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

## Corrections as per examiners report

**Roshilla Ramdin**

**9902650m**

**Johannesburg, August 2012**

Query 1: Sample size to small therefore could not reach definitive conclusions:

Response 1: Excluded all analyses on disease association and allele frequencies from the results as sample size is small, the data are very weak. Removed Table 3.4.1 The genotype frequencies at each of the polymorphic positions, in both the general population and $HIV^+$ samples. In addition that is not what we wanted to achieve in this study. (Page 5 and page 38 on the detailed corrections file)

Query 2: No clear relationship between spoke language and susceptibility

Response 2: The main reason to give the data about language group is that my sample is representative of South African black population. Language groups are a good indication of genetic affinities, representing the two major Bantu migrations into the area. They are better than geographic origin in some parts of SA, because different groups live in the same areas. Thus gave more recent references of this in the introduction (page 50 of detailed corrections)

Query 3: The work seems to be a mix of genetics, molecular biology, and anthropology. None given the in dept discussion they deserve.

Response 3: I have streamlined the thesis to include only the following major areas in the results (defined on pages 47-65 of Chapter 3 of new thesis).

1--Detection of variation and development of genotyping methods

2--Population frequencies in the overall SA black population

Rather than do analyses where I end up saying the sample size was too small for, i have stuck with the two main points mentioned above only. I got comparative data on population frequency from various databases on genomics (pages 30- 39 of the detailed corrections).

This way, I have stuck with the population genetics focus of the work and did not venture into disease association.

Query 3: Intro is lengthy

Response: the introduction has been edited and is shorten to 28 pages from the original 31 pages. The intro is more focussed and unnecessary information has been deleted (pages 10-25 on detailed corrections).

Query 4: List of abbreviations should be carefully edited

Response 4: The list of abbreviation was edited. The commonly known ones were removed (pages 6-8 of detailed corrections)

Query 5: Repetition of known concepts in Materials and Methods Sections

Response 5: Both have been edited so that familiar concepts have been removed (pages 25-30). In addition this allowed that the sections more focussed on the population genetics of the work.

Query 6: Lack of synthesis

Response 6: I have streamlined the thesis in the introduction. In addition i have excluded analyses where the data is weak and focussed more on the population genetics of the work.

Changes were made as per requests by examiner so that thesis was grammatically correct and more concise. In addition there were changes made to the content on the thesis. This helped put my work in context of greater population data which already has been reported.

## Title page:

1. Capitalised all nouns and verbs in the title.

2. Date on the tiles change has been changed from April, 2010 to August, 2012.

## ACKNOWLEDGMENTS:

1. Added ACKNOWLEDGMENTS page iii after the dedications page. The acknowledgment reads **"**I would like to acknowledge the NRF for the funding they have provided for my studies."

## Table of Contents:

1. The table of contents page is now moved to page iv.

2. Table of contents follows from page iv to page vi.

3. Added Acknowledgments page.

4. Deleted the sections 1.4.1.1 Chemokine Receptor CCR5 and 1.4.1.2 Human Leukocyte Antigen (HLA) system.

5. Deleted the sections 3.4.1 Differences between general populations and the HIV + sample groups.

## List of Figures:

1. Deleted Figure 1.2 Schematic diagram of HIV-1 groups. The explanatory section has been removed.

2. Deleted Figure 2.4.2 Diagrammatic representation of PCR Heat block.

3. Deleted Figure 3.1.1.2 Chromatograms showing a sequencing artefact within the upstream non-coding region. This is not essential to thesis.

4. Deleted Figure 3.2.2.1 PCR optimization amplification of exon 4

5. Caption of Figure 2.5 on page vii changed from "pyrogram of variation at two positions" changed to "Pyrogram of variation at codon positions 185 and 186 in exon 4 of Apobec 3G

6. Renamed figures; Figure 1.3 changed to Figure 1.1, Figure 1.5 changed to Figure1.2, Figure 1.6.1 changed to Figure 1.3, Figure 1.8 changed to Figure 1.4, Figure 2.4.3 changed to Figure 2.1, Figure 2.4.3.1 changed to Figure 2.2, Figure 2.4.3.2.1 changed to Figure 2.3, Figure 2.4.3.2.2 changed to Figure 2.4, Figure 2.4.3.2.3 changed to Figure 2.5, Figure 2.5.3.1 changed to Figure 2.6, Figure 3.1.1.1 changed to Figure 3.1, Figure 3.1.1.3 changed to Figure 3.2, Figure 3.2.1 changed to Figure 3.3, Figure 3.2.2.1 changed to Figure 3.5, Figure 3.2.2.2 changed to Figure 3.4, Figure 3.2.3.2 changed to Figure 3.6

## List of Tables:

1. Added Table 3.1 Sequenced data from 2003 and comparative data of other populations retrieved from Ensembl on page 50. This was added as a table of the sites found to be polymorphic, what the bases are, numbers of each genotype, and comparative data from a few other populations would put my data in context to other populations.

2. Deleted Table 3.4.1 The genotype frequencies at each of the four polymorphic positions, in the general population and HIV + samples.

3. Added Table 3.4.1 The estimated haplotype frequencies in JHB, GP and HJ sample sets generated by genotyping data from RFLP and Pyrosequencing genotyping assays in this study, as calculated using PHASE 2.1. (Stephens *et al.*, 2001) on page 66.

4. Deleted Table 4.1 Comparison of population frequencies across three populations groups

5. Renamed Tables; Table 3.5 changed to Table 3.4.2, Table 3.4.3 changed to Table 3.3.3.1, Table 3.4.2 changed to Table 3.3.3, Table 3.1 changed to Table 3.3.2, Table 3.2 changed to Table 3.3.1,

## Abbreviations:

1. Edited the list of abbreviations on pages ix-x to exclude the most common ones that are known. Deleted the following :

- AIDS acquired immune deficiency syndrome
- ATP adenosine triphosphate
- bp base pair
- Cl chlorine
- cm centimeter
- DNA deoxyribonucleic acid
- dNTP dinucleotide triphosphate
- EDTA ethylene diamine tetra-acetic acid
- EtBr ethidium bromide

- g gram
- HCl hydrochloric acid
- HIV human immunodeficiency virus
- kb kilobase
- kDa kilodalton
- Mg magnesium
- $MgCl_2$ magnesium chloride
- mg milligram
- min minutes
- ml millilitre
- mM millimolar
- mRNA messenger RNA
- mtDNA mitochondrial DNA
- NADPH nicotinamide adenine dinucleotide phoshosphate
- NaCl sodium chloride
- NaOH sodium hydroxide
- ng nanogram
- PCR polymerase chain reaction

- RFLP restriction fragment length polymorphism
- RNA ribonucleic acid
- SDS sodium dodecyl sulphate
- soln solution
- Tris tris(hydroxymethyl)aminomethane
- µg microgram
- µl microlitre
- µM micromolar
- V volt

## Abstract:

1. The following paragraph on page xii was removed "Differences in allele and genotype frequencies were also seen between $HIV^+$ individuals and the general population. Thus variation within *APOBEC3G* may well influence an individual's susceptibility to HIV-1 infectivity and/or rate of disease progression but better characterised samples are needed to draw definitive conclusions."

2. The abstract was then edited as follows from "It is known that infectious agents elicit different responses in different individuals which strengthens the view that susceptibility and resistance to infectious diseases has a genetic component. These differences in susceptibility to disease can be observed in populations. *APOBEC3G*

is a member of the cytidine deaminase gene family located on chromosome 22. It is crucial in non-permissive cells as it functions as part of the innate immunity system and is an inhibitor of the HIV-1 accessory protein vif. The goal of the study was to develop genotyping assays and estimate allele frequencies. Thus, genetic variation within *APOBEC3G* was identified and characterized in black South Africans. Indirect genotyping assays were designed to amplify regions within the upstream non-coding region, and in exon 4 of the coding region of the gene. Selected polymorphisms were then genotyped using allele-specific PCR, RFLP-PCR and Pyrosequencing™ assays. Reanalysis of sequence data from 2003 showed numerous SNPs were well represented. Comparison of sequence data at various SNPs showed that allele frequencies were similar to frequencies in other African populations. The only sequenced SNP that deviated from the frequencies in Ensembl was -590. Thus the sequencing was a useful tool for detection of variation. ASA proved to be the least reliable genotyping technique as the minor allele frequency of -571 (0.59) deviated from the published frequency of 0.894 in Africans. RFLP analysis proved more reliable for genotyping -571 and H186R. The minor allele frequency was estimated to be 0.84 and 0.32 for -571 and H186R respectively. The frequency of H186R is similar to published data from An et al (2004) and Reddy et al (2010). If SNPs are in LD they occur together on the same haplotype more often than by chance. Usually SNPs that are in LD are in close proximity. However our data suggests -571 and H186R SNPs which are 5kb apart are not in LD. A LD map of chromosome 22 shows highly variable pattern of LD (Dawson et al, 2002). Widespread regions of nearly complete LD up to 804 kb in length are intermingled with regions of little or undetectable LD. Haplotype analysis showed the most frequent haplotype was GA. This was the most frequent haplotype when the sample types were subdivided according to spoken

language. In comparison to studies from An et al, (2004) *D'* of the two SNPs was estimated at 0.967. The linkage disequilibrium (LD) revealed a non-independence of allele segregation because the loci analyzed were strongly linked in the Apobec 3 G gene. The data are consistent with greater genetic diversity of African populations and can form the basis for further evaluation of the role of variation in this gene in response to HIV. "

## Literature review:

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

1. Sentence 6 on page 1 "This is life threatening as it causes the formation of liver diseases, cancer, tuberculosis and many more" under section 1.1 Human Immunodeficiency Virus (HIV) and Acquired Immune Deficiency Syndrome (AIDS) in Sub-Saharan Africa has been edited to 'This is life threatening as it increases susceptibility to infectious diseases." as it is less repetition than the previous one.
2. The first sentence of paragraph two under section 1.1 Human Immunodeficiency Virus (HIV) and Acquired Immune Deficiency Syndrome (AIDS) in Sub-Saharan Africa on page 1 has been edited from "Statistics reveal that Sub-Saharan Africa has the highest HIV infections" to "Statistics reveal that Sub-Saharan Africa has the highest prevalence of HIV infections". The word prevalence was included to describe the rate of infection of HIV and AIDS was deleted as it is not associated with the prevalence of HIV infections.
3. The following sentence on page 1 "In this region there are 22.5 million adults living with HIV and approximately 1.7 million adults and children had become infected with

the virus in one year (UNAIDS/WHO, 2007). “ “In this region there were 22.5 million adults living with HIV and approximately 1.7 million adults and children had become infected with the virus in one year (UNAIDS/WHO, 2007).”

4. The following sentence on page 1 “Women and girls have become especially vulnerable to HIV/AIDS and in 2007 it is estimated that 61 % of adults in South Africa living with HIV are women (UNAIDS/WHO, 2007).” of paragraph two under section 1.1 Human Immunodeficiency Virus (HIV) and Acquired Immune Deficiency Syndrome (AIDS) in Sub-Saharan Africa was changed to the past tense for grammatical sense and now reads ““Women and girls have become especially vulnerable to HIV/AIDS and in 2007 it is estimated that 61 % of adults in South Africa living with HIV were women (UNAIDS/WHO, 2007).”

## 1.2 HIV diversity in Africa

1. The HIV diversity in Africa section 1.2 “All subtypes of HIV-1 and HIV-2 are found circulating in Africa. Knowing the distribution and characterizing the viruses present in Africa is the only way we can develop effective therapies unique to Africa. DNA sequencing of archival samples from Zaire (now the Democratic Republic of the Congo (DRC)) was used to date the origin of HIV-1 and to discern its evolutionary history (Worobey *et al*, 2008). The archival samples are designated ZR 59 (a blood plasma sample from 1959) and DRC60 (biopsy specimen from female patient in the DRC). The phylogenetic analysis showed that these sequences shared a common ancestor at least 50 years ago because of the short nodal distance between the two. Particularly the DRC60 sequence was found to cluster close to the A subtype ancestral node in the phylogenetic tree while the ZR59 sequence clustered closer to the subtype D (Worobey *et al*, 2008). This indicates that even 50yrs ago group M

strains had evolved into distinct subtypes that were circulating within the populations of this region. The phylogenetic analysis indicates too that there is substantial genetic diversity between the two ancestral sequences. This further confirms that the virus was present in humans long before the epidemic was characterized. An interesting point has been raised that urbanization has played a vital role in the exponential rise of the disease in Africa because the two ancestral sequences cluster with other strains from the same region rather then the same subtype, giving rise to viral lineages which are more diverse within viral subtypes. Thus it was concluded that the diversification of HIV-1 group M viruses began in Kinshasa (Keele *et al*, 2006). The causative agent of HIV-2 is known to be SIV sm (SIV from Sooty Mangabeys). The Sooty Mangabeys naturally inhabit the forest of Senegal east to Ghana. The origin of HIV-2 has not been so contumacious. A natural reservoir of the virus was detected in these monkeys as early as 1989 (Hirsch *et al*, 1989). Unlike HIV-1, each HIV-2 subtype was apparently the result of independent cross-species transmission. The most recent common ancestor of HIV-2 subtype A was dated to be ~1940 and subtype B 1945 in Guinea-Bissau (Lemey *et al*, 2003). Subtypes A and B cause are linked to the epidemic in this region while other subtypes have been identified in singly infected people. Like HIV-1 transmission of the virus has been marked by a period of untraceability followed by an exponential rise in infections. This rise of infections is estimated to occur around the same time as the War of Independence between 1963 and 1974. Once again it is evident that viral epidemics are reliant on socio-economic conditions. HIV-1 group M accounts for more than 95 % of all HIV infection around the world apart from HIV-2 infections in certain parts of Africa (Lemey *et al*, 2003). Group M is composed of numerous different subtypes, each endemic in certain parts of the world, e.g. subtype B is found mainly in Europe, the Americas, Japan and

Australia (Heeney *et al*, 2006). In Southern Africa, subtype B is the most common. HIV -1 subtype C is well documented as being the predominant strain in South Africa itself (Rodenberg *et al*, 2001 & Papathanasopoulos *et al*, 2002). However there is evidence of minor strains such as non-C subtypes and various recombinant subtype viruses also being present in South Africa. In addition HIV-1 subtype B is found to be present in the homosexual population (van Harmelen *et al*, 1997 & Williamson *et al*, 1995)." This explanation of HIV subtypes circulating in Africa has no bearing on the genetic affinities of the populations examined herein and was deleted.

2. Under the section HIV diversity in Africa on page 2 the following was added as it explained the evolutionary history of HIV-1 "There are two types of HIV; HIV-1 and HIV-2. HIV-1 is further divided into groups called Groups M, N, O and P. These groups are further subdivided subtypes A-K. HIV-1 group M accounts for more than 95 % of all HIV infection around the world apart from HIV-2 infections in certain parts of Africa (Lemey *et al*, 2003). All subtypes of HIV-1 are found circulating in Africa. Subtype B is found mainly in Europe, the Americas, Japan and Australia (Heeney *et al*, 2006). The origins of HIV-1 have been very contentious. The first HIV-1 like virus called Simian Immunodeficiency Virus (SIVcpz) was characterised in captive chimpanzees in 1989. Initially the SIVcpz virus was thought not to be responsible for the disease in humans as it could not be determined if wild chimpanzees are naturally infected by the virus. Chimpanzees are divided into four subspecies; the western, Nigerian, central and eastern chimpanzees (Gagneux et al, 1999). Each subspecies occupy a different geographical niche. Sequence analysis of all the SIVcpz strains from captive chimpanzees described by Peeters et al (1989) and Peeters et al (1992) showed that all these strains clustered together within the central chimpanzee subspecies and also formed one cluster within all HIV-1 strains. This is

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

3. The following sentence of section HIV diversity in Africa on page 3 has been changed from "DNA sequencing of archival samples from Zaire (now the Democratic Republic of the Congo (DRC)) was used to date the origin of HIV-1 and to discern its evolutionary history (Worobey *et al*, 2008)." to the following "DNA sequencing of archival samples from Zaire (now the Democratic Republic of the Congo (DRC)) was used to date the origin of HIV-1 and confirmed the evolutionary history (Worobey *et al*, 2008) as this is the epicentre of diversity." as now the evolutionary history has now been clearly and concisely explained in the preceding paragraph.
4. The following sentence from paragraph two of the HIV diversity in Africa on page 3 has been changed from "The phylogenetic analysis showed that these sequences shared a common ancestor at least 50 years ago because of the short nodal distance between the two." to "The phylogenetic analysis demonstrated a short nodal distance between the two." As this sentence is more conies and still coveys the information.
5. In paragraph two of the section HIV diversity in Africa on page 2 changed the following sentence from "This indicates that even 50 yrs ago group M strains had evolved into distinct subtypes that were circulating within the populations of this region." to "This indicates that even 50 yrs ago group M of HIV-1 strains had evolved into distinct subtypes that were circulating within the populations of this region." This change will explain that group M is a subtype of HIV-1.
6. Paragraph 3 on page 3 has been changed from "DNA sequencing of archival samples from Zaire (now the Democratic Republic of the Congo (DRC)) was used to date the origin of HIV-1 and to discern its evolutionary history (Worobey *et al*, 2008)." to "DNA sequencing of archival samples from Zaire (now the Democratic Republic of

the Congo (DRC)) was used to date the origin of HIV-1 and confirmed the evolutionary history (Worobey *et al*, 2008) as this is the epicentre of diversity."

7. The word "then " in sentence one from paragraph 3 under section HIV diversity in Africa has been edited to "than" as it is more grammatically correct.

## 1.3 HIV Life Cycle

1. Under section 1.3 HIV Life Cycle the word "contains" from sentence four on page 5 has been changed to "containing" for grammatical purposes.
2. Under section 1.3 HIV lifecycle on page 5 replaced the word "hi-jacked" with seizes thus the sentence will read as the following "Once the mRNA is processed in the nucleus, it is transported to the cytoplasm where the viruses once again seizes the host protein making processes to produce the viral proteins Env(gp160), Gag, Gag-pol, Vif, Vpr, Vpu, Rev, Tat and Nef." It was replaced as it was not the scientific or correct word to use.
3. Under section 1.3 HIV lifecycle on page 5 replaced the words "such as" was removed thus the sentence will read as the following "Once the mRNA is processed in the nucleus, it is transported to the cytoplasm where the viruses once again seizes the host protein making processes to produce the viral proteins Env (gp160), Gag, Gag-pol, Vif, Vpr, Vpu, Rev, Tat and Nef." The words "such as "are unnecessary as the proteins are being stated.
4. The space between "gp 41" was removed thus the sentence under section 1.3 HIV lifecycle on page 6 reads as follows "The gag and gag-pol polyproteins aggregate near the membrane and interact with plasma membrane and the gp41 present in the membrane. "the space was removed as it was scientifically incorrect to write the protein name with a space in between.

5. The following sentence on page 6 was edited from “As the gag and gag-pol aggregate at the plasma membrane the virion begins to form and begins to be extruded from the host cell membrane.” to “As the gag and gag-pol aggregate at the plasma membrane the virion begins to form and will be extruded from the host cell membrane.” This new sentence is more grammatically sound.

## 1.4 Host proteins involved in HIV

1. The word infection was added to the first sentence of the first paragraph on page 6 as it was omitted after the word HIV.
2. The following two sentences on page 6 were combined to read as one more concise sentence **“**Classical analysis of host genetics and its involvement in viral genetics was studied by making use of candidate gene studies. Numerous AIDS restriction genes that effect susceptibility to viral infection have been discovered via candidate gene approach (Hutcheson *et al*, 2007).” The new sentence reads as follows “Classical analysis of host genetics and its involvement in viral genetics made use of candidate gene studies, leading to the discovery of numerous AIDS restriction genes that effect susceptibility to viral infection (Hutcheson *et al*, 2007).”
3. In the following sentence on page 7 **“**Most candidate gene studies have focused on European populations (or populations of European origin); little information is available for the effects of variation in these genes in Africans where the epidemic is flourishing at an alarming rate.” The word Sub-Saharan black was replaced with Africans as this word is correct.
4. Removed section 1.4.1 Proteins involved in HIV/AIDS pathogenesis, 1.4.1.1 Chemokine receptor CCR5, 1.4.1.2 Human leukocyte Antigen (HLA) system on

pages 8-10. This section of the literature review has been well documented and does not have to be included here.

5. The following sentence on page 7 “This does not give us a complete picture of the disease in Sub-Saharans Africa but is a good platform from which to grow.” has been edited to the following “This does not give us a complete picture of the disease in Africans due to admixture of genes but is a good platform from which to grow.”

6. Inserted the word “Candidate” so the sentence on page 7 now reads “Candidate gene products include chemokine receptors and their variants (CCR5), chemokine receptor ligands (SDF), cytokines (IL), the HLA system and various factors involved in cellular immunity such as TRIM5α , APOBEC3G (Sheehy *et al*, 2000) and 3F.” This word was added as all the above proteins functions have been found by using the candidate gene approach.

7. Removed the following paragraph on page 8 “In 2003 an initial study showed that there is detectable variation in the upstream non-coding region of APOBEC3G (Ramdin, 2003). The definitive role of APOBEC3G in HIV/AIDS pathogenesis could not be clarified as all sequenced study samples were HIV positive and there was no control group against which to make a comparison. Further PCR and sequencing was needed to clarify this issue. In addition there was no detection of heterozygotes at various SNPs locations within any of the sequenced samples, as sequencing was not very reliable. Thus an *in vitro* assay is needed to facilitate rapid heterozygote detection. Thus it is vital that this gene be re-examined in detail. “and placed it at the end of the introduction to put into context the reason for doing the study’

## 1.5 Apobec 3G

1. Referenced Figure 1.2 at the end of sentence one of paragraph two. "*APOBEC3G* has eight exons, all of which are transcribed, arranged in tandem (Jarmuz *et al*, 2002) (Figure 1.2)."

## 1.6 Origin and Evolution of APOBEC Deaminase

**1.** The following sentence in paragraph two on page 10 was changed to "Homology modelling revealed that by removing the sequences of nucleotides termed the gaps from *E. coli* cytidine deaminase (ECCDA) the signature sequence of APOBEC 1 was derived (Chester *et al*, 2000)." From "structure and gene organization were based on *E. coli* cytidine deaminase. Homology modelling revealed that by removing the some sequences of nucleotides termed the gaps from *E. coli* cytidine deaminase (ECCDA) the signature sequence of APOBEC 1 was derived (Chester *et al*, 2000)." As the sentence did not make sense grammatically.

### 1.6.1 Evolution of the Apobec 3G family

1. Referenced Figure 1.3 in the following sentence on page 11 "This duplication gave rise to seven APOBEC3 genes in humans designated APOBEC 3 A, B, C, D, F, G, H (Figure 1.3)."
2. The following sentence on page 12 was edited from "In addition the sequences flanking the *APOBEC 3* are highly conserved in bony fish and chicken suggesting that the human APOBEC 3 locus may have evolved much later in humans" to "In addition

the sequences flanking the *APOBEC 3* are highly conserved in bony fish and chicken suggesting that the human APOBEC 3 locus may have evolved much later in mammals."

## 1.7 HIV-1 viral infectivity factor (vif)

1. The word "found "was deleted from the sentence on page 12 "The C-terminal domain is functionally important and is found bound to many membranes and to be associated with Gag precursors (Khan *et al*, 2001)." As it was grammatically incorrect. The sentence now reads "The C-terminal domain is functionally important and is bound to many membranes and to be associated with Gag precursors (Khan *et al*, 2001)."

## 1.8 The Interaction of APOBEC 3G and Vif

1. In paragraph four the word "effect" in the sentence on page 14 "Although mutations in this motif reduced interaction with the complex, it did not affect the interaction of Vif with APOBEC3G." was changed to "affect" as it was previously the incorrect word to be used

## 1.9 Selection of APOBEC3G and Vif

1. References for the following sentence on page 17 was added, "Comparative sequence data from non human primates such as Old World monkeys (OWM) New World monkeys (NWM) and hominids indicate that APOBEC3G has been under positive selection pressure in primates for at least 33 million years (J,Lui et al, 2010 & L, Sawyer, 2004)."

2. Reference for the following sentence on page 17 was added, "However examination of *APOBEC 3G* reveals that non-synonymous changes are far greater that synonymous changes which is the driving force for the fixation of variants with altered proteins (Meyerson and Sawyer, 2011)."
3. In paragraph 3 sentence two on page 17 "APOBEC3G in Old World monkeys and hominids appears to have diverged from each other 23 million years ago" was edited to "APOBEC3G in Old World monkeys and hominids appear to have diverged from each other 23 million years ago"

## 1.11 Linkage Disequilibrium and haplotypes

1. The sentence "Though, when genes are inherited more often than chance then these genes are said to be in linkage disequilibrium (LD) (VanLiere & Rosenberg, 2008)." has been changed to be more grammatically correct sentence on page 19 "Though, when variation at two sites is inherited together more often than by chance then these genes are said to be in linkage disequilibrium (LD) (VanLiere & Rosenberg, 2008)."

## 1.12 Clinical significance of variation within APOBEC 3G

**1.** "haplotypes" is changed to "haplotype" in the sentence on page 20 paragraph one "This study revealed through haplotype analysis that there are six frequent haplotypes which cause these SNPs to be inherited together."
2. The comma in 2.9 % in sentence "The frequency was 37 % in AA as compared to 2, 9 % in EA (Winkler *et al*, 2004)." was removed as it was an error. The correct percentage is 29 % (page 20).

3. The sentence “29 polymorphisms were identified; these included 14 novel polymorphisms all with the exception of one found within the coding region of APOBEC3G.” was edited so that the number is written out in full as per grammatical rules (page 20).
4. The word then was replaced with than in Sentence in paragraph one page 20 “In contrast the P values for the estimated haplotypes and the arginine to histidine substitution within the AA seropositive sub populations studied was less than 0.024 indicating statistical significance (Winkler *et al*, 2004).” It is grammatically correct.
5. In the following sentence “However there is some consistency amongst the different studies; APOBEC3G shows no association with AIDS progression in Caucasians”, the word “shows” was replaced with past tense adjective “showed”. In addition the word “Caucasians” was replaced with “Europeans”. The sentence now reads “However there is some consistency amongst the different studies; APOBEC3G showed no association with AIDS progression in Europeans.

## 1.13 Origin of Modern Human

1. Under the section 1.13 Origins of Modern Humans on page 21 added the following paragraph “The patterns of LD and Haplotypes have been used as tools for elucidating the genealogical and demographic history of populations. There are two schools of thought. The least popular idea is that race or ethnicity can be used to predict genetic classification (Tishkoff and Kidd, 2004, Jorde and Wooding, 2004, Mountain and Risch; 2004). Historically race has been classified according to biological factors such as skin colour, morphology. This in itself is complicated and not always correct; traits that produce phenotypic differences are

a result of genetic component adapting to the environment in which an individual lives. In contrast ethnic races are clustered in groups but this is a consequence of the geographic expansion of population out of Africa (Tishkoff & Kidd, 2004). This expansion has not given rise to any race specific genes. However the most popular ideology is that LD patterns and population haplotypes are useful in determining the geographical distribution and lineage of populations (Rosenberg et al, 2002, Lane et al, 2002). These studies show that populations that arise from the same geographic region have similar LD patterns suggesting that the origin of human populations is important in assessing the genetic diversity.

2. The sentence on page 22 "It is argued that by roughly 30 000 yrs ago *Homo sapiens* was dominant and gave rise to modern humans. "was edited to "It is argued that by roughly 30 000 yrs ago *Homo sapiens* was dominant in Europe." This was edited as modern human originated in Africa not Europe.
3. The sentence on page 23 was edited from "It is well documented that the first exodus out of Africa to Near East occurred some 60 000 years ago (Cavalli-Sforza *et al*, 1988 & Cavalli-Sforza *et al*, 1992 & Cavalli-Sforza, 2006)." to "The first exodus out of Africa to Near East occurred some 60 000 years ago (Cavalli-Sforza *et al*, 1988 & Cavalli-Sforza *et al*, 1992 & Cavalli-Sforza, 2006)." The word "it is well documented" is redundant.
4. A more recent reference for the following sentence was added. "It is believed that the subsequent migrations to East Asia, Europe & Australia, and South Asia occurred 30 000, 40 000 and 50 000 years ago respectively on page 23.
5. The following sentence within paragraph four on page 23 was edited from "These migrations are confirmed by tracing the mitochondrial and Y-chromosome

lineages" to "These migrations are confirmed by tracing the mitochondrial and Y-chromosome lineages and numerous nuclear markers."

### 1.13.1 Bantu Expansion

1. Sentence four in paragraph one on page 24 was edited from "Therefore detecting the association of SNPs in HIV can be very valuable in Africans because local patterns of LD have been characterized across the genome for Sub-Saharan South Africans (Donfack *et al* 2006)." To "Therefore detecting the association of SNPs in HIV restricting genes can be very valuable in Africans because local patterns of LD have been characterized across the genome for Sub-Saharan South Africans (Donfack *et al* 2006)." It was important to make the distinction that this sentence was in reference to HIV restriction genes.
2. The first sentence on paragraph two on page 24 was edited to remove the word "classically" and now reads as follows "The Bantu are classified as a group of related individuals who originated in West Africa".
3. The seventh sentence in paragraph two on page 24 was edited to the past tense and reads as follows "As the Bantu migrated so too did their languages."
4. Removed the following paragraph from page 24 "The issue of racially biased genes in disease has compounded the quest for answers. There has been some favour given to the theory that certain traits are racially biased (Mountain & Risch, 2004). Historically race has been classified according to biological factors such as skin colour, morphology. This in itself is complicated and not always correct; traits that produce phenotypic differences are a result of genetic component adapting to the environment in which an individual lives. In contrast ethnic races are clustered in groups but this is a consequence of the geographic

expansion of population out of Africa (Tishkoff & Kidd, 2004). This expansion has not given rise to any race specific genes. Nevertheless, the manifestation of certain degrees of the same condition in different ethnic populations does make this assumption attractive."

5. Introduced a new section 1.13.2 Genetic substructure of South African Populations on page 25. The following was added under this section. "Black South Africans compromise approximately 77 % of the country's population. In addition, South Africa has 11 official languages. 9 or the 11 official languages are Bantu speaking languages. These linguistic groups all belong to the Eastern Bantu-speaking group. From Y chromosome data it was shown that these languages cluster into 3 specific groups; Tswana/Sotho, Nguni and Venda language groups. The Nguni group comprises the Zulu, Xhosa, Tsonga/Shagaan linguistic groups. The Sotho/Tswana group comprises the North Sotho, South Sotho and Tswana language groups. Measurements of $F_{st}$, was very low, indicating these population groups although linguistically diverse share more than 98% of their genetic variation, suggesting that they all share a common ancestor. These Linguistic groups also show genetic differences between these populations. The Nguni linguistic groups split in two with the Zulu and Xhosa forming a cluster while the Tsonga and Shangaan form a separate cluster with the Venda. The Sotho /Tswana group clusters midway between the two. This is indicative that migration events also influence underlying genetic diversity."

## 1.14 AIM

1. The aim on page 27 was edited. A new paragraph was a removed from section 1.4 "In 2003 an initial study showed that there is detectable variation in the upstream

non-coding region of APOBEC3G (Ramdin, 2003). The definitive role of APOBEC3G in HIV/AIDS pathogenesis could not be clarified as all sequenced study samples were HIV positive and there was no control group against which to make a comparison. Further PCR and sequencing was needed to clarify this issue. In addition there was no detection of heterozygotes at various SNPs locations within any of the sequenced samples, as sequencing was not very reliable. Thus an *in vitro* assay was needed to facilitate rapid heterozygote detection. Thus it is vital that this gene be re-examined in detail. "and added in the first paragraph of the aim as it made more sense to have it here as it explained the purpose of the study.

# Chapter 2

# Materials and Methods

## 2.1 Sample description

1. The number 79 in the first sentence of the first paragraph on page 29 was written out in full as it is in the beginning of the sentence.
2. The word only on page 29 in "In the present study only 69 of the 71 samples were used for follow-up investigations as material was limited" was removed from the following sentence as it was not necessary. The sentence now reads as "In the present study 69 of the 71 samples were used for follow-up investigations as material was limited"
3. The sentence on page 29 was edited from "In addition 45 samples were collected from the Bantu population at Wits University termed the General Population (GP)

(Table 2.1).” to “In addition 45 samples were collected from staff and students at the University of the Witwatersrand from the Bantu speaking population and termed the General Population (GP) (Table 2.1).” as it was more grammatically correct.

4. The words on page 29 “at the time of collection “was omitted from the sentence “The HIV status of these individuals was unknown at the time of collection” to “The HIV status of these individuals was unknown. “
5. The sentence on page 29 was edited from “Short self-reported patient histories of date of first infection, recent CD 4 count, and secondary illnesses were obtained from JHB and GP sample sets.” To “Short self-reported patient histories of date of first infection, recent CD 4 count, and secondary illnesses were obtained from the JHB sample set.”
6. The Figure 2.7 was moved from beginning of the section under the Sample description on page 30 to the end of the material and methods section on page 46 as it makes more sense as it consolidates the methods and samples used.
7. The following sentence on page 29 “Adding to these 56 samples was received from the HIVNET 028 Study (HIVNET). “was edited from present to past tense. It now reads “Adding to these 56 samples were received from the HIVNET 028 Study (HIVNET).”
8. The following sentence on page 29 was edited to be more grammatically correct. It was edited from “These samples included CD4 counts and viral loads; however origin of ancestry of individuals was lacking.” To “These samples included CD4 counts and viral loads, however information on ancestry of individuals was not available.”

9. The comma in the sentence was removed and replaced with and so the sentence now reads as “Data was also collected on geographic origin of participants from JHB, GP and HJ study individuals.”

## 2.2 DNA extraction

1. The word “the” was removed from the sentence “This DNA was used to detect variation in the *APOBEC3G*.” on page 31 as it is redundant.
2. The following sentence on page 31 was edited grammatically from “The samples were also quantified using the Nanodrop ND-1000.” to “The DNA concentration in the samples was quantified using the Nanodrop ND-1000.”

## 2.4 Detection of Variation in *APOBEC3G*

### 2.4.1 The genotyping of position -571 using allele specific amplification

**1.** The genotyping of position -571 using allele specific amplification on page 32 was removed the description of PCR “The use of thermos table DNA polymerases and automation of the method increases its efficiency. As a result many PCR applications have been developed. These include screening and sequencing of inserts from phages and bacteria and the visualization of single-copy genes. There are three main steps in PCR, denaturation, annealing and extension. Initially double stranded template DNA is denatured. This is termed a hot start which is then followed by a ramp down to the annealing temperature where primers anneal to their target sites. Lastly extension of the primers, with nucleotides, by DNA polymerase is performed. Primer design is crucial in PCR and numerous things have to be taken into consideration in order to yield successful results. Primers

should have GC content between 40 % - 60 %. The higher the GC content the more stable the primers. The primers should have a length of between 18 nucleotides and 30 nucleotides. The primers should not be self complementary or complementary to each other. If primers are self-complementary they will fold back and bind on themselves. If complementary to each other they will bind to each other and not to the annealing sites on the target molecules resulting in inefficient or no amplification"

2. The following sentence on page 33 was edited from "The *APOBEC 3G* was accessed using Pubmed and the accession number (NT011520)." To "The *APOBEC 3G* was accessed using NCBI and the accession number (NT011520) (http://www.ncbi.nlm.nih.gov/gene/60489)." Pubmed was changed to NCBI as that was where the information the URL was added for referencing purposes.

3. In paragraph four the following sentence on page 34 was edited from "The cyclic conditions for the -571C PCR are denaturation at 94 °C for 2 min, this is followed by 35 cycles of denaturation at 94 °C for 30s, annealing temperature at 63.5 C for 25 s, extension at 72 °C for 20 s, the final extension is at 72 °C for 5 min." to "The cyclic conditions for the -571C PCR are denaturation at 94 °C for 2 min, followed by 35 cycles of denaturation at 94 °C for 30s, annealing temperature at 63.5 C for 25 s, extension at 72 °C for 20 s, the final extension is at 72 °C for 5 min."

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

**1.** Removed The PCR master mix reaction on page 34 is the same as for conventional PCR. The difference is that the PCR heat block where the PCR tubes

are placed is divided into different temperature zones which are selected by the user. The PCR is setup for one sample for each temperature zone tested. Essentially the same sample can be amplified using different annealing temperatures at once thereby reducing time. The 50 µl reaction mixture for the PCR consisted of 2X Fermentas Master Mix (0.05 u/µl *Taq* DNA polymerase (recombinant), reaction buffer, 4 mM $MgCl_2$, 0.4 mM of each dNTP (dATP, dCTP, dGTP, and dTTP). Primers were diluted to give a stock concentration of 2 µg/µl (mass (µg) + 0.5 (mass (µg) TE) and a working solution of 20 ng/µl (2 µl stock soln + 198 µl water) and 20 – 100ng of DNA. The cyclic conditions for the gradient PCR is denaturation at 94 °C for 2 min, this is followed by 35 cycles of denaturation at 94 °C for 30s, annealing temperature of between 60 °C - 58 °C for 30 s, extension at 72 °C for 30 s, the final extension is at 72 °C for 5 min. PCR products were electrophoresed on a 1% agarose gel stained with 10µg/ml EtBr in 1X TBE buffer at 7V/cm for 45 min." This was removed as it was one of the initial screening methods before the actual RFLP analysis.

**2.** Removed Figure 2.4.2 on page 35 "Diagrammatic representation of PCR heat block" on page 38 as the description of the gradient PCR has been removed thus the figure does not apply anymore.

**3.** **I**nserted on page 34 the following "The 50 µl reaction mixture for the PCR consisted of 2X Fermentas Master Mix (0.05 u/µl *Taq* DNA polymerase (recombinant), reaction buffer, 4 mM $MgCl_2$, 0.4 mM of each dNTP (dATP, dCTP, dGTP, and dTTP). Primers were diluted to give a stock concentration of 2 µg/µl (mass (µg) + 0.5 (mass (µg) TE) and a working solution of 20 ng/µl (2 µl stock soln + 198 µl water) and 20 – 100ng of DNA" into the paragraph beginning

"Once the gradient PCR established the optimal annealing temperature, this temperature was used to amplify the subsequent samples using," to describe the components of the PCR reaction used.

### 2.5.3 Linkage disequilibrium and haplotype analysis

1. The word "Disequilibrium" on page 43 was changed from capital to "disequilibrium" for consistency in the heading.

2. The word "Disequilibrium" on page 43 was changed from capital to "disequilibrium" for consistency in the first sentence of the first paragraph.

3. The words "lets" is deleted from the following sentence on page 44 "To explain LD, lets consider an example" as it is not necessary.

4. The number in the sentence on page 45 "136 samples with known genotypes at loci -571 and H186R were used for the haplotype analysis" was edited so that the full number is written out at the beginning of the sentence.

## Chapter 3

## Results

## 3.1 Analysis of upstream non-coding region sequences

### 3.1.1 Reanalysis of previously sequenced samples

Sequenced data is compared to the information about population variation, and y allele frequency and LD analysis were compared to other populations. Thus the entire section on page 46 to 51 was changed from "15 samples were sequenced by Inqaba Biotech in 2003.

The samples were sequenced in the forward and reverse directions and the average length of the sequences was 900bp. In 2003 the forward and reverse sequences of each sample were automatically aligned with the putative upstream non-coding region of APOBEC 3G (reference sequence) obtained from the Human Genome Database to form a contig. Each contig was then edited according to the chromatograms. Editing involved examining the forward and reverse chromatograms while simultaneously looking at the sequences in relation to the reference sequence. Thereafter the reference sequence was removed from the contig and a new consensus sequence was created from the forward and reverse sequences. The consensus sequence of each sample was then aligned automatically to form a new contigs. The relative position of the mutations was then noted. In 2003 sequencing of the samples showed that there were numerous point mutations present in the samples (Figure 3.1.1.1). Different combinations of transitions and transversions were present in each sample. Transversions were the most common point mutations in the sample. Six of the eight possible transversions were present in the sequences, T-A, C-A, C-G, T-G, G-C, and A-C. Only three transitions were detected, C-T, A-G, G-A, with C-T being the most frequent transition. Numerous insertions and deletions were also found. SNPs at -972, -963, -960, -881, were not observed in these samples but other new SNP were characterized. During consensus alignment of the samples, two sites were of particular interest. At position -590 all sequences were different from the reference sequence, there being an A-G transition. Furthermore, at position -571 11 of the 18 sequences were different from the reference sequence there being a G-C transversion (Figure 3.1.1.3). Another SNP (G-C) was found at -571 in most of the samples. One of the previously characterized SNP at -91 was also present in some samples. -286 A-G transition was represented in 5 of 10 sequences. In 2003 -91 SNP was found at a position of 91 bases upstream of transcription initiation. Reanalysis demonstrated after robust editing only 6 from initial 15 sequences provided informative data.

Four heterozygotes (GC) and 2 CC homozygotes at position -91 were detected. The allele frequencies could not be estimated for this SNP because of the very small sample size. SNPs at position -163, -166 and -199 were not characterized in 2003 but re-analysis showed it to be well represented in 15 reanalysed samples. Five heterozygotes were observed at each position -163, -166 and -199. Heterozygote genotypes observed for these positions were found in the same samples. Homozygotes for the ancestral alleles (T) according to the dbSNP database were observed at these positions in the remainder of the samples. No homozygote genotypes were observed for the minor alleles at each SNP position. The remaining 10 samples were homozygous for the major allele at all three loci. In subsequent reanalysis, I found that the SNP at -286 was the result of a sequencing artefact as may happen when G is preceded by T in direct sequencing (Figure 3.1.1.2).The SNP at -571 observed during analysis in 2003 was still polymorphic after re-analysis. Of 15 samples 6 were heterozygous, 8 homozygous for the C allele and 1 was homozygous for the G allele. The frequency of the C allele is 0.75 and the frequency of the G allele is 0.25 in this group. The population diversity on the dbSNP shows that in Sub-Saharan populations the allele frequency of the C and G alleles to be 0.917 and 0.083 respectively (www.ensembl.org). This difference will be discussed later. Polymorphism at position -881 were not characterized in the 2003 analysis but were well represented upon re-analysis. 7 heterozygotes and 2 homozygotes for the allele C were observed. However, frequency data could not be ascertained from the sequences as the sample size is too small." to "Fifteen samples were sequenced by Inqaba Biotech in 2003. The samples were sequenced in the forward and reverse directions and the average length of the sequences was 900bp. In 2003 the forward and reverse sequences of each sample were automatically aligned with the putative upstream non-coding region of APOBEC 3G (NCBI reference sequence NT_011520.11) to form a contig. Each contig was then edited according to the chromatograms. Editing involved examining the forward and reverse chromatograms

while simultaneously looking at the sequences in relation to the reference sequence. Thereafter the reference sequence was removed from the contig and a new consensus sequence was created from the forward and reverse sequences. The consensus sequence of each sample was then aligned automatically to form a new contigs. The relative position of the mutations was then noted. In 2003 sequencing showed that there were numerous point mutations present (Figure 3.1). Different combinations of transitions and transversions were present in each sample. Transversions were the most common point mutations. Six of the eight possible transversions were present in the sequences, T-A, C-A, C-G, T-G, G-C, and A-C. Only three transitions were detected, C-T, A-G, G-A, with C-T being the most frequent transition. Numerous insertions and deletions were also found. SNPs at -972, -963, -960, -881, were not observed in these samples but other new SNP were characterized. During consensus alignment of the samples, two sites were of particular interest. At position -590 all sequences were different from the reference sequence, there being an A-G transition. Furthermore, at position -571 11 of the 15 sequences were different from the reference sequence there being a G-C transversion (Figure 3.1.1.3). Another SNP (G-C) was found at -571 in most of the samples. One of the previously characterized SNP at -90 was also present in some samples. -286 A-G transition was represented in 5 of 10 sequences. In 2003 -90 SNP was found at a position of 91 bases upstream of transcription initiation. Reanalysis demonstrated after additional editing only 6 from initial 15 sequences provided informative data. Four heterozygotes (GC) and 2 CC homozygotes at position -90 were detected. The allele frequencies could not be estimated accurately for this SNP because of the very small sample size. However comparative data from Ensembl shows that the allele frequency is relatively the same in the Yoruba from Nigeria (G and C allele frequency is 0.562 and 0.438 respectively) while the frequency of the ancestral allele (C) in Europeans is much less than the frequency of the G allele (Table 3.1). SNPs at position -163, -166 and -199 were not

characterized in 2003 but re-analysis showed them to be well represented in 15 reanalysed samples. Five heterozygotes were observed at each position -163, -166 and -199. Heterozygote genotypes observed for these positions were found in the same samples. Homozygotes for the ancestral alleles according to the dbSNP database were observed at these positions in the remainder of the samples. No homozygote genotypes were observed for the minor alleles at each SNP position. The remaining 10 samples were homozygous for the major allele at all three loci. There is no frequency data available for SNP -163 and -166 for Africans or Europeans. However the data available for SNP -199 shows that the frequency of the ancestral allele (A) in Africans is similar to the estimated values within the study (Table 3.1)."

1. The number "15" in sentence on page 47 "15 samples were sequenced by Inqaba Biotech in 2003." was edited to the full number as it was at the beginning of the sentence.

2. The following sentence on page 47 "In 2003 the forward and reverse sequences of each sample were automatically aligned with the putative upstream non-coding region of APOBEC 3G (reference sequence) obtained from the Human Genome Database to form a contig." was edited to include the reference sequence NCBI identity to the following "In 2003 the forward and reverse sequences of each sample were automatically aligned with the putative upstream non-coding region of APOBEC 3G (NCBI reference sequence NT_011520.11) to form a contig."

3. The word possible was removed from the following sentence on page 47 "Six of the eight possible transversions were present in the sequences, T-A, C-A, C-G, T-G, G-C, and A-C". the new sentence reads as follows "Six of the eight transversions were present in the sequences, T-A, C-A, C-G, T-G, G-C, and A-C."

4. The number 18 in the following sentence on page 48 “Furthermore, at position -571 11 of the 15 sequences were different from the reference sequence there being a G-C transversion (Figure 3.1.1.3).” was changed to 15 as this was the correct number.

5. The word calculated in sentence on page 48 “The allele frequencies could not be calculated for this SNP because of the very small sample size.” Was replaced with estimated as this is the correct word to use.

6. The following was added on page 48 “However comparative data from Ensembl shows that the allele frequency of both alleles is relatively the same in Africans. While the frequency of the ancestral allele (C) in Europeans is much less than the frequency of the G allele (Figure 3.1).” into the paragraph three after sentence four to compare the study data to data available on Ensembl.

7. The following was added at the end of paragraph four page 48 “There is no frequency data available for SNP -163 and -166 for Africans or Europeans. However the data available for SNP -199 shows that the frequency of the ancestral allele (T) in Africans is similar to the estimated values within the study (Table 3.1).”

8. “Frequency data from Ensembl was also not available for this SNP.” On page 50 was added to the end of paragraph six.

9. The following sentences on page 50 were changed from “The frequency of the C allele is 0.75 and the frequency of the G allele is 0.25 in this group.” To “The frequency of the C allele is 0.733 and the frequency of the G allele is 0.267 in this group.” The frequencies have been changed as they were incorrectly calculated previously.

10. The following sentence on page 50 was edited from "The population diversity on the dbSNP shows that in Sub-Saharan populations the allele frequency of the C and G alleles to be 0.917 and 0.083 respectively (www.ensemble.org)." to "The population diversity on the dbSNP shows that in African populations the allele frequency of the C and G alleles to be 0.894 and 0.106 respectively (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=5757463.)"

11. The following sentence on page 50 is edited so that the numbers are written out in full from "7 heterozygotes and 2 homozygotes for the allele C were observed." to "Seven heterozygotes and two homozygotes for the allele C were observed."

12. The following table was added "Table 3.1 Sequenced data from 2003 and comparative data of other populations retrieved from Ensembl" was added to page 51. This was necessary to compare my data to what is in literature.

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

### 3.2.1 The genotyping of position -571 using Allele Specific Amplification

1. The first heading under the results section has been edited to include "using genotyping assays"

2. Under the heading on page 54 added the following explanation of the samples genotyped "Genotyping was not possible in all of the samples of each sample set as the DNA was not available for all samples at the time of testing. Thus only subset of each group was genotyped as indicated in Appendix I."

3. Under section 3.2.1 The genotyping of position -571 using Allele Specific Amplification the first sentence of the first paragraph on page 54 has been edited to read “Allele-specific PCR was used to genotype the insertion at position -571 in 165 samples (69 JHB, 56 HIVNET and 40 GP) (Appendix I).”
3. The second sentence of the second paragraph on page 54 was edited from “The PCR products for each allele of each SNPs were run together on the size same agarose gel to facilitate their correct genotyping (Figure 3.2.1).” to “The PCR products for each allele of each SNP were run together on the same agarose gel to facilitate their correct genotyping (Figure 3.2.1).”
4. The word “will” was deleted from the sentence on page 54 and will now read “While heterozygotes yield a product of the same intensity in both reactions.”
5. The word “will” was deleted from the sentence on page 54 and will now read “The sizes of the products for the different alleles differ allowing accurate genotyping.”
6. The following sentence on page 54 was edited to “GG homozygotes were more frequent than CC homozygotes.”
7. 3.2.2 Detection of SNP -571 using Restriction Fragment Length Polymorphism (RFLP) the first sentence of the first paragraph on page 55 should read “A RFLP-PCR assay was designed to genotype the -571 SNP at position within APOBEC 3G in 136 samples (13 JHB, 91 HIVNET and 32 GP).”
8. Under section 3.2.3 Detection of H186R using pyrosequencing added in a breakdown of the 136 samples used. Thus the sentence on page 57 now reads “Variation in codon 186 within exon 4 was detected in 136 samples (13 JHB, 91 HJ, and 32 GP).
9. Table 3.3.1 A summary of the genotyping data collected at all polymorphic positions using allele-specific PCR, RFLP-PCR and Pyrosequencing™ assays. The genotypes,

numbers of individuals' genotyped, genotype and allele frequencies, $\chi^2$ and P values are given for each position on page 59 was edited to include a breakdown of which samples were included within each genotyping assay.

10. Under section 3.3.2 Estimation of Pair-wise Allelic Linkage Disequilibrium on page 60 the Table 3.3.1 was edited to include comparative data from other populations in the publication from An et al. No LD data was available for any populations in Ensembl as the allele frequencies were very low. This is necessary to show the deviation of D' between my results and what is reported.

11. On page 60 the following was added as an explanation for the lack of European LD data. "Ensembl does not give LD values because allele frequencies are low in European populations. However the D' value for African Americans was 0.967 and 1.000 in European Americans (An et al, 2004)."

12. Under section 3.3 Estimation of Gene frequencies on page 59. I excluded the differences between the general Population and the HIV + groups as the sample sizes are too small and the data is very weak. Hence removed "**3.4.1 Differences between the General Population and HIV$^+$ Groups-** In an effort to establish if a possible association exists between variations in *APOBEC3G* and HIV-1 infectivity, the genotype frequencies at each of the polymorphic positions were compared between the general population and HIV$^+$ sample groups (Table 3.4.1).

13. Removed Table 3.4.1 The genotype frequencies at each of the polymorphic positions, in both the general population and HIV$^+$ samples. The number of individuals genotyped and P values for Fisher's exact test are also given for each group. The removal of the table was necessary as the entire analysis was removed from the thesis

as no conclusions can be drawn from differences between General population and HIV + samples on the effect of Apobec3G.

## 3.3 Estimation of Gene Frequencies

This section on page 58 has been renumbered from 3.4 in the original thesis to 3.3 in new thesis.

### 3.3.1 Differences at -571 and H186R using various genotyping methods

**T**his section has been added and the table 3.3.1(previously Table 3.2) has been moved from Section 3.2.3 Detection of H186R using pyrosequencing in the original thesis to this new sub section.

1. Edited Table 3.3.1. The sample numbers for RFLP and Pyrosequencing for -571 and H186R respectively were changed to 132. In addition the allele and genotype frequencies were also changed as they were incorrect.

### 3.3.2 Differences between the Bantu language groups

This section was renumbered from section 3.4.2 to 3.3.2 in the new thesis on page 61.

1. The following was added on page 61 “The analysis was based on the data from RFLP and pyrosequencing at -571 and H186R as these were the most reliable genotyping methods tested for each SNP. There was a total of 132 samples (Appendix I, Table A 4, A 5 and A 6) which represent 132 individuals within each genotyping method. However the final” to the first paragraph to explain the samples used in the analysis.

## 3.5 Haplotype Analysis

1. Included the following “Analysis of haplotype frequencies (Table 3.5.1) in the pooled data from samples sets JHB, GP and HJ there was a similar frequency between the GG and GA haplotypes of 0.419 and 0.413 respectively. The CG and CA haplotypes also exhibited a similar frequency”

2. Included the following table, Table 3.5.1 as it gives the haplotypes of the whole population and is imperative that it be included in the analysis.

3. Rename Table 3.5.2 from 3.5.1

## 4.0. Discussion

## 4.1. Direct Sequencing

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

2. Expanded paragraph 4 to include “In comparison the frequencies of the African population showed the ancestral allele frequency to be 0.438. The G allele is 0.562. The discrepancies between these two groups is attributed to the detection of specific genotypes. With the direct sequencing from 6 sequences 2 CC and 4 CG

and no GG genotypes were detected. Examining the African population data genotypes CG and GG are present in the population with equal frequency of 0.375 the CC genotype occurs at a slightly lower frequency of 0.250. In the European population the CG and GG genotypes also occur at the same genotype frequency however the numbers of these genotypes detected is higher than in the African population. Therefore the CC genotype is detected at a lower rate in this group. The frequency of minor allele in African Americans was 0.319, 0.340 in European Americans in work of An et al (2004). Furthermore a very similar minor allele frequency of 0.327 was observed in Africans in work of Reddy et al, (2010). These frequencies are very similar to the study frequency of 0.333." this explains the sequencing in relation to other published data.

3. Included the following paragraphs "Only SNP -199 has been characterised before (rrs 34550797). The sequencing data indicate very similar frequency of genotypes as in dbSNP. The GG occurred 42 out of 48 samples in Africans (Yoruba in Ibadan, Nigeria) (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=34550797). In addition the GG genotype also occurred in 22 out of 24 samples (http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=22:39472437-39473437;v=rs34550797;vdb=variation;vf=32183674). The allele frequency of the G allele is 0.938 and 1.000 in Yoruba and Utah residents respectively. In our study the allele frequency is comparable at 0.833 with the Yoruba population. The allele frequency of the minor allele is lower in the Yoruba population (0.062) than in the study (0.167) (http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=22:394

72437-39473437;v=rs34550797;vdb=variation;vf=32183674). Sequencing data shows SNP -199 is tightly linked to SNPs -163, -166. They always occur in the same samples and if one SNP is heterozygous or homozygous for ancestral allele the other two follow the same pattern. An et al (2004) makes reference to this SNP and terms it rare. It was not rare in the sequenced samples though. SNP -571 sequencing data frequencies are different from the Ensembl population data. The frequency of the minor allele is 0.267 as estimated from the sequencing data. In contrast the frequency within Africans is 0.106, 0.100 in Europeans, and 0.03 and 0.087 in Americans and Asians respectively. The frequencies from the study and the Ensembl are vastly different from those of An et al (2004) and Reddy et al (2010). The frequencies in these studies are very similar. For Africans the frequency is 0.0.91 and 0.089 for African Americans and Africans respectively. The European Americans exhibit a frequency of 0.063. SNPs -590 is a fixed polymorphism in our samples. All the sequenced samples deviated from the reference sample at this position. The sequencing data is dissimilar from the Ensembl data (http://www.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=22:39471748-39472748;v=rs17496004;vdb=variation;vf=32167980). In fact the frequency of the "minor" allele was 1 and the ancestral allele was not present in any of the sequenced samples. However the minor allele frequencies of Americans (Mexican Ancestry from Los Angeles USA) (0.03) and Europeans (Utah Residents (CEPH) with Northern and Western European ancestry) (0.07) are comparable to each other. The frequencies of the Yoruba (0.006) and Asians (Han Chinese in Bejing, China) (0.000) were also comparable. The minor allele frequency at SNP -881 for Europeans (Iberian population in Spain, British in

England and Scotland, Finnish in Finland, Toscani in Italia, Utah Residents (CEPH) with Northern and Western European ancestry) and Asians (Han Chinese in Bejing, China, Japanese in Tokyo, Japan, Southern Han Chinese, Chinese Dai in Xishuangbanna, China, Kinh in Ho Chi Minh City, Vietnam) is 0.001 and 0.003 respectively. In comparison the study data the minor allele frequency is 0.389 versus 0.287 for Africans (Yoruba in Ibadan, Nigeria, Luhya in Webuye, Kenya, Gambian in Western Divisons in The Gambia, Mende in Sierra Leone, Esan in Nigeria, Americans of African Ancestry in SW USA, and African Carribbeans in Barbados). Comparing the sequencing data to Ensembl it is clear that direct sequencing does remain the method of choice for detecting variation. The frequency values differed amongst various populations. It is relatively expensive so in the context of this study it was used an exploratory tool. Thereafter, indirect genotyping assays were designed to genotype the study population."

1. The last paragraph of this subsection was edited to "The minor allele frequency at SNP -881 for Europeans and Asians is 0.001 and 0.003 respectively. In comparison the study data the minor allele frequency is 0.389 versus 0.287 for Africans. Comparing the sequencing data to Ensembl it is clear that direct sequencing does remains the method of choice for detecting variation. Although the frequency values did differ they were comparable. However SNP -590, frequencies did deviate from Ensembl. It is relatively expensive so in the context of this study it was used an exploratory tool. Thereafter, indirect genotyping assays were designed to genotype the study population. "from "Direct sequencing remains the method of choice for detecting variation. However it is relatively

expensive so in the context of this study it was used an exploratory tool. Thereafter, indirect genotyping assays were designed to genotype the study population."

## 4.2. Genotyping of -571 and H186R SNP

1. Under section 4. 2 Genotyping of -571 and H186R SNP. The first paragraph is edited and reads as follows "Allele specific amplification did allow the genotyping of some samples of study population at -571. Genotyping was not possible in all of the samples of each sample set as the DNA was not available for all samples at the time of testing. Thus only subset of each group was genotyped as indicated in Appendix I."

2. Added the following sentence to paragraph 2 "In addition the minor allele frequency is 0.59 and it deviates from the sequencing (0.267) and Ensembl data in Africans (0.106), Europeans (0.100), Americans (0.03), and Asians (0.087)."

3. Added the following to the end of paragraph 4 "However SNP -571 minor allele frequency (0.84) is dissimilar from sequencing data (0.267) and Africans (0.106). In this instance it seems that RFLP technique was not reliable in detecting the variation in-571. However it proved very reliable when detecting variation in H186R. The minor allele frequency of the study population was 0.32. This is similar to the frequency of African Americans from the work An et al and Africans in Reddy et al (2010) where the frequencies were 0.368 and 0.307 respectively."

4. Added the following sentence to the last paragraph of this subsection "On the other hand it did not prove dependable in genotyping H186R as the frequency of the minor allele (0.50) deviated from published literature."

5. Deleted the following paragraph as it was not necessary "As the data from the RFLP analysis at -571 and pyrosequencing at H186R were more reliable and reproducible they were used for further analysis. Comparison of the genotyping results with the literature shows some marked differences. This study's frequencies were compared to the Yoruba population from Nigeria and African Americans (AA). The Yoruba and AA data was obtained from dbSNP and Ensembl. The -571 allele frequencies are similar between the Yoruba and this study but the genotype frequencies between the two are very different. The GG genotype has a higher frequency in this study but has a very low frequency in the Yoruba. The CC genotype was considerably higher in the Yoruba than in this study. There was no data available for AA thus no conclusions can be drawn regarding the similarities and/or differences in frequencies between AA and the Bantu population. Nonetheless, Africa has 2000 ethnically diverse populations. This difference in frequencies shows that the Yoruba population, from Nigeria, is different from the Bantu speaking study population. These ethnic differences do affect the genotype frequencies. The genotype frequencies for H186R are different between this study and the Yoruba. However, there remains a heterozygote excess in both populations. The TT genotype has a frequency of 33 % in AA (An *et al*, 2004). In contrast TT genotype within this study population was 20 %. The CC and TT genotype frequencies are analogous within both populations. In contrast the allele frequencies in the Yoruba population are equivalent to those in this

study. However they appear to differ from the allele frequencies of AA. No genotype data was obtained for AA at H186R and it would be interesting to determine if there is still a heterozygote excess in this population group. This heterozygotes excess is not present in the CEPH (Utah residents with Northern or Western Europe ancestry), Han Chinese or Japanese populations studied in the Hap Map. The frequency of the TT genotype for the CEPH, Han Chinese, and Japanese are 0.933, 0.911, and 0.822 respectively. This indicates that selection must be operating with this study population and the Yoruba allowing for the heterozygote excess."

6. The following was removed from the thesis "The identification of candidate genes that influence susceptibility to HIV-1 and the rate of progression to AIDS are well documented. The genotype frequencies at each of the polymorphic positions were compared between the general population and $HIV^+$ sample groups in an effort to establish whether a possible association exists between variation in APOBEC3G and HIV-1 infectivity. Firstly, the heterozygote excess at H186R could be indicative of incorrect genotyping. However, two reliable methods employed showed similar results, suggesting that that this is not a result of incorrect genotyping. The TT genotype was more frequent in the general population than in the HIV+ group. This is expected because more people would be advancing to AIDS and death hence less of this genotype would be found within an HIV + population. This seems to be consistent with a previous study where the TT genotype is associated with faster progression to progression AIDS (An *et al*, 2004). The substitution of histidine (H186) to arginine (186R) results in a change of charge from negative to positive. This change in polarity may influence the

structure of *APOBEC3G* binding to Vif thus may lead to faster progression to AIDS. There also seems to be selection against the GG genotype at -571 in the HIV + group and thereby increasing the frequency of the CC and CG genotypes. The GG genotype of -571 is more frequent in the general population than in the HIV + group. Likewise, as more individuals progress to AIDS and death the GG genotype is reduced. Therefore, there is selection against this genotype in HIV + group. There is no reported clinical implication associated with the GG genotype at -571. The p-values indicate that the frequency distributions at -571 (0.0026) and H186R (0.0245) are significant. This was removed as the analysis is not appropriate given the sample size under consideration.

7. Deleted Table 4.1 as it is not necessary in the analysis as a comparative table was done for the sequencing data.

8. Deleted the following as it is explained in next subsection. “The pair wise allelic Linkage Disequilibrium of -571 and H186R showed that they are not in Linkage Disequilibrium, D' is 0.216. In contrast they appear to be linked in AA and the D' is 0.967 (An *et al*, 2004). This lack of correlation is interesting because these SNPs are relatively close together on the chromosome and the literature suggests that SNPs within 10kb are always in strong Linkage Disequilibrium (Goldstein & Weale, 2001).”

## 4.3 Genetic variation in South Africans

1. The first sentence of the first paragraph on page 74 was edited as follows as i am not discussing the disease association of the SNPs under investigation as the data is weak and sample size was too small. The sentence previously read “However before

conclusions can be drawn about disease association, it is necessary to have at least some understanding of the local demographic history which has helped to outline patterns of genetic variation at this locus" and is replaced with "It is necessary to have an understanding of the local demographic history which has helped to outline patterns of genetic variation at this locus."

2. Edited the first sentence of the second last paragraph on page 75 of the discussion. It was changed from "Haplotype analysis revealed four haplotypes between the SNPs." To "Haplotype analysis revealed four haplotypes between the SNPs within the whole population and when the population was grouped according to different languages."

3. The following paragraph on page 75 was edited from "Zulu speakers comprised 45 % of the population followed by the Xhosa speakers at 17 %. The Tswana, Sotho, Pedi language speakers occur at a frequency of 12 %, 11 % and 7 % respectively. The Tsonga, Venda, Ndebele and languages represent less than 5 % of the sample. Even though this sample was representative of the population, some samples had to be pooled as suggested by Lane *et al* (2002) for subsequent analysis as they would be uninformative if not grouped. Only groups larger than ten individuals were used for analysis and subsequently called macrogroups. " to one concise sentence "The samples were representative of the population (Zulu speakers, Xhosa, Tswana, Sotho, Pedi, Tsonga, Venda, and Ndebele); some samples had to be pooled as suggested by Lane *et al* (2002) for subsequent analysis as they would be uninformative if not grouped. "

4. The following paragraph on page 77 was edited from "The pair wise allelic Linkage Disequilibrium of -571 and H186R showed that they are not in Linkage Disequilibrium, |D'| is 0.216. In contrast they appear to be linked in AA and the |D'| is

0.967 (An *et al*, 2004). This lack of correlation is surprising as the SNPs are just over 5000bp apart. Thus should be in strong linkage equilibrium as suggested by the literature and our own laboratory estimates. However this shows that LD is strongly affect by population history, selection, and sample size. The sample size may not have been sufficient to detect the LD. In addition another compounding reason may be the selection of study participants. These participants were all matched for ethnicity but not other parameters which may have underlying effects on the LD. "to the following "The pairwise allelic Linkage Disequilibrium of -571 and H186R showed that they are not in Linkage Disequilibrium, |D'| is 0.216. In contrast they appear to be linked in AA and the |D'| is 0.967 and 1.000 in EA (An *et al*, 2004). This lack of correlation is surprising as the SNPs are just over 5000bp apart. Thus should be in strong linkage equilibrium as suggested by the literature and our own laboratory estimates. However this shows that LD is strongly affect by population history, selection, and sample size. The sample size may not have been sufficient to detect the LD. In addition another compounding reason may be the selection of study participants. These participants were all matched for ethnicity but not other parameters which may have underlying effects on the LD. As mentioned previously both D' and $r^2$ are measures used to quantify LD. However their interpretation is different. It is thought that while both ranges from 0 to 1, $r^2$ are the more accurate measure. This is so because there is an inverse relationship between $r^2$ and sample size. While the calculated $r^2$ value is small it is not dependant on the sample size. Importantly because genes are linked on the same chromosome does not mean they will be in linkage disequilibrium. It is known that intermediate values of D' from ~0.3 to 0.7 is difficult to interpret as the D' value can be highly variable in pairs of sites that are separated by large distance (Wall &

Pritchard, 2003)." as it provided a more detailed explanation of the importance of the LD in context to my results.

## References

1. Added the following reference for synonymous and non-synonymous changes. Meyerson, N., and Sawyer, S., (2011).Two stepping through time: mammals and The

2. Added the following reference. Bakewell, Oliver and Hein de Haas (2007) African Migrations: continuities, discontinuities and recent transformations. in Patrick Chabal, Ulf Engel and Leo de Haan (eds.) *African Alternatives*. Leiden: Brill: 95-118.
3. Added the following reference: Ingman, M., and Gyllensten, U. 2003. Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Res*. **13:** 1600-6.
4. Added the following reference: Mitchell P (2010) Genetics and southern African prehistory: an archaeological view. Journal of Anthropological Sciences, 88,73–92. This is a more recent reference as to why languages give better indication of genetic affinities.
5. Added following reference: Mitchell, P., Whitelaw, G. 2005. The archaeology of southernmost Africa from c. 2000 BP to early 1800s: A review of recent research. *The Journal of African History* **46:** 209-241. This reference to alluded to the importance of demographic changes in languages and how this influences genetics.

6. Added the following reference: Reddy K., Winkler C.A, Werner L., Mlisana K., Abdool Karim SS, Ndung'u T. 2010. APOBEC3G expression is dysregulated in

primary HIV-1 infection and polymorphic variants influence CD4+ T-cell counts and plasma viral load. *AIDS* 24:195-204.

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