Sequence Analysis of Human Immunodeficiency Virus Type 1

Stephanie Lucas [1,2]
Mentor: Panayiotis V. Benos [1,3]
With help from: David L. Corcoran [4]

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

HIV-1 Facts

- Some Facts as of 2005…
  - There are 38.6 M people in the world living with HIV/AIDS
  - US has 1.2 M out of the 1.3 M living in N. America.
  - Kenya has 1.3 M, and sub-saharan Africa 24.5 M

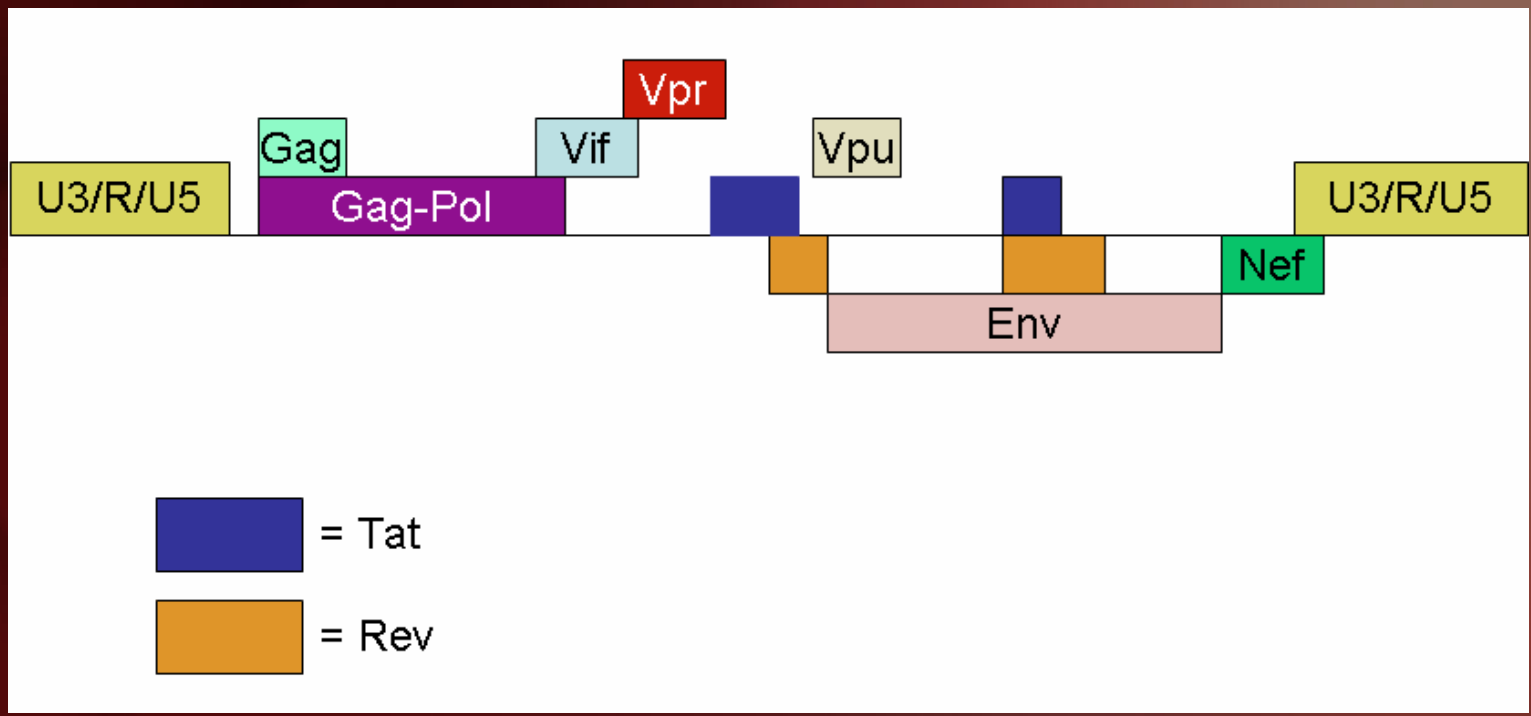
- There is currently no successful therapy.
  Production of vaccine is difficult due to the high mutation rate of the virus.
  The effects are severe both economically (cost of care) and socially .

Purpose

- We will study the evolution of different parts of the HIV-1 genome
- Parts that evolve slower could act as potential vaccine targets

HIV-1 Genome

- Reference Sequence is 9181 base pairs long
- Contains 9 genes

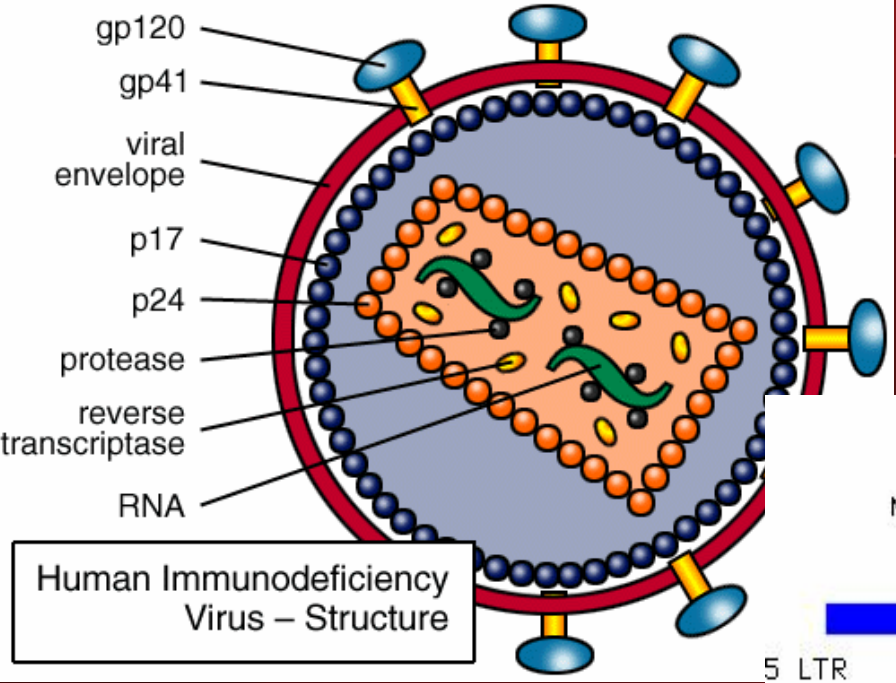


The Genes

- Gag: codes for internal structural proteins and capsid proteins
- Gag-Pol: codes for the three enzymes necessary for replication Vpr
- Vpu: virus protein U
- Tat: transactivator protein
- Rev: regulator of expression of virus protein

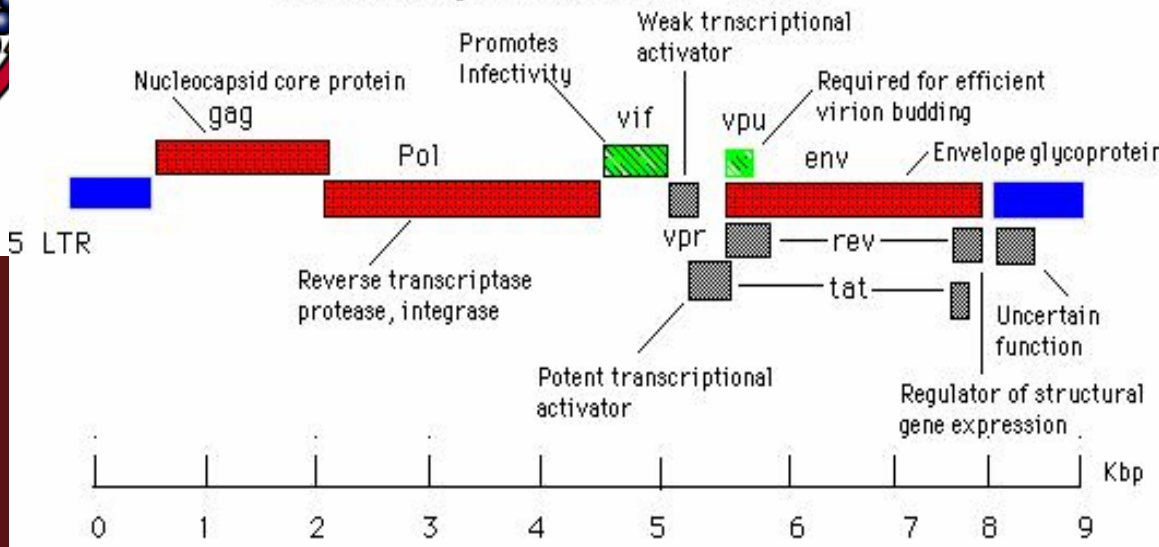
The genes cont'd

- Env: codes for the surface proteins gp120 and gp41 that protrude from the lipid envelope and attach to cellular receptors
- Nef: an enhancing factor
- Vpr: virus protein R
- Vif: virus infectivity factor



Human Immunodeficiency Virus – Structure

Genetic Organiation of HIV-1 virus



Polymerase precursor

Protease
p10

Reverse Transcriptase
p64,p51

Integrase
p32

Gag precursor
p53

Envelope precursor
gp160

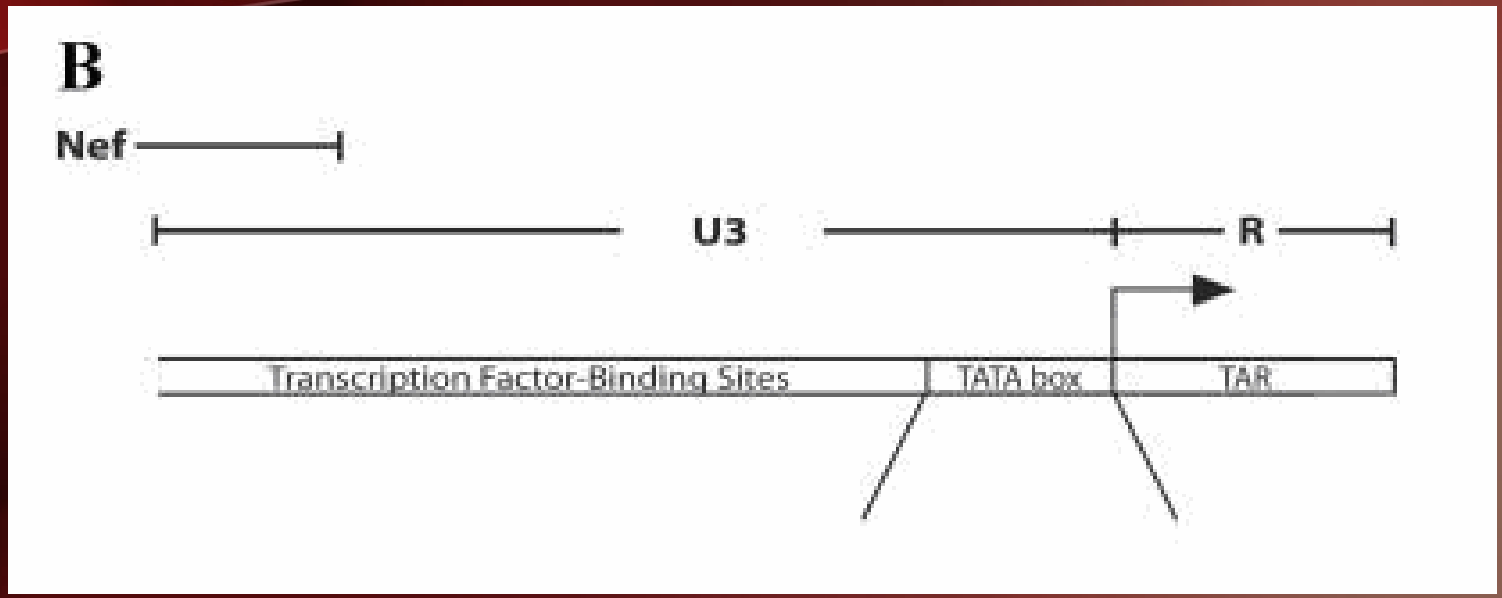
Core
p17 Core
p24

p15
p7
p9

Extracellular protein
gp 120

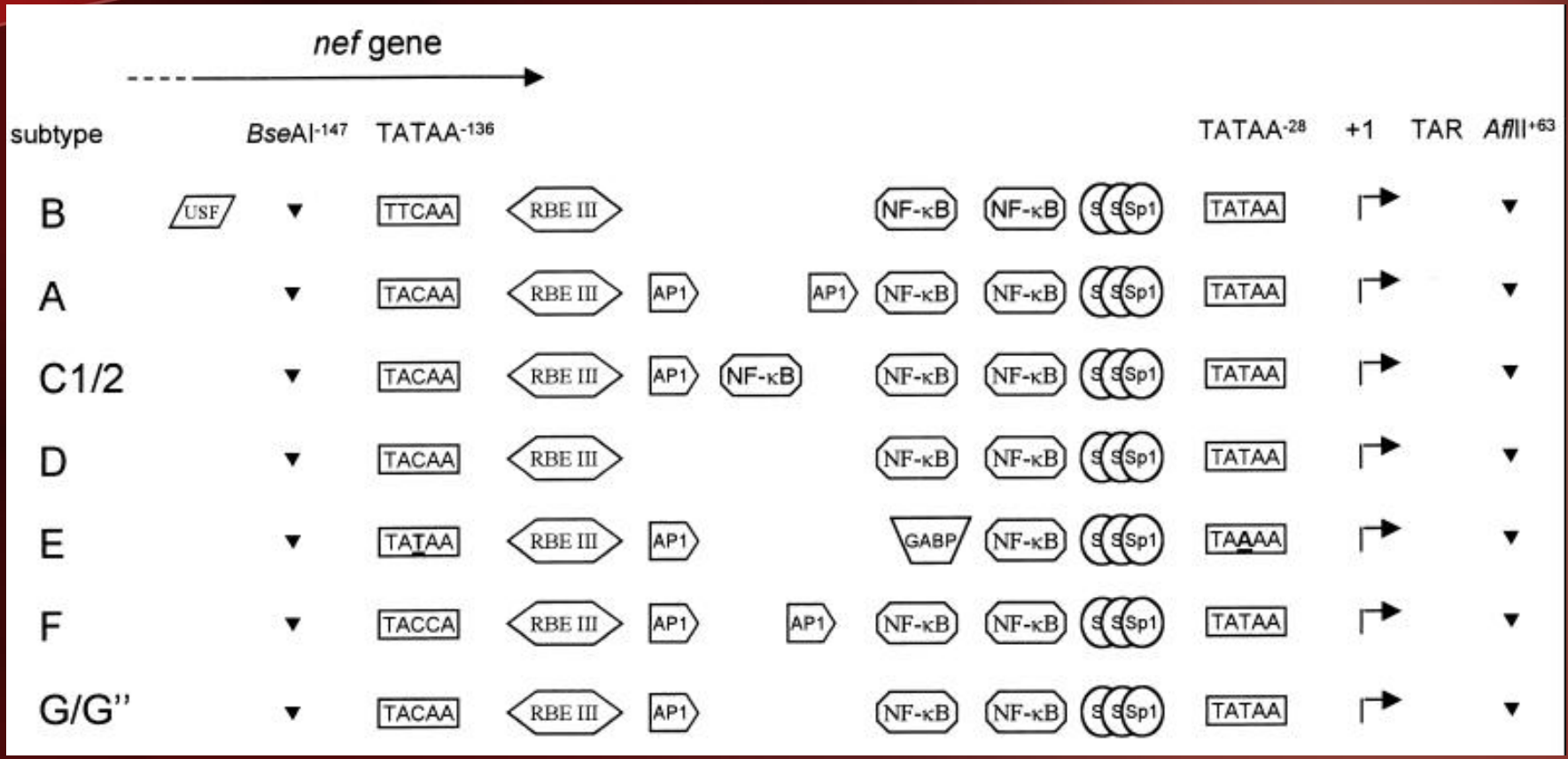
Transmembrane protein
gp41

U3_R_U5 Region



- The U3_R region is the regulatory region and contains most of the transcription factor binding sites

Predicted TFBS



Overview of Research---Our Goals

- To track if there are any differential selective pressure on parts of the genome
- Identify regions of higher/lower variability
- Predict and confirm TFBS within the promoter region

Methods

Obtaining the sequences

a. Looked up the Ref Seq from the database

b. Searching in the public databases yielded 1,183 genomes

c. Split the Ref Seq into individual genes and regulatory regions
- **-coding/ regulatory regions only**

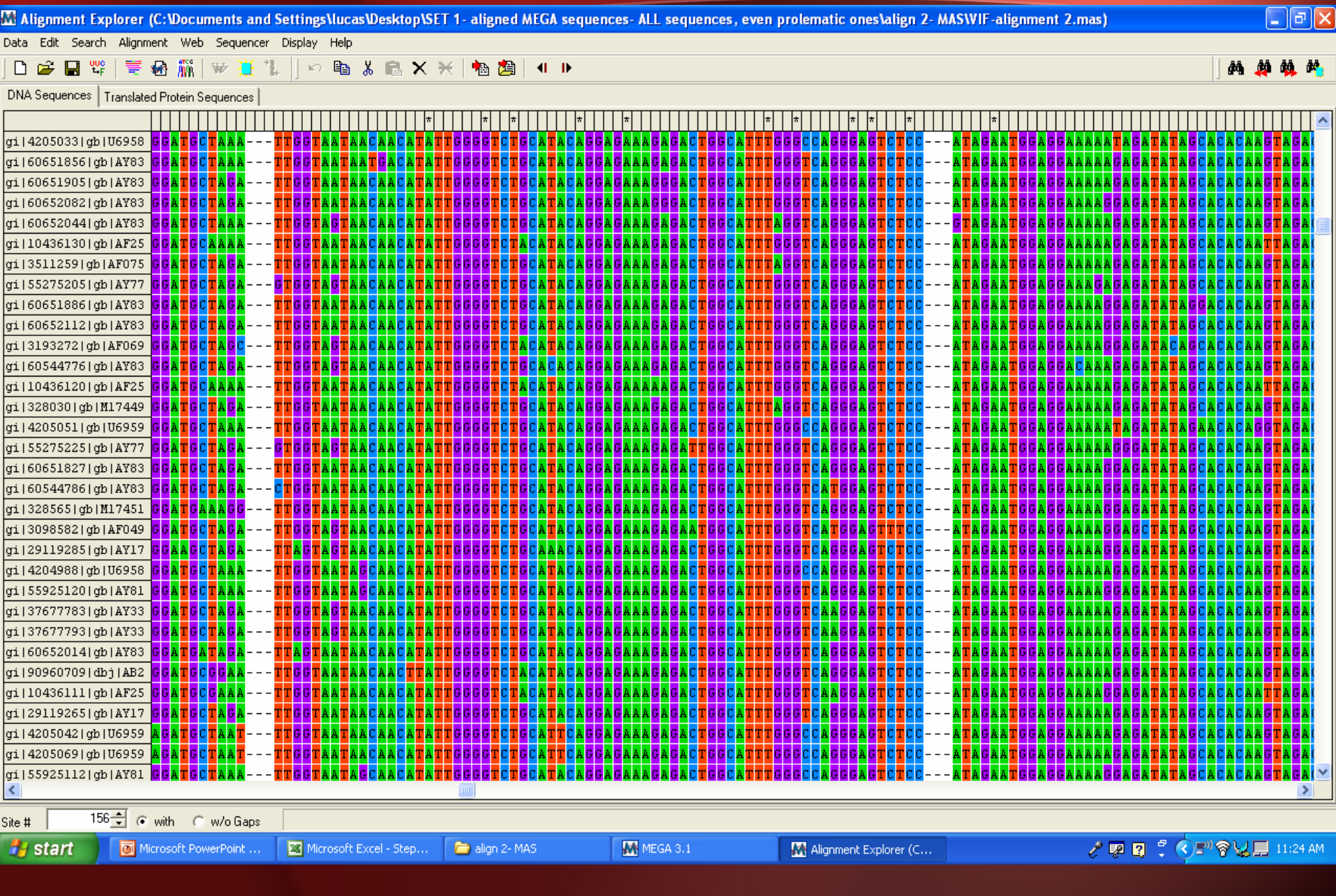
d. Removed overlapping sequences and Start/Stop codons
- **-There are differential constrains within individual bases**
- **-As a consequence, 2 genes were not analyzed- TAT and GAG**
- **-Start/ Stop codons are relatively invariable and may stray the conserved sequence count**

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

The sequences were then ready for analysis...

Example CLUSTAL W



Infinite Sites Neutral Model

- Developed by Kreitman and Hudson (1991)
- Focused on neutral (silent) mutations
  - Removes the functional constraints in order to focus on the genetic drift alone
- Θ= level of polymorphism $\theta = 4N_e\mu$
  - Actual Θ value cannot be found because N and μ are difficult to obtain…

Equations

First we had to calculate the number of segregating sites per silent nucleotide (ps)…

$$p_s = S / n$$

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

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

Where….

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

$$a_k = \sum_{x=1}^{m-1} x^{-k}$$

Calculate the variables…

Example sequence:

	Leu	Gly
Seq #1	CTG	GGC
Seq #2	CTG	GGC
Seq #3	CTA	GGC
Seq #4	CTT	GGC

- The 3rd codon position is a potential silent site

- **Kreitman and Hudson method**
  - **S= 1**
  - **n=2**
  - **Counts whether or not a column of the sequence is a silent site**

Problems…

- Upon further investigation, we discovered that some of the gene sequences were nonfunctional due to mutation or premature stop codons

- We thought that this may influence the results and decided to create 3 sets of data and run each set through the script

Run the test with 3 different sets of data…

SET 1-

All sequences

SET 2- Problematic sequences deleted from each gene

SET 3- Problematic sequences deleted from all genes

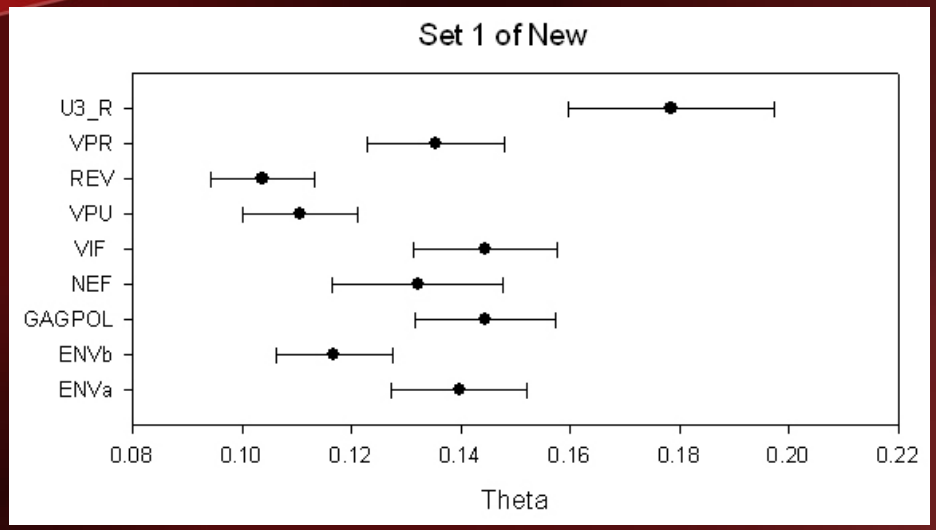
VPR

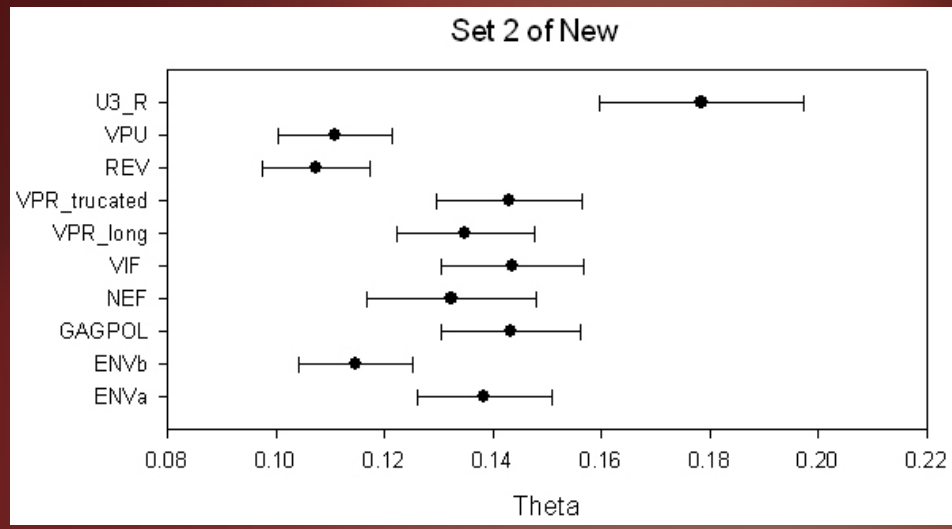
sequence

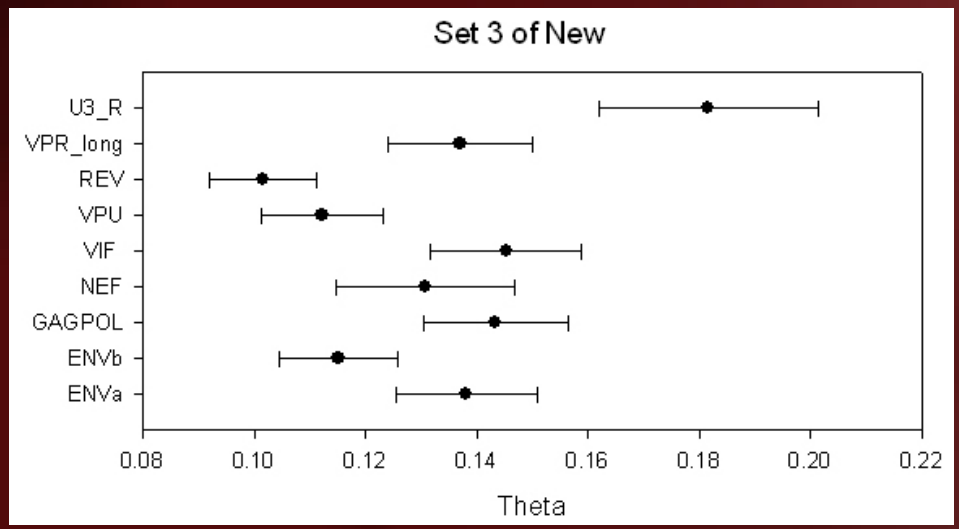
Delete all short sequences

Truncate all sequences

Results


Set 1 of New
U3_R
VPR
REV
VPU
VIF
NEF
GAGPOL
ENVb
ENVa
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
Theta


Set 2 of New
U3_R
VPU
REV
VPR_trucated
VPR_long
VIF
NEF
GAGPOL
ENVb
ENVa
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
Theta


Set 3 of New
U3_R
VPR_long
REV
VPU
VIF
NEF
GAGPOL
ENVb
ENVa
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
Theta

Results Interpreted

- 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)

Comparison with other closely related 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 ratio between species if the mutations are neutral
- Used HIV-2, SIV-1, and SIV-2 in order to test this
- Although they were close relatives, the sequences were too different, and could not be compared

Instead we used the different subtypes of HIV-1 compared to that of Subtype B

- There are 9 subtypes: A, B, C, D, F, G, H, J, K

Methods:

-Obtained complete genomes for each subtypes

-Did similar extraction methods used in HIV-1 sequence

-Do to limited time, analyzed all sequences of each gene of subtype B and a random sequence of a gene per subtype

Results

	Observed			Expected		
ENV						
	Fixed	Polymorph	*residuals*	Fixed	Polymorphi	**G-value**
Nonsynonyn	439	790	*1229*	434.50	794.50	4.523008465 -4.48702125
Synonymous	395	735	*1130*	399.50	730.50	-4.47435055 4.513618736
	834	*1525*	2359			**G-test value** 0.150510788
						p-value 0.698047708
GAGPOL						
	Fixed	Polymorphic				
Nonsynonyn	219	601	*820*	266.24	553.76	-42.7794747 49.20393745
Synonymous	482	857	*1339*	434.76	904.24	49.72230631 -45.9873991
	701	*1458*	2159			**G-test value** 20.3187399
						p-value **6.5555E-06**
NEF						
	Fixed	Polymorphic				
Nonsynonyn	51	128	*179*	55.65	123.35	-4.45066615 4.737231073
Synonymous	46	87	*133*	41.35	91.65	4.902885968 -4.53059894
	97	*215*	312			**G-test value** 1.317703895
						p-value 0.251004736
REV						
	Fixed	Polymorphic				
Nonsynonyn	51	66	*117*	46.58	70.42	4.619723277 -4.27516555
Synonymous	35	64	*99*	39.42	59.58	-4.1594237 4.576460189
	86	*130*	216			**G-test value** 1.523188436
						p-value 0.21713774
VIF						
	Fixed	Polymorphic				
Nonsynonyn	82	134	*216*	76.29	139.71	5.914699402 -5.58826033
Synonymous	72	148	*220*	77.71	142.29	-5.49156862 5.819345174
	154	*282*	436			**G-test value** 1.308431244
						p-value 0.252679029
VPR						
	Fixed	Polymorphic				
Nonsynonyn	25	67	*92*	32.34	59.66	-6.4331233 7.770285908
Synonymous	46	64	*110*	38.66	71.34	7.992447149 -6.94572213
	71	*131*	202			**G-test value** 4.767775269
						p-value **0.02899728**
VPU						
	Fixed	Polymorphic				1.133974299 -1.10602586
Nonsynonyn	39	53	*92*	37.88	54.12	-1.09240413 1.134875083
Synonymous	24	37	*61*	25.12	35.88	**G-test value** 0.140838782
	63	*90*	153			**p-value** 0.707448569

Promoter Regions

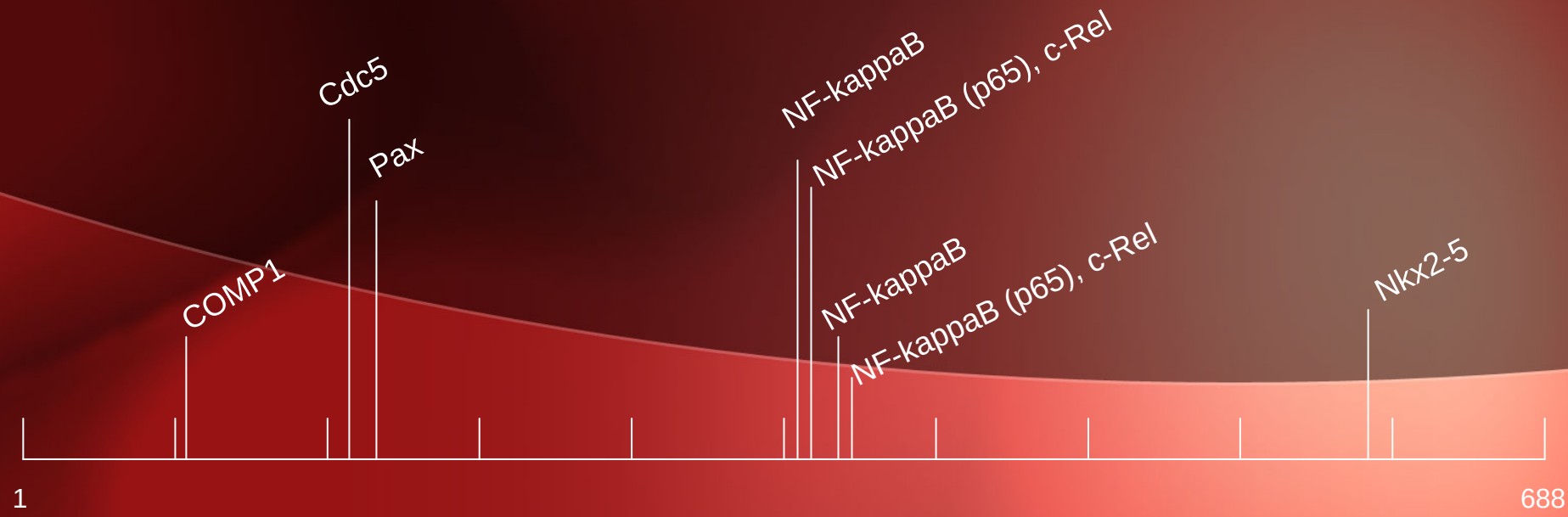
- Although we weren't able to analyze HIV-1 with HIV-2, SIV-1, or SIV-2, we compared the Ref Seq promoter region of each species
  - SIV-2 did not have a promoter region defined
- We used MATCH (part of the TRANSFAC database) to predict TFBS within the LTR/ U3_R regions of each genome.

MATCH

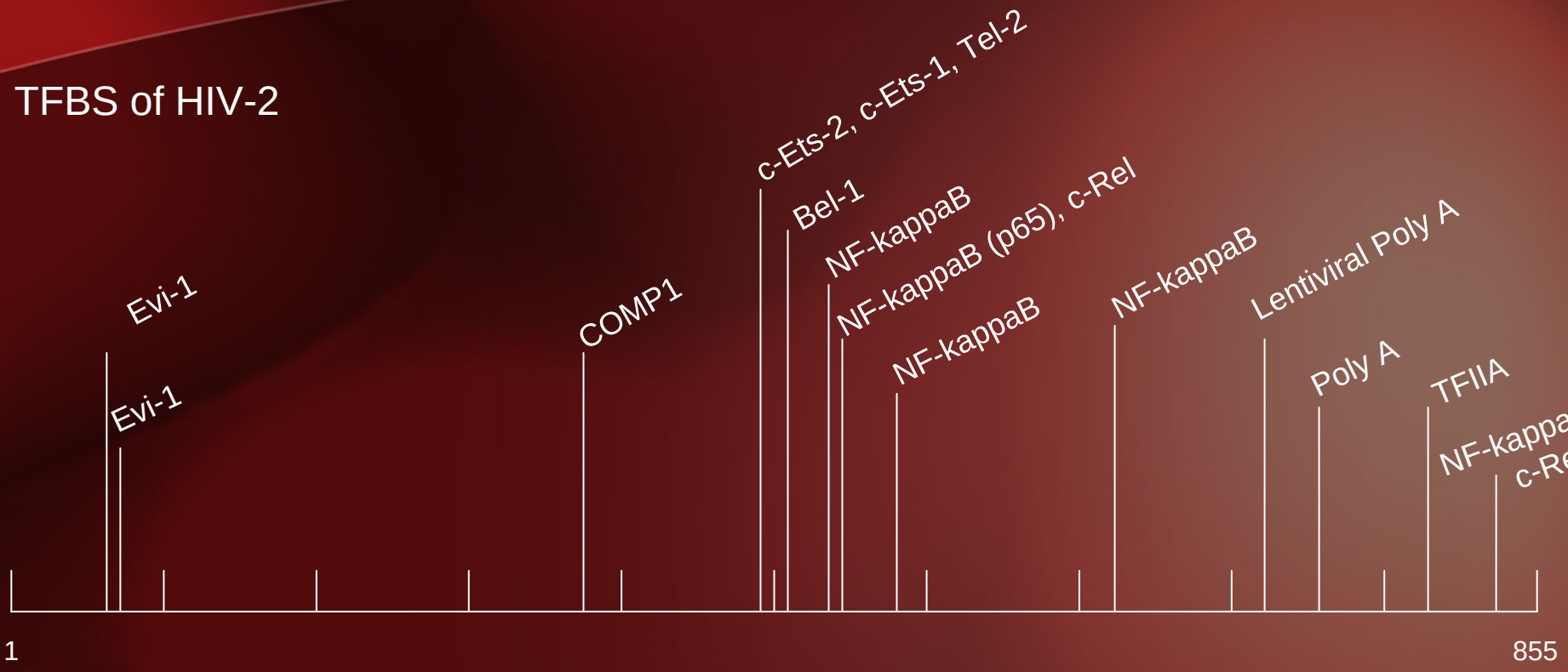
- It uses a scoring matrix generated from known TFBS to predict which TFBS are present within the promoter of the gene
- We predict that there will be many false positives

Results…..

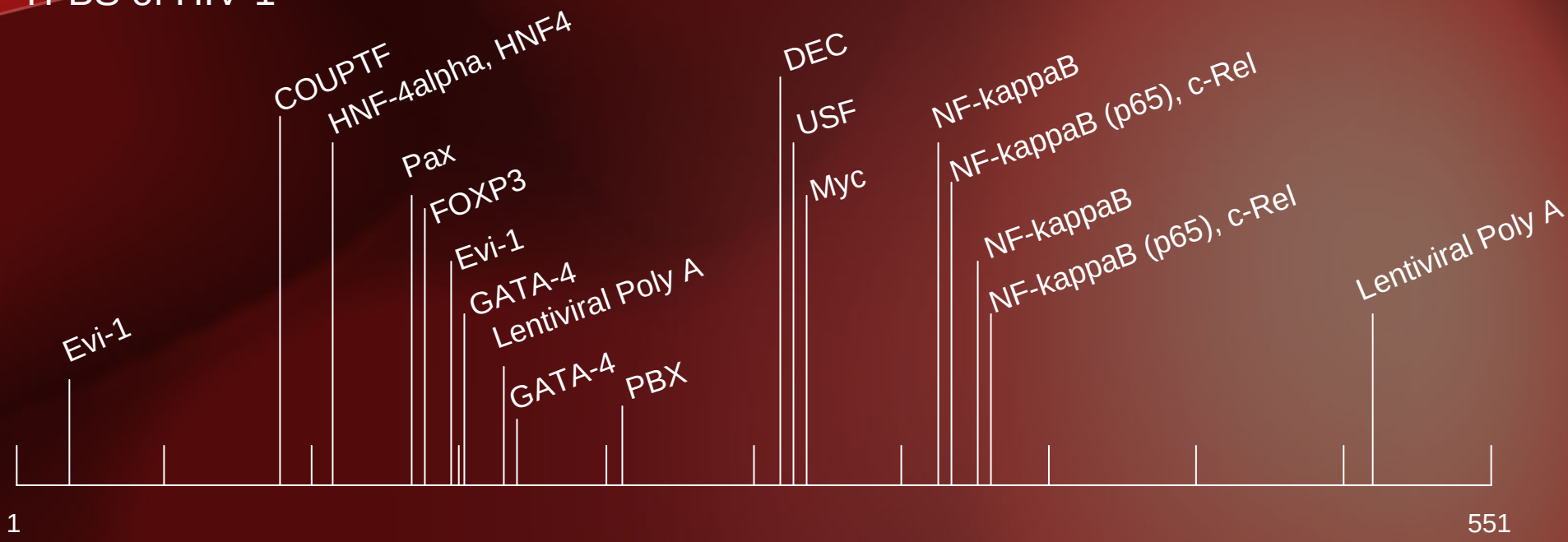
TFBS of SIV-1


Cdc5
Pax
NF-kappaB
NF-kappaB (p65), c-Rel
COMP1
NF-kappaB
NF-kappaB (p65), c-Rel
Nkx2-5
1
688

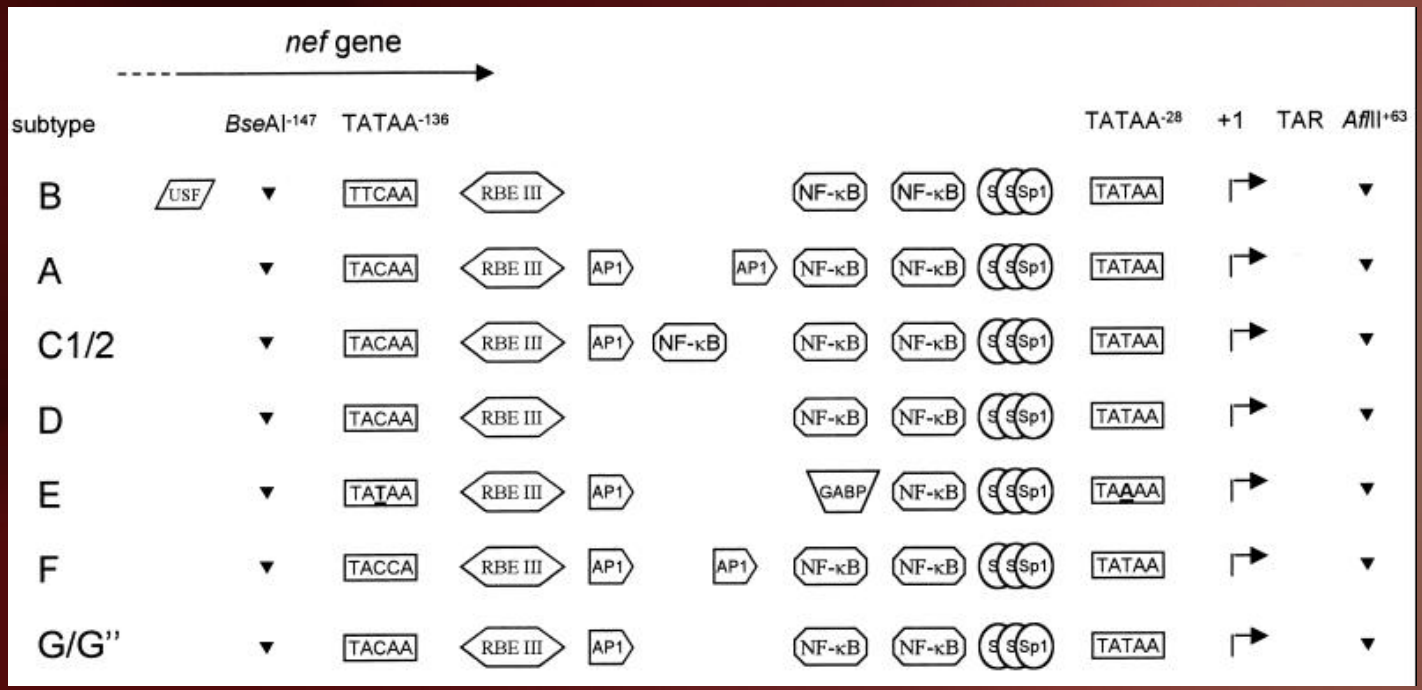
TFBS of HIV-2


Evi-1
Evi-1
COMP1
c-Ets-2, c-Ets-1, Tel-2
Bel-1
NF-kappaB
NF-kappaB (p65), c-Rel
NF-kappaB
NF-kappaB
Lentiviral Poly A
Poly A
TFIIA
NF-kappaB, c-Rel
1
855

TFBS of HIV-1


COUPTF
HNF-4alpha, HNF4
Pax
FOXP3
Evi-1
GATA-4
Lentiviral Poly A
GATA-4
PBX
DEC
USF
Myc
NF-kappaB
NF-kappaB (p65), c-Rel
NF-kappaB
NF-kappaB (p65), c-Rel
Lentiviral Poly A
Evi-1
1
551

Comparison to the known TFBS



- 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).
  - Use a sliding window of about 100 bp overlapping by 50 bp

Future Work

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

Acknowledgements

- Takis Benos
- David Corcoran
- Shaun Mahony
- Judy Wieber and Rajan Munshi
- Dept. Comp. Bio.- University of Pittsburgh
- All BBSI participants
- NIH-NSF

References

- 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.
- 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.
- 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