	CC	172	CC	220	GC	541-36	GG
106	CC	173	CC	221	GC	541-49	GG
109	CC	175	GC	222	GC	541-62	GG
110	CC	176	GC	223	GG	541-73	GG
111	CC	177	GC	224	GG	541-98	GG
112	CC	178	GC	225	GG	614-121	GG
113	CC	179	GC	227	GG	615-107	GG
114	CC	180	GC	228	GG	615-11	GG
115	CC	181	GC	229	GG	615-136	GG
116	CC	182	GC	230	GG	615-15	GG
117	CC	183	GC	231	GG	615-26	GG
118	CC	184	GC	232	GG	615-31	GG
120	CC	185	GC	233	GG	615-325	GG
123	CC	187	GC	234	GG	615-332	GG
124	CC	188	GC	235	GG	615-340	GG
126	CC	189	GC	236	GG	615-358	GG
127	CC	190	GC	237	GG	615-377	GG
131	CC	191	GC	239	GG	615-394	GG
138	CC	192	GC	241	GG	615-406	GG
140	CC	193	GC	243	GG	615-59	GG
141	CC	194	GC	244	GG	615-67	GG
143	CC	195	GC	245	GG	615-78	GG
146	CC	197	GC	521-160	GG	615-80	GG
147	CC	198	GC	521-171	GG	615-93	GG
148	CC	199	GC	521-298	GG	616-17	GG
149	CC	200	GC	521-301	GG	616-42	GG
150	CC	201	GC	521-327	GG	616-445	GG
151	CC	203	GC	521-343	GG	616-453	GG
152	CC	204	GC	536-015	GG	616-457	GG
153	CC	205	GC	536-107	GG	616-472	GG
154	CC	206	GC	536-149	GG	616-486	GG
155	CC	207	GC	536-173	GG	616-499	GG
156	CC	208	GC	536-31	GG	616-503	GG
158	CC	209	GC	541-115	GG	616-91	GG
159	CC	210	GC	541-131	GG	615-366	GG
160	CC	211	GC	541-144	GG		
161	CC	212	GC	541-178	GG		
163	CC	213	GC	541-180	GG		
164	CC	214	GC	541-193	GG		
167	CC	216	GC	541-228	GG		
168	CC	217	GC	541-234	GG		
170	CC	218	GC	541-256	GG		

Raw Data Subpopulation B								
Sample	H186R-Pyro	571-RFLP	Sample	H186R-Pyro	571-RFLP	Sample	H186R-Pyro	571-RFLP
101	CT	GC	243	CC	GC	341	CT	GC
105	CT	GG	245	CT	GG	342	CC	GG
106	CT	GC	300	CT	GG	343	TT	GG
111	CC	GG	301	CC	GG	344	CT	GC
113	TT	GG	302	CC	GG	346	CT	GG
114	TT	CC	303	CC	GC	347	TT	GC
118	CC	GC	304	CC	GC	348	CT	GC
119	CC	GG	305	CC	GC	349	TT	CC
126	CC	GG	306	CT	GG	356	CT	GC
127	TT	GC	307	CC	GG	357	CT	GG
131	CT	GG	308	CC	GG	358	CT	GG
141	CT	GC	309	CT	GG	360	TT	GG
142	CT	GG	310	CT	GG	361	TT	GG
206	TT	GG	311	CT	GG	362	CT	GG
208	TT	GG	312	CT	GC	363	CT	GG
209	CT	GG	313	CT	GG	365	CT	GG
210	CT	GG	314	TT	GC	366	CT	GG
212	TT	GG	315	TT	GG	367	CT	GG
213	CT	GG	316	CT	GG	368	CT	GG
214	CT	GG	317	CC	GG	369	CC	GG
215	CT	GG	318	CT	GG	370	CC	GG
216	CT	GG	319	CT	GG	371	TT	GG
217	CT	GG	320	CT	GG	372	CT	GC
221	CT	GG	321	CT	GG	373	CT	GC
222	CT	GG	322	CT	GG	374	TT	GC
223	TT	GG	323	CT	GG	375	CT	GC
224	TT	GG	324	CT	GC	376	CT	GC
226	CT	GG	325	CT	GC	377	CT	GC
227	CT	GC	326	CT	GC	378	CT	GG
228	CT	GG	327	CC	GC	379	CT	GG
229	CT	GG	328	CT	GC	381	CT	GG
230	CT	GG	329	CT	GC	382	CT	GG
231	CT	GG	330	CT	GC	383	CT	GC
232	CT	GG	331	CT	GC	384	CT	GG
233	CC	GG	332	CC	GC	385	CT	GG
234	CT	GG	333	CT	GC	386	CT	GG
235	TT	GG	334	CT	GC	387	TT	GG
237	CT	GG	335	CC	GC	390	CT	GG
238	TT	GG	336	CT	GC	391	CC	GG
239	CC	GG	337	CT	GC	392	CT	GG
240	TT	GG	338	CT	GC	393	TT	GG
241	TT	GG	339	CT	GG	394	CT	GG
242	TT	GC	340	TT	GG	395	CT	GG
396	CC	GG	397	CC	GG	398	CT	GG
399	TT	GG	400	CT	GG	401	CT	GG
402	CT	GG						

Raw Data Subpopulation C					
Sample	H186R	Sample	H186R	Sample	H186R
105	AA	213	AG	541-62	AG
109	AA	214	AG	541-73	AG
111	AA	215	AG	541-85	AG
112	AA	216	AG	541-131	AG
113	AA	217	AG	541-144	AG
126	AA	218	AG	541-242	AG
147	AA	221	AG	541-256	AG
150	AA	222	AG	541-353	AG
151	AA	223	AG	615-11	AG
152	AA	224	AG	615-26	AG
153	AA	225	AG	615-31	AG
154	AA	226	AG	615-44	AG
155	AA	227	AG	615-59	AG
156	AA	229	AG	615-67	AG
160	AA	230	AG	615-78	AG
161	AA	231	AG	615-93	AG
172	AA	232	AG	615-107	AG
175	AA	233	AG	615-136	AG
176	AA	234	AG	615-325	AG
181	AA	235	AG	615-332	AG
182	AA	236	AG	616-17	AG
183	AA	237	AG	616-42	AG
184	AA	238	AG	616-91	GG
188	AA	239	AG	616-445	GG
189	AA	240	AG	616-457	GG
190	AA	241	AG		
193	AA	242	AG		
194	AA	243	AG		
195	AA	244	AG		
197	AA	245	AG		
198	AA	521-171	AG		
199	AA	521-298	AG		
200	AA	521-301	AG		
201	AA	521-316	AG		
203	AA	521-327	AG		
204	AA	521-343	AG		
205	AA	536-015	AG		
220	AA	536-31	AG		
206	AA	536-107	AG		
208	AA	536-149	AG		
209	AA	536-173	AG		
211	AA	541-36	AG		
212	AG	541-49	AG		