**Associations between HIV Susceptibility and Mutations in the Vif-associated APOBEC3G Proteasomal Complex**

Kevin Emancipator¹*, Jack Michuda¹*, and Sudhir Penugonda²**
Student¹: The Illinois Math and Science Academy, 1500 Sullivan Road, Aurora, Illinois  
Mentor²: Northwestern University-Feinberg School of Medicine, 645 N Michigan Avenue, Suite 900  
*These authors contributed equally  
**Corresponding author: spenugonda@gmail.com

**Abstract**

The purpose of this experiment was to determine whether a host’s single nucleotide polymorphisms (SNPs) in the Vif-associated APOBEC3G Proteasomal Complex influence his/her innate immunity to Human Immunodeficiency Virus-1. The experimenters compared host DNA at specific sites on each patient's genome (SNPs) and the patient's HIV-1 disease progression. They extracted DNA from peripheral blood mononuclear cells. Then, using high-throughput microarray nano-liter plates, a polymerase chain reaction (PCR) was run. The PCR gave a signal reading which the experimenters then translated into a genotype reading. Anonymous patient data concerning each individual's HIV progression were obtained from the National Institute of Health. The experimenters used a univariate analysis along with logistic regression to obtain a P-value, which was then corrected for multiple testing errors. A total of 5,000 patients were analyzed for each of 125 SNPs. After analysis, the experimenters found that the SNPs tested did not have a significant correlation with the viral set-point due to the high q-values that were above .05. Also, there was not a significant correlation between the SNPs tested and the patient CD4+ T-Cell slope. However, the study found an association that trended towards significance between HIV progression and the APOBEC3G associate SNP rs3736685.

**Introduction**

Human immunodeficiency virus (HIV) is a retrovirus that is prevalent throughout much of the world. HIV affects 33.3 million people worldwide and is related to substance abuse and poverty in an area². As with all viruses, HIV replicates intracellularly. It predominantly uses CD4+ T-helper cells to replicate. Upon antigen presentation by major histocompatibility complex (MHC) class I, CD8+ cytotoxic lymphocytes initiate apoptosis of the infected cell. However, CD4+ T-cells are reduced in HIV infection, mitigating helper T-cell response. Consequently, infection with HIV is characterized by a decrease in CD4+ T cell count. As T-helper cells play a vital role in acquired immunity, infection with HIV results in a compromised immune system.

There are two distinct types of the virus, aptly named HIV-1 and HIV-2. HIV-1 is most closely related to viruses found in gorillas and chimpanzees. There is also regional variation amongst the subgroup M which is primarily due to the many genome mutations of the HIV virus. Other subgroups of HIV-1 include group O, which refers to outliers, and group N, which refers to non-M and non-O subgroups. Furthermore, subgroup P refers to a recently analyzed HIV sequence unlike any other HIV virus that bears a stronger resemblance to simian immunodeficiency virus in gorillas than that which is in infected chimpanzees. HIV-2 is far less pathogenic and rarely observed outside of Africa. It is also categorized into several regional subgroups. The study, and all references to HIV, will be concerning HIV-1 group M.

As with many viruses, the HIV genome is fairly simple. The production of three structural and six accessory proteins is encoded within the genome. The six accessory proteins are VPU, REV, TAT, Nef, Vpr, and Vif. VIF, or viral infectivity factor plays a role in the pathogenesis of the HIV virus by degrading the enzyme APOBEC3G by ubiquitination. This process is relevant to the study and will be discussed in further detail.

HIV enters a CD4+ T cell by binding to CD4+ protein on the surface of the cell. The virus then enters the cell. After undergoing reverse transcriptase, newly formed HIV DNA is transported into the cell nucleus and integrates into host DNA. The integrated DNA services as a template for further viral replication, sometimes preceded by a prolonged latency. Protease acts as a catalyst to help the formation of new viruses. After replication, new viruses bud off infected cells into the body.

Apolipoprotein B mRNA-editing, enzyme-catalytic, polypeptide-like 3G (APOBEC3G) is a human enzyme that is part of a family that plays a role in innate viral immunity. APOBEC3G is known to exert antiretroviral immune activity. There are several mechanisms by which APOBEC3G is able to inhibit retrovirus reproduction. By causing deoxyctydine to deoxyuridine mutations in the HIV genome, the enzyme interferes with reverse transcription. Furthermore, it is proposed that a nucleophilic attack causes cytidine deamination. Deamination results in hypermutations G to A in the viral genome and ultimately resulting in an inability to replicate.

Viral infectivity factor (Vif) is a gene that encodes for the VIF protein. This protein increases the pathogenesis of the HIV virus. The VIF protein interferes with host mediated anti-viral defense. Specifically, it neutralizes APOBEC3G, an anti-viral enzyme. VIF inhibits the protein from entering the virion. Furthermore, by binding to APOBEC3G, proteolysis by ubiquitination is initiated, rendering the enzyme ineffective. VIF targets APOBEC3G degradation by removing Cullin5 E3 ubiquitin ligase. By inactivating the APOBEC3G enzyme through VIF, HIV is able...
to successfully attack CD4+ T-cells and cause infection. Without the presence of VIF, HIV is ineffective in infecting cells.

HIV progression to AIDS occurs on average over a ten-year period. However, some cases have been observed where individuals have progressed at much faster and slower rates. The virus may never even progress past stage one in some individuals. Naturally, HIV progression is linked to individual health and availability of medication and treatment options. In addition, studies have found a genetic link to HIV progression. Recent research has proposed various mechanisms that suggest a genetic linkage with progression.

One such mechanism involves an interaction between VIF, the APOBEC3G enzyme, and a complex of proteins. As stated, VIF neutralizes APOBEC3G by ubiquitination and allows HIV to remain pathogenic. The VIF mediated neutralization is mediated by set of recruited human proteins, Cul5, elongins B and C, and Rbx1, VIF forms a Cul5-SCF complex. This complex of proteins determined the effectiveness of VIF in inducing ubiquitination. This protein complex exists within every human but varies slightly due to genetic diversity. Rbx1 is essential to the function of the SCF complex. Over expression of Rbx1 interferes with the proper function of SCF8. In the presence of APOBEC3G, HIV infectivity is reduced by 75% when Rbx1 is over expressed. Furthermore, Cul5 proteasome inhibitors and mutants prevented VIF from degrading APOBEC3G by ubiquitination. To further investigate the effects of the genetic differences in the coding regions for the SCF complex on HIV progression, a large-scale analysis of 125 single nucleotide polymorphisms in high-risk individuals was performed.

Materials and Methods

In the study, mutations and variations in single-nucleotide polymorphisms (SNPs) found in coding (exon) and non-coding (intron) regions of DNA and the effect of these differences on the apolipoprotein B mRNA-editing, enzyme-catalytic, polypeptide-like 3G (APOBEC3G) enzyme and the cullin5 ECS E3 ubiquitin ligase complex were examined. SNPs are biallelic differences in DNA nucleotide bases. There are, on average over 150 polymorphisms for every human gene. That, coupled with the amount of genes that humans possess, makes for a lot of possible combinations. However, the study looked at specific SNPs on candidate genes that would affect how APOBEC3G or the ligase complex functions. SNPs on all genes, even on introns, could potentially influence the coding regions. So, the mutations in these regions could cause potentially beneficial or harmful effects. The human APOBEC3G enzyme has been shown to play an essential role in a host’s innate immunity to viruses such as the human immunodeficiency virus (HIV). The ligase complex plays a role in the ubiquitination of APOBEC3G enzyme, targeting it for proteolysis. In order to obtain the genotypes of the patients, genomic DNA was extracted from peripheral blood mononuclear cells (PBMCs). Enzymes from Life Technologies were added to degrade the membranes and nuclear envelopes of the cells in the cell pellets. After this process, only the cell’s nucleic acid is left behind in the vial. Genomic DNA is isolated after RNA digestion and magnetic bead selection. Specific enzymes hydrolyze the RNA molecules to destroy them, and then magnetic beads that respond to the charge of the DNA were added to extract the DNA from the solution. Finally, the Genomic DNA is then plated on microarrays with desiccated primers and probes placed within nanoliter reaction wells. Many nanoliter polymerase chain reaction (PCR) tests in a high-throughput microarray format to identify the individual SNP signatures were run. The experimenters used the Taq polymerase, which is naturally found in the thermophilic bacterial species Thermus aquaticus, because the enzyme only works in the high temperature range. In this way, they could effectively control the amount of DNA copying by controlling the temperature. Taq-polymerase and dNTPs (nucleotides) are added to the wells in conjunction with genomic DNA. The microarrays are then subjected to multiple cycles of heating and cooling. During the heating phases, the hydrogen bonds in the DNA are broken, then as it cools, the primers that identify the desired region bind. Then, the Taq polymerase binds, and creates a complementary strand going in the direction that the primer designated. After three cycles, the desired segment of DNA is isolated, and after many more trials, many copies of this desired segment are created. The experimenters isolated the regions of DNA that contained the desired SNPs by using specific probes purchased by the laboratory. Then, the experimenters used a Biotrove OpenArray NT Cycler to capture high-resolution images of fluorescent signal. The signals (VIC and FAM) are then quantified and translated to normalized values and stored on a standard flatfile. The OpenArray SNP Genotyping Analysis Software 1.0.5 was used to analyze the clustering resulting from the aggregate signal. This program plots each individual patient’s VIC or FAM signal in a scatter plot. From the plots, the experimenters determined the best-fit lines to make the best clusters. The clusters and the amount of points in the cluster showed the dominance of a certain genotype pairing according to VIC and FAM. Using the anonymous patient data acquired from the National Institute of Health (NIH) on differential SNPs in humans, genetic data with the patient’s matching phenotypes and their HIV progression rates was compared. A comparison between the host’s genotype and the rapidity of their HIV progression was made. To look for significance between the two sets of data, a univariate analysis and multiple variable logistical analyses of the specific variables found significant in the initial univariate analysis was used. Also, a false discovery rate (FDR) assuming independence to correct for multiple comparisons was applied to lessen the risk of obtaining a false positive correlation.

Results

Upon completion of data collection, the experimenters analyzed the single nucleotide polymorphism cluster and point frequencies, meaning the genotype data, with the phenotypic data obtained from the National Institute of Health. The initial p-values found in the experiment had to be corrected using a False Discovery Rate correction (generating q-values). This correction multiplied the initial p-value by the number of SNPs in the study to compensate for potential errors due to multiple testing. Since 125 individual SNPs were tested, the corrected p-values were all naturally very high. In the comparison between the patients’ phenotypes for various viral load set-points, the q-values were all above the threshold of .05. The lowest corrected p-value was about .37 (Fig. 1). In Figure 2, the experimenters found
the q-values in association with the CD4+ T-cell slope of decline within the patients. However, the lowest q-value of the figure was that of rs3736685, and that was .11 (Fig. 2). In Figure 3, the experimenters plotted the probability distributions of the RNA set-point and the inverse normal distribution of the SNP genotype data. From this, a qualitative assessment of the relationship between the two data sets was obtained. The slope of the first graph was nearly one (Fig. 3). In Figure 4, the association between the inverse normal distribution of the SNP genotypes and the CD4+ slopes of the patients was examined. In this figure, the slope was less than one (Fig. 4). Finally, in Figure 5, a linkage disequilibrium distribution of the 125 SNPs was created. The areas of red show SNPs that tend to be inherited together, and the blue show SNPs that do not tend to be present in the same host (Fig. 5).

After analysis, the experimenters found that all SNPs tested did not have a significant correlation with the viral set-point due to the high q-values that were above .05 (Fig. 1). Next, they also found that there was not a significant correlation between the SNPs tested and the patient CD4+ slope. There was one SNP that trended towards significance though, and that was in rs3736685. The corrected p-value after the False Discovery Rate test was 0.11 (Fig. 2). This value trends towards significance and it is one of the polymorphisms that influence APOBEC3G function. This means that there is a possible slight correlation between the SNP genotype of the patient and APOBEC3G function. In Figure 3, qualitative data was used to try to determine the relationship between SNP distribution and the viral RNA set-point. The slope of this graph was one, meaning that the two sets of data were equally dispersed (Fig. 3). In Figure 4, the slope of the probability distribution comparison of the SNPs and the patient CD4+ slopes were determined to have a slope smaller than one, meaning that the genotypes were more dispersed than the phenotypic CD4+ slope data (Fig. 4). Finally, the linkage disequilibrium plot showed that although the majority of the SNPs did not travel together during genetic replication, there were several of them that did, which are shown by the dark red coloration (Fig. 5). The overall trend in the experiment is a lack of significance and little linkage disequilibrium. In either case, though, there are notable outliers to the trend such as rs3736685.
Figure 2. This figure shows a comparison between the single nucleotide polymorphism tested in the study, and the subsequent p-values. The F-test was once again used to correct p-values due to multiple testing (Fig.1).

Figure 3. The quantile-quantile probability plot of the inverse normal distribution of the sample population and the RNA set-point.

Figure 4. The quantile-quantile probability plot of the inverse normal distribution of the sample population (the patients’ genotype) and his CD4+ T-cell slope.
Figure 5 A comparison of the linkage disequilibrium between the various single nucleotide polymorphisms examined in the study. The LD plot shows the relationship between the SNPs included. Areas of high-concordance are redder in color. These areas appear together in a population suggesting that these areas move together from person-person.

Discussion

These SNPs in this study were identified as potentially related to HIV progression. They either encoded for the proteins Cul5, elongins B and C, Rbx1 or APOBEC3G. As stated, Cul5, elongins B and C, and Rbx1 are human proteins that form a complex that is essential to the function of viral infectivity factor (VIF) in degrading APOBEC3G by ubiquitination. SNPs that encoded for the production of any of the five aforementioned proteins were labeled as potentially related to the progression of HIV. These were screened prior to the experiment and chosen to capture the genetic heterogeneity of the proteins associated with the complex. The experimenters were surprised that only one SNP was identified as having somewhat of an association with the progression of HIV. The experimenters had anticipated that the nature of the experiment would allow the finding of multiple SNPs that were significant, especially considering a similar study by the advisor of around 130 SNPs had found nine SNPs associated with HIV progression.

The one SNP that was found to be partially associated with HIV progression as measured in CD4+ T-cell count slope was rs3736685. Measuring HIV progression through CD4+ T-cell count is common considering the virus infects these cells. A diagnosis of AIDS is determined by CD4+ T-cell count. The association between the SNP and progression had an uncorrected p-value of 0.0009. However, the standard of significance was changed in order to adjust for multiple comparison problems. The adjusted q-value was 0.11. Although this is not considered statistically significant, it shows a strong trend towards significance (it is close to .05).

The one SNP trending towards significant association, rs3736685, is one of the genes and encodes for the production of APOBEC3G. The gene can contain a thymine homozygote, cytosine homozygote, or thymine-cytosine heterozygote. A thymine homozygote is most prevalent, occurring in 95% of all Caucasians. 4.5% of all Caucasians express the heterozygote and about 0.5% express a homozygous cytosine genotype. Similar genotypic ratios are present in other racial groups as well. Rather than encoding for proteins that were part of a complex that affects the functionality of VIF in degrading APOBEC3G, this SNP encoded for the protein degraded by the VIF-induced proteasomal complexes.

Although one SNP was found to associate with HIV progression, a greater contribution lies in that the experimenters found 124
SNPs that had no association. There are approximately three billion nucleotides in the human genome, three million of which express genetic variation and hundreds of thousands of which affect phenotypic expression. On that note, there are thousands of SNPs that could potentially affect HIV progression\textsuperscript{11}. It will be a long and arduous process to determine all of the significant SNPs. This project will require collaboration on the part of multiple laboratories and many researchers. The experimenters have contributed to this study by eliminating many from consideration.

Studies have found that VIF interacts with a complex of Cul5, elongins B and C and Rbx1. Variation amongst this protein complex affects the effectiveness of VIF in degrading the APOBEC3G protein, a human enzyme that inhibits HIV replication\textsuperscript{5}. Rs3736685 is a SNP within a gene coding for the APOBEC3G protein and was approaching association with HIV progression. This suggests that not only does the aforementioned complex affect HIV progression, but also that the APOBEC3G enzyme affects it as well.

The future work associated with this study is apparent. Throughout this research, genotypic information for 125 SNPs for five thousand individuals was obtained. As a result, a data set with approximately 625,000 data points was created. Further research would include identifying more SNPs that encode for the proteins APOBEC3G, Cul 5, Rbx1, and elongins B and C and have an association with HIV progression. Screening and identifying these will be tasking in both time and resources, but a collaborative effort amongst researchers is already occurring. As more SNPs are identified that have an association with HIV susceptibility, a more effective anti-viral can be developed\textsuperscript{10}. The individual mutations have been collated into haplotypes and networked into large haplogroups, and a more significant initial association has become apparent with the additive effects of SNPs grouped within genes and amongst genes.

### References


### Acknowledgements

Kevin Emancipator, Jack Michuda, and Sudhir Penugonda