

https://dspace.org/

Genetic variation in TRIM5 in the black South African population

Wingfield, Chyreene Lesley Margaret

2009-07-01T12:01:49Z
http://hdl.handle.net/10539/7059

Downloaded from a DSpace repository. https://dspace.org/

CHAPTER 1

INTRODUCTION

1.1 HIV impact and prevalence

Acquired immune deficiency syndrome (AIDS) is caused by an attack on the immune system, specifically the CD^+ T-lymphocytes, and is the result of human immunodeficiency virus (HIV)-1 infection (Webber 2001). It has been the focus of public health attention for the past two decades and is particularly challenging as the causative virus genome has a very high mutation rate and is constantly changing. AIDS is a severe pandemic that has spread to the entire globe. However, the majority of infected individuals reside in sub-saharan Africa, where 62.5% of HIV infections occur (UNAIDS 2006).

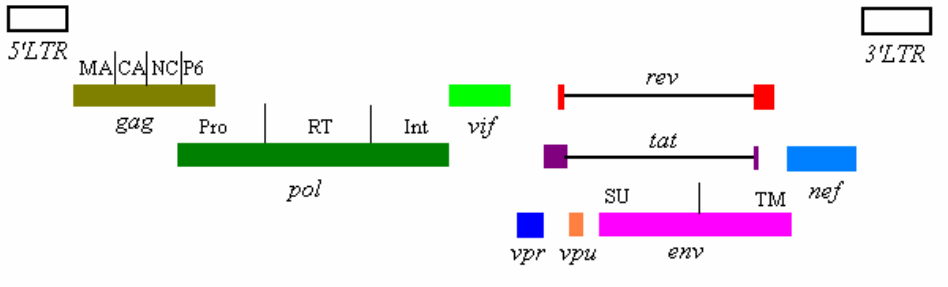
Currently there are approximately 39.5 million people living with AIDS, of which 37.2 million are adults and 2.3 million are children under the age of 15 (UNAIDS 2006). South Africa has one of the largest HIV epidemics in the world by number of HIV infections. In 2002 the prevalence of HIV in South Africa for people aged two to 14 years was 5.6%. This increased to a prevalence of 9.3% for youth, aged between 15 and 24 and as much as 15.5% for adults between the ages of 25 and 49, which is the age bracket of the country's work force, resulting in a loss in productivity, as well as an increase in orphans and child-headed households (Nelson Mandela/HSRC Study of HIV/AIDS 2002). This is confirmed by the age bracket with the highest percentage of AIDS prevalence in 2005 being 25 to 30 years of age and the second highest age bracket is 30 to 34 years of age according

to the 2006 report by UNAIDS. In 2006 there were approximately 2.9 million deaths due to AIDS worldwide and 2.1 million deaths due to AIDS in Sub-saharan Africa which translates to almost 8 000 deaths daily (UNAIDS 2006).

1.2 Basic biology of HIV-1

The genomes of retroviruses and HIV-1 in particular, consist of two identical, coding single strands of RNA that are tightly associated with one another (Figure 1.2.1a). The RNA strands are surrounded by a shell of proteins encoded by the *gag* gene (Figure 1.2.1b). The gag protein of the HIV-1 is initially translated into a 55kD polyprotein. This is then cleaved into four gag proteins, namely the matrix protein, the capsid protein, the nucleocapsid protein and p6, all of which make up the protein shell (Goff *et al* 2004a and Owens *et al* 2003). This core containing the RNA genome also includes the products of the *pol* gene, namely reverse transcriptase, integrase and a protease enzyme. Surrounding this core is a lipid bilayer. Incorporated in the lipid bilayer are clusters of envelope glycoprotein which are encoded by the *env* gene. The *gag*, *pol* and *env* genes, as well as six other genes, are flanked by long terminal repeat (LTR) sequences at either end of the genome. The reading frames overlap one another and the mRNA is spliced in different ways to produce different products (Wang *et al* 2000).

a)



b)

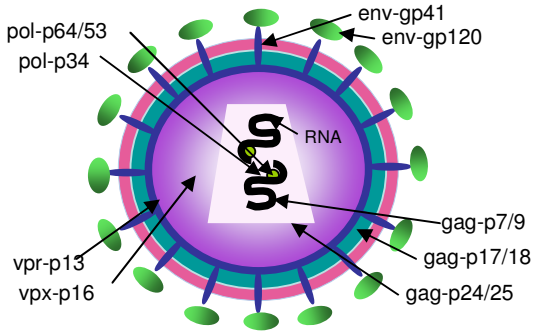


Figure 1.2.1. a) The genome of HIV-1 consists of 9 genes whose products form b) the virion.

The virus enters the host T-lymphocytes or macrophages by binding to a CD4 receptor and a chemokine co-receptor on the surface of the cell mediates the entry of HIV into the cell. Once the virus has entered the cell, the capsid surrounding the viral RNA is broken down and the RNA is reverse transcribed in the cytoplasm. Reverse transcription takes place in a large structure known as the reverse transcription complex (RTC) and involves the synthesis of DNA, copied from the RNA genome by the enzyme reverse transcriptase. The reverse transcriptase enzyme is multifunctional and has three activities, namely RNA-

dependant DNA polymerase activity, RNase H activity and DNA-dependant DNA polymerase activity (Nielsen et al 2005). The viral RNA is then degraded by the RNase H activity of the reverse transcriptase enzyme and a complementary DNA strand is synthesized from the initial strand. This results in a double-stranded DNA product with blunt termini. The DNA remains in a complex known as the preintegration complex (PIC) or provirus where the 3' ends of the viral DNA are cleaved by the viral integrase enzyme (Goff 2004a). Once in the nucleus the viral DNA is integrated into the host genome by the integrase enzyme (Wang *et al* 2000). The integrated viral DNA is now the template for transcription. The LTR sequences, flanking the integrated DNA, include receptors for host transcription factors and the viral DNA undergoes transcription by the DNA-dependant DNA polymerase activity of the reverse transcriptase to produce new viral genomes as well as up to 30 different viral mRNAs for translation into proteins. Splicing of the mRNAs is performed by host splicing machinery and the viral genome assembles with proteins transcribed by viral mRNAs to form new virions.

As HIV-1 has a relatively small genome of only 10kb and consists of nine genes itself, and due to the complexity of its life cycle the virus employs many host proteins in order to replicate. Therefore, many loci in the human genome have the ability to affect the severity of the disease phenotype. These loci are thus referred to as host genetic factors (Goff 2004a, Telenti 2005 and Webber 2001).

1.3 Restriction factors

Retroviral restriction factors are an example of host genetic factors. They are cellular proteins which influence the result of HIV-1 contact or infection. AIDS restriction factors differ in the stage in which they act. They can influence either susceptibility to HIV-1 infection or the progression to AIDS, as well as the stage of the HIV life cycle in which they act (as reviewed by O'Brien and Nelson 2004). This block occurs in a saturable manner and the restriction factors may affect any of the following stages in the retroviral life cycle: entry into the host cell, viral DNA synthesis, movement of viral nucleic acids between the cytoplasm and nucleus and protein expression (Figure 1.3.1). The genes encoding such restriction factors are known as AIDS restriction genes (ARG's) and play an important role in an innate immunity against viral infection (Goff 2004b).

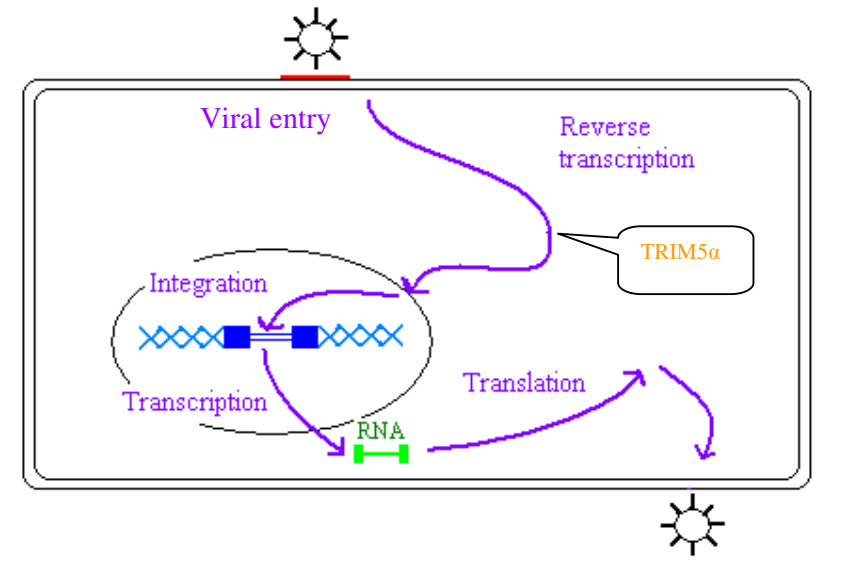


Figure 1.3.1 Diagram of the life cycle of HIV-1. Various stages in the life cycle are susceptible to the action of restriction factors. The block may be to viral entry, uncoating of the capsid, reverse transcription, movement into the nucleus or post-transcriptional (modified from Telenti 2005).

Polymorphisms in host genes that play a role in the viral life cycle may result in differences in susceptibility to HIV-1. AIDS restriction factors may determine whether an HIV infected person will be a 'rapid progressor' or a 'long term nonprogressor'. A so-called rapid progressor will develop AIDS symptoms within 1-5 years, whereas a long term nonprogressor may not develop any AIDS symptoms for up to 20 years (O'Brien and Nelson 2004).

1.4 TRIM5α

One example of an AIDS restriction factor is TRIM5α. For many years it has been evident that primate species differ in their susceptibility to infection by HIV-1. TRIM5α was recently identified as the species-specific mediator of this intrinsic cellular resistance to HIV-1. It was shown to be the factor responsible for an early post-entry block to HIV-1 infection in Old World monkeys (Stremlau *et al* 2004). TRIM5α from African Green monkey (AGM) cell lines is able to restrict HIV-1 and TRIM5α from humans and rhesus monkeys is able to restrict N-MLV (N-tropic murine leukaemia virus) (Yap *et al* 2004) and inhibit infection by EIAV (equine infectious anaemia virus) (Hatziioannou *et al* 2004b).

TRIM5α is an alternatively spliced variant of the *TRIM5* gene, situated on chromosome 11p15 and is approximately 21.5kb in size. It consists of 8 exons, 7 of which are included in the TRIM5α variant (Figure 1.4.1). The *TRIM5* gene codes for 5 other variants, namely TRIM5β, γ, δ, ε and ζ (Reymond *et al* 2001).

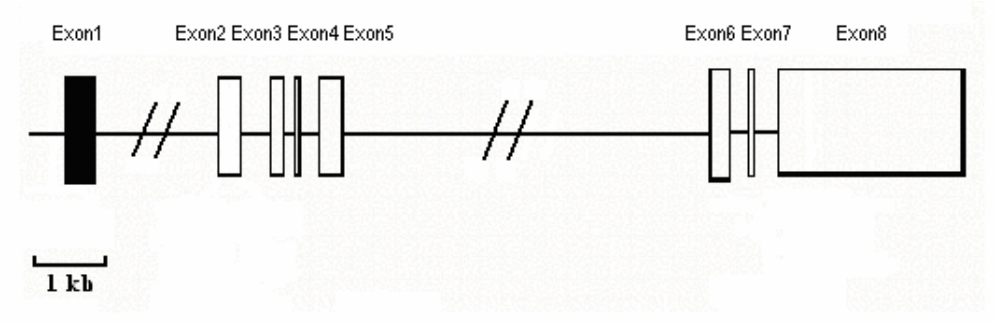


Figure 1.4.1 Relative positions of the introns and exons of TRIM5α in the *TRIM5* gene. Exon 1 (shaded) is not included in the TRIM5α variant.

The tripartite motif (TRIM) family of proteins have a modular structure consisting of three zinc-binding domains, namely a RING motif, a B-Box and a coiled-coil domain. For this reason they are also sometimes referred to as the RBCC family of proteins (Reymond *et al* 2001). The TRIM family of proteins may have one or two B-Box domains that differ in length and consensus sequences and are called B1 and B2. TRIM5α consists of the RING domain, a B2 domain and a coiled-coil domain as well as a SPRY domain (Figure 1.4.2) (Stremlau *et al* 2004). TRIM5 proteins have been found to associate with one another (multimerize) in the cytoplasm and to localize to discrete 'cytoplasmic bodies' (Reymond *et al* 2001).

The expression of TRIM5α, extracted from rhesus monkeys, ($TRIM5\alpha_{rh}$) in HeLa cells which express the correct CD4 receptors for HIV-1, blocks infection by HIV-1 but not MLV. It also inhibits infection by SIV slightly, but not to the extent that it inhibits HIV-1. Thus, $TRIM5\alpha_{rh}$ is capable of blocking infection by HIV-1 in cells of Old World monkeys and to a much lesser extent SIV (Stremlau *et al* 2004). TRIM5α from African green monkey also has the ability to restrict

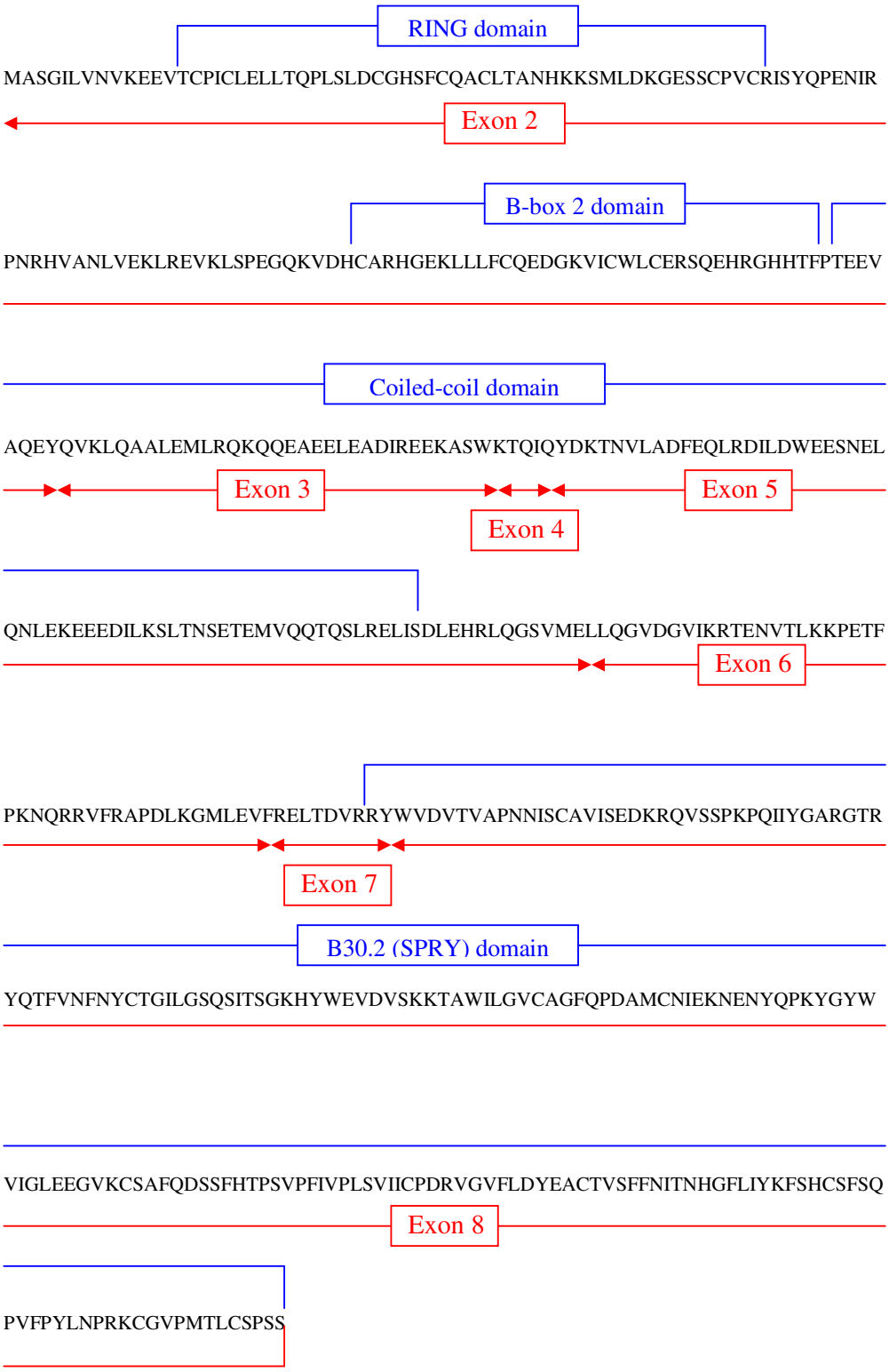


Figure 1.4.2 Human TRIM5α protein with key domains and corresponding exon positions shown.

both HIV-1 and SIV (Besnier *et al* 2002, Cowan *et al* 2002, Munk *et al* 2002 and Hatziioannou *et al* 2004). TRIM5α_{rh} is a requirement for restriction of HIV-1 in Old World monkey cells, demonstrated by short interfering RNA's (siRNA's) specifically targeting TRIM5α_{rh}, resulting in a large increase in HIV-1 infection in these cells (Stremlau *et al* 2004).

The p2 linker region joins the capsid and nucleocapsid domains of the polyprotein in HIV/SIV. A chimeric virus, SIV(HCA-p2)-GFP, which is a recombinant strain of SIV from rhesus macaque (SIV_{mac}) that contains an HIV capsid-p2 domain and expresses green fluorescent protein, is restricted at the levels that HIV-1 would be restricted at, rather than at the level SIV_{mac} would be restricted at, in primate cells. Thus, the target of the block is most likely the p2 domain of the capsid protein (Cowan *et al* 2002, Owens *et al* 2003 and Stremlau *et al* 2004) and the capsid sequences affect viral susceptibility to the restriction that is mediated by TRIM5α from rhesus macaque monkeys (TRIM5α_{rh}), or more specifically, the capsid-p2 region affects viral susceptibility to TRIM5α_{rh} restriction. This specific binding of TRIM5α to the capsid has also been demonstrated in another study (Sebastian *et al* 2005).

TRIM5α_{rh} causes rapid degradation of the HIV-1 capsid upon entry into the cell. The restriction is specific to the capsid protein as the matrix, nucleocapsid and other viral proteins do not undergo this rapid degradation. Binding of TRIM5α to the capsid is a requirement for restriction, and the presence of TRIM5α in the cell correlates with a decrease in the amount of HIV-1 capsid in the cytoplasm.

Therefore it is likely that TRIM5α-mediated restriction of various retroviruses occurs by recognizing the capsid cores and disassembling the capsids (Chatterji *et al* 2006 and Stremlau *et al* 2006). TRIM5α_{rh} either accelerates the decay of viral cDNA or disrupts its synthesis (Stremlau *et al* 2004). In cells expressing TRIM5α_{rh} the viral cDNA levels are very low in comparison with control cells that are not expressing the restriction factor (Stremlau *et al* 2004).

TRIM5α mutants for RING, SPRY and B-box 2 domains have been tested for restriction efficiency. Intact B-box 2 and SPRY domains are necessary for HIV-1 inhibition by TRIM5$\alpha_{rh,}$ whilst an intact RING domain is not an absolute requirement for restriction of HIV-1 by TRIM5$\alpha_{rh.}$ (Javanbahkt *et al* 2005, Stremlau *et al* 2004, 2005 and Sawyer *et al* 2005).

The RING domain has been implicated in the correct localization of TRIM5α proteins and thus may indirectly affect the level of restriction by affecting the levels of protein in the cytoplasm (Javanbahkt *et al* 2005, Stremlau *et al* 2004, 2005a, 2005b and Sawyer *et al* 2005). The coiled-coil domain, which is responsible for homomultimerization of TRIM5α, is necessary for restriction as active TRIM5α is present as a multimeric complex (Perez-Caballero *et al* 2005). However the SPRY domain is probably the most important domain in determining HIV-inhibition by TRIM5α_{rh}, because other members of the TRIM gene family are alternatively spliced in such a way that they lack SPRY domain and they do not have the restrictive properties displayed by TRIM5α_{rh}. This domain, specifically the region between amino acids 332 and 340, plays the largest role in

restriction specificity (Perez-Caballero *et al* 2005, Sawyer *et al* 2005 and Yap *et al* 2005). Interestingly this region between amino acids 332 and 340 of human TRIM5α shows the greatest diversity among sequences of human, rhesus and African green monkey TRIM5α. The SPRY domain is also very important in determining the amount of restrictive activity of TRIM5α and less than 2% change in the human protein can confer strong restrictive activity to HIV-1 (Stremlau *et al* 2005). The functions, either proven or hypothesized, of the different domains in the TRIM5α protein are summarized in figure 1.4.3 below.

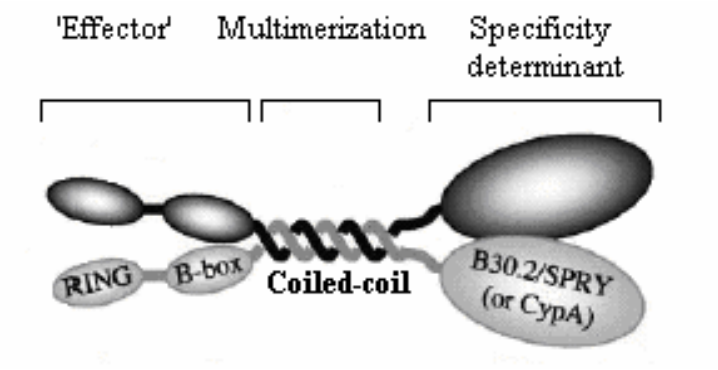


Figure 1.4.3. The functions of the different domains of TRIM5α that have been demonstrated or inferred from genetic evidence. The RING and B-box 2 domains probably effect retroviral restriction, the Coiled-coil domain is responsible for dimerization and the SPRY domain determines species specificity (modified from Perez-Caballero *et al* 2005)

While TRIM5α interacts with and causes the rapid degradation of the HIV-1 capsid in Old World monkeys, the exact mechanism of action of TRIM5α has yet to be discovered. One possible scenario is that TRIM5α binds to the capsid core and directs it to compartments such as lysosomes where it is uncoated and

degraded. Another possible mechanism of action is that TRIM5α provokes core uncoating so that the uncoated capsid molecules are more susceptible to degradation than capsid molecules associated with the core. A third possibility is that TRIM5α facilitates an association of the core with proteases that are then able to take break down the capsid, rendering the inner reverse transcription complex dysfunctional (Chatterji *et al* 2006). It is also uncertain whether the degradation of the capsid is the cause or a result of TRIM5α-mediated restriction.

The strong block that TRIM5α confers to various retroviruses has probably been maintained due to the selective advantage it confers (Stremlau *et al* 2004). There is evidence of positive selection in TRIM5α in primate species. The ratio of non-synonymous (dN) to synonymous substitutions (dS) can be used to determine whether there is positive selection acting on a gene, a value of >1 for this dN/dS ratio indicates positive selection. *TRIM5* has an average dN/dS of 1.1 for the gene with four out of 10 primate species having a dN/dS ratio of >1. There are two clusters of positive selection in the region encoding the SPRY domain of the gene, named variable region 1 (v1) and variable region 2 (v2) (Ortiz *et al* 2006). In other studies 10 out of 22 primate lineages have dN/dS >1 and 12 out of 20 primate lineages showed dN/dS >1 for the whole *TRIM5* gene. This becomes 15 out of 22 primate lineages and 9 out of 20 primate lineages with a dN/dS > 1, respectively, when looking only at the region of the *TRIM5* gene encoding the SPRY domain (Liu *et al* 2005 and Sawyer *et al* 2005). This is evidence of positive selection in the *TRIM5* gene of various primates and especially the SPRY domain, presumably due to selection by exposure to previous retroviruses.

The differences in restriction between different Old World monkey species can most likely be explained by variation in the *TRIM5* genes (Stremlau *et al* 2004). Indeed African green monkey (AGM) TRIM5α has an 18 amino acid insertion that is not present in rhesus macaque TRIM5α. This insertion occurs in the SPRY domain and appears to confer specificity to a broad range of viral capsids (Keckesova *et al* 2004). Similarly AGM TRIM5α has a 20 amino acid duplication in the SPRY domain that is not present in the cynomolgus monkey. A construct containing this duplication in the background of cynomolgous monkey TRIM5α confers the ability to effectively restrict SIV_{mac} (Nakayama *et al* 2005). This again points to the likelihood of the specificity to various retroviruses being determined by the SPRY domain of TRIM5α. Interestingly, the SPRY domain is also found in members of the immunoglobulin superfamily, a family of proteins involved in protection from invading pathogens, so it very well may confer specificity in pathogen detection (Keckesova *et al* 2004).

Variation in other regions of the *TRIM5* gene may also play a role in the differences in restriction mediated by TRIM5α. Single nucleotide polymorphisms (SNPs) in the regulatory region are able to alter signalling or transcription factor acivity, thereby influencing the level of gene expression and thus the level of protein in the cytoplasm (Johnson *et al* 2005). Upstream polymorphisms may also be linked to internal functional polymorphisms. As the B-box 2 domain is necessary for restriction by TRIM5α, polymorphisms in the part of the gene encoding this region may also determine the amount of restriction conferred by

this protein. Variation in the RING and coiled-coil domains may also indirectly affect restriction by affecting the levels of available protein.

Variation has been found that causes significant differences in susceptibility to HIV and certain haplotypes seem to have an effect on the progression of AIDS. There have been three recent studies on various human populations that illustrate these differences to susceptibility and disease progression. One study on European- and African-American populations showed significant association between two alleles in *TRIM5*, which cause amino acid substitutions 43H and 136R, and HIV-seropositive status. TRIM5α variants 43H ad 136Q exhibited better anti-HIV activity in tissue culture than the 43Y and 136R variants, respectively (Javanbakht *et al* 2006). A study on a European-American population detected no significant differences between HIV-positive and seronegative samples at any single polymorphic site, however there was significant association between the two groups for the presence of the haplotype containing the minor allele for the non-synonymous variant R136Q, with the haplotype containing 136Q being elevated in HIV-positive samples (Speelmon *et al* 2006). The differences in the results obtained by the above studies on the relationship between this gene and HIV-susceptibility may be due to differences in sample size as well as differences in the populations used for the studies. In a third study on a large cohort of HIV-positive subjects, none of the common TRIM5α variants had any association with disease progression, however weak association between some haplotypes and disease progression was detected, but this was not significant. This study also shows that at site 136 the Q variant

represents the ancestral variant. Interestingly, none of the variation in the human *TRIM5* gene occurs within the variable regions (v1 and v2) (Goldschmidt *et al* 2006). It seems that while some variants may result in differences in HIV-1 susceptibility, no effects on disease progression have yet been detected.

1.6 Population variation studies of restriction factors

When looking at genetic variation in relation to viral susceptibility it is important to note that this genetic variation is a subset of overall human genetic variation and as such it is important to understand the evolutionary processes underlying genetic variation (Jorde *et al* 2001). One of these processes is a fluctuation in effective population size. At some stage in the past the human population decreased significantly in size and then increased rapidly. This is referred to as a bottleneck and results in a loss of genetic diversity (Jorde *et al* 1998, Jorde *et al* 2001). Another factor that could have an impact on the amount of genetic variation observed in the human population is the time since the species originated. A recent evolutionary origin of the species results in low levels of genetic diversity (Jorde *et al* 2001). Another process that influences the amount of genetic variation in human populations is gene flow (Zietkiewicz *et al* 1997) however, bottlenecks are more important when comparing DNA sequence variation in African populations to other populations (Jorde *et al* 2001 and Akey *et al* 2004).

Genetic evidence supports the notion that the global population arose out of Africa based on mitochondrial DNA, nuclear DNA and Y-chromosome studies.

This is known as the 'recent African origin' (RAO) model. African populations have been shown to have the greatest variation or diversity, as can be seen by the long branch lengths of the cladogram in figure 1.6.1 based on mtDNA. This is most likely due to their population sizes remaining large, whilst other populations experienced bottlenecks and founder effects (Jorde *et al* 1998, Jorde *et al* 2001, Kimmel 1999, Maca-Meyer *et al* 2001, Stoneking and Soodyall 1996; Tishkoff and Williams 2002). Based on the RAO model all variation that exists in other regions of the world is likely to be a subset of the variation contained within the African continent or represents new variants that have arisen since the migration out of Africa (Jorde *et al* 1998; Risch *et al* 2002).

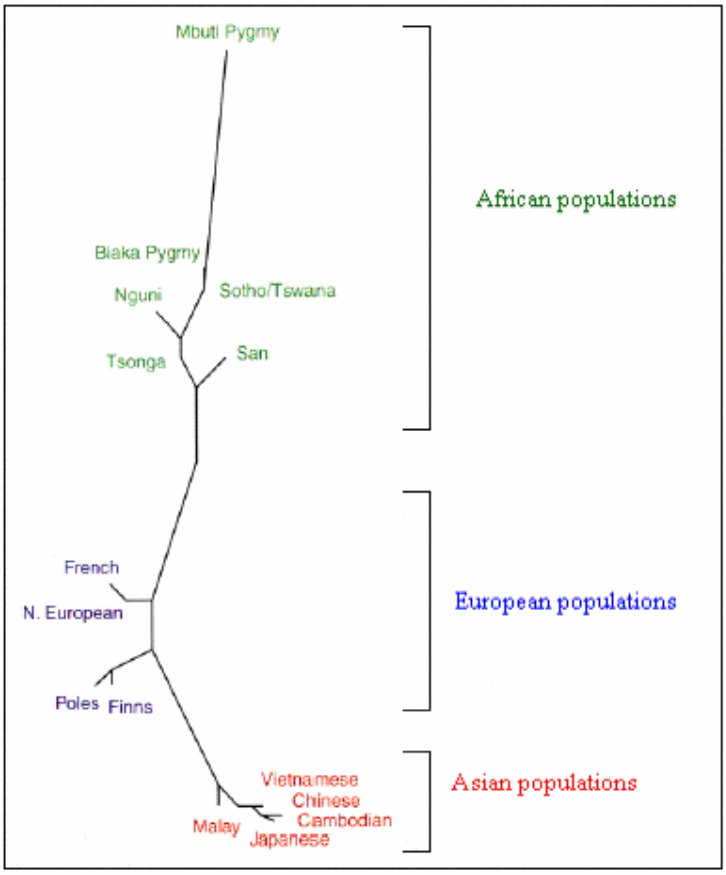


Figure 1.6.1. An unrooted cladogram based on mtDNA from African, European and Asian populations. The African populations have the longest branch lengths, indicating greater genetic diversity within and between these populations. Microsatellite data produces very similar results (modified from Jorde *et al* 1998).

Analysis of mitochondrial DNA and Y-chromosome haplotypes has been used to trace historical migration events. African populations underwent three major migration events before the hypothesized global migration out of Africa (Tishkoff and Williams 2002). The first migration occurred approximately 120 000 years ago from the east of the African continent into central Africa and southern Africa. Populations from these areas, namely the Biaka pygmies and the San respectively, represent the most genetically diverse populations. The second migration

occurred 30 000 to 70 000 years ago from the east of Africa into central and southern Africa, giving rise to the Mbuti pygmies residing in the Democratic Republic of Congo and Bantu speaking populations of west Africa. The third migration was into East Africa, Ethiopa and also out of Africa into Eurasia.

Another, more recent, migration is that of Bantu-speaking populations into South Africa approximately 3 000 years ago, possibly from somewhere in or near Cameroon (Tishkoff and Williams 2002). The black population of South Africa makes up 76.6% of the total population and nine of the eleven official languages are Bantu languages, which are part of the Niger-Congo language family (Lane *et al* 2002). The Bantu languages are hypothesized to have their origins in West Africa 3000 to 5000 years ago. There are two branches of Bantu languages, namely Eastern and Western, and the South African Bantu languages are the Southern branch of the Eastern group. Within the South African Bantu speakers gene frequencies among ethnically defined populations are being affected by increasing movements of people out of rural areas and into urban areas. Linguistic distances between seven South-African Bantu speaking groups are correlated with genetic distances, and geographic distances are also correlated with genetic distances, however linguistic distances are not significantly correlated with geographic distances. This could be explained by linguistic and some genetic differentiation occurring before the groups reached their current locations and further genetic differentiation taking place after they settled at their current locations (Lane *et al* 2002).

The human population size has increased dramatically in the past 50 000 to 100 000 years, long after speciation from *Homo erectus* to *Homo sapiens*, and even more dramatically after the agricultural revolution of the past 10 000 years. This population expansion occurred more rapidly in Asian and European populations than in African populations (Tishkoff and Williams 2002). The effect of such a rapid population expansion is that low-frequency, jointly transmitted alleles or haplotypes may be maintained in the population at a higher frequency than predicted (Tishkoff and Williams 2002 and Watkins *et al* 2003). Non-African populations also likely went through more bottlenecks due to founder effects after the migration out of Africa (Tishkoff and Williams 2002 and Watkins *et al* 2003); in this case some rare alleles are likely to be lost due to a smaller effective population size.

Whilst population relationships and history are studied by using presumably selectively neutral polymorphisms, selected polymorphisms have the ability to shape the variation found in AIDS restriction genes due to the selective advantage that may be conferred by certain genotypes.

Africa is the source of HIV, as is indicated by the similarity between SIV from *Pan troglodytes* residing in Africa and HIV-1 (Paraskevis *et al* 2003 and Keele *et al* 2006). HIV-1 pandemic group M and non-pandemic group N have been traced to distinct chimpanzee communities in Cameroon due to the close relationship between these HIV groups and SIVcpz*Ptt* (SIV from *Pan troglodytes troglodytes*) strains from the south of Cameroon (Keele *et al* 2006).

Alleles at neighbouring loci are often inherited together, a property known as linkage disequilibrium (LD). This definition can also be extended to the non-random association between genetic markers such as SNPs. LD is usually lower in African populations than in non-African populations due to the bottlenecks experienced by non-African populations (Jorde *et al* 2001, Reich *et al* 2001, Ardlie *et al* 2002 and Tishkoff and Verrelli 2003). New mutations are usually in LD with other SNPs in close proximity, however this LD decays over time (Tishkoff and Verrelli 2003). Even though LD is lower in African populations, it is also observed that higher levels of LD are observed in populations that have undergone recent admixture, as is the case with Bantu-speaking populations moving from rural to urban areas (Jorde *et al* 2001).

Selection for disease resistant variants also results in an increase in linkage disequilibrium. Natural selection for an allele results in an increase in LD as it results in the allele attaining a high frequency more rapidly than would normally be expected if it were selectively neutral. Thus LD can be used to detect past natural selection (Ardlie *et al* 2002). Deviations from normal expected genotypic frequencies, determined by using Hardy-Weinberg calculations, can also be used to detect current natural selection, however this is not a robust test as very strong selection is required to produce deviations from Hardy-Weinberg equilibrium.

1.7 Problem identification

The majority of studies on variation in AIDS restriction genes have been done using North-American or European population samples. However, African populations have been neglected when it comes to studying the variation within AIDS restriction genes. As the majority of AIDS sufferers reside in sub-saharan Africa it is important to determine the level of polymorphism within restriction genes affecting susceptibility to HIV and the rate of progression of AIDS.

Polymorphism exists within the *TRIM5*$_{rh}$ gene that may have an effect on disease susceptibility and progression in the case of SIV (Kodama *et al* 2005), and in *TRIM5*$_{hu}$ the same is true (Javanbakht *et al* 2006 and Speelmon *et al* 2006). In African populations, this polymorphism may occur at even greater levels, due to the greater natural variation and a different history of infectious diseases in these populations. It is of great interest to examine the role of *TRIM5* polymorphism in intra-species variability to viral restriction in human populations. However, non-synonymous amino acid changes are often not enough to account for differences in disease susceptibility and progression; also of importance is regulatory variation, as polymorphisms in the promoter region of the gene may have an effect on the amount of protein present or may be linked to polymorphisms within the gene that play a role in the level of restriction. Studying the genetic factors associated with susceptibility to HIV-1 infection and onset of disease symptoms contributes to our knowledge of their influence on this pandemic and may possibly pave the way for more effective treatment at a population level.

In this study SNPs in various regions of the *TRIM5* gene as well as in the upstream non-coding region of *TRIM5* of black South Africans were detected and characterized. Genotypes were compared between HIV-positive samples and general population samples to examine association with susceptibility to HIV-1.

CHAPTER 2

MATERIALS AND METHODS

2.1 Samples

The samples used in this population study consisted of 191 DNA samples obtained from black South Africans infected with HIV and 41 samples obtained from black South Africans with unknown HIV status, which were used as a general population.

One hundred and one HIV-positive blood samples were obtained from participants at the Infectious Disease Clinic at Johannesburg Hospital as well as two blood samples from individuals of unknown HIV status. Twenty seven HIV-positive samples were collected from the Themba Lethu clinic at Helen Joseph Hospital along with one sample from an individual with unknown HIV-status. All the above HIV-positive samples had clinical data, including age, estimated time of HIV infection, most recent $CD4^+$ cell count and whether or not the participant had ever been infected with tuberculosis or had any other HIV-related illness. Sixty three DNA samples extracted from the blood samples of HIV-positive individuals were provided by Dr Clive Gray from the National Institute for Communicable Diseases. These samples were accompanied by information on viral load and $CD4^+$ cell count data. Thirty eight blood samples were collected from black South African staff and students at the University of the Witwatersrand, regardless of HIV status.

Written informed consent was obtained from all participants in this study (Appendix 1). The research protocol and collection of samples was approved by the Human Research Ethics Committee of the University of the Witwatersrand, protocol number M040221 (Appendix 2).

2.2 DNA isolation

DNA isolation from blood samples collected at the Infectious Disease Clinic of the Johannesburg General Hospital was performed by former students of the Molecular and Cell Biology Department, University of the Witwatersrand. DNA isolation from blood samples, collected in ethylenediaminetetraacetic acid (EDTA) tubes, at the Themba Lethu clinic at Helen Joseph Clinic was performed using the same method. Blood tubes were centrifuged at 2500 x g for 15 minutes in order to separate plasma, buffy coats (containing leukocytes) and erythrocytes. DNA was extracted from the leukocytes in the buffy coat using the QIAmp® Blood DNA kit as per manufacturer's instructions (Qiagen). RNase was used to eliminate any contaminating traces of virus or RNA in the sample. DNA extraction was performed in QIAmp spin columns, which contain a DNA-adsorbing silica-gel membrane. DNA was eluted from the membrane in Tris-EDTA (TE) buffer, comprised of 10mM Tris-HCl and 1 mM EDTA, and stored at -20°C. Isolated DNA was electrophoresed on 0.8% agarose gels in Tris-borate, EDTA (TBE) buffer at 7.8V/cm for 1 hour in order to determine the size and estimate the concentration of the DNA. The TBE buffer consisted of 89mM Tris base, 89mM boric acid and 2mM EDTA. This DNA was used for the detection and characterization of variation in the *TRIM5* gene.

2.3 Direct detection of variation in regions of the *TRIM5* gene

Four regions of the *TRIM5* gene were sequenced (Figure 2.3.1) in approximately twenty samples collected from individuals with unknown HIV status (Table 2.3.1).

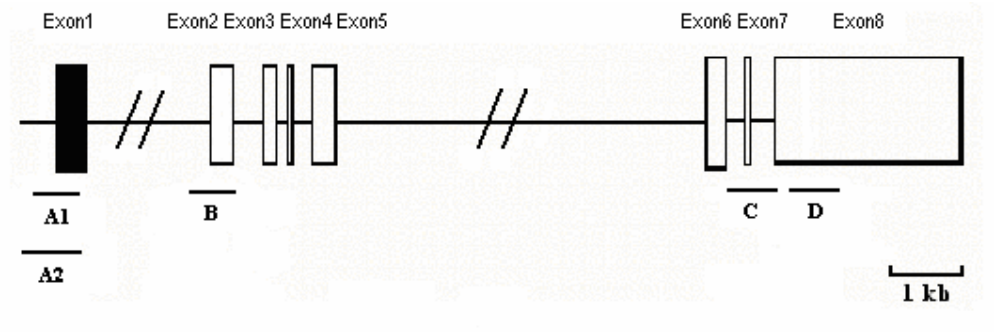


Figure 2.3.1. The *TRIM5* gene, showing regions that were sequenced (A1, A2, B, C and D)

This sample size was sufficient to detect variation that is common in the population. Detection was done by performing PCR amplification of the four regions A1, B, C and D (Table 2.3.1), followed by direct sequencing of the four different PCR products. The regions sequenced were chosen based on the functions of various domains in the gene. Region A1 spans the upstream promoter of the gene and polymorphisms in this region may play a role in the levels of protein produced. It also spans the untranslated exon 1 which may contain polymorphisms that are in linkage disequilibrium with other polymorphisms in the gene that play a role in restriction. Region B spans exon 2 which codes for the RING and B-box 2 domains, both of which are likely to effect

viral restriction. Region C spans part of exon 6, exon 7 and part of exon 8. Exon 7 is in the coding region but does not code for any of the functional domains.

Table 2.3.1. Regions of the gene that are spanned by the PCR primers for sequencing, sizes of the PCR product and samples sequenced for each region.

Primer pairs	Region spanned by sequencing	Region in figure 2.3.1	PCR product size	Samples sequenced
T5upst2-f T5upst-r	Upstream and part of 5'UTR	A2	714bp	206, 207, 209, 210, 211, 212, 217, 222, 226, 229, 230, 231, 232, 233, 235, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245
T5upst-f T5upst-r	Upstream and part of 5'UTR	A1	639bp	106[12], 112[12], 132[12], 138[12], 146[12], 148[12], 159[12], 213, 214, 215, 216, 218, 219, 221, 223, 224, 225, 227, 228
T5ex2-f T5ex2-r	Exon 2	B	659bp	210, 211, 212, 214, 215, 216, 218, 221, 222, 223, 224, 225, 226, 227, 228, 229, 230, 231, 232, 233, 234
T5ex67-f T5ex67-r	Part of exon 6, intron 6, exon 7, intron 7 and part of exon 8	C	713bp	213, 216, 219, 221, 222, 224, 225, 228, 230, 231, 232, 233, 234, 235, 237, 239, 240
T5ex8-f T5ex8-r	Coding part of exon 8	D	603bp	206, 207, 209, 210, 211, 212, 213, 214, 215, 216, 218, 219, 221, 222,223, 224, 225, 228, 229, 230, 232, 234

[1]Samples were sequenced by a former student.

[2]Samples are HIV positive samples, all the others are samples from individuals with unknown HIV status.

The part of exon 8 that was sequenced codes for the beginning of the SPRY domain. The final region sequenced, D, includes the coding part of exon 8 and therefore the entire SPRY domain, which is responsible for viral specificity. An additional 26 samples (Table 2.3.1) were sequenced for the upstream region, A2, as a high number of polymorphisms were detected in this region. A different forward primer was used to detect a polymorphism close to the beginning of the region amplified in the initial samples for this region (Table 2.3.2).

Polymerase chain reaction was carried out in a 50µl reaction volume. The reaction mixture contained 1.25 units of *Taq* DNA polymerase, 2mM $MgCl_2$, 0.8mM of dNTP's, 1µM of each of the forward and reverse primers and 5µg of template genomic DNA. Sequences of the primers used are shown in Table 2.3.2. Primers for PCR and sequencing were designed with the use of the online tool Primer3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). Primer specificity was checked by running a BLAST search (http://www.ncbi.nlm.nih.gov/BLAST/) of the primer sequence against the human genome sequence in order to guarantee that only the target DNA sequence would be amplified. All primers used were synthesized by Inqaba Biotechnical Industries (Pty) Ltd and were resuspended in Tris-EDTA (TE) buffer made up with 10mM Tris-HCl and 1mM EDTA.

Table 2.3.2. Names given to primers, sequences and orientation of all primers used for PCR and sequencing. PCR annealing temperatures of all primer pairs used. PCR product sizes produced by PCR with primer pairs.

Name	Primer Sequence	Orientation	Region in fig 2.3.1	Ann. temp.[1]	Size[2]
T5upst-f	5'TGCAGCTCCTTTTCTGTTCATA3'	Forward	A1	54.9°C	639bp
T5upst-r	5'CAGTTTTGTCCCAAAGACTTCC3'	Reverse			
T5upst2-f	5'CCCTCTCCAGATAAATTTCTAGC3'	Forward	A2	55.9°C	714bp
T5upst-r	5'CAGTTTTGTCCCAAAGACTTCC3'	Reverse			
T5ex2-f	5'GGGATCTGTGAACAAGAGGAAC3'	Forward	B	54.9°C	659bp
T5ex2-r	5'CTGTAATTGGGTGAAATGCAAA3'	Reverse			
T5ex67-f	5'GAGAGTGTTTCGAGCTCCTGAT3'	Forward	C	57.7°C	713bp
T5ex67-r	5'GTTTCGGAGAGCTCACTTGTCT3'	Reverse			
T5ex8-f	5'GACAGTGGCTCCAAACAACAT3'	Forward	D	57.6°C	603bp
T5ex8-r	5'AGGGGCTGAGTGTGTAAGAAGG3'	Reverse			

[1]PCR annealing temperature of the primer pair

[2]Size of the PCR product produced by amplification

PCR amplification consisted of one denaturation cycle for 2 minutes at 94°C, followed by 35 cycles of denaturation for 30 seconds at 94°C, annealing of primers for 45 seconds at annealing temperatures given in Table 2.3.2 and extension for 60 seconds at 72°C. This was then followed by a final extension for 5 minutes at 72°C. The presence and size of the PCR product was confirmed by electrophoresing 5µl of the PCR product on a 1% agarose gel in Tris-borate EDTA (TBE) buffer at 7.8V/cm for 45 minutes with 2µg GeneRuler™ 1kb DNA Ladder (Fermentas Life Sciences). Gels were visualised under UV light using the

UVP BioDoc-It™ system. The PCR product was purified and sequenced by Inqaba biotec. Two samples of each PCR product were sequenced in both directions in order to determine the best primer to be used for sequencing of the additional 18 samples. Sequences were aligned to a *TRIM5* reference sequence obtained from the Ensembl genome browser (http://www.ensembl.org) and analysed using the computer software tool Sequencher™ 4.5. Segments at the ends of the sequence, where signal peaks overlapped, were removed and differences between sequences were examined more closely to determine whether they were artefacts of sequencing or true reflections of sequence variation.

2.4 Identification of known variation

Known single nucleotide polymorphisms (SNP's) were identified using the SNP database of Ensembl (http://www.ensembl.org). The Human GeneSNPview of the transcript ENST00000380034 was used to detect known SNPs as well as the the genomic sequence of the gene ENSG00000132256, with all variation shown, in the Human GeneSeqview.

2.5 Indirect SNP detection

Four SNPs were chosen to be genotyped by indirect methods (Table 2.5.1). One of the SNPs that is common in the upstream region at position -5116, two polymorphisms in exon 2 of the *TRIM5* gene which result in non-synonymous substitutions in the RING and Coiled-coil domains of the TRIM5α protein, respectively, and a fourth polymorphism that causes an amino acid change in the SPRY domain were genotyped. All three of these domains have been implicated

in binding specificity of TRIM5α to HIV, especially the SPRY domain (Stremlau *et al* 2004). Two different methods of indirect detection were employed, namely allele specific amplification (ASA) (Zetterquist and Olerup 1992) and PCR-restriction fragment length polymorphism (PCR-RFLP) (Nomura *et al* 1991 and Mercier *et al* 1992). PCR-RFLP was used for sites where the polymorphism introduced or abolished a restriction endonuclease recognition site. ASA was used if the polymorphism did not result in either the introduction or destruction of a restriction endonuclease recognition site, as with the polymorphism at site 127, or if restriction yielded products which were too small to be resolved by agarose gel electrophoresis, as with the polymorphism at site -5116.

Table 2.5.1. Positions and locations of SNP's genotyped by two indirect methods, Allele specific amplification (ASA) and PCR-Restriction Fragment Length Polymorphism (PCR-RFLP), as well as the amino acid changes they confer.

SNP	Position	Location	Alleles	Amino acid change	Method
1	-5116	Upstream	G/A	None	ASA
2	127	Exon 2	C/T	H43Y	ASA
3	407	Exon 2	G/A	R136Q	PCR-RFLP
4	15323	Exon 8	C/T	P479L	PCR-RFLP

2.5.1 Allele specific amplification (ASA)

Allele specific amplification (ASA) makes use of two different primers specific for each allele of the polymorphism. The primers were designed so that the 3' terminal end of primer A is specific for, and therefore binds to, allele A and similarly the 3' terminal end of primer B is specific for allele B (Zetterquist and Ollerup 1992). A common primer was designed to amplify in the opposite direction, towards primers A and B (Figure 2.5.1.1). Primer sequences are given in table 2.5.1.1. If there is mismatch between the primer and the target sequence, there was no PCR amplification and obviously no product. Thus, heterozygous samples (AB) at this locus should yield a PCR product for PCR reactions performed with both Primer A and B and homozygous samples should yield a PCR product with one of the primers only.

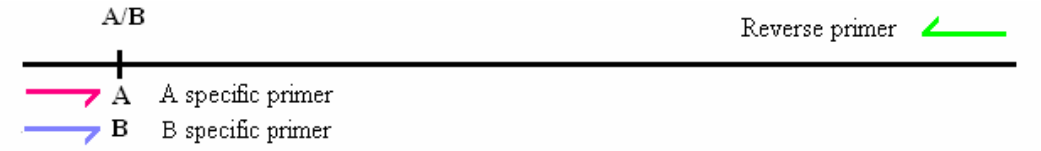


Figure 2.5.1.1. Diagram showing how primers bind to and amplify two different products based on the allele present at the 3' end of the primer binding site.

PCR reactions were carried out at a 10µl volume and contained 0.25 units of *Taq* DNA polymerase, 2mM $MgCl_2$, 0.8mM dNTPs, 1µM of primer A or primer B, 1µM reverse primer and 1µg template genomic DNA. A control PCR reaction was also carried out at a 10µl volume containing 0.25 units of *Taq* DNA polymerase, 2mM $MgCl_2$, 0.8mM dNTPs, 1µM of primer A or primer B and 1µM

reverse primer. Controls were performed in which template genomic DNA was replaced with dH_2O. This was done in order to eliminate the possibility of DNA contamination from a source other than the genomic DNA sample.

The allele specific amplification was optimized for samples that had known nucleotide sequences based on direct sequencing results. PCR amplification was carried out at an initial denaturation temperature of 94°C for two minutes, followed by 35 cycles of denaturation for 30 seconds, annealing of primers for 45 seconds at temperatures given in table 2.5.1.1 for primer pairs and extension for 60 seconds at 72°C, followed by a final extension cycle for 5 minutes at 72°C.

Table 2.5.1.1. Primer sequences and annealing temperatures for the primers used for ASA.

Site	Primers	Allele amplified	Primer sequence	Annealing temperature[1]
-5116	TRIM-5116A-f (Primer A)	A allele	5'AATACTTGGCTG GGTTAATCTA3'	54.0°C
	TRIM-5116G-f (Primer B)	G allele	5'AATACTTGGTGG GTTAATCTG3'	54.9°C
	T5upst-r (Reverse primer)		5'CAGTTTTGTCCC AAAGACTTCC3'	
127	T5ex2C-f (Primer A)	C allele	5'GCATGCCTCACT GCAAACC3'	54.7°C
	T5ex2T-f (Primer B)	T allele	5'GCATGCCTCACT GCAAACT3'	55.6°C
	T5ex2-r (Reverse primer)		5'CTGTAATTGGGT GAAATGCAAA3'	

[1] Optimized annealing temperature for specific forward primer with common reverse primer

Genotyping was performed by scoring the presence versus absence of a 436bp band on a 1% agarose gel for the polymorphic site -5116 and the presence versus absence of a 489bp band on a 1% agarose gel for polymorphic site 127, corresponding to the presence or absence of the allele. Gels were electrophoresed in TBE buffer consisting of 89mM Tris-borate and 2mM EDTA with a GeneRuler™ 1kb DNA ladder (Fermentas Life Sciences) for 45 minutes at 7.8V/cm. Gels were visualized using the UVP BioDoc-It™ system.

2.5.2 Polymerase chain reaction- Restriction fragment length polymorphism (PCR-RFLP)

PCR-restriction fragment length polymorphism (PCR-RFLP) relies on the introduction or removal of a restriction endonuclease recognition site by the replacement of a single nucleotide. PCR was first performed in order to amplify a region surrounding the SNP site and, where possible, a second control restriction site. The PCR product was then digested by the appropriate enzyme, which was determined by analysing the sequence information already obtained. The restriction products were electrophoresed on an agarose gel to determine whether the fragment had been cut by the restriction enzyme (Figure 2.5.2.1) (Nomura *et al* 1991 and Mercier *et al* 1992).

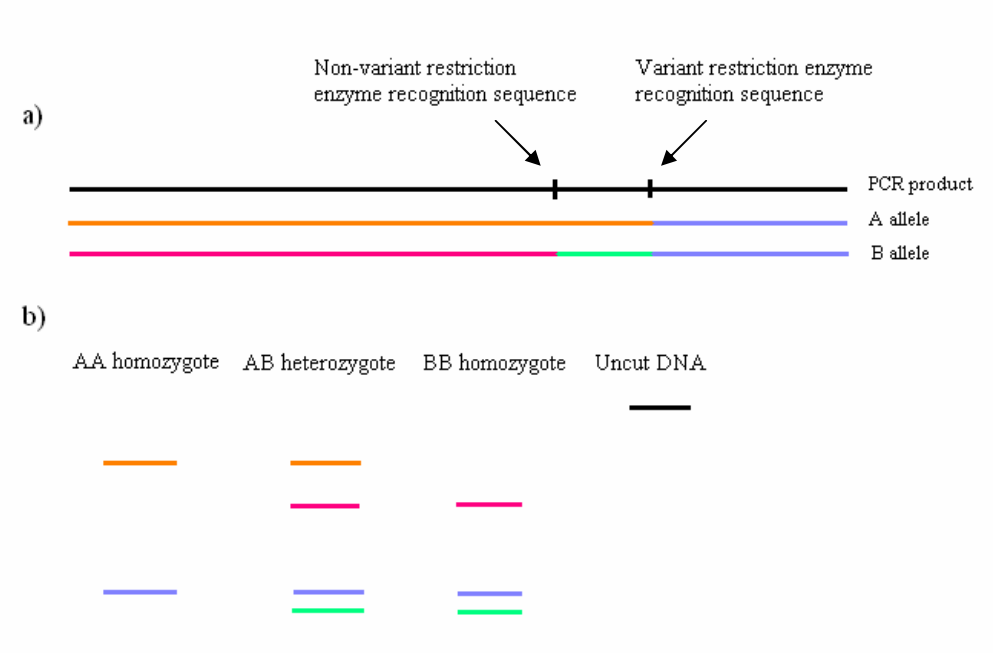


Figure 2.5.2.1. **T**he products of PCR-RFLP. a) Products of PCR. The variant restriction enzyme recognition sequence is the one introduced by the B allele SNP or abolished by the A allele SNP, whilst the non-variant restriction enzyme recognition sequence is a control site. b) Products of restriction as seen on an agarose gel.

PCR was carried out in a 20µl reaction volume. The reaction mixture contained 0.5 units *Taq* DNA polymerase, 2mM $MgCl_2$, 0.8mM dNTPs, 1µM of each of the forward and reverse primers and 2µg template genomic DNA. PCR thermocycling consisted of an initial denaturation at 94°C for 5 minutes, followed by 35 cycles of denaturation for 30 seconds at 94°C, annealing of primers for 45 seconds at various temperatures depending on GC content of primers (Table 2.5.2.1), extension for 1 minute at 72°C and a final extension for 5 minutes at 72°C. Presence of PCR product was confirmed by electrophoresing 5µl on a 1%

agarose gel in TBE buffer consisting of 89mM Tris-borate and 2mM EDTA along with the appropriate marker for determining the fragment size.

Restriction was carried out in a 30µl reaction volume. The reaction mixture contained 75-150ng PCR product, determined visually by gel electrophoresis, 7 units of restriction endonuclease and 1X appropriate buffer (Table 2.5.2.1). Restriction digests were incubated for 8 hours at the optimum temperature of the restriction endonuclease (Table 2.5.2.1). Restriction products were electrophoresed on 2% agarose gels in TBE buffer consisting of 89mM Tris base, 89mM boric acid and 2mM EDTA along with the appropriate marker for determining the fragment size (Figure 2.5.2.2).

Table 2.5.2.1. Sequences of primers and annealing temperature of reactions for PCR-RFLP. Restriction endonucleases, buffers and incubation temperatures used for restriction digests.

Primer sequence and orientation[1]	SNP	Annealing temperature	PCR product size	Restriction endonuclease[2]	Restriction buffer[2]	Incubation temperature
5'GGGATCTGTGAACAAGAGGAAC3' (Fw) 5'CTGTAATTGGGTGAAATGCAAA3' (Rv)	127	54.9°C	659bp	*Sma*I	Buffer Tango	30°C
5'GACAGTGGCTCCAAACAACAT3' (Fw) 5'AGGGGCTGAGTGTGTAAGAAGG3' (Rv)	15323	57.6°C	603bp	*FspB*I	Buffer Tango	37°C

[1] Primers labelled Fw are forward primers and Rv are reverse primers
[2] All restriction endonucleases and buffers provided by Fermentas Life Sciences

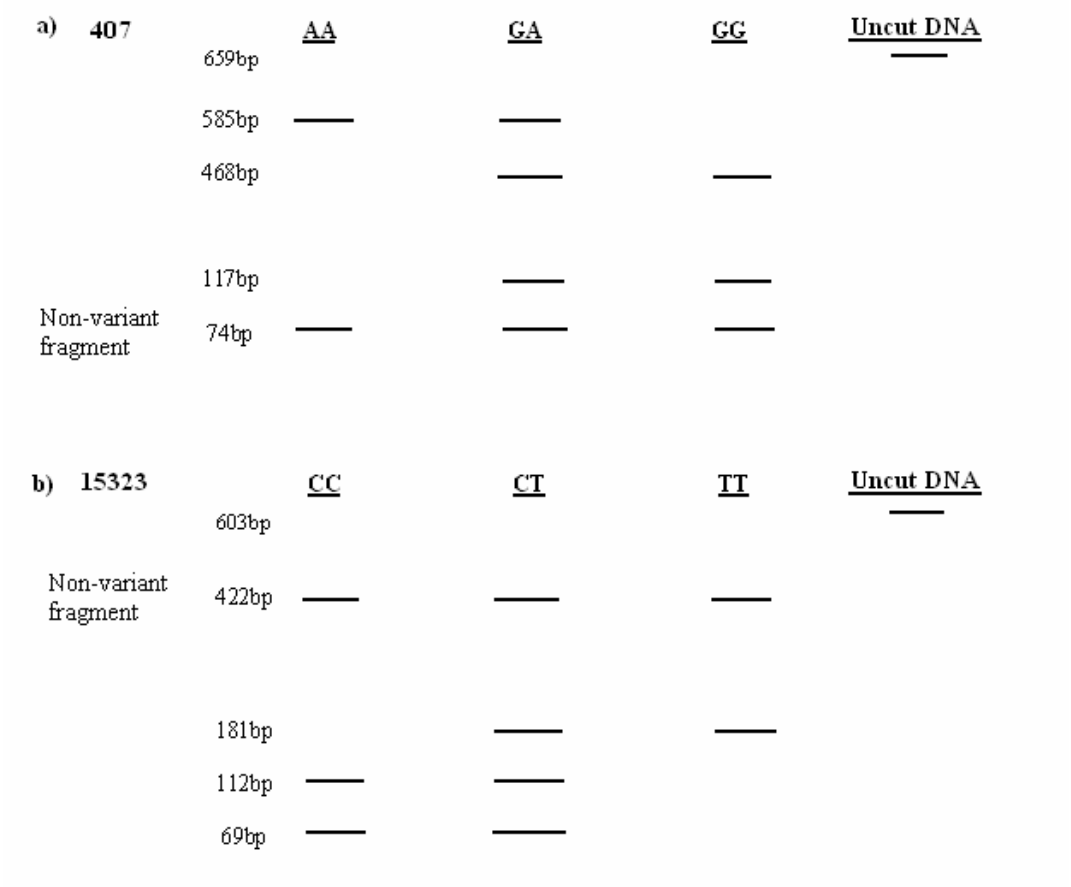


Figure 2.5.2.2. Sizes of restriction fragments obtained for various genotypes for each of the SNP's genotyped by PCR-RFLP. a) Size of fragments obtained by restriction of 659bp fragment for genotyping at site 407 and b) Size of fragments obtained by restriction of 603bp fragment at site 15323.

2.6 Data Analysis

2.6.1 Allele and genotype frequency determination

Genotype and allele frequencies were determined by counting the number of alleles and genotypes and calculating the frequency of each allele type and genotype in the population.

In a population with the sample size given by n, if a locus has two alleles, namely A and a, the frequency of the A allele is p and the a allele is q. The possible genotypes at this locus are AA, Aa and aa. If we let the number of individuals carrying each of the genotypes be P, Q and R respectively, then the allele frequencies, p and q, can be calculated as follows:

$$p = \frac{2P + Q}{2n}$$

$$q = \frac{2R + Q}{2n}$$

where, $p + q = 1$

Observed genotypic frequencies for AA, Aa and aa are calculated as follows:

$$\text{Frequency of } AA = \frac{P}{n}$$

$$\text{Frequency of } Aa = \frac{Q}{n}$$

$$\text{Frequency of } aa = \frac{R}{n}$$

2.6.2 Calculation of Hardy-Weinberg equilibrium

The observed genotype frequencies were compared to the expected frequencies according to the Hardy-Weinberg principle. This was done in order to determine whether the allele and genotype frequencies deviate from expected proportions of the Hardy-Weinberg principle for the other analyses. The Hardy-Weinberg principle states that, if an infinitely large population is mating randomly, with all the members of the population breeding and producing the same number of offspring and no occurrence of mutation, natural selection or migration in or out of the population:

a locus with alleles *A* and *a* at frequencies *p* and *q*, should give the following genotypic frequencies: $f(AA) = p^2$, $f(Aa) = 2pq$ and $f(aa) = q^2$ where $p^2 + 2pq + q^2 = (p + q)^2 = (1)^2 = 1$ (Falconer and Mackay 1996 and Crow 1986).

Once the allelic frequencies *p* and *q* were determined the expected genotypic frequencies were calculated using the Hardy-Weinberg equation above. Deviations of observed genotypic frequencies from expected genotypic frequencies according to Hardy-Weinberg equilibrium were determined using the chi-squared (χ^2) test:

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

where o = observed number and e = expected number

If the χ^2 value showed significant deviation from zero (P value < 0.05) this was considered significant evidence of deviation from Hardy-Weinberg proportions.

2.6.3 HIV-1 susceptibility association

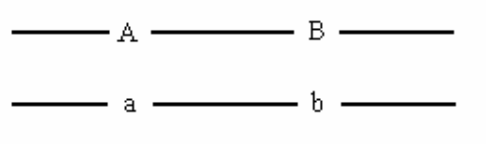
Genotype numbers were compared, for each genotype, between HIV-positive and general population samples to determine whether there were significant differences between these two groups in their susceptibility to HIV-1 infection. The comparison was done using a Chi-square test with the software tool SAS.

2.6.4 Calculation of linkage disequilibrium

Genes that are non-randomly associated are said to be in linkage disequilibrium (LD); this occurs when alleles at adjacent loci are found occurring more frequently together than would be predicted under random segregation. Pairwise linkage disequilibrium was calculated with the use of the computer software tool Linkage Disequilibrium Analyzer (LDA) version 1.0 (Ding *et al*, 2003). Linkage disequilibrium was calculated separately for the polymorphisms detected in the upstream region by sequencing and for the polymorphisms genotyped by indirect methods.

The most simple measure of linkage disequilibrium is denoted by the symbol *D* and is the product of the frequencies of the alleles segregating randomly subtracted from the observed frequency of the haplotype with the two alleles in

question (Lewontin 1964). More simply put, consider two loci (A and B) on the same chromosome with alleles A, a and B, b:



Thus, the possible gametic genotypes produced by this individual are AB, Ab, aB and ab. If the haplotype in question is that consisting of the alleles A and B (P_{AB}). The linkage disequilibrium for the gametic type AB is given by the equation:

$D = P_{AB} - P_A \times P_B$

Where P_{AB} is the frequency of the gamete with alleles A and B. P_A is the frequency of allele A at the one locus and P_B is the frequency of allele B at the other locus. If $D > 0$, allele A is associated with allele B, if $D = 0$, alleles are randomly associated and if $D < 0$, alleles A and b are associated and alleles a and B are associated.

Quantitatively, this statistic is not very accurate as it depends on allele frequencies (Jorde 2000). Thus a more useful measure is $|D'|$. This value is obtained by the equation:

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

where D_{max} is the maximum value D can be, given the values of the allele frequencies (Lewontin 1964). A value of 1 for $|D'|$ indicates complete LD between the two alleles. Disruption of ancestral LD results in a value less than 1 for $|D'|$. The allele frequency also has no effect on the measurement of LD using

this statistic (Lewontin 1964, Jorde 2000 and Weiss and Clark 2002), however the |*D'*| measure is inflated at low (<50) sample sizes or in cases where one allele frequency is extreme (Teare *et al* 2002).

Another measure of linkage disequilibrium is the square of the correlation coefficient between loci A and B, or r^2, where:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}$$

Under selectively neutral evolution, at equilibrium, the value of r^2 can be calculated using the simple formula:

$$r^2 = \frac{1}{(4Nc + 1)}$$

where N is the population size and c is the recombination rate per nucleotide. This measure of linkage disequilibrium is more dependant on allele frequency and usually gives a lower value for LD than |*D'*| does (Jorde 2000 and Weiss and Clark 2002), however it is not sensitive to sample size. The linkage disequilibrium between polymorphism pairs was measured using both the |*D'*| measure, and the r^2 statistic as both of these measures are appropriate measures of LD and under many circumstances are almost identical (Devlin and Risch 1995).

2.6.5 Haplotype analysis

Adjacent alleles occuring together within blocks of linkage disequilibrium are referred to as haplotypes (Reich *et al* 2001). Alternatively put, haplotypes are sets of alleles or markers proximal to one another on a chromosome that display a

tendency to be transmitted jointly (Tishkoff and Williams 2002). Haplotypes and haplotype frequencies were estimated using the computer software tool PHASE version 2.1 that makes use of Gibbs sampling for phase reconstruction (Stephens *et al* 2001).

Haplotypes and haplotype frequencies were estimated separately for the samples genotyped by sequencing as there was information on the genotypes at eight sites, and the samples for which there was only genotype information for the four sites genotyped by indirect methods. There was a sample size of 38 for samples sequenced for the upstream region and 170 for those genotyped by indirect methods. Only samples with one or no missing data points were included in the haplotype analysis.

In addition haplotypes, resolved to a probability difference of >0.15 for the haplotype pair, were compared between HIV-positive and general population samples. There was a sample size of $n = 94$ for the HIV-positive samples and $n = 33$ for the general population samples.

CHAPTER 3

RESULTS

3.1 DNA Isolation

Genomic DNA was successfully extracted from the buffy coat of centrifuged blood (Figure 3.1.1). Different concentrations of DNA, based on different yields from the blood samples, were reflected in the varying intensities of the bands on a 0.8% agarose gel. High molecular weight DNA obtained by extraction from the buffy coat of the blood samples was observed on the low percentage gels. There was little evidence of degradation of DNA or the presence of RNA. Differences in the intensity of the bands were either due to lower yield because of the efficiency of extraction or differences in the $CD4^+$ count in the blood samples.

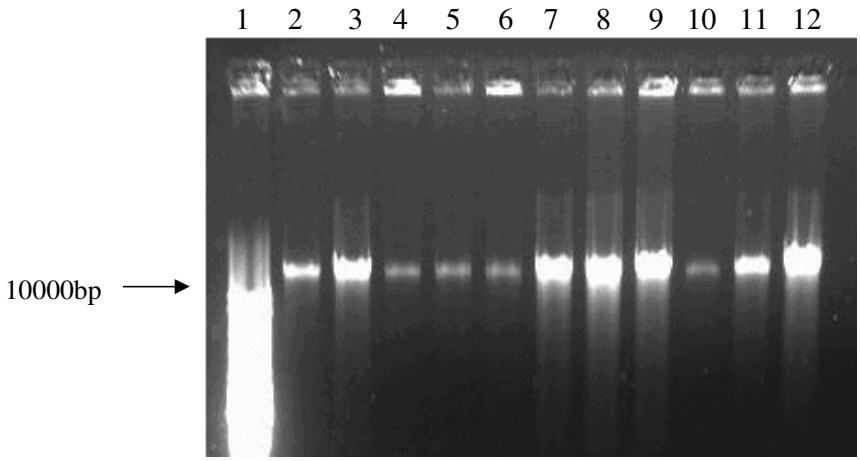


Figure 3.1.1. Extracted genomic DNA used in sequencing and genotype analysis. Genomic DNA was separated on a 0.8% agarose gel. Lane 1 contains a 1Kb DNA ladder and lanes 2-12 each contain 5µl of genomic DNA solution, extracted from 11 blood samples.

3.2 Direct sequencing

All sequence traces of polymorphisms detected by directed sequencing are given in Appendix 3 and the genotypes of all samples genotyped by direct sequencing are given in Appendix 4.

3.2.1 Upstream region

Sequencing of the upstream region 639 bp PCR product which was obtained by amplification using primers T5upst-f and T5upst-r (Figure 3.2.1.1) resulted in 432bp of good sequence with well defined signal peaks for 19 samples. Five polymorphisms (Figure 3.2.1.2) were detected between -5253bp and -4821bp upstream of the start of translation (in exon 2). These include four polymorphisms that have previously been detected and can be found in the NCBI dbSNP (reference numbers in Table 3.2.1.1) and one novel polymorphism. The previously detected polymorphisms are at sites -5116, -4998, -4904 and -4876 and the novel polymorphism is at site -4819 (Figure 3.2.1.2)

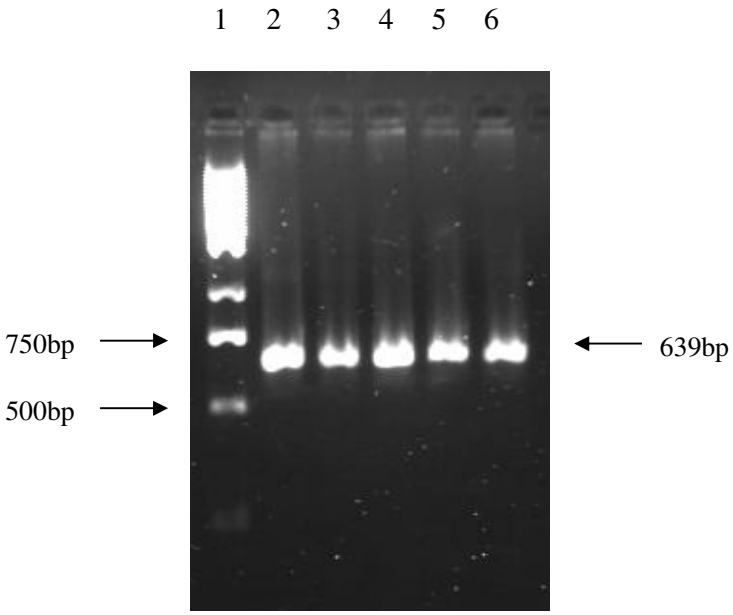


Figure 3.2.1.1. PCR product of the upstream region of the *TRIM5* gene. The 639bp PCR product of the upstream region of the *TRIM5* gene obtained by amplification using primers T5upst-f and T5upst-r was separated on a 1% agarose gel. Lane 1 contains a 1Kb DNA ladder. Lanes 2-6 contain the 639bp fragment obtained by PCR amplification with primers T5upst-f and T5upst-r.

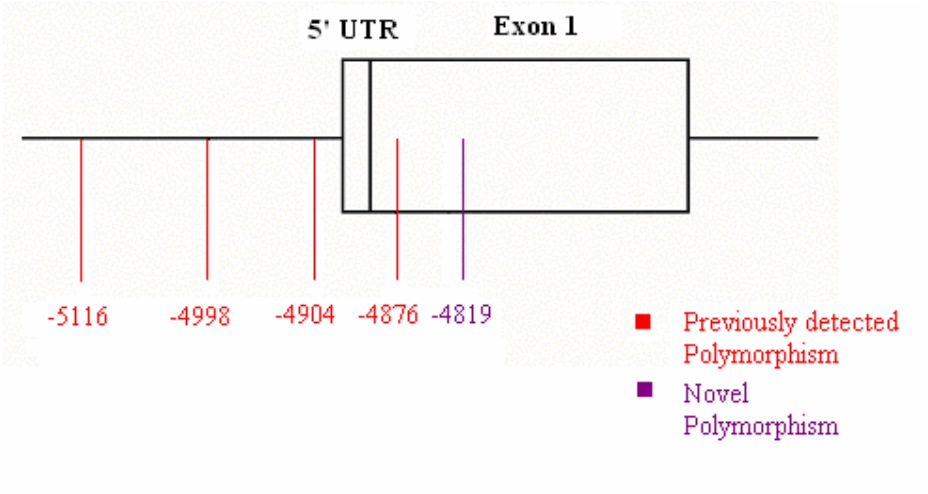


Figure 3.2.1.2. Positions of polymorphisms identified by sequencing the upstream region of the *TRIM5* gene. Five polymorphisms were identified by sequencing the upstream region of the *TRIM5* gene. This region includes Exon 1, which is not incorporated in the alternatively spliced α variant of the *TRIM5* protein, as well as the 5' untranslated region (UTR) and approximately 430bp upstream of the 5' UTR. The polymorphisms at sites -5116, -4998, -4904 and -4876 were in the NCBI dbSNP database and the polymorphism at site -4819 was a previously unrecorded polymorphism.

The polymorphisms at sites -5116 and -4904 were found to be common (Table 3.2.1.1) while the polymorphisms at sites -4998, -4876 and -4819 were rare, with no homozygotes for the minor allele and few heterozygous samples.

Table 3.2.1.1. SNPs found by sequencing a 639 bp fragment of the upstream region of the *TRIM5* gene.

Site	Alleles	# samples	Genotype numbers			SNP reference[1]
			GG	**GA**	**AA**	
-5116	G/A	19	6	3	10	rs3802981
			GG	**GC**	**CC**	
-4998	G/C	19	17	2	0	rs16934387
			CC	**CT**	**TT**	
-4904	C/T	19	11	6	2	rs3802980
			TT	**TC**	**CC**	
-4876	T/C	19	14	5	0	rs16934386
			GG	**GA**	**AA**	
-4819	G/A	19	16	3	0	-

[1] SNP reference from NCBI dbSNP

Due to the large number of polymorphisms detected in this region, genotyping was performed by sequencing an additional 26 samples for an upstream region of 714bp amplified using primers T5upst2-f and T5upst-r (Figure 3.2.1.3). This was done in order to be able to analyse these polymorphisms with respect to the presence or absence of linkage disequilibrium. Good sequence, with well defined signal peaks, was obtained for 416bp spanning from position -5276bp to -4860bp. No additional SNPs were observed in these sequences. Allele and genotype frequencies obtained by direct detection of variation are given in Table 3.4.1.

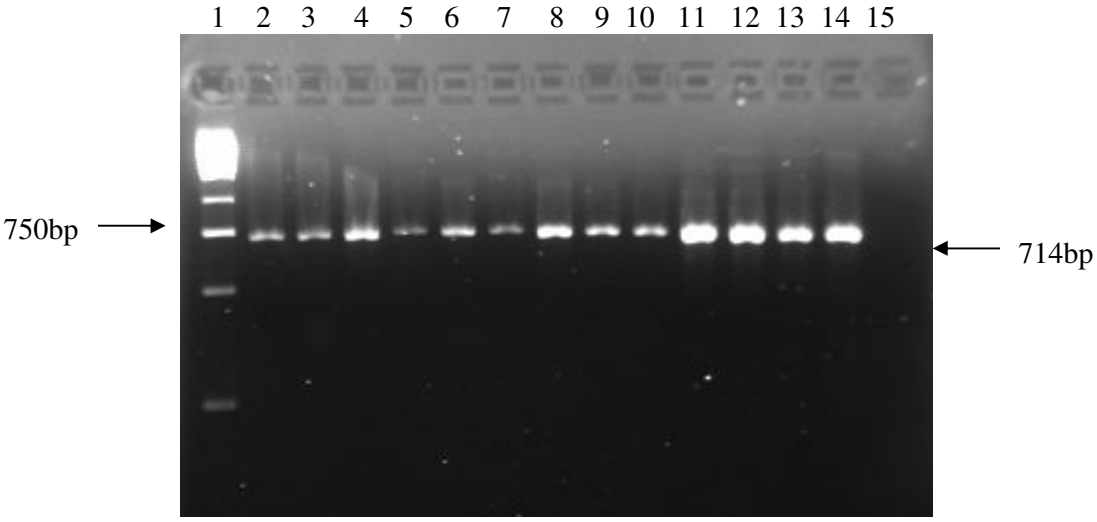


Figure 3.2.1.3. PCR product of the upstream region of the *TRIM5* gene amplified. The 714bp PCR product of the upstream region of the *TRIM5* gene obtained by amplification using primers T5upst2-f and T5upst-r was separated on a 1% agarose gel. Lane 1 contains a 1Kb DNA ladder. Lanes 2-14 contain the 714bp fragment obtained by PCR amplification with primers T5upst2-f and T5upst-r. Lane 1 contains a 1Kb DNA ladder. Lane 15 contains a control in which template DNA was replaced with dH_2O.

3.2.2. Exon 2

Sequencing of 21 samples for the 659 bp PCR product spanning exon 2 (Figure 3.2.2.1) gave 545bp of sequence with well resolved signal peaks. Analysis of this sequence revealed 3 polymorphisms between the region spanning from base pair position -47 to 498 in the gene.

Known polymorphism at site -2 was found to be common (Table 3.2.2.1), while the known polymorphisms at sites 127 and 407 were rare with no homozygotes for the minor allele and only three out of 21 samples with both alleles.

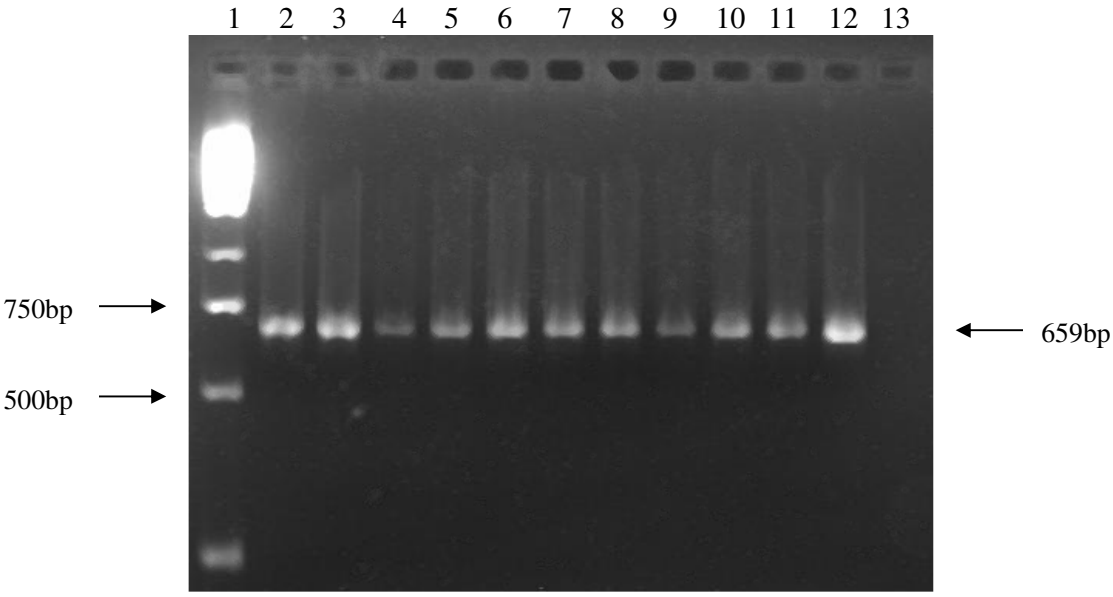


Figure 3.2.2.1. PCR product of the exon 2 region of the *TRIM5* gene. The 659bp PCR products of the exon 2 region of the *TRIM5* gene were obtained by amplification using primers T5ex2-f and T5ex2-r and were separated on a 1% agarose gel. Lane 1 contains a 1Kb DNA ladder. Lanes 2-12 contain the PCR product of amplification using primers T5ex2-f and T5ex2-r. Lane 13 contains a control in which template DNA was replaced with dH_2O.

Table 3.2.2.1. SNPs found by sequencing a 659bp fragment of the exon 2 region of the *TRIM5* gene.

Site	Alleles	# samples	Genotype numbers			SNP reference[1]
			CC	**CG**	**GG**	
-2	C/G	20	12	4	4	rs3824949
			CC	**CT**	**TT**	
127	C/T	21	18	3	0	rs3740996
			GG	**GA**	**AA**	
407	G/A	21	18	3	0	rs10838525

[1] SNP reference number from the Ensembl SNP database

3.2.3 Exon 7

Sequencing of 17 samples for the 713 bp PCR product, obtained using primers T5ex67-f and T5ex67-r, spanning exon 7, intron 7 and part of exon 8 (Figure 3.2.3.1) revealed no variation. Well defined peaks were obtained for the 470bp region spanning from base pair position 14223 to 14693.

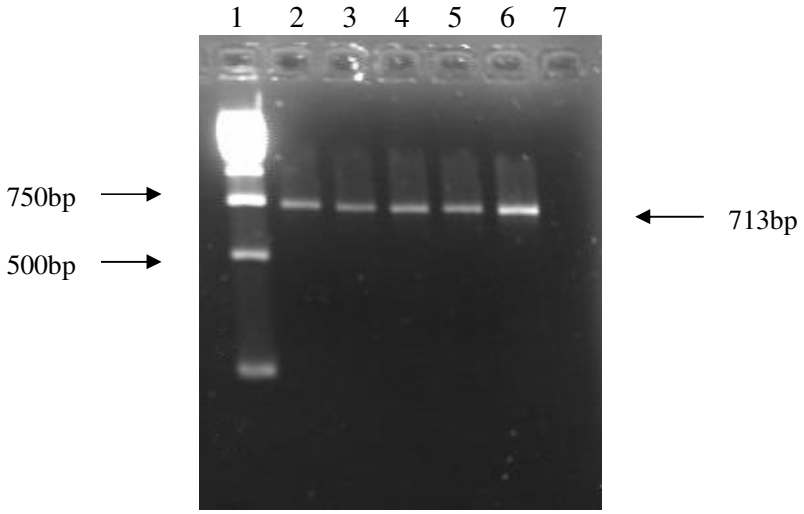


Figure 3.2.3.1. PCR product of the exon 7 region of the *TRIM5* gene. The 713bp PCR products of the exon 7 region of the *TRIM5* gene were obtained by amplification using primers T5ex67-f and T5ex67-r and were separated on a 1% agarose gel. Lane 1 contains a 1Kb DNA ladder. Lanes 2-6 contain the products of PCR amplification by primers T5ex67-f and T5ex67-r. Lane 7 contains a control in which template DNA was replaced with dH_2O.

3.2.4 Exon 8

Sequencing of the 603bp PCR product for the region of exon 8, excluding the 3'UTR (Figure 3.2.4.1) gave 418bp of sequence with well defined peaks between base pairs 14906 and 15324 in the *TRIM5* gene. There was 1 polymorphism at site 15323 in 21 samples. The polymorphism has alleles C and T and is rare, with 4 heterozygotes in 21 samples and 17 homozygotes for the C allele. No homozygotes for the T allele were observed. The polymorphism has the identity rs7104422 in the NCBI dbSNP.

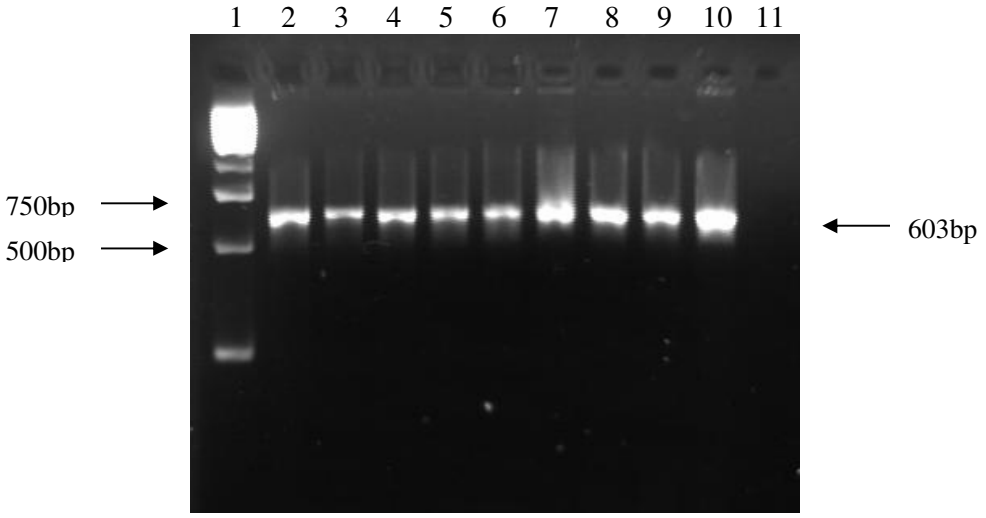


Figure 3.2.4.1. PCR product of the exon 8 region of the *TRIM5* gene. The 713bp PCR products obtained by amplification using primers T5ex8-f and T5ex8-r, which span exon 8, excluding the 5' untranslated region, were separated on a 1% agarose gel. Lane 1 contains a 1Kb DNA ladder. Lanes 2-10 contain the PCR products of amplification by primers T5ex8-f and T5ex8-r. Lane 11 contains a control in which template DNA was replaced with dH_2O.

3.3 Indirect SNP detection

PCR-RFLP and allele specific amplification were performed in order to detect the four SNPs at sites -5116, 127, 407 and 15323, which were found by sequencing. All genotypes of samples at these sites are given in Appendix 5.

3.3.1 Polymorphic site -5116

Allele specific amplification was performed on all 222 samples in order to detect alleles at polymorphic site -5116. PCR with primer pair TRIM-5116A-f and T5upst-r yielded a 436bp product if the A allele was present at this site. PCR with

primer pair TRIM-5116G-f and T5upst-r yielded a 436bp product if the G allele was present (Figure 3.3.1.1).

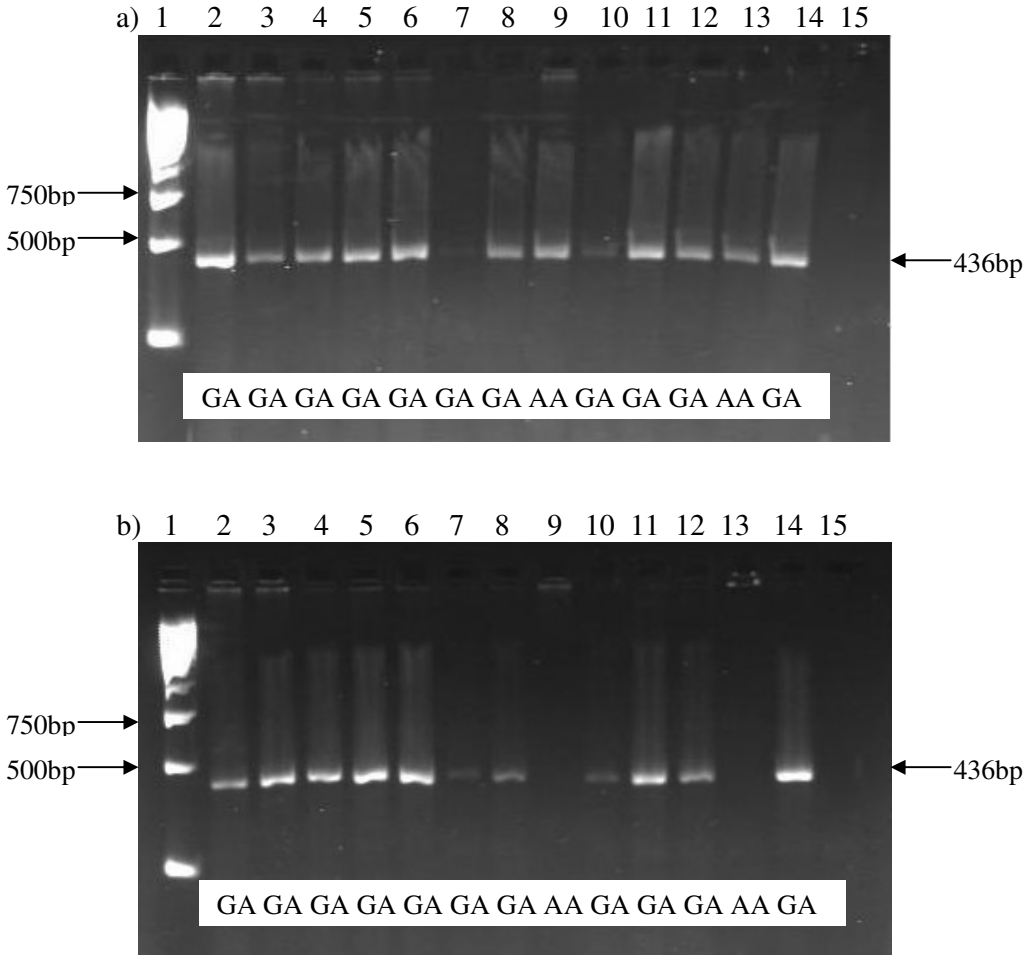


Figure 3.3.1.1. The products of allele specific amplification at polymorphic site -5116. An example of the product of allele specific amplification of a 436bp region upstream of the *TRIM5* gene, amplified using the following primer pairs: a) TRIM-5116A-f and T5upst-r, specific for the A allele and b) TRIM-5116G-f and T5upst-r, specific to the G allele. Lane 1 in both gels contains a 1kb DNA ladder. Lanes 2-14 in both gels contain the same samples which were amplified by the different primer pairs and lane 15 contains no-template controls for the different primer pairs used. Lanes 2-8 as well as lanes 10, 11, 12 and 14 show similar intensity bands for both primer pairs, therefore these samples are heterozygotes. Lanes 9 and 13 have PCR product for the primer pair used in a) but not that used in b), so they represent AA homozygotes.

3.3.2 Polymorphic site 127

Allele specific amplification was used to detect the polymorphism at site 127. If the C allele was present, amplification with the C specific primer (T5ex1C-f) and the reverse primer, T5ex1-r, yielded a 489bp band on the gel. Similarly the presence of the T allele was determined by the presence of a 489bp band on the gel upon amplification with the T specific primer (T5ex1T-f) and the reverse primer (Figure 3.3.2.1).

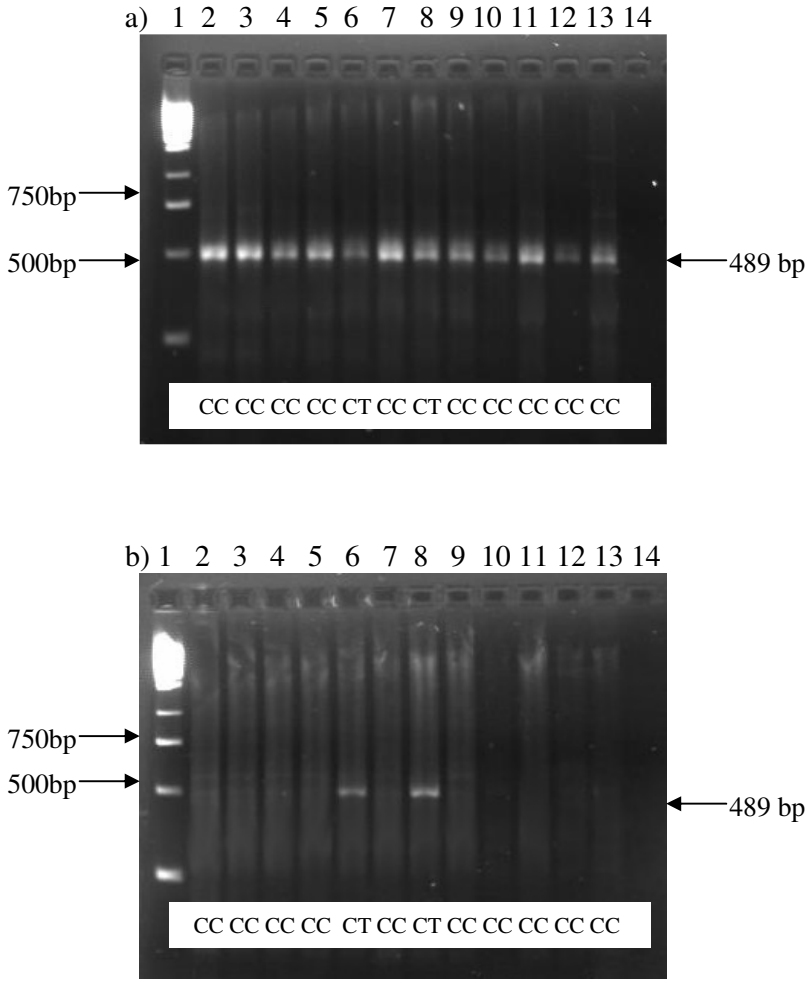


Figure 3.3.2.1. The products of allele specific amplification at polymorphic site 127. An example of the product of allele specific amplification of a 489bp region of exon1 in the *TRIM5* gene amplified using primer pairs a) T5ex1C-f and T5ex1-r specific to the C allele and b) T5ex1T-f and T5ex1-r specific to the T allele. Lane 1 in both gels contains a 1kb DNA ladder. Lanes 2-13 in both gels contain the same samples which were amplified by the different primer pairs and lane 14 in both gels contains no-template controls for the different primer pairs used. Samples in lanes 2-13 in gel a) all have the C allele and samples in lanes 6 and 8 in gel b) have the T allele at this site. Therefore the samples in lanes 6 and 8 in both gels are CT heterozygotes and the samples in lanes 2, 3, 4, 5, 7, 9, 10, 11, 12 and 13 are CC homozygotes.

3.3.3 Polymorphic site 407

To detect alleles at polymorphic site 407 PCR-RFLP was performed. PCR yielded a 659bp product (Figure 3.3.3.1). Restriction of the PCR product with restriction endonuclease *Sma*I resulted in different size fragments present on a gel based on the alleles present in the sample at this site. The presence of the A allele at this site resulted in 585bp and 74bp fragments and the presence of the G allele resulted in 468bp, 117bp and 74bp fragments (Figure 3.3.3.2).

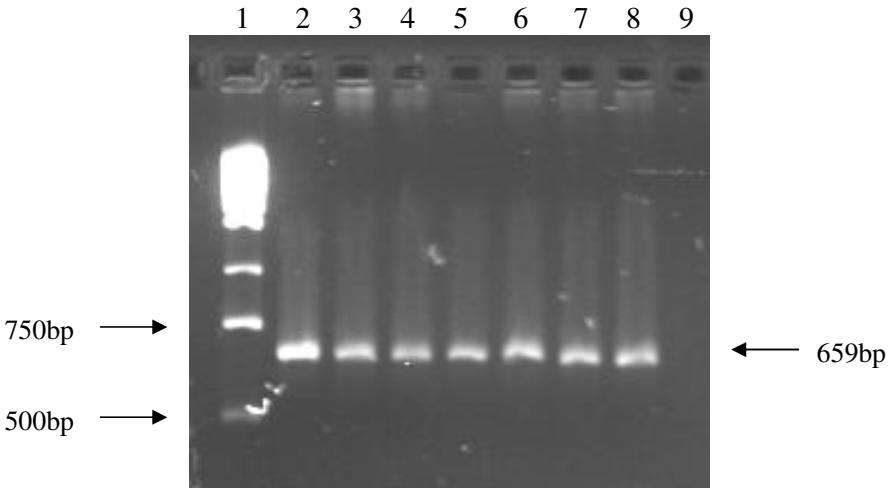


Figure 3.3.3.1. A 1% agarose gel of the 659bp PCR product. Lane 1 contains a 1Kb DNA ladder. Lane 9 contains a control in which template DNA was replaced with dH_2O.

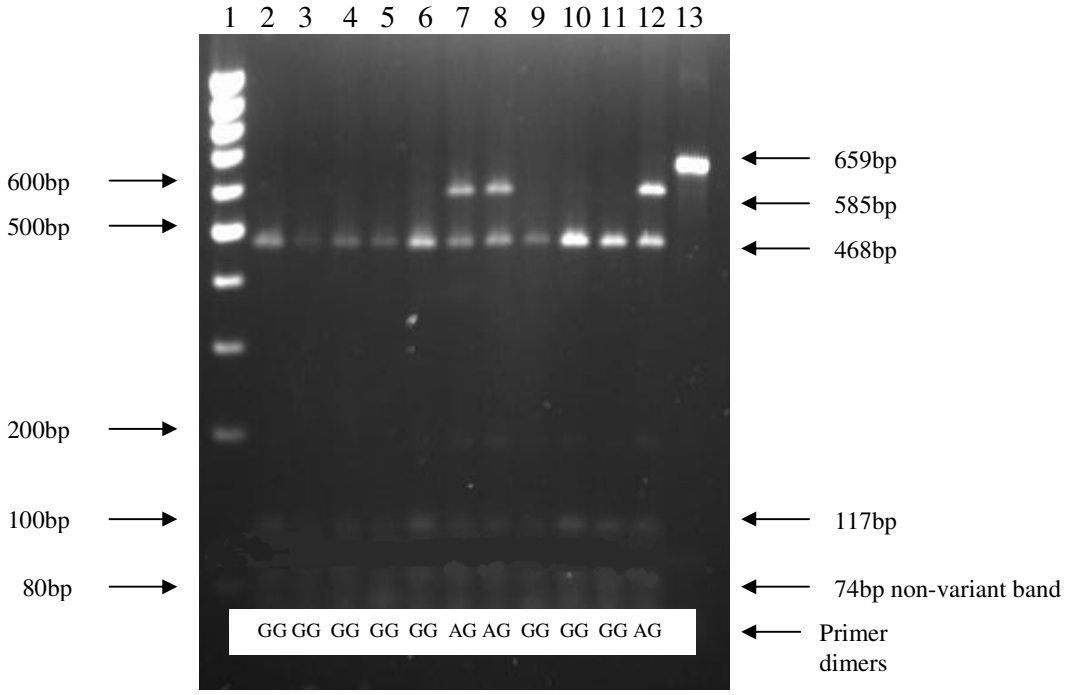


Figure 3.3.3.2. The restriction products of digestion by *Sma*I. The restriction products of digestion of the 659bp PCR product by restriction endonuclease *Sma*I were resolved on a 1% agarose gel. Lane 1 contains a low range DNA ladder. The samples in lanes 2-6 and 9, 10 and 11 only have the 468 bp band. The samples in lanes 7, 8 and 12 have 468bp and 585bp fragments. Lane 13 contains an undigested 659bp PCR product control.

3.3.4 Polymorphic site 15323

PCR-RFLP was performed on all samples to detect the SNP at site 15323, which has alleles C and T. PCR resulted in a 603bp product (Figure 3.3.4.1). Restriction of the PCR product with *FspB*I resulted in different size fragments present on an agarose gel based on the alleles present at this site. The presence of the C allele at this site resulted in 422bp, 112bp and 69bp fragments and the presence of the T allele resulted in 422bp and 181bp fragments (Figure3.3.4.2).

The 422bp fragment is the non-variant fragment and is present on all gels, indicating complete digestion of the PCR product by the restriction endonuclease.

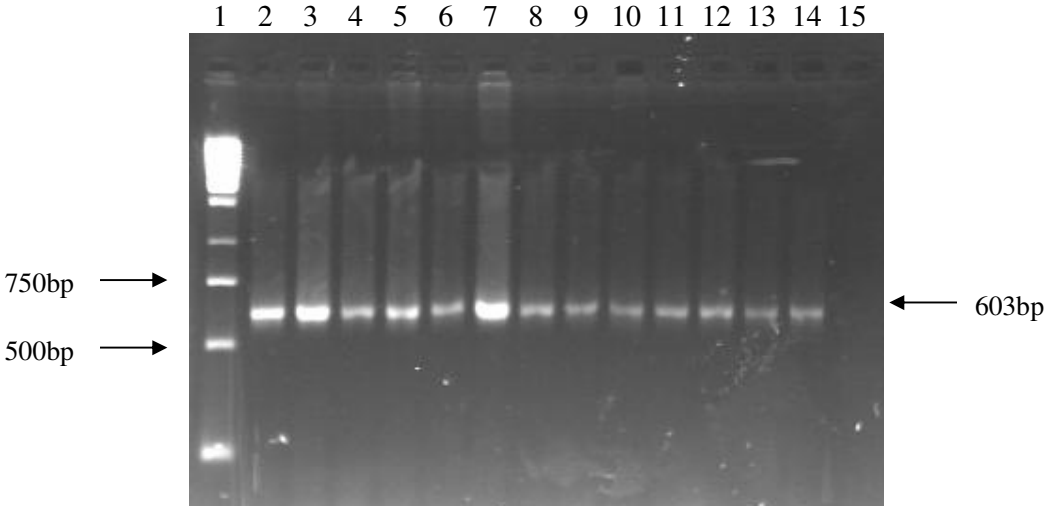


Figure 3.3.4.1. A 1% agarose gel of the 603bp PCR product of exon 8. Lane 1 contains a 1Kb DNA ladder. Lane15 contains a control in which template DNA was replaced with dH_2O.

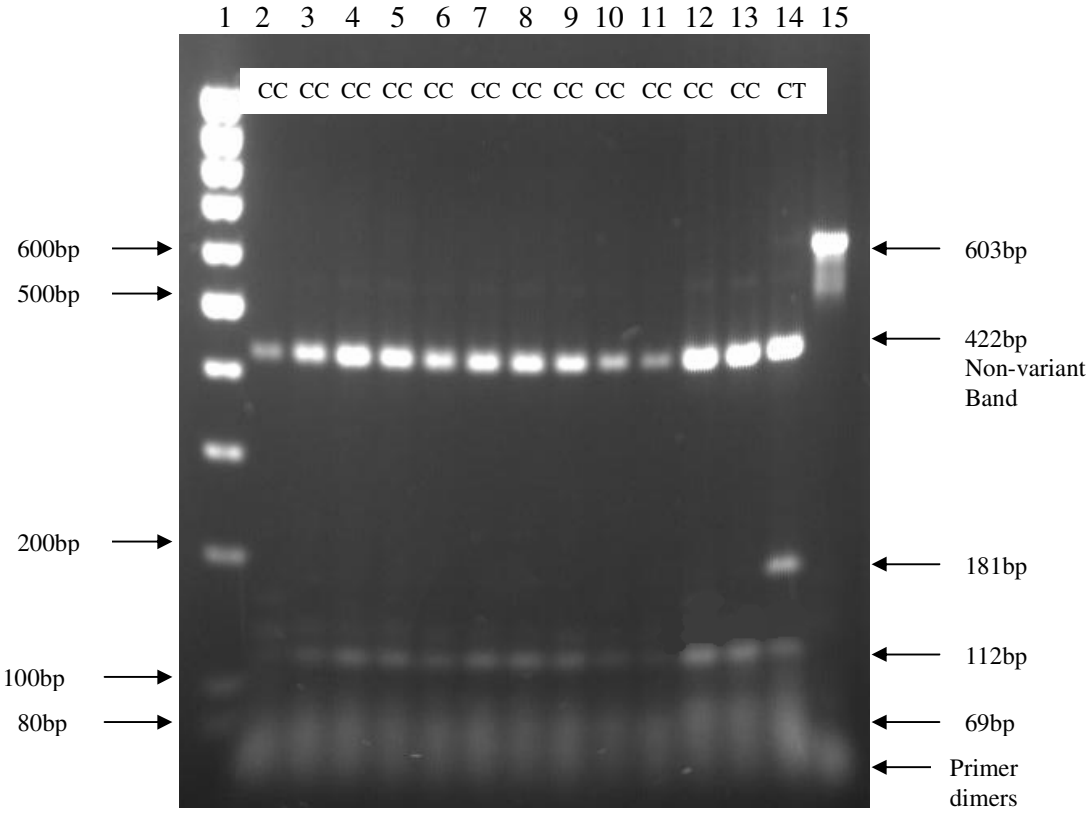


Figure 3.3.4.2. The restriction products of digestion by *FspB*I. A 1% agarose gel showing the products of restriction of the 603bp fragment containing polymorphic site 15323 by restriction endonuclease *FspB*I. Lane 1 contains a low range DNA ladder. The samples in lanes 2-13 have the 422bp, 112bp and 69bp bands and thus represent samples with the CC genotype at this site. The sample in lane 14 has 422bp, 181bp, 112pb and 69bp fragments and therefore contains both C and T alleles. It should however be noted that the 69bp fragments are not easily distinguishable from the primer bands. Lane 15 contains an undigested 603bp control.

3.4 Allele frequencies and Hardy-Weinberg equilibrium

Allele frequencies were determined separately for HIV^+ samples, general population samples and the entire study population. χ^2 and P-values were calculated and no significant deviation from Hardy-Weinberg equilibrium was observed at any of the sites (Table 3.4.1). There was, however, an overall low frequency of homozygotes for the minor allele at sites 407 and 15323.

Table 3.4.1. Results of genotyping of all polymorphic sites, by direct and indirect methods, showing the observed genotype numbers, allele frequencies and χ^2 values for a fit to Hardy-Weinberg equilibrium.

Site	Alleles	Group	# samples	Observed genotype numbers			Allele frequency	χ^2 value (1df)	P-value
				GG	**GA**	**AA**			
-5116	A/G	HIV$^+$	150	32	87	31	0.497	2.5991	0.1069
		Gen	39	9	15	15	0.577	3.8425	0.0500
		Total	189	41	102	46	0.513	0.9224	0.3368
				GG	**GC**	**CC**			
-4998	G/C	**Total**	38	31	7	0	0.908	0.3911	0.5317
				CC	**CT**	**TT**			
-4904	C/T	**Total**	37	22	10	5	0.730	3.667	0.0555
				TT	**TC**	**CC**			
-4876	C/T	**Total**	38	30	6	2	0.868	3.6304	0.0567
				GG	**GA**	**AA**			
-4819	G/A	**Total**	35	30	5	0	0.929	0.2071	0.649
				CC	**CT**	**TT**			
127	C/T	HIV$^+$	155	102	51	2	0.823	2.5107	0.1131
		Gen	40	32	6	0	0.875	0.2630	0.6081
		Total	195	134	57	2	0.833	2.2610	0.1327
				GG	**GA**	**AA**			
407	G/A	HIV$^+$	129	112	16	1	0.930	0.2548	0.6137
		Gen	38	34	4	0	0.947	0.1173	0.7320
		Total	167	146	20	1	0.934	0.1200	0.7290
				CC	**CT**	**TT**			
15323	C/T	HIV$^+$	126	107	19	0	0.925	0.8378	0.3600
		Gen	33	28	5	0	0.924	0.2217	0.6377
		Total	159	135	24	0	0.925	1.0596	0.3033

3.5 HIV-1 susceptibility association

A χ^2 test was used to determine whether there are significant differences between the genotype numbers of the HIV-positive samples and the general population samples. The test was performed under a dominant model and a recessive model in order to compensate for the relatively low frequencies of homozygotes at some of the sites. The dominant model groups the two genotypes containing the major allele together before the comparison is made between the HIV-positive and general population samples. The recessive model groups the two genotypes containing the minor allele together before comparing the HIV-positive and general population samples. The test was performed using a co-dominant model for the polymorphism at site 15323 as there are only two classes of genotype information for this site. The co-dominant model compares all three genotypes between HIV-positive and general population samples.

Table-wide significance was determined according to the sequential Bonferroni technique (Rice 1989). A P value below 0.005 was taken to indicate significant differences between HIV-positive and general population genotype numbers for the 7 comparisons made in table 3.5.1. Significant deviation was not observed between the HIV^+ and general population genotype numbers at any of the sites, -5116, 127, 407 and 15323, under any of the models.

Table 3.5.1. Genotype number differences between HIV$^+$ samples and general population samples as determined by χ^2 analysis.

Site	Genotype numbers					Co-dominant model		Dominant model		Recessive model	
						χ^2 value[3]	P-value	χ^2 value[4]	P-value	χ^2 value[4]	P-value
-5116		**GG**	**GA**	**AA**	**Total**						
	HIV$^+$	49	75	23	**147**	-	-	4.4001	0.0359	1.035	0.309
	Gen	8	15	16	**39**						
127		**CC**	**CT**	**TT**	**Total**						
	HIV$^+$	103	53	2	**158**	-	-	0.5115	0.4745	5.876	0.0153
	Gen	34	6	0	**40**						
407		**GG**	**GA**	**AA**	**Total**						
	HIV$^+$	113	17	1	**131**	-	-	0.292	0.5891	0.2687	0.6042
	Gen	34	4	0	**38**						
15323[5]		CC	CT	**Total**							
	HIV$^+$	109	19		**128**	0.002	0.9647	-	-	-	-
	Gen	28	5		**33**						

[1] Dominant model refers to G-dominant at site -5116
[2] Recessive model refers to A-dominant at site -5116
[3] χ^2 value determined to 1 degrees of freedom for site 15323
[4] χ^2 value determined to 1 degree of freedom for sites -5116, 127 and 407
[5] There were no homozygotes for the minor allele at site 15323 for the general population samples or the HIV$^+$ samples

3.6 Linkage disequilibrium analysis

Pairwise linkage disequilibrium (LD) was calculated for samples that had genotype information at all eight polymorphic sites (n = 38) and separately for all the samples at the four sites genotyped by indirect methods (n = 175), as there is a larger sample size with genotype information at these four sites. In addition, pairwise linkage disequilibrium was determined for HIV-positive samples only (n = 137). Table-wide significance was determined for all linkage disequilibrium analyses using the sequential Bonferroni technique (Rice 1989).

Linkage disequilibrium analysis of the alleles at all eight sites showed nonrandom association as can be seen by the high r^2 and |D'| values (Table 3.6.1). P values below 0.002 were taken to reflect significant linkage disequilibrium. However, the high |D'| values observed, for example |D'| values of 1, are most likely due to the small sample size of 38 and the low frequency of certain alleles, rather than a true reflection of linkage disequilibrium. This is because a |D'| value of 1 means that only 3 of the possible four combinations are present in the population. |D'| also tends to be inflated at low sample sizes (Teare *et al* 2002) so the r^2 statistic is a better determinant of LD for the samples with information at all 8 polymorphic sites, as this subset has a small sample size.

|D'| values are significantly different from zero for the polymorphism at site -5116 and all the other upstream polymorphisms except the polymorphism at site -4876 (Table 3.6.1). This may be due to their close proximity to one another as these polymorphisms can all be found within 300bp of each other. Significant linkage

disequilibrium was also observed between polymorphisms at site -4998 and site 127 as well as between polymorphisms at sites -4998 and 15323. It may be that there is significant linkage disequilibrium between these polymorphism pairs, however this significance should be viewed with caution due to the small sample sizes. The same can be said for the significance of the linkage disequilibrium observed between polymorphism pairs at sites -4904 and 15323 (Table 3.6.1).

Table 3.6.1. Levels of linkage disequilibrium between all 8 polymorphic sites, given as a coefficient of correlation (r^2) and as |D'|, with a χ^2 statistical test of the |D'| measure of LD. Polymorphism pairs[1] with significant P-values are indicated in bold text in the table.

	Polymorphism pairs[1]	*N*	r^2	|D'|	$\chi 2$	P-value
1	**-5116,-4998**	**38**	**0.1621**	**1**	**12.159**	**0.0005**
2	**-5116,-4904**	**37**	**0.5426**	**0.92**	**40.1497**	**0**
3	-5116,-4876	38	0.0077	0.1778	0.5747	0.4484
4	**-5116,-4819**	**35**	**0.1365**	**1**	**9.8286**	**0.0017**
5	-4998,-4904	37	0.0093	0.33	0.6693	0.4133
6	-4998,-4876	38	0.0011	0.1471	0.0733	0.7866
7	-4998,-4819	35	0.0202	0.5385	1.3505	0.2452
8	**-4998,127**	**35**	**0.2382**	**1**	**17.8676**	**0**
9	-4998,407	30	0.0154	1	1.1682	0.2798
10	**-4998,15323**	**30**	**0.2596**	**0.5625**	**18.9519**	**0**
11	-4904,-4876	37	0.0023	0.0476	0.1655	0.6841
12	-4904,-4819	34	0.0038	1	0.2605	0.6098
13	-4904,127	34	0.012	1	0.8165	0.3662
14	-4904,407	29	0.0579	1	4.3403	0.0372
15	**-4904,15323**	**29**	**0.2381**	**1**	**17.1429**	**0**
16	-4876,-4819	35	0	0.0286	0.0028	0.9577
17	-4876,127	35	0.006	0.275	0.4021	0.526
18	-4876,407	30	0.0054	0.3556	0.3596	0.5487
19	-4876,15323	30	0.0128	1	0.9359	0.3333
20	-4819,127	32	0.0164	1	1.1967	0.274
21	-4819,407	27	0.1041	0.4643	7.0764	0.0078
22	-4819,15323	27	0.012	1	0.8165	0.3662

[1] Refers to polymorphic sites between which pairwise comparisons were done

For all the samples with genotype information at the four sites genotyped by indirect methods, generally low levels of linkage disequilibrium were detected between polymorphisms (Table 3.6.2), as can be observed from the low r^2 values. P values below 0.01 indicate that |D'| is significantly greater than zero. The |D'| value of pair 6 does not reflect this low disequilibrium due to the absence of homozygotes for the minor allele at site 15323 and the presence of only one homozygote for the minor allele at site 407. There is evidence of significant linkage disequilibrium between alleles at sites 127 and 407.

Table 3.6.2. Levels of linkage disequilibrium between polymorphic sites genotyped, given as a coefficient of correlation (r^2) and as |D'|, with a χ^2 statistical test of the |D'| measure of LD. The bold text indicates the polymorphic sites between which significant linkage disequilibrium was found.

	Polymorphism pairs[1]	***N***	**r^2**	**|D'|**	**χ2 value**	**P-value**
1	A-5116G, C127T	153	0.0035	0.1406	1.1868	0.276
2	A-5116G, G407A	137	0.0021	0.1788	0.691	0.4058
3	A-5116G, C15323T	140	0.008	0.3239	2.5608	0.1095
4	**C127T, G407A**	**149**	**0.0184**	**0.2148**	**6.13**	**0.0133**
5	C127T, C15323T	152	0	0.039	0.0065	0.9356
6	G407A, C15323T	135	0.0057	1	1.7984	0.1799

[1] Refers to polymorphic sites between which pairwise comparisons were done

For all the HIV-positive samples, with genotype information at the four sites genotyped by indirect methods, overall low levels of linkage disequilibrium were

detected between polymorphic sites (Table 3.6.3), as can be observed from the low r^2 values. P values < 0.01 were taken to indicate significant linkage disequilibrium. There is no evidence of significant linkage disequilibrium between any of the polymorphisms. Again, a |D'| value of 1 was observed for pair six, however this is not significant and is probably due to the low frequency of the T allele.

Table 3.6.3. Levels of linkage disequilibrium between polymorphic sites in HIV-positive samples given as a coefficient of correlation (r^2) and as |D'|, with a χ^2 statistical test of the |D'| measure of LD.

	Polymorphism pairs[1]	***N***	**r^2**	**|D'|**	**χ2 value**	**P-value**
1	A-5116G, C127T	118	0.001	0.0733	0.2684	0.6044
2	A-5116G, G407A	104	0.0116	0.3904	2.9571	0.0855
3	A-5116G, C15323T	109	0.0056	0.2633	1.3841	0.2394
4	C127T, G407A	114	0.0174	0.2197	4.5261	0.0334
5	C127T, C15323T	120	0.0009	0.2434	0.221	0.6383
6	G407A, C15323T	105	0.0061	1	1.4862	0.2228

[1] Refers to polymorphic sites between which pairwise comparisons were done

3.7 Haplotype analysis

Haplotypes and haplotype frequencies were analysed for all eight polymorphisms (Table 3.7.1). Haplotypes and haplotype frequencies were analysed separately for the four polymorphisms genotyped by indirect methods for all the samples, as there is a larger samples size with genotype information for these sites (Table 3.7.2). All haplotyping was performed with the use of the computer software tool, Phase version 2.1.1 (Stephens *et al* 2001).

Haplotyping analysis of all eight polymorphic sites gave nine haplotypes in 24 samples, with Hap 1 being the most common with a frequency of 0.420551 and Hap 9 being the least common with a frequency of 0.031707. The haplotypes all have the G allele at site 407 (position 7). Eight of the nine possible haplotypes have the major allele at site 127 (position 6), site 15323 (position 8) and site -4998, with only one haplotype containing the alternate allele. This is most likely due to the frequencies of these alleles at these sites. Similarly, at site -4876 seven of the nine haplotypes contain the G allele. In general the haplotype frequencies obtained by haplotype analysis with Phase seem to be a reflection of the allele frequencies, this is likely due to the small sample size.

Table 3.7.1. Haplotypes and haplotype frequencies, given by Phase, for all eight polymorphic sites in 24 samples.

Name	Haplotype	Frequency
Hap 1	AGCTGCGC	0.420551
Hap 2	GGTTGCGC	0.115521
Hap 3	AGCTGTGC	0.063386
Hap 4	GGTTACGC	0.059316
Hap 5	AGCTGCGT	0.053140
Hap 6	GCTTGCGC	0.050804
Hap 7	GGCCGCGC	0.041667
Hap 8	GGCTGCGC	0.033269
Hap 9	AGCCGCGC	0.031707

For the four polymorphisms genotyped by indirect methods, eleven of the sixteen possible haplotypes were detected in 124 samples. Hap 1 is the most common, with a frequency of 0.380261 and Hap 11 is the least common, with a frequency of 0.004559 (Table 3.7.2). Again, the haplotype frequencies seem to be a reflection on the allele frequencies more than an indication of inherited haplotypes due to selection.

Table 3.7.2. Haplotypes and haplotype frequencies, given by Phase, for the four polymorphic sites genotyped by indirect methods in 128 samples.

Name	Haplotype	Frequency
Hap 1	ACGC	0.380261
Hap 2	GCGC	0.357464
Hap 3	GTGC	0.061827
Hap 4	ATGC	0.061582
Hap 5	ACGT	0.035026
Hap 6	GCGT	0.028710
Hap 7	ACAC	0.025551
Hap 8	GCAC	0.023744
Hap 9	ATAC	0.010793
Hap 10	ATGT	0.005400
Hap 11	GTAC	0.004559

HIV-positive samples and general population samples were compared for haplotype pairs that were resolved to a probability difference of >0.15. That is, for each sample with more than one possible haplotype, a list of the possible haplotypes was given, with the probability of each haplotype. Only haplotypes with a probability greater than 0.15 more than the other haplotype probabilities were used for this comparison. Haplotypes were compared between the HIV-positive and general population samples in order to determine which haplotypes were more common in each group and if their frequencies differed between groups (Table 3.7.3). Hap 1 is most common in the HIV-positive samples and Hap 2 is most common for the general population group, whilst Hap 5 and Hap 9 are not present in the general population samples despite occurring in the HIV-positive samples at frequencies of 0.011 and 0.005 respectively. It should also be noted that a larger proportion of the haplotypes in the HIV-positive samples were

not able to be resolved than in the general population sample group and that this may result in the discrepancies seen in the comparison, along with the relatively small sample size of the general population group compared with that of the HIV-positive group.

Table 3.7.3. Haplotypes and haplotype frequencies for the haplotype pairs, resolved to a probability difference of >0.15, for the four polymorphic sites genotyped by indirect methods in 93 HIV-positive samples 31 general population samples.

Name	**Haplotype**	**Frequency in HIV$^+$**	**Frequency in general population**
Hap 1	GCGC	0.296	0.306
Hap 2	ACGC	0.290	0.403
Hap 3	GTGC	0.027	0.016
Hap 4	ATGC	0.022	0.016
Hap 5	GCAC	0.011	0
Hap 6	ACGT	0.005	0.032
Hap 7	ACAC	0.005	0.016
Hap 8	GCGT	0.005	0.016
Hap 9	ATAC	0.005	0
Unresolved		0.333	0.194

CHAPTER 4
DISCUSSION

Direct sequencing of the four regions, spanning a total of 2689bp in the *TRIM5* gene, revealed nine polymorphisms in an average of 20 samples. Three polymorphisms, at sites -5116, -4904 and -2, are common, with a minor allele frequency close to 0.15 and six of the polymorphisms are rarer ($q<0.15$). Five polymorphisms were detected in the upstream region of the gene, four of which were known and one that has not previously been detected, at site -4819. Sequencing of exon 2 revealed three previously detected polymorphisms. No variation was detected in exon 7 despite the fact that there is one recorded polymorphism in this region, in intron 7 at site 14345, according to the NCBI dbSNP. One polymorphism was detected in exon 8.

Methods devised for the indirect detection of four SNPs were successful. Two different methods of indirect detection were used, namely allele specific amplification (ASA) and polymerase chain reaction restriction fragment length polymorphism (PCR-RFLP). Where the single nucleotide polymorphism (SNP) introduced or disrupted an endonuclease recognition sequence PCR-RFLP was used, unless the fragments of restriction were too small to be resolved on an agarose gel as in the case of the polymorphism at position -5116, where ASA was used. ASA was also used for the indirect detection of the polymorphism at site 127 as it conferred no change to any endonuclease restriction recognition sequence. There was no evidence of deviation from Hardy-Weinberg equilibrium

for any of the five polymorphisms genotyped by sequencing 40 samples or the four sites genotyped by indirect methods in more than 150 samples.

Significant differences in genotype proportions were not observed between HIV-positive and general population samples at any of the four sites genotyped for all the samples. No significant differences between of the genotype numbers for HIV-positive and general population samples at non-coding site -5116 were observed. There were also no significant differences observed between genotype numbers for the two population groups for coding sites 127, 407 and 15323, which code for variants H43Y, R136Q and P479L respectively. This implies that none of the variation studied has a protective effect on the population studied. However, as the general population samples are few and not persistent long term high-risk seronegative samples, it is premature to say that there is no protective effect conferred by any of these variants.

In general, low levels of linkage disequilibrium were observed for the four polymorphisms. This is consistent with the low levels of LD observed in African populations. However, linkage disequilibrium was observed between polymorphisms at sites 127 and 407 in the analysis of all the samples. No linkage disequilibrium was observed between polymorphisms at any of the sites in the analysis on the HIV-positive samples. A |D'| value of 1 for a polymorphism pair means that two or three haplotypes are present in the population, and not all of the four possible haplotypes for the pair. This was observed between polymorphisms at sites 407 and 15323 and was confirmed by haplotype analysis, with the A allele

at site 407 and the T allele at site 15323 not being present in the same haplotype in any samples. However, this is not significant and is possibly due to the absence of homozygotes for the minor T allele at site 15323 and the presence of only 1 homozygote for the A allele at site 407 rather than a result of complete linkage between the G and C alleles. Another study found that the A allele at site 407 can occur in the same haplotype as the T allele at site 15323, albeit at a very low frequency (<0.1) (Goldschmidt *et al* 2006).

Haplotyping analysis showed that there are two haplotypes with high frequencies, namely GCGC and ACGC, with frequencies of 0.3575 and 0.3803 respectively. This is most likely due to the low frequency of rare alleles at all of the sites except site -5116.

Six of the polymorphisms which were detected are in non-coding regions, whilst three of the polymorphisms detected result in amino acid substitutions. These non-synonymous changes, namely H43Y, R136Q and P479L, occur in the RING, coiled-coil and SPRY domains of the TRIM5α protein respectively.

A recent study on the effects of *TRIM5* polymorphisms on susceptibility to HIV-1 shows that the TRIM5α protein containing a tyrosine at amino acid position 43, which corresponds to the T allele at site 127, is less efficient at HIV-1 restriction in tissue culture that that containing a histidine residue in this position, which corresponds to the C allele at site 127 (Javanbakht *et al* 2006). This implies a protective effect conferred by the C allele, which was not observed by

comparing genotype numbers between HIV-positive and general population samples.

The study by Javanbakht *et al* (2006) showed an elevated frequency of the T allele at site 127 in HIV-1-seronegative African American individuals (n = 302) as well as in high-risk exposed HIV-1 uninfected African American individuals (n = 77), in comparison to HIV-1 seroconverters (n = 282). They therefore suggested that this allele may provide an HIV-1-protective effect (Javanbakht *et al* 2006). No significant difference in genotype numbers between HIV-positive and general population samples was observed at this site. Furthermore, another study conducted on European American males having sexual intercourse with males showed no significant difference in allele frequencies between high risk exposed seronegative (n = 96) and HIV-infected (n = 140) individuals at this site (Speelmon *et al* 2006), which is in agreement with our findings. The discordant results between the study by Javanbakht *et al* (2006) and this study may be explained by differences in the populations studied and in the sampling of the populations, as that study was on African-Americans, whereas this study was on Black South Africans and their study had access to high-risk uninfected individuals as well as HIV-seronegative individuals, whereas this study only makes use of a relatively small group of samples with unknown HIV status. The differences of the results in the Javanbakht *et al* (2006) and Speelmon *et al* (2006) studies may be explained by the differences in the population groups studied.

The data produced by Javanbakht and colleagues also suggests that the A allele frequency at site 407 is elevated in HIV-1-seronegative (n = 420) and high-risk exposed uninfected (n = 80) African American individuals in comparison with seroconverters (n = 295) and that this allele may also provide an HIV-inhibitive effect (Javanbakht *et al* 2006). The study by Speelmon *et al* (2006), however, shows a contradictory result. They show that there is no significant difference between allele frequencies in the HIV-uninfected (n = 96) and HIV-positive (n = 140) groups in a European American population, as our data also reflects, but found that the haplotype containing the A allele had an elevated frequency in HIV-positive individuals (Speelmon *et al* 2006). This elevated frequency of the haplotype containing the A allele in HIV-positive individuals was not observed in this data. These results show that the haplotypes containing the A allele have a frequency of 0.021 in HIV-positive individuals and 0.03 in the general population samples.

Another study investigated the role of TRIM5α variants on disease progression using a large cohort of 979 samples (Goldschmidt *et al* 2006). They found no association between any of the TRIM5α variants and disease progression. They did, however, note that none of the common variation detected occurs within the blocks of positive selection in the primate lineage, v1 and v2 (Goldschmidt *et al* 2006). There was also no variation detected in these variable regions by sequencing this region in a black South African population. As these regions of positive selection represent parts of the gene that may be functionally important in retroviral resistance, it is surprising that no variation is detected within these

regions, while there is variation in other parts of the gene. The variation that this and other studies have detected in the rest of the gene may therefore have been selected for by past viral epidemics that were not related to HIV.

The Javanbakht *et al* (2006) study did not detect the polymorphism at position -5116 while neither the Speelmon *et al* (2006) or Goldschmidt *et al* (2006) studies found any association of the SNP at site -5116 with HIV-susceptibility. This study did not reveal a significant difference between genotype numbers for the HIV-positive sample population and the general population with respect to site -5116, however, differences in the haplotype frequencies for haplotypes containing the A and G alleles at this site were observed between HIV-positive and general population samples. The function of this polymorphism is not known, however it may regulate levels of TRIM5α protein in the cell. It is also in linkage disequilibrium with the other upstream polymorphisms of the gene at sites -4998, -4904 and -4819. Alternatively, it may be in linkage disequilibrium with another polymorphism in the gene that was not investigated in this study. It is also possible that it is not a single variant that has a protective effect against HIV but rather a combination of variants that determine susceptibility to HIV. However, it should be noted that the sample size of resolved haplotypes in the general population which was used for comparison between haplotype frequencies was very small and the differences revealed may be an artefact of this.

Another discrepancy between our data and that of Javanbakht *et al* (2006) is that they found SNP's at sites 127 and 407 to be in complete negative linkage

disequilibrium, that is the T allele at site 127 and the A allele at site 407 never occurring together in the same haplotype. This was not the case with our data, as can be seen in samples 185 and 311. Sample 185 has the genotypes TT and GA and sample 311 has genotypes CT and AA at sites 127 and 407 respectively, thus it is possible for the T and A alleles to occur on the same haplotype. However, this study did detect the existence of significant linkage disequilibrium between these two polymorphic sites.

Interestingly, despite the fact that more variation was expected in this African population than that which has been reported for other populations, there were seven polymorphisms that have previously been detected according to submissions to the Ensembl SNP database that were not present in any of the samples that sequenced in this study. These SNPs occur in the upstream region at position -5321 (rs16934390), the 5' untranslated region of exon 1 at position -4811 (rs28381978), exon 2 at sites 334 (rs11601507) and 369 (rs35216582), intron 7 at site 14345 (rs34525757) and two SNPs in exon 8 at sites 15017 (rs35852130) and 15142 (rs28381981). The polymorphism at site -5321 has a minor allele frequency of 0.075 in Yoruba mother-father-child trios and 0.087 in 23 African American individuals. This occurs at a frequency that one would have expected to be detectable by sequencing the number of samples sequenced in this study. However, it may occur in the South African black population at a lower frequency. The polymorphism at site -4811 was detected with a minor allele frequency of 0.02 in a sample population of 48 Swiss Caucasians, so it is not improbable that it was not detected in the study population due to its low

frequency. However, the absence of this polymorphism in these samples may also be due to population differences between populations that it has been detected in and African populations. At site 334 the polymorphism was not detected in any of the populations genotyped by the HapMap project and the SNP was validated by multiple submissions to dbSNP but no frequency information was supplied. The polymorphisms at sites 369, 14345 and 15017 had no frequency data and no information on the validation of these SNPs so it was not possible to determine whether the inability to detect them in this study is due to the sample size or differences between populations. At site 15142 the SNP had a minor allele frequency of 0.08 in 48 Swiss individuals and 0.04 in an unknown population of 7832 individuals. Again, differences in the frequencies of the polymorphism in different populations may explain this, depending on the origin of the latter sample population.

The polymorphisms in exon 2 both result in amino acid substitutions as do those in exon 8. The three other studies that resequenced in order to detect variation were able to detect some of these polymorphisms that are present in the Ensembl database; Javanbakht *et al* (2006) detected rs11601507, Speelmon *et al* (2006) detected rs28381975 and rs11601507 and Goldschmidt *et al* (2006) detected rs28381975. In addition, all the studies detected rs28381981, a polymorphism at site 15142, in the SPRY domain of the *TRIM5* gene (Goldschmidt *et al* 2006, Javanbakht *et al* 2006 and Speelmon *et al* 2006); however not one of them was able to detect the polymorphism at site 15323 (rs7104422), also a non-synonymous substitution in the SPRY domain.

The SNP at site 15142, which does not occur in any of our samples, causes a change in amino acid 419 from histidine to tyrosine; the minor allele is present in approximately 1% of an African American population and 5-6% of three Caucasian populations (Javanbakht *et al* 2006 and Speelmon *et al* 2006). Similarly, the polymorphism at site 334 that was not present in any of our samples causes amino acid 112 to change from valine to phenylalanine; the minor allele is present in approximately 1% of an African American population and in 7% of two European American populations (Javanbakht *et al* 2006 and Speelmon *et al* 2006). Thus, it is not all that surprising that these polymorphisms were not detected by sequencing 20 samples of the black South African population, as they occur at such a low frequency in the African American population compared to European American populations. They may in fact be European population-specific polymorphisms as African -American populations have 6.8% - 22.5% European genetic ancestry (Parra *et al* 1998). The polymorphisms at sites 334 and 15142 were found to have no significant differences between HIV-positive and -negative samples (Javanbakht *et al* 2006 and Speelmon *et al* 2006).

The differences between this study and those by Goldschmidt *et al* (2006), Javanbakht *et al* (2006) and Speelmon *et al* (2006) may be explained by sample size differences. Approximately 150 HIV-positive samples and 40 general population samples were genotyped in this study. The Goldschmidt study had 979 HIV-seroconverter study participants and the Javanbakht study had sample sizes of 251-295 for HIV-seroconverters, 282-420 for HIV-seronegative and 74-

80 for high-risk exposed uninfected groups, depending on the polymorphism genotyped. The Speelmon study had a sample size of 95-96 for high-risk exposed seronegative individuals and 96-140 for HIV-1 infected individuals, depending on the polymorphism genotyped. In addition it must be taken into account that in this study the general population samples are not as informative as HIV-negative and high risk seronegative samples when looking at a possible protective effect of certain alleles or haplotypes.

However, these discrepancies may also be reconciled by genetic differences between African Americans, black South Africans and European Americans. It is likely that if seronegative individuals were genotyped and these comparisons were made, rather than making the comparisons between seronegative and general population samples, these differences may be reconciled. This is due to the fact that there is a high prevalence of HIV in the black South African population, in 2002 it was estimated at 12.9% (Nelson Mandela/HSRC study of HIV/AIDS 2002), thus the general population group in this study may contain as many as 5 HIV-positive samples. It would also help to have a larger population size to genotype in order to further investigate these differences.

It is not possible to definitively say whether there is more variation in this gene in the black South African population than in European or African American populations. Sequencing a larger sample size may help to better answer this question. There is also no conclusive evidence that any of the polymorphisms characterized in this study are directly or indirectly involved in the restriction of

HIV-1 through regulation of the gene, linkage disequilibrium with another polymorphism in the gene, or in combination with other variants. The results of this study, taken together with those from other similar studies (Goldschmidt *et al* 2006 and Speelmon *et al* 2006), do not point to a protective effect conferred by variants of TRIM5α to HIV-1. To investigate this further, it would be helpful to have larger sample sizes and a sample group of high-risk seronegative individuals for comparisons. The haplotypes present in the population were not completely resolved, however there may be haplotypes acting outside of the regions investigated that play a role in HIV inhibition. Again, genotyping of a larger sample size may help to better resolve the haplotypes, alternatively molecular haplotyping could be performed. As no correlation has been found between any of the polymorphisms in the *TRIM5* gene and disease progression, it would be interesting to investigate this with the use of long term cohorts with good disease progression data. It would also be interesting to compare more populations to one another with regard to the amount of variation and susceptibility to HIV to determine whether there really are major differences in the *TRIM5* gene between populations.

CONCLUSIONS

Until recently the effect of *TRIM5* polymorphism on HIV had not been explored and this is the first study on *TRIM5* polymorphism in a black South African population. Here, a novel polymorphism in the upstream region of the *TRIM5* gene has been reported at site -4819 and the presence of eight other polymorphisms has been confirmed. However, there are reported low frequency polymorphisms that were not detected and sequencing of a larger population size may be able to detect these.

No significant differences were observed between HIV-positive and general population groups for genotype numbers. Generally low levels of linkage disequilibrium were detected as is consistent with African populations, but significant linkage disequilibrium was observed between alleles at sites -5116 and 407.

Haplotyping analysis gave inconclusive results; the haplotypes were not completely resolved due to the low frequency of the minor allele at three of the sites genotyped, however there is a difference in the most common haplotypes for HIV-positive and general population samples, although this not statistically supported.

Future work on black South African populations includes molecular haplotyping, the use of larger population sizes and the establishment of long term cohorts with samples from high risk long term seronegative individuals, for comparison with

data from HIV-positive individuals. It would also be worthwhile to perform an analysis of the function of polymorphisms in the upstream non-coding region.

APPENDIX 1

Patient Information and Consent Form

PATIENT INFORMATION & CONSENT FORM

"Population Genetics of HIV resistance in southern Africa"

PATIENT NUMBER _____

Dear Patient,

We, in the Departments of Molecular and Cell Biology and the Clinical HIV Research Unit at the University of the Witwatersrand want to gain a better understanding of the variation in people's responses to HIV and AIDS. Some people in other parts of the world have genes that make them respond to HIV in different ways. We want to examine variation in these genes in people of southern African origin. Your blood and your medical history will allow us to determine what genetic variation is common and what effects that variation has on how fast HIV positive people become sick with AIDS.

We would be very grateful if you would donate 5 ml (1 tsp) of blood, drawn from an arm vein only once. Slight discomfort may be associated with drawing blood; this may include pain, swelling or bruising at the site of puncture. We also ask that you allow us to have access to your medical records at this clinic.

POTENTIAL BENEFITS

There will be no immediate benefit to you by participating in this study. You will be assisting in collecting accurate information about patients infected with HIV. We hope the data will help us to assess long term trends and design effective treatments for people in southern Africa.

VOLUNTARY PARTICIPATION

You may choose to take part in this study and you may decide to withdraw from the study at any time.

PRIVACY

We will keep your records confidential. Your blood samples are encoded with a number and your identity will only be known to the principal investigator and your doctor.

COST OF PARTICIPATION

There will be no charge for the laboratory tests or procedures which are required for your participation in the study. The study institution (University of the Witwatersrand, School of Molecular and Cell Biology) may compensate you for any travel expenses incurred by you to attend the clinic up to R20. There is no other form of compensation for your participation.

CONTACT NUMBERS

The investigator in charge of this study is Prof Tracy McLellan who may be reached for answers to questions concerning this study at 011 717-6357.

If you have questions about patient rights, you may contact The Committee for Research on Human Subjects, telephone (011) 717-1234.

If there is an emergency, you may contact the doctor, Dr P MacPhail, at 278-8800.

CONSENT

I, _____ have voluntarily agreed to take part in this research study. I have read this consent form (or had it read to me by the social worker) and have had an opportunity to discuss this with my doctor. All questions have been answered to my satisfaction and I fully understand my participation in this study. I will be given a copy of this consent form to keep.

_____	_____	_____
Signature of Patient	Date	Time
_____	_____	_____
Signature of Witness	Date	Time
_____	_____	_____
Signature of Investigator	Date	Time

PATIENT QUESTIONNAIRE

"Population Genetics of HIV Resistance in Southern Africa"

Patient Number _____ File number _____

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

YOU

Place of birth _____ Province or country _____

Home language _____

YOUR PARENTS

Your mother
Place of birth _____
Province or country _____
Home language _____

Your father
Place of birth _____
Province or country _____
Home language _____

YOUR GRANDPARENTS

Your mother's mother
Place of birth _____
Province or country _____
Home language _____

Your father's mother
Place of birth _____
Province or country _____
Home language _____

Your mother's father
Place of birth _____
Province or country _____
Home language _____

Your father's father
Place of birth _____
Province or country _____
Home language _____

Medical information

Year of birth _____

HIV + since _____

Most recent CD4+ count _____ Date _____

Other conditions: _____

Tuberculosis? _____

APPENDIX 2

University Of the Witwatersrand Ethics Clearance Certificate

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 McLellan

CLEARANCE CERTIFICATE	**PROTOCOL NUMBER M040221**
PROJECT	Population genetics of resistance to HIV in Southern Africa
INVESTIGATORS	Prof T McLellan
DEPARTMENT	Molecular & Cell Biology
DATE CONSIDERED	04.02.27
DECISION OF THE COMMITTEE*	Approved unconditionally

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

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

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof T McLellan

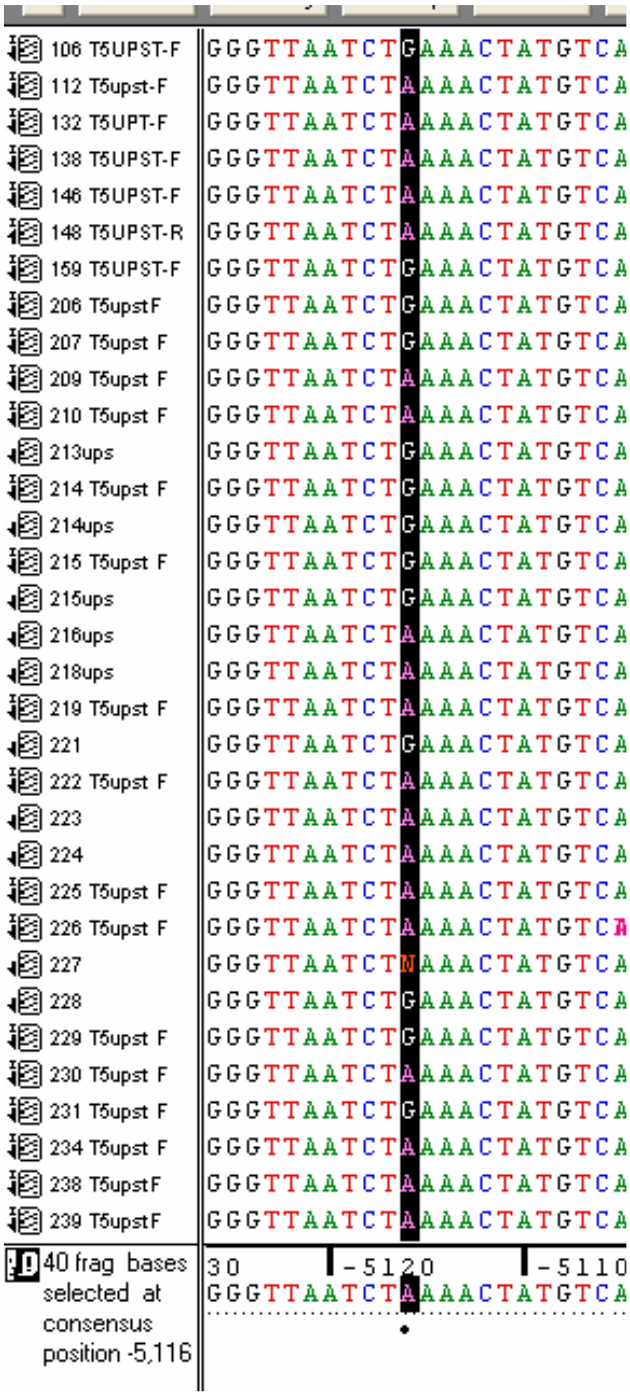
DECLARATION OF INVESTIGATOR(S)

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

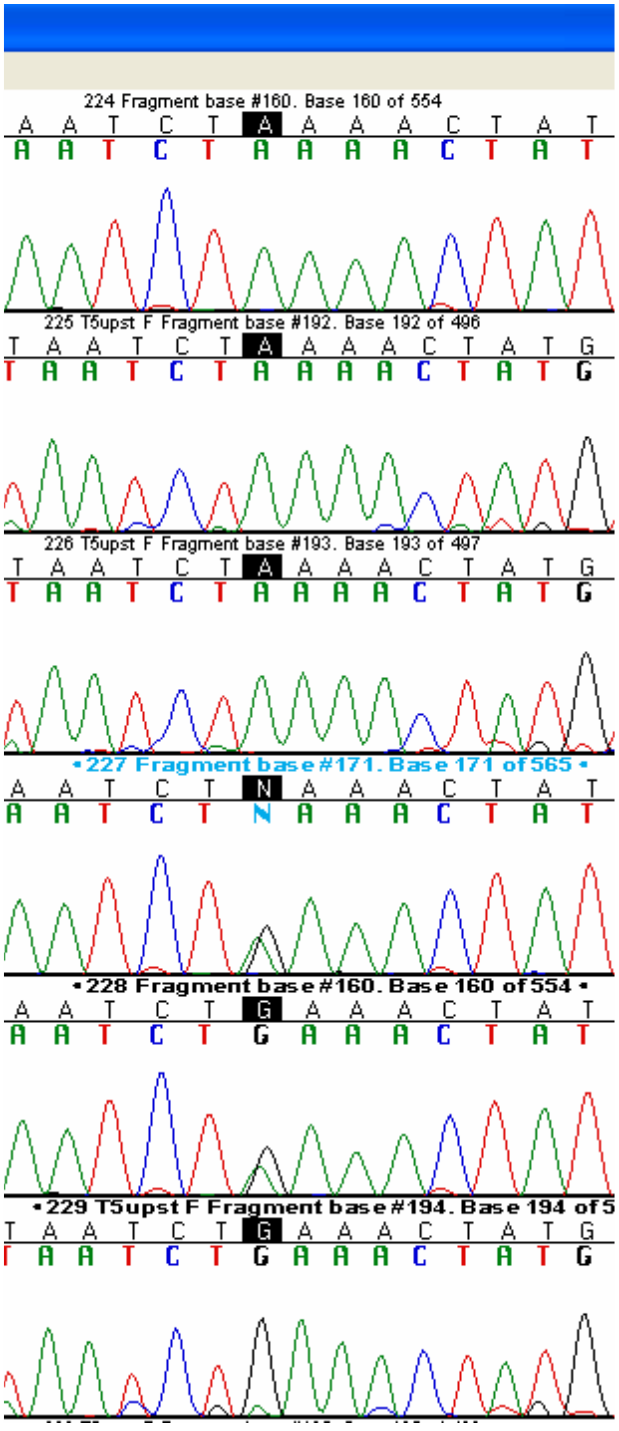
PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

APPENDIX 3

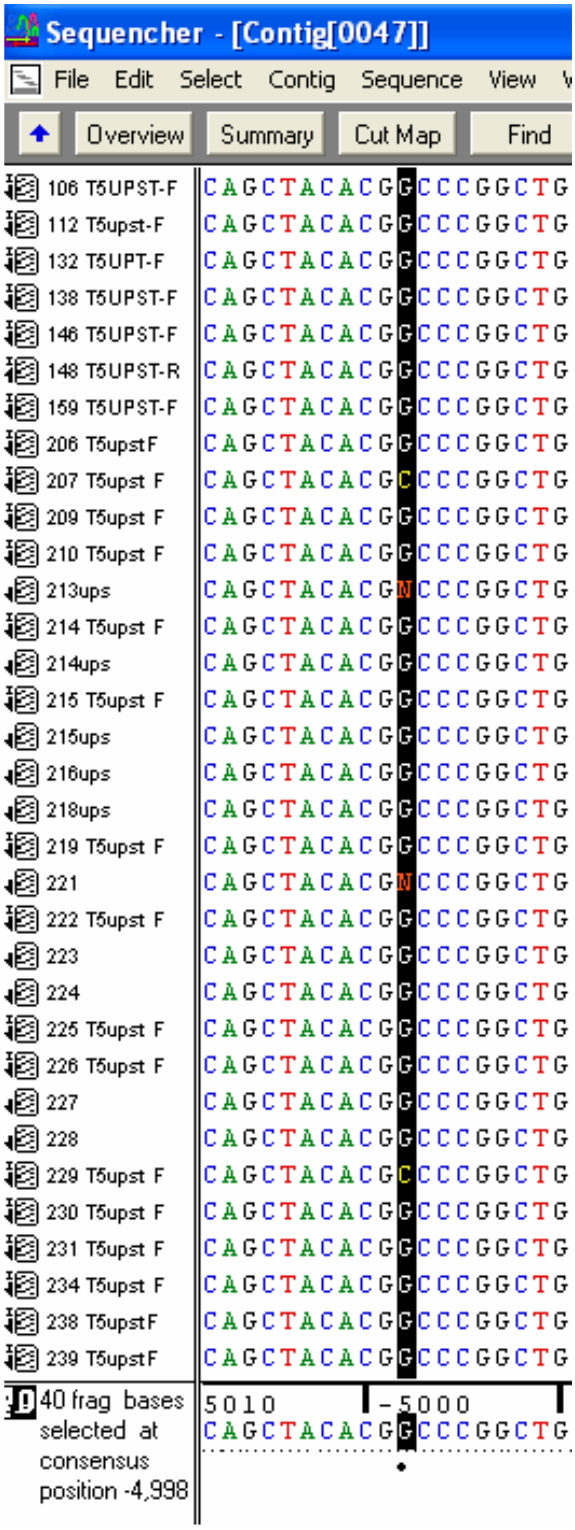
Sequencher contig sequence view of polymorphism at site -5116



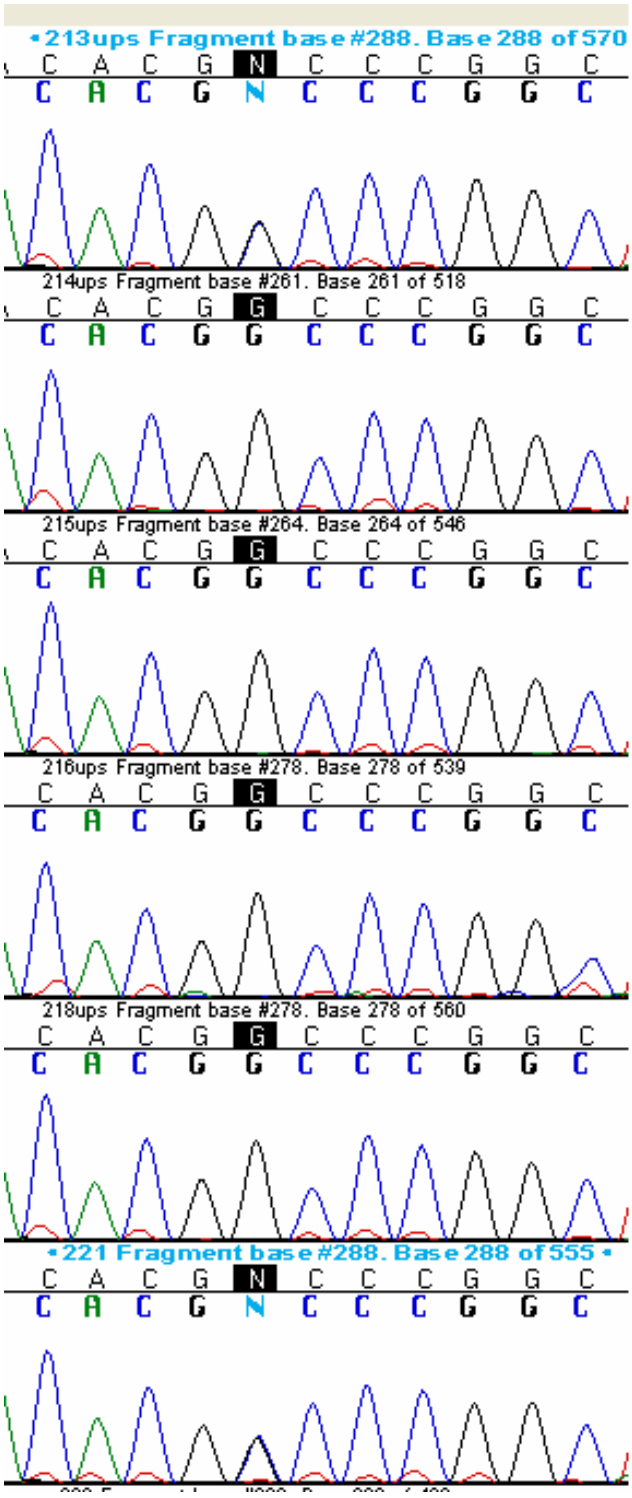
Sequencher chromatogram view of polymorphism at site -5116



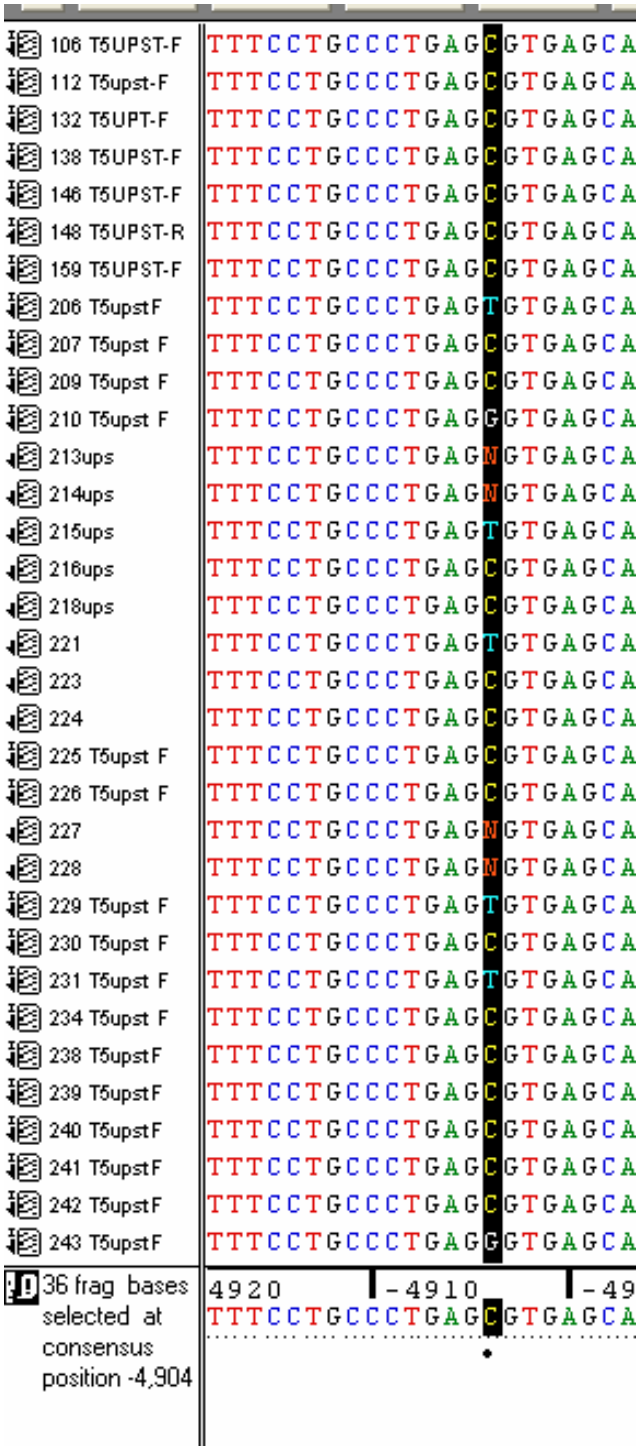
Sequencher contig sequence view of polymorphism at site -4998



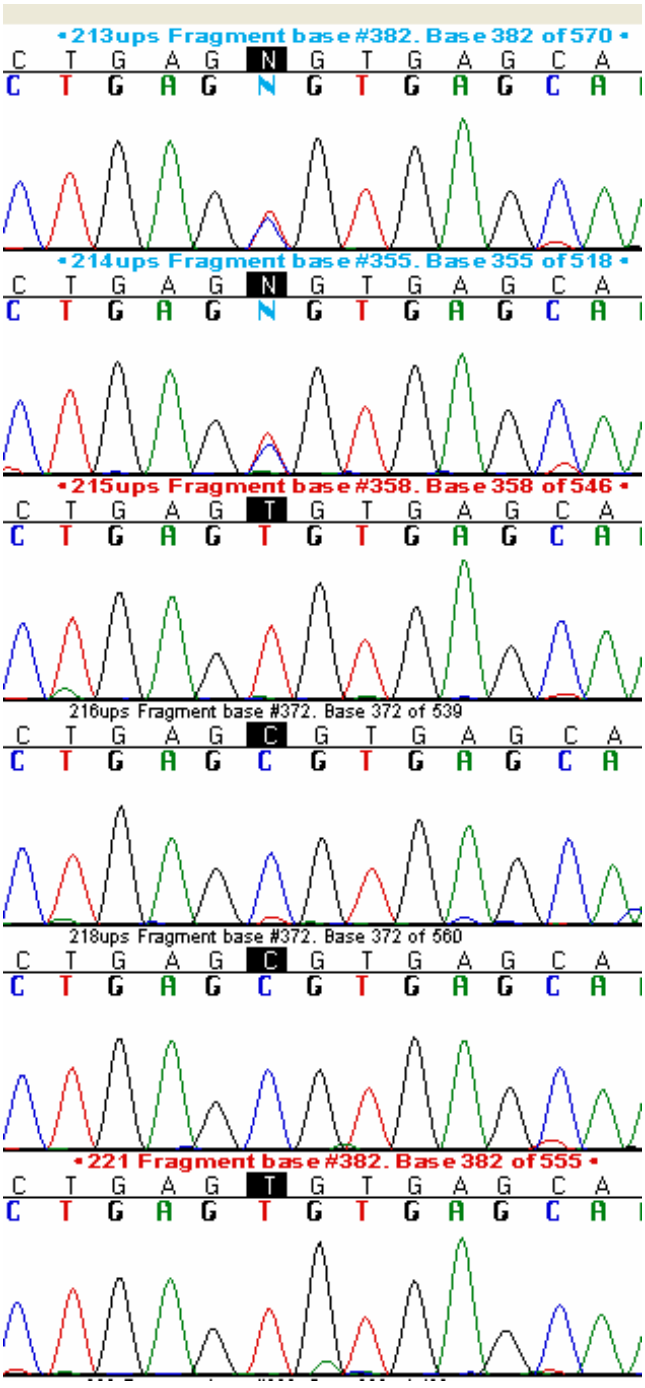
Sequencher chromatogram view of polymorphism at site -4998



Sequencher contig sequence view of polymorphism at site -4904



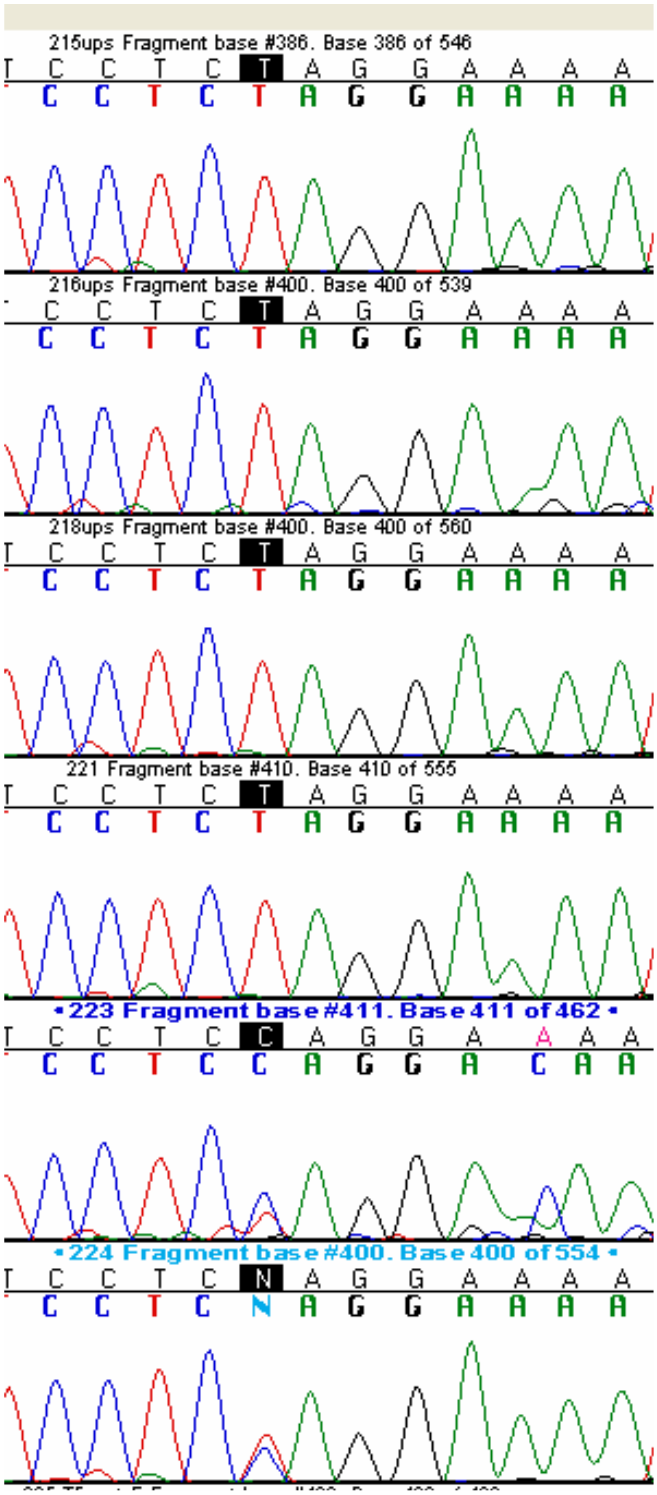
Sequencher chromatogram view of polymorphism at site -4904



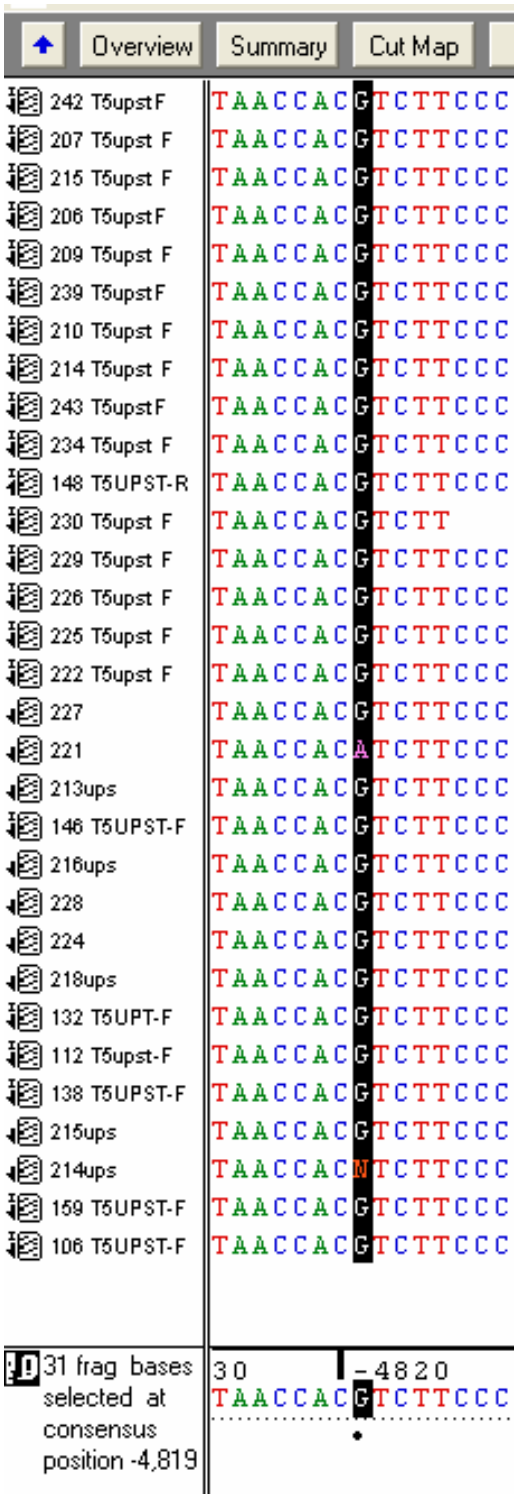
Sequencher contig sequence view of polymorphism at site -4876



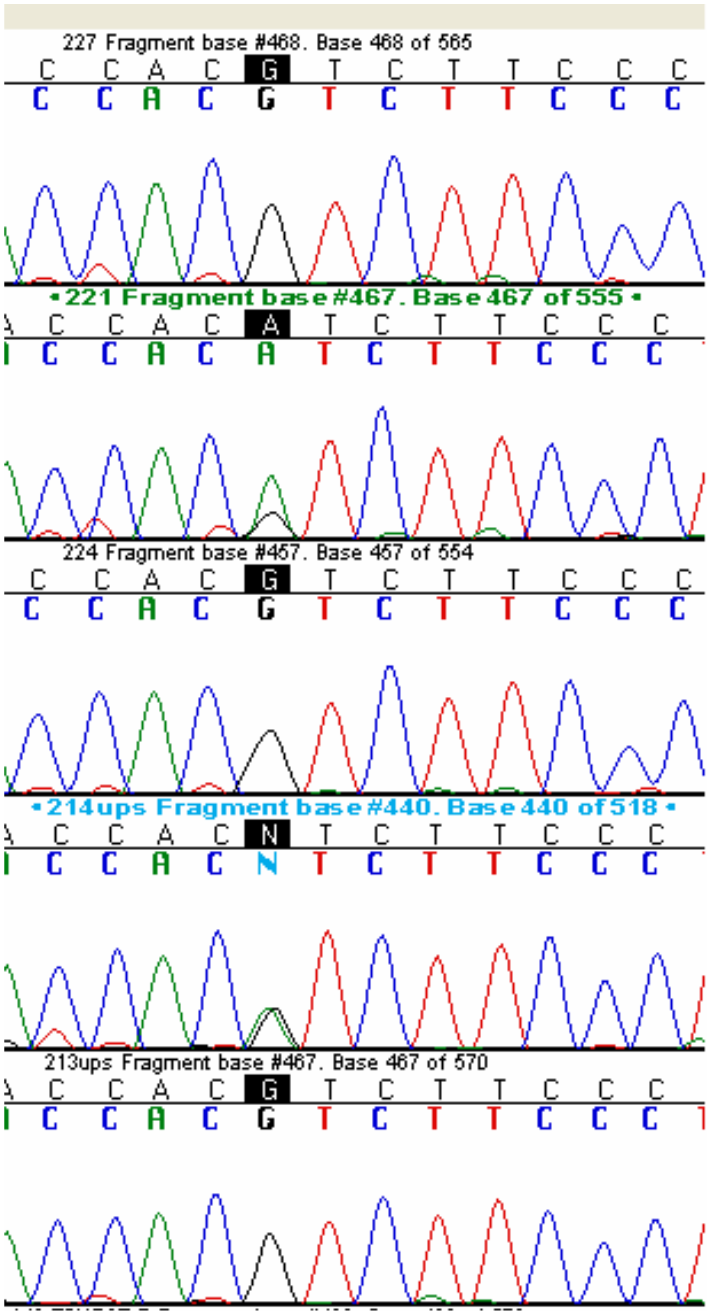
Sequencher chromatogram view of polymorphism at site -4876



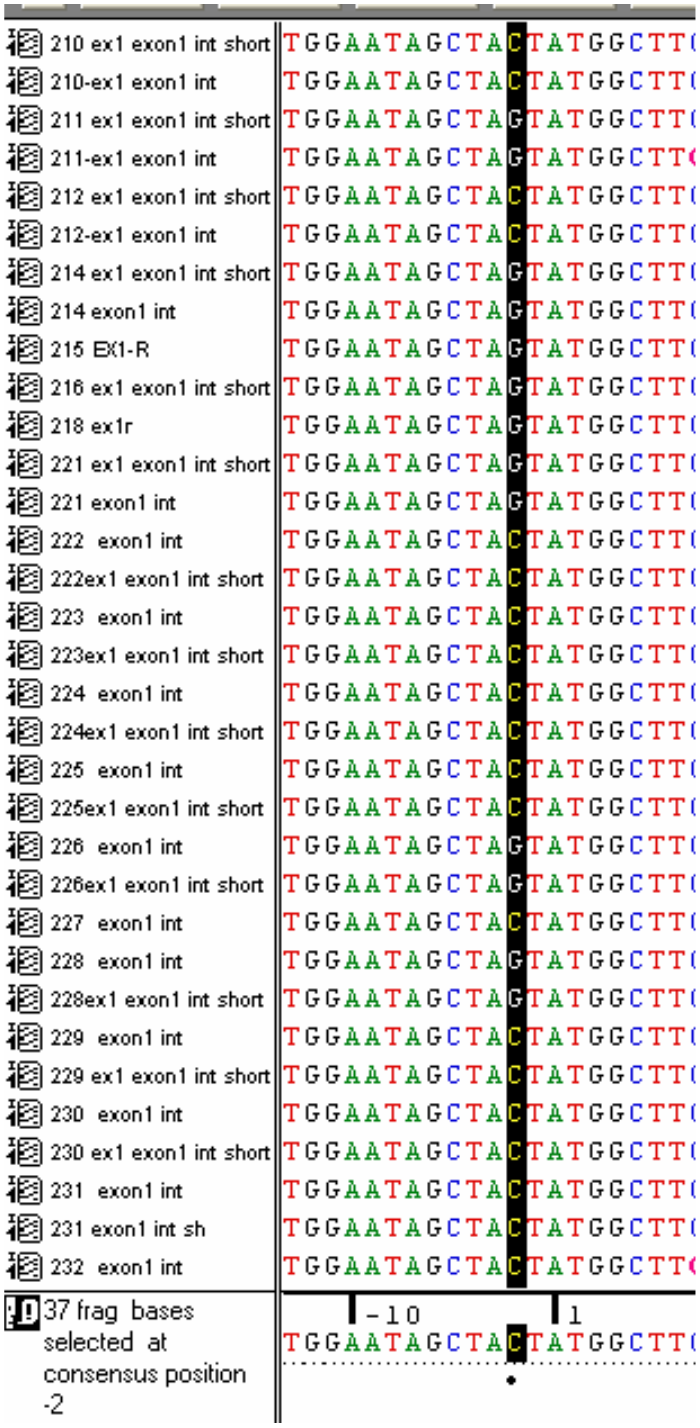
Sequencher contig sequence view of polymorphism at site -4819



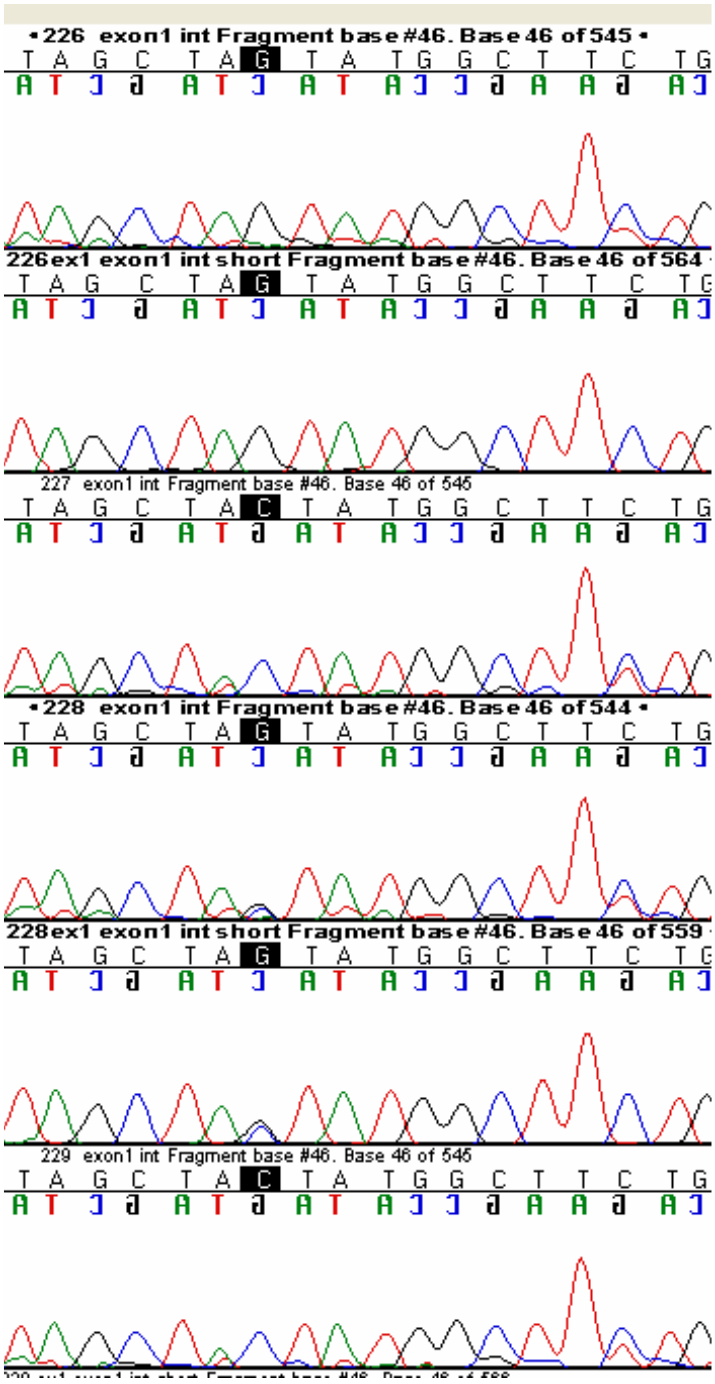
Sequencher chromatogram view of polymorphism at site -4819



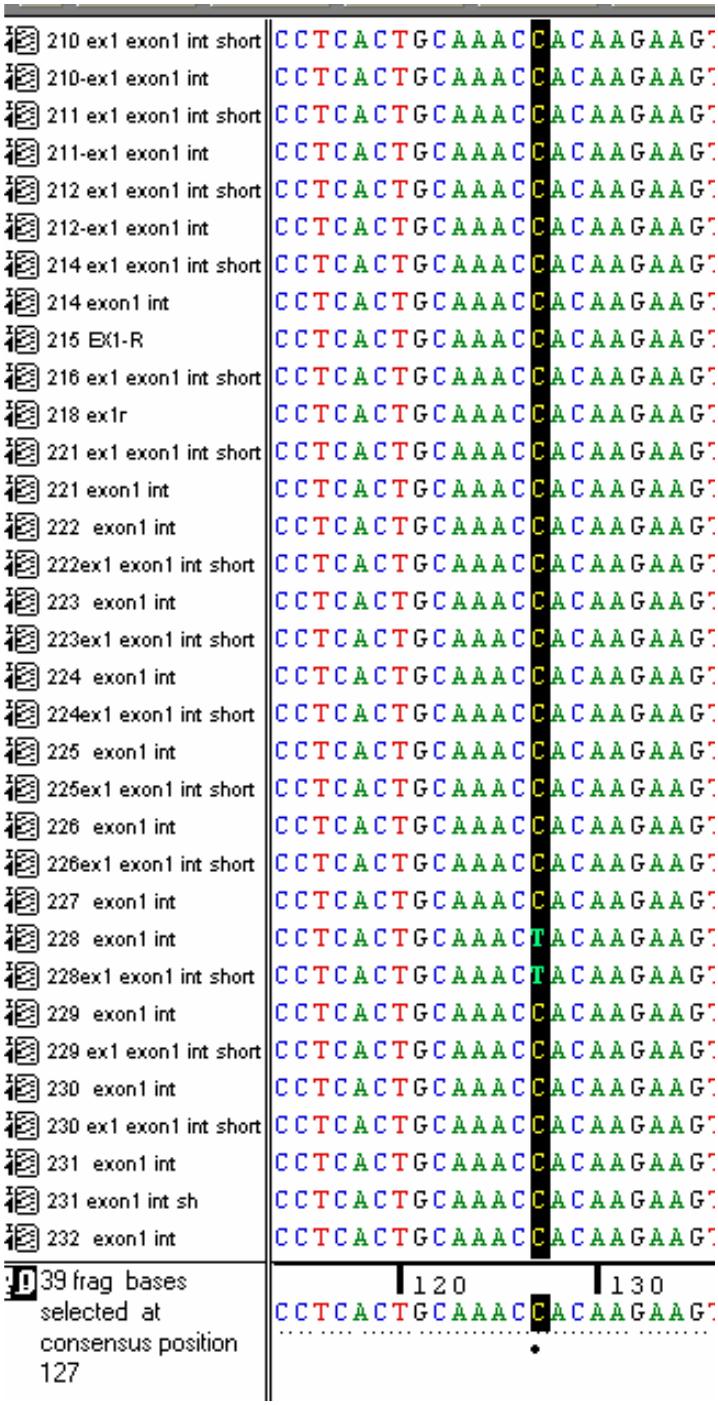
Sequencher contig sequence view of polymorphism at site -2



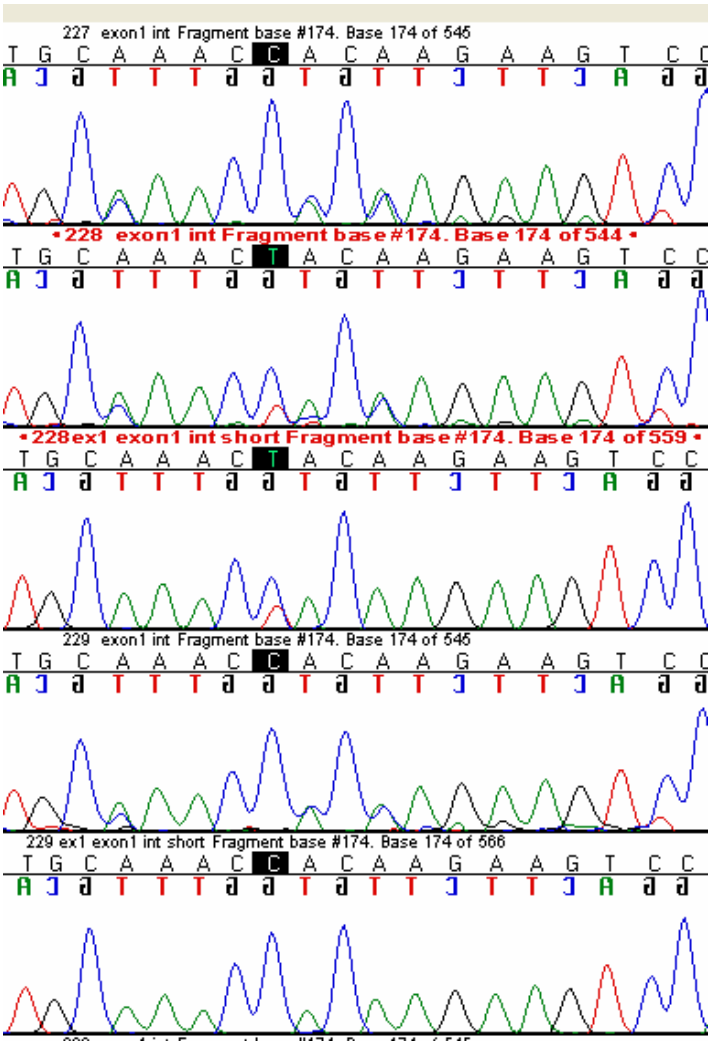
Sequencher chromatogram view of polymorphism at site -2



Sequencher contig sequence view of polymorphism at site 127



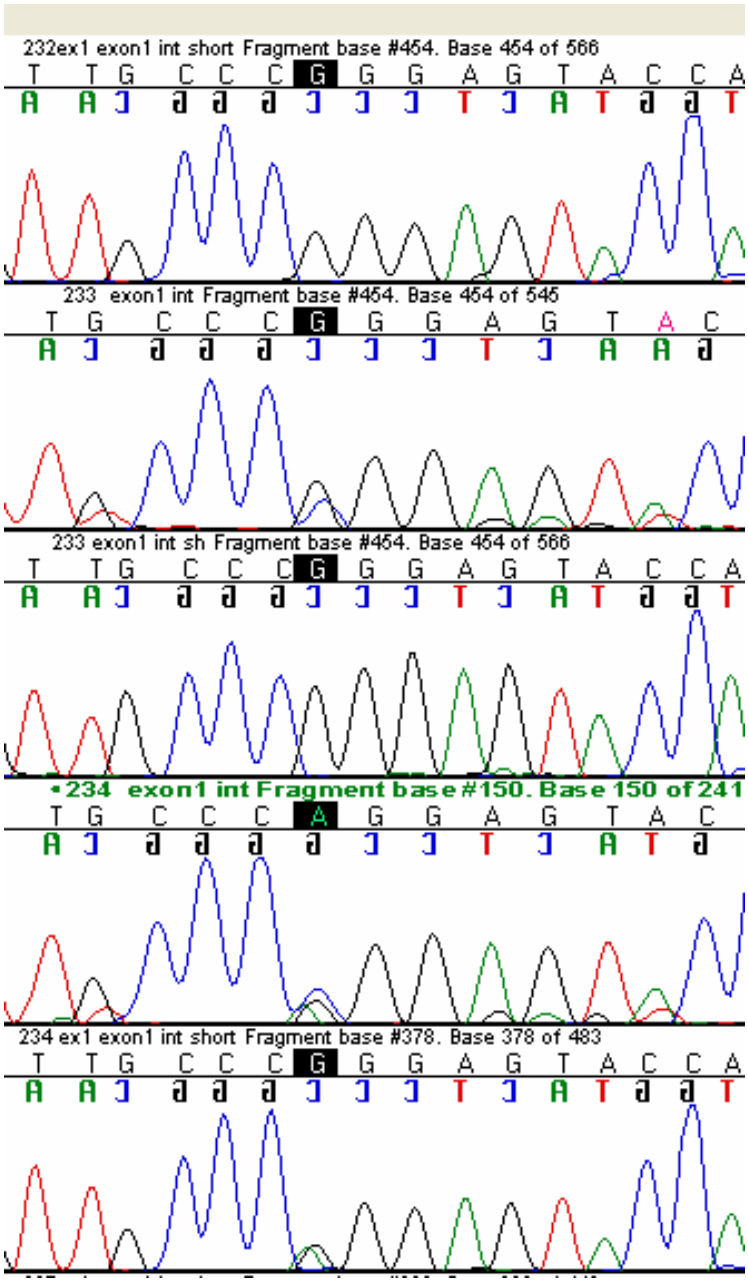
Sequencher chromatogram view of polymorphism at site 127



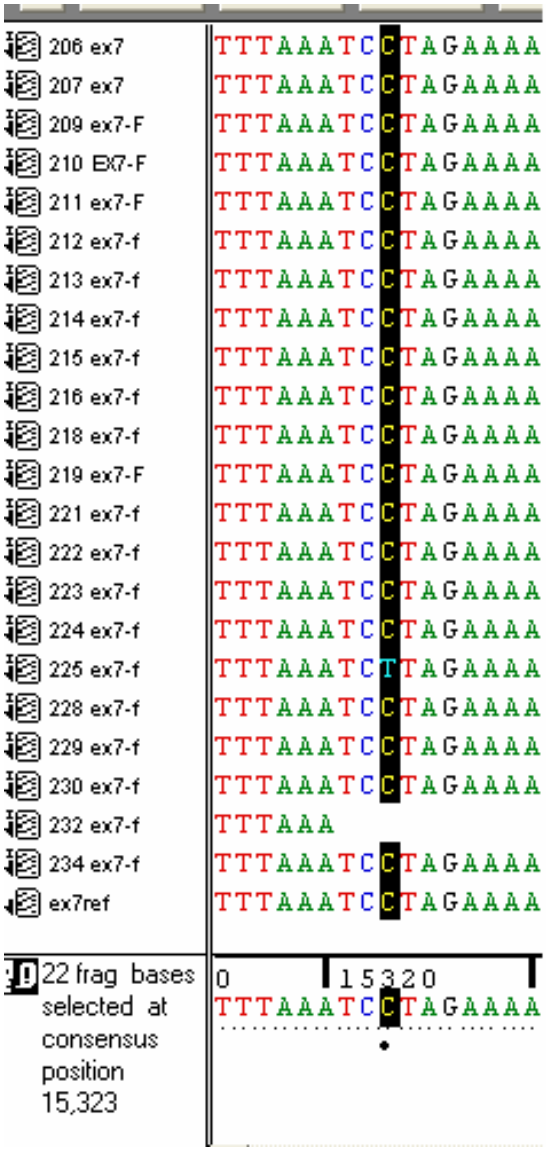
Sequencher contig sequence view of polymorphism at site 407



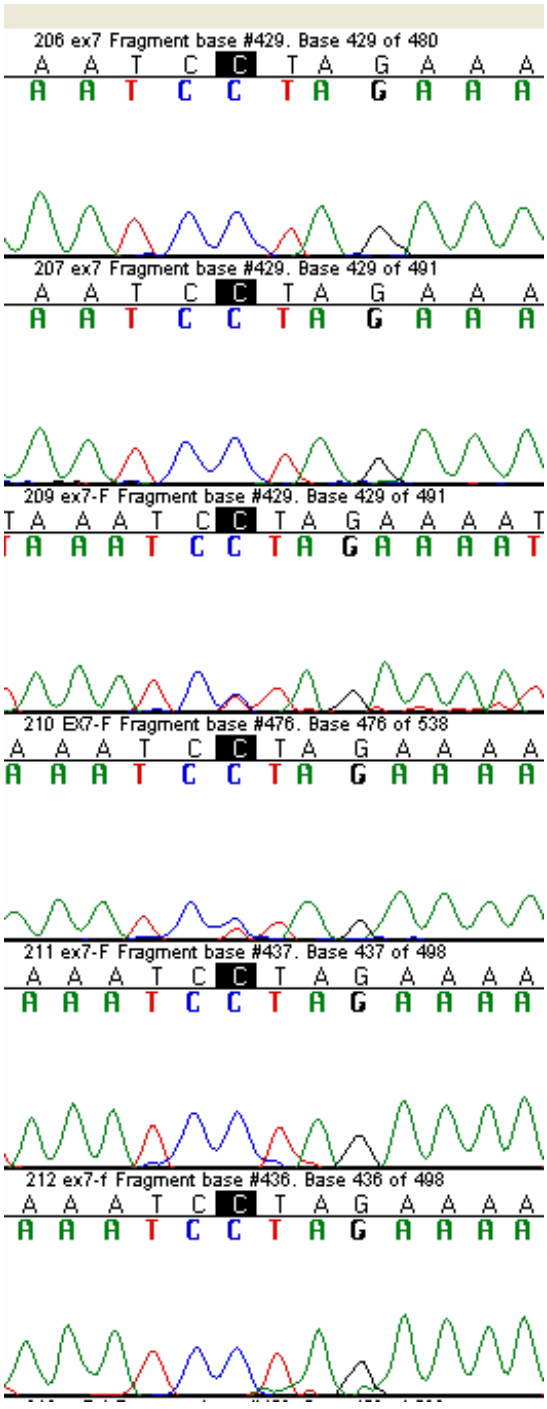
Sequencher chromatogram view of polymorphism at site 407



Sequencher contig sequence view of polymorphism at site 15323



Sequencher chromatogram view of polymorphism at site 15323



APPENDIX 4

Genotype information obtained by sequencing of the upstream region of the *TRIM5* gene.

Samples	-5116	-4998	-4904	-4876	-4819
106	GA	GG	CC	CT	GG
112	AA	GG	CC	TT	GG
132	AA	GG	CC	TT	GG
138	AA	GG	CC	CT	GG
146	GA	GG	CT	CT	GG
148	AA	GG	CC	TT	GG
159	GA	GG	CT	TT	GA
206	GG	GC	TT	TT	GA
207	GA	GC	CT	TT	
209	AA	GG	CC	TT	GG
210	GA	GG	CC	TT	GG
211	AA	GG	CC	TT	GG
213	GA	GC	CT	TT	GG
214	GG	GG	CT	TT	GA
215	GG	GG	TT	TT	GG
216	AA	GG	CC	TT	GG
217	AA	GG	CC	CT	GG
218	AA	GG	CC	TT	GG
219	AA	GG	CC	TT	GG
221	GG	GC	TT	TT	GA
222	AA	GG	CC	TT	GG
223	AA	GG	CC	CT	GG
224	AA	GG	CC	CT	GG
225	AA	GG	CC	TT	GG
226	AA	GG	CC	TT	GG
227	GA	GG	CT	TT	GG
228	GA	GG	CT	TT	GG
229	GG	GC	TT	TT	GA
231	GG	GG	TT	TT	
235	GG	GG	CC	CC	GG
237	NN	GG	CC	TT	GG
239	AA	GG	CC	TT	GG
240	GA	GG	CT	TT	GG
241	GA	GG	CC	TT	
242	AA	GG	CT	TT	GG
243	GA	GC	CT	TT	GG
244	GG	GC		TT	GG
245	GG	GG	CC	CC	GG

APPENDIX 5

Genotype information obtained by indirect genotyping of four single nucleotide polymorphisms. General population samples are shown in bold text.

Samples	-5116	127	407	15323
101	GG	CC		
105	GA	CC		
106	GA	CC		
108	AA	CC	GG	
111	AA			
112	AA			
113	AA			CC
114	GA	CC		
116	GG			
117		TT		
119	GG	CT		CC
120	AA	CT		
123	AA	CT		
124		CC		
125	GA	CT	GG	CC
126	GA	CC	GG	CT
127	GG	CT		
128	GA	CC	GA	CC
129	GA			
130	AA	CC		
131	GG		GG	CC
135			GG	
137		CC		CC
138	AA			
142	AA	CT		
145	GA	CT	GG	CC
146	GA	CC		CC
147	GA	CT	GG	CC
148	AA	CT	GG	CC
149	GA	CT	GG	CC
150	GA	CC	GG	CC
151	AA	CT	GG	CC
152		CT	GG	CC
153	GA	CC	GG	CT
154	GA	CT	GG	CC
155	GA	CT	GG	CC
156	GA	CC	GG	CC
157	GA	CT	GG	CC
158	GA	CT	GG	CC
159		CT	GG	CC
160	GA	CC	GG	CC
161	GG	CC	GG	CC
162	GA	CT		
163	GA	CC		CC
164		CC	GG	CT

166		CT	GA	CC
167		CC		CC
169	GA	CC	GG	CC
170	GG	CT		CC
171	GA	CC	GG	CC
172	GG	CC	GG	CC
173		CC		CT
175	GA	CC	GG	CC
176	GG	CC		
177		CC	GG	CC
179	GG	CC		CC
180	AA	CC	GG	CC
181	AA	CC	GG	CC
182	GA	CC		CC
183	AA	CT	GG	CC
184	GA	CC	GG	CC
185		TT	GA	
186				CC
187	GA	CT	GA	CC
188	GA	CC	GG	CT
189	GA	CC	GG	CT
190	GA	CT	GG	CC
191	GA	CC	GG	CC
192	GA	CC	GG	CC
193	GA	CC	GG	CC
194	GA	CT	GG	CC
195	GA	CC	GG	CT
196	GA	CC	GG	
197	GA	CC	GG	CC
198	GA	CT	GG	CC
199	GA	CC	GG	CC
200	GG	CC	GG	CC
201	GG	CT	GG	
203	AA	CC	GG	CC
205	GA	CT	GA	CC
143			**GG**	
144	**GA**	**CT**	**GA**	**CC**
206	**GG**	**CT**	**GG**	**CC**
207	**GA**	**CT**	**GG**	**CC**
209	**AA**	**CT**	**GG**	**CT**
210	**GA**	**CC**	**GG**	**CT**
211	**AA**	**CC**	**GG**	**CC**
212	**GA**	**CC**	**GG**	**CC**
213	**GA**	**CC**		**CC**
214	**GG**	**CC**	**GG**	**CC**
215	**GG**	**CC**	**GG**	**CC**
216	**AA**	**CT**	**GG**	**CC**
217	**AA**	**CC**		
218	**AA**	**CC**	**GG**	**CC**
219	**AA**	**CC**	**GG**	**CC**
221	**GG**	**CC**	**GG**	**CT**
222	**AA**	**CC**	**GG**	**CC**

223	**AA**	**CC**	**GG**	**CC**
224	**AA**	**CC**	**GA**	**CC**
225	**AA**	**CC**	**GG**	**CT**
226	**AA**	**CC**	**GG**	
227	**GA**	**CC**	**GG**	**CC**
228	**GA**	**CC**	**GG**	**CC**
229	**GG**	**CC**	**GG**	**CC**
230	**AA**	**CC**	**GG**	**CC**
231	**GG**	**CC**	**GA**	
232	**GA**	**CC**	**GG**	**CC**
233	**GA**	**CC**	**GG**	
234	**GA**	**CC**	**GA**	**CC**
235	**GG**	**CC**	**GG**	**CC**
236	**GG**	**CC**	**GG**	**CC**
237		**CC**	**GG**	
238	**AA**	**CC**	**GG**	
239	**AA**	**CC**	**GG**	**CC**
240	**GA**	**CC**	**GG**	**CC**
241	**GA**	**CT**	**GG**	**CC**
242	**AA**	**CC**	**GG**	**CT**
243	**GA**	**CC**	**GG**	**CC**
244	**GA**	**CC**	**GG**	**CC**
245	**GG**	**CC**		**CC**
312	**AA**	**CC**	**GG**	
525 028	GA	CT	GG	
525 160	GG	CC	GG	
525 171	GA	CC		
525 298	GA	CC	GG	CT
525 301	GG	CC	GG	CT
525 316	GA	CC	GG	
525 327	GG	CC	GG	
525 343	GG	CC	GG	
536 015	GA	CT	GG	
536 031	AA	CC	GA	
536 107	GG	CC	GG	
536 121	GA			
536 149	GA	CC	GG	
536 173	GA	CC	GG	CC
541 036	GA		GG	CC
541 049	GA	CC	GA	CC
541 062	GA	CC	GG	CC
541 098	GA	CC	GG	CC
541 115	GA	CC	GG	CC
541 131	GA	CC	GG	CC
541 144	GA	CC	GG	CC
541 178	GA	CC	GG	CT
541 180	GA	CC	GG	CC
541 193	GG	CC	GA	CC
541 228	GA	CC		CC
541 234		CC		CC
541 242		CC	GG	CC
541 256		CC		CC

541 353	GA	CT		CT
615 015	GA	CC	GG	CT
615 026	GA	CT	GG	CT
615 031	GA	CT	GA	CT
615 044	AA	CT	GG	CC
615 059	GA	CT	GG	CT
615 067		CT		CT
615 078	AA	CC	GG	CC
615 080	GG	CC	GG	CC
615 093	GA	CC	GA	CC
615 107	GA	CC	GA	CC
615 110	GA	CC	GG	CC
615 121	GA	CC		CC
615 136	AA	CC	GG	CT
615 325	GA	CC	GG	CC
615 332	GA	CT	GG	CT
615 340	AA	CC	GG	CC
615 358	AA	CC		CC
615 366	GA	CC	GA	CC
615 377	GA	CC	GG	CC
615 381	AA	CC	GG	CC
615 394	GA	CC	GG	CC
615 406	GG	CC	GA	CC
616 017	GA	CC	GG	CC
616 042	AA	CC	GG	CC
616 091	AA	CC		CC
616 445	GA	CC	GG	CC
616 453	AA	CC	GG	CC
616 457	GA	CC	GG	CC
616 472	GG	CC	GG	CC
616 486	GG	CC	GG	CC
616 499	GA	CC	GG	CC
616 503	GA	CC	GG	CC
304	GA	CC	GG	CC
305	GA	CT	GG	CC
306	GA		GG	CT
307		CC	GG	CC
308	AA	CC	GG	CC
309	GG	CT	GG	CC
310	GA	CT	GG	
311	AA	CT	AA	CC
313	GG	CC	GG	CC
314	GG	CC	GG	CC
315	GA	CC	GG	
317	GG	CT	GG	CC
318	GA	CT	GG	CC
319	AA	CC	GG	CC
320	AA	CC	GG	
321	GG	CT	GG	CC
322	GG	CT	GG	CC
323	GA	CT	GG	CC
324	AA	CT	GG	

325	GG		GG	
326	GA	CC	GG	
328	GA	CT	GA	CC
329	GA	CT	GA	
330	GG	CT	GG	
331	GG	CC	GG	CC
332	GG	CT	GG	CC
333	AA	CT	GA	CC

REFERENCES

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., and Kruglyak, L. 2004. Population history and natural selection shape patterns of genetic variation in 132 Genes. *PLoS Biology* **2**:1591-1599.

Ardlie, K.G., Kruglyak, L., and Seielstad, M. 2002. Patterns of linkage disequilibrium in the human genome. Nature Reviews **3**: 299-399.

Besnier, C., Takeuchi, Y., and Towers, G. 2002. Restriction of lentivirus in monkeys. *Proceedings of the National Academy of Sciences of the USA* **99**:11920-11925.

Chatterji, U., Bobardt, M.D., Gaskill, P., Sheeter, D., Fox, H., and Gallay, P.A. 2006. TRIM5α accelerates degradation of cytosolic capsid associated with productive HIV-1 entry. *The Journal of Biological Chemistry*. **281**:37025-37033.

Clark, P.A., and O'Brien, K. 2003. Fighting AIDS in Sub-Saharan Africa: is a public-private partnership a viable paradigm?. *Medical Science Monitor* **9**:ET28-39.

Cowan, S., Hatziioannou, T., Cunningham, T., Muesing, M.A., Gottlinger, H.G., and Bieniasz, P.D. 2002. Cellular inhibitors with Fv1-like activity restrict human and simian immunodeficiency virus tropism. *Proceedings of the National Academy of Sciences of the USA*. **99**:11914-11919.

Crow, J.F. 1986. Basic concepts in population, quantitative, and evolutionary genetics. W.H. Freeman, New York.

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

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

Falconer, D.S., and Mackay, T.F.C. 1996. Introduction to quantitative genetics. Longman, Essex.

Galvani, A.P., and Slatkin, M. 2003. Evaluating plague and smallpox as historical selective pressures for the CCR5-Δ32 HIV-resistant alelle. *Proceedings of the National Academy of Sciences of the USA*. **100**:15276-15279.

Goff, S.P. 2004a. Genetic control of retrovirus susceptibility in mammalian cells. *Annual Reviews Genetics*. **38**:61-85.

Goff , S.P. 2004b. Retrovirus restriction factors. *Molecular Cell*. **16**:849-859.

Goldschmidt, V., Bleiber, G., May, M., Martinez, R., Ortiz, M., Telenti, A., and The Swiss HIV Cohort Study. 2006. Role of common human TRIM5α variants in HIV-1 disease progression. *Retrovirology*. **3**:54.

Hatziioannou, T., Cowan, S., Goff, S.P., Bieniasz, P.D., and Towers, G.J. 2003. Restriction of multiple divergent retroviruses by Lv1 and Ref1. *The EMBO Journal*. **22**:385-394.

Javanbakht, H., Diaz-Griffero, F., Stremlau, M., Si, Z., and Sodroski, J. 2005. The contribution of RING and B-box 2 domains to retroviral restriction mediated by monkey TRIM5α. *The Journal of Biological Chemistry*. **280**:26933-26940.

Javanbakht, H., An, P., Gold, B., Petersen, D.C., O'Huigin, C., Nelson, G.W., O'Brien, S.J., Kirk, G.D., Detels, R., Buchbinder, S., *et al*. 2006. Effects of human TRIM5α polymorphisms on antiretroviral function and susceptibility to human immunodeficiency virus infection. *Virology*. **354**:15-27.

Johnson, A.D., Wang, D., and Sadee, W. 2005. Polymorphisms affecting gene regulation and mRNA processing: Broad implications for pharmacogenetics. *Pharmacology and Therapeutics*. **106**:19-38.

Jorde, L.B., Bamshad, M., and Rogers, A.R. 1998. Using mitochondrial and DNA markers to reconstruct human evolution. *BioEssays*. **20**:126-136.

Jorde, L.B. 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Research*. **10**:1435-1444.

Jorde, L.B., Watkins, W.S., and Bamshad, M.J. 2001. Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics*. **10**:2199-2207.

Keckesova, Z., Ylinen, L.M.J., and Towers, G.J. 2004. The human and African green monkey TRIM5α genes encode Ref1 and Lv1 retroviral restriction factor activities. *Proceedings of the National Academy of Sciences of the USA*. **101**:10780-10785.

Keele, B.F., Van Heuwerswyn, F., Li, Y., Bailes, E., Takehisa, J., Santiago, M.L., Bibollet-Ruche, F., Chen, Y., Wain, L.V., Liegeois, F., *et al*. 2006. Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. *Science*. **313**:523-526.

Kimmel, M. 1999. Population dynamics coded in DNA: genetic traces of the expansion of modern humans. *Physica A*. **273**:158-168.

Kodama, T., Kawahara, T., Wahlberg, B., Hartman, A., Rajakumar, P. and Murphy-Corb, M. 2005. Polymorphism in rhesus macaque TRIM5 associated with an innate susceptibility to SIV infection and disease progression. *12th conference on retroviruses and opportunistic infections*. Paper # 151LB.

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

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

Liu, H., Wang, Y., Liao, C., Kuang, Y., Zheng, Y., and Su, B. 2005. Adaptive evolution of primate TRIM5α, a gene restricting HIV-1 infection. *Gene*. **362**: 109-116.

Maca-Meyer, N., Gonzalez, A. M., Larruga, J. M., Flores, C. and Cabrera, V. M. 2001. Major genomic mitochondrial lineages delineate early human expansions. *BMC Genetics* **2**:13

Mercier, B., Ferec, C., Dufosse, F., and Huart, J.J. 1992. Improvement in HLA-DQB typing by PCR-RFLP: introduction of a constant restriction site in one of the primers for digestion control. *Tissue antigens*. **40**:86-89.

Münk, C., Brandt, S.M., Lucero, G., and Landau, N.R. 2002. A dominant block to HIV-1 replication at reverse transcription in simian cells. *Proceedings of the National Academy of Sciences of the USA*. **99**:13843-13848.

Nakayama, E.E., Miyoshi, H., Nagai, Y., and Shioda, T. 2005. A specific region of 37 amino acid residues in the SPRY (B30.2) domain of African Green monkey TRIM5α determines species-specific restriction of simian immunodeficiency virus SIVmac infection. *Journal of Virology*. **79**:8870-8877.

NELSON MANDELA/HSRC STUDY OF HIV/AIDS, South African National HIV Prevalence, Behavioural Risks and Mass Media. Household Survey 2002.

Nielsen, M.H., Pedersen, F.S., and Kjems, J. 2005. Molecular strategies to inhibit HI-1 replication. *Retrovirology*. **2**:10.

Nomura, N., Inoko, H., Kato, S., Arimori, S., Ota, M., and Tsuji, K. 1991. PCR-RFLP: a new HLA-DNA typing method tested in bone marrow transplantation. *Transplant proceedings*. **23**:431-433.

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

Ortiz, M., Bleiber, G., Martinez, R., Kaessmann, H., and Telenti, A. 2006. Patterns of evolution of host proteins involved in retroviral pathogenesis. *Retrovirology*. **3**:11.

Owens, C.M., Yang, P.C., Göttlinger, H., and Sodroski, J. 2003. Human and Simian Immunideficiency Virus Capsid Proteins Are Major Viral Determinants of Early, Postentry Replication Blocks in Simian Cells. *Journal of Virology*. **77**:726-731.

Paraskevis, D., Lemey, P., Salemi, M., Suchard, M., Van de Peer, Y., and Vandamme, A., 2003. Analysis of the evolutionary relationships of HIV-1 and SIVcpz sequences using bayesian inference: Implications for the origin of HIV-1. *Molecular Biology and Evoloution*. **20**:1986-1996.

Parra, J.E., Marcini, A., Akey, J., Martinson, J., Batzer, M.A., Cooper, R., Forrester, T., Allison, D.B., Deka, R., Ferrell, R.E., *et al*. 1998. Estimating African American admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*. **63**:1839-1851.

Perez-Caballero, D., Hatziioannou, T., Yang, A., Cowan, S., and Bieniasz, P.D. 2005. Human tripartite motif 5α domains responsible for retrovirus restriction activity and specificity. *Journal of Virology*. **79**:8969-8978.

Perron, M.J., Stremlau, M., Song, B., Ulm, W., Mulligan, R., and Sodroski, J. 2004. TRIM5α mediates the postentry block to N-tropic murine leukemia viruses in human cells. *Proceedings of the National Academy of Sciences of the USA*. **101**:11827-11832.

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

Reymond, A., Meroni, G., Fantozzi, A., Merla, G., Cairo, S., Luzi, L., Riganelli, D., Zanaria, E., Messali, S., Cainarca, S., *et al*. 2001. The tripartite motif family identifies cell compartments. *The EMBO Journal*. **20**:2140-2151.

Rice, W.R. 1989. Analyzing tables of statistical tests. *Evolution*. **43**:223-225.

Risch, N., Burchard, E., Ziv, E., and Tang, H. 2002. Categorization of humans in biomedical research: genes, race and disease. *Genome Biology*. **3**:comment 2007.1-2007.12.

Sawyer, S.L., Wu, L.I., Emerman, M., and Malik, H.S. 2005. Positive selection of primate TRIM5α identifies a critical species-specific retroviral restriction domain. *Proceedings of the National Academy of Sciences of the USA*. **102**:2832-2837.

Sayah, D.M., Sokolskaja, E., Berthoux, L., and Luban, J. 2004. Cyclophilin A retrotransposition into *TRIM5* explains owl monkey resistance to HIV-1. *Nature*. **430**:569-573.

Sebastian, S., and Luban, J. 2005. TRIM5α selectively binds a restriction-sensitive retroviral capsid. *Retrovirology*. **2**:40.

Shisana, O., and Simbayi, L.C. 2002. Nelson Mandela/HSRC study of HIV/AIDS: South African nation HIV prevalence, behavioural risks and mass media: household survey. Human Sciences Research Council. Cape Town

Speelmon, E.C., Livingston-Rosanoff, D., Li, S.S., Vu, O., Bui, J., Geraghty, D.E., Zhao, L.P., and McElrath, M.J. 2006. Genetic Association of the Antiviral Restriction Factor TRIM5α with Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology*. **80**: 2463-2471.

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

Stoneking, M., and Soodyall, H. 1996. Human evolution and the mitochondrial genome. *Current Opinion in Genetics and Development*. **6**:731-736.

Stremlau, M., Owens, C.M., Perron, M.J., Kiessling, M., Autissier, P., and Sodroski, J. 2004. The cytoplasmic body component TRIM5α restricts HIV-1 infection in Old World monkeys. *Nature*. **427**:848-853.

Stremlau, M., Perron, M., Welikala, S., and Sodroski, J. 2005. Species-specific variation in the B30.2 (SPRY) domain of TRIM5α determines the potency of human immunodeficiency virus restriction. *Journal of Virology*. **79**:3139-3145.

Teare, M.D., Dunning, A.M., Durocher, F., Rennart, G., and Easton, D.F. 2002. Sampling distribution of summary linkage disequilibrium measures. *Annals of Human Genetics*. **66**:223-233.

Telenti, A. 2005. Adaption, co-evolution, and human susceptibility to HIV-1 infection. *Infection, Genetics and Evolution*. **5**:327-334.

Tishkoff, S.A., and Verrelli, B.C. 2003. Role of evolutionary history on haplotype block structure in the human genome: implications for disease mapping. *Current Opinion in Genetics and Development*. **13**:569-575.

UNAIDS. 2006. AIDS epidemic update, December .

Wang, W., Chen, M., Chuang, C., Jeang, K., and Huang, L. 2000. Molecular biology of human immunodeficiency virus type 1. *Journal of Microbiology, Immunology and Infection*. **33**:131-140.

Watkins, W.S., Rogers, A.R., Ostler, C.T., Wooding, S., Bamshad, M.J., Brassington, A.E., Carroll, M.L., Nguyen, S.V., Walker, J.A., Prasad, B.V.R., *et al*. 2003. Genetic variation among world populations: Inferences from 100 Alu insertion polymorphisms. *Genome Research*. **13**:1607-1618.

Webber, J. 2001. The pathogenesis of HIV-1 infection. *British Medical Bulletin*. **58**:61-72.

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

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

Yap, M.W., Nisole, S., Lynch, C., and Stoye, J.P. 2004. Trim5α protein restricts both HIV-1 and murine leukemia virus. *Proceedings of the National Academy of Sciences of the USA*. **101**:10786-10791.

Zetterquist, H., and Ollerup, O.1992. Identification of the *HLA-DRB1*04*, -*DRB1*07*, and -*DRB1*09* alleles by PCR amplification with sequence-specific primers (PCR-SSP) in 2 hours. *Human Immunology*. **34**:64-74.

Zheng, Y., and Peterlin, B.M. 2005. Intracellular immunity to HIV-1: newly defined retroviral battles inside infected cells. *Retrovirology*. **2**:25.

Zietkiewicz, E., Yotova, V., Jarnik, M., Korab-Laskowska, M., Kidd, K.K., Modiano, D., Scozzari, R., Stoneking, M., Tishkoff, S., Batzer, M., *et al*. 1997. Nuclear DNA diversity in worldwide distributed human populations. *Gene*. **205**:161-171.