# VARIATION IN *ABCB1* AND ITS EFFECT ON IMMUNE RECOVERY WITH ANTIRETROVIRALS

**Ingrid Marie Du Plooy**

A thesis submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the Degree of Doctor of Philosophy.

Johannesburg, 2011

**TABLE OF CONTENTS**

## DECLARATION

I declare that this research is my own, unaided work. It is being submitted for the Degree of Doctor of Philosophy at the University of the Witwatersrand, Johannesburg, South Africa. It has not been submitted before for any degree or examination at any other University.

…….

Ingrid Marie Du Plooy

…….…day of ……..2011

## ACKNOWLEDGEMENTS

Firstly, I would like to express my huge gratification to my supervisor, Tracy McLellan for her support, guidance and patience over the last few years, as well as my co-supervisors, Eric Dabbs and Collet Dandara, for their valuable input.

I would like to thank all the members of the Population Genetics Lab from the last 5 years for their help and friendship; members of the Microbial Genetics Lab for helping with restriction digest troubleshooting; members of the Fly Lab for help with the RNA work and general help with Lab work; members of the Cell Biology Lab for help with the RNA work and members of the Plant Biotech Lab for help with the Q-PCR work.

I would also like to thank Dr. Francois Venter for his help in collecting samples at Johannesburg General Hospital, Prof. Patrick MacPhail, Dr. Thapelo Maotoe & Lynn McNamara for their help in collecting samples at the Themba Lethu Clinic at Helen Joseph Hospital in 2005 and 2009, and Clive Gray at NICD for the Cohort samples.

My thanks go to Inqaba Biotech for all the sequencing they did for me, to Dr. Steven Orzack & Andries Oelofse for their help with the haplotype analysis, to Dr. Mhairi Maskew for retrieving the data from the TherapyEdge-HIV$^{TM}$ database, and to Dr. Nicole Soranzo for providing her genotypic data for reanalysis.

I would like to thank all the individuals who donated their blood for DNA and/or RNA isolation, without whom the project would not have been possible; the AIDS Research Institute, National Research Fund and Postgraduate Merit Award from the University of the Witwatersrand for funding.

Lastly I would like to say a huge ‘Thank You’ to my husband, the rest of my family and my friends for all their love, support and encouragement.

## ABSTRACT

The *ABCB1* gene encodes P-glycoprotein, a transmembrane protein that regulates the efflux of drugs in the cells and may affect the response to antiretroviral drugs. *ABCB1* polymorphisms affect the function or expression of P-gp. The 3435T allele has been associated with decreased protein production, but is in linkage disequilibrium with other polymorphisms. HIV is prevalent in Southern Africa, and characterization of *ABCB1* variation may provide insight into its role in antiretroviral immune response. The aim was to determine if there was any association between *ABCB1* variation, relative mRNA levels and immune response. Seven known polymorphisms were characterized for linkage disequilibrium and haplotype analysis, regions upstream of the gene were sequenced for bioinformatic analysis, the relative amounts of mRNA were determined, and $CD4^+$ and viral load data was analyzed for association. Sequencing revealed six novel variations: T-137G, C-233T and G-298A upstream of exon 1, T108G and G153A in exon 2, and A111G in intron 26. The frequencies of the -129T (0.85), 1236T (0.70), 2677G (0.77), IVS 25+3050G (0.86), IVS 25+5231T (0.51), 3435C (0.88) and IVS 26+80T (0.89) polymorphisms were different and LD was lower compared to other populations. The haplotype frequencies were different to other populations and the genetic structure was probably a result of multiple recombination or mutation events. The viral load counts at the second measurement after baseline (time point 2) were significantly different from baseline for the 2677GG and 2677GA genotypes, and the -129T allele was associated with a lower proportional decrease in viral load at

the second measurement. The IVS 25+3050GG, 3435CC and IVS 26+80TT genotypes have been associated with lower mean relative mRNA levels. In conclusion, the genetic structure of the southern African populations is different from other populations and that genetic association and functional studies derived from other populations would be irrelevant in this population. A larger sample size and functional studies would be required to attempt to resolve the molecular mechanisms of the *ABCB1* gene and to confirm the findings of association between *ABCB1* polymorphisms and immune response.

## LIST OF ABBREVIATIONS

| | |
|---|---|
| ABCB1 | ATP-binding cassette family, subfamily B gene 1 |
| AIDS | Aquired immune deficiency syndrome |
| ANOVA | Analysis of variation |
| ART | Antiretroviral therapy |
| ATP | Adenosine triphosphate |
| B2M | Beta-2-microglobulin gene |
| bp | Base pair |
| cDNA | Copy DNA |
| $C_t$ | Threshold cycle |
| CYP2D6 | Cytochrome P450 2D6 gene |
| CYT2C9 | Cytochrome 2C9 gene |
| DF | Degrees of Freedom |
| DNA | Deoxyribonucleic acid |
| EDTA | Ethylenediaminetetraacetic acid |
| GAPDH | Glyceraldehyde 3-phosphate dehydrogenase |
| GWAS | Genome wide association studies |
| HAART | Highly active antiretroviral therapy |
| HIV | Human immunodeficiency virus |
| iMED | Inverted multiple start site element downstream 1 |
| LD | Linkage disequilibrium |
| LDA | Linkage disequilibrium analyser |
| MDR1 | Multidrug resistance gene 1 |

| | |
|---|---|
| mRNA | Messanger RNA |
| NNRTI | Non-nucleoside reverse transcriptase inhibitor |
| NRTI | Nucleoside (or nucleotide) reverse transcriptase inhibitor |
| PCR | Polymerase chain reaction |
| P-gp | P-glycoprotein |
| PI | Protease inhibitor |
| RFLP | Restriction fragment length polymorphism |
| RNA | Ribonucleic acid |
| RQ | Relative quantity |
| SNP | Single nucleotide polymorphism |
| TPMT | Thiopurine methyltransferase gene |
| U | Units (of enzyme) |
| UTR | Untranslated region |
| VKORC1 | Vitamin K epoxide reductase complex subunit 1 |

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1: INTRODUCTION

## 1.1 Pharmacogenetics

Pharmacogenetics is the investigation of the relationship between genetic variation and individual differences in drug response. Although there are strict regulations for clinical trials that must be followed to ensure drugs are effective and safe, some individuals may have adverse reactions to the drugs, or the drugs may have no effect (Roses 2000a). Variation in genes encoding proteins, such as drug-metabolizing enzymes, transporters and receptors, influence the response of an individual to different drugs (Gwee et al. 2003). The goal of pharmacogenetics is to improve the understanding of how genetic variation can influence drug response, to ensure treatment is effective without the adverse side effects, and to find new methods of treating or preventing diseases (Schmith et al. 2003; Javitt and Hudson 2007; Marsh 2008; Löscher et al. 2009).

Genetic variation is an important factor in individual response to certain drugs and their susceptibility to the toxicity of these drugs (Stanger et al. 2003; Martin et al. 2004). Knowing how the variation causes these changes and being able to determine individual variation will be important for individualizing medicine, based on the genetic make-up of an individual, and increasing the rate of drug development (Roses 2000b).

SNPs in the human population are of particular interest for studying human evolution, determining the mechanisms of maintaining genetic variability in the population and for identifying genes associated with complex diseases (Yu et al. 2002). SNPs and other genetic factors provide information on how an individual responds to a pathologic disease and how drugs can be made to suit the individual. Genetic variation in drug-metabolizing enzymes, transporters, receptors and other genetic factors influence the response of an individual to different drugs (Gwee et al. 2003).

The most common type of polymorphism (sequence variation) found in the genome are single nucleotide polymorphisms (SNPs), which are single base substitutions found between individuals. They are found on average every 500 nucleotides in a genome (Van Tassell et al. 2008). Most polymorphisms found in the genome do not change the structure or function of the protein, but some do, through amino acid changes. SNPs are characterized according to their location and the effect they have on the expressed protein. Random SNPs are the most common and occur in introns of genes and between genes. Coding SNPs are found in the exons of genes and non-coding SNPs occur outside the coding region of a gene in the promoter or 5′ or 3′ untranslated regions (UTR). Many non-coding SNPs are also regulatory SNPs, which have a role in controlling the expression of genes and evolutionary change. Synonymous SNPs are found in exons and do not cause a change in the amino acid sequence. They usually occur at the third (wobble) position of the codon (Cargill et al. 1999; Collins et al. 1999; Knight 2004).

Human genetics uses two approaches: one approach uses populations of individuals with a disease to determine genes that make those individuals more susceptible to that disease; the second approach uses DNA sequence databases to identify genes and gene families that can be screened for their possible role in a disease (Roses 2000b).

Single genes involved in rare diseases include the cystic fibrosis gene (located on chromosome 7) and a gene on chromosome 4 that has been related to Huntington's disease. Variation in a number of genes has been associated with immune response: variation in the thiopurine methyltransferase gene (TPMT), has been associated with toxicity with azathiopurine drugs, variation in the Cytochrome P450 2D6 (CYP2D6) gene has been associated with the outcome of tamoxifen treatment in breast cancer (Marsh 2008), and variation in the Cytochrome 2C9 (CYT2C9) and vitamin K epoxide reductase complex subunit 1 (VKORC1) genes have been associated with warfarin sensitivity or resistance (Javitt and Hudson 2007).

## 1.2 HIV and Antiretrovirals

The Human Immunodeficiency Virus (HIV), a retrovirus, causes Acquired Immune Deficiency Syndrome (AIDS) by invading cells with $CD4^+$ receptors, including macrophages and helper T cells. HIV encodes its own reverse transcriptase, which it uses to produce DNA from its RNA. The DNA is then incorporated into the genome of the infected cell, where it is replicated when the cell replicates its own DNA. The replicated viral DNA is then translated into viral proteins for viral assembly, followed by viral release and cell death (Lal et al. 2005).

Antiretroviral drugs consist of five groups: nucleoside (or nucleotide) reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs), integrase inhibitors, fusion inhibitors and protease inhibitors (PIs), which bind and inhibit the HIV proteins reverse transcriptase and protease. NRTIs, such as zidovudine (AZT), are compounds that bind the HIV reverse transcriptase at the substrate binding site and inhibit HIV replication. NNRTIs, such as nevirapine and efavirenz, bind the HIV reverse transcriptase at a non-substrate binding site to inhibit HIV replication (Travel et al. 1999). PIs, such as ritonavir, nelfinavir and indinavir, prevent the cleavage of the GAG and GAG-POL precursors of HIV protease to form functional proteins (including reverse transcriptase, integrase and protease), preventing HIV maturation (Flexner 1998; Travel et al. 1999; Dickinson et al. 2005; Pozio et al. 2005). Combining NRTIs, NNRTIs and PIs in antiretroviral treatment leads to the improved inhibition of

both HIV reverse transcriptase and HIV protease, thereby increasing $CD4^+$ T cell count, decreasing of viral load and delaying disease progression (Henry et al. 1998). Highly active antiretroviral therapy (HAART) is a combination therapy using at least three antiretroviral drugs: two NRTIs, one NRTI and one or two PIs (Van Heeswijk et al. 2002).

HIV-1 protease is essential to the virus for its replication and the maturation of viral particles. Protease inhibitors prevent the post-translational cleavage of the long peptide precursors and thereby inhibit viral assembly (Dickinson et al. 2005; Pozio et al. 2005). The eleven current protease inhibitors are amprenavir, ritonavir, nelfinavir, darunavir, indinavir, saquinavir, lopinavir, norvir, atazanavir, fosamprenavir and tipranavir (UNAIDS 2008).

The high mutation rate of the virus has necessitated the use of several drugs simultaneously to prevent drug resistance. The combination therapy has some disadvantages, such as the serious side effects, response rate, cost and the strict protocol that must be followed. Because of the disadvantages of combination therapy and the prevalence of HIV infections, other areas are being investigated in the fight against the AIDS pandemic (Lal et al. 2005). These include research into AIDS restricting genes and antiretroviral drug transporters.

### 1.3 *ABCB1* and P-glycoprotein

The multidrug resistance-1 gene (*ABCB1* or *MDR1*) is the first of seven members of the ATP-binding cassette family, subfamily B. The gene was first identified in cancer cells that showed resistance to a variety of structurally unrelated drugs, and was therefore named multidrug resistance gene (Sukhai et al. 2000). It was observed that cancer chemotherapy failed due to overexpression of *ABCB1* (Labialle et al. 2002).

The *ABCB1* gene, which is located on chromosome 7q21.1, is approximately 210 kilobases (kb) in size and is composed of a core promoter region preceeding 29 exons (Bodor et al. 2005). *ABCB1* homologs have been reported in a variety of organisms, including humans, fruit flies, yeast and some plants. A large number of single nucleotide polymorphisms (SNPs) has been found throughout the gene, in both coding and non-coding sequences (Schaeffeler et al. 2001).

*ABCB1* functions to regulate the movement of drugs, peptides and foreign substances (xenobiotics) in the cell. It does this by encoding P-glycoprotein, a 170 kDa transmembrane transporter (Taniguchi et al. 2003; Tang et al. 2004), that acts as an efflux pump. The P-glycoprotein is a phosphorylated and N-terminally glycosylated energy-dependant efflux pump made up of 1280 amino acids (Schwab et al. 2003). P-glycoprotein is composed of two homologous, symmetrical domains, each domain comprising six transmembrane regions and an ATP-binding motif. ATP hydrolysis is performed to provide the energy

required to actively pump a variety of hydrophobic, amphipathic compounds against a steep gradient, from inside the cell to the extracellular domain (Kim et al. 2001). Membrane transporters play a role in providing drug targets and physiological processes, such as maintaining homeostasis in cells and organisms. They all have similar secondary structures, multiple membrane-spanning regions separated by intracellular and extracellular loops (Leabman et al. 2003).

The substrates that bind to the transporter have a broad specificity (Bodor et al. 2005) and include hormones, plant-derived chemicals (Stenger et al. 2003) and more than 50 commonly used drugs, including HIV-1 protease inhibitors. The affinity with which the substrates bind P-glycoprotein depends on their hydrogen bonding and their affinity for lipids (Soranzo et al. 2004; Zhu et al. 2004). The mechanisms by which P-glycoprotein acts on HIV immune recovery are unclear, although a few suggestions have been made (Owen et al. 2005). There maybe a relationship between P-gp and the expression of HIV-relevant proteins that affect the cell's susceptibility to infection (Owen et al. 2004), P-gp might act directly on the virus by affecting viral replication or infection (Lee et al. 2000), or it might act on other proteins and cholesterol (Xiao et al. 1998; Speck et al. 2002).

P-glycoprotein is expressed in a variety of tissues, such as the kidney, liver, small intestine and brain, and cells of the immune system, including mature macrophages, natural killer cells, antigen presenting dendritic cells and T and B lymphocytes. The function of the protein is to protect the body from toxins by excreting them into the bile, urine and intestinal lumen, to regulate oral

absorption of drugs, to help in the defence against viral infections and to play a role in steroid metabolism (Hoffmeyer et al. 2000; Schaeffeler et al. 2001).

### 1.4 *ABCB1* Promoter Region

The *ABCB1* gene has two promoters and two transcriptional start sites. The first promoter, located in exon 1 -434 to +1 in relation to exon 1b, does not have a TATA box, is undefined and believed to be cryptic (Cornwell et al.1993; Scotto et al. 2003; Takane et al. 2004). Cornwell et al (1993) reported that the region did not have a high CpG content, whereas Takane et al. (2004) reported that the region is CpG-rich. CpG sites in promoter regions are sites for DNA methylation, which is the most common form of DNA modification in eukaryotes (Takane et al. 2004). Hypomethylation of the *ABCB1* gene results in elevated gene expression (Takane et al. 2004).

The promoter is regulated by an initiator element (Scotto et al. 2003). A 13 bp region surrounding the proximal initiator element (-6 to +7) has been related to accurate initiation *in vitro* and the sequence downstream of the initiator site has been shown to be important in the efficiency of promoter activity (Cornwell et al.1993). A G-rich region in the proximal promoter (-121 to -88) has been shown to bind to transcriptional activators and repressors (Cornwell et al.1993).

The second promoter, which is in exon 2, drives gene expression in both normal and tumour tissues (Cornwell et al.1993; Wang et al. 2006). The translational start site is situated in exon 3. As the translational start site occurs in exon 3, exons 1 and 2 are often combined as exons 1a and 1b, making exon 3 also known as exon 2 (Wang et al. 2006).

This upstream region contains a Y-box or inverted CCAAT-box (-79 to -75), which is a binding site for the transcription factor NE-Y, and a GC-box (-56 to -43), which is a binding site for the transcription factors Sp1 and Sp3. Both these elements are required for efficient transcriptional regulation and are commonly found in TATA less promoters (Taniguchi et al. 2003; Scotto et al. 2003; Takane et al. 2004). The promoter region also contains an inverted multiple start site element downstream 1 or iMED1 (-105 to -100), which has been associated with the activation of *ABCB1* in drug-resistant cells and transcription is decreased by 60% when iMED1 is functionally disrupted (Scotto et al. 2003; Labialle et al. 2002). Transcription of the *ABCB1* gene is also repressed by the binding of p53 to any of the numerous sites (Scotto et al. 2003) (Figure 1).

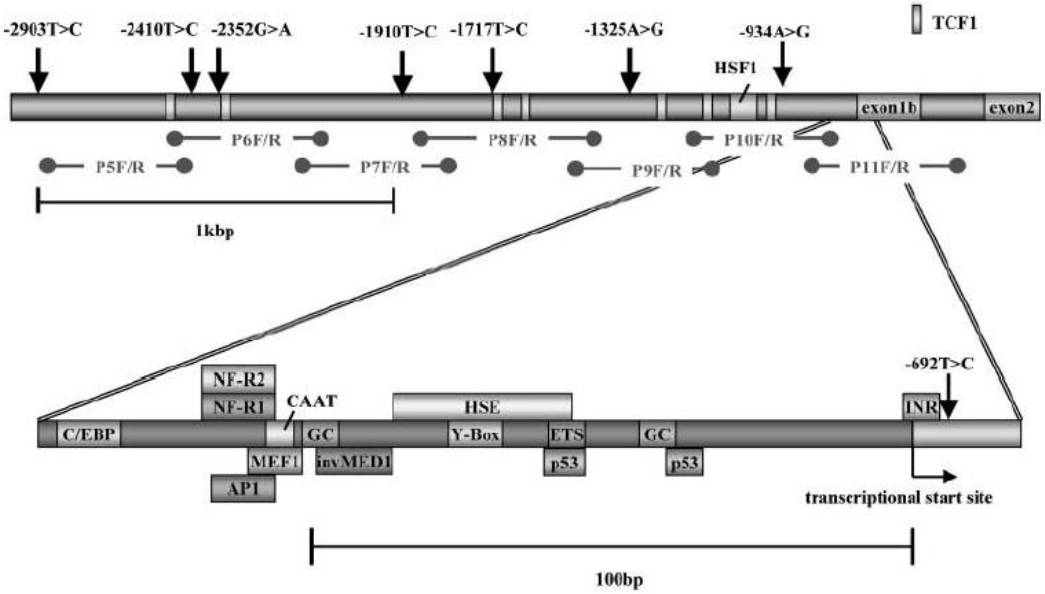


Figure 1: The location of regulatory elements within the *ABCB1* promoter region upstream of the transcriptional start site in exon 1b (adapted from Taniguchi et al. 2003).

Numerous elements have been identified as regulators of *ABCB1* transcription. Variation that interferes with the functioning of these elements or that generates other regulatory elements can play a role in the regulation of *ABCB1* gene expression (Wang et al. 2006).

Mutations in the -110 G-box (-110 to -103) in the proximal promoter region have been shown to increase promoter activity and inhibit the formation of a specific nuclear protein complex, suggesting that the region normally functions as a transcriptional repressor binding site (Cornwell et al.1993). Mutation or deletion of the -50 G-box (-61 to -43) decreases promoter activity, suggesting that the site normally enhances promoter activity, and inhibits the binding of the transcription

factor Sp1 to the -50 G-box, which is required for efficient promoter activity (Cornwell et al.1993). Mutations in the *ABCB1* promoter region, especially in the Y-box, have been shown to reduce the activity of the promoter in some drug-sensitive and drug-resistant human cell lines (Goldsmith et al. 1993).

Two specific SNPs (T-129C and A-41G) in the promoter region affect promoter activity and are associated with differential gene expression. In the presence of the major alleles (-129T and -41A) these SNPs have been associated with a decrease in promoter activity, but in the presence of the minor allele (-129C and -41G) the SNPs have been associated with an increase in promoter activity (Labialle et al. 2002; Scotto 2003; Taniguchi et al. 2003; Wang et al. 2006).

## 1.5 *ABCB1* variation and pharmacogenetics

Variation in *ABCB1* results in differences in P-glycoprotein expression in normal tissues, leading to inter-individual differences in drug response. Site-directed mutagenesis studies have been performed to introduce nucleotide changes in highly conserved regions of *ABCB1*, resulting in profound changes in P-glycoprotein expression and function (Pauli-Magnus and Kroetz 2004). Variation in the *ABCB1* gene has been associated with a variety of disorders, such as renal cell carcinoma, Parkinson's disease, inflammatory bowel diseases (e.g. ulcerative colitis and Crohn's disease), there is some debate as to the association in epilepsy (Stanger et al. 2003; Martin et al. 2004; Leschziner et al. 2007), and there has also been associated with $CD4^+$ cell recovery with antiretroviral therapy (Stanger et al. 2003; Martin et al. 2004).

To date, about 600 polymorphic sites have been identified in the almost 210 kb *ABCB1* gene. Of these polymorphic sites, the majority are found in the non-coding regions and only about 47 are found in exons. The most commonly studied SNP is the synonymous C3435T transition in exon 26 (http://www.ensembl.org, ID: ENSG00000085563), which has associated with a change in the expression and function of P-glycoprotein. The SNP results in a synonymous substitution in a codon which codes for isoleucine in P-glycoprotein, which results in a decrease in the expression of the protein. A large number of genotype-phenotype studies have been performed and many have yielded

conflicting results, for which there are a number of reasons, and the molecular basis of the *ABCB1* SNP at 3435 is still not well understood.

The *ABCB1* 3435TT genotype has been shown to result in a conserved silent ("wobble") polymorphism and corresponds to a decreased level of P-glycoprotein expression, while the 3435CC and 3435CT genotypes show comparatively higher and intermediate levels of P-glycoprotein expression, respectively (Hoffmeyer et al. 2000; Kim et al. 2001; Stanger et al. 2003). Hoffmeyer et al. (2000) concluded that the synonymous C3435T polymorphism in exon 26 was the causal site for functional variation, since the TT genotype corresponded with the lowest P-glycoprotein levels, whereas the CC genotype correlated with the highest P-glycoprotein levels. Decreased expression of P-glycoprotein causes less of the drugs to be pumped out of the cell or tissue (Fellay et al. 2002) thus, allows the drug to exert its effects for a long time at a concentration that inhibits viral replication.

Many studies have focused solely on this one polymorphic site, but Soranzo et al. (2004) questioned the importance of this site acting alone. They found three intronic polymorphisms with high levels of linkage disequilibrium with the 3435 locus, IVS 25+G3050T and IVS 25+T5231C in intron 25, and IVS 26+T80C in intron 26 and suggested that the C3435T site was probably not solely responsible for the functional variation, but that a number of linked polymorphic sites were probably the cause of the differences seen in the expression levels of P-glycoprotein.

Strong linkage disequilibrium has been found within the entire *ABCB1* gene, but especially between the SNP in exon 26 and certain other SNPs (G2677T in exon 21, G3050T and T5231C in intron 25, T80C in intron 26), making it difficult to determine which SNP is responsible for the change in protein expression. The SNP in exon 21 causes an amino acid change from alanine to serine at position 893 in the peptide. The serine variant has been shown to enhance the transportation function of P-glycoprotein when compared to the alanine variant (Kim et al. 2001). Both SNPs in exons 12 and 26 are synonymous, meaning that they do not change the encoded amino acids glycine and isoleucine, respectively (Taniguchi et al. 2003). The 2677T and 3435T alleles form the *ABCB1**2 haplotype (Kroetz et al. 2003). A number of haplotypes, groups of linked alleles that are inherited as a unit, have been found involving polymorphic sites within the *ABCB1* gene. The observed effects may be due to a single SNP (G2677T/A or C3435T) or a haplotype involving these SNPs (Allabi et al. 2005; Ardlie et al. 2002, Ding et al. 2003; Shifman et al. 2003; Bleiber et al. 2004; Takane et al. 2004).

Leschziner et al. (2007) genotyped three *ABCB1* polymorphisms (C1236T, G2677T/A and C3435T) in epilepsy patients and found no association with any of the SNPs and multidrug resistance. They suggested that the C3435T polymorphisms may not be the functional variant, but rather that it is in linkage disequilibrium, non-random association, with another site, which may be the functiona variant, and that gene-wide tagging might reveal the true finctional variant. The role of the variation in *ABCB1* in immune response is still unknown

and to facilitate the understanding of how variation is associated with immune response, the genetic structure and patterns of linkage disequilibrium are required (Leschziner et al. 2006). The use of tagged SNPs to reduce the number of sites required for genotyping will only be useful in populations with high levels of linkage disequilibrium, such as Caucasian populations, but might not be useful in populations with low levels of linkage disequilibrium, such as African populations (Leschziner et al. 2006).

A change in the tertiary conformation of P-glycoprotein could be the basis for the reduced functionality. This could be due to shortened mRNA molecules caused by the inactivation of splice donor sites by a synonymous polymorphism, which results in a premature stop codon or exon splicing (Kimchi-Sarfaty et al. 2007). Although the 1236T and 3435T SNPs alleles are synonymous substitutions, they have been shown to result in a change in the mRNA structure, whereas the 2677T nonsynonymous substitution does not (Wang and Sadée 2006). *In vitro* transfections showed that only the 3435T allele decreased mRNA levels by affecting mRNA stability.

Wang et al. (2005) found an association between the 3435C allele and higher mRNA expression levels, while the 3435T allele was associated with a decrease in mRNA stability, possibly due to effects on the secondary structure of the mRNA molecule (Moriya et al. 2002; Nakamura et al. 2002).

Single base sustitutions can cause differences in mRNA folding, which can influence splicing, processing or transcriptional regulation of the resulting protein. Reports by Kimchi-Sarfaty et al. (2007) have shown that it may not be the presence of the G2677T polymorphism that is causal, but the presence of the C3435T polymorphism in combination with one or two other SNPs due to the observation of different tertiary structures between the wild-type and haplotype P-glycoprotein (C1236T-G2677T-C3435T).

In individuals suffering from the diseases associated with variation in *ABCB1* (carcinomas, Parkinson's disease, inflammatory bowel diseases and epilepsy), the 3435T allele was found to present at a higher frequency than the 3435C allele. The 3435T allele has also been associated with a decrease in drugs being pumped out of $CD4^+$ cells and shown to increase the recovery of these cells during antiretroviral therapy (Stanger et al. 2003) and reduce susceptibility to HIV infection (Schaeffeler et al. 2001). Antiretroviral therapy (ART) involves the use of a variety of drugs and drug classes to combat the virus. The use of multiple drugs results in different drug-host and drug-drug interactions, to which genetic variations also contribute (Martin et al. 2004). HIV-1 infected individuals have shown significant variation in the effect of antiretroviral therapy, their plasma drug concentrations, the rate of immune recovery and their response to toxins. In a population of HIV positive individuals treated with antiretrovirals, such as efavirenz and nelfinavir, those with the 3435TT genotype demonstrated better $CD4^+$ T cell recovery and decreased viral load compared to those with the

3435CC genotype (Telenti et al. 2002; Nasi et al. 2003), further supporting a role for the C3435T in drug response.

Transporters also play a role in the protection of the host from viral infections. P-glycoprotein has been associated with the glycolipid-rich membrane domains, known as lipid rafts, which are sites for viral fusion and release. P-glycoprotein overexpression affects viral fusion and release, resulting in decreased susceptibility of $CD4^+$ cells to HIV infection, thus the reduction in P-glycoprotein expression that has been attributed to the 3435TT genotype (Fellay et al. 2002; Hulgan et al. 2003; Bleiber et al. 2004; Tang et al. 2004) could actually be associated with increased risk of infection. There have been suggestions that the C3435T polymorphic site may not be the causal site, but that another site, undetected as a result of linkage disequilibrium, may be the cause of the functional variation. Three intronic sites have been identified to have strong linkage disequilibrium with the C3435T site in Europeans and may be additional candidates for the influence of P-glycoprotein expression (Soranzo et al. 2004).

The polymorphism or combination of polymorphisms responsible for the observed functional variation, as well as the molecular basis of the *ABCB1* gene are still unknown. Linkage disequilibrium between SNPs makes it difficult to determine which SNP is responsible for the observed phenomenon, as the SNPs cannot be analysed seperately.

## 1.6 *ABCB1* and Antiretrovirals

Many human genes and other factors, such as sex, age and environmental determinants, influence the progression to AIDS and the response to antiretroviral therapy. The use of antiretrovirals, such as NNRTIs can result in adverse effects or cumulative toxicity, which can be life-threatening (Wilson et al. 2001; Telenti et al. 2002). Nevirapine and efavirenz have been associated with hepatotoxicity in patients with high and low $CD4^+$ T cell counts, respectively (Ritchie et al. 2006). It is therefore important to understand the mechanisms of drug toxicity in order to minimize the adverse effects and increase the efficacy of antiretroviral therapy, or to find drugs that do not have a toxic effect (Fellay et al. 2002).

There are differences between populations and also between individuals, which might affect the efficacy or toxicity of the drugs. These differences include variation in metabolism and transport of drugs. The best time to start therapy as well as the most effective drug combination might be based on a particular genotypic profile (Telenti et al. 2002; Ritchie et al. 2006). Genetic tests might be used to predict an individual's response to specific drugs and thereby allow the drug combination to be tailored to the individual's genetic makeup (Wilson et al. 2001).

Variation in the *ABCB1* gene results in different concentrations of protease inhibitors and can be used to predict the rate of recovery of the immune system after the start of antiretroviral therapy, using increased $CD4^+$ T cells as an

indication (Telenti et al. 2002). Although NNRTIs are not substrates of P-glycoprotein, there have been suggestions that P-glycoprotein is involved in the disposition of nevirapine and efavirenz, but these findings remain controversial. P-glycoprotein expression has been linked to intracellular nevirapine concentrations (Störmer et al. 2002) and the inhibition of intestinal P-glycoprotein is predicted to increase the oral bioavailability and decrease the clearance of protease inhibitors (von Heeswijk et al. 2002).

It may be that P-glycoprotein can alter the clinical course of HIV infection in a mechanism, independent of its role in drug transportation (Winzer et al. 2005). Overexpression of P-glycoprotein in $CD4^+$ T cells has been shown to result in the inhibition of HIV replication, increased efficacy of antiretrovirals and decreased viral fusion and release (Fellay et al. 2002). During the progression of the HIV disease $CD4^+$ T cell numbers decrease primarily as a result of an increase in cell loss, rather than a decrease in cell production. The $CD4^+$ T cell turnover may be influenced by P-glycoprotein expression (Fellay et al. 2002; Haas et al. 2003).

The 3435T allele has been associated with a decreased risk of hepatotoxicity with NNRTI-containing antiretroviral drug regimes and a more favourable virological response to antiretroviral drug regimes that include efavirenz (Ritchie et al. 2006). The 3435CC genotype has been associated with a decreased response to combination antiretroviral therapy (Fellay et al. 2002) as well as earlier virological failure with the use of protease inhibitors (Brumme et al. 2003). The 3435TT genotype has been associated with an increase in $CD4^+$ T cells in

patients receiving antiretroviral therapy (Haas et al. 2003) as well as lower plasma drug concentrations of nelfinavir and efavirenz (Fellay et al. 2002).

In 149 drug-naïve HIV positive patients no evidence of influence of the *ABCB1* genotype on the efficacy of antiretrovirals was found (Nasi et al. 2003). In 411 subjects no correlation was found between *ABCB1* expression and permissiveness or HIV infection, and no modification of disease progression by *ABCB1* genotypes was reported (Bleiber et al. 2004). Winzer et al. (2005) reported no significant difference in the response (virological or immunological) of 72 HIV positive patients with respect to the 3435 or 2677 genotypes, or the 2677- 3435 haplotype.

A correlation has been reported between the expression of *ABCB1*, P-glycoprotein and plasma drug concentrations (Fellay et al. 2002). There was an immunological benefit in the presence of the 3435TT genotype and lower levels of P-glycoprotein (measured by flow cytometry), presumably due to the enhanced presence of antiretroviral drugs in cells infected by HIV-1 (Fellay et al. 2002). This evidence shows a potential molecular basis for interindividual differences in the rate of $CD4^+$ T cell recovery with antiretroviral therapy, but there is still controversy surrounding the association between *ABCB1* variation, P-glycoprotein expression and function and plasma drug concentrations (Fellay et al. 2002; Winzer et al. 2005).

## 1.7 Population Differences

Studies on the variation of the *ABCB1* gene and the effects of P-glycoprotein have been done for Europeans, Asians, Americans, African Americans and West Africans, but not for South Africans, where HIV is highly prevalent. The frequency of common mutations and the possibility of population-specific SNPs both play a role in the drug response and success of drug therapy in infected individuals. The SNPs in exons 12, 21 and 26 have been found in all populations, but the frequencies at which they occur are different (Hoffmeyer et al. 2000; Kim et al. 2001; Schaeffeler et al. 2001; Fellay et al. 2002; Tang et al. 2002; Telenti et al. 2002; Gwee et al. 2003; Tang et al. 2004) as shown by the frequencies of the 3435T allele of approximately 0.5 in Caucasians, 0.4 to 0.6 in Asians, and approximately 0.8 in Africans. When considering the SNPs in exon 12, 21 and 26, different haplotypes are observed in different populations (Schaeffeler et al. 2001; Tang et al. 2002; Tang et al. 2004; Zhu et al. 2004).

In the West African population the frequency of the 3435C allele was found to be much higher frequency than in any other population, possibly due to natural selection against inflammatory bowel diseases, which is decreased in the presence of the 3435C allele (Schaeffeler et al. 2001).

## 1.8 African Origin

There are two major hypotheses for the origin of modern humans. The first is the multiregional hypothesis, which suggests that modern humans evolved from more primitive forms over a period of a million years in different areas of the world. The second is the African replacement hypothesis, which suggests that modern humans originated in Africa and migrated throughout the world to replace the primitive human forms completely (Jorde et al. 1998; Reed and Tishkoff 2006). However, the high level of diversity found in Africans gives more support to the second hypothesis that modern humans first arose in Africa and then colonized other parts of the world.

Variation in mitochondrial and nuclear DNA has shown the African populations to be the most variable and genetically diverse of all populations. Mitochondrial DNA and Y-chromosome haplotype have also shown all non-African lineages have a sub-set of the genetic diversity found in Africans, possibly due to genetic bottlenecks that occurred during the migration out of east Africa. Mitochondrial DNA and Y-chromosome haplotype lineages provided a model for the 3 main lineages that migrated within Africa. The first lineage (L1) is thought to be the most ancient and is present in the San population (South Africa) and the Biaka Pygmies (Central African Republic), the two most genetically diverse populations in Africa. The second lineage (L2) is present in the Mbuti Pygmies (Democratic Republic of Congo) and in the West African Bantu-speaking populations. The third lineage (L3) is widely distributed throughout East Africa

and is rare in the sub-Saharan region of Africa (Jorde et al. 1998; Tishkoff and Williams 2002; Yu et al. 2002; Watkins et al. 2003; Reed and Tishkoff 2006; Tishkoff et al. 2009). South Africa's population comprises mostly groups that are off shots of the Bantu population which fall under the bigger Niger-Congo group (Tishkoff et al. 2009).

The Bantu expansion, migration and spread, occurred towards the south and east around 3,000 to 5,000 years ago from West Africa in the region of present day Nigeria and Cameroon into sub-Saharan Africa in two directions (Guthrie 1962; Phillipson 1993). The south western expansion followed a mostly coastal and ravine route, whereas the eastern expansion involved the movement and settlement of farmers in present day Uganda. From there, the Great Lake region, the expansion continued in two waves. The one wave moved south along the coast, while the other moved south through eastern Zimbabwe (Philipson 1993; Newman 1995; Beleza et al. 2005; Reed and Tishkoff 2006).

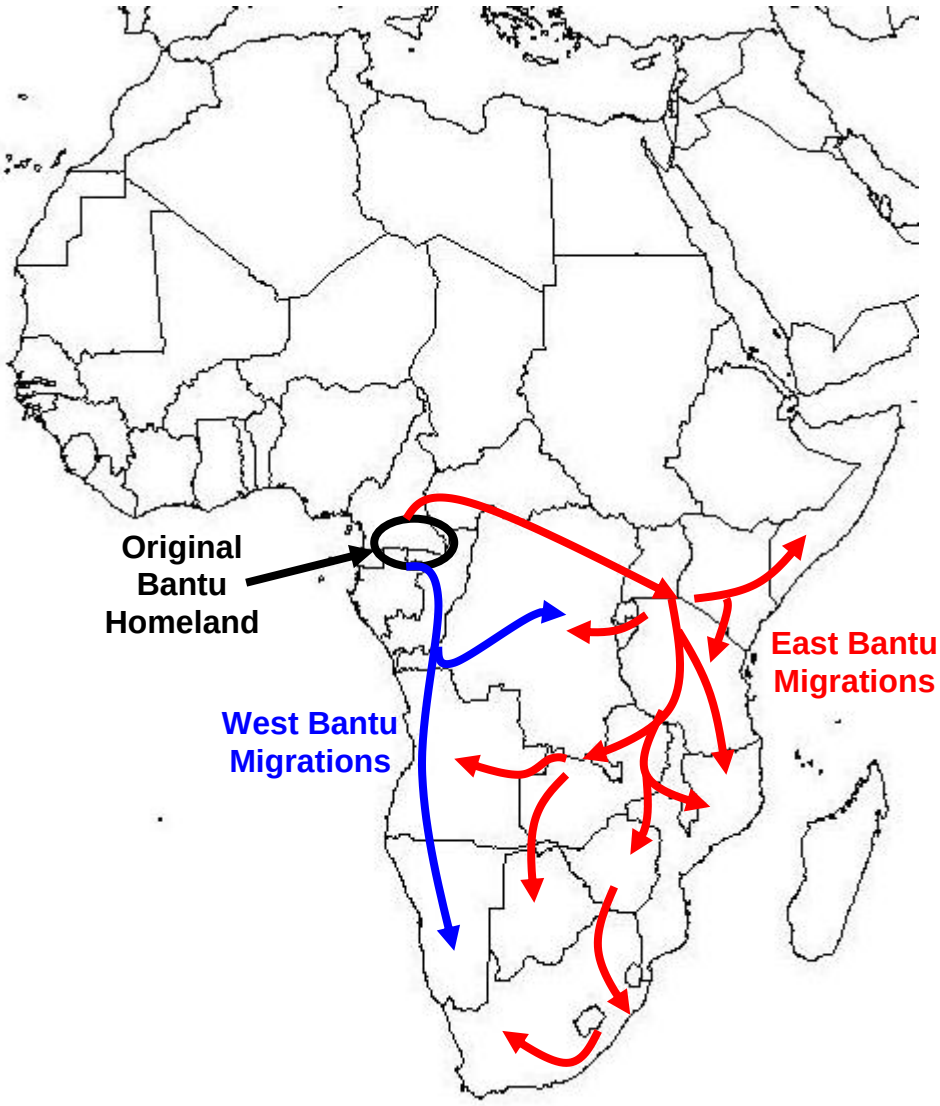


Figure 2: Diagram of the east and west Bantu migrations from West Africa into southern Africa (based on data from Beleza et al. 2005 and Reed and Tishkoff 2006).

Bantu speakers can be broadly classified on linguistic grounds. The Western Bantu languages are older and more diverse, whereas the Eastern Bantu languages are more recently derived (Nurse 1982; Vansina 1984). The East Bantu languages include a Sotho/Tswana language group (Southern Sotho, Northern

Sotho or Pedi, and Tswana), an Nguni language group (Zulu, Xhosa and Tonga or Shangaan) or Venda (Lane et al. 2002). There are about 400 Bantu languages, spoken by people in sub-Saharan Africa (Greenberg 1963; Guthrie 1962) and of the 11 official languages in South Africa, nine are Bantu languages. According to the National Census of 2001 the language distribution across South Africa is as follows: Zulu (23.8 %), Xhosa (17.6 %), Pedi (9.4 %), Tswana (8.2 %), Southern Sotho (7.9 %), Tsonga (4.4 %), Swati (2.7 %), Venda (2.3 %) and Ndebele (1.6%). The remaining 22.1 % of South Africans speak European languages (English and Afrikaans).

South Africa has experienced an increase in the movement of people from rural to urban areas, often creating novel urban communities. This results in the mixing of different ethnic populations and alters of gene frequencies in these populations (Lane et al. 2002). The genetic structure of seven South African Bantu speaking groups showed that the seven groups shared 99% of their genetic variation, based on autosomal and Y-chromosome haplotype markers (Lane et al. 2002). This showed that although there had been a development of different languages, the different groups had all come from a common ancestor and the separation, during which the languages developed, had not been for a very long time.

The demographics of human history are complex, due to populations migrating out of ancestral Africa into new regions, where they experienced isolation, migration, expansion or bottlenecks, all of which influence the genetic cmposition of the population (Ardlie et al. 2002). African populations show

tremendous diversity in the more than 2,000 distinct ethnic groups. The sub-Saharan populations show an especially distinct difference in linkage disequilibrium patterns as opposed to non-African populations. They have more haplotypes and lower levels of linkage disequilibrium between alleles. Their high level of genetic variation and their demographic history make African populations useful for fine-mapping complex diseases (Jorde et al. 1998; Tishkoff et al. 2002; Reed and Tiahkoff 2006; Tishkoff et al. 2009). The multiple types of genetic systems (autosomal, mitochondrial and Y-chromosomal DNA) have shown that there is more genetic diversity in Africans than in Europeans or Asians and that Africans have the largest total number of alleles as well as the largest number of unique alleles (Jorde et al. 1998).

## 1.9 HIV/AIDS in South Africa

At the end of 2009 it was estimated that 33.3 million people were living with HIV globally and that during that year approximately 2.6 million new HIV infections occurred with 1.8 million people dying of AIDS-related illnesses. It was estimated that 34% of people living with HIV were from ten countries in southern Africa, and that these ten countries counted for 31% of new infection and 34% of AIDS-related deaths that year. In sub-Saharan Africa 22.5 million people were estimates to be living with HIV (+/- 5.6 million in South Africa) and 1.8 million new infections (+/- 1.5 million in South Africa) and 1.3 million AIDS-related deaths (+/- 310,000 in South Africa) occurred in 2009 (UNAIDS 2010 Global Report).

Based on the South African National HIV Survey 2008 (Shisana et al. 2009), which was a population-based household survey involving almost 49 million individuals over the age of 2 years, the HIV prevalence in South Africa is estimated to affect 10.9 % of the population (5.2 million people), compared to the 11.4 % in 2002 and 10.8 % in 2005. KwaZulu Natal has the highest provincial estimated HIV prevalence (15.8 %), Gauteng is fifth on the list (10.3 %) and the Western Cape has the lowest estimated HIV prevalence (3.8 %).

The interpretation of the prevalence data has become increasingly complex, as there is an increase in the access to antiretroviral therapy, which increases the

prevalence of HIV, as HIV-related deaths are reduced and more people are aware of their HIV status (Shisana et al. 2009).

South Africa has the world's largest antiretroviral therapy program, but it also has the largest epidemic and access to treatment could be better. At the end of 2009 it was estimated that only 970,000 of HIV-infected individuals (17%) were receiving treatment. This was due to the delay and slow pace of treatment delivery, as well as a doubt in the efficacy of the drugs by many South Africans (WHO 2008; UNAIDS 2010 Global Report). There have been improvements in treatment delivery recently, but the cost remains problematic.

## 1.10 Aim

Previous work involved genotyping the *ABCB1* IVS 25+ G3050T and IVS 25+T5231C SNPs in the southern African Bantu population, revealing different frequencies from values reported for other populations (Taniguchi et al. 2003; Soranzo et al. 2004). A region containing the C3435T and IVS 26+T80C polymorphic sites was sequenced in seventeen samples to determine the variation in this region. The allele frequencies for the C3435T and IVS 26+T80C SNPs were determined and novel variation was found. The frequency of the 3050T, 3435T and 80C alleles were lower than those recorded for a European population. The frequency of the 5231C allele was approximately the same as that in the European population (Taniguchi et al. 2003; Soranzo et al. 2004). The 3435C allele frequency in the South African population sample was compared with frequencies found in other populations. It is clear that the European populations have the lowest 3435C allele frequency and the African populations have the highest 3435C allele frequency.

Nineteen samples from the general population samples were successfully sequenced for the 5′ upstream untranslated region, revealing one a novel polymorphism. The number of samples sequences and genotypes is small and a greater sample number would produce more reliable data.

Variation in the human genome can have an effect on drug interactions and drug response, where the resonse to an ordinary drug can be an advese reaction or a lack of response altogether. The variation in these genes and their effect on antiretroviral drugs needs to be examined in the black South African population, where a high prevalence of HIV and an increasing availability of treatment mean that research in this field is essential.

Pharmacogenetics can tailor the treatment to the genotypes of an individual. This may not be possible in the South African health care system, but there may be specific genotypes that are present in this population, either specific variation or different frequencies, that would alter the drug response to that seen in other populations. There is a need to understand the interactions between genetic factors and drug response, so that recommendations can be made for drug dosage and combination, to ensure the highest efficacy and lowest toxicity of antiretroviral drugs in the South African population.

The aim of the project was to characterize variation in *ABCB1* in the black South African population and to determine if any association could be found between that variation and immune response.

To achieve this aim additional DNA samples were collected and the identified SNPs genotyped, followed by an analysis to estimate linkage disequilibrium and infer haplotypes. RNA was obtained during the last sample collection, in order to

obtain relative mRNA level data. $CD4^+$ T cell count and viral load data was obtained for patients sampled at the Themba Lethu clinic.

Two sets of association studies were performed: The first tested the association between SNPs or sets of SNPs and relative mRNA level data to determine if there was any association between variation in the gene and variation in expression levels. The second tested the association between SNPs or sets of SNPs and the relative change in $CD4^+$ T cell count and viral load between the different time intervals, to determine any association between variation in the gene and immune response.

# CHAPTER 2: MATERIALS & METHODS

## 2.1 Samples

Genomic DNA was obtained from a variety of sources for a broad study to characterize variation in genes involved in different aspects of HIV infection, including HIV drug metabolism and transportation in Bantu-speaking individuals living in the Johannesburg area. The sources of genomic DNA were whole blood collections from individuals from the university community and HIV clinics, and genomic DNA from the HIVNET 028 provided by Dr. Clive Gray.

The initial collection of whole blood for DNA extraction comprised 42 volunteer staff and students recruited from the University of the Witwatersrand, who were chosen as Bantu-speakers and whose HIV status was unknown (labelled General Population).

Numerous whole blood collections at HIV clinics over a number of years resulted in genomic DNA from 105 individuals from the Infectious Disease Clinic at Johannesburg Hospital, 94 individuals from the Themba Lethu Clinic at the Helen Joseph Hospital in 2006, and 104 individuals from the Themba Lethu Clinic at the Helen Joseph Hospital in Johannesburg. Patients were approached to volunteer for the study, based on language and being HIV positive. The Themba Lethu clinic was chosen because of their extensive database of clinical data,

which was accessed at a later stage of the study, and the willingness of patients to participate in studies.

The whole blood from the 2009 collection at the Themba Lethu Clinic at the Helen Joseph Hospital was also used for the RNA extractions.

Individuals were chosen for blood collection based on language, as the study was to focus on Bantu-speakers. All samples were obtained under informed consent (Ethics consent forms, permission letter from the Helen Joseph Hospital, patient information, questionnaires and consent forms for the 2006 and 2009 collections are given in Appendices IV, V and VI). From the Patient Questionnaire we were able to determine the place of birth and home language of the subject, parents and grandparents. The blood was drawn by a phlebotomist into EDTA-coated tubs for DNA and RNA extractions.

The Patient Questionnaires revealed that all nine South African languages were represented in the individuals sampled (Table 1). In the group, 67 % of the subjects reporting a single language in 3 generations, 30 % reported more than one language (most of whom were born in Gauteng), and 3 % did not know what languages were spoken by their relatives. The concentrations of the previously collected DNA samples were determined using the NanoDrop® ND-1000 version 3.3 spectrophotometer.

Table 1: Frequency of the nine Bantu languages found in the samples collected from the Infectious Disease Clinic at Johannesburg Hospital, Themba Lethu Clinic at the Helen Joseph Hospital and the University of the Witwatersrand community.

| **Bantu Language** | **Frequency** |
|---|---|
| Zulu | 0.41 |
| Tswana | 0.13 |
| Xhosa | 0.12 |
| Sotho | 0.11 |
| Pedi | 0.08 |
| Tsonga | 0.06 |
| Ndebele | 0.05 |
| Swati | 0.02 |
| Venda | 0.02 |

Genomic DNA was also provided by Dr. Clive Gray from 64 recently HIV-1-infected individuals within approximately 2 years of seroconversion as part of the HIVNET 028 study from five clinical sites in southern Africa: Johannesburg (n=9) and Durban (n=21), South Africa; Lusaka, Zambia (n=18); Harare, Zimbabwe (n=6) and Blantyre, Malawi (n=10). The concentration of the HIVNET samples is known to be 20 ng/µl (Morris et al. 2002).

## 2.2 DNA and RNA Extractions from Whole Blood

The whole blood from each collection was used for DNA extraction and the whole blood from the last collection was also used for RNA extraction.

The QIAGEN QIAamp® (Germantown, USA) DNA Blood Mini Kit was used to extract genomic DNA from the leukocyte-rich buffy coat of whole blood obtained from the patients according to the manufaturer's instructions (Appendix I). The DNA was electrophoresed on a 0.8 % agarose gel in the presence of ethidium bromide to check for integrity and stored at -20$^{O}$C.

TRIzol LS was used to extract whole RNA from the blood obtained from the patients (Appendix I). The whole procedure was performed in a fume hood cabinet using RNase-free plastic- and glassware, and RNaseAway® to avoid RNA degradation. The RNA solution was aliquoted into amounts of 20 µl and stored at -70$^{O}$C. After the extraction, 10 µl of RNA was loaded on to a 1 % agarose gel to determine the presence and quality of RNA.

## 2.3 DNA Sequencing

Sequencing was performed to characterize the variation in the upstream untranslated region, using randomly selected DNA samples. The 5′ upstream untranslated region from exon 1 was chosen, as this region contains regulatory elements and may provide insight into regulation in gene transcription. The 5′ upstream untranslated region from exon 3 was also chosen for sequencing, as transcription starts at exon 3 and the region upstream from that may also contain regulatory elements, in addition to a well characterized polymorphism (T-129C).

Polymerase chain reaction (PCR) to amplify the region involved an initial denaturation of two minutes at $94^{O}$C, followed by 25 to 30 cycles of three stages: a 15 second denaturation at $94^{O}$C, a 30 second annealing stage at a primer-specific amplification, a 30 second elongation at $72^{O}$C, a final five minute elongation stage at $72^{O}$C and cooling to $4^{O}$C. PCR conditions were optimised using 10 µl reactions, followed by increasing the reaction volume to 50 µl. Each reaction contained 2X PCR Master Mix (Fermentas), 1 µM concentration of each primer, 10 µl of nuclease-free water and 5 µl (approximately 25 ng) of DNA. A control reaction was set up by replacing the DNA with 5 µl of nuclease-free water. Successful amplification was determined by electrophoresing the PCR product on a 2% agarose gel in the presence of ethidium bromide.

A 774 bp fragment of the 5′ upstream untranslated region of exon 1 was amplified using the forward primer MDR15URF and the reverse primer 5UTRR2

and an annealing temperature of 53$^{O}$C (Table 2). A 648 bp region surrounding exon 2, also known as exon 1b and containing the T-129C locus, was amplified using the primers published by Gwee et al. (2003) with forward primer E1Fwd and the reverse primer E1Rev and an annealing temperature of 53$^{O}$C (Table 3).

The samples were then purified and concentrated prior to being sent for sequencing using Sure Clean from Bioline. The sample was then resuspended in distilled water, followed by electrophoresis on a 2% agarose gel to ensure that the PCR fragment had not been lost during purification.

The samples were then sent to Inqaba Biotec for sequencing. Two samples were amplified in duplicate to initially determine which primer produced a better sequence. This primer was then used to sequence the remaining samples. Sequences were returned electronically as chromatographs, which were then analysed and aligned to the reference sequence (obtained form Ensembl – gene number: ENSG00000085563) using the Sequencher® Version 4.5 (Gene Codes Corporation, 2005) program.

A number of bioinformatic tools were used to analyse the sequences of the two promoter regions. The promoter sequences were entered into each program twice, once with the major allele at each SNP and once with the minor allele at each SNP, in order to determine if the variation changed the promoter region.

PROSCAN Version 1.7 (Prestridge 2000) predicts promoter elements by comparing them with eukaryotic polymerase II promoter sequences. The current version of the program recognizes close to 70% of primate promoter sequences and has a false positive prediction rate of one in 14,000 base pairs (bp). FPROM (Solovyev 1997; Solovyev 2001; Solovyev and Shahmuradov 2003) predicts potential transcription start sites using functional motifs and sequence composition. This program recognizes 80% of TATA promoter sequences with a false positive prediction rate of one in 2,000 bp, and recognizes 50% of TATA-less promoter sequences with a false positive prediction rate of one in 650 bp.

NSITE searches DNA sequences for regulatory motifs using two databases. The REGSITE database contains 1,500 functional motifs for genomic or promoter sequences. Ghosh is the Animal Transcription Factor Database, containing 8,458 functional motifs for genomic or promoter sequences. TFSEARCH (Akiyama) searches DNA sequences for transcription factor binding sites using the TFMATRIX (Heinemeyer 1998) transcription factor binding sites profile database.

Table 2: Sequences and annealing temperatures for primers used in sequencing and genotyping PCR reactions.

| Name | Sequence (5′-3′) | Size (bp) | $T_a$ ($^{O}$C) |
|---|---|---|---|
| MDR15URF | CTCATTTGAAGGTCTTCCCAGT | 22 | 61 |
| 5UTRR2 | TAGGGAGTTATTTCAAAGTTTTTAT | 25 | 55 |
| E1Fwd | GGTGTTAGGAAGCAGAAAG | 19 | 58 |
| E1Rev | ACTATCCACGCCTCAAGA | 18 | 58 |
| EXON12FC | GTCCTGGTAGATCTTGAAGGGC | 22 | 65 |
| EXON12FT | GTCCTGGTAGATCTTGAAGGGT | 22 | 63 |
| EXON12R | ATTTAGCATAAGGACAAGCTATCTC | 25 | 60 |
| MDR-9 | TGCAGGCTATAGGTTCCAGG | 20 | 62 |
| MDR-10 | GTTTGACTCACCTTCCCAG | 19 | 60 |
| MDR-10a | TTTAGTTTGACTCACCTTCCCG | 22 | 61 |
| ASA3050F | TGGTTCTAAGGTTCGGTGA | 20 | 58 |
| ASA3050RT | CCTTTGTATCTAATTTTGCATTA | 23 | 54 |
| ASA3050RG | CCTTTGTATCTAATTTTGCATTC | 23 | 56 |
| ASA5231FT | TCCAAAGGATGATCTGTTTT | 20 | 54 |
| ASA5231FC | TCCAAAGGATGATCTGTTTC | 20 | 56 |
| ASA5231R | TTCCCTCTCCCACAAGAC | 18 | 60 |
| GT3435F | CTCTTGTTTTCAGCTGCTT | 19 | 56 |
| AE26I26R | TCCCAGAAATGTTCCTCTCT | 20 | 58 |
| AE26I26F | TGACAGTTCCTCAAGGCATA | 20 | 58 |

## 2.4 Genotyping

Seven SNPs in the *ABCB1* gene were chosen to be genotyped in all samples in order to obtain a large dataset for characterizing the variation in the gene.

The UCSC (http://genome.ucsc.edu/cgibin/hgTracks?position=chr7:87132948872295 06&hgsid=154228320&knownGene=pack&hgFind.matches=uc011khc.1), Ensmbl (http://www.ensembl.org/Homo_sapiens/Gene/Summary?g=ENSG00000085563) and NCBI (http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=hum_chr.inf&query=ABCB1) databases were used to determine the sequences surrounding the chosen SNPs and their frequency in other populations. The seven polymorphic sites chosen were T-129C in exon 2, C1236T in exon 12, G2677T/A in exon 21, G3050T and T5231C in intron 25, C3435T in exon 26 and T80C in intron 26, all of which had all been previously associated with variable protein expression, or were in high linkage disequilibrium with the C3435T site. The relative distances between the seven loci were determined (Figure 3).

Two methods of genotyping were used: PCR (as described in the sequencing section) followed by restriction digest or, in the absence of a restriction site, allele-specific amplification. Both methods of amplification involved a 10 μl PCR reaction of 2X PCR Master Mix, 1 μM concentration of each primer, 2 μl of nuclease-free water and 1 μl (approximately 5ng) of DNA. A control reaction was set up using 5 μl of nuclease-free water instead of DNA.

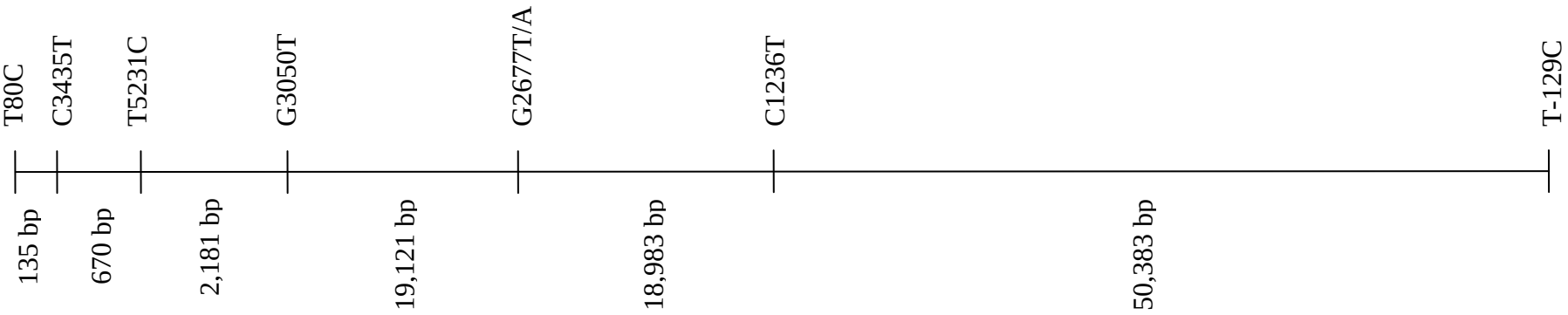


Figure 3: The relative distances between the seven *ABCB1* polymorphisms that were chosen to be genotyped.

For the amplifications followed by restriction digest, each sample was amplified and one sample was amplified in duplicate to use as an undigested control. Amplification was confirmed by electrophoresis of 5 µl on a 2% agarose gel in the presence of ethidium bromide.

Restriction digest reactions of 10 µl were set up with 5 U of restriction enzyme, 1X enzyme-specific buffer, 3.8 µl of nuclease-free water and 5 µl of PCR product. Undigested controls were set up replacing the restriction enzyme with nuclease-free water. To confirm the function of the enzyme, most restriction digetions were designed to include at least one restriction control site, where the amplified DNA would be cut, regardless of the genotype. The digestions were electrophoresed on agarose gels in the presence of ethidium bromide to determine the genotypes.

For allele-specific amplification three primers are required, either one forward and two reverse primers, or two forward primers and one reverse primer. For the two forward or reverse primers, each primer is specific for one of the two possible alleles. The amplification reaction is performed in duplicate, each using on of the allele-specific primers. The PCR products are electrophoresed and analysed to determine the genotypes. The successful amplification of one reaction only indicates a homozygote of the successful allele-specific reaction, whereas the successful amplification of both reactions indicates a heterozygote.

For all polymorphic sites genotyped, blind replicates of a random selection of 10% of the samples were performed to confirm the reproducibility of the methods and to blindly test the accuracy of genotyping.

To genotype the T-129C polymorphism the same 648 bp region was amplified as that used to sequence the 5′ upstream region of exon 3 using the primers E1Fwd and E1Rev, an annealing temperature of $53^{O}$C (Table 2) and the protocol designed by Gwee et al. (2003). Amplification was confirmed by electrophoresis on a 2% agarose gel in the presence of ethidium bromide. The fragment was subjected to restriction digestion at $37^{O}$C for 2 hours using the enzyme *Msp*A1I, which recognises the restriction site CMG/CKG, where M is either A or C and where K is either G or T. The restriction site was created in the presence of the -129C allele and three control restriction sites were also included (Figure 4a). The genotypes were determined from the elctrophoresis of the restriction digest on an ethidium bromide containing 4% agarose gel.

The C1236T polymorphism was genotyped using allele-specific amplification of a 766 bp fragment using two forward primers (EXON12FC and EXON12FT), one common reverse primer (EXON12R) and annealing temperatures of $57^{O}$C for the C allele and $56^{O}$C for the T allele (Table 2). Ampification was confirmed and genotypes allocated by electrophoresis on a 2% agarose gel containing ethidium bromide.

The G2677T/A polymorphism was genotyped using the protocol from Sigmund et al. 2002, which separates the genotyping into G2677A and G2677T. The G2677A polymorphism was genotyped by amplification a 220 bp fragment using the forward primer, MDR-9, the reverse primer, MDR-10 and an annealing temperature of $56^{O}$C (Table 2). Amplification was confirmed by electrophoresis on an ethidium bromide containing 2% agarose gel. The fragment was restricted at $65^{O}$C using the enzyme *Bsr*I, which recognises the restriction site ACTGGN/. In the presence of the A allele, the *Bsr*I restriction site was created and the fragment was restricted (Figure 4b). Electrophosesis on a 4% agarose gel with ethidium bromide was performed to confirm and analyse restriction digestion.

The 224 bp fragment for genotypinc the G2677T polymorphism was amplified using the forward primer MDR-9, the reverse primer MDR-10a and an annealing temperature of $57^{O}$C (Table 2), and confirmed by electrophoresis on a 2% agarose gel in the presence of ethidium bromide. The fragment was restricted at $37^{O}$C using the enzyme *Ban*I, which recognises the restriction site G/GYRCC, where Y is either C or T and R is either A or G. This restriction site is not naturally occurring in the *ABCB1* exon 21 sequence, but amplification with the reverse primer MDR-10a resulted in a mutation being incorporated into the amplified fragment to provide a restriction site for *Ban*I in the presence of the G allele (Figure 4c). Restrictions were analysed after electrophoresis on a 4% agarose gel containing ethidium bromide. Once the two polymorphisms were genotyped individually the data was combined to result in the final expected genotypes.

The IVS 25+G3050T SNP was genotyped by allele-specific amplification of a 336 bp fragment using one common forward primer (ASA3050F), two reverse primers (ASA3050RT and ASA3050RG) at annealing temperatures of 51.1$^{O}$C for the T allele and 52$^{O}$C for the G allele (Table 2). Electrophoresis on an ethidium bromide containing 2% agarose gel was used to confirm amplification and assign genotypes.

Allele-specific amplification of a 293 bp fragment was used to genotypes the IVS 25+T5231C SNP, using two forward primers (ASA5231FT and ASA5231FC), each specific for one allele, one common reverse primer (ASA5231R) at annealing temperatures of 47$^{O}$C for the T allele and 48$^{O}$C for the C allele (Table 2). Samples were allocated genotypes after electrophoresis on a 2% agarose gel containing ethidium bromide.

The 477 bp fragment containing the C3435T SNP was amplified using the forward primer GT3435F, the reverse primer AE26I26R and an annealing temperature of 47$^{O}$C (Table 2). Electrophoreseis on a 2% agarose gel in the presence of ethidium bromide, was used to confirm amplification, for restriction at 37$^{O}$C using the enzyme *Mbo*I, which recognises the restriction site /GATC. In the presence of the C allele, the *Mbo*I restriction site was created, the fragment was restricted, and control restriction site was also included to verify the function of the enzyme (Figure 4d).

The IVS 26+T80C SNP wasgenotyped by amplification of a 651 bp fragment using the forward primer AE26I26F and the reverse primer AE26I26R at an annealing temperature of $53^{O}$C (Table 2). Amplification was confirmed by electrophoresis on an ethidium bromide containing 2% agarose gel. The fragment was restricted at $37^{O}$C using the enzyme *Alw*26I, which recognises the restriction site GTCTC$(N)_1$/. In the presence of the C allele, the *Alw*26I restriction site was created and the fragment was restricted. Two control restriction sites were also included to verify the function of the enzyme (Figure 4e). Genotypes were assigned after electrophoresis on a 4% agarose gel in the presence of ethidium bromide.

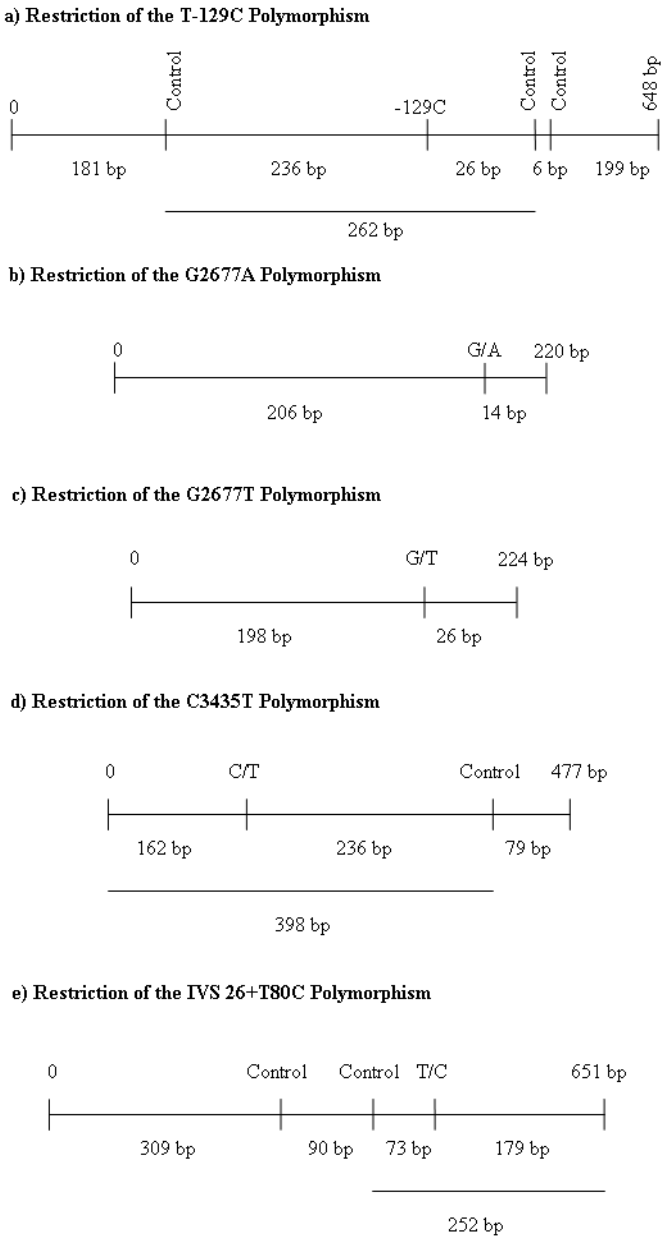


Figure 4: Restriction patterns for the T-129C, G2677A, G2677T, C3435T and IVS 26+T80C polymorphisms based on the size of the amplified fragment, the position of the restriction site created by one allele and removed in the presence of the other allele, and the presence and position of any control restriction sites. The DNA was restricted using the following restriction enzymes: *Msp*A1I, *Bsr*I, *Ban*I, *Mbo*I and *Alw*26I, respectively.

The genotypic data was entered into a spread-sheet and the bioinformatic tool Linkage Disequilibrium Analyser® (LDA) Version 1.0 (Ding et al. 2003) was used to test for Hardy-Weinberg. This program is a java-based Graphic User Interface and provides an estimation of linkage disequilibrium statistics; it performs statistical tests and predicts linkage disequilibrium for autosomal SNPs (Ding et al. 2003). The Hardy-Weinberg equilibrium ($p^2 + 2pq + q^2 = 1$) was used to determine if the allele and genotype frequencies, for autosomal chromosomes, were in equilibrium, meaning that there is no heterozygous excess or selection for a specific allele or genotype. This program performs the Chi squared test to test whether the data fits the Hardy-Weinberg equilibrium.

This test is a non-parametric, statistical test for two sets of data. It was used to test whether the observed and expected genotype frequencies, were similar enough for us to say that the observed data does not occur at random. The Chi squared equation ($\chi^2 = \Sigma\ (o-e)^2/e$) and the degrees of freedom (df = n-1) for a specific threshold or error (5%) are used to determine the p value. If the p value falls above the error threshold ($p \geq 0.05$) then we can accept the hypothesis that the observed data is not significantly different from the expected results. If the p value falls below the threshold ($p < 0.05$) then we have to reject the hypothesis that the observed and expected data are not significantly different from one another (Roshner et al. 1990).

The linkage disequilibrium between a two-locus haplotype can be measured using *D*, expressed as differences between the experimental and expected

frequencies in terms of random segregation. *D* is calculated using the frequency of the haplotype formed by the alleles A and B ($P_{AB}$) and the frequencies of each allele undergoing independent assortment ($P_A$ and $P_B$), such that $D = P_{AB} - P_A \times P_B$. The numerical value of *D* is meaningless for comparisons or determining the strength of the disequilibrium. Therefore the Lewontin's coefficient $|D'|$ ($|D'| = D/D_{max}$) and Pearson correlation $r^2$ ($r^2 = D/P_A \times P_a \times P_B \times P_b$) are used to quantify the degree of linkage disequilibrium.

When $|D'| = 1$ and $r^2 = 1$, then there is no recombination (complete linkage disequilibrium), when $|D'| < 1$, then there is recombination (linkage disequilibrium is destroyed), and when $|D'| > 1$, then there is no clear interpretation. The Lewontin's coefficient ($|D'|$) is strongly dependant on the sample size, often leading to conflicting results, since $|D'|$ increases for small population sizes. The LDA software was also used to calculate the linkage disequilibrium between two alleles. This program was used to calculate the Lewontin's coefficient $|D'|$ and the Pearson correlation $r^2$ from the genotypic data (Lewontin 1964; Hill and Robertson 1968, Shifman et al. 2003; Reich et al. 2001; Ardlie et al. 2002; Nsengimana et al. 2004).

The samples with complete genotypic data were used for the haplotype analysis, performed using the PHASE version 2.1.1 (Stephens et al. 2001; Stephens and Sheet 2005). PHASE was run ten times to determine the reproducibility of the output (Stephens et al. 2001; Stephens and Sheet 2005). The expected and inferred frequencies were compared for the haplotypes that were most frequent in

our data set. The frequencies of the commonly found *ABCB1* haplotypes, involving the C1236T, G2677T/A and C3435T polymorphisms, were compared to haplotype frequencies from other populations.

A reduced median network of common haplotypes was constructed using the Network Version 4.510 (Dec. 2008) program designed by Bandelt et al. (1995). This program is found on the Fluxus Engineering website and is able to construct networks from genotypic or binary data.

## 2.5 cDNA Synthesis using Reverse Transcription of RNA

For all RNA samples the total RNA was reverse transcribed into single-stranded cDNA for relative quantification, using the High-Capacity cDNA Reverse Transcription Kit (Applied Biosystems) according to the manufacturer's instructions (Appendix I). The process requires the binding of reverse transcriptase enzyme to the RNA molecule. This enzyme facilitates the binding of the random primers to the RNA and extension of these primers to form the cDNA molecule using dNTPs.

This reaction occurs under optimum conditions, maintained by a buffer. The random primers ensure that synthesis of the first cDNA strand proceeds at an equal efficiency for all RNA molecules present. An RNase inhibitor is added to ensure that the RNA molecules present are not digested by RNase. The PCR product was not electrophoresed, since the reaction only synthesised the first cDNA strand without amplifying it, and therefore these would only be a few copies of the cDNA, not enough to view on an agarose gel. The newly synthesised cDNA was used for the quantitative PCR reactions.

## 2.6 Calibrator Normalised Relative Quantification of mRNA

The synthesized cDNA was used for the relative quantification of mRNA, which was performed using the TaqMan® Gene Expression Assay (Applied Biosystems) according to manufacturer's instructions, using TaqMan® MGB probes. These probes contain a reported dye (6-FAM), which is linked to the 5′ end of the probe, a minor groove binder, which allows the melting temperature of the probe to be increased for the design of shorter probes, and a nonfluorescent quencher at the 3′ end of the probe. The probes bind at an exon-exon boundary in the *ABCB1* cDNA and the reporter dye is removed during the elongation step in the PCR reaction to fluoresce.

For quantification of the levels of mRNA the Calibrator Normalised Relative Quantification ($\Delta\Delta C_t$) method was chosen. No standard curve was required as the target gene and reference gene (housekeeping gene), were compared to one another and no standand curve was needed. We also chose this method for the determination of relative amounts of mRNA, as these would be compared with the genotypes and haplotypes to determine if these was any association between them. The quantification reaction is carried out for each sample for the target gene (*ABCB1*) and a reference gene (human beta-2-microglobulin (B2M) gene). The crossing points can then be compared to determine the target/reference ratio.

An additional sample (HIV negative in our case) is chosen to be the calibrator. The calibrator is also quantified for both the target and reference genes and is

used to normalise the samples in a particular PCR run, as well as between different runs. The normalised ratio is then calculated as the target/reference ratio for the sample, divided by the target/reference ratio for the calibrator. Each sample was amplified in real time with detection for the target and reference genes five times to ensure reliable results for analysis.

The target reaction used a TaqMan Gene Expression Assay (hs01067802_ml), which consisted of a set of primers that span the exon-exon junctions, preventing genomic DNA from amplifying, and specifically amplifying an 84 bp region of the *ABCB1* cDNA. Of the many available assays for the *ABCB1* gene, this one spans a region that includes exon 26. The primer sequences, and therefore the specific region, were proprietary and could not be obtained.

The endogenous control assay amplified the human beta-2-microglobulin (B2M) gene as a reference gene. This gene was chosen as it is expressed uniformly in all cells. The popular GAPDH gene was not chosen as a reference gene as this gene is not as uniformly expressed as was previously thought (Zhong and Simons 1999).

The 7500 Software Version 2.0.3 (Applied Biosystems) was used to analyse the relative quantification data. The software calculated the $C_t$ values for the target and reference genes and calculates the means for the repeats. The $C_t$ value is the threshold cycle, which is the cycle number at which the generated fluorescence

crosses the threshold line. The average $C_t$ values were determined from the five repeats and the standard deviation calculated for each sample average.

The $\Delta C_t$ value is then calculated by subtracting the $C_t$ (reference) from the $C_t$ (target) for the test samples and the calibrator. The $\Delta\Delta C_t$ is then calculated by subtracting the $\Delta C_t$ (calibrator sample) from the $\Delta C_t$ (test sample). The fold difference value between the sample and calibrator, or relative quantity (RQ) is given by the $2^{-\Delta\Delta Ct}$ value. The $\Delta\Delta C_t$ for the calibrator is set to 1, so that all the sample fold differences are represented as a fold difference of the calibrator. These values were then depicted as a bar chart with standard deviations.

## 2.7 Association Studies

The genotypic and haplotypic data was used for association studies with the clinical data was obtained from the Themba Lethu Clinic at Helen Josep Hospital and the relative quantificatiug data, to determine if any association could be found between variation in the gene and the level of transcription and/ or drug response.

The data obtained from the Thamba Lethu Clinic at Helen Josep Hospital included the baseline $CD4^+$ and viral load counts with dates, the $CD4^+$ and viral load counts at time point one (as close to six months as possible) with dates, and the $CD4^+$ and viral load counts at time point two (as close to twelve months as possible) with dates. The viral load data was not as complete as the $CD4^+$ data, as viral load is not always measured at clinical visits. The number of days between the baseline measurement and the two time points were established, as visits to the clinic were not always strictly at six and twelve months. It was suspected that the data would not show distinct time points, but a merger between time points one and two. The $CD4^+$ cell counts or viral load counts were plotted against the days since baseline.

The clinical data and genotypic data files were merged and the data was sorted for each SNP by genotype. Line graphs were then constructed to show the increase in $CD4^+$ or decrease in viral load counts for the genotypes of each SNP at the three time points, baseline, time point one and time point two. The SAS

9.1.3 software (SAS Institute Inc. Cary NC, USA) was used to perform the statistical analyses for association. The analysis of variance (ANOVA) and non-parametric Kruskal-Wallis tests were chosen based on the papers by Fellay et al. (2002) and Nasi et al. (2003), who used these tests for their association studies. Fellay et al. (2002) found an association between the 3435TT genotype and low P-gp expression and low plasma drug levels.

The mean $CD4^+$ counts and viral load counts were claculated for each genotype at each polymorphism. The mean $CD4^+$ counts and viral load counts were then claculated for combinations of SNPs. These were chosen to analyse each of the SNPs with the C3435T SNP, and based on linkage disequilibrium data. The C1235T-G2677T/A-C3435T haplotype was also chosen as a combination. These were used to compare the increse in $CD4^+$ count or decrease in viral load for the genotypes of the combined SNPs and to compare the combined SNPs with the individual ones.

The GLM proceedure was then used to perform an analysis of variance (ANOVA) test for association between the individual genotypes and the difference in $CD4^+$ counts at baseline and time 1, between the individual genotypes and the difference in $CD4^+$ counts at baseline and time 2, between the individual genotypes and the difference in viral load counts at baseline and time 1, and between the individual genotypes and the difference in viral load counts at baseline and time 2. This is a one-way test for determining the difference in

means of the dependent variables, by sorting them by levels of the independent variable. The proceedure was then repeated using the SNP combinations.

The proportional increase in $CD4^+$ count and proportional decrease in viral load count was calculated for both time points. The proportional increase or decrease was calculated by subtracting the baseline $CD4^+$ count or viral load count from the $CD4^+$ count or viral load count at time point one or two, and dividing this by the baseline $CD4^+$ count or viral load count. This was done for a better comparison the proportional increase or decrease as opposed to the actual $CD4^+$ or viral load counts. The mean proportional increase in $CD4^+$ count or decrease in viral load count was determined for each genotype of each SNP and for the SNP combinations. This data was presented graphically and used to repeat the GLM procedure in SAS, to determine any association between the proportional increase or decrease and the different genotypes.

The mean relative quantification (RQ) values corresponding to each genotypes of each SNP were calucluated and presented graphically for comparison. The mean RQ values for the SNP combinations were also calculated and presented graphically for comparison betwwen genotypes and between the single and combined SNPs. A t-test was performed to determine any association between the RQ values and genotypes. This allows the testing of significant difference between the sample means and the hypothesised value. In our analysis the hypothesised value was one, as the RQ values were being compared to the calibrator value.

A non-parametric version of the ANOVA test, the Kruskal-Wallis test, was performed in SAS to test for association between the individual genotypes and RQ values. This test only allowed for two variables to be compared and therefore the GLM ANOVA test was performed in SAS to determine and association between the different genotypes of the SNPs or SNP combinations, the RQ values and the mean proportional difference in $CD4^+$ counts at time point 1 and time point 2, and between the different genotypes of the SNPs or SNP combinations, the RQ values and the mean proportional difference in viral load counts at time point 1 and time point 2.

# CHAPTER 3: RESULTS

## 3.1 Samples

For the 42 general population samples, the average DNA concentration was 51.8 ng/µl, for the 105 HIV positive samples from the Infectious Disease Clinic at Johannesburg Hospital, the average concentration was 46.5 ng/µl, the 64 HIV Net samples were at a concentration of 20 ng/µl and the 207 blood samples that were collected from the Themba Lethu HIV Clinic at the Helen Joseph Clinic in 2006 and 2009 had an average concentration of 9.4 ng/µl. The genomic DNA was diluted to working solutions with concentrations of approximately 5 ng/µl. From the samples collected in 2009, all individuals whose DNA and RNA were successfully extracted were on the regiment IA antiretrovirals, which all have efaviranze as one of the combination drugs.

## 3.2 DNA and RNA Extraction from Whole Blood

DNA was successfully extracted from all samples and run on 0.8% agarose gels to verify the presence and purity of DNA (Figure 5). Of the 104 samples for which the RNA extraction was performed, only 28 samples were successful (Figure 6).

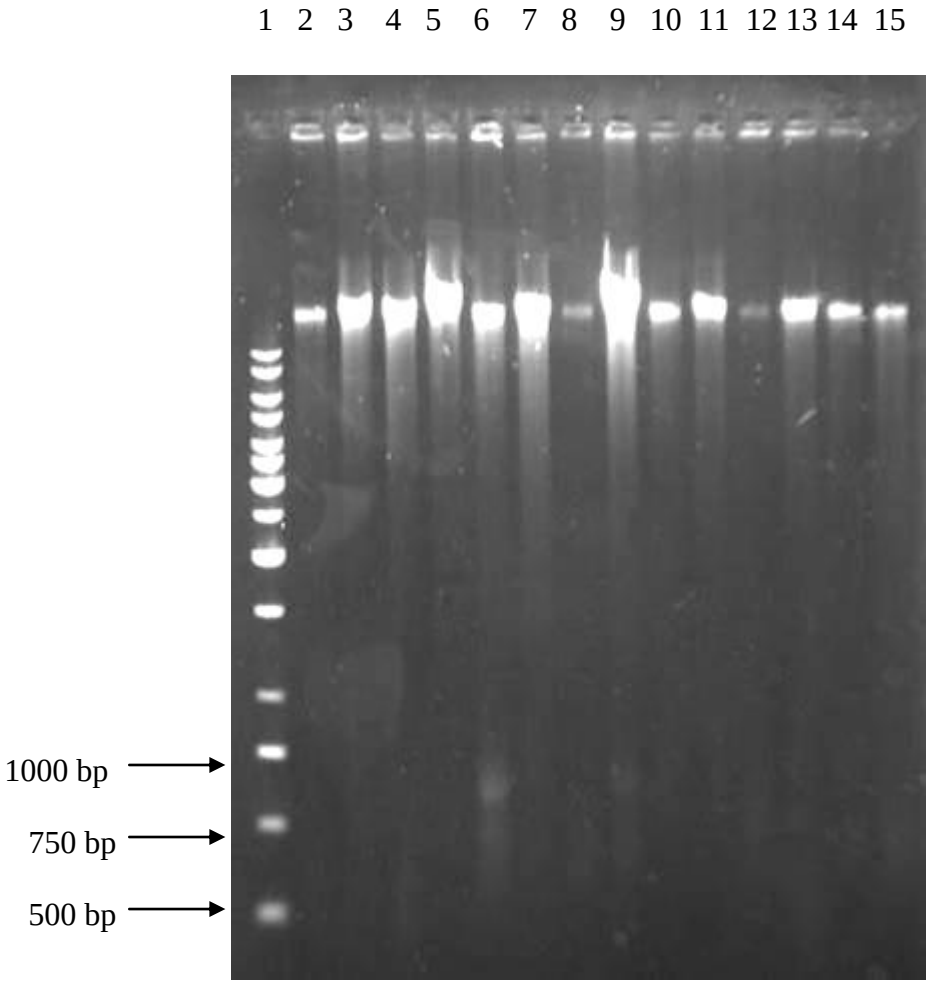


Figure 5: Agarose gel (0.8%) showing integrity of DNA extraction from whole blood of HIV positive patients. The agarose gel was also used to determine if there was RNA contamination, which was not present in any of the above samples. Lane 1 containes a 1 kb molecular weight marker as a reference, and lanes 2 to 15 contain DNA samples from different patients. The concentrations are different due to the different amounts of white blood cells obtained from the patients.

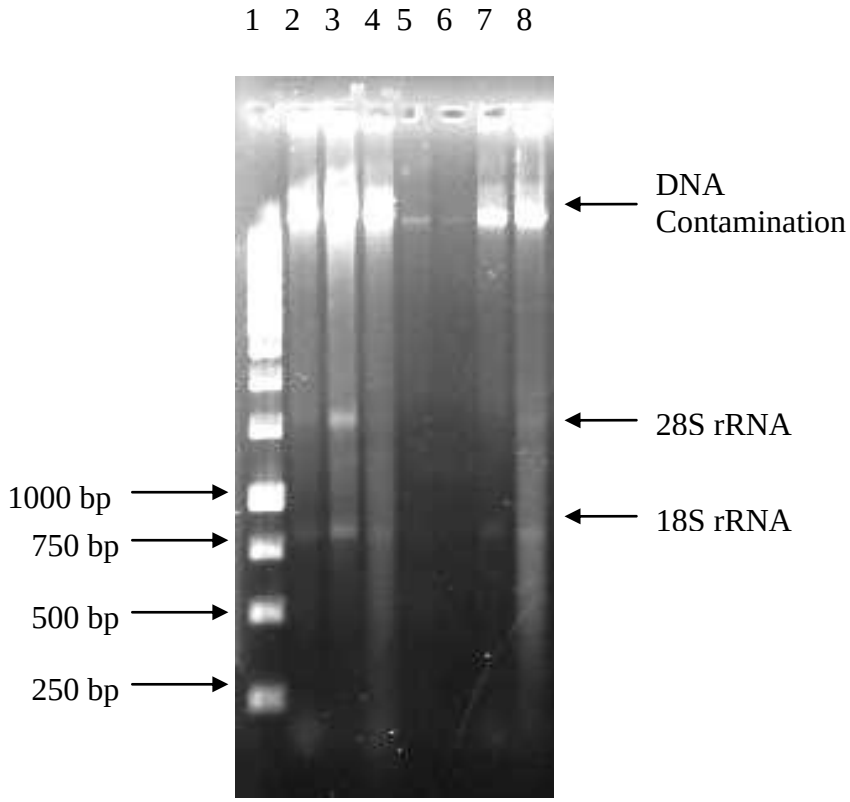


Figure 6: Integrity of RNA extraction from whole blood HIV positive samples visualized on a 2% agarose gel. Lane 1 contains a 1 kb molecular weight marker and lanes 2 to 8 contain successful RNA extractions from different patients. All the extracted RNA samples had genomic DNA contamination, but for the samples that extracted successfully bands for both the 18S and 28S rRNA were visible. An attempt was made to remove the contaminated DNA, but this resulted in the loss of all nucleic material in the sample. The removal of DNA contamination was not necessary though, due to the nature of the primers used for quantitative PCR.

### 3.3 Sequencing

The PCR product of the amplification of the region upstream of exon 1 was electrophoresed to confirm sucessful amplification of the 774 bp fragment (Figure 7).

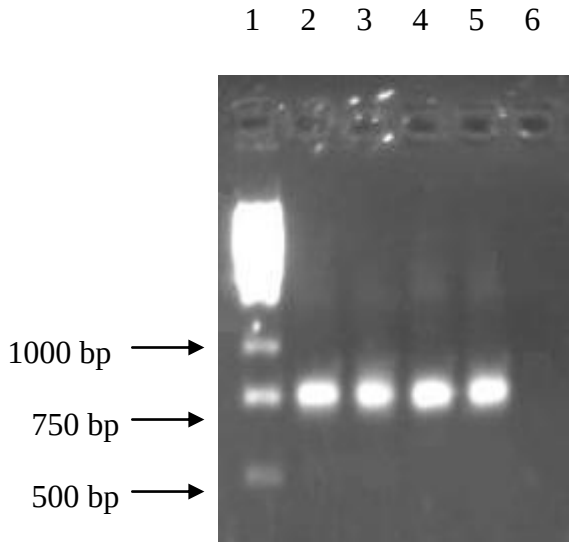


Figure 7: Confirmation of the region upstream of exon 1 by electrophoresis on a 2% agarose gel in the presence of ethidium bromide by a single band. The size of the 774 bp fragment was confirmed by a standand graph ($R^2$ value = 0.9995).

The PCR products for the amplification of the exon 2 region were electrophoresed to determine if amplification was successful. Successful amplification was seen as a single band 648 bp in size (Figure 8).

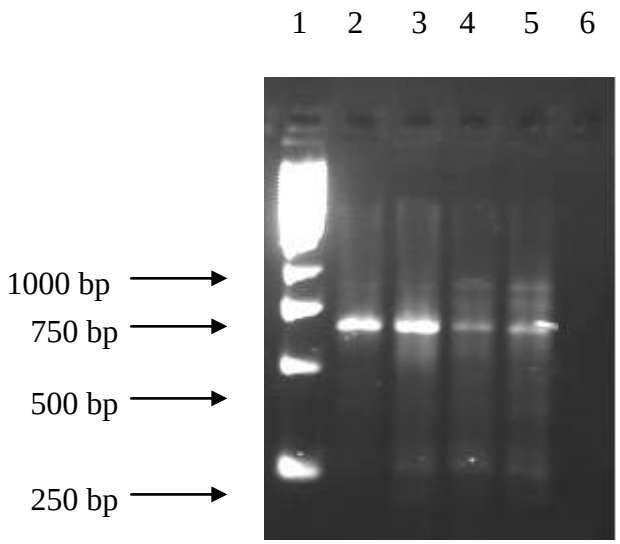


Figure 8: Electrophosesis on an ethidium bromide containing 2% agarose gel to confirm amplification of the fragment containing exon 2 (1b). The 648 bp fragment was confirmed by a standard curve ($R^2 = 0.9992$).

The 5′ untranslated region upstream from exon 1 was sequenced in 32 samples, revealing 4 polymorphisms: at positions A-137C, T-196C, G-223A and G-298A, numbered from exon 1 (Table 3). For the T-196C SNP both homozygotes and the heterozygote were found, for the A-137C, G-223A and G-298A polymorphisms only one homozygote and the heterozygote were detected.

The upstream region from exon 2 (also known as 1b) was sequenced, as transcription only begins with exon 3 and one polymorphic site in particular has been analysed in this region (T-129C). Twenty samples were sequenced and 5 polymorphic sites were observed: A-60T and A-41G in intron 1, T108G, G153A and T-129C in exon 2 (Table 3). All the samples showed homozygozity for only

one allele. For the T-129C site the minor allele frequency is the highest, whereas in the other polymorphic sites only one out of the 20 samples had the variation, all of which were in the form of heterozygotes.

Table 3: Summary of the analysis from the sequence data from the two upstream regions of the *ABCB1* (87133175 – 87342611)

| **Upstream Region** | **No.** | **SNP** | **Status** | **rs Number** | **Minor Allele Frequency** |
|---|---|---|---|---|---|
| Exon 1 (87343026 – 87342553) | 32 | A-137C | Novel | - | 0.01 (C) |
| | | T-196C | Known | rs4148727 | 0.13 (C) |
| | | G-233A | Novel | - | 0.04 (A) |
| | | G-298A | Novel | - | 0.06 (A) |
| Exon 2 (87230607 – 87229959) | 20 | A-60T | Known | rs113117 | 0.05 (T) |
| | | A-41G | Known | rs11126 | 0.05 (T) |
| | | T108G | Novel | - | 0.05 (G) |
| | | G153A | Novel | - | 0.05 (A) |
| | | T202C | Known | rs3213619 | 0.2 (C) |

All novel SNPs were submitted to GenBank and have the following accession numbers: A-137C (JF775576), G-223A (JF775577), G-298A (JF775578), T108G (JF775579) and G153A (JF775580).

The sequence data was used for bioinformatic analysis of both upstream regions. The sequence with the major alleles for all SNPs found and the sequence with the minor alleles for all SNPs found were used for the analysis and compared (Tables 4 & 5).

Table 4: Comparison of outputs from various bioinformatic promoter analysis programs for the sequence upstream of exon 1 with major or minor alleles

| **Program** | **Sequence with major alleles** | **Sequence with minor alleles** |
|---|---|---|
| PROSCAN | No promoter region predicted | No promoter region predicted |
| FPROM | No promoter/ enhancers predicted | No promoter/ enhancers predicted |
| NSITE | 6 human motifs found (REGSITE)<br>11 human motifs found (Ghosh) | 6 human motifs found (REGSITE)<br>12 human motifs found (Ghosh) |
| TFSEARCH | 87 vertebrate TF binding sites found | 89 vertebrate TF binding sites found |

The PROSCAN, FPROM and PROMOTER 2.0 programs did not predict any promoter elements in the region upstream of exon 1. The additional motif found by NSITE using the Ghosh database in the sequence containing the minor alleles was a Lymphokine CS motif (GGGGTTTCAC) in the presence of the -137G allele.

Two additional transcription factor binding sites were found by TFSEARCH in the sequence containing the minor alleles. The -223T allele creates a CdxA binding site (TATTTTT) which is absent in the presence of the -223C allele. The -137G allele creates an AML-1a binding site (TGGGGT) which is absent in the presence of the -137T allele.

Table 5: Comparison of outputs from various bioinformatic promoter analysis programs for the sequence upstream of exon 2 with major or minor alleles

| **Program** | **Sequence with major alleles** | **Sequence with minor alleles** |
|---|---|---|
| PROSCAN | No promoter region predicted | No promoter region predicted |
| FPROM | No promoter/ enhancers predicted | No promoter/ enhancers predicted |
| NSITE | 8 human motifs found (REGSITE)<br>14 human motifs found (Ghosh) | 8 human motifs found (REGSITE)<br>14 human motifs found (Ghosh) |
| TFSEARCH | 48 vertebrate TF binding sites found | 48 vertebrate TF binding sites found |

The PROSCAN and FPROM programs did not predict any promoter elements in the region upstream of exon 3. The NSITE program, using the Ghosh database, found 12 motifs identical in both sequences. The sequence with the major alleles also has the alpha-1-proteinase inhibitor (GGGGCGTGGGCTGAG) and *ABCB1* (GGGCTGAGCA) motifs which are both due to the presence of the 153G allele

and are not found in the sequence with the minor allele as the 153G allele is replaced with the 153A allele.

The sequence with the minor alleles also has a P-LAP (CTCATTCGAGCAGCG) motif due to the presence of the minor 202C (-129C) allele and, which is absent in the sequence with the major 202T (-129T) allele. The minor 108G allele creates the TCR-Vbeta-8.1 (TCTGTGGGGA) motif, which is not present in the sequence with the major 108T allele.

TFSEARCH found the same number of transcription factor binding sites in both sequences, but two polymorphisms altered the sequence from one transcription factor binding site to another. The major -60A creates a binding site for CdxA (CATAA), but the minor -60T allele changes this to a binding site for Nkx-2 (TACTTAA). The major 180T allele is situated in the binding site for AML-1a (TGTGGT), but the minor 180G allele changes this to a binding site for MZF 1 (TGTGGGGA).

### 3.4 Genotyping

The products of the PCR and restriction digestion reactions were eletrophoresed on agarose gels containing ethidium bromide to confirm amplification and restriction digestion, and to ensure correct fragment sizes after amplification and digetion (Figures 9 to 16). The sizes of the fragments were estimated by comparing the bands with the bands of known size of the molecular weight marker and confirmed using standard curves. The reliability of the standard curve is determined by the $R^2$ value, with a value close to 1 being the most reliable. The agarose gels following allele-specific amplification or restriction digestion were used to assign genotypes to the samples.

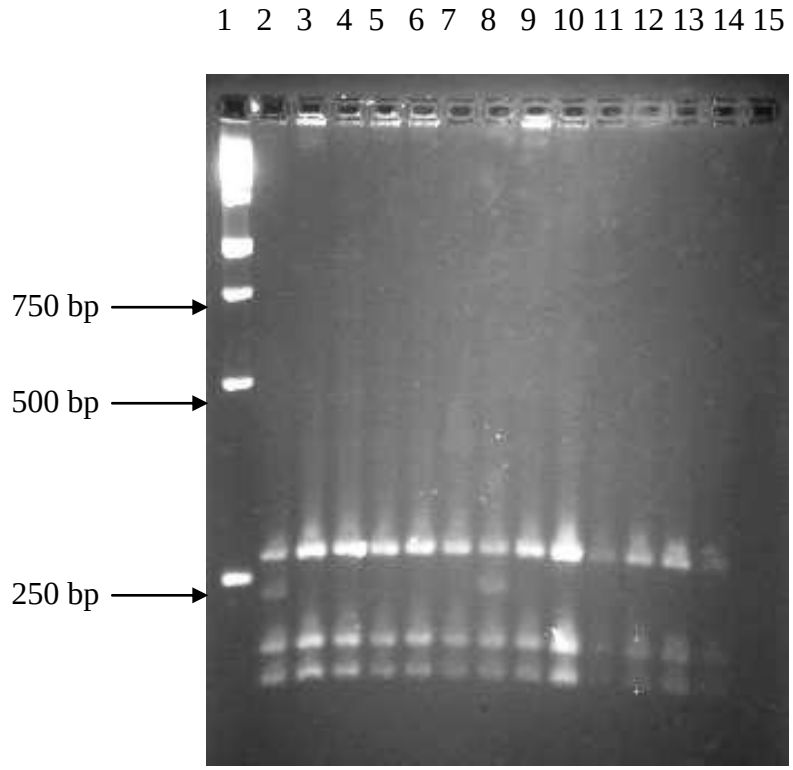


Figure 9: Restriction digestion of the 648 bp PCR fragments amplified for the exon 2 T-129C region, by *Msp*A1I confirmed by the presence of bands after electrophoresis on a 4% agarose gel. Lane 1 contains the 1 kb molecular weight marker, lanes 2 to14 contains restricted DNA for different samples and lane 15 contains the undigested control sample. The PCR amplified fragment sizes and restricted fragment sizes were confirmed using standard curves ($R^2$ = 0.9995 and $R^2$ = 0.988 respectively).

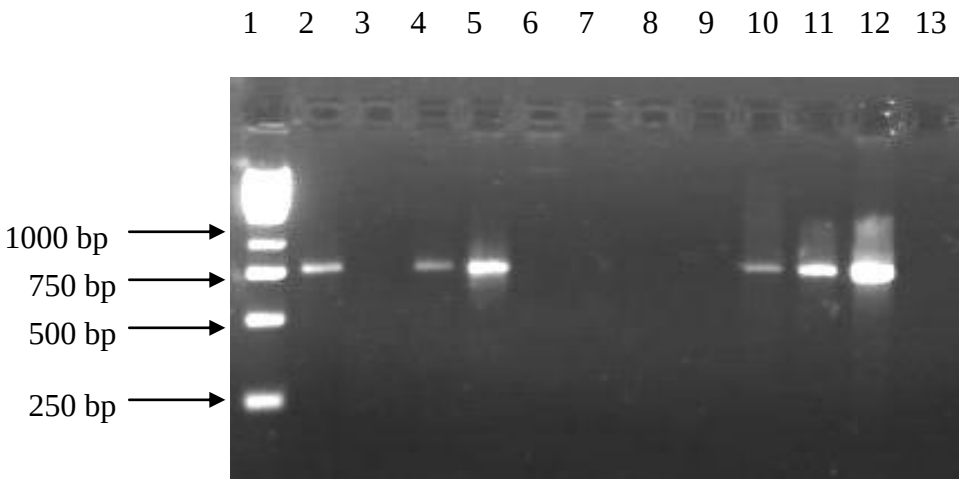


Figure 10: Allele-specific amplification of the exon 12 C1236T fragment confirmed by the presence of bands after electrophoresis on a 2% agarose gel in the presence of ethidium bromide. Lane 1 contains a 1 kb molecular weight marker, lanes 2 to 6 contain samples amplied for the T allele, lane 7 contains the control sample for T allele amplification, lanes 8 to 12 contain samples amplified for the C allele and lane 13 contains the control sample for C allele amplification. The sample sizes were confirmed by a standatd curve ($R^2 = 0.9898$).

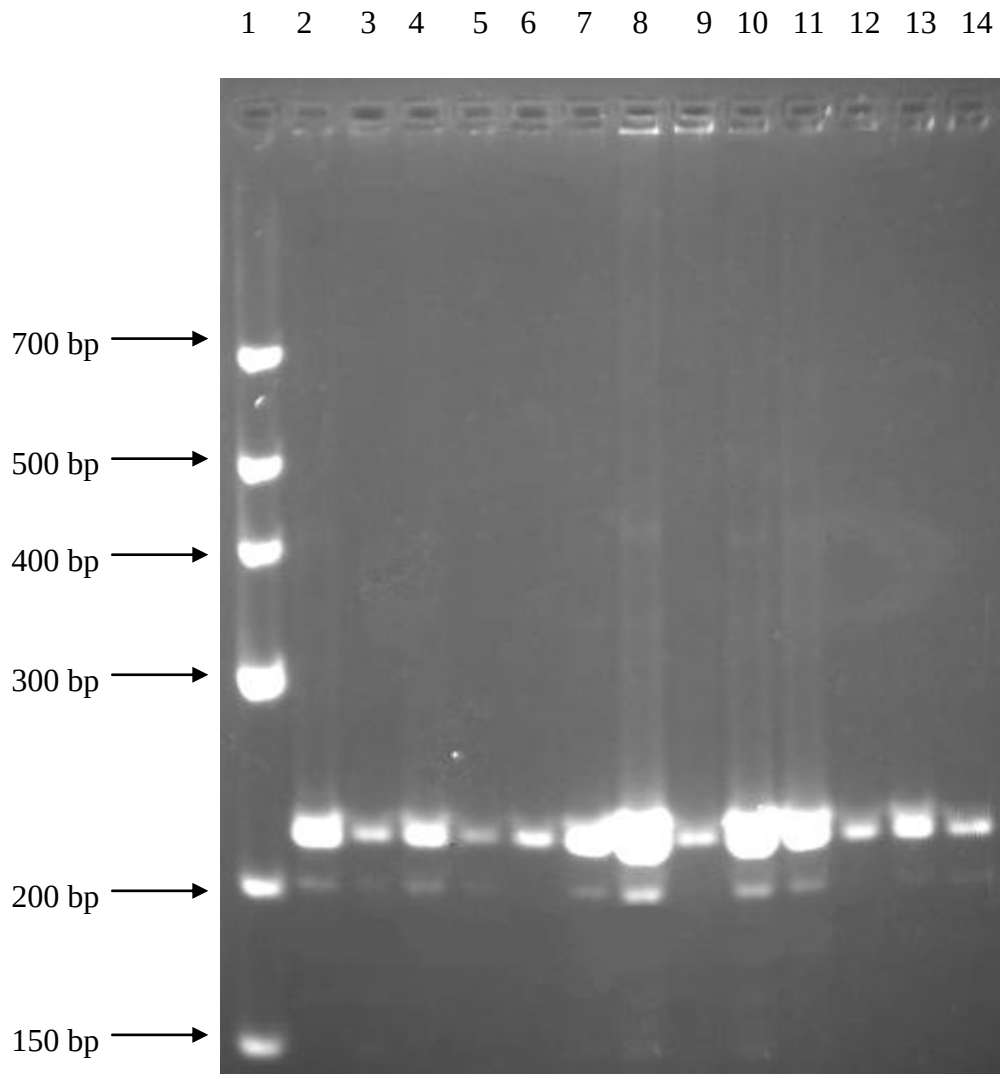


Figure 11: Restriction digest of samples amplified for the 220 bp exon 21 G2677A fragment by *Bsr*I, confirmed by the presence of bands after electrophoresis on a 4% agarose gel in the presence of ethidium bromide. Lane 1 contains a low range molecular weight marker, lanes 2 to 12 contain restriction digestions for different samples and lane 13 contains the undigested control sample. Amplified and digested fragment sizes were confirmed by standard curves ($R^2 = 0.9811$ and $R^2 = 0.9996$, respectively).

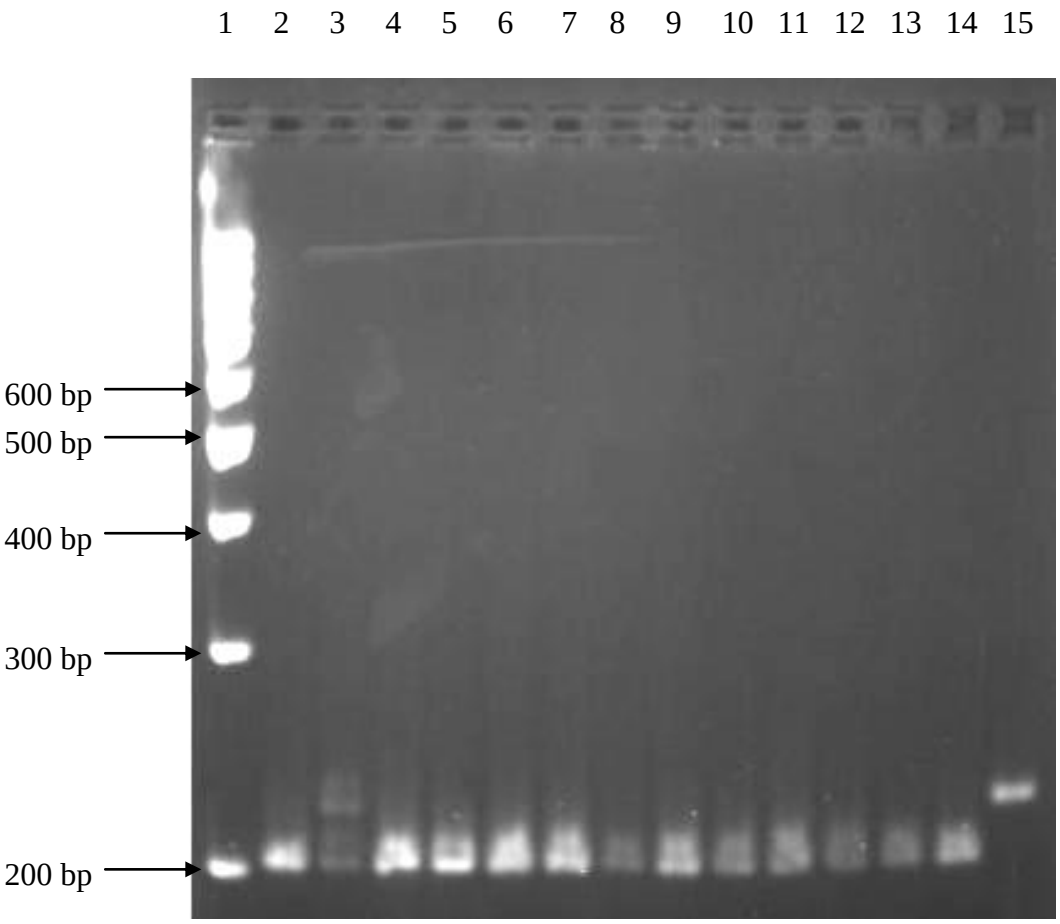


Figure 12: Electrophoresis on an ethidium bromide containing 4% agarose gel to confirm restriction digestion by *Ban*I of samples amplified for the 224 bp exon 21 G2677T fragment. Lane 1 contains a 100 bp molecular weight marker, lanes 2 to 14 contain restriction digestions of different samples, and lane 15 contains an undigested control sample. Amplification and restriction fragment sizes were confirmed using standard curves ($R^2$ = 0.9996 and $R^2$ = 0.9904, respectively).

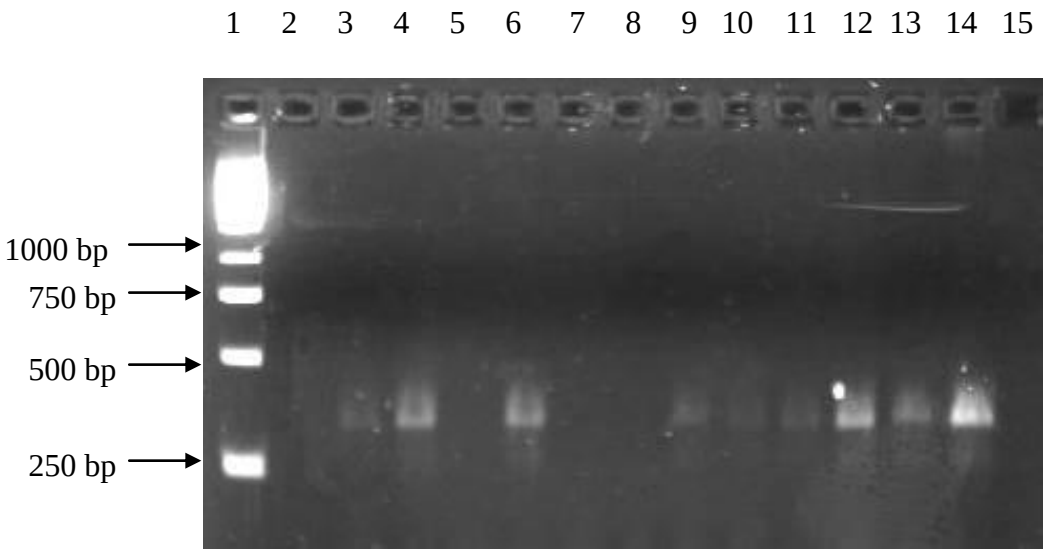


Figure 13: Allele-specific amplification of the 336 bp IVS 25+G3050T fragment, confirmed by the presence of a single band after electrophoresis on a 2% agarose gel in the presence o ethidium bromide. Lane 1 contains a 1 kb molecular weight marker, lanes 2 to 7 contain PCR products of different samples amplified for the T allele, lane 8 contains the control for the T allele reaction, lanes 9 to 14 contain the PCR products of different samples amplified for the G allele, and lane 15 contains the G allele control. The fragment sizes were confirmed using a standard curve ($R^2 = 0.9996$).

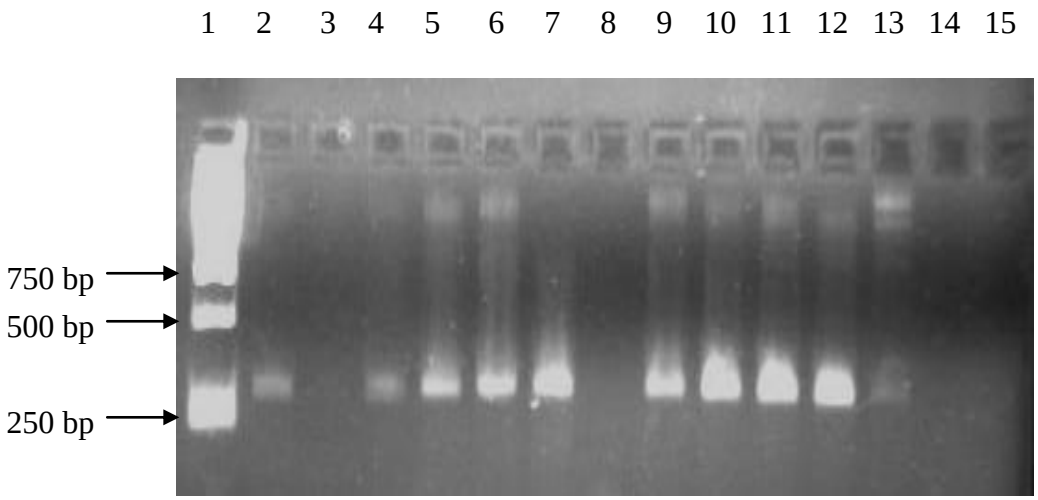


Figure 14: Confirmation of allele-specific amplification of the 293 bp IVS 25+T5231C fragment by electrophoresis on an ethidium bromide containing 2% agarose gel. Lane 1 contains a 1 kb molecular weight marker, lanes 2 to 7 contain PCR products of different samples amplified for the T allele, lane 8 contains the T allele control reaction, lanes 9 to 14 contain the PCR products of different samples amplified for the C allele, and lane 15 contains the C allele control reaction. Fragment sizes were confirmed using a standard curve ($R^2 = 0.9968$).

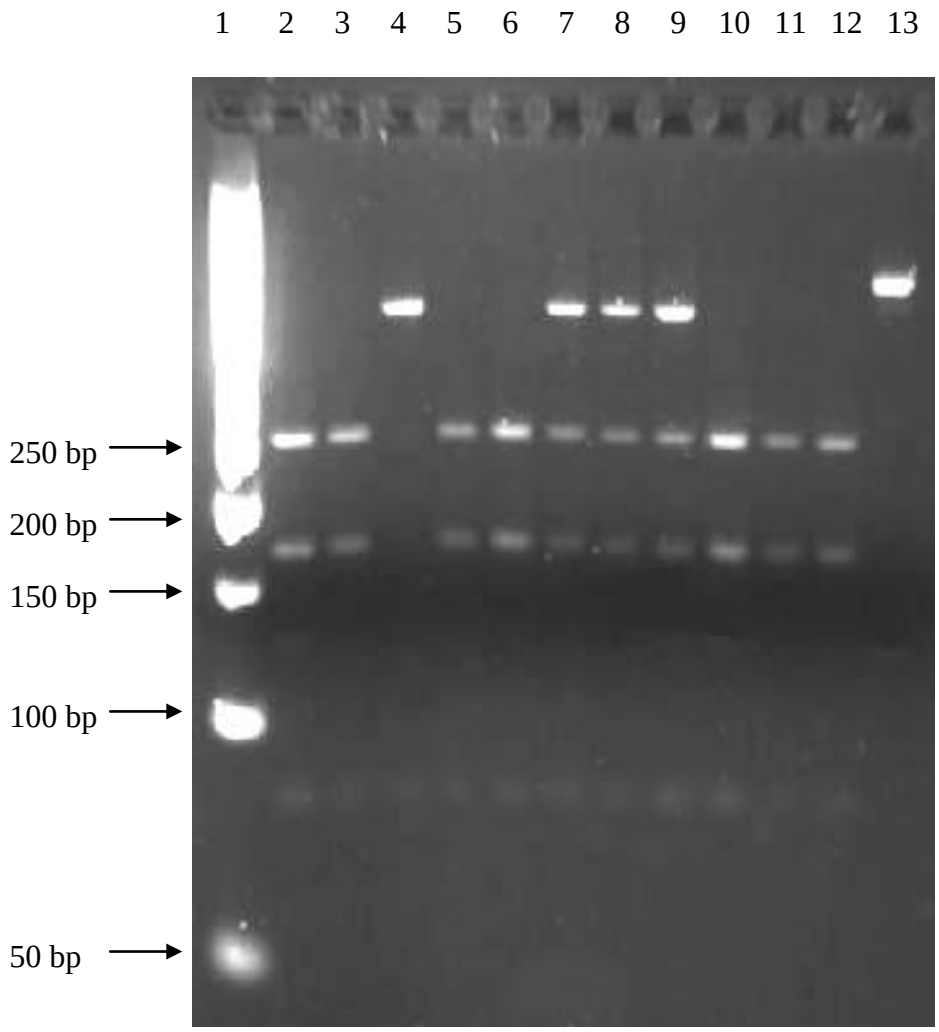


Figure 15: Restriction digest of samples amplified for the 447 bp exon 26 C3435T fragment by *Mbo*I, confirmed by the presence of bands after electrophoresis on a 2% agarose gel containing ethidium bromide. Lane 1 contains a 50 bp molecular weight marker, lanes 2 to 12 contain the restriction digestions of different samples, and lane 13 contains the undigested control sample. The amplification and restriction digestion fragment sizes were confirmed using standard curves ($R^2$ = 0.9998 and $R^2$ = 0.9965, respectively).

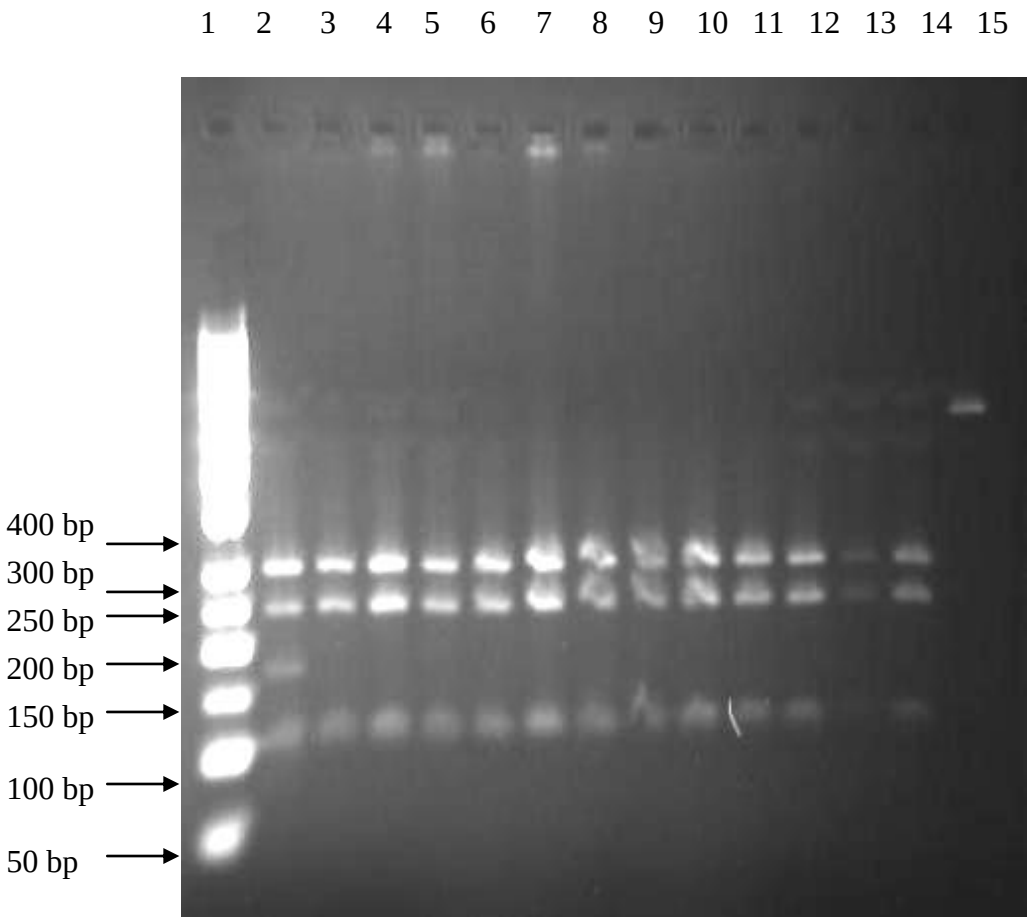


Figure 16: Restriction digest of samples amplified for the 652 bp IVS 26+T80C region by *Alw*26I, confirmed by the presence of bands after electrophoresis on a 2% agarose gel in the presence of ethidium bromide. Lane 1 contains a 50 bp molecular weight marker, lanes 2 to 14 contain restriction digestions of different samples, and lane 15 contains the undigested control sample. Amplification and restriction digestion fragment sizes were confirmes using standard curves ($R^2 = 0.9996$ and $R^2 = 0.9865$, respectively).

Between 354 and 385 samples were successfully genotyped for the seven polymorphic sites (Table 6). The allele frequencies were calculated from the genotypes and compared to allele frequencies from European populations (Soranzo et al. 2004 and Takane et al. 2004). The genotypic data and the following analysis thereof were done by first separating the general population samples from those of the HIV positive samples, and then for all samples together. The data for the general population was not significantly different from that of the HIV positive samples, which was why the data was chosen to be presented for all samples together.

The minor allele frequencies are lower than those found in the studies on Europeans, except for the 5231 locus, which shows a very similar minor allele frequency, and the T-129C locus which has a higher minor allele frequency. The frequencies for the 1236T and 3435T alleles are much lower in the southern African population compared to the European population (Soranzo et al. 2004). The $\chi^2$ test was performed using the genotyping frequencies and showed no significant deviation from Hardy-Weinberg equilibrium, except for the 2677 locus (Table 6).

The $\chi^2$ test was not the most appropriate test for the G2677T/A polymorphism, as more than two alleles are possible, resulting in some genotype groups with small numbers, for which the $\chi^2$ test is can not accurately calculate the $\chi^2$ and P values for. The Fisher's exact test is able to handle SNPs with more than two possible alleles and produced a P value was 0.195 (> 0.05), when the 2677T and 2677A

allele frequencies were combined, indicating no significant deviation from the expectations of Hardy Weinberg equilibrium.

All the linkage disequilibrium values for the Lewontin's coefficients and Pearson correlations were found to be less than 1 and greater than zero, except for the Pearson correlation between G2677T/A and C3435T, and between G2677T/A and IVS 26+T80C, which were found to be zero (Table 7). The three highest $|D'|$ values were found between C1236T and IVS 26+T80C (0.75), C1236T and C3435 (0.74), and between T-129C and C3435T (0.69). The lowest $|D'|$ value was between G2677T/A and IVS 26+T80C (0.01). The three highest $r^2$ values were seen between C3435T and IVS 26+T80C (0.39), C3435T and IVS 25+3050 (0.13) and between IVS 26+T80C and IVS 25+3050 (0.12). The lowest $r^2$ value was seen between T-129C and C1236T (0.0001).

The linkage disequilibrium values obtained from our data were compared to the values obtained by Soranzo et al. 2004 for Europeans. The $r^2$ values for our data are much lower than those for Europeans (Table 8). The $|D'|$ values for all but the 1236-3435 and 1236-IVS 26+80 pairs were greater in the data from Soranzo et al. (2004). The linkage disequilibrium with the T-129C site could not be compared as this site was not included in the Soranzo et al. (2004) study.

Table 6: Genotypic data with test for Hardy-Weinberg equilibrium.

| Locus | dbSNP rs# | Chr. Position | No | DF | Observed Genotype Number | | | Hardy-Weinberg Equilibrium $\chi^2$ | P | MAF (My data) | MAF (Europeans [a]) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| T-129C (UTR) | rs3213619 | 87230193 | 354 | 2 | 253 (TT) | 95 (TC) | 6 (CC) | 0.87 | >0.05 | 0.15 (C) | 0.02 (C) |
| C1236T (Exon 12) | rs1128503 | 87179601 | 385 | 2 | 33 (CC) | 163 (CT) | 189 (TT) | 0.13 | >0.05 | 0.30 (C) | 0.47 (T) |
| G2677T/A (Exon 21) | rs2032582 | 87160618 | 376 | 3 | 216 (GG) | 136 (GA) | 12 (GT) | 9.44 | <0.05 | 0.23 (A/T) | 0.47 (A/T) |
| | | | | | 0 (AA) | 1 (TT) | 11 (AT) | | | | |
| IVS 25+G3050T (Intron 25) | rs1002204 | 87141497 | 374 | 2 | 281 (GG) | 81 (GT) | 12 (TT) | 2.62 | >0.05 | 0.14 (T) | 0.40 (T) |
| IVS 25+ T5231C (Intron 25) | rs4437575 | 87139316 | 385 | 2 | 104 (TT) | 188 (TC) | 93 (CC) | 0.91 | >0.05 | 0.49 (C) | 0.42 (C) |
| C3435T (Exon 26) | rs1045642 | 87138645 | 375 | 2 | 290 (CC) | 79 (CT) | 6 (TT) | 0.02 | >0.05 | 0.12 (T) | 0.42 (C) |
| IVS 26+T80C (Intron 26) | rs2235048 | 87138511 | 362 | 2 | 285 (TT) | 74 (TC) | 3 (CC) | 1.47 | >0.05 | 0.11 (C) | 0.43 (C) |

[a] Soranzo et al. 2004

Table 7: Linkage disequilibrium estimates for all polymorphic pairs.

| | T-129C | C1236T | G2677T/A | IVS 25+G3050T | IVS 25+T5231C | C3435T | IVS 26+T80C |
|---|---|---|---|---|---|---|---|
| T-129C | - | **0.0001** | **0.0005** | **0.005** | **0.0008** | **0.01** | **0.004** |
| C1236T | **0.04** | - | **0.007** | **0.02** | **0.0002** | **0.03** | **0.03** |
| G2677T/A | **0.09** | **0.23** | - | **0.002** | **0.004** | **0.00** | **0.00** |
| IVS 25+ G3050T | **0.43** | **0.55** | **0.06** | - | **0.03** | **0.13** | **0.12** |
| IVS 25+ T5231C | **0.07** | **0.02** | **0.11** | **0.46** | - | **0.01** | **0.03** |
| C3435T | **0.69** | **0.74** | **0.06** | **0.38** | **0.30** | - | **0.39** |
| IVS 26+T80C | **0.40** | **0.75** | **0.01** | **0.39** | **0.51** | **0.65** | - |

**Lewontin's coefficient (|*D′*|)**

**Pearson correlation ($r^2$)**

Table 8: Linkage disequilibrium estimates for all polymorphic pairs from the data from Soranzo et al. (2004).

| | C1236T | G2677T/A | IVS 25+G3050T | IVS 25+T5231C | C3435T | IVS 26+T80C |
|---|---|---|---|---|---|---|
| C1236T | - | **0.213** | **0.587** | **0.349** | **0.190** | **0.199** |
| G2677T/A | **0.565** | - | **0.350** | **0.786** | **0.901** | **1.000** |
| IVS 25+G3050T | **0.896** | **0.623** | - | **0.563** | **0.399** | **0.327** |
| IVS 25+T5231C | **0.788** | **0.934** | **0.877** | - | **0.795** | **0.800** |
| C3435T | **0.566** | **1.000** | **0.728** | **0.892** | - | **0.909** |
| IVS 26+T80C | **0.583** | **1.000** | **0.651** | **0.944** | **1.000** | - |

**Lewontin's coefficient (|*D′*|)**

**Pearson correlation ($r^2$)**

A total of 383 samples were used for the haplotype analysis. This number was obtained by including only samples, for which at least four of the seven polymorphisms had been successfully genotyped. Based on the alleles present in the data set, there were 152 possible haplotypes, and 56 of these were detected in the data set. PHASE was able to resolve 38% of the haplotype pairs, with 62% remaining ambiguous.

The 10 most common haplotypes in our data set with the frequencies inferred by PHASE, using the haplotype population frequencies, and the expected frequencies obtained from our allele frequencies, where haplotype 1 is the most common (Table 9). All ten runs resulted in the exact same data. The most common haplotyes with respect to positions -129, +1236, +2677, IVS 25+3050, IVS 25+5231, +3435, IVS 26+80 were TTGGTCT (15,5%), TTGGCCT (12,7%), TCGGCCT (12,5%), TCGGTCT (10,6%), TTAGCCT (10,1%), CTGGCCT (4,8%), TTGTTCT (3,4%), CTGGTCT (3,4%), TTGTCTC (2,9%), TTAGTCT (2,0%).

The most commonly reported *ABCB1* haplotype is 1236-2677-3435. The frequencies of various phases of this haplotype were compared in a number of populations, where the South African data represents our data, compared to data for the Benin population in West Africa (Allabi et al 2005), African Americans (AfAm) (Kim et al. 2001; Tang et al. 2002), Caucasians (Cauc) (Kim et al. 2001; Tang et al. 2002) and Chinese (Tang et al. 2002) (Figure 17).

Table 9: Comparison of the inferred and expected frequencies of the 10 most common haplotypes.

| Locus / Haplotype | -129 T/C | 1236 C/T | 2677 G/T/A | IVS 25+ 3050 G/T | IVS 25+ 5231 T/C | 3435 C/T | IVS 26+ 80 T/C | Frequency Inferred | Expected |
|---|---|---|---|---|---|---|---|---|---|
| 1 | T | T | G | G | T | C | T | 0.155 | 0.160 |
| 2 | T | T | G | G | C | C | T | 0.127 | 0.172 |
| 3 | T | C | G | G | C | C | T | 0.125 | 0.088 |
| 4 | T | C | G | G | T | C | T | 0.106 | 0.081 |
| 5 | T | T | A | G | C | C | T | 0.101 | 0.054 |
| 6 | C | T | G | G | C | C | T | 0.048 | 0.035 |
| 7 | T | T | G | T | T | C | T | 0.034 | 0.030 |
| 8 | C | T | G | G | T | C | T | 0.034 | 0.030 |
| 9 | T | T | G | T | C | T | C | 0.029 | 0.016 |
| 10 | T | T | A | G | T | C | T | 0.020 | 0.037 |

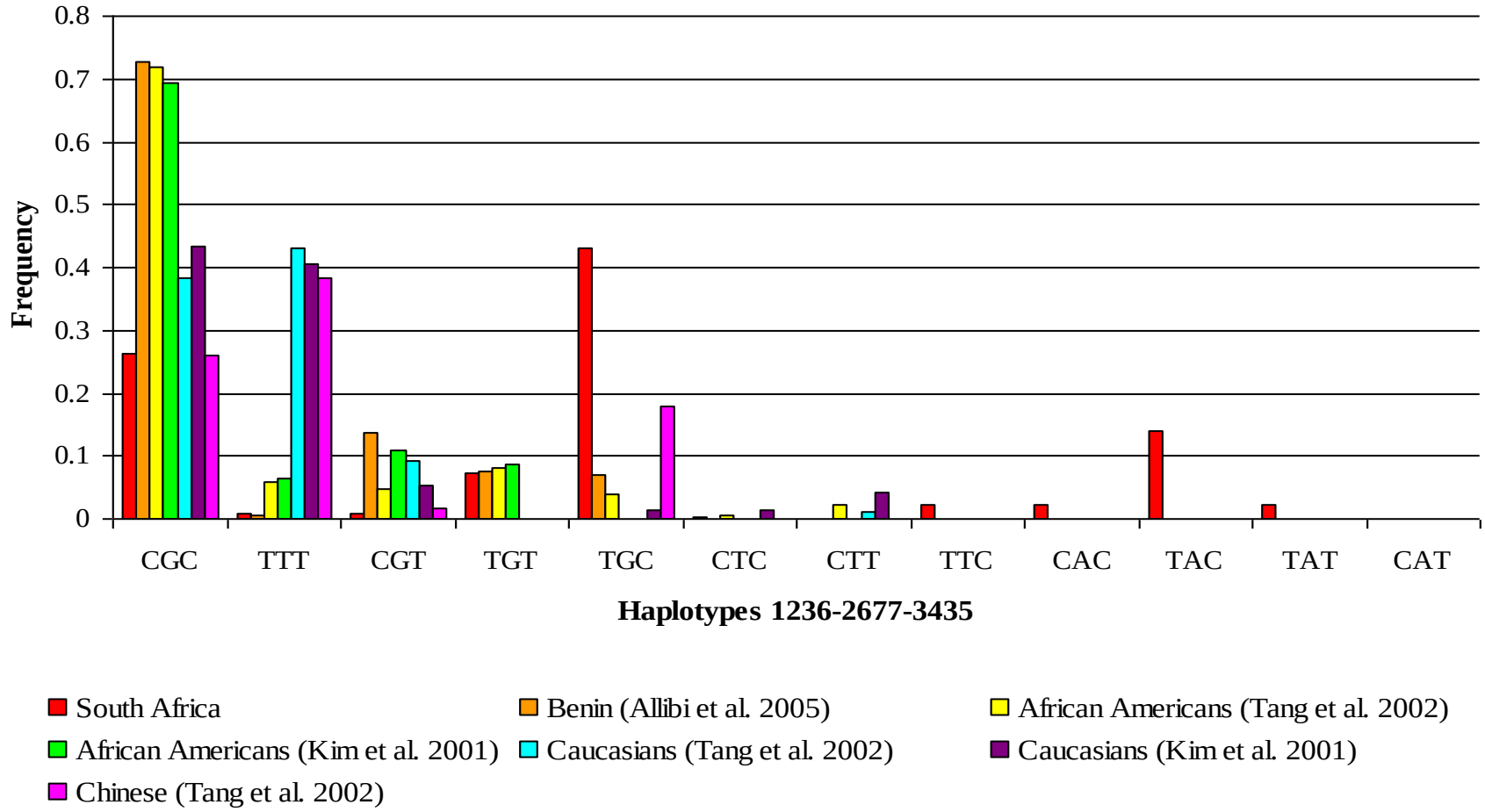


Figure 17: Bar chart for the comparison of the common *ABCB1* 1236-2677-3435 haplotype in various populations.

The CGC phase has been found to be most common in West African and African American populations, whereas the TTT phase has been found to be most common in non-African populations. The CGC occurs at a much lower frequency than would be expected, as we would expect the data to be similar to that of other African-originating populations. The TTT phase is at a very low frequency, as expected since those three alleles are at a low frequency in our data set but at a high frequency in European populations, but is also lower than the frequencies observed in West Africans and African Americans. For the CGT, CTC, CTT, TTC, CAC, and TAT phases, the frequencies are very low. For the TGT phase the frequency is approximately the same as for the other populations of African origin. The TGC phase occurs at a frequency of 43% and is the most common phase of this haplotype in our data. It is at a much higher frequency then in any other population, with the next highest frequency being about 18% in the Chinese. The TAC phase occurs at a frequency of about 15% and the CAT phase was not found at all in our data.

Two reduced median networks were constructed for the 10 most common haplotypes (Table 9). The first was generated using the haplotypes of all seven polymorphisms (Figure 18). The second was generated using the haplotypes of the six closest polymorphisms and excluding the T-129C polymorphisms to determine any difference in the pattern of the network seen by removing the most distant variant (Figure 19). The ancestral haplotypes, based on those present in other primate sequences (www.ensemble.org), (TCGGTCT and CGGTCT) are coloured green.

There is only one base difference between most of the ten haplotypes. With all seven SNPs there is one haplotype that is three base changes different from two other haplotypes. The most common change is at the 5231 locus and the least common changes are at the 3435 and 80 loci. With only six SNPs there are three haplotypes that are two base changes different from other SNPs. The most common changes are at the 3050 and 5231 loci.

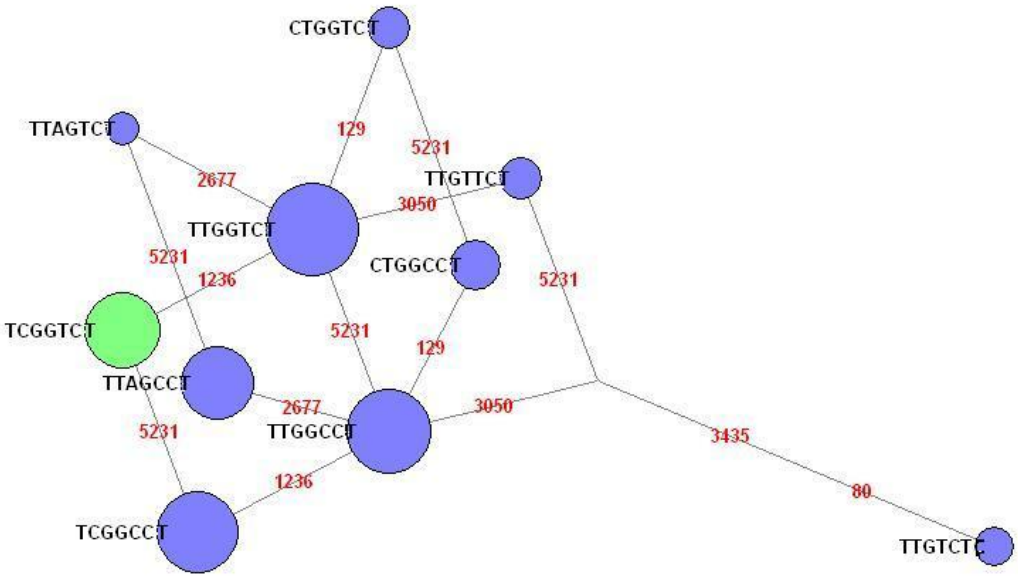


Figure 18: Reduced Median Network for the 10 most common haplotypes for all seven polymorphisms.

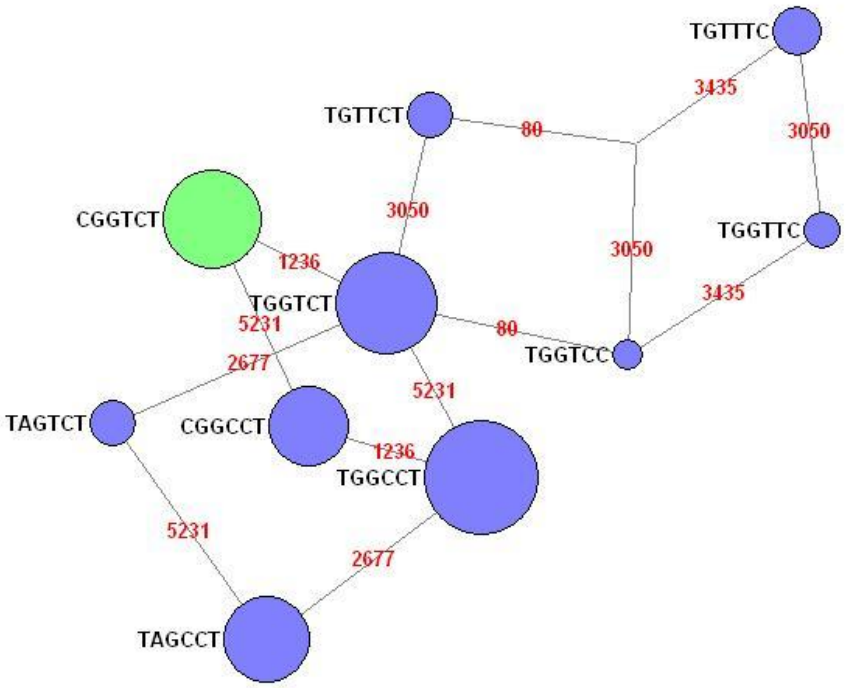


Figure 19: Reduced Median Network for the 10 most common haplotypes for the six closest polymorphisms.

### 3.5 Calibrator Normalised Relative Quantification of mRNA

The $C_t$ values for each repeat were used to determine the mean $C_t$ values, the $\Delta C_t$, $\Delta\Delta C_t$ and RQ ($2^{-\Delta\Delta Ct}$) values with standard deviations for each sample (Appendix III). The positive and negative deviation was calculated for the RQ values for each sample, as well as for the mean RQ (Table 10).

The sample RQ values are compared to the calibrator, whose value is set to one. Sample 1180 has the highest RQ value, but also has the greatest standard deviation, whereas sample 1179 has the lowest RQ value and a very small standard deviation (Figure 20).

The t-test of the RQ values produced a t value of -4.09 and a p value of 0.0004, which is less than the significance threshold of $\alpha = 0.05$, indicating significant differences between the RQ values.

Table 10: RQ Values with standard deviation for 28 samples and mean RQ.

| Sample | RQ ($2^{-\Delta\Delta C}$) | Positive deviation | Negative deviation |
|---|---|---|---|
| 1036 | 0.23 | 0.23 | 0.12 |
| 1038 | 0.61 | 0.23 | 0.16 |
| 1057 | 0.85 | 0.51 | 0.33 |
| 1066 | 0.48 | 0.39 | 0.22 |
| 1081 | 0.42 | 0.13 | 0.1 |
| 1087 | 0.24 | 0.17 | 0.11 |
| 1106 | 0.35 | 0.25 | 0.14 |
| 1129 | 0.3 | 0.26 | 0.14 |
| 1137 | 0.96 | 0.71 | 0.41 |
| 1141 | 0.25 | 0.17 | 0.1 |
| 1150 | 0.37 | 0.42 | 0.2 |
| 1153 | 0.68 | 0.45 | 0.27 |
| 1156 | 1.05 | 0.33 | 0.25 |
| 1158 | 0.86 | 0.21 | 0.17 |
| 1169 | 0.85 | 0.18 | 0.15 |
| 1173 | 0.78 | 0.39 | 0.26 |
| 1174 | 0.59 | 0.4 | 0.23 |
| 1175 | 0.41 | 0.67 | 0.25 |
| 1176 | 0.71 | 0.16 | 0.12 |
| 1178 | 1.42 | 0.35 | 0.28 |
| 1179 | 0.2 | 0.08 | 0.06 |
| 1180 | 1.66 | 1.84 | 0.88 |
| 1182 | 0.86 | 0.22 | 0.17 |
| 1184 | 0.86 | 0.15 | 0.12 |
| 1185 | 1.16 | 0.25 | 0.21 |
| 1191 | 0.98 | 1.04 | 0.5 |
| Mean | 0.70 | 0.39 | 0.28 |

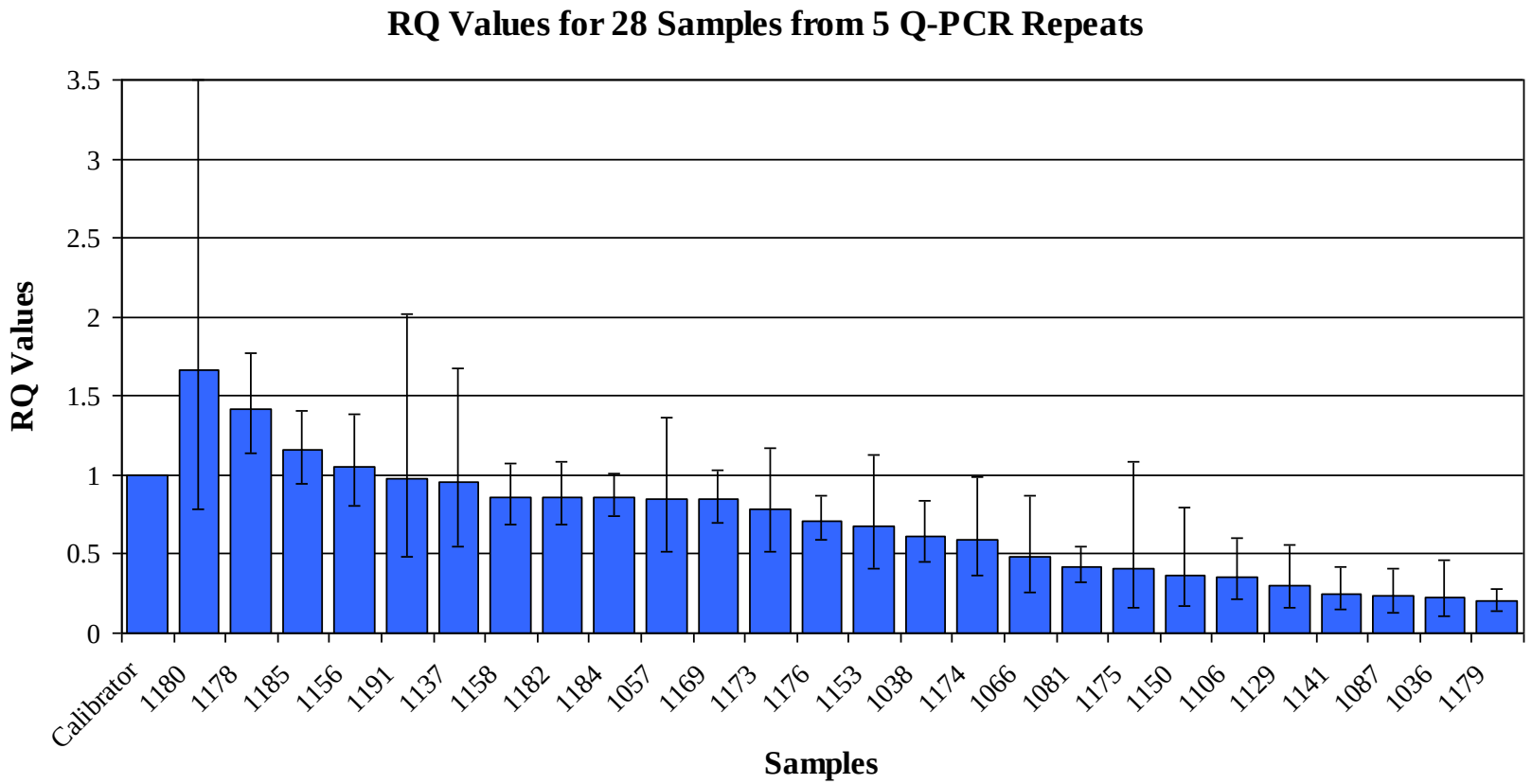


Figure 20: Bar chart of RQ values and standard deviations for 28 samples, generated from five repeats of relative quantification using cDNA. RQ values are compared to the Calibrator, set to a value of one, ordered from the highest to lowest RQ value.

### 3.6 Association Studies

The $CD4^+$ counts for the 151 available samples were compared at three time points, baseline, time point one, and time point two, given in days since baseline. In terms of the clinic's database, time point one is supposed to be as close to six months as possible (+/- 180 days) after baseline readings, and time point two is supposed to be as close to twelve months as possible (+/- 360 days) after baseline readings. The viral load data was not as complete as the $CD4^+$ data and was only available for 46 samples. We plotted the $CD4^+$ counts against the number of days since baseline and instead of producing clear time intervals, the time points were located over a range of days, often overlapping between time points one and two (Figure 21).

The same comparison was made using the viral load data. Due to the high values at baseline compared to time points one and two, the log values of viral load were used. The viral load counts were plotted against the number of days since baseline and a wide range of time points was observed (Figure 22).

Inconsistency in patient visits to the clinic produces an overlap in the data points for the two time points, instead of distinct groupings around 180 days and 360 days. This is pattern is observed for both the $CD4^+$ count and viral load data. There are fewer data points for the viral load data, as this is not measured as consistently at clinic visists as $CD4^+$ counts.

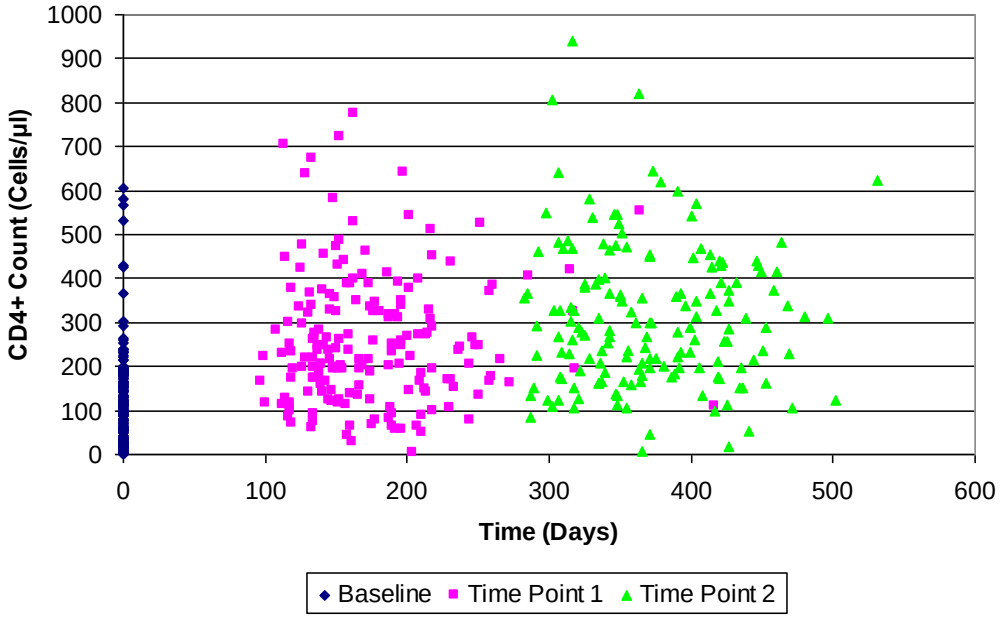


Figure 21: $CD4^+$ counts for three time points: baseline (time zero), time point one (as close to six months or 180 days as possible) and time point two (as close to twelve months or 360 days as possible). Instead of distinct clusters around 180 and 360 days, the data points form a large cluster over a large time interval and there is an overlap between data points from time points one and two.

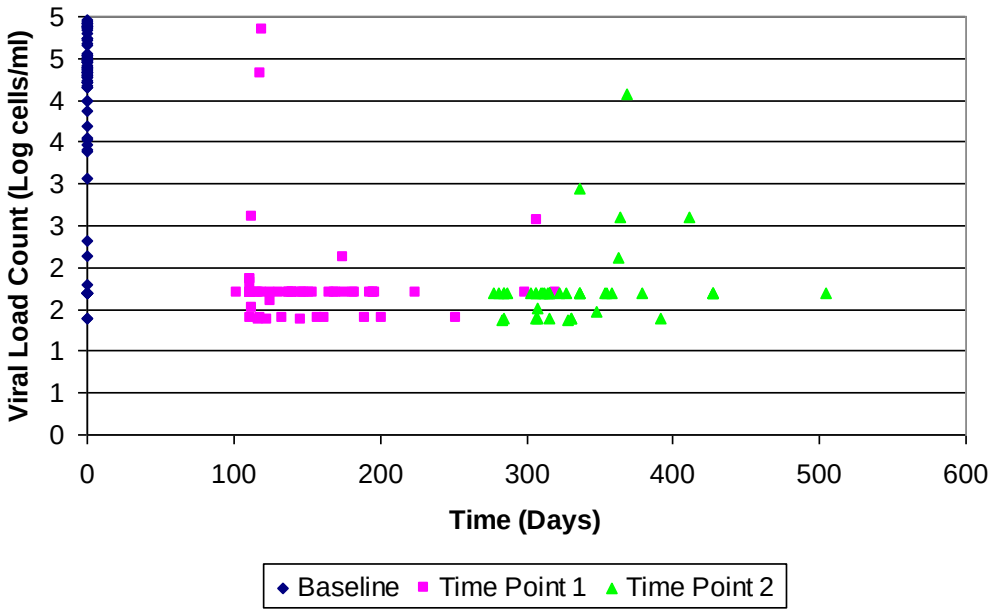


Figure 22: Log viral load counts for three time points: baseline (time zero), time point one (as close to six months or 180 days as possible) and time point two (as close to twelve months or 360 days as possible). Instead of distinct clusters around 180 and 360 days, the data points form a large cluster over a large time interval and overlap between the time points one and two.

The mean $CD4^+$ counts were grouped into the different genotypes of each polymorphism to compare the increase in mean $CD4^+$ count between each genotype at the three time points: baseline, time point one and time point two (Figure 23). The haplotypes that were resolved by PHASE were too few to include in the association studies, so the mean $CD4^+$ counts were also grouped into combinations of SNPs, based on linkage disequilibrium data, and the increase compared between each genotype combination. The combination of C1236T, G2677T/A and C3435T was also used to analyse the mean $CD4^+$ count data for each genotype combination (Figure 24).

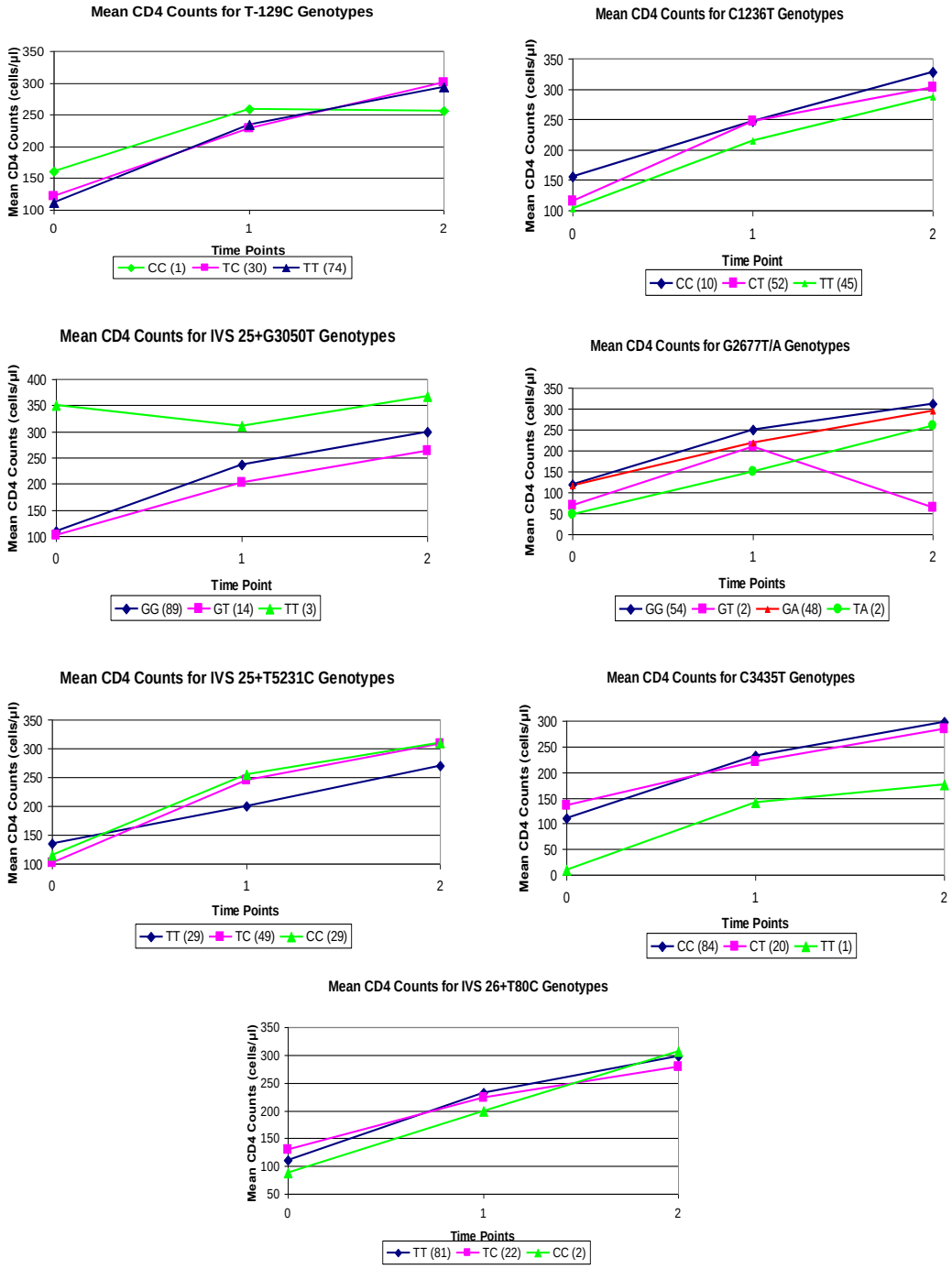


Figure 23: Increase in mean $CD4^+$ counts for the genotypes of each of the T-129C, C1236T, G2677T/A, IVS 25+G3050T, IVS 25+T5231C, C3435T and IVS 26+T80C polymorphisms at baseline, time point one and time point 2.

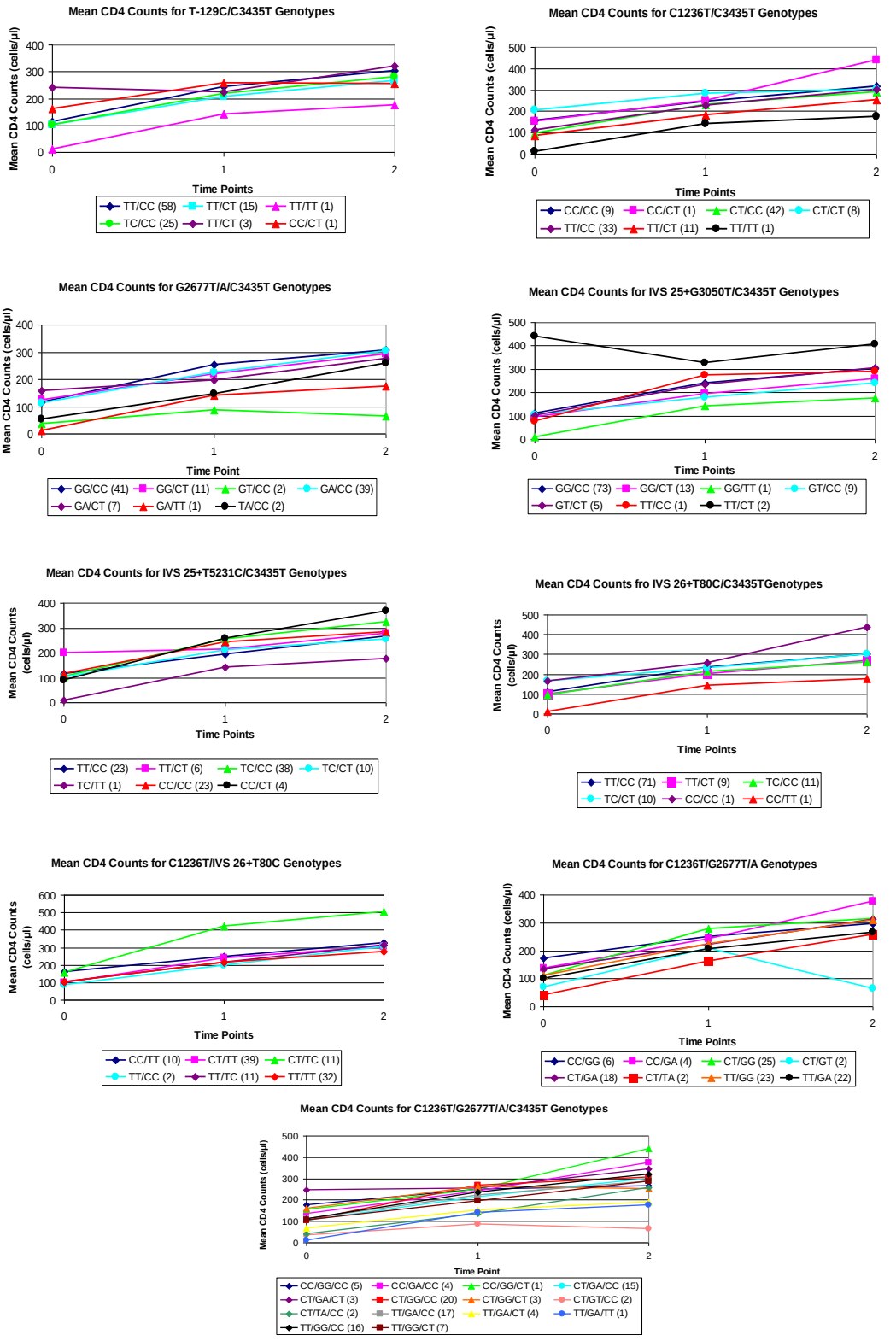


Figure 24: Increase in mean CD4$^+$ counts for the genotype combinations of each of the SNP combinations T-129C/C3435T, C1236T/C3435T,

G2677T/A/C3435T, IVS 25+G3050T/C3435T, IVS 25+T5231C/C3435T, IVS 26+T80C/C3435T, C1236T/ IVS 26+T80C, C1236T/G2677T/A and C1236T/G2677T/A./C3435T polymorphisms at baseline, time point one and time point 2.

An ANOVA test was performed to test for association between the genotypes and the difference in $CD4^+$ counts at baseline and time 1, and the difference in viral load counts at baseline and time 2, for the individual SNPs (Table 11). All P values were above the level of significance $\alpha = 0.05$, indicating that there was no association for any of the SNPs at either time point.

The ANOVA test between the genotypes and the difference in $CD4^+$ counts at baseline and time 1, the difference in $CD4^+$ counts at baseline and time 2 for the SNP combinations showed no association for any of the SNP combinations. The order of SNPs for sorting the data was altered to determine any change in the results, but there was still no detectable association using SNP combinations. The ANOVA test was run a second time, this time without the outlying samples. All P values were above 0.05, indicating no significant difference and no association.

A student's t test showed that there was a table-wide significant difference between the p values from the ANOVA test ($p = 0.017$).

Table 11: Results of the ANOVA test for testing the association between the SNPs and mean $CD4^+$ count differences at the first and second time points.

| Variables | DF | F value | p value |
|---|---|---|---|
| T-129C; Time1 | 15 | 1.26 | 0.3061 |
| T-129C; Time2 | 8 | 0.54 | 0.8071 |
| C1236T; Time1 | 21 | 0.95 | 0.5491 |
| C1236T; Time2 | 12 | 0.94 | 0.5429 |
| G2677T/A; Time1 | 21 | 0.71 | 0.7663 |
| G2677T/A; Time2 | 12 | 1.04 | 0.4757 |
| IVS 25+G3050T; Time1 | 7 | 0.56 | 0.7796 |
| IVS 25+G3050T; Time2 | 5 | 0.44 | 0.8235 |
| IVS 25+T5231C; Time1 | 22 | 1.17 | 0.3849 |
| IVS 25+T5231C; Time2 | 16 | 0.58 | 0.8317 |
| C3435T; Time1 | 10 | 0.72 | 0.6970 |
| C3435T; Time2 | 8 | 1.69 | 0.1761 |
| IVS 26+T80C; Time1 | 11 | 1.68 | 0.1345 |
| IVS 26+T80C; Time2 | 8 | 1.36 | 0.2855 |

The mean viral load counts were grouped into the different genotypes of each polymorphism to compare the decrease in mean viral load count between each genotype (Figure 25) and each genotype combination for the SNP combinations (Figure 26) at the three time points: baseline, time point one and time point two.

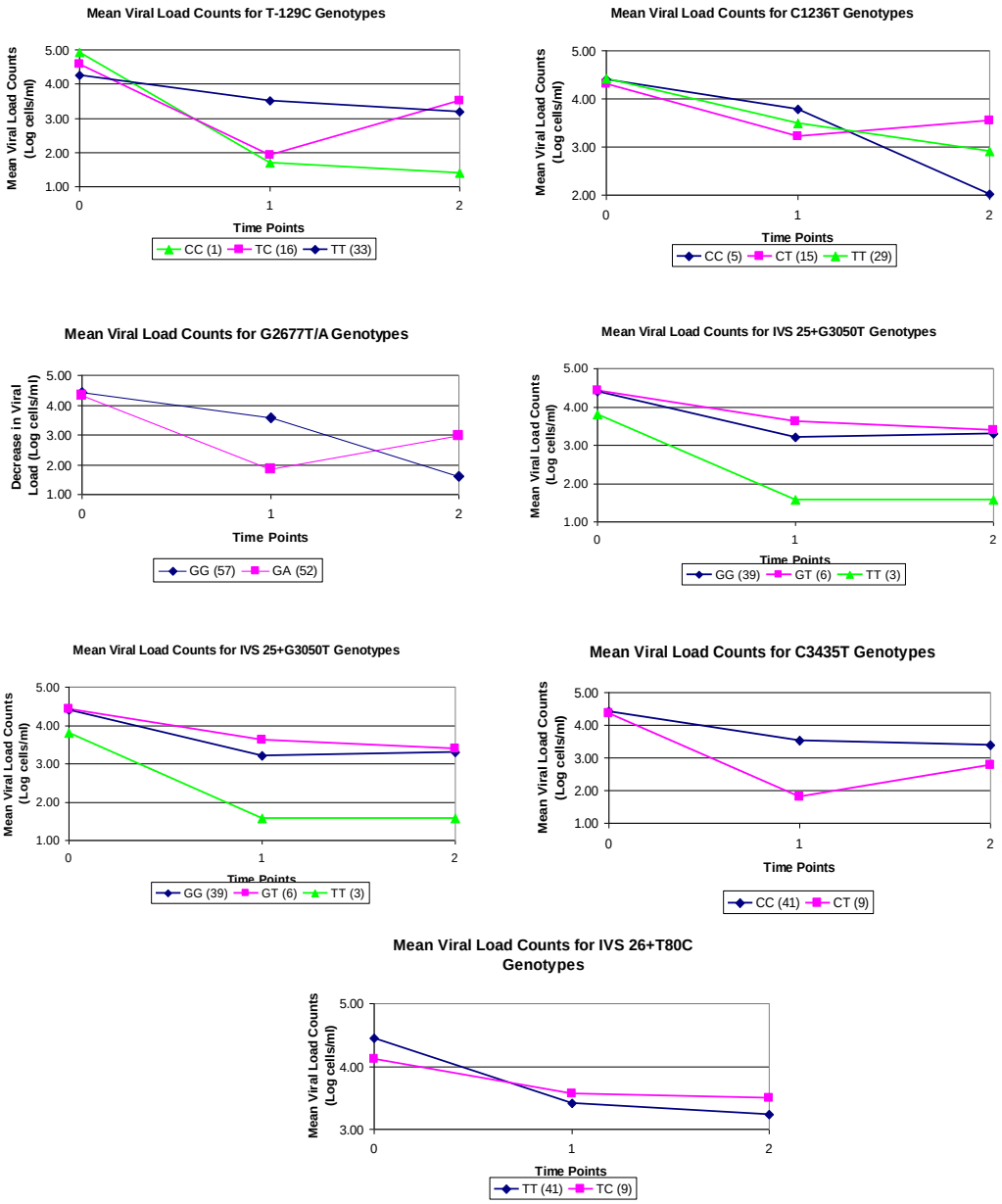


Figure 25: Decrease in mean viral load counts for the genotypes of each of the T-129C, C1236T, G2677T/A, IVS 25+G3050T, IVS 25+T5231C, C3435T and IVS 26+T80C polymorphisms at baseline, time point one and time point 2.

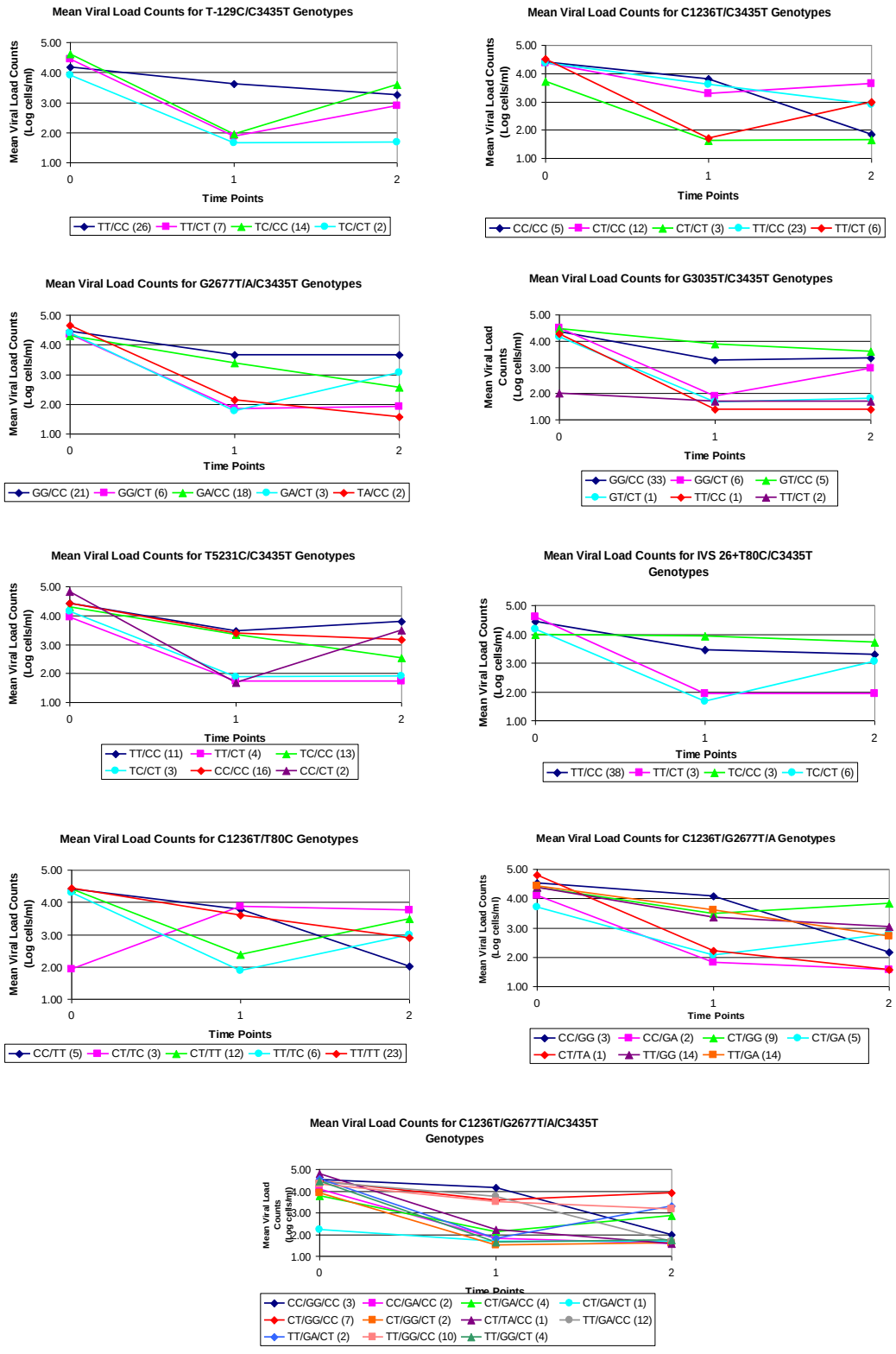


Figure 26: Decrease in mean viral load counts for the genotype combinations of each of the SNP combinations T-129C/C3435T, C1236T/C3435T,

G2677T/A/C3435T, IVS 25+G3050T/C3435T, IVS 25+T5231C/C3435T, IVS 26+T80C/C3435T, C1236T/ IVS 26+T80C, C1236T/G2677T/A and C1236T/G2677T/A./C3435T polymorphisms at baseline, time point one and time point 2.

An ANOVA test was performed to test for association between the genotypes and the difference in viral load counts at baseline and time 1, and the difference in viral load counts at baseline and time 2, for the individual SNPs (Table 12). There was no data available for the second time points for the T-129C, C1236T, IVS 25+G3050T and IVS 26+T80C SNPs, so association cound not be tested for. For the IVS 25+T5231C and C3435T SNPs the P values were above the significance level of $\alpha = 0.05$ but the P value for the G2677T/A polymorphism was 0.0121, indicating a significant difference in the mean viral load count at the second time point. Only two genotypes were found to correlate with the viral load data, 2677GG and 2677GA, the two most common genotypes for this polymorphism. At the fist time point the viral load counts at each genotype have decreased, but at the second time point the viral load counts for the 2677GA genotype have increased significantly compared to the 2677GG genotype.

The ANOVA test between the genotypes and the difference in viral load counts at baseline and time 1, the difference in viral load counts at baseline and time 2 for the SNP combinations showed no association for any of the SNP combinations. The order of SNPs for sorting the data was altered to determine any change in the results, but there was no detectable association using SNP combinations. The

ANOVA test was run a second time, this time without the outlying samples. All P values were above 0.05 for the first time point, indicating no significant difference and no association, and the P value for the G2677T/A polymorphism at time point two remained 0.0121.

A student's t test showed that there was a table-wide significant difference between the p values from the ANOVA test for the mean viral load counts ($p < 0.0001$).

Table 12: Results of the ANOVA test for testing the association between the SNPs and mean viral load count differences at the first and second time points.

| Variables | DF | F value | p value |
|---|---|---|---|
| T-129C; Time1 | 3 | 0.33 | 0.8009 |
| C1236T; Time1 | 3 | 0.24 | 0.8683 |
| G2677T/A; Time1 | 2 | 0.64 | 0.5507 |
| G2677T/A; Time2 | 2 | 81.70 | 0.0121 |
| IVS 25+G3050T; Time1 | 2 | 0.32 | 0.7375 |
| IVS 25+T5231C; Time1 | 3 | 0.05 | 0.9835 |
| IVS 25+T5231C; Time2 | 2 | 0.09 | 0.9238 |
| C3435T; Time1 | 1 | 0.51 | 0.4908 |
| C3435T; Time2 | 1 | 0.02 | 0.8956 |
| IVS 26+T80C; Time1 | 2 | 0.04 | 0.9633 |

The mean proportional increase in $CD4^+$ counts for each genotype (Figure 27) and genotype combination (Figure 28) and mean proportional decrease in viral load counts each genotype (Figure 29) and genotype combination (Figure 30) were calculated at each of the two time point.

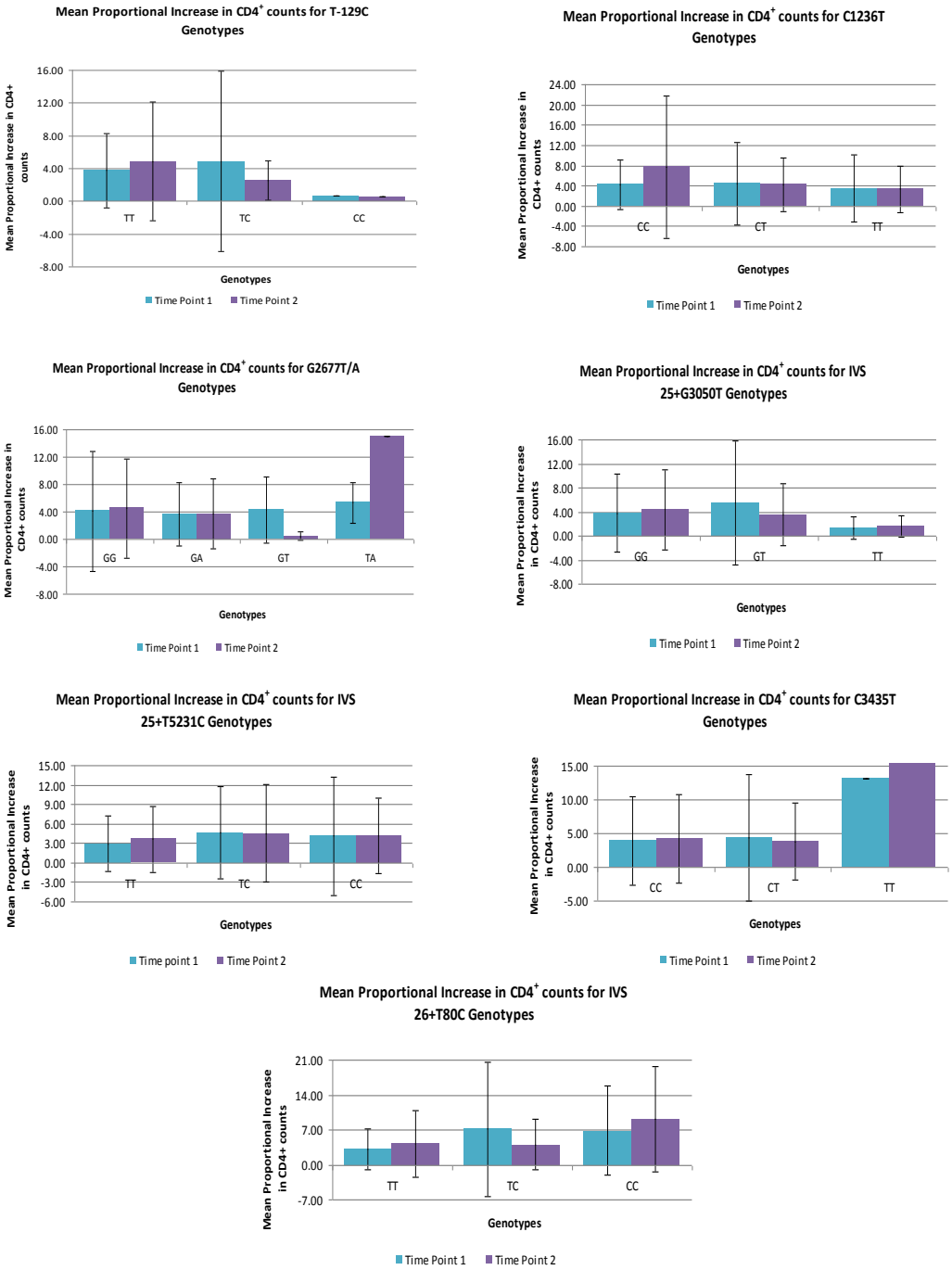


Figure 27: Mean proportional increase in $CD4^+$ count for each genotype at each polymorphism at time points one and two.

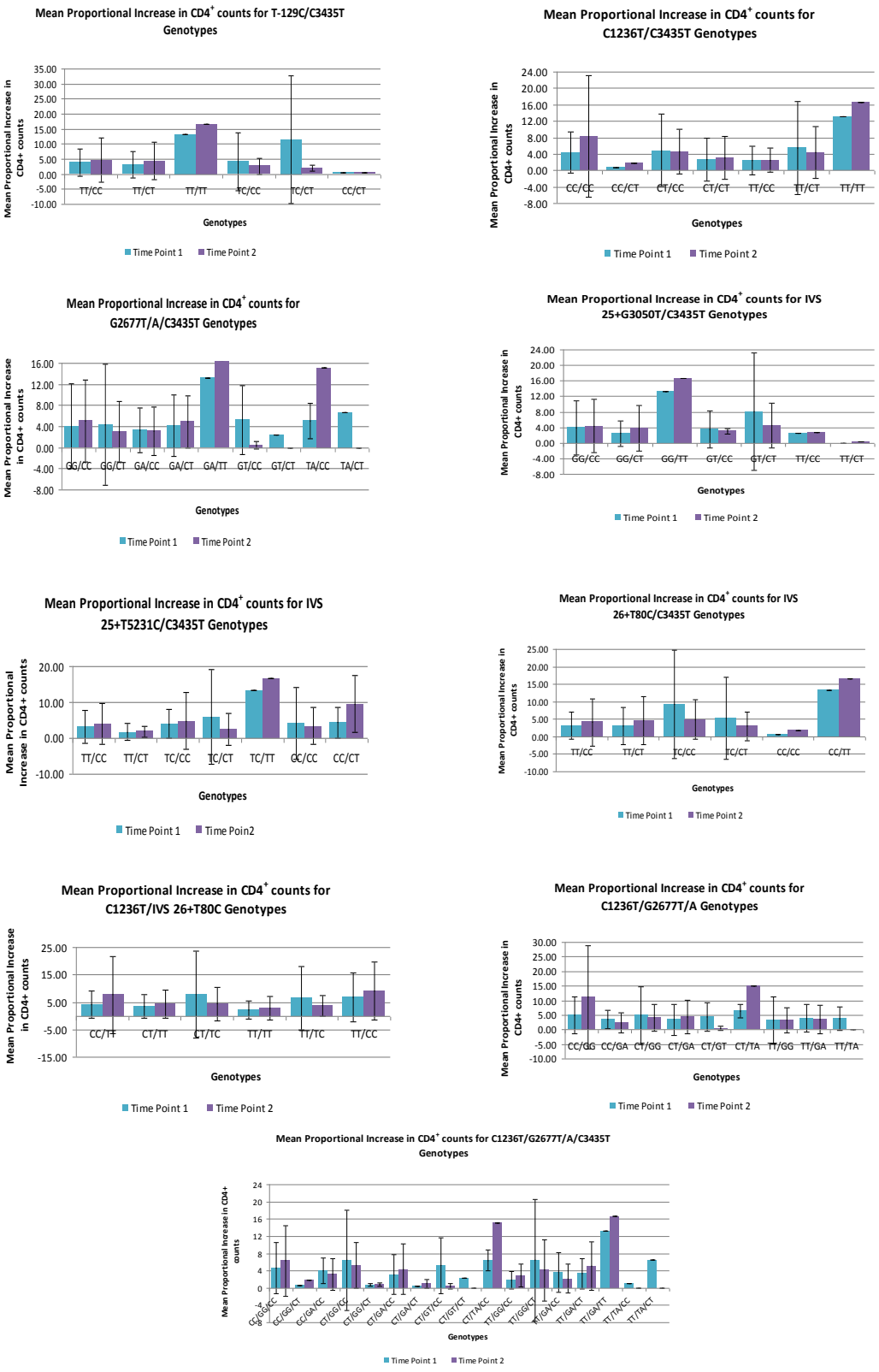


Figure 28: Mean proportional increase in $CD4^+$ count for each genotype combination for each polymorphism combination at time points one and two.

The Kruskal-Wallis test was performed to determine any association between genotypes and the mean proportional increase in $CD4^+$ count at each time point, instead of the mean values for the absolute difference between baseline and the two time points. All SNPs had P values above the significance value of $\alpha = 0.05$ (Table 13). There was no association detected for any of the SNP combinations using the proportional increases. A student's t test showed that there was a table-wide significant difference between the p values from the Kruskal-Wallis test ($P < 0.0001$). The Kruskal-Wallis test was repeated, omitting the outlying samples, yielding no change in the P values.

Table 13: Results of the Kruskal-Wallis test for association between the SNPs and the mean proportional increase in $CD4^+$ count at the first and second time points.

| **Variables** | **DF** | **$\chi^2$ value** | **p value** |
|---|---|---|---|
| T-129C; Time1 | 2 | 0.9196 | 0.6314 |
| T-129C; Time2 | 2 | 1.2446 | 0.5367 |
| C1236T; Time1 | 2 | 0.8569 | 0.6515 |
| C1236T; Time2 | 2 | 0.9072 | 0.6353 |
| G2677T/A; Time1 | 3 | 3.2451 | 0.3554 |
| G2677T/A; Time2 | 3 | 4.8799 | 0.1808 |
| IVS 25+G3050T; Time1 | 2 | 0.6067 | 0.7383 |
| IVS 25+G3050T; Time2 | 2 | 0.9763 | 0.6138 |
| IVS 25+T5231C; Time1 | 2 | 4.3688 | 0.1125 |
| IVS 25+T5231C; Time2 | 2 | 0.6974 | 0.7056 |
| C3435T; Time1 | 2 | 3.3430 | 0.1880 |
| C3435T; Time2 | 2 | 2.5727 | 0.2763 |
| IVS 26+T80C; Time1 | 2 | 0.9239 | 0.6301 |
| IVS 26+T80C; Time2 | 2 | 0.8400 | 0.6571 |

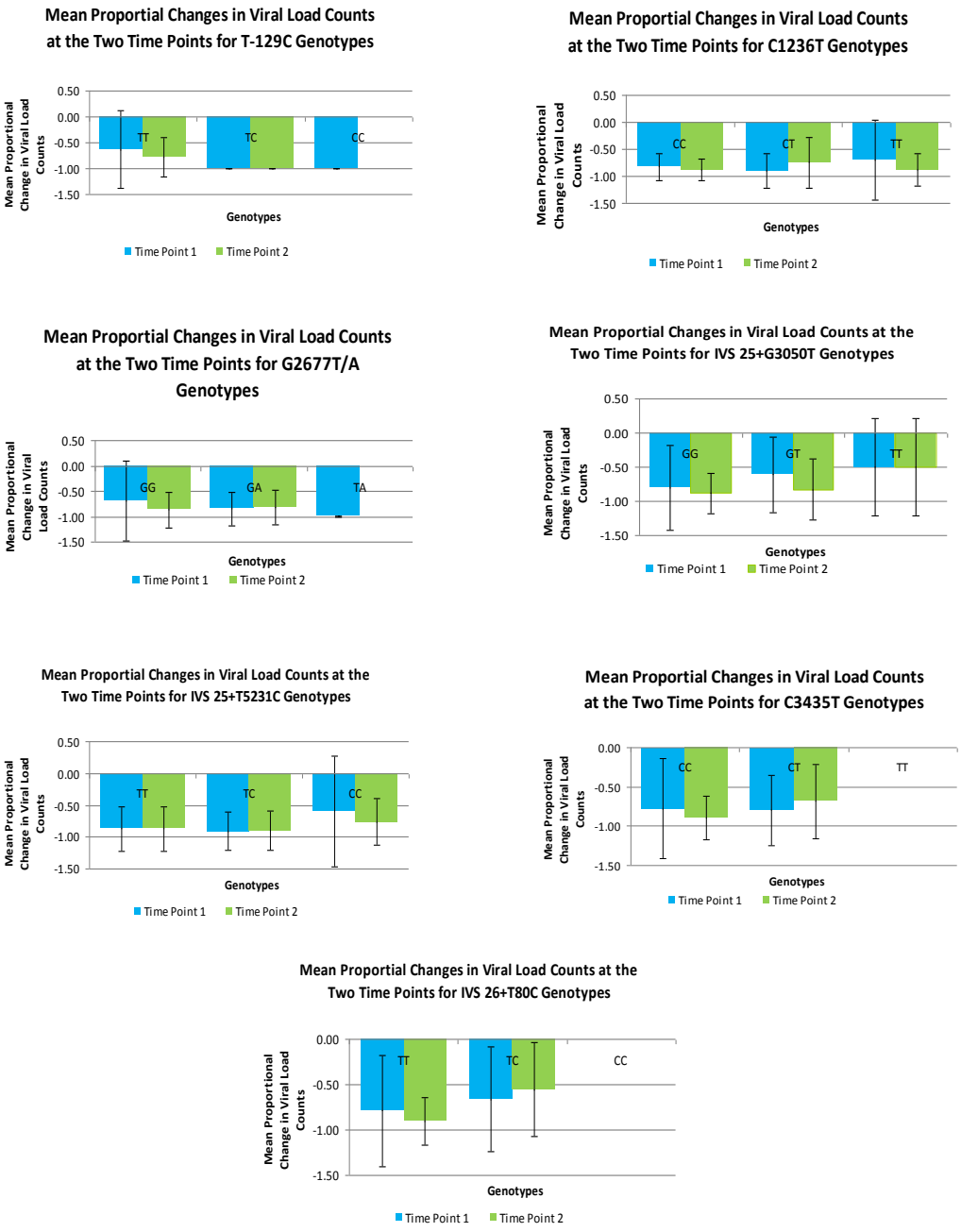


Figure 29: Mean proportional decrease in viral load count for each genotype at each polymorphism at time points one and two.

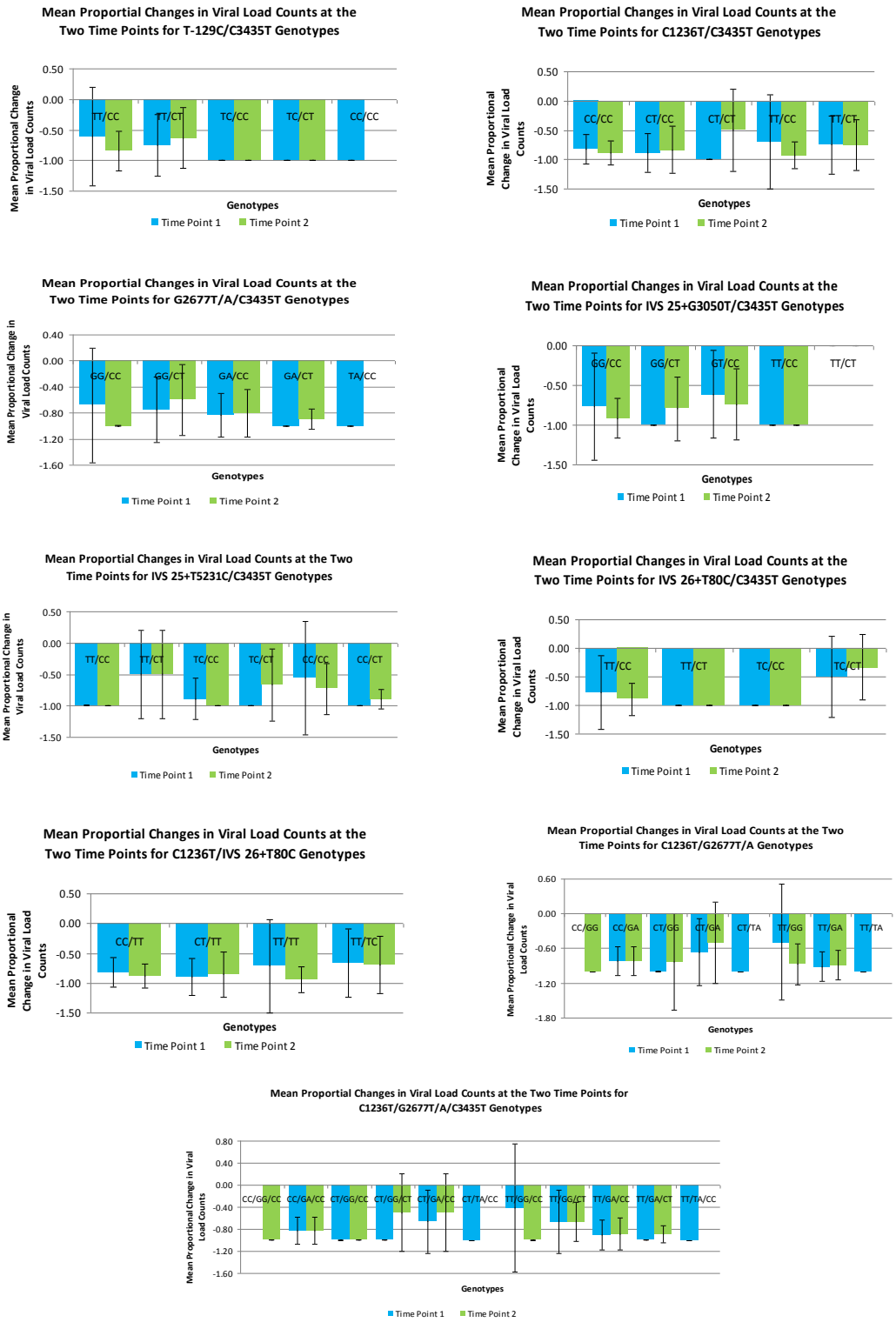


Figure 30: Mean proportional decrease in viral load count for each genotype combination for each polymorphism combination at time points one and two.

The was performed to determine any association between genotypes and the mean proportional increase in $CD4^+$ count at each time point, instead of the mean values for the absolute difference between baseline and the two time points (Table 14). The P values for all SNPs at both time points were above the significance value of $\alpha = 0.05$, except for the T-129C SNP, where the P value for the first time point was determined to be 0.0029. There was no association detected for any of the SNP combinations using the proportional increases. A student's t test showed that there was a table-wide significant difference between the p values from the Kruskal-Wallis test ($P = 0.0053$).The Kruskal-Wallis test was repeated without the outlying samples and no significant difference in P values was found.

Table 14: Results of the Kruskal-Wallis test for association between the SNPs and the mean proportional increase in viral load count at the first and second time points.

| **Variables** | **DF** | **$\chi^2$ value** | **p value** |
|---|---|---|---|
| T-129C; Time1 | 2 | 11.6661 | 0.0029 |
| T-129C; Time2 | 1 | 1.4913 | 0.2220 |
| C1236T; Time1 | 2 | 0.7068 | 0.7023 |
| C1236T; Time2 | 2 | 0.7726 | 0.6796 |
| G2677T/A; Time1 | 2 | 2.7645 | 0.2510 |
| G2677T/A; Time2 | 1 | 0.7271 | 0.3938 |
| IVS 25+G3050T; Time1 | 2 | 0.3899 | 0.8229 |
| IVS 25+G3050T; Time2 | 2 | 0.6465 | 0.7238 |
| IVS 25+T5231C; Time1 | 2 | 1.3316 | 0.5139 |
| IVS 25+T5231C; Time2 | 2 | 0.6395 | 0.7263 |
| C3435T; Time1 | 1 | 0.0631 | 0.8017 |
| C3435T; Time2 | 1 | 0.5136 | 0.4736 |
| IVS 26+T80C; Time1 | 1 | 0.0353 | 0.8509 |
| IVS 26+T80C; Time2 | 1 | 2.5296 | 0.1117 |

The mean RQ values corresponding to each genotype of each SNP were calculuated (Figure 31) as were the mean RQ values for the SNP combinations (Figure 32). As there were only 28 samples that could be used for cDNA synthesis and subsequent quantitative PCR, not all genotypes were present in those samples.

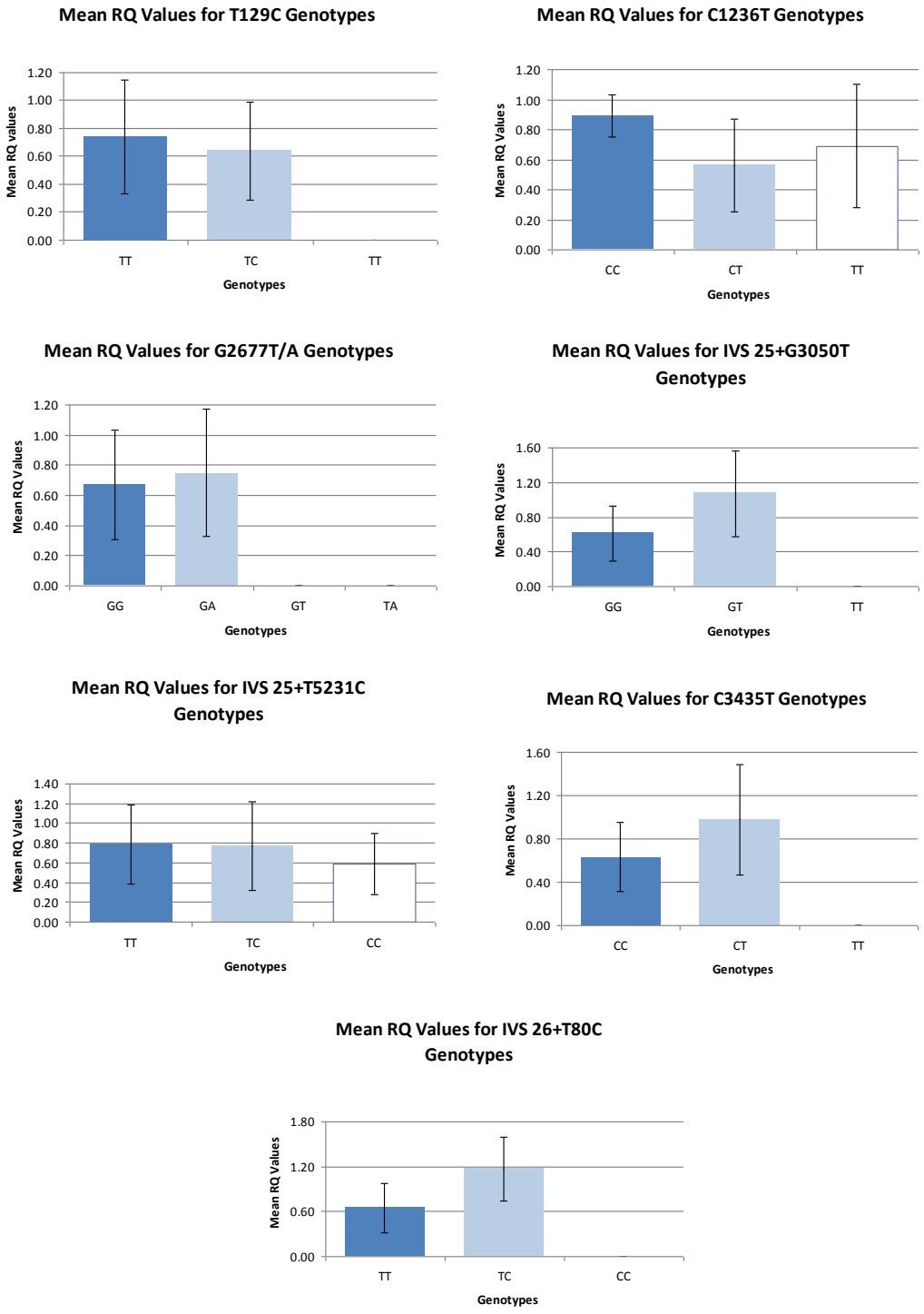


Figure 31: Mean RQ values for each genotype of each of the seven polymorphisms.

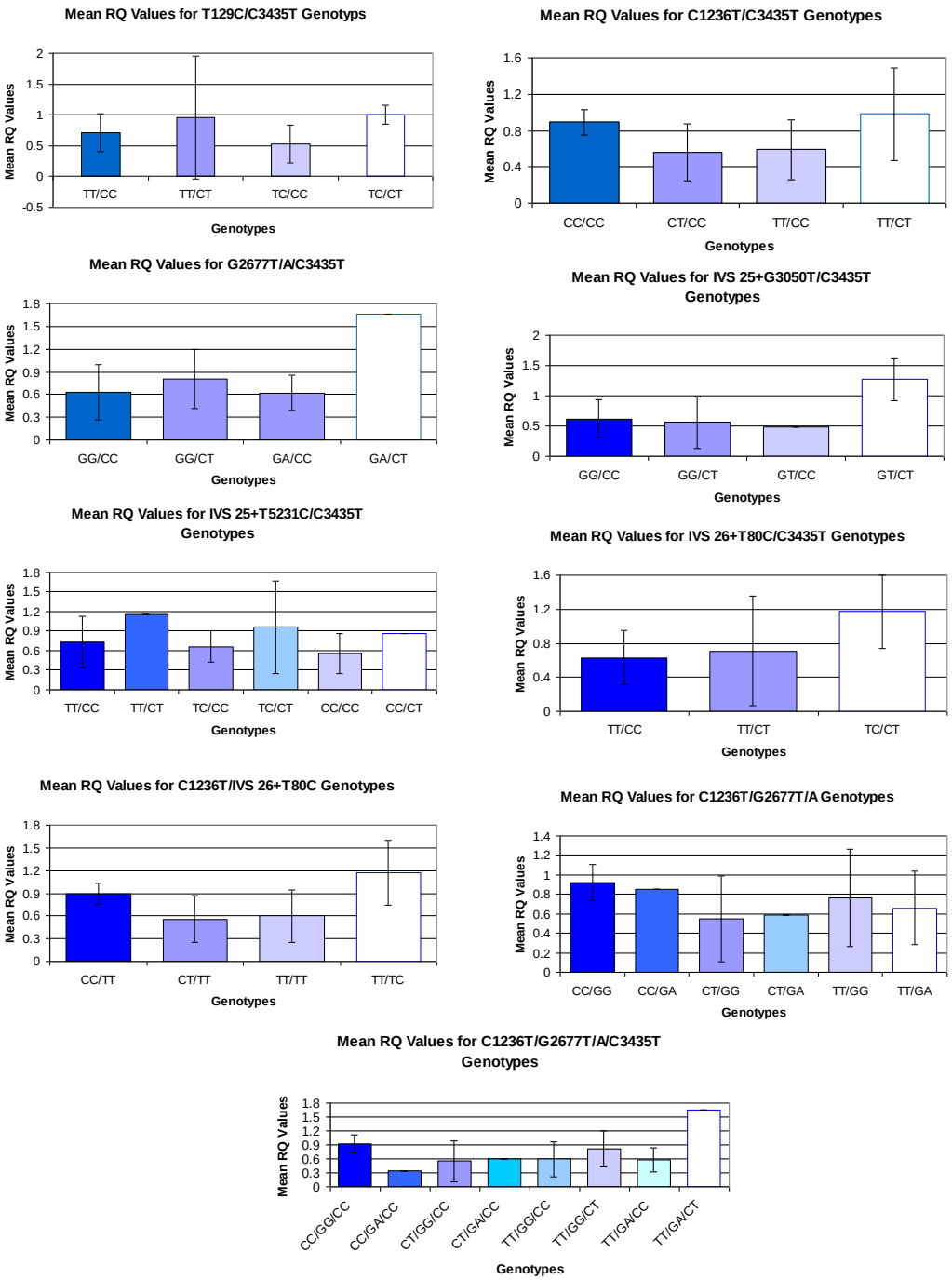


Figure 32: Mean RQ values for each genotype combination for each of the polymorphism combinations.

The Kruskal-Wallis test showed that there are no significant associations for the test for association between the SNPs T-129C, C1236T, G2677T/A and IVS 25+T5231C and the mean RQ values, the P values were above the significance value of 0.05. The Kruskal-Wallis test for association between the two genotypes found for each of the SNPs IVS 25+G3050T, C3435T and IVS 26+T80C and the mean RQ values, the P values were below the significance value of $\alpha = 0.05$ (Table 15). A student's t test showed that there was a table-wide significant difference between the p values from the Kruskal-Wallis test ($p < 0.0001$).

Table 15: Results of the Kruskal-Wallis test for association between the genotypes of the SNPs and the RQ values.

| Variables | DF | $\chi^2$ | p-value |
|---|---|---|---|
| T-129C; RQ | 1 | 0.0067 | 0.9349 |
| C1236T; RQ | 2 | 1.7911 | 0.4084 |
| G2677T/A; RQ | 1 | 0.1666 | 0.6831 |
| IVS 25+G3050T; RQ | 1 | 5.0570 | 0.0245 |
| IVS 25+T5231C; RQ | 2 | 0.8377 | 0.6578 |
| C3435T; RQ | 1 | 3.8890 | 0.0486 |
| IVS 26+T80C; RQ | 1 | 5.1077 | 0.0238 |

# CHAPTER 4: DISCUSSION

## Regulation of *ABCB1* expression

Sequencing of both upstream non-coding reagions resulted in the detecteion of novel variation and only some of the known Polymorphisms. This was in line with our expectations of finding different variation in the South African population, based on previous studies in other African populations (Tishkoff et al. 2009). Of the two well studied SNPs in the upstream region of *ABCB1*, only the T-129C polymorphism was confirmed by sequencing, and later genotyped, whereas the A-41G was not detected as polymorphic in this population. The presence of known polymorphisms at different allele frequencies, novel variation and lack of variation at some known polymorphic sites confirms that African populations have a different level of variation compared to other populations. The undetected polymorphisms may be present in the population, but at very low frequencies.

From the bioinformatic analysis it is clear that variation in the upstream region can alter the structure of the regulatory elements. In the region upstream of exon 1 no promoter was predicted and a transcriptional start site is not likely in either the forward or reverse strands. The -137G allele creates a Lymphokine CS motif and an AML-1a transcription factor binding site, while the -223C allele creates a CdxA transcription factor binding site.

In the region upstream of exon 3 there was marginal promoter prediction from PROMOTER 2.0, which could be due to CpG islands in the upstream region. There is a strong correlation between CpG islands and vertebrate promoters (Ohler et al. 2000; Ohler et al 2001) and the region upstream of exon 3 may contain a number of CpG islands (Takane et al 2004). PROSCAN or FPROM did not predict any promoters. The 153A allele disrupts both an alpha-1-protinase inhibitor motif and an *ABCB1* motif. The 180G allele creates a P-LAP motif and changes an AML-1a transcription factor binding site to a MZF binding site. The -60T allele changes a CdxA transcription factor binding site to an Nkx-2 binding site.

The impact of the novel variation and changes to the motifs and transcription factor binding sites in the upstream regions is not clear, as not much is known about the regulation of expression of the gene, but the variation may result in altered transcription. The transcription factors regulating *ABCB1* transcription have not been well studied and the removal or addition of binding sites could impact gene regulation. It is unclear at this point what impact the creation of new transcription factor binding sites is likely to have on the regulation of the gene and it is also unclear if any of the detected promoter elements influence the regulation of transcription of this gene and what impact the changes these SNPs may have on that regulation. Functional studies need to be conducted to determine the effects, if any, of the detected variation in the two upstream regions of the *ABCB1* gene.

**Genetic Structure in the *ABCB1* gene**

The results have shown very different allele frequencies in the southern African Bantu speaking population for all but the IVS 25+T5231C polymorphism (T allele = 0.51), which has a similar allele frequency to that seen in Caucasians (T allele = 0.58) (Soranzo et al. 2004). The increased frequency of the 1236T allele (0.70 in this study, compared to 0.47 in the European study by Soranzo et al. 2004) is obvious in the analysis of the C1236T-G2677T/A-C3435T haplotype, where the TGC phase has the highest frequency, as opposed to the CGC phase, which was expected to have the highest frequency based on allele frequencies in the West African and African American populations (Kim et al. 2001; Tang et al. 2002; Allabi et al. 2005). The impact of the higher 1236T allele is not known, as all other populations show the 1236C allele to be at the highest frequency of the two.

The 3435C allele was found at a much higher frequency (0.88) than in the European populations (0.42, Soranzo et al. 2004), which we expected due to studies done previously on West Africans (Allabi et al. 2005). The high frequency of the 3435 C allele in Africans has been suggested to be due to selection for this allele, as the high prevalence of this allele corresponds with a lower prevalence in inflammatory bowel disorders (Schaeffeler et al. 2001). No alternative suggestions or explanations have been found.

Studies by Hoffmeyer et al. (2000) and Kimchi-Sarfaty et al. (2007) showed that the silent 3435 C to T polymorphism could change the folding and function of the P-glycoprotein. Silent single nucleotide changes could have an effect on the mRNA structure, the protein folding pathway, protein translation and protein structure (Fung and Gottesman 2009).

The differences in allele frequancies are the direct cause of the very low levels of linkage disequilibrium and different haplotypes and haplotype frequencies for this gene in the southern African population. Soranzo et al. 2004 published linkage disequilibrium values for a number of SNPs paired with C3435T for a European population. When we compared these with the linkage disequilibrium values for the same pairs in our population, the $|D'|$ and $r^2$ values were much lower. The lower values in the South African population were expected, but the very high values in the European population studied by Soranzo et al. (2004) could be due to a small sample size (48 different samples), or the generally higher levels of linkage disequilibrium seen in other populations compared to African populations. Notably there is virtually no linkage disequilibrium between the G2677T/A locus and either of the C3435T or IVS 25+T80C loci, which have almost complete linkage in the Caucasian population depicted in the data by Soranzo et al. (2004).

It is difficult to compare LD between sites or between populations, as there is no perfect measure of LD (Lewontin, 1988) and factors such as population size and allele frequency may influence LD measures (Hendrick, 1987; Lewontin, 1988).

Our values for $|D'|$ and $r^2$ are, however, greatly different and much lower than the values reported for other populations. It is not only population processes that have led to the low LD in this gene.

Variation in linkage disequilibrium estimates and allele frequencies in different populations is due to a number of factors, including genetic drift, migration, population structure and natural selection (Ardlie et al. 2002; Shifman et al. 2003). A number of linkage disequilibrium measures are available, but caution needs to be exercised when using a particular measure. Two commonly used measures are $|D'|$ and $r^2$ ($\Delta^2$). The range of $|D'|$ is independent of the allele frequencies, whereas all other measures are dependant on allele frequency (Hendrick 1987; Lewontin 1988).

The measure $|D'|$ is preferred as it can take a maximum value of 1 if there is complete linkage disequilibrium, meaning that any one allele is only seen with one allele at the other locus and only three of the possible four haplotypes are seen (Teare et al. 2002), however, $|D'|$ is dependant on sample size, with a small increase in low $|D'|$ values seen as the sample size decreases (Teare et al. 2002).

The measure $r^2$ can reach a maximum of 1 when there is complete linkage disequilibrium, and when the two polymorphisms have equal allele frequencies. Intermediate values of $r^2$ can be interpreted, but only when the two polymorphisms have equal allele frequencies (Devlin and Risch 1995; Jorde 2000; Reich et al. 2001; Ardlie et al. 2002; Teare et al. 2002). There is no perfect

measure of linkage disequilibrium, making it difficult to compare populations with different allele frequencies (Lewontin 1988). African populations show especially distinct differences as opposed to non-African populations, as they have more haplotypes and lower levels of linkage disequilibrium between alleles (Jorde et al. 1998; Reich et al. 2001, Tishkoff and Williams 2002; Tishkoff et al. 2009).

The genotypic data was used for the haplotype analysis in PHASE. There is much debate as to the reproducibility of the output obtained from any haplotype analysis software. There are many haplotype inferral programs available that often result in multiple solutions and haplotype inferrals are not usually assessed (Orzack et al. 2003). The programs are usually written using data from populations with European ancestry and may not be relevant to non-European populations. The programs are also unable to deal with heterozygotes and low levels of linkage disequilibrium (Orzack et al. 2003). For this reason, ten runs were performed independently of each other, and analysed to determine the reproducibility of the outputs. All ten outputs showed the exact same data, validating the output for further analysis. The majority of the ten most common haplotyes contained the 1236T, 2677G, 3435C and IVS 26+80T alleles, which is not surpising, as these four alleles occur at very high frequencies (0.7, 0.77, 0.88 and 0.89, respectively).

The most common phases seen in the ten most common haplotypes are the IVS 25+G3050T-C3435T-IVS 26+T80C GCT phase, which occurs in eight out of the

top ten, and the G2677T/A-C3435T- IVS 26+T80C GCT phase, which is present in seven of the top ten. None of the other loci show any pattern of occurrence with other loci.

The most commonly studied haplotype in *ABCB1* is C1236T-G2677T/A-C3435T. The frequencies of the different phases were compared with those found in a number of other studies. The TTT phase is expected to be the most common in non-African populations and the lowest in African populations, which was confirmed by our data, whereas the CGC phase is expected to be the most common in African populations and the least common in non-African populations. From our data is seen that the CGC phase frequency is much lower than in any other populations. Based on studies on West African and African American populations, we expected the frequency of this phase to higher than the observed frequency. The reason for the lower frequency of this phase and the higher frequency of the TGC phase is the difference in the C1236T allele frequencies, where the 1236T allele is found at a much higher frequency then any other population.

The TGC phase occurs at the highest frequency in our data (~ 43%) with all other populations showing a very low frequency of this phase with the next highest frequency seen in Chinese (~ 18%). For the remaining phases of this haplotype the frequencies are very low or they were not found, except for the TAC phase, which occurred at a frequency of ~ 15%.

In their 2009 review article, Fung & Gottesman looked at the possible accumulative effects that the C1236T, G2677T/A and C3435T polymorphisms might have on protein function. The 3435C>T mutation may affect the co-translational folding of certain proximal amino acids, which are located in the second ATP-binding domain and may reduce the ATP-binding affinity of P-glycoprotein (Loo and Clarke 1995; Muller et al. 1996; Szakacs et al. 2000; Urbatsch et al. 2000; Urbatsch et al. 2001). The affect of the 1236C>T mutation is also synonymous and does not affect mRNA stability, protein expression or protein function (Wang et al. 2005). It is possible that this site may also affect the co-translational folding of certain proximal amino acids, which are located in the first ABC domain, and may affect ATP-binding and ATP hydrolysis (Fung and Gottesman 2009; Kim et al. 2006; Sakurai et al. 2007).

The 2677G>T mutation has been suggested to impact drug transport and possibly affect drug-induced ATPase activity (Fung and Gottesman 2009; Sakurai et al. 2007). Fung & Gottesman (2009) suggest that all three mutations may each only have a small impact on the function of the protein, but added together as a haplotype, they may be additive and have a greater effect on protein folding and subsequent function. The TTT haplotype is found at a frequency of about 40% in Caucasians, but is very rare in our population. This again suggests that any advancement made in this field of *ABCB1* research is irrelevant to the southern African and probably other African populations as well.

The high frequency of the TGC phase is very likely to be due to selection for this phase in our population, but this finding is not unexpected, as the allele frequencies at these three loci are quite different from other populations. The 1236C allele in our population is present at a frequency of 30%, compared to the 53% in Caucasians. This results in this allele being the minor allele in our population but the major allele in Caucasians. Similarly the 3435T allele is present at a frequency of 12% in our population, but at 58% in Caucasian. This allele is also the minor allele in our population but the major allele in Caucasians.

The frequency of the 2677G allele is higher in our population (77%) than in the Caucasian population (53%). These differences in the allele frequencies manifest themselves in the differences in the frequencies of the phases seen for this haplotype in our analysis. The levels of linkage disequilibrium between these three sites are also different from the Caucasian population. Between C1236T and G2677T/A and between G2677T/A and C3435T our data shows lower levels of linkage compared to the Caucasian population. Between the C1236T and C3435T loci the linkage seen in our data is slightly higher than that in the Caucasian population. These differences may also impact the frequency of the phases of this haplotype.

Laboratory-based or manual haplotyping is expensive and time consuming. The PHASE program is regarded to be quite accurate, based on studies using non-African populations but computationally expensive. Haplotypes determined by manual haplotyping are more informative than genotypes, but inferred haplotypes

are often less informative than genotypes due to possible uncertainties in phasing calculations and ambiguous haplotypes (Malhi 2002; Orzack et al 2003; Balding 2006). Our haplotype analysis produced a large number of ambiguous haplotypes for the samples, as a result of the reduced levels of linkage disequilibrium seen in the population. Manual haplotyping would be an effective method of determining the accuracy of the haplotype predictions in the Southern African population and to resolve the ambiguous haplotypes. The size of the *ABCB1* gene would be problematic, but the closer SNPs could be used as a starting point.

The most common haplotypes could have originated only through either multiple mutations or recombination between SNPs. Other groups have investigated haplotype evolution (Kroetz et al. 2003; Tang et al. 2004) and found some evidence for recombination using other SNPs and other populations.

Reduced median networks were constructed for all seven sites, as well as the six sites without the T-129C site, as this one is the most distant. For both networks it is clear that multiple mutations and recombination events have occurred to produce the haplotype phases found in the southern African population. The squares in the network that are produced between different phases show that multiple single mutations or recombination events took place. In the network constructed for the seven polymorphisms the multiple single mutations recombinations involve the T-129C, C1236T, G2677T/A, IVS 25+G3050T and IVS 25+T5231C sites, with the IVS 25+T5231C polymorphisms being the most common variant. Only one phase involved single mutations or recombinations of

both the C3435T and IVS 26+T80C polymorphisms. This is also indicative of the two sites being linked.

In the network constructed for only six polymorphisms there are multiple single mutations and recombination events involving all six sites. The IVS 25+G3050T and IVS 25+T5231C polymorphisms are the most common variants between the phases. Both networks confirm that a single functional site could not be extrapolated based on haplotypes, due to the presence of multiple mutations and recombination events in the southern African population.

The genotyping data from the general population samples and those from HIV positive individuals showed no significant deviation, so no comparison was possible for the genotype or haplotype data. The samples were limited to the number of volunteers who were prepared to donate blood, and although the study was limited to Bantu-speaking South Africans living in the Johannesburg area, the study could be expanded to a much larger sample size and the individual language groups studied.

**Quantitative mRNA levels**

The use of real-time for quantifying mRNA expression levels has not always been reproducible from one paper to another (Bustin 2002), as there are a number of aspects that can change from one experiment to another. Human error can cause variability due to the handling of reagents and the amounts of reagents

added for the reaction. The endogenous controls have been re-examined and a number of genes used as endogenous controls (e.g. GAPDH) for their consistant expression, have been found to be expressed differently in some tissues or under certain conditions. Suggestions have been made to include the use of multiple endogenous controls for the samples to be normalized against to ensure more consistant and believable results (Schmittgen and Zakrajsek 2000; Bustin 2002).

The quantitative PCR data showed variable relative amounts of mRNA compared to the endogenous control. The t-test showed no significant difference in the RQ values in the 28 samples, meaning that no particular sample (or samples) have significantly higher or lower RQ values than the other samples. Quantitative PCR is used to quantify the steady state mRNA levels only, and is not an indication of the levels of transcription or mRNA stability. The RNA levels cannot be used to infer information about the protein production in the cell (Bustin 2002).

**Association studies**

Population association studies are useful for identifying patterns of differences in polymorphisms in individuals with different disease states. These patterns could then be used to identify alleles that enhance the risk of the disease or alleles that could be effective in the prevention or treatment of the disease. The problem with this lies in the fact that the human genome is so large, that any patterns that are suggestive of causal polymorphisms may have arisen by chance and not be selected for (Balding 2006; Manolio et al. 2009).

Fellay et al. (2002) found association between the 3435 TT genotype and higher $CD4^+$ cell counts and a better decrease in viral load after six months, but Nasi et al. (2003) found that the C3435T polymorphism did not have any effect on response to therapy over the same time intervale. A meta-anlysis by Leschziner et al. (2006) found no association between *ABCB1* polymorphisms and drug reactions.

The association studies aimed to establish if there was significant association between any particular genotype of any polymorphism (or combination of polymorphisms) and either relative mRNA values, an increase in $CD4^+$ or a decrease in viral load. For the mean viral load data, there was a significant difference (P = 0.0121) between the mean viral load counts at time point two for the G2677T/A polymorphism, with the 2677GA genotype (52 samples) showing a significantly higher mean viral load count than the 2677GG genotype (57 samples). An association was found between the decrease in viral load and the T-129C SNP, as the P value was below 0.05 for the first time point and the TT genotype showed the lowest mean proportional decrease and was also present at the highest frequency of the T-129C genotypes.

For the IVS 25+G3050T, C3435T and IVS 26+T80C SNPs the P values were below 0.05 for the association between the SNP genotypes and the mean RQ values. For all three SNPs the heterozygotes showed higher mean RQ values than the major allele homozygote, and the minor allele homozygote was not present in

any of the samples. This could indicate that the major allele homozygotes are associated with lower relative mRNA levels than the heterozygotes.

The number of samples (twenty-eight) and number of genotypes present for the study (none of the minor alleles homozygotes were found for comparison), suggests that not enough data was available for any association seen to be truely meaningful. A larger number of samples with a better representation of genotypes would be required to determine any true association. Table-wide significance studies showed significant differences between the P values in all the association studies tables.

The lack of association could also be due to the fact that the patients, that were sampled, were primarily on efavirenz-based ART. Efavirenz is not a substrate of P-gp and the association between the ART responses of individuals taking efavirenz has not been substancially proven. The lack of association found in the study could be an indication that there may, in fact, be no association between the effect of efavirenz and variation in *ABCB1* or P-gp.

Another fact to consider is the adherence issue of patients receiving ART in South Africa. Many patients are embarrassed to admit to their families and friends that they are HIV-positive and will not take their medication when they fear detection by family and friends, resulting in irregular adherence to their treatment. Patients do not always stick to their regular appointment dates and a high deviation exists between the dates of their clinical appointment and the date

they visit the clinic. After their visit to determine the baseline viral load and $CD4^+$ counts, patients are meant to have these measured every 6 months, but due to the irregularity of the visit to the clinic, the first two measurements are named time point one and two. This could also have affected the data for association studies. A larger sample size, with individuals that adhere to their medication and visit the clinic regularly, and protein expression data would be required to gain a better understanding of the effects of the polymorphisms and any possible associations.

Genome wide association studies (GWAS) have become popular for identifying common causal variants throughout the human genome and inferring other genotypes based on the identified variables. This requires more than 30,000 well-chosen polymorphisms, relies on the presence of high levels of linkage disequilibrium within the genome, for the inferral of genetic structure for association studies (Balding 2006, Manolio et al. 2009; Need and Goldstein 2009). GWAS aim to provide data that can be used to predict disease risk, to improve the understanding of complex diseases for improved treatment and to identify subclasses of similar diseases (Gibson and Goldstein 2007).GWAS would allow screening of the entire genome to identify the major variation conferring susceptibility on common diseases, for clinical advances and personalizing medicine, and is thought to be the potential scientific realization of major projects such as the Human Genome and the HapMap projects (Gibson and Goldstein 2007; McCarthy et al. 2008).

GWAS have their limitations, as the majority of studies have been conducted in populations of European ancestry, with no focus on populations of recent African ancestry, where genetic variation is the greatest and the linkage disequilibrium is the lowest (McCarthy et al. 2008; Manolio et al. 2009).

Our analysis of the genetic structure of the *ABCB1* gene has shown that the variation in the southern African population is different from that of other populations, even compared to the West African populations, and has very low levels of linkage disequilibrium, making the use of genome wide association studies impractical or impossible if this is a valid indication of the entire genome. Genetic studies will have to rely on the manual characterisation of the genes by genotyping or possibly by whole-genome sequencing and cannot rely on data obtained from studies on other populations. For GWAS to be successful there needs to be extensive research into the genetic structure and linkage disequilibrium of different populations, as well as admixed populations, where different patterns of linkage disequilibrium are also ovserved (McCarthy et al. 2008; Price et al. 2008). Due to the small sample size available for the association studies, it seems unlikely that any of the polymorphisms are associated with immune response at either of the time points and no association was found between the C3435T polymorphism and immune response, indicating that the *ABCB1* gene may not be involved in immune response with antiretroviral therapy. More functional studies are required to add to the information that will determine the molecular mechanism behind the variation in the *ABCB1* gene.

There is some debate as to the usefulness of studying a particular population, as ‘race’ and ‘ethnicity’ are largely social categories and generalisations made about a particular population could be detrimental (Mountain & Risch 2004; Sankar et al. 2007). It has been suggested that research be conducted with a focus on individuals rather than on a population (Feldman et al. 2003). The genomes of two southern African individuals (Archbishop Desmond Tutu, a South African Bantu-speaker, and !Gubi, a Namibian hunter-gatherer) have recently been sequences, confirming the different genetic structure and highlighting the need to study more southern African genomes (Ledford 2010). We feel that there is enough evidence to show that the southern African population is sufficiently genetically different from other populations to warrant studying their variation (Jorde et al. 1998; Reich et al. 2001; Tishkoff and Williams 2002). Due to this difference in variation, also shown in our data, functional studies performed in other populations may not be applicable to this population. Also, due to the high prevalence of HIV and poverty in sub-Saharan Africa studies based on the individual may not be financially or logistically possible.

# CHAPTER 5: CONCLUSIONS

This study is the beginning of the characterization of the variation in the *ABCB1* in the black South African population, which has again proven that variation in this population is different compared to non-African populations, and that there is an importance in characterizing variation in this population. A number of novel SNPs and some known SNPS were detected in the two regulatory regions sequences cause changes to regulatory elements found in those regions, but the impact of these changes are not known as the promoter regions are well characterized and the regulation of the *ABCB1* gene is not well understood.

The genotypic data obtained from seven SNPs, thought to of some level of functional importantance, clearly shows that variation in the *ABCB1* gene in the southern African population is very different from that in other populations, even when compared to other African populations. This affects the allele frequencies, the linkage disequilibrium values as well as the frequency of haplotypes and the frequency of haplotype phases. The genetic structural differences warrent larger studies to characterize variation in this population, to improve to healthcare in this population.

There may be some association between the -129 TT genotype and a lower mean proportional decrease in viral load counts at a later stage of treatment, and there may be an association between each of the major allele homozygotes of the IVS 25+G3050T, C3435T and IVS 26+T80C polymorphisms with lower mean

relative mRNA levels. Due to the small sample size it is unlikely that any true association exists between the *ABCB1* polymorphisms and immune response, but a larger sample size would be required to confirm this. Work is still needed to determine the means by which variation in this gene results in the altered presence or function of P-glycoprotein.

The differences in the genetic structure of the gene indicate that genome wide association studies and other functional studies based on other populations will not be applicable to the southern African population. This population is so genetically diverse, that genome wide association studies may not be possible in this population, due to both the amount of variation between this population and other populations, as well as within the population itself. As a result, population-based determination of an individual's likely response to certain drugs, as well as population- or individual-based drug therapy may not be logistically or financially possible in the southern African population.

No significant association was found between variation in *ABCB1* and immune recovery with HIV therapy. A larger study, both in sample number and in the number of sites characterized, combined with functional studies is needed to better understand the mechanisms by which variation in *ABCB1* affect the functional variation of P-glycoprotein, as well as the mechanisms by which variation in P-glycoprotein are associated with an individuals response to antiretroviral therapy.

## REFERENCES


Akiyama Y.: "TFSEARCH: Searching Transcription Factor Binding Sites" (http://www.rwcp.or.jp/papia/)

Allabi, A. C., Horsmans, Y., Issaoui, B., and Gala, J.-L. 2005. Single nucleotide polymorphisms of ABCB1 (*MDR1*) gene and distinct haplotype in a West Black African population. *Eur. J. Clin. Pharmacol.* **61:** 97-102.

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

Balding, D. J., 2006. A tutorial on statistical methods for population association studies. *Nat. Rev. Genet.* **7:** 781-791.

Bandelt, H. J., Forster, P., Sykes, B. C., and Richards, M. B. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* **141:** 743-753.

Berruet, N., Sentenac, S., Auchere, D., Gimenez, F., Farinotti, R., and Fernandez, C. 2005. Effect of efavirenz on intestinal p-glycoprotein and hepatic p450 function in rats. *J. Pharm. Pharmaceut Sci.* **8**(2): 226-234.

Beleza, S., Gusmão, L., Amorim, A., Carracedo, A., and Salas, A. 2005. The genetic legacy of western Bantu migrations. *Hum. Genet.* **117:** 366-375.

Bleiber, G., May, M., Suarez, C., Martinez, R., Marzolini, C., Egger, M., and Telenti, A. 2004. *MDR1* genetic polymorphism does not modify either cell permissiveness to HIV-1 or disease progression before treatment. *J. Infect. Dis.* **189:** 583-596.

Bodor, M., Kelly, E. J., and Ho, R. J. 2005. Characterization of the human *MDR1* gene. *AAPS* **7**(1): E1-E5.

Brumme, Z. L., Dong, W. W. Y., Chan, K. J., Hogg, R. S., Montaner, J. S. G., O'Shaughnessy, M. V., and Harrigan, P. R. 2003. Influence of polymorphisms within $CX_3CR1$ and MDR1 genes on initial antiretroviral therapy response. *AIDS* **17**(2): 210-208.

Bustin, S. A., 2002. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J. Mol. Endo.* **29:** 23-39.

Cargill, M., Altshuler, D., Ireland, J., Sklar, P., Ardlie, K., Patil, N., Shaw, N., Lane, C. R., Lim, E. P., Kalyanaraman, N., Nemesh, J., Ziaugra, L., Friedland, L., Rolfe, A., Warrington, J., Lipshutz, R., Daley, G. Q., and Lander, E. S. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22:** 231-238.

Collins, A., Lonjou, C., and Morton, N. E. 1999. Genetic epidemiology of single-nucleotide polymorphism. *PNAS* **96**(26): 15173-15177.

Cornwell, M. M. & Smith, D. E. (1993) Sp1 activates the MDR1 promoter through one of two distinct G-rich regions that modulate promoter activity. *J. Biol. Chem.* **268**(26), 19505-19511.

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

Dickinson, L., Robinson, L., Tjia, J., Khoo, S., and Back, D. 2005. Simultaneous determination of HIV protease inhibitors amprenavir, atazanavir, indinavir, lopinavir, nelfinavir, ritonavir and saquinavir in human plasma by high-performance liquid chromatography-tandem mass spectrometry. *CHROMB*-14176.

Ding, K., Zhou, K., He, F., and Shen, Y. 2003. LDA- a java-based linkage disequilibrium Analyzer. *Bioinformatics* **19**(16): 2147-2148.

Emigh, T. H. 1980. A comparison of tests for Hardy-Weinberg equilibrium. *Biometrics* **36**(4): 627-642.

Fellay, J., Marzolini, C., Meaden, E. R., Back, D. J., Buclin, T., Chave, J.-P., Decosterd, L. A., Furrer, H., Opravil, M., Pantaleo, G., Retelska, D., Ruiz, L., Shinkel, A. H., Vernazza, P. C., Eap, B., And Telenti, A. 2002. Response to antiretroviral therapy in HIV-1 infected individuals with allelic variants of the multidrug resistance transporter 1: A pharmacogenetics study. *Lancet* **359:** 30-36.

Flexner, C. 1998. HIV-1 protease inhibitors. *New Eng. J. Med.* **338**(18): 1281-1292.

Förster, V. T. 1948. Zwischenmolekulare Energienwanderung und Fluoreszenz. *Annals of Physics* **2:** 55-75.

Fung, K. L. and Gottesman, M. M. 2009. A synonymous polymorphisms in a common MDR1 (ABCB1) haplotype shapes protein function. *BBA* **1794:** 860-871.

Gibson, G., and Goldstein, D. B. 2007. Human genetics: The hidden text of genome-wide associations. *Curr. Biol.* **17**(21): R929-932.

Ginzinger, D. G. 2002. Gene quantification using real-time quantitative PCR: An emerging technology hits the mainstream. *Exp. Haematol.* **30**: 503-512.

Goldsmith, M. E., Madden, M. J., Morrow, C. S. & Cowan, K. H. (1993) A Y-box consensus sequence is required for basal expression of the human multidrug resistance (mdr1) gene. *J. Biol. Chem.* **268**(8): 5856-5860.

Greenberg, J. H. 1963. The languages of Africa. Bloomington: Indiana University Press.

Guthrie, M. 1962. Some developments in the prehistory of the Bantu origins. *J. Afr. Hist.* **3:** 273-282.

Gwee, P.-C., Tang, K., Chua, J. M. Z., Lee, E. J. D., Chong, S. S., and Lee, C. G. L. 2003. Simultaneous genotyping of seven single-nucleotide polymorphisms in the *MDR1* gene by single-tube multiplex minisequencing. *Clin. Chem.* **49**(4): 672-676.

Haas, D. W., Wu, H., Li, H., Bosch, R. J., Lederman, M. M., Kuritzkes, D., Landay, A., Connick, E., Benson, C., Wilkinson, G. R., Kessler, H., and Kim, R. B. 2003. MDR1 gene polymorphisms and phase 1 viral decay during HIV-1 infection: An adult AIDS clinical trial group study. *J. AIDS* **34:** 295-298.

Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A. E., Kel, O. V., Ignatieva, E. V., Ananko, E. A., Podkolodnaya, O. A., Kolpakov, F. A., Podkolodny N. L. & Kolchanov, N. A. 1998. Databases on Transcriptional Regulation: TRANSFAC, TRRD, and COMPEL. *Nucleic Acids Res.* **26,** 364-370.

Hendrick, P. W. 1987. Gametic disequilibrium measures: Proceed with caution. *Genetics* **117,** 331-341.

Henry, K., Erice, A., and Tierney, C., Balfour, H. H. Jr., Fischl, M. A., Kmack, A., Liou, S.-H., Kenton, A., Hirsch, M. S., Phair, J., Martinez, A., and Kahn, J. O. 1998. A randomized, controlled, double-blind study comparing the survival benefits of four different reverse transcriptase inhibitor therapies (three-drug, two-drug and alternating drug) for the treatment of advanced AIDS. *J. AIDS Hum. Retroviral.* **19:**339-349.

Hill, W. G., and Robertson, A. 1968. Linkage disequilibrium in finite populations. *TAG* **38**(6): 226-231.

Hoffmeyer, S., Burk, O., Von Richter, O., Arnold, H. P., Brockmöller, J., Johne, A., Cascorbi, I., Gerloff, T., Roots, I., Eichelbaum, M., and Brinkmann, U. 2000. Functional polymorphism of human multidrug resistance gene: Multiple sequence variations and correlation of one allele with p-glycoprotein expression and activity *in vivo*. *PNAS* **97**(7): 3473-3478.

Horinouchi, M., Sakaeda, T., Nakamura, T., Morita, Y., Tamura, T., Aoyama, N., Kasuga, M. And Okumura, K. 2002. Significant genetic linkage of MDR1 polymorphisms at positions 3435 and 2677: Functional relevance to pharmacogenetics of digoxin. *Pharmaceut. Res.* **19**(10), 1581-1585.

Hulgan, T., Donahue, J. P., Hawkins, C., Unutmaz, D., D'aquila, R. T., Raffanti, S., Nicotera, F., Rebeiro, P., Erdem, H., Rueff, M., and Haas, D. W. 2003. Implications of T-cell p-glycoprotein activity during HIV-1 infection and its therapy. *J. AIDS* **34**(2): 119-126.

Javitt, G. H. and Hudson, K. 2007. The right prescription for personalized genetic medicine. *Future Med.* **4**(2): 115-118.

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

Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E., Seielstad, M.T., and Batzer, M. A. 2000. The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. *Am. J. Hum. Genet.* **66:** 979-988.

Kim, L. W., Peng, X. H., Sauna, Z. E., FitzGerald, P. C., Xia, D., Muller, M., Nandigama, K. and Ambudkar, S. V. 2006. The conserved tyrosine residues 401 and 1044 in ATP sites of human P-glycoprotein are critical for ATP binding and hydrolysis: evidence for a conserved subdomain, the A-loop in the ATP-binding cassette. *Biochemistry* **45:** 7605-7616.

Kim, R. B., Leake, B. F., Choo, E. F., Dresser, G. K., Kubba, S. V., Schwarz, U. I., Taylor, A., Xie, H.-G., Mckinsey, J., Zhou, S., Lan, L.-B., Schuetz, J. D., Schuetz, E. G., and Wilkinson, G. R. 2001. Identification of functionally variant *MDR1* alleles among European Americans and African Americans. *Clin. Pharmacol. Ther*. **70:**189-199.

Kimchi-Sarfaty, C., Oh, J. M., Kim, I.-W., Sauna, Z. E., Calcagno, A. M., Ambudkar, S. V. & Gottesman, M. M. 2007. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315,** 525-528.

Knight, J. C. 2005. Regulatory polymorphisms underlying complex disease traits. *J. Mol. Med.* **82:** 97-109.

Kroetz, D. L., Pauli-Magnus, C., Hodges, L. M., Huang, C. C., Kawamoto, M., Johns, S. J., Stryke, D., Ferrin, T. E., Deyoung, J., Taylor, T., Carlson, E. J., Herskowitz, I., Giacomini, K. M., and Clarke, A. G. 2003. Sequence diversity and haplotype structure in the human ABCB1 (*MDR1*, multidrug resistance transporter) gene. *Pharmacogenetics* **13**(8): 481-494.

Labialle, S., Gayet, L., Marthinet, E., Rigal, D., and Baggetto, L. G. 2002. Transcription regulation of the human *MDR1* gene at the level of the inverted MED-1 promoter region. *Ann. NY Acad. Sci.* **973:** 468-471.

Lakowicz, J. R. 1983. Energy Transfer. In *Principles of Fluorescence Spectroscopy*, Plenum Press, New York, p. 303-339.

Lal, R. B., Chakrabarti, S., and Yang, C. 2005. Impact of genetic diversity of HIV-1 on diagnostics, antiretroviral therapy and vaccine development. *Indian J. Med. Res.* **121:** 287-314.

Lane, A. B., Soodyall, H., Arndt, S., Ratshikhopha, 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. *Am. J. Phys. Anthrop.* **119:** 175-185.

Leabman, M. K., Huang, C. C., DeYoung, J., Carlson, E. J., Taylor, T. R., de la Cruz, M., Johns, S. J., Stryke, D., Kawamoto, M., Urban, T. J., Kroetz, D. L., Ferrin, T. E., Clarke, A. G., Risch, N., Herskowitz, I., Giacomini, K. M., and Pharmacogenetics of Membrane Transporters Investigators, 2003. Natural variation in human membrane transporter genes reveals evolutionary and functional constraints. *PNAS* **100**(10): 5896-5901.

Ledford, H. 2010. African yields two full human genomes. *Nature* **463**(18): 853.

Lee, C. G., Ramachandran, M., Jeange, K. T., Martin, M. A., Pastan, I., and Gottesman, M. M. 2000. Effect of ABC transporters on HIV-1 infection: inhibition of virus production by the MDR1 transporter. *FASEB J.* **14:** 516-522.

Leschziner, G. D., Andrew, T., Pirmohamed, M., and Johnson, M. R. 2006. *ABCB1* genotype and PGP expression, function and therapeutic drug response: a critical review and recommendations for future research. *Pharmacogenom. J.* **7:** 154-179.

Leschziner, G. D., Andrew, T., Leach, J. P., Chadwick, D., Coffey, A. J., Balding, D. J., Bentley, D. R., Pirmohamed, M., and Johnson, M. R. 2007. Common ABCB1 polymorphisms are not associated with multidrug resistance in epilepsy using a gene-wide tagging approach. *Pharmacogenet. Genom.* **17:**217-220.

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

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

Loo, T. W., and Clarke, D. M. 1995. Membrane topology of a cystein-less mutant of human P-glycoprotein. *J. Biol. Chem.* **270:** 843-848.

Löscher, W., Klotz, U., Zimprich, F., and Schmidt, D. 2009. The clinical impact of pharmacogenetics on the treatment of epilepsy. *Epilepsy* **50**(1): 1-23.

Malhi, R. S., Eshleman, J. A., Greenberg, J. A., Weiss, D. A., Schultz Shook, B. A., Kaestle, F. A., Lorenz, J. G., Kemp, B., M., Johnson, J. R., and Smith, D. G. 2002. The structural diversity within new world mitochondrial DNA haplogroups: Implications for the prehistory of North America. *Am. J. Hum. Genet.* **70:** 905-919.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, L. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittmore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haine, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. 2009. Finding the missing heredity of complex diseases. *Nature* **461:** 747-753.

Marsh, S. 2008. Pharmacogenetic: Global clinical markers. *Pharmacogenetics* **9**(4): 371-373.

Martin, A. M., Nolan, D., Gaudieri, S., Phillips, E., and Mallal, S. 2004. Pharmacogenetics of antiretroviral therapy: Genetic variation of response and toxicity. *Pharmacogenomic*s **5**(6): 643-655.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J.,, Ioannidis, J. P. A., and Hirschhorn, J. N. 2008. Genome-wide association studies for complex diseases: Consensus, uncertainty and challenges. *Nat. Rev.* **9:** 356-369.

Moriya, Y., Nakamura, T., Horinouchi, M., Sadaeka, T., Tamura, T., Aoyama, N., Shirakawa, T., Gotoh, A., Fujimoto, S., Matsuo, M., Kasuga, M., and Okumura, K. 2002. Effects of polymorphisms of MDR1, MRP1, and MRP2 gene on their mRNA expression levels in duodenal entercytes of healthy Japanese subjects. *Biol. Pharm. Bull.* **25:** 1356-1359.

Morris, L., Taylor, N., Oceguera, L., Bures, R., Gray, C., Sheppard, H., Hanson, C., Montefiori, D., and the Internation Conference on AIDS, 2002. Extensive cross-nationalization of an HIV-1 subtype C vaccine strain by HIV-1 subtype C sera from four countries in southern Africa: the HIV NET 028 study. *Int. Conf. AIDS* **14**.

Muller, M., Bakos, E., Welker, E., Veradi, A., Germann, U. A., Gotterman M. M., Morse, B. S., Roninson, J. B. and Sarkadi, B. 1996. Altered drug-stimulated ATPase activity in mutants of the human multidrug resistance protein. *J. Biol. Chem.* **271:** 1877-1883.

Nakamura, T., Sadaeka, T., Horinouchi, M., Tamura, T., Aoyama, N., Shirakawa, T., Matsuo, M., Kasuga, M., and Okumura, K. 2002. Effect of the mutation (C3435T) at exon 26 of the MDR1 gene on expression level of MDR1 messenger ribonucleic acid in duodenal enterocytes of healthy Japanese subjects. *Clin. Pharmacol.* **71:** 297-303.

Nasi, M., Borghi, V., Piniti, M., Bellodi, C., Lugli, E., Maffei, S., Troiano, L., Richeldi, L, Mussini, C., Esposito, R., and Cossarizza, A. 2003. *MDR1* C3435T genetic polymorphism does not influence the response to antiretroviral therapy in drug-naïve HIV-positive patients. *AIDS* **17**(11): 1696-1698.

Need, A. C, and Goldstein, D. B. 2009. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* **25**(11): 489-494.

Newman, J. L. 1995. The peopling of Africa: A geographic interpretation. Yale University Press, New Haven.

Nsengimana, J., Baret, P., Haley, C. S., and Visscher, P. M. 2004. Linkage disequilibrium in the domesticated pig. *Genetics* **166:** 1395-1404.

Nurse, D. 1982. Bantu expansion into East Africa: linguistic evidence. In: Ehret, C., Posnasky, M. (editors) The archaeological and linguistic reconstruction of African history. Berkeley: University of California Press, p. 161-174.

Ohler, U., Stemmer, G., Harbeck, S. & Niemann, H. 2000. Stochastic models of eukaryotic promoter regions. *Pacific Symp. Biocomp.* **5:**377-388.

Ohler, U., Niemann, H., Liao, G.-C. & Rubin, G. M. 2001. Joint modelling of DNA sequence and physical properties to improve eukaryotic promoter recognition. *Bioinf.* **17:**S199-S206.

Orzack, S. H., Gusfield, D., Olson, J., Nesbitt, S., Subrahmanyan, L., and Stanton, Jr., V. P. 2003. Analysis and exploration of the use of rule-bases algorithms and consensus methods for the inferral of haplotypes. *Genetics* **165:** 915-928.

Owen, A., Chandler, B., Bray, P. G., Ward, S. A., Hart, C. A., Back, D. J., and Khoo, S. H. 2004. Functional correlation of P-glycoprotein expression and genotype with expression of the human immunodeficiency virus type 1 co receptor CXCR4. *J. Virol.* **78:** 12022-12029.

Owen, A., Chandler, B., and Back, D. J. 2005. The implications of P-glycoprotein in HIV: friend or foe? *Fund. & Clin. Pharmacol.* **19:** 283-296.

Pauli-Magnus, C., and Kroetz, D. L. 2004. Functional implications of genetic polymorphisms in the multidrug resistance gene *MDR1* (*ABCB1*). *Pharmaceut. Research* **21**(6): 904-913.

Phillipson, D. W. 1993. African archaeology. Cambridge University Press, Cambridge.

Pozio, E., and Morales, M. A. G. 2004. The impact of HIV-protease inhibitors on opportunistic parasites. *TRENDS Parasit.* **21**(2): 58-63.

Prestridge, D. S. 2000. Computer Software for eukaryotic promoter analysis. *TF Prot.* **130:**265-295.

Price, A. L., Weale, M. E., Patterson, N., Myers, S. R., Need, A. C., Shianna, K. V., Ge, D., Rotter, J. I., Torres, E., Taylor, K. D., Goldstein, D. B., and Reich, D. 2008. Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.* **83:** 127-147.

Reed, F. A., and Tishkoff, S. A. 2006. African human diversity, origins and migrations. *Curr. Opin. Genet. Devel.* **16:** 597-605.

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

Ritchie, M. D., Haas, D. W., Motsinger, A. A., Donahue, J. P., Erdem, H., Raffanti, S., Rebeiro, P., George, A. L., Kim, R. B., Haines, J. L., and Sterling, T. R. 2006. Drug transporter and metabolizing enzyme gene variants and nonnucleoside reverse-transcriptase inhibitor hepatotoxicity. *Clin. Infect. Dis.* **43:** 779-782.

Roses, A. D. 2000. Pharmacogenetics and future drug development and delivery. *Lancet* **355:** 1358-1361.

Roses, A. D. 2000. Pharmacogenetics and the practice of medicine. *Nature* **405**(13): 857-865.

Roshner, B. 1990. Fundamentals of Biostatistics, 3rd edition, pp. 328-330 and 341. PWS-KENT Publishing Company.

Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., Willey, D. L., Hunt, S. E., Cole, C. G., Cogill, P. C., Rice, C. M., Ning, Z., Rogers, J., Bentley, D. R., Kwok, P.-Y., Mardis, E. R., Yeh, R. T., Schultz, B., Cook, L., Davenport, R., Dante, M., Fulton, L., Hillier, L., Waterston, R. H., McPherson, J. D., Gilman, B., Schaffner, S., Van Etten, W. J., Reich, D., Higgins, J., Daly, M. J., Blumenstiel, B., Baldwin, J., Stange-Thomman, N., Zody, M. C., Linton, L., Lander, E. S., and Altshuler, D. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928-933.

Sakurai, A., Onishi, Y., Hirano H., Seigeuret, M., Obanayama, K., Kim, G., Liew, E. L., Sakaeda, T., Yoshiura, K., Niikawa, N., Sakurai, M. and Ishikawa, T. 2007. Quantitative structure-activity relationship analysis and molecular dynamics simulation to functionally validate nonsynonymous polymorphisms of human ABC transporter ABCB1 (P-glycoprotein/MDR1). *Biochemistry* **46:** 7678-7693.

Schaeffeler, E., Eichelbaum, U., Penger, A., Asante-Poku, S., Zanger, U. M., and Schwab, M. 2001. Frequency of C3435T polymorphism of *MDR1* gene in African people. *Lancet* **358:** 383-384.

Schmith, V. D., Campbell, D. A., Sehgal, S., Anderson, W. H., Burns, D. K., Middleton, L. T., and Roses, A. D. 2003. Pharmacogenetics and disease genetics of complex diseases. *Cell. Mol. Life Sci.* **60:** 1636-1646.

Schmittgen, T. D., and Zakrajsek, B. A., 2000. Effect of experimental treatment on housekeeping gene expression: validation by rel-time, quantitative RT-PCR. *J. Biochem. Biophys. Methods* **46:** 69-81.

Schwab, M., Eichelbaum, M., and Fromm, M. F. 2003. Genetic polymorphisms of the human *MDR1* drug transporter. *Ann. Rev. Pharmacology & Toxicology* **43:** 285-307.

Scotto, K. W. 2003. Transcriptional regulation of ABC drug transporters. *Oncogene* **22:**7496-7511.

Shifman, S., Kuypers, J., Kokoris, M., Yakir, B., and Darvasi, A. 2003. Linkage disequilibrium patterns of the human genome across populations. *Human Mol. Genet.* **12**(7): 771-776.

Shisana, O., Rehle, T., Simbayi, L. C., Zuma, K., Jooste, S., Pillay-Van-Wyk, V., Mbele, N., Van Zyl, J., Parker, W., Zungu, N. P., Pezi, S., and the SABSSM III implementation Team, 2009. South African National HIV prevalence, incidence, behaviour and communication survey 2008: A turning tide among teenagers? Cape Town: HSRC Press.

Solovyev, V. V. & Salamov, A. 1997. The Gene-Finder computer tools for analysis of human and model organisms genome sequences. *Proc. Int. Conf. Intell. Sys. Biol.* **5,**294-302.

Solovyev, V. V. 2001. Statistical approaches in eukaryotic gene prediction. *Handbook of statistical genetics* (eds D. J. Balding et al.) p.83-127, John Wiley & Sons, Ltd. New York.

Solovyev, V. V. & Shahmuradov, I. A. 2003. PromH: promoters identification using orthologous genomic sequences. *Nucleic Acids Res.* **31**(13):3540-3545.

Soranzo, N., Cavalleri, G. L., Weale, M. E ., Wood, N. W., Depondt, C., Marguerie, R., Sisodiya, S. M., and Goldstein, D. B. 2004. Identifying candidate causal variants responsible for altered activity of ABCB1 multidrug resistance gene. *Genome Res.* **14:** 1333-1344.

Speck, R. R., Yu, X. F., Hildreth, J., Flexner, C. 2002. Differential effects of p-glycoprotein and multidrug resistance protein-1 on productive human immunodeficiency virus infection. *J. Infect. Dis.* **186:** 332-340.

Stenger, M., and Kim, R. B. 2003. Perspective- Drug transporters in HIV therapy. *Trop. HIV Med.* **11**(4): 136-139.

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

Stephens, M., and Sheet, P. 2005. Accounting for decay of linkage disequilibrium in haplotype inference and missing data imputation. *Am. J. Human Genet.* **76:** 449-462.

Störmer, E., Von Moltke, L. L., Perloff, M. D., and Greenblatt D. J. 2002. Differential modulation of P-glycoprotein expression and activity by non-nucleoside HIV-1 reverse transcriptase inhibitors in cell culture. *Pharmaceut. Res.* **19**(7): 1038-1045.

Sukhai, M., and Piquette-Miller, M. 2000. Regulation of the multidrug resistance genes by stress signals. *J. Pharm. Pharmaceut. Sci*. **3**(2): 268-280.

Szakacs, G., Ozvegy, C., Bakos, E., Sarkadi, B., and Veradi, A. 2000. Transition-state formation in ATPase-negative mutants of the human MDR1 protein. *Biochem. Biophys. Res. Commun.* **276:** 1314-1319.

Takane, H., Kobayashi, D., Hirota, T., Kigawa, J., Terakawa, N., Otsubo, K., and Ieiri, I. 2004. Haplotype-oriented genetic analysis and functional assessment of promoter variants in the *MDR1* (ABCB1) gene. *J. Pharmaceut. Exp. Therapeut.* **311**: 1179-1187.

Tang, K., Ngoi, S. M., Gwee, P. C., Chua, J. M. Z., Lee, E. J. D., Chong, S. S., and Lee, C. G. L. 2002. Distinct haplotype profiles and string linkage disequilibrium at the *MDR1* multidrug transporter gene locus in three ethnic Asian populations. *Pharmacogenetics* **12**(6): 437-450.

Tang, K., Wong, L. P., Lee, E. J. D., Chong, S. S., and Lee, C. G. L. 2004. Genomic evidence for recent positive selection at the human *MDR1* gene locus. *Hum. Mol. Genet.* **13**(8): 783-797.

Taniguchi, S., Mochida, Y., Uchiumi, T., Tahira, T., Hayashi, K., Takagi, K., Shimada, M., Maehara, Y., Kuwano, H., Kono, S., Nakano, H., Kuwano, M., and Wada, M. 2003. Genetic polymorphisms at the 5' regulatory region of multidrug resistance 1 (*MDR1*) and its association with interindividual variation of expression level in the colon. *Mol. Cancer Ther.* **2:** 1351-1359.

Teare, M. D., Dunning, A. M., Durocher, F., Rennart, G., and Easton, D. F. 2002. Sample distribution of summary linkage disequilibrium measures. *Ann. Hum. Genet.* **66:** 223-233.

Telenti, A., Aubert, V., and Spertini, F. 2002. Individualizing HIV treatment-Pharmacogenetics and immunogenetics. *Lancet* **359:** 722-723.

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A., A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A. T., Kotze, M. J., Lema, G., Moore, J. H., Mortensen, H., Nymbo, T. B., Omar, S. A., Powell, K., Pretorius, G. S., Smith, M. W., Thera, M. A., Wambebe, C., Weber, J. L., Williams, S. M. 2009. The genetic structure and history of Africans and African Americans. *Science* **324:** 1035-1044.

Tishkoff, S. A., and Williams, S. M. 2002. Genetic analysis of African populations: Human evolution and complex disease. *Nature Rev.* **3:** 611-621.

Travel, J. A., Miller, K. D., and Masur, H. 1999. Guide to major clinical trials of antiretroviral therapy in human immunodeficient virus-infected patients: Protease inhibitors, NNRTIs and NRTIs. *Clin. Infect. Dis.* **28:**643-676.

Urbatsch, I. L., Gimi, K., Wilke-Mounts, S., and Senior, A. E. 2000. Investigation of the role of glutamine-471 glutamine-1114 in the two catalytic sites of P-glycoprotein. *Biochemistry* **39:** 11921-11927.

Urbatsch, I. L., Wilke-Mounts, S., Gimi, K., and Senior, A. E. 2001. Purification and characterization of N-glycosylation mutation mouse and human P-glycoprotein expressed in *Pichia pastoris* cells. *Arch. Biochem. Biophys.* **388:** 171-177.

Van Heeswijk R. P. G., Veldkamp, A. I., Mulder, J. W., Meenhorst, P. L., Lange, J. M. A., Beijnen, J. H., and Hoetelmans, R. M. W. 2002. Combination of protease inhibitors for the treatment of HIV-1-infected patients: A review of pharmacokinetics and clinical experience. *Antivir. Ther.* **6:** 210-229.

Vansina, J. 1984. Western Bantu expansion. *J. Afr. Hist.* **25:** 129-145.

Van Tassell, C. P., Smith, T. P. L., Matukumalli, L. K., Taylor, J. F., Schnabel, R. D., Lawley, C. T., Haudenschild, C. D., Moore, S. S., Wesley C Warren, W. C., and Sonstegard, T. S. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5:** 247-252.

Walensky, R. P., Wolf, L. L., Wood, R., Fofana, M. O., Freedberg, K. A., Martinson, N. A., Paltiel, D., Anglaret, X., Weinstein, M. C., Losina, E., and the CEPAC International Investigators, 2002. When to start antiretroviral therapy in resource-limited settings. *Ann. Internat. Med.* **151**(3): 157-166.

Wang, B., Ngoi, S., Wang, J., Chong, S. S., and Lee, C. G. L. 2006. The promoter region of the MDR1 gene is largely invariant, but different single nucleotide polymorphism haplotypes affect MDR1 promoter activity differently in different cell lines. *Mol. Pharmacol.* **70**(1): 267-276.

Wang, D., Johnson, A. D., Papp, A. C., Kroetz, D. L., and Sadee, W. 2005. Multidrug resistance polypeptide 1 (MDR1, ABCB1) variant 3435C>T affects mRNA stability. *Pharmacogenet. Genom.* **15:** 693-704.

Wang, D. & Sadée, W. 2006. Searching for polymorphisms that affect gene expression and mRNA processing: Example ABCB1 (MDR1). *AAPS J.* **8**(3), E515-E520.

Wang, D. G., Fan, J.-B., Siao, C.-J., Berno, A., Young, P., Sapolosky, R., Ghandour, G., Perking, N., Winchester, E., Spencer, J., Kruglyak, L., Stein, L., Hsie, L., Topaloglou, T., Hubbell, E., Robinson, E., Mittmann, M., Morris, M. S., Shen, N., Kilburn, D., Rioux, J., Nusbaum, C., Rozen, S., Hudson, T. J., Lipshutz, R., Chee, M., and Lander, E. S. 1998. Large-scale identification, mapping and genotyping of single-nucleotide polymorphisms in the human genome. *Science* **280:** 1077-1082.

Watkins, W. S., Rogers, A. R., Ostler, C. T., Wooding, S., Bamshad, M. J., Brassinger, A.-M. E., Carrol, M. L., Nguyen, S. V., Walker, J. A., Prasad, B. V. R., Reddy, P. G., Das, P. K., Batzer, M. A., and Jorde, L. B. 2003. Genetic variation among world populations: Inferences from 100 *Alu* insertion polymorphisms. *Genome Res.* **13**(7): 1607-1618.

Wilson, J. F., Weale, M. E., Smith, A. C., Gratrix, F., Fletcher, B., Thomas, M. G., Bradman, N., and Goldstein, D. B. 2001. Population genetic structure of variable drug response. *Nat. Genet.* **29:** 265-269.

Winzer, R., Langmann, P., Zilly, M., Tollmann, F., Schubert, J., Klinker, H., and Weissbrich, B. 2005. No influence of the P-glycoprotein polymorphisms *MDR1* G2677T/A and C3435T on the virological and immunological response in treatment naïve HIV-positive patients. *Ann. Clin. Micro. Antimicrob.* **4:** 3-9.

World Health Organisation (WHO), 2008. Towards universal access: scaling up priority HIV/AIDS interventions in the health sector: Progress report 2008.

Xiao, L., Rudolf, D. L., Owen, S. M., Spira, T. J., and Lal, R. B. 1998. Adaption to promiscuous usage of CC and CXC-chemokine co receptors in vivo correlates with HIV-1 disease progression. *AIDS* **12:** F137-143.

Yu, N., Chen, F.-C., Ota, S., Jorde, L. B., Pamilo, P., Patthy, L., Ramsey, M., Jenkins, T., Shyue, S.-K., and Li, W.-H. 2002. Larger genetic differences within Africans than between Africans and Eurasians. *Genetics* **161:** 269-274.

Zhong, H., and Simons, J. W., 1999. Direct comparison of GAPDH, β-actin, cyclophilin, and 28S rRNA as internal standards for quantifying RNA levels under hypoxia. *Biochem. Biophys. Res. Comm.* **259:** 523-526.

Zhu, D., Taguchi-Nakamura, H., Goto, M., Odawara, T., Nakamura, T., Yamada, H., Kotaki, H., Sugiura, W., Iwamoto, A, and Kitamura, Y. 2004. Influence of single nucleotide polymorphisms in multidrug resistance-1 gene on cellular export of Nelfinavir and its clinical implications for highly active antiretroviral therapy. *Antivir. Ther.* **9**(6): 929-935.

## BIOINFORMATIC TOOLS:

UCSC: http://genome.ucsc.edu

Ensembl: http://www.ensembl.org

NCBI: http://www.ncbi.nlm.nih.gov

Fluxus Engeneering (NETWORK 4.510): http://www.fluxus-engeneering.com

PROSCAN Version 1.7: http://www-bimas.cit.nih.gov/molbio/proscan/

FPROM:http://www.softberry.ru/berry.phtml?group=programs&subgroup=promoter&topic=fprom

NSITE:http://www.softberry.ru/berry.phtml?group=programs&subgroup=promoter&topic=nsite

TFSEARCH: http://www.cbrc.jp/research/db/TFSEARCH.html

# APPENDIX I

A) Instructions for DNA extraction using the QIAGEN QIAmp DNA Blood Mini Kit

The leukocyte-rich buffy coat was obtained by centrifuging the blood vial at 2,500 x g for 10 minutes at room temperature. The buffy coat was removed and 200 µl added to 20 µl of protinase K, 4 µl of RNase A and 200 µl of lysis buffer (buffer AL) in a 1,5 ml microcentrifuge tube. An incubation at $56^{O}C$ for 10 minutes ensured protein and RNA digestion, as well as cell lysis. The sample was briefly centrifuged for 15 sec at 8,000 x g at room temperature to remove the drops of liquid from the lid. Ethanol (200 µl) was added to precipitate the DNA, the sample was mixing by pulse-vortexing for 15 seconds and then briefly centrifugation at 8,000 x g at room temperature.

The sample was added to the spin column in a 2 ml collection tube, centrifuged for 1 min at 8,000 x g at room temperature in order to bind the DNA to the column. The column was placed in a new collection tube and the filtrate, with the enzymes, buffer and ethanol, discarded. The DNA was washed by adding 500 µl of wash buffer (buffer AW1), followed by centrifugation at 8,000 x g for 1 min at room temperature. The column was placed in a new collection tube and the filtrate discarded. The DNA was washed a second time by adding 500 µl of wash buffer (buffer AW2), followed by centrifugation at full speed (13,400) at room temperature for 3 min to wash the DNA and a further 1 minute to eliminate buffer

carry over. The column was placed in a clean 1,5 ml microcentrifuge tube, the filtrate was discarded, 200 µl of storage buffer (buffer AE) added, followed by incubation at room temperature for 1 min. The sample was centrifuged at 8,000 x g for 1 min at room temperature to elute the DNA. The sample was stored in buffer AE instead of water to avoid hydrolysis.

B) TRIzol LS extraction of whole RNA

The blood vials were centrifuged at 2,500 x g for 10 minutes at room temperature to obtain the leukocyte-rich buffy coat. The buffy coat was removed and 250 µl was added to 750 µl of TRIzol LS. The samples are vortexed to homogenise the solutions and 200 µl of chloroform was added. The tubes were shaken vigorously by hand for 15 sec and incubated at room temperature for 2-15 min. Centrifugation at 12,000 x g at $4^{O}C$ for 15 min separated the solution into two phases. The top layer contained the RNA and was transferred to a clean tube. The lower layer contained DNA, proteins and other cell material and was discarded.

To resuspend the RNA 500 µl of isopropyl alcohol was added to the tube, which was then incubated at room temperature for 10 min and centrifuged at 12,000 x g at $4^{O}C$ for 10 min, to produce a pellet of RNA. The supernatant was removed and the RNA pellet was washed once with 1 ml of 75 % ethanol. The ethanol was added to the pellet, the tube was briefly vortexed and then centrifuged at 7,500 x g at $4^{O}C$ for 5 min. To resuspend the RNA, the ethanol was removed and the

pellet was air-dried. RNase-free water was added to the pellet and the tubes were incubated at 55-60$^{O}$C for 10 min.

C) High-Capacity cDNA Reverse Transcription Kit

A 2X reverse transcription (RT) master mix was produced using 2 µl of 10X RT Buffer, 0.8 µl 25X dNTP mix, 2 µl 10X RT random primers, 1 µl reverse transcriptase, 1 µl RNase inhibitor and 3.2 µl nuclease-free water to make a volume of 10 µl. These 10 µl were added to 10 µl of RNA, briefly centrifuged and placed in a thermocycler. The reverse transcription reaction involved one cycle of a ten minute random primer annealing step at 25$^{O}$C, a 120 minute extension step at 37$^{O}$C, a final primer denaturation step at 85$^{O}$C for five seconds and a hold step at 4$^{O}$C.

D) Calibrator Normalised Relative Quantification of mRNA

A reaction mix was set up for a 20 µl reactions for the samples and a calibrator for the target gene, consisting of 10 µl of 2 X TaqMan® Universal Master Mix, 1 µl of 20 X TaqMan Gene Expression Assay (containing the primes and probes), 7 µl of nuclease-free water and 2 µl of cDNA. A reaction mix for the reference gene was set up for a 20 µl reaction for the samples and a calibrator, consisting of 10 µl of 2 X TaqMan® Universal Master Mix, 1 µl of 20 X endogenous control assay (containing the primes and probes), 7 µl of nuclease-free water and 2 µl of cDNA.

Samples were placed in the Applied Biosystems 7500 Real Time PCR System and underwent the following reaction: The first hold stage was at $52^{O}$C for 2 minutes, followed by a second hold stage at $95^{O}$C for 10 minutes, and then 40 cycles of the cycling stage ($95^{O}$C for 15 seconds and $60^{O}$C for 1 minute), where the fluorescence was measure at the end of the $60^{O}$C step.

## APPENDIX II: GENOTYPING DATA

| Sample # | Exon 2 | Intron 12 | Exon 21 | Intron 25 | | Exon 26 | Intron 26 | Haplotype Pairs[a] |
|---|---|---|---|---|---|---|---|---|
| | T-129C | C1236T | G2677T/A | G3050T | T5231C | C3435T | T80C | |
| 101 | NN | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 104 | NN | CT | NN | GG | TC | CC | TT | |
| 105 | TC | TT | GA | GG | TC | CC | TT | |
| 106 | TC | TT | GG | GG | TC | CC | NN | |
| 109 | TT | CC | GG | GT | TC | CT | TC | TCGGCCT, TCGTTTC |
| 110 | NN | CT | GA | GG | TC | CC | TT | |
| 111 | TT | CT | GG | GT | CC | CT | TC | |
| 112 | TC | CT | GG | NN | TT | CC | TT | |
| 113 | TC | TT | GA | GT | TT | CT | TC | |
| 114 | TC | CT | GA | GG | CC | CC | NN | |
| 115 | TC | CT | GG | GG | TC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 116 | NN | CT | NN | GG | TC | NN | NN | |
| 117 | NN | CT | NN | GG | TC | CC | NN | |
| 118 | TT | TT | GG | GT | TT | TT | CC | TTGGTTC, TTGTTTC |
| 119 | TT | CT | GA | GG | CC | CC | NN | |
| 120 | TT | CT | GG | GT | TC | CT | TC | |
| 122 | TT | TT | GG | GG | TC | CC | NN | TTGGTCT, TTGGCCT |
| 123 | TT | CT | GA | GG | TC | CC | TT | |
| 124 | NN | CC | NN | GG | CC | CC | TT | TCGGCCT, TCGGCCT |
| 125 | TT | TT | GA | GG | TC | CC | TT | |
| 126 | TC | CT | GA | GT | TC | CC | TT | |
| 127 | TC | CT | GA | GG | TC | CC | TT | |
| 128 | NN | CT | NN | GG | TC | TT | NN | |
| 129 | NN | CT | GG | GT | TT | CC | TT | |
| 130 | TT | CT | GG | GT | TC | CC | TT | |
| 131 | TT | CT | GG | GG | TC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 137 | NN | NN | NN | GG | TC | CC | TT | |
| 138 | NN | TT | GA | GG | TT | CC | NN | TTGGTCT, TTAGTCT |
| 139 | NN | TT | NN | NN | CC | CC | NN | |
| 141 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 142 | TC | CT | GA | GG | TC | CC | TT | |
| 143* | TT | CT | GA | GT | CC | NN | TT | |
| 144* | TT | CT | GG | GG | TC | CC | TT | |
| 145 | TC | CT | GG | GG | TC | CT | TC | |
| 146 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 147 | TT | CT | GG | GT | TC | CT | TC | |
| 148 | TC | TT | GG | GG | TT | CC | TT | TTGGTCT, CTGGTCT |
| 149 | TT | CT | GG | GT | CC | CC | TT | |
| 150 | TC | TT | GA | GG | CC | CC | TT | |
| 151 | TT | TT | GA | GG | TC | CC | TT | |
| 153 | TC | TT | GG | GG | TT | CC | TT | TTGGTCT, CTGGTCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 154 | TT | TT | GA | GG | TC | CC | TT | |
| 155 | TT | TT | GA | GT | CC | CC | TT | |
| 156 | TT | TT | GG | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 157 | TT | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 158 | TC | TT | GG | GG | TT | CC | TT | TTGGTCT, CTGGTCT |
| 159 | TT | CT | GA | GG | CC | CC | TT | |
| 160 | TC | CT | GG | GG | TT | CC | TT | |
| 161 | TT | TT | GG | GT | TC | CC | TT | TTGGCCT, TTGTTCT |
| 162 | NN | TT | GA | NN | TT | CC | TT | TTGGTCT, TTAGTCT |
| 163 | TT | TT | GG | GT | TC | CT | TC | TTGGCCT, TTGTTTC |
| 164 | NN | TT | NN | GT | TT | CC | TT | TTGGTCT, TTGTTCT |
| 166 | NN | TT | GA | GT | TT | CT | TC | |
| 167 | NN | CC | NN | GT | TT | CC | TT | |
| 168 | NN | NN | NN | GT | TC | CC | NN | |
| 169 | TC | TT | GG | GT | TT | CT | TC | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 170 | TT | TT | GA | GT | TT | CT | TC | |
| 171 | TT | TT | GT | GG | TT | CT | TC | |
| 172 | TT | CT | GA | GT | TT | CC | TT | |
| 173 | NN | TT | NN | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 175 | TT | CT | GA | GG | TT | TC | NN | |
| 176 | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 177 | NN | TT | NN | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 178 | NN | CC | GG | GG | TT | CC | TT | TCGGTCT, TCGGTCT |
| 179 | TT | CC | GG | GT | TT | CC | TT | TCGGTCT, TCGTTCT |
| 180 | TT | TT | GG | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 181 | CC | TT | GG | GG | CC | CC | TT | CTGGCCT, CTGGCCT |
| 182 | TC | TT | GG | GT | TT | CT | TC | |
| 183 | TT | CC | GG | GG | TC | CC | TT | TCGGTCT, TCGGCCT |
| 184 | TC | TT | GA | GT | TC | CC | TT | |
| 185 | NN | TT | GG | TT | NN | NN | NN | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 187 | TT | TT | GG | NN | TC | CC | TT | TTGGTCT, TTGGCCT |
| 188 | CC | CT | GA | GG | TC | CC | TT | |
| 189 | NN | CT | GG | GG | TC | CC | TT | |
| 190 | TT | CT | GG | GG | TC | CC | TT | |
| 191 | TT | TT | GG | GG | TT | CT | TC | TTGGTCT, TTGGTTC |
| 192 | TT | TT | GG | GG | TC | CC | NN | TTGGTCT, TTGGCCT |
| 193 | NN | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 194 | CC | TT | GA | GG | TC | CC | TT | |
| 195 | TT | TT | GT | GG | CC | CC | TT | TTGGCCT, TTTGCCT |
| 196 | TT | CT | GG | GT | TC | CC | TT | |
| 197 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 198 | TC | TT | GA | GG | TC | CC | TT | |
| 199 | TT | TT | GT | GG | TC | CC | TT | |
| 200 | TT | TT | GT | GG | TC | CC | TT | |
| 201 | TC | CT | GG | GT | TT | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 203 | TT | CT | TA | GT | TC | CC | TT | |
| 205 | TC | TT | GA | GG | CC | CC | TT | |
| 206* | TC | TT | GG | GG | TC | CC | TT | |
| 207* | TT | TT | GG | TT | TC | CC | TT | TTGTTCT, TTGTCCT |
| 208* | TT | CT | GA | GG | TC | CT | TT | |
| 209* | TT | TT | GG | GT | TT | TT | TC | TTGGTTT, TTGTTTC |
| 210* | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 211* | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 212* | TC | TT | GG | GT | TT | CT | TC | |
| 213* | TT | TT | GG | GG | CC | CT | TT | TTGGCCT, TTGGCTT |
| 214* | TT | TT | GG | GT | TC | CT | TC | TTGGCCT, TTGTTTC |
| 215* | TT | TT | GG | GT | TC | CT | TC | TTGGCCT, TTGTTTC |
| 216* | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 217* | TC | CT | GG | GG | TT | CC | TT | TTGGTCT, TTGGCCT |
| 218* | TT | TT | GG | GG | TC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 219* | TC | TT | GG | GT | TC | CT | TC | |
| 222* | TT | TT | GA | GG | TC | CC | TT | |
| 223* | TT | CT | GA | GG | TC | CC | TT | |
| 224* | TT | TT | GA | GT | TC | CT | TC | |
| 225* | TT | TT | GG | GT | TC | CC | TT | TTGGCCT, TTGTTCT |
| 226* | TT | TT | GG | GT | TT | CT | TC | TTGGTCT, TTGTTTC |
| 227* | TC | TT | GA | GG | TC | CC | TT | |
| 228* | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 229* | TT | CT | GG | GG | TC | CC | TT | |
| 230* | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 231* | TT | CT | GG | GT | TC | CC | TT | |
| 232* | TT | CC | GG | GG | TC | CC | TT | TCGGTCT, TCGGCCT |
| 233* | TT | CC | GG | GG | TT | CC | TT | TCGGTCT, TCGGTCT |
| 234* | TT | CT | GA | GG | TT | CC | TT | |
| 235* | TC | CT | GG | GG | TC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 236* | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 237* | CC | TT | GG | GG | TC | CC | TT | CTGGTCT, CTGGCCT |
| 238* | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 239* | TT | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 240* | TC | CT | GG | GG | TC | CC | TT | |
| 241* | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 242* | TT | CT | GG | GG | TC | CC | TT | |
| 243* | TT | TT | GG | GT | TC | CT | TC | TTGGCCT, TTGTTTC |
| 244* | TT | CT | GG | GG | TC | CC | TT | |
| 245* | TT | CT | GG | GG | TC | CC | TT | |
| 300 | TC | CC | GA | GG | TC | CC | TT | |
| 301 | TC | CT | GA | GG | TT | CC | TC | |
| 302 | TC | TT | GA | GG | TC | CC | TC | |
| 304 | TC | CC | GA | GG | TT | CC | TT | |
| 305 | TC | CC | GG | GG | TC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 307 | TT | CT | GG | GG | TT | CC | TC | |
| 308 | TC | CT | GG | GG | TT | CC | TT | |
| 309 | TT | CC | GG | GG | CC | CC | TT | TCGGCCT, TCGGCCT |
| 310 | TT | CT | GA | GG | CC | CC | TT | |
| 311 | TT | TT | GG | GT | TT | CT | TC | TTGGTCT, TTGTTTC |
| 312* | NN | CT | GG | GG | CC | CT | TT | |
| 313 | TT | CT | GA | GG | TC | CC | TT | |
| 314 | TT | CT | GA | GG | CC | CT | TT | |
| 316 | TT | CC | GG | GG | TC | CC | TT | TCGGTCT, TCGGCCT |
| 317 | TT | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 318 | TT | CC | GG | GG | TC | CC | TT | TCGGTCT, TCGGCCT |
| 319 | TT | CT | TA | GT | TC | CC | TC | |
| 320 | TC | CT | GG | GG | TT | CC | TT | |
| 321 | TC | CT | GA | TT | TT | CT | TC | |
| 322 | TT | TT | GA | TT | CC | TT | TC | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 323 | TT | CT | GG | GG | CC | CC | TC | |
| 324 | TT | CT | GG | GG | CC | CC | TC | |
| 325 | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 326 | TT | TT | GT | TT | TT | CC | TT | TTGTTCT, TTTTTCT |
| 327 | TT | CC | GA | GG | CC | CC | TT | TCGGCCT, TCAGCCT |
| 328 | TT | CT | GT | GG | TT | CC | TT | |
| 329 | TT | CT | GG | GG | TC | CC | TC | |
| 330 | TT | TT | GG | TT | CC | CC | TT | TTGTCCT, TTGTCCT |
| 331 | TT | CT | GG | GG | TC | CC | TC | |
| 332 | TC | CT | GT | GT | TC | CC | TC | |
| 333 | TT | TT | GG | GT | TT | CT | TT | TTGGTCT, TTGTTTT |
| 334 | TC | CT | GG | GG | TT | CC | TT | |
| 335 | TT | TT | GG | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 336 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 337 | TC | CT | TA | GG | TC | CC | TC | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 338 | TC | CT | GA | GG | TC | CC | TT | |
| 339 | TT | CT | GT | GG | TC | CC | TT | |
| 340 | TC | CT | TA | GG | TC | CC | TT | |
| 341 | TT | CT | GA | GT | TC | CT | TT | |
| 342 | TC | CT | TA | GG | TC | CC | TT | |
| 343 | TT | CT | GA | GG | TC | CC | TC | |
| 344 | TT | CT | GG | GT | TT | CC | TT | |
| 345 | TT | CT | GT | GT | CC | CT | TT | |
| 346 | TT | CC | GG | GG | TC | CC | TT | TCGGTCT, TCGGCCT |
| 347 | TT | CT | GA | GG | CC | CC | TT | |
| 348 | TT | CT | GG | GT | TT | CT | TT | |
| 349 | TT | CT | GA | GG | TT | CC | TT | |
| 350 | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 352 | TT | TT | GG | GT | TT | CC | TT | TTGGTCT, TTGTTCT |
| 353 | NN | TT | GG | GT | TC | CC | TT | TTGGCCT, TTGTTCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 354 | TT | CT | GG | GG | TC | CC | TT | |
| 355 | TT | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 356 | TT | CT | GA | GT | TT | CC | TT | |
| 358 | TT | CT | GA | TT | CC | CT | TT | |
| 359 | TT | CT | GG | GG | TC | CC | TT | |
| 360 | TT | CT | GG | GG | TC | CC | TT | |
| 361 | TT | CT | GG | GG | TC | CT | TC | |
| 362 | TT | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 363 | TC | CT | GA | GG | TC | CC | TT | |
| 364 | TT | CC | GA | GG | CC | CC | TT | TCGGCCT, TCAGCCT |
| 366 | TT | CC | GG | GG | TC | CC | TT | TCGGTCT, TCGGCCT |
| 367 | TT | CT | GG | GG | TC | CC | TT | |
| 368 | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 369 | NN | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 370 | NN | CC | GG | GG | CC | CC | TT | TCGGCCT, TCGGCCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 371 | TC | CT | GG | GG | TC | CC | TT | |
| 372 | TT | CT | GA | GG | TC | CC | TT | |
| 373 | TT | CT | GG | GG | TC | CT | TT | |
| 374 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 375 | TC | CT | GA | GG | TC | CC | TT | |
| 377 | TC | CT | GG | GG | TC | CT | TT | |
| 378 | CC | CT | GG | GG | TT | CT | TC | |
| 379 | TT | CT | GG | GT | TT | CC | TC | |
| 380 | TC | TT | GG | GT | TT | CC | TC | |
| 381 | TT | TT | GA | GT | TT | CC | TC | |
| 382 | TT | CT | GG | GG | TC | CC | TC | |
| 383 | TT | CT | GG | GG | CC | CC | TC | |
| 384 | TT | CC | GA | GG | TC | CC | TC | |
| 385 | TT | CT | GA | GG | TC | CC | TT | |
| 386 | TC | TT | GA | TT | TC | CT | TC | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 387 | TT | CT | GA | GG | TT | CC | TT | |
| 388 | TT | CT | GA | GG | TC | CC | TT | |
| 389 | TT | TT | GA | GG | CC | CT | TT | |
| 390 | TT | TT | GG | GT | TT | CC | CC | TTGGTCC, TTGTTCC |
| 391 | NN | TT | GG | GT | TT | CC | TT | TTGGTCT, TTGTTCT |
| 392 | TC | CT | GG | GG | TT | CC | TT | |
| 393 | TT | CT | GA | GG | TC | CC | TT | |
| 394 | TT | TT | GA | GG | TT | CT | TT | |
| 395 | TT | CC | GA | GG | TC | CC | TT | |
| 396 | TT | TT | GA | GT | TT | CT | TT | |
| 397 | NN | TT | GG | NN | TT | CC | TC | TTGGTCT, TTGGTCC |
| 398 | TT | TT | GA | GT | TT | CC | TT | |
| 399 | TT | CC | GG | GG | TC | CT | TT | |
| 400 | NN | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 401 | NN | TT | GA | NN | NN | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 402 | TT | CT | GA | GG | TC | CC | TT | |
| 1000 | TT | CT | GA | GG | CC | CT | TC | |
| 1001 | TT | TT | GG | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 1002 | TT | CT | GA | GG | CC | CC | TT | |
| 1003 | TT | CT | GA | GT | TC | CT | TT | |
| 1008 | TT | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 1011 | TT | CT | GA | GG | TC | CT | TT | |
| 1019 | TT | TT | GA | GG | TC | TT | CC | |
| 1021 | TT | CT | GA | GG | TC | CT | TC | |
| 1023 | TC | CC | GG | GG | CC | CC | TT | TCGGCCT, CCGGCCT |
| 1025 | TT | CC | GG | GG | CC | CC | TT | TCGGCCT, TCGGCCT |
| 1027 | CC | CT | GG | GG | CC | CC | TT | CTGGCCT, CCGGCCT |
| 1033 | TT | CT | GA | GG | TT | CC | TT | |
| 1034 | TT | TT | GA | GG | TT | CT | TC | |
| 1035 | TC | TT | GA | GT | TT | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1036 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 1038 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 1040 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1046 | TC | TT | GG | GG | TC | CC | TT | |
| 1047 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 1049 | TT | TT | GA | GG | TC | CC | TT | |
| 1050 | TT | TT | GG | GG | CC | CT | TT | TTGGCCT, TTGGCTT |
| 1051 | TT | CT | GA | GT | CC | CC | TT | |
| 1053 | TC | TT | GA | GG | CC | CC | TT | |
| 1056 | TC | CT | GA | GG | TC | CC | TT | |
| 1057 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1058 | TT | CT | GA | GG | TC | CC | TT | |
| 1062 | TC | CC | GA | GG | TT | CC | TT | |
| 1064 | TT | TT | GA | GG | TC | CT | TC | |
| 1065 | TT | CT | GA | GG | CC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1066 | TT | TT | GA | GT | CC | CC | TT | |
| 1067 | TT | TT | GA | GG | TC | CT | TT | |
| 1068 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1072 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1073 | TT | TT | GA | GG | TT | CT | TC | |
| 1074 | TT | TT | GA | GG | CC | CT | TC | |
| 1075 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1076 | NN | TT | GG | GT | TC | CT | TC | |
| 1079 | TC | NN | GG | NN | NN | CC | TT | |
| 1080 | TC | CC | GG | GG | CC | CC | TT | TCGGCCT, CCGGCCT |
| 1081 | TC | TT | GG | GG | TC | CC | TT | |
| 1084 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1086 | TC | TT | GA | GG | CC | CC | TT | |
| 1087 | TC | CT | GG | GG | CC | CC | TT | |
| 1089 | TT | TT | GA | GG | TC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1091 | TT | TT | GA | GG | TC | CC | TT | |
| 1092 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1094 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 1097 | TC | TT | TA | GG | TC | CC | TT | |
| 1098 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1100 | TC | CT | GG | GG | TC | CC | TT | |
| 1102 | TT | TT | TA | GG | TT | CT | TC | |
| 1103 | TT | CT | GG | GG | TC | CT | TC | |
| 1104 | TT | TT | GA | GG | TC | CC | TT | |
| 1105 | TT | CT | GA | GG | CC | CC | TT | |
| 1106 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1108 | TC | TT | GA | GT | TT | CC | TT | |
| 1109 | TC | TT | GA | GT | TT | CT | TC | |
| 1111 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 1112 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1115 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 1117 | TC | TT | GG | GG | TT | CC | TT | TTGGTCT, CTGGTCT |
| 1119 | TC | TT | GA | GG | TT | CC | TT | |
| 1121 | TT | TT | GG | TT | TT | CT | TC | TTGTTCT, TTGTTTC |
| 1122 | TT | CC | GG | GG | TT | CC | TT | TCGGTCT, TCGGTCT |
| 1124 | TT | TT | GA | GG | TC | CC | TT | |
| 1125 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 1129 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 1131 | TC | CT | GG | GG | TC | CC | TT | |
| 1132 | TC | CT | GG | GG | CC | NN | NN | |
| 1134 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 1136 | TT | CT | GG | GG | CC | CC | TT | TTGGCCT, TCGGCCT |
| 1137 | TT | TT | GG | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 1139 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 1140 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1141 | TT | TT | GG | GG | TC | CT | TT | |
| 1142 | TC | TT | GG | GG | TC | CC | TT | |
| 1143 | TT | CC | GG | GG | TT | CC | TT | TCGGTCT, TCGGTCT |
| 1146 | TC | CT | GG | GG | CC | CC | TC | |
| 1150 | TC | TT | GA | GG | TC | CC | TT | |
| 1151 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 1153 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 1155 | TT | TT | GA | GG | TC | CC | TT | |
| 1156 | TC | CC | GG | NN | CC | CC | TT | TCGGCCT, CCGGCCT |
| 1157 | TC | CT | GG | GG | CC | NN | NN | |
| 1158 | TC | TT | GG | GG | CC | CT | TC | |
| 1161 | TC | CT | GA | GG | TT | CC | TT | |
| 1162 | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 1165 | TT | TT | GA | GG | TC | CC | TT | |
| 1166 | TT | CC | GA | GG | TT | CC | TT | TCGGTCT, TCAGTCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 1168 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 1169 | TC | CC | GA | GG | TC | CC | TT | |
| 1173 | TT | CC | GG | GG | TC | CC | TT | TCGGTCT, TCGGCCT |
| 1174 | TT | CT | GA | GG | TT | CC | TT | |
| 1175 | TC | TT | GG | GG | CC | CC | TT | TTGGCCT, CTGGCCT |
| 1176 | TT | TT | GG | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 1178 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 1179 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 1180 | TT | TT | GA | GT | TC | CT | TC | |
| 1182 | TC | TT | GA | GG | TT | CC | TT | |
| 1184 | TT | CT | GG | GG | TC | CC | TT | |
| 1185 | TC | TT | GG | GT | TT | CT | TT | |
| 1187 | TT | CT | GA | GG | TT | CC | TT | |
| 1188 | TC | TT | GA | GG | TC | CT | TT | |
| 1191 | TC | TT | GG | GT | TC | CT | TC | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 525-00002-8 | TT | CT | GG | NN | TC | NN | TT | |
| 525-00016-0 | TT | TT | GG | GT | TC | CC | TT | |
| 525-00017-1 | NN | TT | GG | GT | TT | CC | TT | TTGGTCT, TTGTTCT |
| 525-00029-8 | TC | CC | GA | GT | TT | NN | NN | |
| 525-00030-1 | TC | TT | GG | GG | TT | NN | TT | TTGGTCT, CTGGTCT |
| 525-00031-6 | TT | TT | GG | GG | TT | CC | TC | TTGGTCT, TTGGTCC |
| 525-00032-7 | TT | CT | GT | TT | TT | CC | TT | |
| 525-00034-3 | TT | CT | GG | GG | TC | CC | TT | |
| 536-00001-5 | TT | CT | GG | GG | TC | CC | TT | |
| 536-00003-1 | TT | TT | GA | GG | TC | CC | NN | |
| 536-00010-7 | TT | CT | GG | GT | TC | TC | NN | |
| 536-00012-1 | NN | CT | GG | NN | CC | CC | NN | TTGGCCT, TCGGCCT |
| 536-00014-9 | TT | CT | GA | GG | TC | NN | TT | |
| 536-00017-3 | TT | CT | GG | GG | TC | CC | TT | |
| 541-00003-6 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 541-00004-9 | TT | TT | GG | GG | TC | TC | TC | TTGGTTC, TTGGCCT |
| 541-00006-2 | TT | CT | GA | GG | CC | CC | TT | |
| 541-00008-5 | NN | TT | GG | NN | TT | NN | NN | |
| 541-00009-8 | TT | TT | GA | GT | TC | CC | TT | |
| 541-00011-5 | TT | TT | GA | GG | CC | CC | TT | TTGGCCT, TTAGCCT |
| 541-00013-1 | TT | TT | GA | GG | TC | TC | TT | |
| 541-00014-4 | TT | TT | GA | GG | TC | TC | TT | |
| 541-00017-8 | TT | CT | GG | GG | TC | TC | TT | |
| 541-00018-0 | TT | CT | GG | GT | TC | CC | TT | |
| 541-00019-3 | TT | CT | GG | GG | TC | CC | TC | |
| 541-00022-8 | TT | TT | GT | GT | TC | TC | TC | |
| 541-00023-4 | TT | TT | GG | GT | TC | NN | NN | |
| 541-00024-2 | TT | TT | GG | GG | TC | CC | NN | TTGGTCT, TTGGCCT |
| 541-00025-6 | TT | TT | GG | GT | TC | TT | NN | |
| 541-00035-3 | TT | TT | GG | GG | TC | NN | TT | TTGGTCT, TTGGCCT |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 615-000-015 | TT | CT | GG | GG | TC | CC | TT | |
| 615-00002-6 | TT | CT | GG | GG | TT | CC | TT | TTGGTCT, TCGGTCT |
| 615-00003-1 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 615-00004-4 | TT | CT | GG | GG | TC | TC | TC | |
| 615-00005-9 | TT | CT | GG | GT | TC | TC | TC | |
| 615-00006-7 | TT | TT | GG | TT | TC | CC | NN | TTGTTCT, TTGTCCT |
| 615-000-078 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 615-00008-0 | TT | CT | GA | GT | TC | CC | TT | |
| 615-00009-3 | TC | CT | GG | GT | TC | CC | TT | |
| 615-00010-7 | TT | TT | GT | GT | TC | TC | TC | |
| 615-00011-0 | TT | TT | GG | GG | TC | TC | TC | TTGGTTC, TTGGCCT |
| 615-00012-1 | TT | TT | GG | GG | TC | CC | TT | TTGGTCT, TTGGCCT |
| 615-00013-6 | TT | TT | GG | GG | CC | CC | TT | TTGGCCT, TTGGCCT |
| 615-00032-5 | TT | TT | GG | GG | TC | TC | TC | TTGGTTC, TTGGCCT |
| 615-00033-2 | TT | TT | TA | GT | TC | TC | TC | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 615-00034-0 | TT | TT | GG | GG | TT | CC | TT | TTGGTCT, TTGGTCT |
| 615-00035-8 | TT | TT | GG | GG | TT | CC | TC | TTGGTCT, TTGGTCC |
| 615-00036-6 | TT | CT | GA | NN | TC | CC | TT | |
| 615-000-377 | TT | TT | TT | TT | TC | TC | TC | TTTTTTC, TTTTCCT |
| 615-00038-1 | TT | TT | GA | GG | TC | CC | TT | |
| 615-00039-4 | TT | CT | GG | NN | TC | TC | TC | |
| 615-00040-6 | TT | TT | TA | GT | TC | TC | TC | |
| 616-00001-7 | TT | CT | TA | GG | TC | CC | TT | |
| 616-000-042 | TT | TT | GA | GG | TC | CC | TT | |
| 616-00009-1 | TT | CT | GA | GG | TC | NN | NN | |
| 616-00044-5 | TT | CT | GG | GG | TC | CC | TT | |
| 616-00045-3 | TC | CT | GG | NN | TT | CC | TT | |
| 616-00045-7 | TT | TT | GG | GG | TT | TC | TT | TTGGTCT, TTCCTTT |
| 616-00047-2 | TC | TT | GA | GT | TC | TC | NN | |
| 616-00048-6 | TC | TT | GA | GG | TC | CC | TT | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 616-000-499 | TT | TT | TA | GG | CC | CC | TT | TTAGCCT, TTTGCCT |
| 616-00050-3 | TT | CT | GG | GG | TC | CC | TT | |

Key:

NN No data

a Only unambiguous haplotypes are shown. No haplotype indicated where PHASE could not resolve the haplotype.

* General Population samples

## APPENDIX III: CALIBRATOR NORMLIZED RELATIVE QUANTITATIVE PCR DATA

| Sample | $C_t$ 1 | $C_t$ 2 | $C_t$ 3 | $C_t$ 4 | $C_t$ 5 | Ave $C_t$ | SD | $\Delta C_t$ | SD | $\Delta\Delta C_t$ | SD | RQ | RQ Min | RQ Max |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1036T | 32.84 | 33.02 | 32.78 | 32.65 | 32.93 | 32.85 | 0.14 | 12.79 | 1.01 | 2.14 | 1.01 | 0.23 | 0.46 | 0.11 |
| 1036R | 21.09 | 21.21 | 19.34 | 19.23 | 19.39 | 20.05 | 1.00 | | | | | | | |
| 1038T | 32.31 | 31.92 | 31.61 | 31.86 | 31.98 | 31.94 | 0.25 | 11.36 | 0.45 | 0.70 | 0.45 | 0.61 | 0.84 | 0.45 |
| 1038R | 20.80 | 20.54 | 20.55 | 21.00 | 19.99 | 20.58 | 0.38 | | | | | | | |
| 1057T | 30.48 | 30.66 | 30.84 | 31.20 | 30.49 | 30.73 | 0.30 | 10.90 | 0.69 | 0.24 | 0.69 | 0.85 | 1.36 | 0.52 |
| 1057R | 20.41 | 20.59 | 19.43 | 19.19 | 19.56 | 19.84 | 0.62 | | | | | | | |
| 1066T | 34.62 | 35.25 | 33.86 | 34.24 | 33.74 | 34.34 | 0.61 | 11.71 | 0.86 | 1.06 | 0.86 | 0.48 | 0.87 | 0.26 |
| 1066R | 23.48 | 22.64 | 21.87 | 22.29 | 22.85 | 22.63 | 0.60 | | | | | | | |
| 1081T | 33.15 | 32.90 | 32.62 | 32.84 | 33.14 | 32.93 | 0.22 | 11.90 | 0.38 | 1.25 | 0.38 | 0.42 | 0.55 | 0.32 |
| 1081R | 21.44 | 21.21 | 20.68 | 21.04 | 20.77 | 21.03 | 0.31 | | | | | | | |
| 1087T | 34.64 | 34.24 | 34.85 | 33.60 | 33.81 | 34.23 | 0.53 | 12.74 | 0.82 | 2.09 | 0.82 | 0.24 | 0.41 | 0.13 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1087R | 22.13 | 22.18 | 21.12 | 21.07 | 20.91 | 21.48 | 0.62 | | | | | | | |
| 1106T | 34.26 | 34.93 | 33.40 | 34.15 | 34.19 | 34.18 | 0.54 | 12.16 | 0.77 | 1.51 | 0.77 | 0.35 | 0.60 | 0.21 |
| 1106R | 22.31 | 22.68 | 21.95 | 21.98 | 21.19 | 22.02 | 0.55 | | | | | | | |
| 1129T | 33.68 | 33.85 | 32.61 | 32.94 | 32.85 | 33.19 | 0.55 | 12.38 | 0.89 | 1.73 | 0.89 | 0.30 | 0.56 | 0.16 |
| 1129R | 21.77 | 21.24 | 20.16 | 20.67 | 20.18 | 20.80 | 0.70 | | | | | | | |
| 1137T | 32.39 | 31.72 | 31.57 | 31.85 | 31.24 | 31.75 | 0.42 | 10.71 | 0.8 | 0.06 | 0.8 | 0.96 | 1.67 | 0.55 |
| 1137R | 21.93 | 21.39 | 21.01 | 20.71 | 20.14 | 21.04 | 0.68 | | | | | | | |
| 1141T | 33.83 | 34.10 | 34.55 | 33.37 | 33.48 | 33.87 | 0.48 | 12.67 | 0.75 | 2.02 | 0.75 | 0.25 | 0.42 | 0.15 |
| 1141R | 21.40 | 22.05 | 20.95 | 20.51 | 21.06 | 21.20 | 0.57 | | | | | | | |
| 1150T | 34.62 | 33.18 | 33.59 | | 34.17 | 33.89 | 0.64 | 12.09 | 1.09 | 1.43 | 1.09 | 0.37 | 0.79 | 0.17 |
| 1150R | 23.27 | 21.93 | 21.01 | 21.40 | 21.39 | 21.80 | 0.88 | | | | | | | |
| 1153T | 32.01 | 31.44 | 31.25 | 31.32 | 31.53 | 31.51 | 0.30 | 11.21 | 0.73 | 0.56 | 0.73 | 0.68 | 1.13 | 0.41 |
| 1153R | 21.08 | 20.92 | 19.50 | 20.05 | 19.96 | 20.30 | 0.67 | | | | | | | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1156T | 31.32 | 31.07 | 31.04 | 30.77 | 30.98 | 31.04 | 0.20 | 10.58 | 0.39 | -0.07 | 0.39 | 1.05 | 1.38 | 0.80 |
| 1156R | 20.09 | 20.16 | 20.65 | 20.88 | 20.50 | 20.45 | 0.33 | | | | | | | |
| 1158T | 33.12 | 32.82 | 32.61 | 32.66 | 32.70 | 32.78 | 0.20 | 10.87 | 0.32 | 0.22 | 0.32 | 0.86 | 1.07 | 0.69 |
| 1158R | 22.17 | 21.75 | 21.64 | 21.79 | 22.18 | 21.91 | 0.25 | | | | | | | |
| 1169T | 32.06 | 32.04 | 31.74 | 32.02 | 32.02 | 31.98 | 0.13 | 10.89 | 0.28 | 0.23 | 0.28 | 0.85 | 1.03 | 0.70 |
| 1169R | 21.26 | 20.88 | 20.89 | 20.98 | 21.44 | 21.09 | 0.25 | | | | | | | |
| 1173T | 31.93 | 31.25 | 31.53 | 31.76 | 31.84 | 31.66 | 0.28 | 11.02 | 0.59 | 0.36 | 0.59 | 0.78 | 1.17 | 0.52 |
| 1173R | 20.36 | 19.92 | 20.78 | 20.92 | 21.25 | 20.65 | 0.52 | | | | | | | |
| 1174T | 34.94 | 34.45 | 34.21 | 34.69 | 34.90 | 34.64 | 0.31 | 11.40 | 0.74 | 0.75 | 0.74 | 0.59 | 0.99 | 0.36 |
| 1174R | 22.44 | 23.92 | 22.63 | 23.38 | 23.80 | 23.23 | 0.67 | | | | | | | |
| 1175T | 36.96 | 34.61 | 33.24 | 34.25 | 34.43 | 34.70 | 1.37 | 11.93 | 1.39 | 1.27 | 1.39 | 0.41 | 1.08 | 0.16 |
| 1175R | 22.55 | 22.96 | 22.74 | 22.77 | 22.84 | 22.77 | 0.15 | | | | | | | |
| 1176T | 33.02 | 32.93 | 32.62 | 32.88 | 32.91 | 32.87 | 0.15 | 11.14 | 0.28 | 0.48 | 0.28 | 0.71 | 0.87 | 0.59 |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1176R | 21.70 | 21.50 | 21.61 | 21.73 | 22.13 | 21.73 | 0.24 | | | | | | | |
| 1178T | 30.58 | 30.25 | 30.37 | 30.41 | 30.43 | 30.41 | 0.12 | 10.15 | 0.32 | -0.51 | 0.32 | 1.42 | 1.77 | 1.14 |
| 1178R | 20.70 | 19.93 | 20.02 | 20.31 | 20.32 | 20.26 | 0.30 | | | | | | | |
| 1179T | 35.46 | 34.95 | 34.53 | 34.91 | 35.40 | 35.05 | 0.38 | 12.96 | 0.48 | 2.31 | 0.48 | 0.20 | 0.28 | 0.14 |
| 1179R | 22.48 | 22.12 | 21.72 | 21.89 | 22.23 | 22.09 | 0.30 | | | | | | | |
| 1180T | 31.92 | 31.25 | 30.83 | 30.96 | 30.96 | 31.18 | 0.44 | 9.93 | 1.08 | -0.73 | 1.08 | 1.66 | 3.50 | 0.78 |
| 1180R | 21.17 | 22.92 | 20.37 | 20.62 | 21.21 | 21.26 | 0.99 | | | | | | | |
| 1182T | 31.89 | 31.95 | 31.87 | 31.93 | 32.33 | 31.99 | 0.19 | 10.87 | 0.33 | 0.22 | 0.33 | 0.86 | 1.08 | 0.69 |
| 1182R | 21.44 | 20.99 | 20.87 | 20.94 | 21.39 | 21.12 | 0.27 | | | | | | | |
| 1184T | 30.35 | 30.65 | 30.27 | 30.51 | 30.72 | 30.50 | 0.19 | 10.87 | 0.22 | 0.21 | 0.22 | 0.86 | 1.01 | 0.74 |
| 1184R | 19.59 | 19.50 | 19.62 | 19.63 | 19.83 | 19.63 | 0.12 | | | | | | | |
| 1185T | 32.07 | 31.62 | 31.51 | 31.60 | 31.82 | 31.72 | 0.22 | 10.44 | 0.28 | -0.21 | 0.28 | 1.16 | 1.41 | 0.95 |
| 1185R | 21.36 | 21.16 | 21.05 | 21.46 | 21.38 | 21.28 | 0.17 | | | | | | | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1191T | 33.91 | 33.85 | 31.50 | 33.55 | 33.56 | 33.28 | 1.01 | 10.68 | 1.04 | 0.03 | 1.04 | 0.98 | 2.02 | 0.48 |
| 1191R | 22.29 | 22.89 | 22.46 | 22.54 | 22.79 | 22.59 | 0.25 | | | | | | | |
| CalibratorT | 30.13 | 30.15 | 30.11 | 30.15 | 30.26 | 30.16 | 0.06 | 10.66 | 0.48 | 0.00 | 0.48 | 1.00 | | |
| CalibratorR | 20.27 | 19.04 | 19.23 | 19.35 | 19.65 | 19.51 | 0.48 | | | | | | | |

## APPENDIX IV: ETHICS CLEARANCE CERTIFICATE FOR 2006 AND 2009 COLLECTIONS

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

**UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG**

Division of the Deputy Registrar (Research)

**HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)**
R14/49 Dandara

| | |
|---|---|
| **CLEARANCE CERTIFICATE** | **PROTOCOL NUMBER M080124** |
| **PROJECT** | Pharmacogenetics profiling: establishing the baseline frequencies of variants of drug or xenobiotic metabolizing..... |
| **INVESTIGATORS** | Dr C Dandara |
| **DEPARTMENT** | Molecular & Cellular Biology |
| **DATE CONSIDERED** | 20080125 |
| **DECISION OF THE COMMITTEE*** | Approved unconditionally |

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

**DATE** 08.02.05 **CHAIRPERSON** .....
(Professor P E Cleaton Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Not applicable

---

**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 V: PERMISSION LETTER FROM HELEN JOSEPH HOSPITAL TO CONDUCT SAMPLE COLLECTION IN 2009

Gauteng Department of Health

Helen Joseph Hospital

PERMISSION FOR RESEARCH

DATE: 04 September 2008

NAME OF RESEARCH WORKER: Dr C Dandara

CONTACT DETAILS OF RESEARCHER (INCLUDE ALTERNATE RESEARHER):
School of Molecular and Cell Biology
University of the Witwatersrand
Tel: 011 717 6366 / 084 995 5010

TITLE OF RESEARCH PROJECT The role of pharmacogenetics on the response to treatment using antiretroviral drugs in South Africa

OBJECTIVES OF STUDY (Briefly or include a protocol):
The effectiveness of most drugs in use is affected by metabolism by a group enzymes commonly referred to as drug metabolizing enzymes. Unfortunately, this metabolism of these differs between different people and also between different ethnic groups due to a phenomenon referred to as "genetic polymorphisms". In addition, most drugs that are used for example antiretroviral drugs, anti-malarial drugs, and anti-tuberculosis drugs, would have been developed after being tested in other populations that do not have the same genetic polymorphisms as our populations. So the main objective to determine the role of pharmacogenetics in treatment for HIV and TB in the South African context by evaluating the effects of carrying genetic variants of CYP2B6, CYP3A4/, ABCB1 (MDR1), PXR and NAT2, whose gene products are involved in the metabolism of commonly used antiretroviral drugs.

METHODOLOGY (Briefly or include a protocol):
Prospective patients who are able to satisfy all the required criteria will be recruited. Finding the correct sample size is important because if too many samples are taken, this is a waste of resources while collection of too few samples may result in false positive results. We plan to recruit 600 patient volunteers on ARV's.

CONFIDENTIALITY OF PATIENTS MAINTAINED: Yes

COSTS TO THE HOSPITAL: NIL

APPROVAL OF HEAD OF DEPARTMENT:

APPROVAL OF CRHS OF WITS UNIVERSITY: Yes

SUPERINTENDENT PERMISSION:

Signature: Date: 22/09/08

Subject to any restrictions: 180 COPIES TO THE HOSPITAL

## APPENDIX VI: PATIENT INFORMATION AND CONSENT FORM FOR THE 2006 AND 2009 COLLECTIONS

### PATIENT INFORMATION & CONSENT FORM 2006

"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 INFORMATION AND CONSENT FORM 2009

## INFORMATION LEAFLET

STUDY TITLE: **Role of pharmacogenetics on the response to treatment using antiretroviral drugs in South Africa**

INVESTIGATOR: Dr Collet Dandara

INSTITUTION: University of Witwatersrand, East Campus Braamfontein, Johannesburg.

**Introduction**

Hello, I am _____*, a researcher in the School of Molecular and Cell Biology, University of Witwatersrand. I am doing a research as part of contribution to improving HIV medication and invite you to participate in this research project. I would be very grateful for your participation. Before agreeing to participate, it is important that you read the following explanations of the purpose of the study, the study procedures, benefits, risks, discomforts, and precautions as well as your right to withdraw from the study at any time. This information leaflet is to help you to decide if you would like to participate. You should fully understand what is involved before you agree to participate.

**Purpose of the study**

You have been diagnosed as suffering from infection by human immunodeficiency virus (HIV) and you have been put on treatment with antiretroviral drugs. Like any other treatment, some patients respond well while others do not respond so well. The aim of the study is to determine the role of what we inherit from our parents in the differences in response to antiretroviral medication. If you decide to participate, you will not be asked to take any additional drugs except what your doctors have prescribed for your condition. However, you will be monitored for six months on your response to treatment.

**Procedures**

If you agree to participate, your medical information including demographic information will be accessed from the clinic files and you will be required to complete two questionnaires, one asking for information on your ethnicity and the second questionnaire evaluating your adherence to treatment.

A qualified nurse will draw 5 ml (one teaspoon) from you for the extraction of DNA that will be used for genetic analysis and another 2.5 mL of blood for plasma drug analysis. Two hours later we will draw an additional 2.5 mL of blood for drug analysis. Possible side effects which may be associated with obtaining a blood sample include pain, bruising, light-headedness and on rare occasions infection. Precautions will be taken to avoid these difficulties. The entire procedure should take approximately 15 minutes.

**Sample collection schedule**

After consenting to participate in the study, the following schedule will be followed for collection of study samples from you.

| Time | |
|---|---|
| Interview | • 5 mL (one teaspoon) blood sample for genetic analysis<br>• 2.5 mL (half-teaspoon) blood sample for drug analysis |
| 2 hours after interview | • 2.5 mL (half-teaspoon) blood sample 2 hours after a previous sample for drug analysis |

**Unforeseen risks and potential benefits**
You will continue on your regular medication, you should however contact your prescribing doctor immediately if you experience any side effects. There are no direct benefits for you arising from participating in this study, however, the findings may help future patients in the choice and doses of their medication.

**Rights as a participant in the study:** It is up to you to decide whether or not to take part in this study. Your participation is voluntary. If you decide to take part you are still free withdraw from the study at any time and without giving reason. This will not affect the standard of care that you receive or the relationship you have with your doctor. You are free to refuse to participate or withdraw your consent at any time. The results of the tests will be made available to you on your request. The results will be strictly confidential and if the data are published your name will not be used. There will be no costs to you from the study and there are no risks attached to this study.

**Ethical approval:** *This clinical study protocol has been submitted to the University of Witwatersrand, Human Research Ethics Committee (HREC) and written approval has been granted by that committee.*

**Source of additional information:** For the duration of the study you are still under the care of your prescribing doctor whom you should contact at any time should you feel that any of your symptoms are causing you problems. For questions relating to the study you can contact me at any time on the following cellular number: 084 9955010

If you agree to participate, please sign the consent form.

Thank you

Dr Collet Dandara

# INFORMED CONSENT

**Participant number:** _____

I hereby confirm that *_____ has explained to me about the nature, conduct, benefits and risks of the study "**Role of pharmacogenetics on the response to treatment using antiretroviral drugs**". I have also received, read and understood the above written information (patient information leaflet and informed consent) regarding the study. I am aware that the results of the study, including personal details regarding my sex, age, ethnicity, marital status and medical information will be anonymously processed into a study report. I may at any stage, without prejudice, withdraw my consent and participation in the study. I have had sufficient opportunity to ask questions and (of my own free will) declare myself prepared to participate in the study. I am **willing/not willing** (delete the inappropriate) to be contacted for follow-up studies.

_____

Name and Surname (Capital letters) Signature

Date: _____

**Study doctor and person administering the consent**

I, _____ ***(person administering consent form)***,
herewith confirm that the above patient has been fully informed about the nature, conduct and risks of the study

_____

Name (Capital letters) Signature Date

WITNESS

_____

Name (Capital letters) Signature Date

**Name of member of the researcher team who will be administering the informed consent.*

## INFORMED CONSENT FOR LEGAL GUARDIANS

**(Applicable where participants are incapable of giving consent).**

**Participant number:** _____

I, the undersigned, *_____, has fully explained the information leaflet to the patient,

_____

and/or his/her legal guardian_____

The account I have given explained both the possible risks and benefits of the study. The patient and/or his/her legal guardian understand these. The patient and/or his/her legal guardian indicated that he/she understands that the patient will be free to withdraw from the study at any time for any reason and without jeopardizing his/ her subsequent treatment.

I hereby certify that, the patient and/or his/her legal guardian acting on his/her behalf has agreed to participate. I am **willing/not willing** (delete the inappropriate) to be contacted for follow-up studies

PATIENT

_____

Name and Surname (Capital letters) Date

STUDY DOCTOR AND PERSON ADMINISTERING THE CONSENT

_____

Name (Capital letters) Signature Date

LEGAL GUARDIAN

_____

Name (Capital letters) Signature Date

WITNESS

_____

Name (Capital letters) Signature Date

**Name of member of the researcher team who will be administering the informed consent.*

## APPENDIX VII: PATIENT QUESTIONNAIRE FOR THE 2006 AND 2009 COLLECTIONS

# PATIENT QUESTIONNAIRE 2006

"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** | **Your father** |
|---|---|
| Place of birth _____ | Place of birth _____ |
| Province or country _____ | Province or country_____ |
| Home language _____ | Home language _____ |

**YOUR GRANDPARENTS**

| **Your mother's mother** | **Your father's mother** |
|---|---|
| Place of birth _____ | Place of birth _____ |
| Province or country _____ | Province or country_____ |
| Home language _____ | Home language _____ |

| **Your mother's father** | **Your father's father** |
|---|---|
| Place of birth _____ | Place of birth _____ |
| Province or country_____ | Province or country_____ |
| Home language _____ | Home language _____ |

**MEDICAL INFORMATION**

Year of birth _____

HIV + since _____

Most recent CD4+ count _____ Date _____

Other conditions: _____

Tuberculosis?_____

**PATIENT QUESTIONNAIRE 2009**

**APPENDIX A:** **ETHNICITY QUESTIONNAIRE**

"Role of pharmacogenetics on the response to treatment using antiretroviral drugs"

**Participant number:** _____

We need information about you, your parents and your grand parents in order to determine if there is any genetic difference in the metabolism of drugs between different Southern African subpopulations. Please try to be as accurate as possible. The quality of data will depend on the truth of your response to the questions below. It is better to leave out information than to give us something that might be wrong.

**1. ABOUT YOUSELF**

Gender _____ Date of birth: _____

Place of birth _____ Province or country_____

Home language _____ _____

Do you smoke tobacco? Yes / no (circle the appropriate) Do you consume alcohol? Yes / no (circle the appropriate)

**2. ABOUT YOUR PARENTS**

**Your mother**

Place of birth _____

Province or country_____

Home language _____

**Your father**

Place of birth _____

Province or country_____

Home language _____

**3. ABOUT 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 _____

Appendix B: **ARV TREATMENT ADHERENCE QUESTIONNAIRE**

The questions that follow are to help us understand the levels of the medicines that we find in your blood today. The medicines you have taken during the past few days, and the times that you took them can affect the levels of the antiretroviral (ARV) medicines that we are measuring today. We understand that it is very difficult to never miss a dose, and that it is difficult to remember when you have missed as dose. But this information is important, so please make an effort to remember, and be as accurate as possible; take the time you need.

| **Participant number** | | | | | | | **Today's date:** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|

Q1: a) What ARV medicine did you take yesterday?
b) How many tablets did you take?
c) What time did you take this medication?

Q2: a) What ARV medicine did you take 2 days ago?
b) How many tablets did you take?
c) What time did you take this medication?

Q3: a) What ARV medicine did you take 3 days ago?
b) How many tablets did you take?
c) What time did you take this medication?

List of the ARVs you have taken. Next to the medication name, list the number of tablets that you took each day and the time(s) that you took them: yesterday, 2 days ago and 3 days ago.
* Use the generic name, or the trade name if you prefer.

| Name of ARV medicine* | YESTERDAY | | 2 DAYS AGO | | 3 DAYS AGO | |
|---|---|---|---|---|---|---|
| | Number of tablets | time | Number of tablets | time | Number of tablets | time |
| e.g. | 5 | | 0 | | 5 | |
| 1. | | | | | | |
| 2. | | | | | | |
| 3. | | | | | | |
| 4. | | | | | | |

Q4: During the past 3 days, on how many days have you missed taking your ARV medicines?

| none | 1 | 2 | 3 |
|---|---|---|---|

Q5: What other medications including complimentary medicine have you taken in the last 3 days. Next to the medication name, list the number of tablets in each dose and the time(s) that you took them: yesterday, 2 days ago and 3 days ago.
* Use the generic name, or the trade name if you prefer.

| Name of medicine* | YESTERDAY | | 2 DAYS AGO | | 3 DAYS AGO | |
|---|---|---|---|---|---|---|
| | Number of tablets in each dose | Number of Doses/day | Number of tablets in each dose | Number of Doses/day | Number of tablets in each dose | Number of Doses/day |
| e.g. Vit B | 1 | 2 | 2 | 3 | 0 | 0 |
| 1. | | | | | | |
| 2. | | | | | | |
| 3. | | | | | | |
| 4. | | | | | | |
| 5. | | | | | | |
| 6 | | | | | | |
| 7. | | | | | | |
| 8. | | | | | | |
| 9. | | | | | | |
| 10. | | | | | | |

## APPENDIX VIII: ETHICS CLEARANCE CERTIFICATE FOR USE OF THE DATABASE

**UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG**

Division of the Deputy Registrar (Research)

**HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)**
R14/49 Heitkamp

| | |
|---|---|
| **CLEARANCE CERTIFICATE** | **PROTOCOL NUMBER M081116** |
| **PROJECT** | The Effect of MDR1 Variation of Initial Immune Recovery with Antiretrovirals |
| **INVESTIGATORS** | Miss I Heitkamp |
| **DEPARTMENT** | Molecular & cell Biology |
| **DATE CONSIDERED** | 08.11.28 |
| **DECISION OF THE COMMITTEE*** | |

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

**DATE** **CHAIRPERSON** .....
(Professor P E 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 10004, 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 IX: PERMISSION LETTER FROM HELEN JOSEPH HOSPITAL FOR USE OF THE DATABASE

University
of the Witwatersrand,
Johannesburg

Clinical HIV Research Unit

SECRETARIAT: Suite 176, Private Bag x2600, Houghton 2041, South Africa • Tel: +27-11-276-8800 • Fax: +27-11-482-2130

Date: **18/12/2008**

This is to certify that **Ingrid Heitkamp** has been granted permission to use data from the Themba Lethu Clinic cohort stored on TherapyEdge-HIV™ for a research project titled:

**The effects of MDR1 variation on initial immune recovery with antiretrovirals**

The standard operating procedures for the Clinical HIV Research Unit have been complied with.

Dr Mhairi Maskew

Professor AP MacPhail.

# APPENDIX X: SOP FOR RESEARCH AT THEMBA LETHU CLINIC



## Standard Operating Procedure: Process to be followed for granting access to the Themba Lethu HIV/AIDS Comprehensive Care Management and Treatment and TB site for the purposes of research

Requests from individuals wanting to make use of the facilities at the Themba Lethu Clinic for the purposes of research must follow the following procedure and sign this document:

1. A written request must be sent to the Regulatory Manager at the Clinical HIV Research Unit, Department of Medicine, Helen Joseph Hospital (Marlene Naidoo; manaidoo@witshealth.co.za).
   - 1.1. The request should include:
     - 1.1.1. Names of individuals who will be conducting the research and their affiliations.
     - 1.1.2. Information on whether the research is for degree purposes and details of the institution that will grant the degree.
     - 1.1.3. A proposal detailing the objectives of the research, how the research will be conducted, methods of data collection and analysis and public distribution of results.
     - 1.1.4. A schedule indicating the proposed time-line for the start and duration of the study.
   - 1.2. The Regulatory Manager will communicate the requests to Dr Ian Sanne, researchers in the Clinical HIV Research Unit and the Head of the Themba Lethu Clinic.
     - 1.2.1. A copy of the proposal will be submitted to the Head of the Department of Medicine at Helen Joseph Hospital.

2. A consensus decision of the Clinical HIV Research Unit and the Department of Medicine at Helen Joseph Hospital on whether to grant access to the clinic will be required in all cases.
   - 2.1. The Regulatory Manager of the CHRU will ensure that the relevant documentation required by the Helen Joseph Hospital is completed and submitted to the office of the CEO of Helen Joseph Hospital for approval.
   - 2.2. Once permission has been agreed by the above groups, the decision will be communicated to the applicant by the Regulatory Manager of the CHRU.

3. By signing this SOP the applicant agrees to the following statements of assurance:
   - 3.1. All references to the data either in public verbal presentation or in print must credit the Clinical HIV Research Unit, Right to Care and the Department of Medicine at Helen Joseph Hospital.

3.2. Should the applicant wish to use the data for analysis beyond the originally submitted proposal, a further written request will be required and the procedures outlined above followed.

3.3. A final draft of the results of the research will be submitted to the CHRU for information and comment preferably before it is submitted for publication or public presentation.

3.4. The authorship of papers or presentations arising from the research will be decided using internationally recognized criteria, and must recognize the CHRU and Department of Medicine researchers appropriately, as well as any international researcher/s involved.

3.5. The applicant will also be required to submit a copy of the final product resulting from use of the data set to the Regulatory Manager at the CHRU and to the Head of the Department of Medicine at Helen Joseph Hospital.

3.6. The data derived from the study will not be shared, copied or provided to anyone other than the person/s outlined in the proposal.

4. Where the research involves the analysis of data in the Themba Lethu Clinical Cohort stored on TherapyEdge in addition to the clinical study outlined in 1 above, the requirements of the SOP "Process to be followed for granting access to data for research on the Themba Lethu Clinical Cohort and other Cohorts stored on the TherapyEdge database" shall be followed in addition to this SOP.

5. All research at Themba Lethu Clinic must have the formal written approval of the Human Research Ethics Committee (Medical) of the University of the Witwatersrand.

5.1. Copies of documents submitted for the purpose of obtaining approval from the above committee must be lodged with the Regulatory Manager at CHRU prior to the start of any study.

5.2. A copy of the formal approval of the protocol from the Human Research Ethics Committee (Medical) of the University of the Witwatersrand must be lodged with the Regulatory Manager at CHRU prior to the start of any study.

5.3. The applicant will be responsible for obtaining approval from any other authorities or Internal Review Boards as may be necessary.

6. If possible, the applicant should make an oral presentation to the Clinical HIV Research Unit at the start of the investigation and again once it has been concluded. This will be part of the regular academic programme of the Clinical HIV Research Unit.

7. A file will be maintained in the offices of the Clinical HIV Research Unit for all correspondence in this regard. In particular:

   7.1. Correspondence documenting the approval process as out lined above

   7.2. The signed agreement of the applicant (this SOP).

   7.3. A copy of the final product resulting from use of the data.

8. Requirements for authors in the CHRU

   - The first author is decided by the person who developed the idea and that first author must write the majority of the first draft of the paper.
   - First, second and senior author to decide who else to invite to be an author
   - Invited authors have 2 weeks to respond to an invitation unless there are extenuating circumstances
   - To be included in the authorship ALL authors must be actively involved in writing the paper or AT LEAST provide a critical review* of the final draft of the paper PLUS fulfil 2 of the following criteria

| | |
|---|---|
| CONCEPT | Idea for research or article framing the hypothesis |
| DESIGN | Planning methods to generate results |
| SUPERVISION | Oversight & responsibility for organisation, course of project and manuscript |
| RESOURCES AND MATERIAL | Money, equipment, space, personnel Biological materials, reagents, patients |
| DATA COLLECTION/ANALYSIS | Doing experiments, managing patients, organising or reporting data |
| ANALYSIS/INTERPRETATION | Making sense of & presenting results |
| LITERATURE SEARCH | Responsibility for this function |
| WRITING | Creating all or most of the manuscript |
| OTHER | Novel contributions |

*Critical review is defined as

- Input into the content of the paper in terms of the design, methodology, writing style, proofing etc.
- Providing critical input and editing of the final manuscript.

I have read and accept these conditions.

Applicant: [signature]

Date: 16 MARCH 2009.