Sequence Analysis of Human Immunodeficiency Virus Type 1

Stephanie Lucas[1,2], Panayiotis V. Benos[1,3], and David L. Corcoran[4]

1 Bioengineering and Bioinformatics Summer Institute, Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA 15213
2 Department of Biology, University of San Francisco, San Francisco, CA 94117
3 Department of Computational Biology, University of Pittsburgh, Pittsburgh, PA 15213
4 Departments of Computational Biology and Human Genetics, University of Pittsburgh, Pittsburgh, PA 15213

Abstract

Human Immunodeficiency Virus Type 1 is known for its ability to evolve very quickly, which leads to increased studies in its evolution. Using its genomic sequence, a comparison in observed and theoretical synonymous sites was used to determine whether or not each gene is undergoing differential selective pressure (Kreitman and Hudson, 1991). HIV-1 subtype B was also compared to the other subtypes in order to test if the differences between sunbtypes are neutral (McDonald and Kreitman (1991). HIV-1's regulatory region (U3/R) was also analyzed to test the levels of polymorphisms present within this region. The regulatory regions of HIV-1, SIV-1, and SIV-2 were then used to predict potential transcription fact binding sites within the promoter region. These tests provide us with information about any evolutionary constraints within the HIV-1 virus potentially helping us understand the structure of the regulatory regions as well as determining which areas of the virus are most essential.

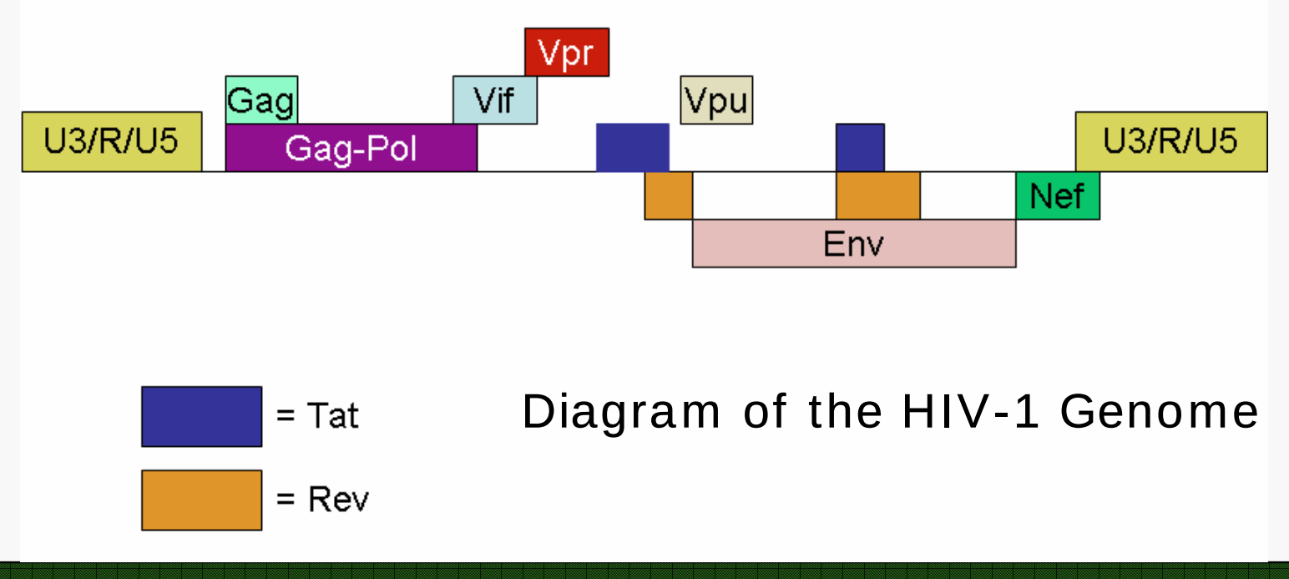


Diagram of the HIV-1 Genome

Introduction

Currently there are 38.6 M people in the world who live with HIV/ AIDS1. Despite all of the research done with HIV, there is still no cure. Similarly, vaccines are very difficult to produce since the virus is known for its high mutation rates.

Our research focuses on understanding what evolutionary forces are acting on the virus. We plan to 1) track if there are any differential selective pressures on parts of the genome, 2) identify regions of higher/lower variability, and 3) predict and confirm TFBS within the promoter region. Areas that evolve slower can potentially act as vaccine targets. Prediction and confirmation of TFBS can also give us more information about the virus' regulatory regions.

In order to test if there are any selective pressures on the virus, we will employ two types of tests- both of which take into account ϴ (= levels of polymorphisms). The infinite sites neutral model compares different areas within 1 species, and focuses on ϴ as silent polymorphisms. We hypothesize that ϴ will be equal across all genes since they are synonymous mutations. We would also expect that the U3/R region would have lower or at least similar ϴ values (since the model assumes all regulatory sites are silent. Deviation from ϴ assumes that the particular gene is under either selective constraint or lack there of.

The other test compares the levels of polymorphisms between species (or in our case between subtypes of the HIV-1 virus). The idea is that the ratio of nonsynonymous to synonymous mutations within a species (polymorphisms) should be the same as the ratio between species if the mutations are neutral. This can be done by using a derivative of the chi-square test. If they are not similar, we suspect that there are some differences in the evolution of the gene between subtypes.

Lastly, we were able to use the regulatory sequences of HIV-1, HIV-2, and SIV-1 to predict TFBS by using a simple program, MATCH. By doing so, we could confirm published TFBS of HIV-1 as well as compare the regulatory regions between species. We expect that those TFBS conserved between species are important in the regulation of the virus.

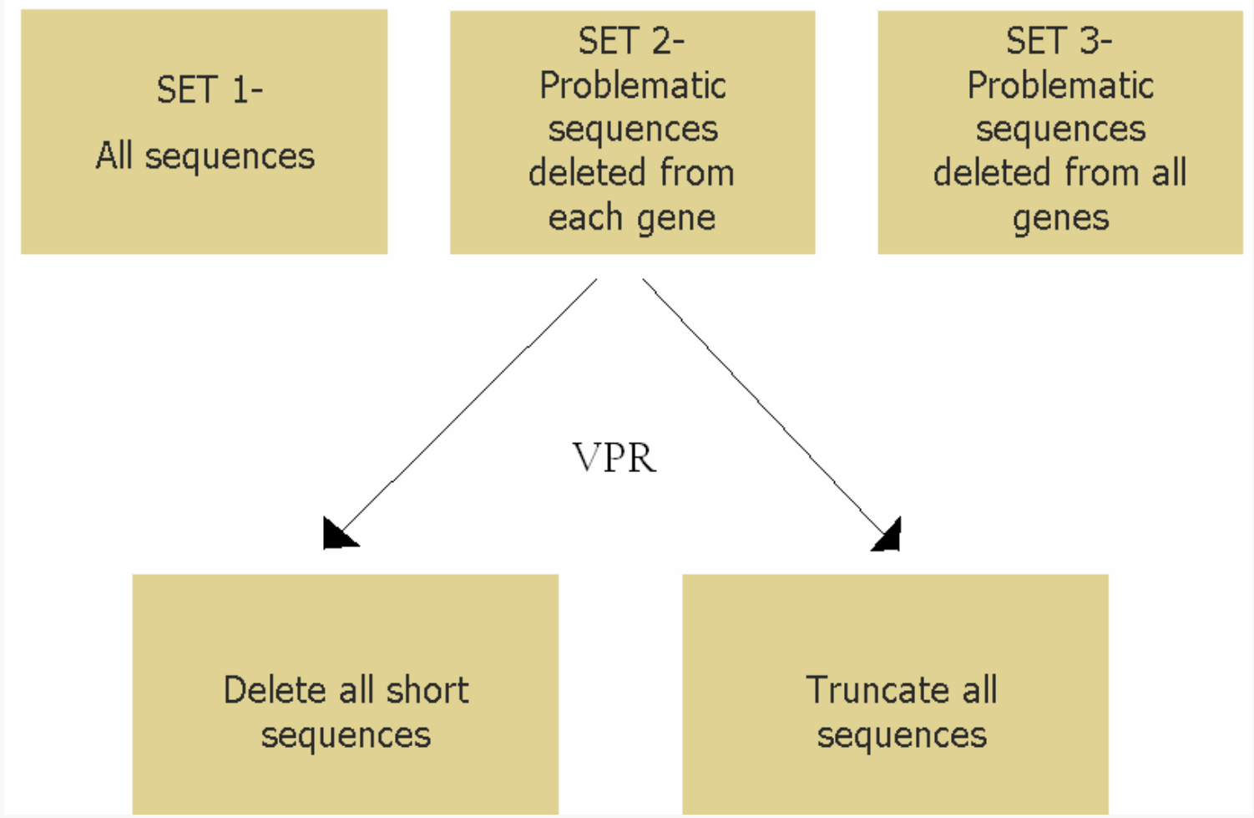
[1] UNAIDS (http://www.unaids.org/en/HIV_data/2006GlobalReport/default.asp)presented in: http://hivinsite.ucsf.edu/global?page=cr-01-00&post=2&cid=US#General%20HIV/AIDS

Methods

1) Obtaining the sequences

a. Looked up the Ref Seq from the database
b. Searching public databases yielded 1,183 sequences
c. Split the Ref Seq into individual genes and regulatory regions
  -coding/ regulatory regions only
d. Remove overlapping sequences and Start/Stop codons
  -There are differential constrains within individual bases
  -As a consequence, 2 genes were not analyzed- TAT and GAG
  -Start/ Stop codons are relatively invariable and may stray the conserved sequence count
e. Did a BLAST search against the 1,183 genomes to extract out each gene from the sequences and remove identical sequences- left with about 200 sequences

f. Align with Clustal W using the MEGA software package


SET 1- All sequences
SET 2- Problematic sequences deleted from each gene
SET 3- Problematic sequences deleted from all genes
VPR
Delete all short sequences
Truncate all sequences

Due to problematic sequences that could potentially influence the value of ϴ, we separated them out into 3 different sets.

2) Infinite Sites Neutral Model

-Kreitman and Hudson (1991)

$$p_s = S/n$$

$$Var(p_s) = [E(p_s)/n] + a_2 \cdot \theta$$

Where....
S= # silent segregating sites
n= # possible silent sites

$$a_2 = \sum_{x=1}^{m-1} x^{-2}$$

We can estimate ϴ...

$$\hat{\theta} = p_s / a_1$$

$$Var(\hat{\theta}) = \mathrm{var}(p_s)/a_1{}^2$$

Actual ϴ:

$$E(p_s) = a_1 \cdot \theta$$

Where...

$$\theta = 4N\mu$$

$$a_1 = \sum_{x=1}^{m-1} x^{-1}$$

-ϴ is difficult to obtain and because N (population size) and μ (rate of mutation per silent site) are difficult to obtain

3) Evolution rates compared between species

-McDonald and Kreitman (1991)

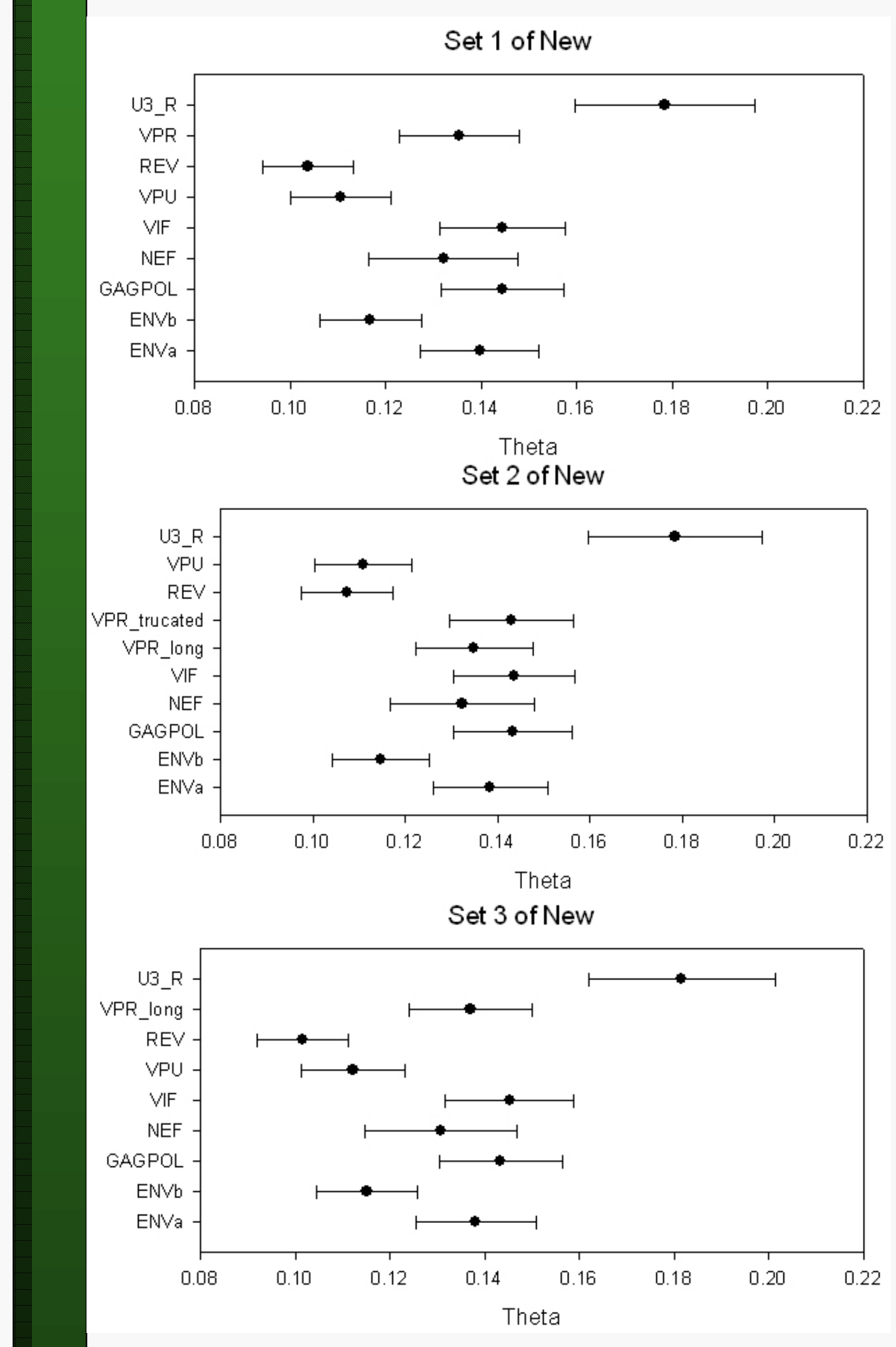
The idea is that the ratio of nonsynonymous to synonymous mutations within a species (polymorphisms) should be the same as the same ratio between species if the mutations are neutral

4) Predict TFBS within the promoter regions of HIV-1, HIV-2, and SIV-1

We used MATCH (part of the TRANSFAC database) to predict TFBS within the LTR/ U3_R regions of each genome.

Results

Data using script similar to Kreitman paper:



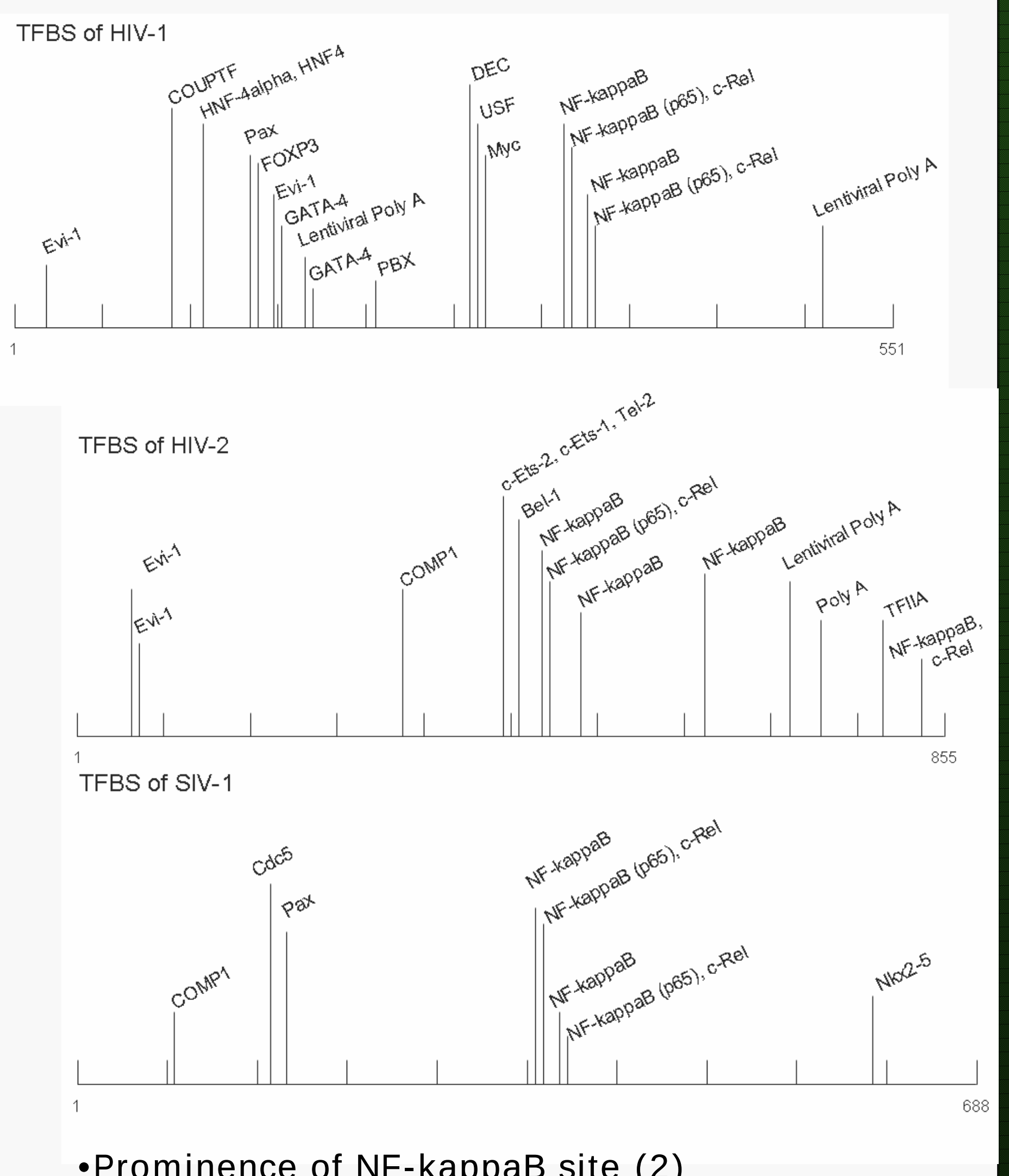
•U3_R region is less constrained than the other regions (due to higher ϴ values)
•VPU, REV, ENVb are more constrained than other genes (due to lower ϴ values)

G-Test Values

	G-value	
ENV	4.523008465	-4.48702125
	-4.47435055	4.513618736
	G-test value	0.150510788
	p-value	0.698047708
GAGPOL	-42.7794747	49.20393745
	49.72230631	-45.9873991
	G-test value	20.3187399
	p-value	**6.5555E-06**
NEF	-4.45066615	4.737231073
	4.902885968	-4.53059894
	G-test value	1.317703895
	p-value	0.251004736
REV	4.619723277	-4.27516555
	-4.1594237	4.576460189
	G-test value	1.523188436
	p-value	0.21713774
VIF	5.914699402	-5.58826033
	-5.49156862	5.819345174
	G-test value	1.308431244
	p-value	0.252679029
VPR	-6.4331233	7.770285908
	7.992447149	-6.94572213
	G-test value	4.767775269
	p-value	**0.02899728**
VPU	1.133974299	-1.10602586
	-1.09240413	1.134875083
	G-test value	0.140838782
	p-value	0.707448569

•GAGPOL and VPR are sequences are significantly different.

Graphs of predicted TFBS using MATCH:



•Prominence of NF-kappaB site (2)
•We would've liked to see the difference in ϴ values across the promoter region. This would confirm and better prove TFBS (the lower ϴ values= the more conserved the sequence). We would use a sliding window of about 100 bp overlapping by 50 bp.

Conclusion

Certain areas of the HIV-1 genome are found to have differential selective pressure, suggested by the difference in ϴ.

TFBS (such as NF-kappaB) have been predicted, with relative confidence by comparison to published data.

This study, though, has more work to be done. We plan to...

1. More accurately compare the subtypes of HIV-1
2. Further develop the new test used to calculate ϴ
3. Calculate ϴ values for different regions of the promoter region to better prove TFBS

Acknowledgements

The national BBSI program (http://bbsi.eeicom.com) is a joint initiative of the NIH-NIBIB and NSF-EEC, and the BBSI @ Pitt is supported by the National Science Foundation under Grant EEC-0234002.

1) Panayiotis V. Benos and lab members: David Corcoran and Shaun Mahony

2) Department of Computational Biology, University of Pittsburgh: Judy Wieber and Rajan Munshi

References

http://uhavax.hartford.edu/bugl/hiv.htm#types

Jeeninga, Rienk E., Hoogenkamp, Maarten, Armand-Ugon, Mercedes, de Baar, Michel, Verhoef, Koen, Berkhout, Ben. "Functional Differences between the Long Terminal Repeat Transcriptional Promoters of Human Immunodeficiency Virus Type 1 Subtypes A through G". Journal of Virology 74:8 (2000): 3740-3751.

Van Opijnen, Tim, Kamoschinski, Joost, Jeeninga, Rienk E., Berkhout, Ben. "The Human Immunodeficiency Virus Type 1 Promoter Contains a CATA Box instead of a TATA Box for Optimal Transcription and Replication." Journal of Virology 78:13 (2004): 6883-6890.

Kreitman M. and Hudson R. "Inferring the Evolutionary Histories of the *Adh* and *Adh-dup* Loci in *Drosophila melanogaster* From Patterns of Polymorphism and Divergence." Genetics 127 (1991): 565-582.

McDonald, John H. and Kreitman Martin. "Adaptive protein evolution at the *Adh* locus in *Drosophila.*" Nature 351 (1991): 652-654.