Signatures of Mutational Processes in HIV

Shefali Qamar and Robert Missirian

Institute of Computational Biovirology
University of California, Santa Cruz

Abstract

Motivation: Our objective is to generate and classify significant signatures of mutational processes that influence the high incidence of single-nucleotide polymorphisms observed in the HIV virus. We hope to contribute to the development of effective antiviral drugs for HIV by studying the evolution of these process-based signatures and attributing the signatures to known agents.

Results: The expected result is to see signatures indicating the activity of viral HIV reverse transcriptase and the human APOBEC mechanism in comparisons to the reference genome.

Contact: sqamar@ucsc.edu; ramissir@ucsc.edu

Project Github: https://github.com/shefaliqamar/HIVMutationSignatures

1 Introduction

HIV is a retro-virus that causes over 1 million deaths annually. While drugs that keep HIV-infected patients in stable condition now exist, they can be foiled by mutations in the retroviral genome that lead to drug resistance. A high mutation rate is the foundation of HIV's versatility, and thus understanding the mechanisms which create the mutations is essential to understanding the virus itself.

In recent research, the Ludmil Alexandrov lab at the University of California, San Diego used a multitude of machine learning techniques, namely source separation by non-negative matrix factorization and k-means clustering, to generate unique signatures for different mutational processes in cancer. These processes included smoking, exposure to UV light, etc. for various types of cancer. Each signature is represented by a set of mutation types which the specific mutational process is likely to induce. [2] This method has successfully attributed about half of the significant signatures to a known mutational agent. To generate these signatures, the method uses mutation data from genomes of patients diagnosed with cancer. Application of this technique to HIV mutational data may provide similar results.

There are several dissimilarities which arise between HIV and Cancer. HIV has a much higher rate of mutation, as high as 4.1 × 10-3 substitutions per site per year (s/s/y) [5]. It is also a virus, so we are inspecting RNA instead of DNA. Methods developed for HIV must account for this difference.

For HIV, there exist two known mutational agents which may yield signatures in such an analysis: APOBEC and reverse transcriptase. The human APOBEC mechanism is an editing enzyme which identifies and attacks viral RNA by inducing single nucleotide polymorphisms in sequences. [8] Reverse transcriptase is the error-prone polymerase used in viral RNA replication. Unlike DNA polymerase, RNA polymerase cannot proofread the transcribed strand for accuracy, which induces mutations by incorrectly replicating viral RNA. [1]

The aim of our analysis is to utilize HIV patient information gathered by next-generation sequencing in order to generate mutational signatures and gain insight into HIV's mutational processes. We hope to provide this analysis as a basis for further studies to identify significant mutational processes which create the variants of HIV undergoing selection toward drug resistance.

2 Methods

2.1 Project Goal and Structure

The goal of our project was to find statistically significant, stable mutational signatures in HIV and associate them with their corresponding mutational processes. HIV presents unique challenges to obtaining useful signatures, as it has a high mutation rate that varies in different parts of the genome. [6] Furthermore, patient studies have different sections of the genome sequenced, resulting in potential confounding of the data. However, the processes in HIV are likely more evident, as its only known major mutational processes are APOBEC and reverse transcriptase. By optimizing the data to use partial genomes inside which every base pair has been sequenced, we hope to generate meaningful mutation signatures. We tried to accomplish this goal by following this procedure:

1. Mutation Frequency Extraction & Laplace Smoothing
2. Localization
3. SigProExtractor: Blind source separation by non-negative matrix factorization and k-means clustering
4. Analysis of stability and reproducibility
5. Comparison to Literature: Verify our results with existing identified mutation signatures

2.2 Mutation Frequency Extraction & Laplace Smoothing

Our first step in the process of extracting mutation signatures was parsing through our patient data in order to produce large frequency tables of mutation types, called genome catalogues. These tables would eventually feed into the Alexandrov Lab's provided SigProfilerExtractor tool.

The patient data which we were working with included for each patient: an HIV genome nucleotide sequence from the segment of the genome which coded for viral reverse transcriptase, a study number, etc.. Unfortunately, the portion of the genome for which the sequence was provided varied from study to study and often did not align with other studies.

The sequenced portion of each patient genome was compared to the given reference genome in order to identify single nucleotide polymorphisms (SNP) such as A>G, C>T, etc.. The frequency of each SNP was recorded in context of the two nucleotides surrounding it in the reference genome (mutational context, ex. CAT>CTT for an A>T SNP). These were then populated into the patient's genome catalogue. [2] Due to the single-stranded form of RNA, we hypothesized that the 6-nucleotide-substitution- types assumption made for double-stranded DNA would not generalize well, and a 12-SNP-types alphabet would be required for RNA (ex. A>G != T>C in RNA SNPs). Formatting forced us to discount SNPs not in the predefined DNA alphabet.

Optimally, we would have a large enough dataset to find a significant amount of each type of SNP in each possible context. The tool into which we place the genome catalogue also requires that each frequency has a certain minimum value. [3] However, this was not the case, and we found that our data was very sparse due to the small dataset. (Although we had data on over 1750 HIV patients, the length of the genome was very short compared to that of a cancer patient's. [4] Thus, SNPs were much fewer, even given HIV's high mutation rate.)

To fix this issue, we employed Laplace smoothing over the frequencies. Laplace smoothing is a machine learning technique commonly used to solve the problem of zero-probability for predictive models such as artificial intelligence by pretending to see each possible value a few more times.

$$P_{LAP,k}(x) = \frac{c(x) + k}{N + k|X|}$$

Equation 1a: Laplace's estimate formula. x is any given SNP in a given context type (ex. A>G, CAT). c(x) is the count of x: how often we see this SNP in the genome. N is the total number of mutations in the genome. k is the strength of the prior.

2.3 Localization

As aforementioned, the sequenced portion of the genome varied from study to study and often did not align with the sequenced portion in other studies. Thus, we created several genome catalogues split by study in order to maximize information gain from the SNP frequencies. This has the added benefit of taking into account the fact that different mutational processes may act on different sections of the genome, and are thus more clearly able to express signatures when a patient study including one of those regions is examined.

2.4 SigProfilerExtractor

The SigProfilerExtractor tool is readily available from the Alexandrov lab, built to extract mutation signatures from mutational catalogues. In order to do this, it performs two machine learning techniques to separate the frequencies into distinct signatures:

1. Blind source separation by non-negative matrix factorization
2. K-means clustering

Blind source separation is a technique used to separate out signals from a mixture of sources. Non-negative matrix factorization (NMF) accomplishes this in context of our model for mutational signatures. By iterating over multiple samples, NMF is able to create a model of mutation signatures and frequencies of mutations within that signature, only given information about the overall genome catalogue.

$$\begin{bmatrix} m_1^1 & m_2^1 & \cdots & m_{G-1}^1 & m_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ m_1^K & m_2^K & \cdots & m_{G-1}^K & m_G^K \end{bmatrix} \approx \begin{bmatrix} p_1^1 & p_2^1 & \cdots & p_{N-1}^1 & p_N^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ p_1^K & p_2^K & \cdots & p_{N-1}^K & p_N^K \end{bmatrix}$$

$$\times \begin{bmatrix} e_1^1 & e_2^1 & \cdots & e_{G-1}^1 & e_G^1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ e_1^N & e_2^N & \cdots & e_{G-1}^N & e_G^N \end{bmatrix}$$

Equation 1b: M = mutation in genome catalogues [1...k]; P= Probability of mutational process with signature [1...n] to cause mutation [1...k]; E= Number of mutations with process signature [1...n] in genome catalogue [1..G]. [2]

Given matrix M as the mutational catalogue, NMF predicts matrices P and E. This is run multiple times over each patient. [3]

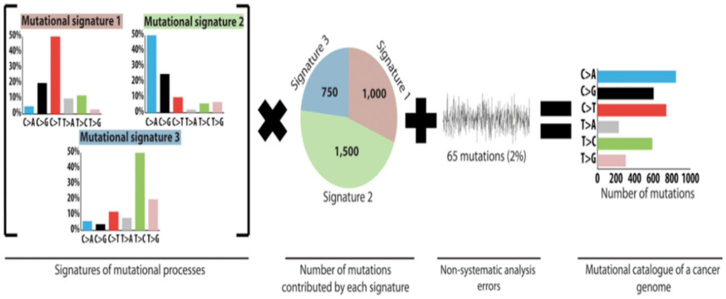

Figure 2a: Visualization of NMF for mutation signatures. Mutation signatures multiplied in correct proportions add up the overall catalogue. (Alexandrov) [2]

Once the signatures were extracted, the tools runs k-means clustering, a method of analyzing clusters of data, in order to determine the stability of each signature. It does this by clustering results of NMF to determine reproducibility of each signature, which it then denotes as stability. This is an extra step taken to ensure accuracy of the results.

3 Results

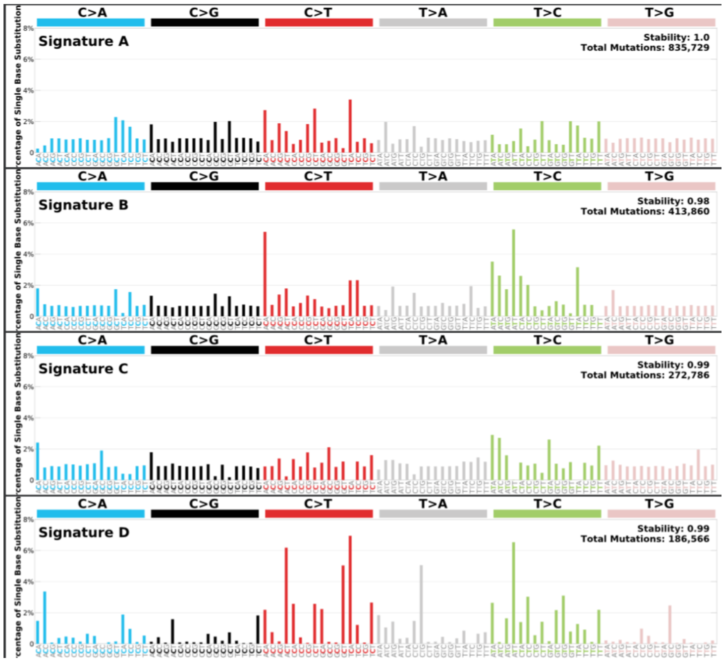

Figure 3a: Mutation signatures over all HIV patient genome catalogues. (Laplace smoothing with k = 0.3.)

3.1 Analysis of stability and reproducibility

SigProExtractor does not provide us with the number of mutation signatures present across the genome catalogues. Rather, it presents us with a suggested solution and signature extractions for numbers of signatures *n* = 1, 2, 3, ...10*.* It runs NMF on these, adjusting the matrix sizes to n fit n processes. In the end, we are left with a graph denoting the stability of each number of signatures. We want to maximize the stability, or average signature reproducibility. We also want to minimize the reconstruction error, which is the irreducible error between the original genome catalogue and a simulated catalogue generated by summing over all mutation signatures. [2]

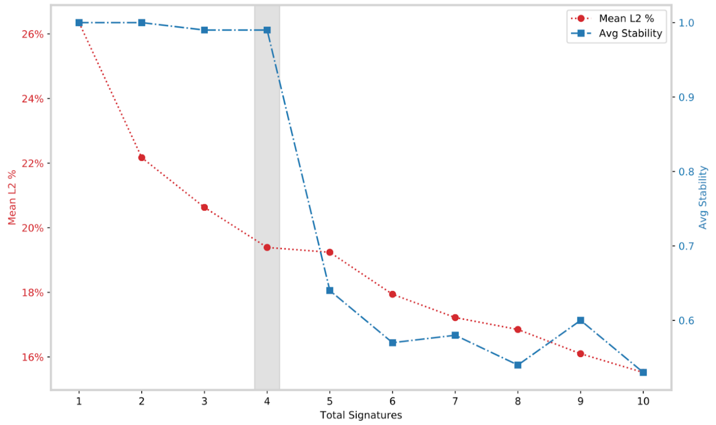

Figure 3b: Selection plot for mutation signatures over all HIV patient genome catalogues. (Laplace smoothing with k = 0.3.). Red line = reconstruction error, Blue line = average signature reproducibility. Optimal number of signatures is 4, thus this was chosen for Figure 3a as the result for this data.

3.2 Comparison to Literature

The last step of each run is to verify our results with existing identified mutation signatures. From literature, we extracted versions of the signature of the human APOBEC mechanism and try to match it to one of our produced signatures.

3.3 Bugs Fixed

Initially, we obtained two highly reproducible, independent mutation signatures. When we attempted to cross-reference, the results were very promising, as we found matches for both mutation signatures in Alexandrov's cancer signatures. However, questions remained. How did the signatures line up when we only graphed half of our mutations due to the limited alphabet in the Alexandrov code? Furthermore DNA mutations count either as the actual substitution or it's complement, depending on whether the substitution is one of the six in the graph. There are sufficient differences in the properties of our HIV data to warrant scepticism of the immediate clarity we saw in the results.

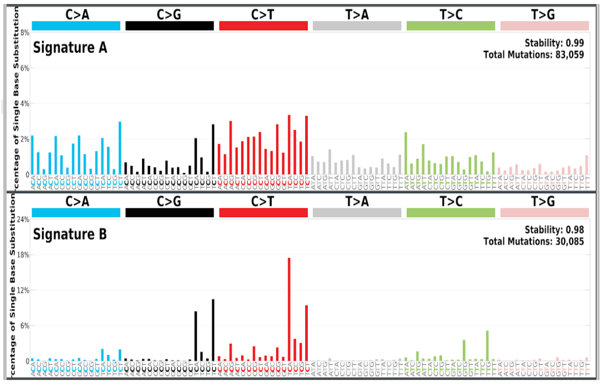

Figure 3c: Our first mutation signatures

When we ran SigProfiler on preprocessed data in confounding ways, we realized that regardless of what data we ran the code on, the results for 2 signatures and above always came out the same. For example, we got the signatures from Figure 3.1 even when we added the mutations in each sample to the sum of all the previous samples. The perfectly matching signatures were the result of a bug, as they were generated on hard-coded example data despite the real data being passed in as a program argument. The real signatures actually look like Figure 3d, which we obtained after fixing the bug.

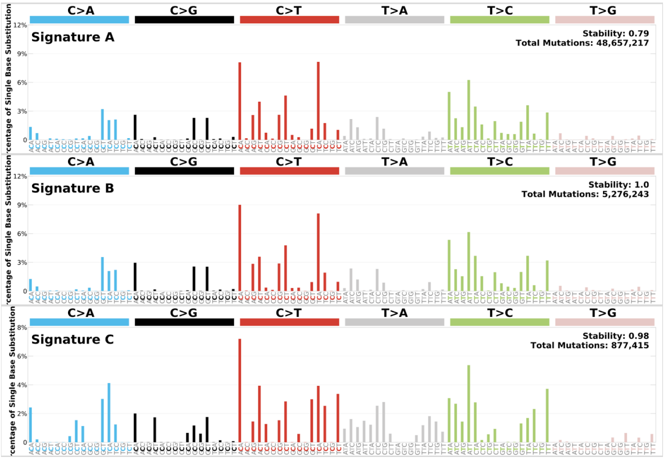

Figure 3d Real signatures from cumulative addition.

We also encountered a bug that only allowed us to run 143 samples, but it only occurred when run on sparse data. As a workaround, we first attempted to run the analysis on single studies with some success. From this approach, we obtained at least two reasonably independent signatures from a set of 40 samples in a study that sequenced patients' entire genomes. This data represents the clearest signatures we obtained using limited datasets.

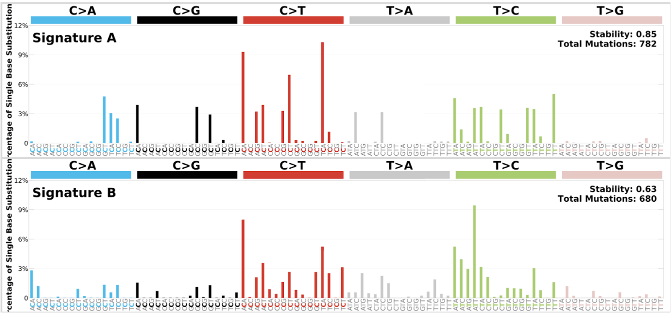

Figure 3e: 40 patients from a full genome study.

Since this bug severely limited our sample size, we tried several workarounds to the increased density of the genome catalogue, ultimately culminating in the Laplace Method, which proved more effective.

4 Discussion

Despite bug-related difficulties, our results give substantial insight into understanding mutational processes in HIV as well as the interpretability of the signatures themselves.

4.1 Interpreting Signatures

Firstly, we generated a wide variety of mutation signatures in our many attempted methods, the graphs of which have interesting properties. Comparing the mutation signatures in the upper section, we see that the spikes fall mostly in the same places across each signature. This suggests that the signatures are structurally dependent, and seem to be describing the same process. The bottom signature is smoother and seems to have a few more differences, but we can also see that it describes a lot less data. This falls into a pattern that we also see in the lower section of Figure 4a, where the bottom signature retains most of the structural patterns. Again, it is more filled in due to an increase in the frequency of some base pairs that are at negligible levels in the upper signatures.

Comparing the upper and lower sets of signatures in **Figure 4a**, both sets seem to have similar, although not identical, features. To begin, there are substantially more mutations in the C>T and T>C categories than in the surrounding categories. Structures that we see in common are a blue spike on the right side of the C>A categories; the three spikes in the C>G category of the first bar of the upper set; a red spike at the start of the C>T, followed by three islands separated by spaces; the descending pattern from the first bar of T>C to the third, then by a spike in the fourth; An isolated bar at the end of T>C preceded by an island of three with a space in between; unsubstantial activity in T>G.

The ubiquitous nature of these structures suggests that they describe a collective property of all the mutations, not one specific signature. This implies a larger force at play such as the polymerase style SNP generation by reverse transcriptase.

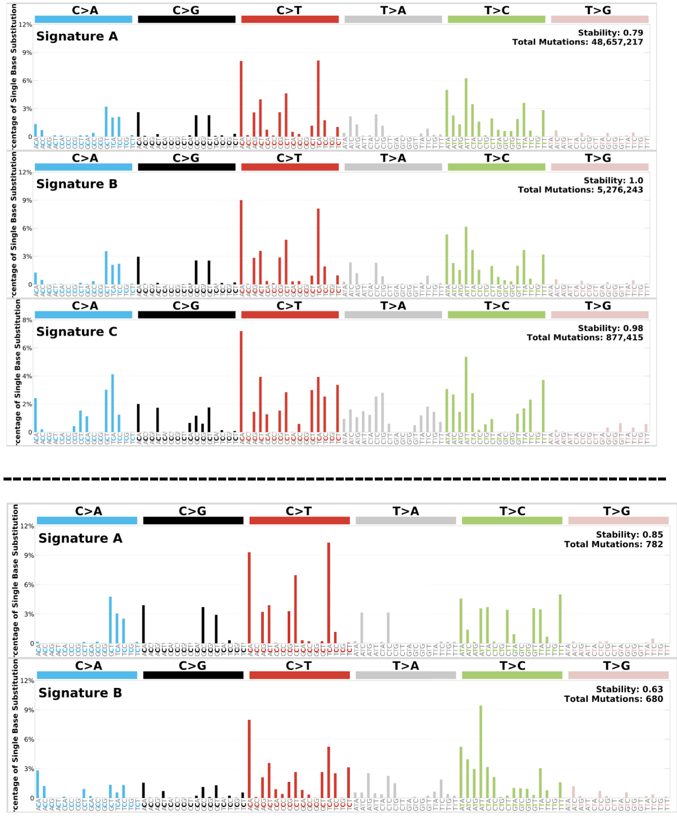

Figure 4a: Comparison between early results

4.2 Profiling HIV mutational processes

In contrast, the Laplace Smoothing results demonstrate the same patterns, but distributed with structural individuality across the signatures. For example, we still see spikes at ACA, CCT, ACT, and TCA, but each spike is present in at most two signatures. We also see the previously mentioned pattern in the first four bars of T>C, but it only occurs in one signature. The spike in the fourth bar is present in two. Generally, if a feature is present in two signatures, it is one of their only similarities. Thus, it is apparent that NMF has finally extracted distinct processes from the data. Furthermore, the stability of each signature is .98 and above, which is a higher standard than our earlier attempts. The only structural similarity across three signatures seems to be a certain base level of frequency, due to the smoothing function.

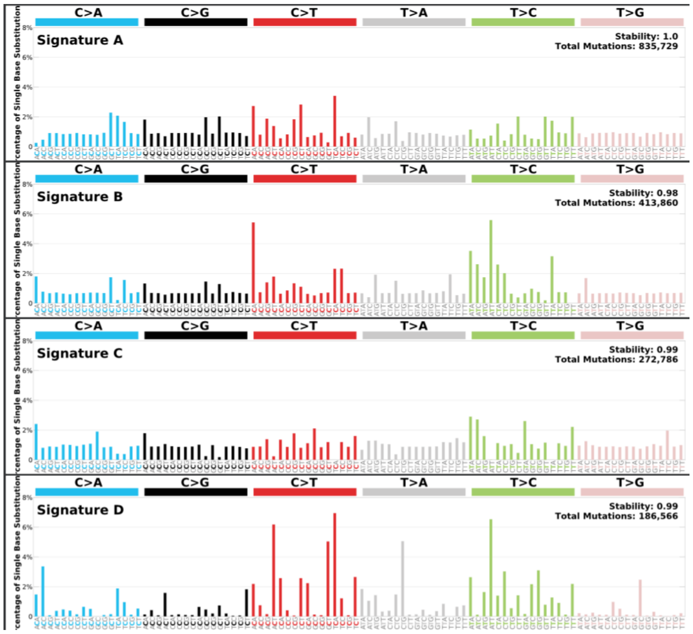

Figure 4b: See figure 3a.

It is possible that in our previous attempts to circumvent the sample limiting bug to run the program on the full dataset, we confounded the original patterns in the data. With smoothing, the data patterns could be preserved while still filling in the zeros with small values. Even now, the structural similarity between the base frequencies of the top three signatures suggests that the data was still confounded slightly, just not enough to wash away the real signatures.

4.3 Identifying HIV mutational processes

Despite obtaining the signatures to a high degree of accuracy, it will be non-trivial to map them to known processes. It is biologically dubious to compare our signatures derived from RNA directly those derived from DNA, as mutation signatures generally are, although is likely some correlation. After all, a DNA mutation is typically grouped into either its actual substitution or its complement, depending on which is in the table, as both are equally valid. In contrast, an RNA mutation is unique, since as RNA is single-stranded, it has no complement attached. In our current embodiment, half of the RNA mutations aren't even included in the table. Due to these confounding factors, we will probably need to go by literary accounts of each enzyme to map our signatures to mutational agents. If other researchers find it useful to use mutational signatures specifically on RNA, we might be able to comparatively deduce the identity of the mutation signatures in conjunction with existing literature, as has been done for the Alexandrov signatures.

4.4 Avenues for further research

Beyond our existing results, it is probable that related further research can extend our understanding of the processes described by these signatures. It is possible that we can even identify the signatures with further experimentation.

Currently, we have created a tool that automatically segments the genome into subdivisions and only catalogues the samples that are fully sequenced in each region. Unfortunately, we ran out of time to run the analysis on these catalogues, but when we do, it will allow us to break down variations in the mutational processes affecting each part of the genome.

By filtering for specific strings contained in the genome, we could potentially identify any mutation signatures that are caused by APOBEC. This would require the alignment number and contents of the sequence vif, which is known to prevent APOBEC from acting on certain strains of HIV [7][9]. If we can determine whether a specific genome exhibits high vif by enumerating the mutation rate observed within each patient genome, we could run the analysis on only high-vif genes so that the remaining signatures are caused by something other than APOBEC, and the missing signatures are caused by APOBEC. The only mutational agent we are studying for HIV other than APOBEC is reverse transcriptase. Thus, if only one signature remains, we could assume that signature to be reverse transcriptase and the rest to be subprocesses of APOBEC [9].

Changing the mutational contexts to include more base pair combinations would yield more accurate and informative results, but it would also present challenges even beyond our existing issue with the sparsity of the data. Expanding the alphabet in this way involves editing a large portion of the SigProfilerExtractor tool's codebase, including how it reads and interprets input data, which would take a significant amount of further time to completely implement.

5 Conclusion

We have made substantial progress in identifying and fixing critical bugs as well as generating four statistically significant signatures. Furthermore, we have several simple avenues for further research, and a general path to long-term validation of our results via comparison to literature.

Acknowledgements

We would like to thank Jay Kim and Dr. Manel Camps from the University of California at Santa Cruz, and S. M. Ashiqul Islam from the University of California, San Diego.

References

1. Abram, M. E., et al. "Nature, Position, and Frequency of Mutations Made in a Single Cycle of HIV-1 Replication." *Journal of Virology*, vol. 84, no. 19, 21 Oct. 2010, pp. 9864–9878., doi:10.1128/jvi.00915-10.
2. Alexandrov, et al. "Deciphering Signatures of Mutational Processes Operative in Human Cancer." *Cell Reports*, vol. 3, no. 1, 31 Jan. 2013, pp. 246–259., doi:10.1016/j.celrep.2012.12.008.
3. Alexandrov, et al. "Abstract IA11: Signatures of Mutational Processes in Human Cancer." *DNA Repair Gene Mutations in Cancer Genomes*, 22 Aug. 2017, doi:10.1038/nature12477.
4. Alexandrov, Ludmil B, and Michael R Stratton. "Mutational Signatures: the Patterns of Somatic Mutations Hidden in Cancer Genomes." *Current Opinion in Genetics & Development*, vol. 24, 2014, pp. 52–60., doi:10.1016/j.gde.2013.11.014.
5. Andrews, Sophie M., and Sarah Rowland-Jones. "Recent Advances in Understanding HIV Evolution." F1000Research, vol. 6, 2017, p. 597., doi:10.12688/f1000research.10876.1.
6. Cuevas, et al. "Extremely High Mutation Rate of HIV-1 In Vivo." *PLOS Biology*, vol. 13, no. 9, 2015, doi:10.1371/journal.pbio.1002251.
7. Delviks-Frankenberry, Krista A., et al. "Minimal Contribution of APOBEC3-Induced G-to-A Hypermutation to HIV-1 Recombination and Genetic Variation." *PLOS Pathogens*, vol. 12, no. 5, 2016, doi:10.1371/journal.ppat.1005646.
8. Kim, Eun-Young, et al. "Human APOBEC3 Induced Mutation of Human Immunodeficiency Virus Type-1 Contributes to Adaptation and Evolution in Natural Infection." *PLoS Pathogens*, vol. 10, no. 7, 31 July 2014, doi:10.1371/journal.ppat.1004281.
9. Okada, Ayaka, and Yasumasa Iwatani. "APOBEC3G-Mediated G-to-A Hypermutation of the HIV-1 Genome: The Missing Link in Antiviral Molecular Mechanisms." *Frontiers in Microbiology*, vol. 07, 2016, doi:10.3389/fmicb.2016.02027.