EFFECTS OF SEQUENCING RANGE AND DEPTH ON THE INTERFERENCE OF FITNESS EFFECTS OF MUTATIONS IN HIV EVOLUTION

By

Eric K. Ngo

A capstone project submitted for
Graduation with University Honors

May 30, 2020

University Honors
University of California, Riverside

APPROVED

_____
Dr. John Barton
Department of Physics and Astronomy

_____
Dr. Richard Cardullo, Howard H Hays Jr. Chair, University Honors

Abstract

Infection with human immunodeficiency virus (HIV) leads to serious illness for millions, attacking their immune system and rendering it weak and vulnerable to opportunistic infections. Due to its high mutability, HIV is capable of evolving drug resistance and escaping from immune control. Studies have shown that HIV can accumulate mutations that allow it to escape recognition from CD8+ T cells. At the same time, mutations also occur in parts of the viral genome which are not targeted by T cells. We estimated the fitness effects of HIV mutations by analyzing viral sequencing data collected longitudinally from 13 different patients using previously developed computational methods. In particular, we compared fitness effects estimated from a smaller number of half-genome-length sequences and from deeper, focused sequencing of the HIV envelope protein. The typical absolute difference between selection coefficients was small, suggesting that different data sources give broadly similar results. However, a minority of genetic variants differed substantially in their estimated effects on fitness. This work will provide insights into effects of sequencing range and depth on the inference of natural selection from genetic data, and a better understanding of the fitness landscape of HIV.

Table of Contents

Acknowledgment

I would like to first thank my principal investigator and faculty mentor, Dr. John Barton, for his unwavering support and guidance throughout my capstone project and undergraduate career. This was an unfamiliar area of research for me, and he has patiently taught me analytic physics throughout this project. I would also like to thank graduate student, Yunxiao Li, for teaching me how to code in Python and for his willingness to work the long hours with me on this project. I also want to express my gratitude to Dr. Louis Santiago for introducing me into the world of research and proving me the necessary skills and knowledge to execute a project. Furthermore, I thank Dr. Stephanie Dingwall for her mentorship and wisdom that she has imparted upon me throughout my undergraduate career. Lastly, I want to thank my friends and my family for their endless love and support throughout this journey. I am honored to have met so many wonderful people that have helped me grow and become the person I am today.

Rationale

The creation of a global HIV vaccine is currently blocked due in part its viral diversity. Its reverse transcriptase frequently generates mutations, allowing different strains of HIV to be created in an individual and cause further immune escape. The difference between strains of HIV from multiple infected individuals makes the genome difficult to study as each viral strain provides new information in its evolution of escaping certain drug resistance. The complexity of HIV evolution makes it challenging to understand how different mutations affect the virus, and how to limit its ability to escape from the immune system or drug treatment. In most cases, complete viral genome data may not even be available to study, giving only certain parts of the genome such as the Env region, the viral gene that encodes the protein responsible for viral entry. Expression of this gene allows for HIV to target and attach onto CD4+ T cells (T helper cells), in order to enter the target cell and replicate.

Literature Review

HIV causes the immune system disfunction that leads to AIDS, which leads to opportunistic infections of the immune system. HIV is dangerous because it infects the CD4+ T cells, which are important for defending against infectious viruses and diseases. The article *The T-Cell Response to HIV* by Bruce Walker and Andrew McMichael talks more in depth for the functions of the CD4+ and CD8+ T cells in HIV infection and potential counter measures to the virus. They layout a view of the pathogenesis HIV with clear signals in how the CD8+ T cells initially try to control the virus, but the effects of HIV still render the cell functions dysfunctional. In the case of CD4+ T cell, HIV actively infects CD4+ T cells and eliminates them from the body directly, rendering it useless. Bruce Walker and Andrew McMichael also explore the effects of mutations on the viral fitness. They look at the relationship between the

viral load and the class I restricted cytotoxic T lymphocytes (CTLs), since they were viewed as a major cause for the virus progression. It was observed that viral escape from CTL response was done through multiple mechanisms and mutations. The rate of CTL efficacy was impacted by the mutations induced during viral replication. It was shown that "responses to Env, which readily tolerates mutations without affecting the viral fitness, is associated with higher viral loads."

Evidence has shown that HIV can escape recognition from CD8+ T cells which can kill infected cells. *Vertical T cell immunodominance and epitope entropy determine HIV-1 escape* by Michael K.P. Liu and others examined aspects of T cell functions such as immunodominance, which is "the magnitude of the response relative to the total response within each patient at a given time point (Liu, 2012, p. 4)," and their effects on virus control. The studies done have shown data on how certain factors of the T-cells such as functionality, exhaustion, and receptor responses have each played separate roles in how HIV responds to T cell pressures. One data set explored the effects of the HLA B*5801-restricted T cell response from different patients on the Gag 240 – 249 epitopes, showing that some patients had a faster escape compared to others. An explanation for time differences was that there were "upstream Gag mutations, H219Q and I223V that occurred and facilitated a more rapid escape." This shows that the viral sequence background can affect escape dynamics.

The viral sequencing data I collected compares the analysis between methods of looking at the 3 prime half-genomes of HIV versus only the envelope region of the virus through the usage of the selection coefficients, which quantify how efficiently the virus can replicate. The article *Resolving genetic linkage reveals patters of selection in HIV-1 evolution"* by Muhammad S. Sohail, Raymond H.Y. Louie, Matthew R. McKay, and John P. Barton talks about using Marginal Path Likelihood (MPL) as a method of investigating patterns of selection in HIV-1

evolution through the usage of genetic time series data. Through statistical physics, a probable evolutionary path of the virus could be obtained through looking at the "probability of changes in the mutant allele frequencies between successive generations." This method of study was used in my research to compare the selection coefficients between the 3'-half-genome and envelope region.

Methods

Gathering Experimental Data. For this experiment, we obtained fourteen different patient's viral genome data from the hiv.lanl.gov website. We chose these specific patients because the regions of the virus targeted by T cells were experimentally determined in *Vertical T cell immunodominance and epitope entropy determine HIV-1 escape*, which in principle would allow us to better understand how the virus evolves to escape T cell responses. We obtained two sets of viral data, one containing a larger breadth of sequencing (3' half-genome) and the other containing the envelope region. The viral sequencing data of each patient contains the nucleotide sequences and the label composing of the name, sampling year, days from first sample or days from seroconversion (the development of detectable antibodies in the blood and officially becoming HIV-positive), and the HXB2 Reference Sequence. The days from first sample represents the number of days between the first sample received to the current sample while the days from seroconversion represents the number of days between the patient's seroconversion and the day that the sample was taken.

Aligning Experimental Data. The viral sequencing data obtained from the hiv.lanl.gov website was unordered, requiring re-organization of the data through the coding language, Python. Through CHAVI.py, the viral data was time ordered based on a combination of the days from seroconversion and days from first sample. While the days from seroconversion allows

tracking of the genome when there is HIV-positivity, not all sequences are paired with this information. In order to use as much of the viral data as possible, the time information found in viral data containing "Days from Infection" and "Fiebig stage" were used to infer the time information. For example, for patient CH040, viral data CH40E-N14 contained both time information however, viral data CH40E_TF6 only contained "Fiebig stage" information. Since the Fiebig stage and Days from Seroconversion matched with CH40E-N14, the "Days form Infection" is inferred and able to incorporate this data into the time ordering. Once both the complete genome viral data and envelope region data were time ordered, the data needed to be converted into MPL analysis format.

Setting Data for MPL Analysis. MPL looks at how mutations on certain sites would affect the overall fitness of the sequence or the protein. MPL analysis needs the sequences to be time ordered and for the mutant alleles to be identified and represented in a numerical format. This was done for the ordered genome data of the full genome and only envelope. Afterwards, the newly formatted data was sent to Professor Barton to be analyzed through MPL to receive a new data set containing the epitope, frequency, s_MPL (selection coefficients via MPL method), and s_Sl (the selection coefficients at the Single Locus).

Comparing Envelope only data vs. Full Genome data. The s_MPL (selection coefficients) of each patient allows for interactions of a specific region of the genome against the rest of the genome sites. Python was used to compare difference in s_MPL between the envelope only genomic data against the full (or majority) genomic data. The selection coefficients of the data sets were categorized as deleterious ($s \leq -.003$), neutral ($-.003 \leq s \leq .003$), and beneficial ($s > .003$). Only selection coefficients with matching HXB2 tags between the two data sets were taken for analysis. The selection coefficients of the envelope and half-genome were compared to

view changes in mutation category as well as taking the difference between the two selection coefficients. Data points containing large changes in selection coefficients (≥ .01) and mutation categorial changes were taken into account for further data analysis. To view the differences in data, a scatter plot was made using Matlab to compare the selection coefficients between the two data sets. Data points deviating from the line of best fit were used for analysis. Selected data points were then manually checked from the MPL data.

Further analysis of Data Comparison. Using Matlab and Python, the difference in selection coefficient data analyzed. This was done by checking the average difference in selection coefficients for each patient, creating a histogram of the data based on the selection coefficients and absolute value selection coefficients. Frequency trajectories of the data were also taken to check whether the number of time points influenced selection coefficient differences.

Results

Envelope Mutation Dynamic	Three Prime Mutation Dynamic	MPL Envelope	MPL ThreePrime	MPL Difference
-1	1	-0.01052526	0.009328968	-0.019854228
0	0	0	0	0
-1	-1	-0.0059675	-0.00275255	-0.00321495
0	0	0	0	0
-1	0	-0.00362858	-0.000374273	-0.00325431
0	0	0	0	0
-1	-1	-0.00681777	-0.001655903	-0.005161866
...	...	...	...	...

Table 1. *Patient 040 early selection coefficient data.* The MPL of the ThreePrime and Envelope are compared and are labeled as either deleterious (-1), 0 (neutral), or (1) beneficial. Differences between

the selection coefficients were taken. Large selection coefficients were noted as well if there were differences in the two data sets' mutation dynamics. One large difference in selection was noticed with Envelope being deleterious while Three-Prime being beneficial, with a difference of 0.19854228.

Envelope Mutation Dynamic	ThreePrime Mutation Dynamic	MPL Envelope	MPL ThreePrime	MPL Difference
0	0	0	0	0
0	0	0.014043778	0.018143939	-0.004100161
0	0	0	0	0
-1	-1	-0.016387756	-0.003719948	-0.012667808
0	0	0	0	0
0	0	0.002426757	0.011574301	-0.009147544
0	0	0	0	0
...	...	...	...	...

Table 2. *Patient 077 early selection coefficient data.* The MPL of the Three Prime and Envelope are compared and are labeled as either deleterious (-1), 0 (neutral), or (1) beneficial. Differences between the selection coefficients were taken. Large selection coefficients were noted as well if there were differences in the two data sets' mutation dynamics. No changes in the mutation dynamic can be seen in the early data sets for Patient 077.

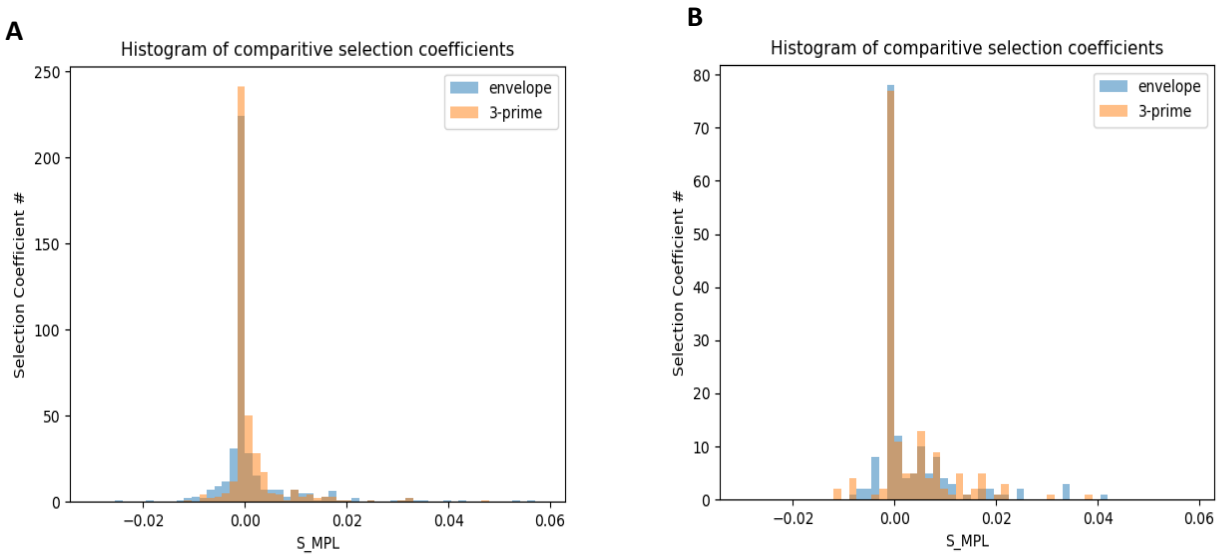

Figure 1. *Patient 040 and 077 Histogram.* Figure A. Patient 040 Histogram. Figure B. Patient 077 Histogram. Compares the number of selection coefficients that are similar or different in the selection

coefficient range. Results show large similarities and amounts of selection coefficient that are overlapping around the -0.01 – 0.02 range in Figure A. Highest peak was 250 selection coefficients around the 0.00 selection coefficient number in Figure A. In Figure B, results show large similarities and amounts of selection coefficient that are overlapping around the 0.00 – 0.02 range. Highest peak was 77 selection coefficients around the 0.00 selection coefficient number in Figure B. Displays how similar the same selection coefficients are between the two data sets.

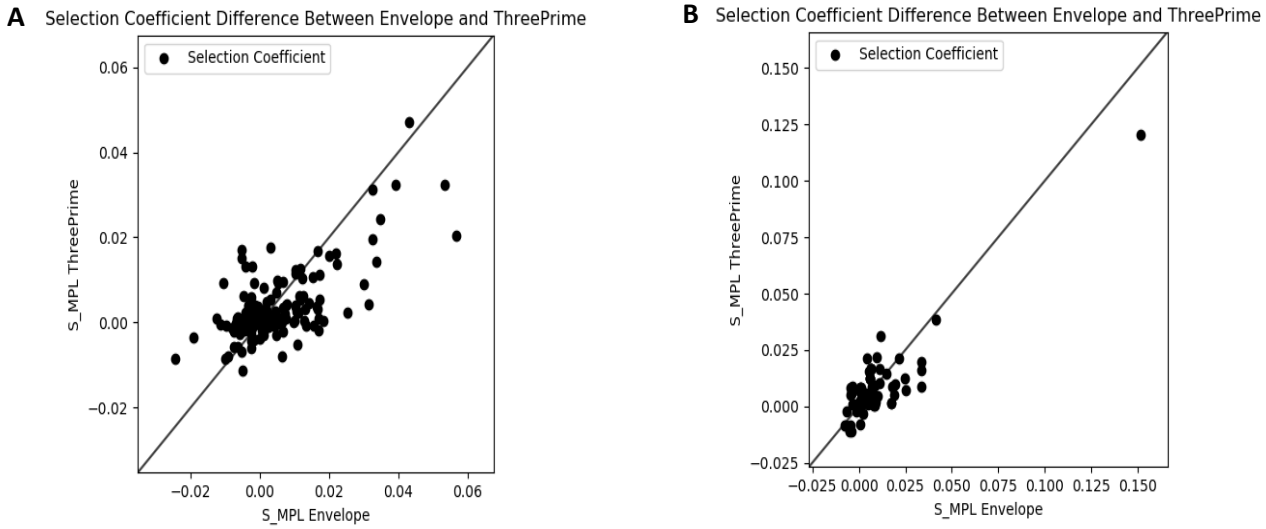

Figure 2. *Patient 040 and 077 Scatterplot*. Figure A. Patient 040 Scatterplot. Figure B. Patient 077 Scatterplot. Comparing the data between the selection coefficients found between the two data sets. The line of best fit represented how closely similar the selection coefficients were to each other. Points deviating far from the line of best fit correlated to large selection coefficient differences and potentially differences in mutation dynamics. At least four points per patient were noted off and recorded to see if they were in T- cell epitopes. In Figure A., later amounts of differences were found deviating more from the envelope selection coefficient compared to its Three Prime counterpart. In Figure B., only one major point was found to be an indicator for large selection coefficient differences.

Patient ID	s_MPL Envelope	s_MPL 3' Prime	HXB2	Epitope
40	0.056	0.021	7204	N/A
40	0.053	0.0329	6659	N/A
40	-0.024	-0.0088	6408	N/A
40	-0.019	-0.0027	6491	N/A

58	0.0650658	0.00347789	8463	N/A
58	0.0401	0.019	8733	N/A
58	-0.016	-0.004	6507	N/A
58	0.022	0.005	6918	N/A
77	0.1516	0.1204	7285	**QF - RNKTIVF**
77	0.033	0.0082	8722	**DRVIEELQR**
77	0.012	0.0317	7978	N/A
470	0.0266	0.06	8462	N/A
470	0.0447	0.028	7035	N/A
470	0.029	0.015	7971	N/A
470	-0.01	0.006	8262	N/A
607	N/A	N/A	N/A	N/A

Table 3. *Table for Patient 040,058,077, and 470.* From the scatterplots (Figure V and VI), the data points with large differences in selection coefficients were noted. These points were then looked up in the raw data to find for their HXB2 tag number and if they were in any T-cell epitopes. Patient 40,58, and 470 were not found to be in any T-cell epitopes but Patient 77 had mutations that were found in two epitopes, QF-RNKTIVF and BDRVIEELQR*)*

Patient ID	s_MPL Envelope	s_MPL 3' Prime	HXB2	Epitope
131	0.095	0.0667	7227	N/A
131	-0.0114	0.022	8075	N/A
131	-0.002	-0.018	6278	N/A
131	0.074	0.093	7458	N/A
159	0.092	0.05	7136	N/A
159	0.048	0.0157	6494	N/A
159	-0.00758	0.025	7619	N/A
159	-0.03	-0.01	6478	N/A
159	0.01	-0.01	6539	N/A
256	0.059	0.04	8443	N/A
256	0.047	0.03	8446	N/A
256	0.024	0.035	6356	N/A

256	0.02	0.01	8456	**RMRSIRLVN**
42	-0.036	-0.0137	7112	N/A
42	-0.003	0.02	6710	N/A
42	0.00086	0.02	6356	N/A
42	-0.0299	-0.008	6422	N/A

Table 4. *Table for Patients 131, 159, 256, and 42.* Same measurement used in Table 3. It was shown that Patient 131, 159, and 42 were not found to be in any T-cell epitopes. However, Patient 256 was found to have a nucleotide in T-cell epitope RMRSIRLVN.

Patient ID	s_MPL Envelope	s_MPL 3' Prime	HXB2	Epitope
162	0.052	0.085	6924	N/A
162	0.028	0.0016	8699	N/A
162	0.032	0.012	6716	N/A
185	-0.007	0.0186	7071	N/A
185	0.0146	-0.00769	7072	N/A
185	-0.0299	-0.0128	8253	N/A
198	0.0348	0.00375	7048	N/A
198	0.068	0.048	7600	N/A
198	0.011	0.02	6513	N/A
164	0.0658	0.0406	6238	N/A
164	0.00548	0.0346	7686	N/A
164	0.056	0.032	6237	N/A
164	0.034	0.014	7239	N/A
164	0.000858	0.0194	7423	N/A

Table 5. *Table for Patients 162, 185, 198, and 164.* Same measurement used in Table 3 and 4. It was shown that all the patients in this list had no nucleotides in any T-cell epitopes.

Patient ID	s_MPL Envelope Mean	s_MPL ThreePrime Mean
40	0.001695242	0.001400572
58	0.004740439	0.005625471
77	0.004285821	0.004151239
470	0.000116129	-0.000252279
607	0.000730688	0.000586915
131	0.000847309	0.000687304
159	0.002353156	0.002310948
256	0.00075123	0.000635126
42	-0.000481474	-0.00036323
162	0.00113357	0.000468066

185	0.000812725	0.000826896
198	0.001865285	0.0013399
164	0.000504414	0.000574074

Table 6. *Average Selection Coefficients of Patients.* Data showed that between the Envelope and Three-Prime, the mean selection coefficients were very similar to each other. This shows that measurements either taken from the Envelope or Three-Prime would not have a drastic difference.

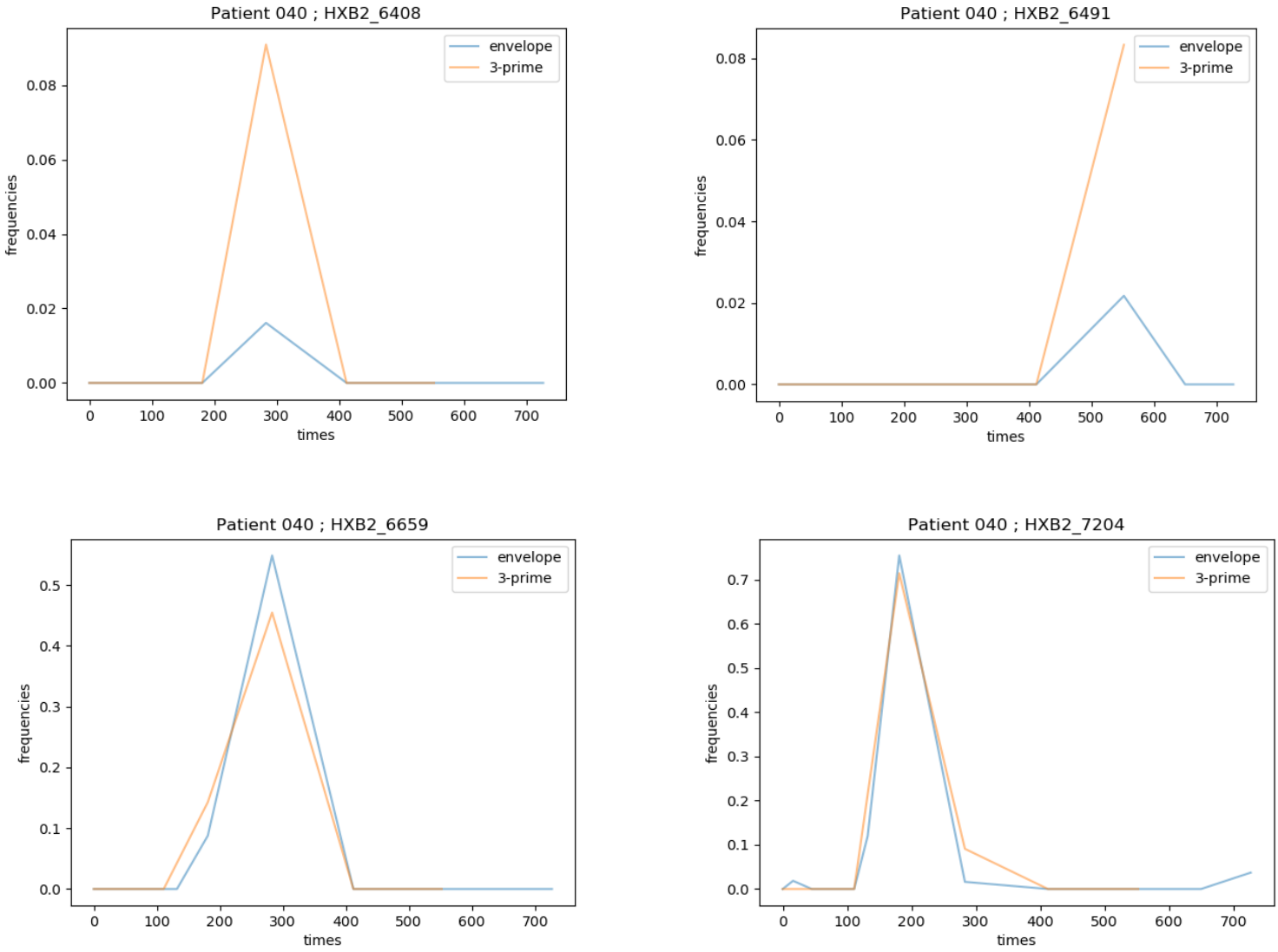

Figure 3. *Frequency Trajectories for Patient 040.* Frequency trajectories were measured to view whether time measurement influenced the selection coefficient differences. Patient 040 had uneven time measurements for two data sets as the envelope data had more time measurements compared to the Three-Prime data set. In HXB2 6408 and 6491, the Three Prime data points were found to have x8 increase in

frequency at time 300 and 500 respectively compared to the envelope frequency. However, overall the shape of the envelope and Three-Prime frequencies were very similar, meaning that they both had the same trend. Results showed that even though the time measurements were different between the two data sets, the overall trend of frequency was kept, meaning that time difference was not a cause for selection coefficient differences.

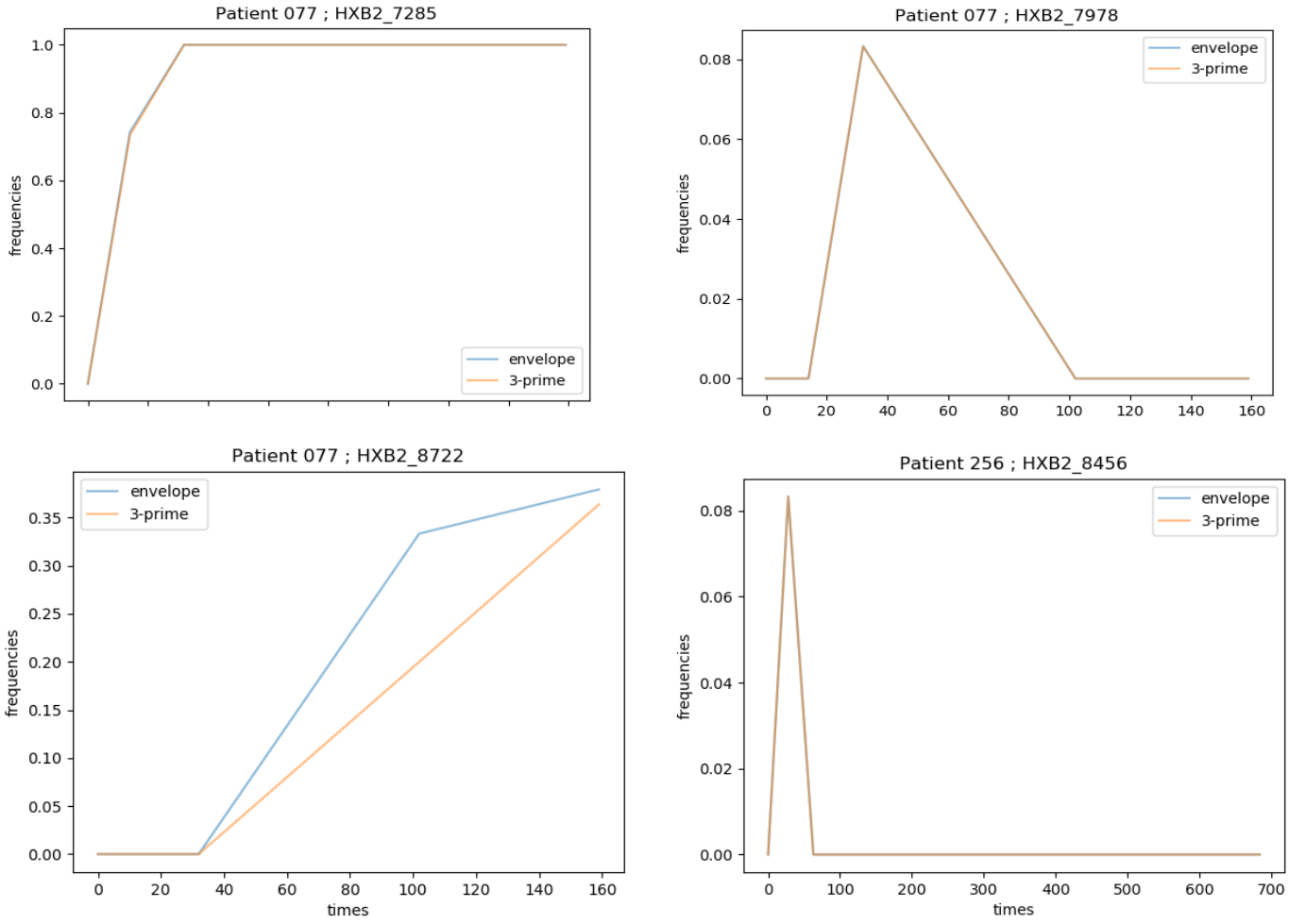

Figure 4. *Frequency Trajectories for Patient 77 and 256 HXB2 tags found to be in T-cell epitopes.* Frequency trajectories were measured to view whether time measurement influenced the selection coefficient differences. For Patient 77 and 256, envelope and Three-Prime time measurements were measured at the same time points. Patient 77 was found to have HXB2 tag 7285 and 8722 to be in T-cell epitopes. HXB2-tag 7285 was found to have identical frequency trajectories for the Three-Prime

and Envelope. HXB2-tag 8722 had a deviation at time 100 but had an overall similar shape. Patient 77 HXB2-7978 was found to be in T-cell epitopes and was found to also have identical frequency trajectories for the two data sets. Patient 256 HXB2-tag 8456 was found to be in a T-cell epitope as well. However, it was shown to have identical frequency trajectories for the two data sets and no deviations in points as well. Patient 77 and 256's HXB2-tag 7285 and 8456 respectively showed that being in T-cell epitopes showed no change in frequency trajectories but Patient 77 HXB2-tag 8722 deviation at time 100 is unknown. Further studies must be conducted to understand if the deviation at the time point is due to being in a T-cell epitope or other reasons.

Discussion

The 3-prime-half-genome and envelope-specific region data represents two different views of the same evolutionary process. The envelope-specific region represents nucleotides found in the envelope region while the 3-prime-half genome data contained both the envelope nucleotides and other viral nucleotides as well. When comparing the selection coefficients between the patient's 3'-half-genome data and the envelope-specific region, some selection coefficients' large differences. Patient 040 in Table 1 can be seen having one nucleotide inferred to be deleterious in the envelope while the same nucleotide in the three-prime data was inferred to be beneficial. Patient 077 (Table 2) showed no changes in mutation dynamics in the early parts of the data.

In order to compare the two data sets' selection coefficients, a histogram was created to visualize the data. Figure 1 displayed patients 040 and 77 histograms and showed that the measured selection coefficients overlapped each other, showcasing how similar the data was. In both patients the overlap in data can be seen in around the -0.01 to .02 range. Some outliers can be seen in the .04 range. The overlapping in selection coefficients noted in the two patients were commonly seen in the other patient data sets as well. The average selection coefficients of the

two data sets for the patients were also taken (Table 6) and shown to be nearly identical to one another with an average difference of .0004. Patient 58 had the largest mean difference of .000885 while patient 159 had the lowest mean difference of .00042.

For Figure 2, the 3'-half-genome selection coefficients on the Y-axis and the envelope region selection coefficients were plotted on the X-axis. This created a scatterplot which was used for a better visual analysis for the selection coefficient differences. The line of best fit represented the how similar the two data sets were to each other. The points deviating far from the line of best fit were recorded as the points were nucleotides that had the largest selection coefficient difference between the two. At least four points for each patient were recorded and were search in the data sets if they were in T cell epitopes. These epitopes are the T cells binding to small regions of the viral genome and replaced with a mutated version that would allow for escape from immune response. Tables 3,4, and 5 showed the recordings of the Patients with the T-cell epitopes. While majority of the patients were not found to be in any T-cell epitopes, patient 77 had nucleotides in epitope *QF-RNKTIVF* and *DRVIEELQR* and patient 256 had a nucleotide in epitope *RMRSIRLVN*.

No further study was done to see if there could be linkage effects with an epitope but a possible reason for large changes in selection coefficients could be due to genetic hitchhiking. This is the phenomena of when a neutral gene will experience a change due to being closely linked to a selected gene. It could be plausible as the large changes in selection coefficients could have been caused by the variants being close or near a region that contain strong mutations even though they possessed no effect on viral fitness.

In some data sets, non-identical longitudinal time measurements could have been taken, creating unreliable measurements in data. Because of this, we checked to see whether different

measured time points influenced the selection coefficient difference. A graph was created with the x-axis as the time measurements and the y-axis as the frequency to get a trajectory of the frequency movement along the data's time scale (Figures 3 and 4). Patient 077, 470, and 256 half genome and envelope region data were measured at the same time points and showed identical frequencies. However, Patient 040 had different time points measured and the frequency at 300 days showed an 8x difference but an identical shape. This meant that the unequal time measurements of the data did not affect the trendline for the virus in patient 77, potentially stating that they reached the same end result. Figure 6 showed the data points of patient 77 and 256 that were in the T-cell epitopes. The data showed that there were equal time points in both data sets and had identical frequencies and trajectory trends. Patient 77 HXB2_8722 had a different time point at 300 days for the envelope and three prime frequency, but also had similar shapes, correlating to a similar frequency trajectory as well. Since the three prime and envelope viral sequencing data came from the same individuals, it is important to note that any differences found in the comparison was entirely due to sampling and not any underlying genetic cause.

Results from both studies showed that restricting the range of sequence data that can be looked at will not have a dangerous effect on the results. This shows that the depth of sequencing data is broadly similar to each other and either cases of looking at a narrow and deep or wide and shallow data would potentially produce similar results. Consistency was shown between the global and regional scales of the data. This is important because it shows that data can still be used for analysis even if the data was not the complete genome. While a check was done to see if there were variants in epitopes in the envelope region between measurements, I was unable to investigate whether these variants with large differences were strongly linked with escape

mutations or not to prove if there was linkage disequilibrium or simply genetic hitchhiking that was occurring. Future studies can be done by linking selected mutations from inside the envelope with epitopes outside of the envelope to view whether the escape mutation was outside of the sequencing region and the neutral variants that had large selection coefficient differences but not in T- cell epitopes could be linked to the outside escape mutations. Through a trial and error method of manipulating the selection coefficients data to look at different reasons for the large selection coefficient differences, a theoretical evolutionary path of the mutations can be created to provide further understanding of the fitness landscape of HIV.

References

Liu, M. K., Hawkins, N., Ritchie, A. J., Ganusov, V. V., Whale, V., Brackenridge, S., Li, H., Pavlicek, J. W., Cai, F., Rose-Abrahams, M., Treurnicht, F., Hraber, P., Riou, C., Gray, C., Ferrari, G., Tanner, R., Ping, L. H., Anderson, J. A., Swanstrom, R., CHAVI Core B, … Goonetilleke, N. (2013). *Vertical T cell immunodominance and epitope entropy determine HIV-1 escape.* The Journal of clinical investigation, 123(1), 380–393. https://doi.org/10.1172/JCI65330

Sohail, M. S., Louie, R. H. Y., McKay, M. R., & Barton, J. P. (2019, January 1). *Resolving genetic linkage reveals patterns of selection in HIV-1 evolution.* Retrieved from https://www.biorxiv.org/content/10.1101/711861v1

Walker B, McMichael A. *The T-cell response to HIV.* Cold Spring Harb Perspect Med. 2012;2(11):a007054. Published 2012 Nov 1. doi:10.1101/cshperspect.a007054