

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

The Selfish Organization of the Overlapping HIV-1 Genes tat and rev

Permalink

<https://escholarship.org/uc/item/3bv5p99v>

Author

Fernandes, Jason D.

Publication Date

2013

Peer reviewed|Thesis/dissertation

The Selfish Organization of the Overlapping HIV-1 Genes tat and
rev

by

Jason Dionisio Fernandes

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Pharmaceutical Sciences and Pharmacogenomics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2013

by

Jason Dionisio Fernandes

ACKNOWLEDGMENTS

My collaborators on this work Tyler Faust, Dave Crosby, Sumanjit Mann, Nicolas Strauli, Rob Nakamura and Ryan Hernandez all contributed immensely to this work. Tyler and Dave could easily have published their own separate first author papers with the effort and datasets they contributed here. Tyler contributed the majority of the work on Tat, while Dave did most of the virology.

Alan Frankel has taught me more about science and how to be a good scientist than he probably realizes. Also thanks to my qualifying and thesis committee members: Keith Yamamoto, Nevan Krogan, Jeff Cox, Patsy Babbitt, Raul Andino and Andrej Sali.

All the members of the Frankel lab have been my scientific family, but a special thanks to some of those not mentioned so far: Matt Daugherty, Bhargavi Jayarman, Gwen Jang, Mariana Tioni, Elizabeth Quezada, Ivan D'Orso and David Booth. Also thanks to the Yamamoto and Andino labs, for helpful conversations and use of a yellow bucket.

My friends from graduate school especially, Howard Horng, Stacy Musone, Jen Yokoyama, Dan Sirkis, and Somayeh Ahmadiantehrani have been extremely supportive throughout my graduate career. Michael Hicks and Melissa Calton barely deserve mention. To anyone else I've forgotten here, but who's reading this...apologies and I'm honestly shocked you are reading my thesis.

To the future students building on this project and trying to make sense of this thesis: Good luck! I've left information with the raw data, and you can always contact me in

the likely event that no one else in the lab knows what the heck I was doing. Also, it's not a bad idea to read the actual paper which postdates this thesis.

To my family: I love you all very much.

The Selfish Organization of the Overlapping HIV-1 Genes *tat* and *rev*

Jason Dionisio Fernandes

Abstract

Many genomes contain overlapping genes in which nucleotides are shared in alternative open reading frames. Despite the fact that this phenomena occurs quite frequently in viruses, the manner in which shared nucleotide sequence impacts protein evolution has not yet been fully elucidated. We propose two simple models: 1) A “selfish” arrangement in which one protein dictates the coding sequence of the other and 2) A “compromising” model in which a group of nucleotides encodes critical residues in both frames, thereby forming a viral Achilles’ heel that must satisfy multiple selection pressures.

To test these models, we examine human immunodeficiency virus type 1 (HIV-1) which contains substantial coding overlap including the essential, regulatory genes *tat* and *rev*. Tat and Rev play known and critical roles in the production and processing of the viral RNA, making them ideal candidates for study. Here we combine statistical analyses of existing patient data, a residue-resolution, functional dissection of Tat and Rev, and directed evolution experiments that decouple the overlapped regions to demonstrate that the selfish organization of the *tat/rev* overlap minimizes the constraining effects of overlapped reading frames. This arrangement of overlapped genes may extend to other overlapped genes and also represent an important

mechanism for the creation of genomic novelty and post-transcriptional regulation in small genomes.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS.....	iii
ABSTRACT.....	v
TABLE OF CONTENTS.....	vii
LIST OF TABLES.....	viii
LIST OF FIGURES AND ILLUSTRATIONS	viii
CHAPTER 1: Introduction	1
Introduction	1
Figure Legend	12
Figures	14
CHAPTER 2: Analysis of Patient Datasets	17
Introduction	18
Methods	19
Results and Discussion	21
Summary.....	22
Figure Legend	29
Tables	30
Figures	33
CHAPTER 3: Functional Screening of Tat and Rev.....	37
Introduction	38
Methods	38
Results and Discussion	40
Summary.....	45
Figure Legend	48
Tables	50
Figures	52
CHAPTER 4: Development of a Decoupled, Directed-Evolution Viral Platform.....	59
Introduction	60
Methods	60
Results and Discussion	64
Summary.....	69
Figure Legend	72
Figures	74
CHAPTER 5: Conclusions, Impact and Future Directions.....	81
Introduction	82
Discussion.....	82
Summary.....	88

Figure Legend	93
Figures	94
CHAPTER 6: Proteomic Characterization of Rev.....	96
Introduction	97
Methods	99
Results and Discussion	102
Summary.....	107
Figure Legend	111
Tables	113
Figures	115
CONCLUDING REMARKS	120
APPENDIX A: List of Commonly Used Abbreviations.....	121
APPENDIX B: Supplementary Figures.....	122
Figure Legend	122
Figures.....	125
APPENDIX C: Supplementary Tables	135

LIST OF TABLES

Chapter 2

Table 1: Number of Sequences Used in the Analysis of Each Protein	30
Table 2: Entropy Statistics Across the HIV-1 Genome	31
Table 3: Classification of Residues by Relative Entropy	32

Chapter 3

Table 2: Reclassification of Residues Using Reporter Data	50
---	----

Chapter 6

Table 1: Top Candidates from Proteomic Screen	114
Table 2: Top Candidates from Filtered Analysis	115
Table 3: TMED Proteins Identified in Screen	115
Table 4: Top Candidates from SILAC Experiments	115

Appendix C

Table C1: Raw Data for Library Selection	111
--	-----

LIST OF FIGURES AND ILLUSTRATIONS

Chapter 1

Figure 1: Interpretation of the Genetic Code	14
Figure 2: Organization of the HIV-1 Genome	14
Figure 3: Schematic of the Tat/Rev overlap	15
Figure 4: Function of Tat and Rev	16

Chapter 2

Figure 1: Shannon Entropy Distribution Across the HIV-1 Genome.....	33
Figure 2: Entropy Analysis of the HIV-1 Exome.....	34
Figure 3: Entropy Heat Map of Tat and Rev	35
Figure 4: Classification of Sites by Relative Entropy.....	36

Chapter 3

Figure 1: Schematic of the Reporter Assays	52
Figure 2: Alanine Scanning of Tat and Rev	53
Figure 3: Structural Heat Map of Tat	54
Figure 4: Structural Heat Map of Rev.....	55
Figure 5: Relative Entropies and Activities.....	56
Figure 6: Comparison of Relative Activities of Shared Residues	58

Chapter 4

Figure 1: Scheme for the design of uncoupled viruses	74
Figure 2: Flowchart for Evolution Experiments	74
Figure 3: Replication Rates of Uncoupled Viruses	75
Figure 4: Time Course for Viral Replication.....	75
Figure 5: Diversity of Proviral Libraries	76
Figure 6: Selection of Viruses with Known Relative Fitnesses...	78
Figure 7: Heatmaps of Directed Evolution Libraries	79

Chapter 5

Figure 1: Inhibition of Viral Replication by Rev Overexpression	94
Figure 2: Model for Fitness Advantages of an Overlap.....	95

Chapter 6

Figure 1: Purifications of Rev and Known Interactors	116
Figure 2: Characterization of Rev-TMED Interactions	117
Figure 3: Effect of Overexpression of Candidates in Functional Assay	119
Figure 4: Co-localization studies of Rev and TMEDs	120

Appendix B

Figure B-1: The Gentic Code	101
Figure B-2: Groupings of Amino Acids by Characteristics.....	101
Figure B-3: Unfiltered Protein Alignments	102
Figure B-4: Mean Pairwise Distance for Tat and Rev	104
Figure B-5: Synonymous Rates of Evolution.....	105
Figure B-6: Non-synonymous Rates of Evolution	106
Figure B-7: Representative Flow Cell Image and Q-Score Distribution	107
Figure B-8: Biochemical Implications of Different Overlapped Architectures.....	108
Figure B-9: Phylogenetic History of Retroviruses	109
Figure B-10: Organization of tat and rev in Other Viruses	110

Chapter 1

Introduction

Introduction to Overlapped Genes

The genetic code is interpreted as triplets of bases; as a result any DNA sequence can encode protein information in one of six reading frames (Figure 1, Appendix C1).

Although transcription and post-transcriptional machinery typically interact with a nucleic acid to translate only a single frame, overlapping genes, in which genes share nucleotides but encode different protein products, have been found to naturally occur in a wide variety of species from bacteriophage to mammals^{1,2}. Indeed, coding overlap is quite common in viruses with estimates of 56-75% of viruses containing at least a single instance of overlap^{3,4}, and potentially thousands of overlapping genes in mammals². Despite the relative frequency of this phenomenon, much remains unknown about the relative costs and benefits of this arrangement of genes.

Evolutionary Origins of Overlaps

Overlaps occur via the “overprinting” of an ancestral gene⁵. Overprinting can occur when an alternate start codon arises internally in a gene or when an existing stop codon is lost and causes one gene to extend into the reading frame of another. However the set of selection pressures that lead to the creation of overlaps remains disputed. Pressures commonly cited include: the need for genomic novelty, constraints on genome size, and mechanisms of post-transcriptional regulation^{3,4,6-9}.

The genomic novelty model suggests that overlaps are an important aspect of *de novo* gene creation, a process which allows an organism to explore entirely new regions of protein-coding evolutionary space⁷. Existing genes have several advantages as sites

for *de novo* gene creation as they exist in parts of the genome which are actively transcribed and translated. Thus a simple change in splicing, translation initiation, ribosomal slippage, or stop read-through can take an existing nucleic acid sequence and produce a novel protein that can subsequently be shaped by evolutionary pressures. Moreover, overlapping proteins tend to have unique structural properties that tend towards intrinsic disorder and allow a more diverse sampling of protein space than the classic well-structured folds found so often in single-frame genes^{7,8}.

The genome size model suggests that factors such as polymerase fidelity, and physical packaging considerations, such as cell or capsid size, constrain genome size. Indeed, there is a general positive correlation between genome size and proportion of overlap, however the underlying factor driving this relationship appears to vary³. For instance, although the constraint between cell/capsid size and genome length makes intuitive sense¹⁰, correlations between overlap proportion and capsid size appears to be largely driven by small, icosahedral viruses and less true for those with flexible capsids⁴.

Similarly, the relationships between polymerase error rate and genome size remains controversial. Several studies have shown a negative correlation between these two factors as large genomes with high error rates are susceptible to lethal point mutations (e.g. the creation of unwanted stop codons)^{3,6}. Under this model, overlaps can have protective effects as they allow smaller genomes and thus fewer lethal point mutations per cycle; however this benefit may be offset if that the chance of lethality doubles due to the presence of constraints from the other frame. Simple simulation studies have shown this protective effect of overlaps on mutation rate; however these studies

underestimate the effect of point mutations on a protein as they are not able to estimate the effect of each mutation on protein function³. Moreover the relationship between polymerase error rate and overlap proportion appears to vary depending on the sets and types of viruses examined⁴.

The post-transcriptional regulation model suggests that the arrangement of the genes in shared nucleotides may serve to regulate more complex biological networks via coexpression and coevolution. Several recent studies in both humans and viruses have provided exciting examples that hint at this behavior. In all retroviruses, the ratio of the overlapping viral proteins Gag and Gag-Pol is fixed at approximately 10:1 and disruption of this ratio, in either direction, inhibits replication. In some viruses this ratio is maintained by stop-codon read-through, while in others it is the product of ribosomal slippage. Thus it appears that, despite disparate molecular mechanisms, overlapped frames are essential in regulating essential protein ratios in all retroviruses¹¹. This behavior is also seen in the RNA virus influenza A where the ratio of the NS1 and NEP proteins is tightly controlled by regulation of the splicing of two overlapped frames¹². Control of this ratio is essential in maintaining a “molecular clock” that correctly times the course of viral replication. Even in higher order organisms, overlapping proteins appear to have complex interactions with one another: the human protein ALEX physically interacts with its overlapping partner XL α 5^{13,14}, and the overlapping genes INK4 and ARF both regulate tumor suppression pathways¹⁵.

Each of the selective forces discussed in these models is likely to play important roles in shaping the evolution a dual-coding frame region. However the degree to which each force impacts the residue and codon-level evolution of each alternative reading frame remains largely unknown. For that reason, we decided to explore the effect of evolution in overlapped genes in a representative and experimentally tractable system.

HIV-1 as a Test Case for Studying Overlaps

HIV-1 contains 8 distinct areas of coding overlap (Figure 2) constituting ~8% of its entire genome. This is only slightly higher than the viral average of ~5%, but fairly typical for the lentiviruses as a whole. Moreover, HIV-1's genome is well mapped, sequence datasets are readily available, and most of its genes are well studied with established functions. Together, these factors allow us to investigate how selection pressures acting on a single, overlapped gene have the potential to impact the evolution of its offset counterpart.

Additionally, HIV-1 is a pathogen of particular scientific interest as it is the causative agent of Acquired Immunodeficiency Syndrome (AIDS) a disease which, as of 2011, afflicts more than 34 million worldwide and leads to nearly 1.7 million deaths a year¹⁶. Identifying potential weak points under multiple selection pressures, and understanding the evolutionary constraints of this virus is a critical area of scientific research.

Introduction to Rev and Tat

The overlapped HIV-1 genes *rev* and *tat* pose a particularly suitable pair for the study of overlapped genes. Both are essential in viral replication and their retroviral orthologues have previously served as models in statistical analyses of overlapped evolution^{17,18} (Figure 3). Tat is a transcriptional activator responsible for transcriptional elongation at the HIV-1 promoter via its interactions with host transcription machinery and the viral RNA trans-activation response element (TAR)^{19,20} (Figure 4). Rev is responsible for the nuclear export of partially and unspliced viral RNAs that encode essential late-stage viral proteins. Rev binds a viral RNA element known as the Rev Response Element (RRE) in a highly cooperative manner and then guides these intron-containing RNAs through the nuclear pore via interactions with host nuclear export machinery²¹⁻²³ (Figure 4). In addition to their extensive nucleotide overlap, both Tat and Rev have well characterized functional domains and partial crystal structures (Figure 3)²⁴⁻²⁶. Curiously their partial crystal structures do not share any nucleotides, suggesting that the viruses may make structural and functional tradeoffs. The potential for either coupling or conflict between the two proteins, given their genetic arrangement and established molecular mechanisms, makes them intriguing candidates for detailed study.

Models for Evolution of Tat and Rev

Two simple models can be posited when examining the evolutionary consequences of overlapped genes at the residue level. The first model is a “selfish” model in which one gene dominates and effects the amino acid composition of the other with no known benefit to the second gene. The second model is a “compromising” model in which

both proteins share nucleotides that encode functionally beneficial amino acids in both frames. These occurrences could be thought of as an evolutionary “Achilles’ heel”; targeting such a position with a therapeutic would drastically hinder escape as any mutation would have to satisfy functional constraints in both frames.

To differentiate between selfish and compromising regions of Tat and Rev we use three major approaches each of which is the subject of its own chapter. Chapter 2 focuses on patient datasets that we examine for signatures of conservation and attempt to attribute the conservation to a particular frame. Through this analysis we are able to determine regions of the gene which appear to act selfishly and which ones appear to compromise. In Chapter 3 we perform comprehensive alanine scanning of both proteins to determine the functional contributions of each side chain to each protein. This allows us to validate the classifications of the selfish and compromising sites from Chapter 2. Surprisingly, we find that all sites in Tat and Rev act selfishly and compromising signatures of conservation are likely driven not by function, but by requirements of the genetic code. In Chapter 4, we develop a system to decouple the overlap and perform directed evolution selection experiments to show that conserved sites that do not appear to be functionally important can in fact freely mutate. Those sites that are restricted by function, however, remain conserved though their codon usage patterns differ from the patient datasets. Taken together these findings suggest that, on a residue level, evolution of the individual genes has been selfish and thereby avoided the pitfall of an Achilles’ heel. Analysis of literature studies of other viruses suggests that this arrangement of overlapped genes is not unique to HIV-1, but is likely a

more general outcome of the evolution of overlapped frames. Finally, Chapter 6 provides a more focused examination of Rev function through an examination of Rev's interactions with host factors.

References

1. Sanger, F. *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–695 (1977).
2. Sanna, C. R., Li, W.-H. & Zhang, L. Overlapping genes in the human and mouse genomes. *BMC Genomics* **9**, 169 (2008).
3. Belshaw, R., Pybus, O. G. & Rambaut, A. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research* **17**, 1496–1504 (2007).
4. Chirico, N., Vianelli, A. & Belshaw, R. Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences* **277**, 3809–3817 (2010).
5. Sabath, N., Wagner, A. & Karlin, D. Evolution of viral proteins originated de novo by overprinting. *Molecular Biology and Evolution* (2012). doi:10.1093/molbev/mss179
6. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral Mutation Rates. *Journal of Virology* **84**, 9733–9748 (2010).
7. Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R. & Karlin, D. Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. *Journal of Virology* **83**, 10719–10736 (2009).
8. Kovacs, E., Tompa, P., Liliom, K. & Kalmar, L. Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 5429–5434 (2010).
9. Keese, P. K. & Gibbs, A. Origins of genes: “big bang” or continuous creation? *Proceedings of the National Academy of Sciences of the United States of America* **89**, 9489–9493 (1992).
10. Gregory, T. R. The bigger the C-value, the larger the cell: genome size and red blood cell size in vertebrates. *Blood cells molecules diseases* **27**, 830–843 (2001).
11. Shehu-Xhilaga, M., Crowe, S. M. & Mak, J. Maintenance of the Gag/Gag-Pol Ratio Is Important for Human Immunodeficiency Virus Type 1 RNA Dimerization and Viral Infectivity. *Journal of Virology* **75**, 1834–1841 (2001).
12. Chua, M. a, Schmid, S., Perez, J. T., Langlois, R. a & Tenover, B. R. Influenza a virus utilizes suboptimal splicing to coordinate the timing of infection. *Cell reports* **3**, 23–9 (2013).

13. Nekrutenko, A., Wadhawan, S., Goetting-Minesky, P. & Makova, K. D. Oscillating Evolution of a Mammalian Locus with Overlapping Reading Frames: An XLaS/ALEX Relay. *PLoS Genetics* **1**, e18 (2005).
14. Klemke, M., Kehlenbach, R. H. & Huttner, W. B. Two overlapping reading frames in a single exon encode interacting proteins—a novel way of gene usage. *The European Molecular Biology Organization Journal* **20**, 3849–3860 (2001).
15. Szklarczyk, R., Heringa, J., Pond, S. K. & Nekrutenko, A. Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 12807–12812 (2007).
16. Amfar.org. HIV Statistics: Worldwide. (2013). at <<http://www.amfar.org/About-HIV-and-AIDS/Facts-and-Stats/Statistics--Worldwide/>>
17. McGirr, K. M. & Buehring, G. C. Tax & rex: overlapping genes of the Deltaretrovirus group. *Virus Genes* **32**, 229–239 (2006).
18. McGirr, K. M. & Buehring, G. C. tax and rex Sequences of bovine leukaemia virus from globally diverse isolates: rex amino acid sequence more variable than tax. *Journal of veterinary medicine B Infectious diseases and veterinary public health* **52**, 8–16 (2005).
19. D’Orso, I. & Frankel, A. D. HIV-1 Tat: Its Dependence on Host Factors is Crystal Clear. *Viruses* **2**, 2226–2234 (2010).
20. Ott, M., Geyer, M. & Zhou, Q. The Control of HIV Transcription: Keeping RNA Polymerase II on Track. *Cell host microbe* **10**, 426–35 (2011).
21. Fernandes, J., Jayaraman, B. & Frankel, A. The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNABiol* **9**, 6–11 (2012).
22. Pollard, V. W. & Malim, M. H. The HIV-1 Rev protein. *Annual review of microbiology* **52**, 491–532 (1998).
23. Hammarskjold, M. & Rekosh, D. A long-awaited structure is rev-ealed. *Viruses* **3**, 484–92 (2011).
24. Tahirov, T. H. *et al.* Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature* **465**, 747–751 (2010).

25. Daugherty, M. D., Liu, B. & Frankel, A. D. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nature Structural & Molecular Biology* **17**, 1337–1342 (2010).
26. DiMattia, M. A. *et al.* Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 5810–5814 (2010).

Figure Legend

Figure 1: Interpretation of the Genetic Code

All six possible reading frames for a given sequence of double stranded DNA. Reading frames can start with the first (0), second (+1) or third (+2/-1) nucleotide listed on the sense stand. Each frame has reverse complement analog also read in the 5' → 3' direction. Boxed codons representing the translation product of the labeled frames are shown with the DNA sequence.

Figure 2: Organization of the HIV-1 Genome

Structure of the HIV-1 genome: HIV-1 consists of 9 genes, 8 of which have coding overlap (gray boxes) with at least one other protein. Only *nef* does not overlap with a coding region while *rev* is entirely contained within overlapping regions. One 3-way coding region (blue box) exists in which *rev* (red), *tat*(blue) , and *env* are all translated.

Figure 3: Schematic of the Tat/Rev overlap

A) Genomic organization of *tat* and *rev*. Both proteins are encoded by two overlapping exons (intron represented with dashed lines). Additionally the 5' end of *tat* overlaps with the 3' end of *vpr* and the 3' end of *rev* is entirely contained within *env*. The second exon of both genes overlaps with *env*. Nucleotide numbering (HXB2) and protein sequence (NL4-3) are shown. Respective reading frames and codon arrangements are

shown in the inset. B) Functional organization of Tat and Rev. Previously characterized domains are annotated on 2D representations of the protein. The protein maps are staggered to accurately represent the nucleotide overlap. OD indicates the Rev oligomerization domains. C) Structural organization of Tat and Rev. Crystal structures of Tat (left, residues 1-49) and Rev (right, residues 9-70) with the domains highlighted in B) colored.

Figure 4: Function of Tat and Rev

A) In the absence of Tat: an inactive form of the host elongation factor PTEFb is bound at the promoter and transcriptional elongation is defective. When Tat is present inhibitory portions of the complex are ejected through interactions with TAR and PTEFb. PTEFb then becomes active, correctly phosphorylating Pol II and allowing productive elongation. B) In the absence of Rev: full-length transcripts are spliced in the nucleus by the cellular splicing machinery prior to export to the cytoplasm. These fully spliced transcripts encode early-stage HIV proteins, including Rev. Rev enters the nucleus via its NLS where it binds the RRE and allows unspliced and partially spliced RNAs to circumvent splicing and enter the cytoplasm via an interaction with the host export factor Crm1. Once in the nucleus these RNAs can be translated or packaged into virions.

Figure 1

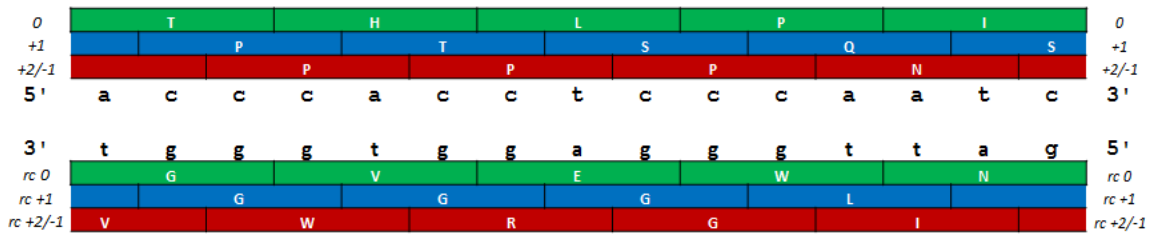


Figure 2

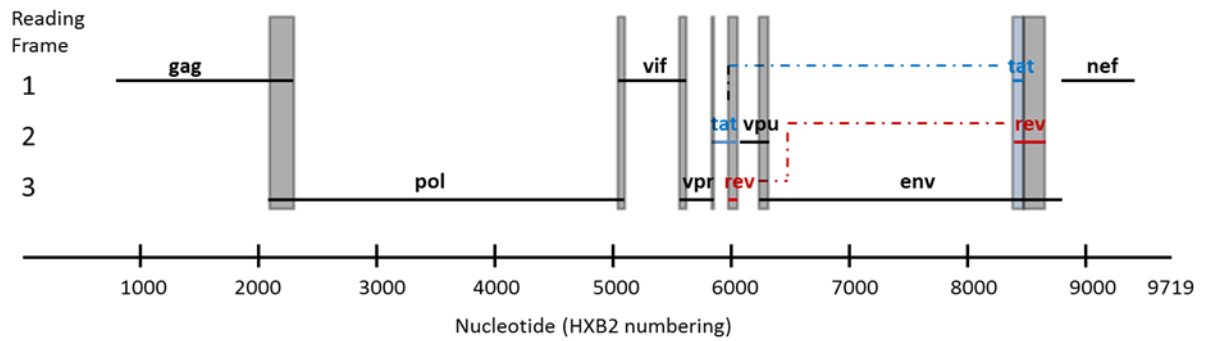
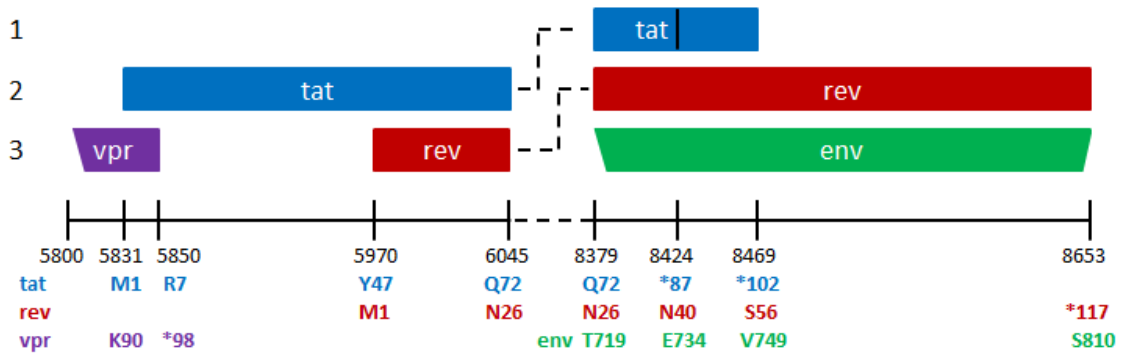
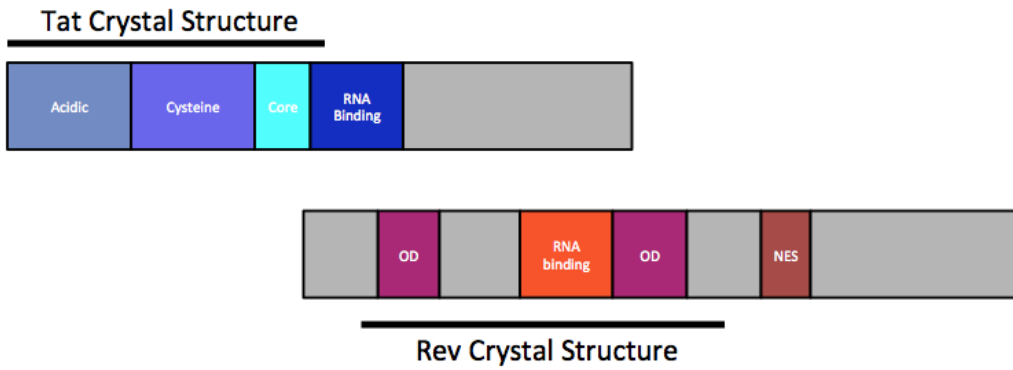


Figure 3

A.



B



C

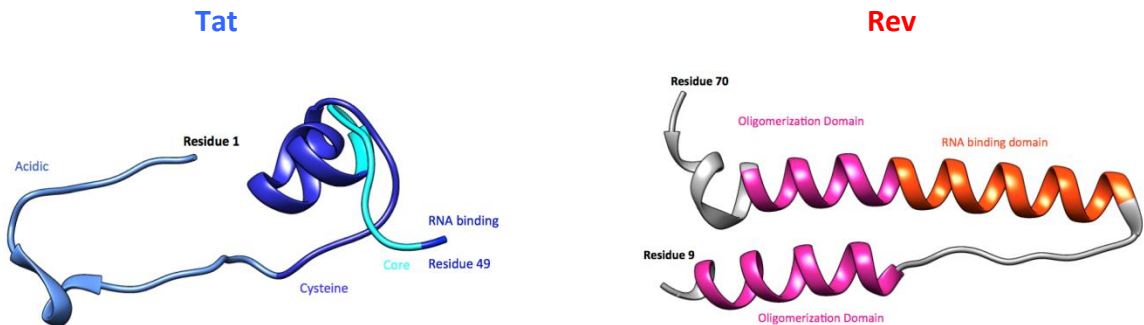
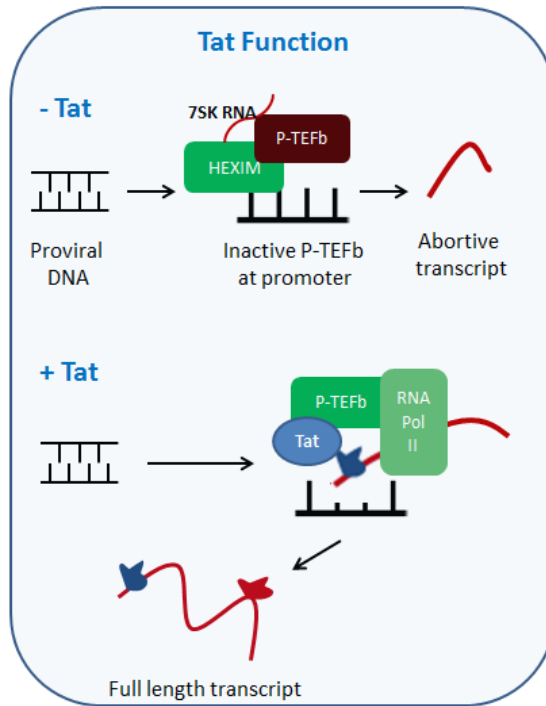
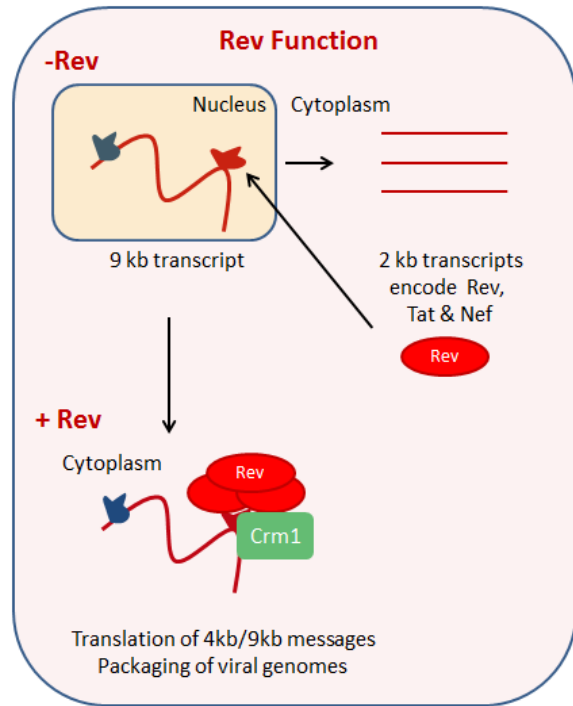


Figure 4

A



B



Chapter 2:

Analysis of Patient Datasets

Introduction

The massive effort to eradicate HIV-1 has resulted in the availability of a vast amount of patient sequence data. Most notably the Los Alamos National Laboratory provides a sequence compendium of every sequence isolate deposited in Genbank since 1987¹.

We sought to examine these datasets to see if we could find signatures of protein selection in overlapping frames and attribute that signature to a particular frame. We first performed this analysis on the entire HIV-1 genome to determine general distinguishing features of overlaps before focusing on the *tat/rev* overlap.

Challenges in Estimating Selection of Overlapped Reading Frames

Analyses of selection from sequence data is traditionally performed by calculating the rate of non-synonymous (dN) versus synonymous substitutions (dS). However, interpretations of these studies in overlapping areas is difficult as most evolutionary models require assumptions, such as a constant rate of neutral substitution, that are not true in these regions. Although techniques exist for analyzing overlapped genes, these strategies are not able to accurately generate site-specific rates²⁻⁴. Despite these caveats, evolutionary analyses of the entire HIV-1 genome report the surprising findings that overlapped areas have relatively poor conservation, a high rate of synonymous substitution, and a high rate of positive selection^{5,6}. A more recent study using molecular clock estimates appears to dispute these findings: the authors find that overlapped regions experience a lower rate of evolution than non-overlapped regions⁷. However the authors caution that these correlations are weak as their models required

the elimination of a significant portion of overlapped areas. Moreover, their datasets follow the evolutionary dynamics of whole genomes from 15 individuals infected from the same source as opposed to the other studies which reflect a much larger degree of viral diversity from thousands of sequences. Because of these complications we sought to use simple metrics (alternative metrics provided in Appendix C.3-6) to provide a coarse picture of the evolutionary landscape of HIV-1 overlaps that would then be further refined by experimental evidence.

Methods

Generation of Patient Datasets for Analysis

The Los Alamos HIV-1 Sequence Database (hiv.lanl.org) provides curated, high-quality alignments for each HIV-1 protein¹. The sequences represented in these alignments represent a highly diverse range of proviruses isolated from infected patients and are filtered to remove sequences with poor quality reads, large insertions and multiple frameshifts. After filtering, these sequences represent approximately 1000 sequences for each protein (Table 1). Each region of each protein was then divided into areas of one, two, or three coding overlaps and statistics generated for each region. Residues which flank overlapped areas were discarded in the analysis. Gaps were eliminated and numbering made to match the HXB2 reference sequence. Additionally for the *rev* and *tat* exons nucleotide sequences were downloaded so that comparisons could be made at both the codon and amino acid level (Appendix C3).

Shannon Entropy Provides a Metric for Sequence Analysis of Overlaps

In order to examine evolutionary characteristics of the overlap compared to the rest of the exome, we calculated the Shannon entropy of each position of the alignment.

Shannon entropy is a well-defined information theory metric:

$$H_n(p_1, p_2, \dots, p_n) = - \sum_{i=1}^{21} p_i \ln(p_i)$$

where H_n represents the Shannon entropy of a position in the alignment, n . For each of these positions there is a probability p_i of amino acid i (there are 21 states; ones for each amino acid, an additional state for the codon). This means that, for any given position in the alignment the Shannon entropy of a site has a minimum value of 0 at a perfectly conserved position, and a maximum value of 3.04 at a perfectly distributed position. Conceptually, Shannon entropy reflects the uncertainty of the identity of a residue given an alignment and provides a more sophisticated metric than sequence conservation, yet requires no assumptions about selection or substitution rate.

Furthermore this metric has successfully been used to quantify diversity in viral sequences, to identify interaction surfaces on proteins, and to analyze overlapping reading frames⁸⁻¹⁰. We performed this analysis for each protein in HIV-1 using the full filtered web alignment (Table 2, Figure 1).

Overlapped Regions in HIV-1 Have a Higher Mean Entropy than Single Coding Regions

Of the seven genes that contain both single and multiple frame regions, five showed higher entropies in the overlapped areas (Figure 1). Indeed, looking globally, multiple coding regions had a higher mean entropy than single coding regions (Figure 2).

Furthermore, this pattern held for Tat, Rev (comparing two and three coding regions) and Env, which are the areas of particular interest in this work. For detailed analysis of the Rev-Tat overlap, filtered alignments containing nucleotide information from (HXB2 numbering) 5831-6045(tat exon 1) and 8379-8653 (rev exon 2) were downloaded, and the analysis was repeated, using both codon based statistics and the appropriate open reading frames (Appendix C3). We also computed global entropy means for each area of overlap (1 frame, 2 frame and 3 frames) and compared their distributions to random sampling of the whole exome. The 2 frame region had a statistically significant different entropy distribution than the whole exome ($p = 0.02$), while the 3 frame region, despite its low number of datapoints, approached significance ($p = 0.07$) (Figure 2).

Classification of Residues by Relative Entropy

We defined a concept of a “relative entropy” in which the raw entropy of each residue was normalized by the mean of the global single frame entropy. We reasoned that the one frame mean entropy represented an average of values coming from functionally required sites (low entropy sites) as well as those sites undergoing neutral drift or active selection all driven by pressures from the single coding frame. Therefore sites with an entropy below the mean should have a higher-than-average probability of playing an important functional role. This assumption was supported by the observation that overlapped regions appeared to have a more complex set of pressures coming from multiple frames, as seen in their overall higher median entropy and broad quartiles (Figure 1, Table 1). Furthermore, the use of a relative entropy normalized by the mean

has been successfully used to assess the functional importance of individual residues in PDZ binding domains¹¹. In following with this line of reasoning, we inferred that those residues whose side chains appeared to be making substantive contributions to protein function, would likely have a relative entropy less than one (as they bear a signature of conservation equivalent to, or better than, the single frame equivalent). This analysis was performed for each site in Tat and Rev (Figure 3).

Pairing of Overlapped Residues

We paired residues from Tat and Rev that shared two nucleotides; for instance, M1 of Rev (atg) shares two nucleotides (at) with Y47 of Tat (tat). Alternative groupings results in a commonality of only 1 base and conveys minimal information as this nucleotide is the wobble base of Rev. Grouping the residues in two nucleotide manner, we graphed Rev relative entropy vs. Tat relative entropy and classified residues according to their positioning within logarithmic quadrants around (1,1) (Figure 4). Sites in Quadrant 1 (S_{none}) have high entropies in both proteins and represent non-conserved positions that do not appear to be critical for either protein's function. Residues in Quadrants 2 (S_{Tat}) and 4 (S_{Rev}) are "selfish" sites that appear to be making functional contributions exclusively to either Tat or Rev respectively. Quadrant 3 (S_{both}) on the other hand represents the "compromising" Achilles' heel residues: sites in this quadrant are conserved at the amino acid level in both proteins. A complete listing of residues and their absolute entropies is given in Table 3.

Results & Discussion

Coding Overlap Appears to be a Relatively Weak Restraint on Protein Evolution

Conventional thinking about overlaps has posited that the presence of an additional selection pressure via an offset reading frame should drive higher conservation than a single reading frame^{7,12}. Indeed, this behavior has been observed in regions of coding/non-coding overlap in HIV-1, where RNA structure can act as an additional restraint on protein evolution¹³. These results suggest that, in HIV-1, this conventional model is incorrect. In independent studies, Snoeck et al. found a high rate of positive selection in overlapped areas while Ngandu et al. found a high rate of synonymous substitution in overlapped areas^{5,6}. These seemingly conflicting results can perhaps be resolved by our observations which demonstrate that an overall low amount of conservation means that a high rate of neutral substitution in one frame will almost certainly lead to a high rate of non-synonymous substitution in the other, thereby giving the appearance of positive selection. These results would be in keeping with the overall higher entropy observed in multiple reading frames.

Interestingly these findings may also reflect a difference between the structural organization of overlapped and non-overlapped proteins. Two studies have found that protein structural considerations are far more likely to drive conservation than the number of reading frames utilized^{5,14}. This finding suggests that the traditional manifestation of protein function via structure may be very different than in a multiple reading frame than a single frame one; indeed, most overlaps appear to code intrinsically disordered proteins and therefore have a high mutation tolerance¹⁵.

Conversely, well –structured, single reading frame regions are much more susceptible to loss of function via point mutation as demonstrated by mutational studies of HIV capsid¹⁶. Thus the lower rate of conservation observed in these dual frame regions may represent a greater mutational tolerance conferred by the structural properties of overlapped proteins.

Classification of Sites as Selfish or Compromising

Pairing the sites in Tat and Rev allows us to identify candidate compromising and selfish positions solely based on their relative entropies. The sites classified as S_{Tat} and S_{Rev} appear to be the most confident classifications as they have only one apparent conservation signature. Only one site appears in the S_{Tat} grouping: R56 of Tat which is part of Tat's functionally required ARM. The corresponding Rev residue, E10, is poorly conserved and has no known functional role, making the dedication of these nucleotides to Tat unsurprising. Similarly the residues identified as selfish for Rev (Table 3) map to known Rev domains (the Rev ARM and OD) and do not correspond to known Tat domains. Although it is tempting to dismiss the category of residues classified in S_{none} as simply unused "genetic potential", it is important to note that they may be important for Env, or have functions that are currently unknown and insensitive to detection by this technique. Indeed a few residues known to be important for Rev structure and function (V16, L18, I55) appear to miss our significance cutoff as they have simple requirements of hydrophobicity as opposed to the requirement of a particular amino acid.

Most curious though, are the potential compromising residues, S_{both} , which show strong conservation in both frames. Most of these residues have well defined roles in either the Tat ARM (R49, K50, K51, R52, R55) or Rev ARM (N40, R42, R43, R46), but their cognate residues in the Rev N-terminus and C-terminus have more controversial roles. The Rev N-terminus has been proposed to interact with several host factors including hnRNP A1 and hnRNP Q, while deletion of Tat's C-terminus appears to slightly hinder viral replication in cell culture^{17,18}. However, although the residues in S_{both} are good candidates for compromising sites, it is still possible that the low entropy is driven only by a requirement in one frame. In essence they may still be selfish positions that give the appearance of functional conservation simply because of restrictions imposed by the offset genetic code. Therefore to fully investigate the role of these residues, we require a more complete understanding of each residue's role in protein function: the subject of Chapter 3.

Summary & Conclusions

By examining the entire HIV-1 genome using Shannon entropy as a statistical measure of variability, we found that overlapped regions had a higher overall variability than single frame regions. This counter-intuitive result suggests that the structural constraints imposed by ordered proteins are, in fact, greater than the imposition of an additional reading frame on an already poorly structured protein. Using the 1-frame region mean as a normalization factor for our entropy values, we classified shared sites in the overlap

as important to both, neither, or one of the proteins. We found several sites that, based solely on relative entropy calculations, appeared to contribute to both proteins.

References

1. Kuiken C, Foley B, Leitner T, Apetrei C, Hahn B, Mizrahi I, Mullins J, Rambaut A, Wolinsky S, K. B. *HIV Sequence Compendium 2010*. (Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 10-03684., 2010).
2. Sabath, N., Landan, G. & Graur, D. A Method for the Simultaneous Estimation of Selection Intensities in Overlapping Genes. *PLoS ONE* **3**, 7 (2008).
3. Hein, J. & Støvlbaek, J. A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. *Journal of Molecular Evolution* **40**, 181–189 (1995).
4. De Groot, S., Mailund, T., Lunter, G. & Hein, J. Investigating selection on viruses: a statistical alignment approach. *BMC Bioinformatics* **9**, 304 (2008).
5. Snoeck, J., Fellay, J., Bartha, I., Douek, D. C. & Telenti, A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology* **8**, 87 (2011).
6. Ngandu, N. K. *et al.* Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. *Virology Journal* **5**, 160 (2008).
7. Alizon, S. & Fraser, C. Within-host and between-host evolutionary rates across the HIV-1 genome. *Retrovirology* **10**, 49 (2013).
8. Pan, K. & Deem, M. W. Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. *Journal of the Royal Society Interface the Royal Society* **8**, 1644–1653 (2011).
9. Stewart, J. J. *et al.* A Shannon entropy analysis of immunoglobulin and T cell receptor. *Molecular Immunology* **34**, 1067–82 (1997).
10. Zaaijer, H. L., Van Hemert, F. J., Koppelman, M. H. & Lukashov, V. V. Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. *The Journal of general virology* **88**, 2137–2143 (2007).
11. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–42 (2012).
12. Hughes, A. L., Westover, K., Da Silva, J., O'Connor, D. H. & Watkins, D. I. Simultaneous Positive and Purifying Selection on Overlapping Reading Frames of

the tat and vpr Genes of Simian Immunodeficiency Virus. *Journal of Virology* **75**, 7966–7972 (2001).

13. Sanjuán, R. & Bordería, A. V. Interplay between RNA structure and protein evolution in HIV-1. *Molecular Biology and Evolution* **28**, 1333–1338 (2011).
14. Woo, J., Robertson, D. L. & Lovell, S. C. Constraints on HIV-1 diversity from protein structure. *Journal of Virology* **84**, 12995–13003 (2010).
15. Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R. & Karlin, D. Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. *Journal of Virology* **83**, 10719–10736 (2009).
16. Rihn, S. J. *et al.* Extreme genetic fragility of the HIV-1 capsid. *PLoS pathogens* **9**, e1003461 (2013).
17. Hadian, K. *et al.* Identification of a Heterogeneous Nuclear Ribonucleoprotein-recognition Region in the HIV Rev Protein. *The Journal of Biological Chemistry* **284**, 33384–33391 (2009).
18. Neuveut, C., Scoggins, R. M., Camerini, D., Markham, R. B. & Jeang, K.-T. Requirement for the second coding exon of Tat in the optimal replication of macrophage-tropic HIV-1. *Journal of Biomedical Science* **10**, 651–660 (2003).

Figure 1: Shannon Entropy Distribution across the HIV-1 Genome. Box and whiskers plots (min, max, median) are shown for each area of overlap (1, 2 or 3). The leftmost boxes represent the 3' end of the genome and the rightmost end represents the 3' end. For each overlapped area, entropies calculated from the same nucleotide sequence but different coding frames are shown. Single frame regions are colored red and three frame regions are colored purple. Overlapped areas tend to have higher median entropy as well as a broader range.

Figure 2: Entropy Analysis of the HIV-1 Exome. Box and whiskers plots for the global behavior of 1, 2 and 3 frame reading regions of the HIV-1 genome; 2 and 3 frame distributions represent the union of entropies from all residues entirely within the overlapping areas. Median, max, min, 1st and 3rd quartile values are all shown, and the mean value for each distribution is marked with an X. The means rises in a statistically significant manner in the 2 frame regions ($p = 0.02$).

Figure 3: Entropy Heat Map of Tat and Rev. Heat map showing the relative entropies of Tat and Rev residues in an overlapped context. Red indicates a low entropy (high conservation) while dark blue indicates a high entropy (low conservation).

Figure 4: Classification of Sites by Relative Entropy. Entropy scatter plot of residue pairs (sharing 2 nt, i.e. Rev M1 and Tat Y47) in the *tat/rev* overlap. The pairs can be grouped into four classes (colored) depending on their relative entropy in both proteins. Residues appear to either being selfish selected for one protein (S_{Rev} or S_{Tat} , i.e. quadrants II and IV), both (S_{both} i.e. Quadrant III) or neither (S_{none} , Quadrant I).

Gene	Number of Sequences Used
<i>Gag</i>	3120
<i>Pol</i>	1805
<i>Env</i>	2869
<i>Rev</i>	1542
<i>Tat</i>	1443
<i>Vif</i>	2086
<i>VpU</i>	2713
<i>VpR</i>	1819
<i>Nef</i>	2707

Table 1: Number of Sequences Used in the Analysis of Each Protein

1 frame	nt start	nt end	aa start	aa end*	avg entropy	standard deviation	frame	Length (nt)
Gag	790	2084	1	431.66	0.31	0.41	1	1294
Pol	2293	4229	69.33	986.33	0.19	0.30	3	1936
Vif	5097	5558	18.66	172.66	0.43	0.47	1	461
Vpr	5620	5830	20.33	90.66	0.35	0.46	3	210
Tat	5851	5969	6.66	46.33	0.45	0.59	2	118
Vpu	6062	6224	1	54.33	0.71	0.55	2	162
Env	6311	8378	28.66	718	0.59	0.68	3	2067
Env	8654	8795	809.66	857	0.61	0.48	3	141
Nef	8797	9417	1	207	0.48	0.50	1	620

2 frames	nt start	nt end	aa start	aa end*	avg entropy	standard deviation	frame	Length (nt)
Gag/Pol	2085	2292						207
Gag			431.66	501	0.58	0.51	1	
Pol			1	69.33	0.63	0.46	3	
Pol/Vif	5041	5096						55
Pol			985.33	1004	0.21	0.27	3	
Vif			1	18.66	0.12	0.16	1	
Vif/Vpr	5559	5619						60
Vif			172.66	193	0.51	0.26	1	
Vpr			1	20.33	0.26	0.26	3	
Vpr/Tat	5831	5850						19
Vpr			90.66	97	0.39	0.37	3	
Tat			1	6.66	0.26	0.24	2	
Tat/Rev	5970	6045						75
Tat			46.33	71.66	0.76	0.57	2	
Rev			1	25.33	0.50	0.51	3	
Vpu/Env	6225	6310						85
Vpu			54.33	83	0.81	0.52	2	
Env			1	28.66	0.83	0.48	3	
Rev/Env	8470	8653						183
Rev			55.66	116	0.76	0.47	2	
Env			748.33	809.66	0.58	0.44	3	

3 frames	nt start	nt end	aa start	aa end*	avg entropy	standard deviation	frame	Length (nt)
Tat/Rev/Env	8379	8469						90
Tat			71.66	102	0.84	0.48	1	
Rev			25.33	55.66	0.48	0.53	2	
Env			719	748.33	0.71	0.47	3	

Table 2: Summary of Entropy Statistics Across the HIV-1 Genome

Summary table of entropy statistics and region definitions of areas of overlap for the HIV-1 genome. Numbering follows the HXB2 reference strain. *Stop codon included in numbering. Fractional residues indicate how much of the codon is included in the overlap (i.e. 71.66 means residue 72 has 2 nt in the region).

S_{none} (High Tat, High Rev)					
Tat			Rev		
Pos	Entropy	AA	Pos	Entropy	AA
53	1.13	R	7	1.30	D
54	1.29	Q	8	2.18	S
57	3.09	R	11	3.42	E
60	3.41	Q	14	3.76	R
61	3.31	N	15	2.32	T
62	2.20	S	16	1.45	V
63	3.43	Q	17	1.09	R
64	3.57	T	18	3.19	L
67	4.15	A	21	3.64	L
74	3.97	T	28	3.05	P
76	1.57	Q	30	3.97	N
77	4.08	S	31	2.50	P
78	2.12	R	32	2.28	E
85	2.10	K	39	1.84	R
93	2.58	R	47	2.51	E
96	1.90	E	50	1.32	R
97	2.12	T	51	1.51	Q
99	2.18	P	53	4.18	H
100	5.03	F	54	3.59	S
101	3.04	D	55	1.48	I

S_{both} (Low Tat, Low Rev)					
Tat			Rev		
Pos	Entropy	AA	Pos	Entropy	AA
47	0.45	Y	1	0.14	M
48	0.13	G	2	0.13	A
49	0.17	R	3	0.19	G
50	0.17	K	4	0.17	R
51	0.15	K	5	0.50	S
52	0.38	R	6	0.48	G
55	0.13	R	9	0.34	D
65	0.99	H	19	0.21	I
66	0.11	Q	20	0.79	K
71	0.97	K	25	0.50	S
72	0.06	Q	26	0.08	N
73	0.49	P	27	0.09	P
79	0.55	G	33	0.17	G
82	0.61	T	36	0.35	Q
83	0.22	G	37	0.12	A
86	0.57	E	40	0.17	N
88	0.69	K	42	0.14	R
89	0.53	K	43	0.15	R
91	0.39	V	45	0.10	W
92	0.36	E	46	0.23	R

S_{Tat} (Low Tat, High Rev)					
Tat			Rev		
Pos	Entropy	AA	Pos	Entropy	AA
56	0.34	R	10	1.44	E

S_{Rev} (High Tat, Low Rev)					
Tat			Rev		
Pos	Entropy	AA	Pos	Entropy	AA
58	3.17	A	12	0.22	L
59	1.29	H	13	0.56	I
68	3.54	S	22	0.47	L
69	2.88	L	23	0.24	Y
70	1.95	S	24	0.48	Q
75	2.16	S	29	0.08	P
80	2.09	D	34	0.83	T
81	2.16	P	35	0.09	R
84	2.31	P	38	0.65	R
87	1.29	*	41	0.25	R
90	1.01	K	44	0.09	R
94	2.11	E	48	0.38	R
95	1.23	T	49	0.20	Q
98	1.54	D	52	0.46	I
102	2.78	*	56	0.94	S

Table 3: Classification of Residues by Relative Entropy. Grouping of each site by relative entropy. The position, identify and relative entropy values are shown for each site.

Figure 1

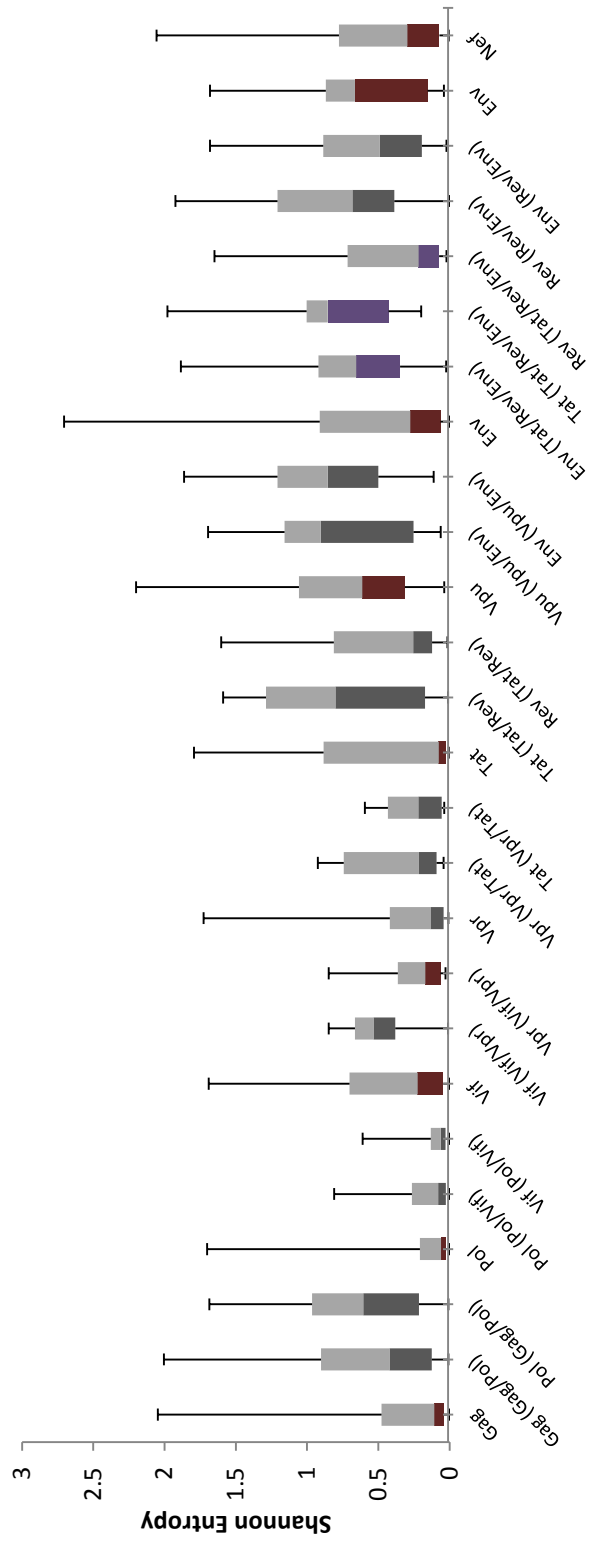
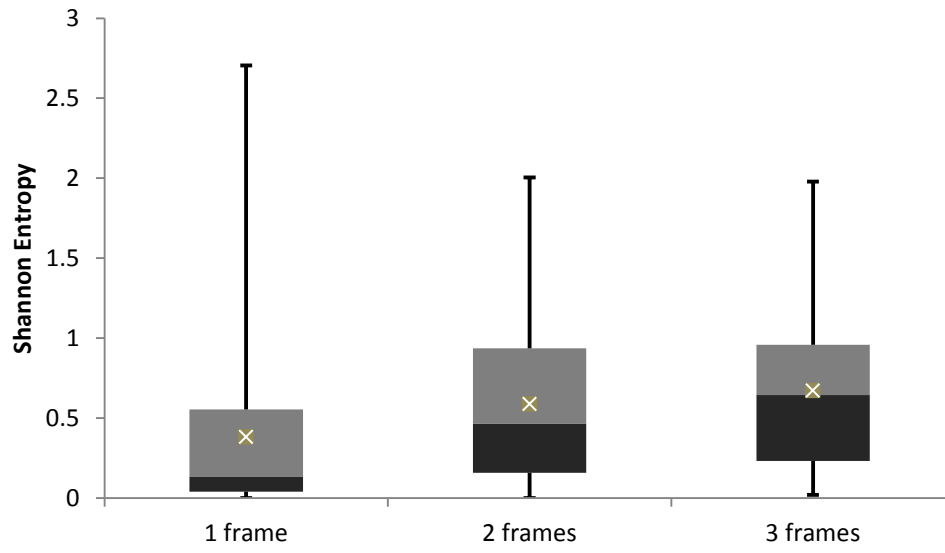


Figure 2



	Whole Exome	One gene	Two genes	Three genes
Mean entropy	0.42	0.38	0.59	0.67
p-value	-----	0.47	0.02*	0.07

Figure 3

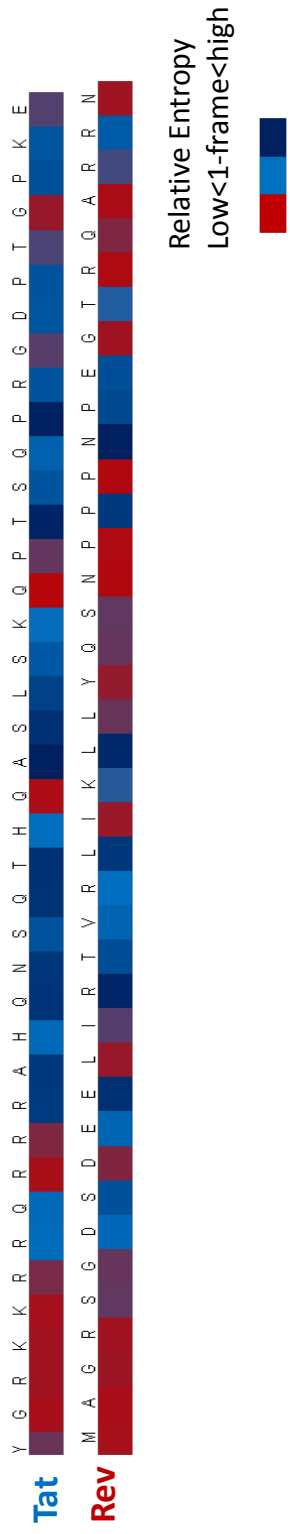
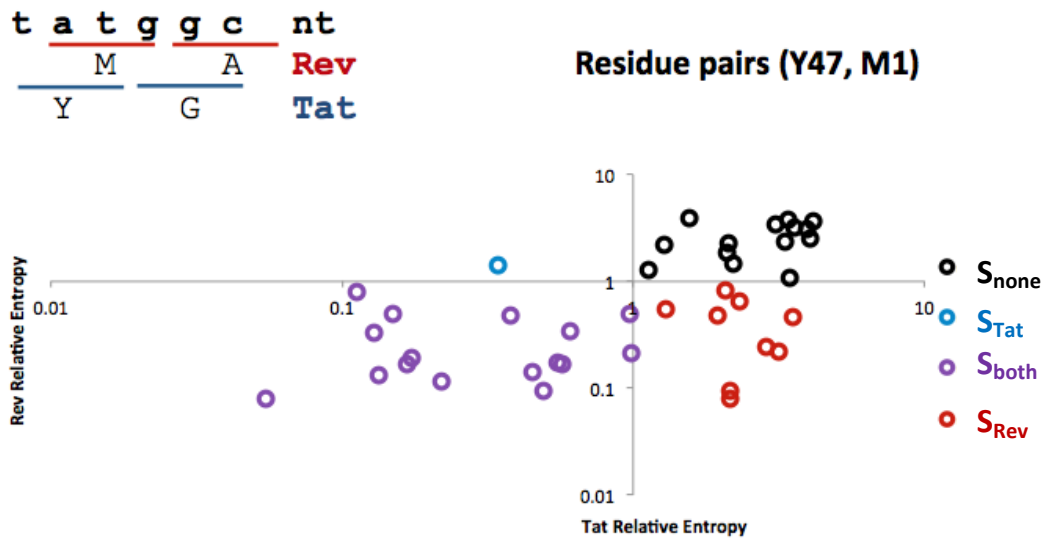


Figure 4



Chapter 3

Functional Screening of Tat and Rev

Introduction

Tat and Rev have well established functions in the viral replication cycle. Tat acts as a transcriptional activator allowing productive elongation from the HIV promoter through interactions with the cellular elongation factor PTEFb and the viral RNA element TAR¹⁻⁷. Rev allows intron containing mRNAs that encode essential viral proteins to circumvent the cellular splicing machinery and enter the cytoplasm where they can be translated or packed into virions^{8,9}. Because of their essentiality in HIV-1 replication, these proteins have been subjected to extensive mutagenesis in a wide variety of contexts^{6,10-12}. However, to date, there has been no systematic and comprehensive scan that examines the contributions of each residue to each protein's function. In this chapter, we perform such a screen and use the results to validate the classifications of residues from Chapter 2.

Alanine Scanning Provides a Mechanism of Measuring Side Chain Contributions

Point alanine mutations have proven useful in generating high resolution epitope maps of proteins as these mutations minimize disruptions to global protein structure while minimizing a single side chain¹³. Alterations in function and protein structure can therefore be directly linked to a single site on the protein. We generated such maps for Tat and Rev using cell based reporters specific to their established functions.

Methods

Generation of Tat and Rev point mutants

In order to generate every possible for Tat and Rev we performed whole plasmid site-directed mutagenesis on mammalian expression plasmids containing sequence corresponding to mature RNAs of *tat* and *rev*. Reference Tat was cloned in a pcDNA4-TO-CStrep backbone while Rev was cloned similarly in a pcDNA4-TO-Nstrep vector. To create each alanine point mutant we performed 200 independent site directed mutagenesis reactions (one for each position in each protein) in which the targeted codon had its first two bases changed to GC to encode an alanine. Each mutagenesis oligonucleotide had 15 flanking nucleotides on each side of the mutation site. Initiator methionines and reference alanines (Rev: A2, A37, A68, Tat: A21, A42, A58, A67) were left unmutated. Each mutant was sequenced and tested for expression by Western blot.

Cell-based Reporter Assays for Tat and Rev

For the Tat reporter assay, the Tat mutant and a renilla control were transfected into a previously described cell line harboring an integrated copy of firefly luciferase under control of the HIV-1 LTR¹⁴. 1 ng of Tat DNA was transfected in a 96 well dish; 24 hours later cells were lysed, and renilla and luciferase values were measured using a luminometer. For the Rev assays 250 ng of pCMVgagpolRRE were cotransfected with 20 ng of a Rev mutant into 293 cells in a 24 well format¹⁵. Transfections were carried out using Polyjet (Signagen) in triplicate for each mutant. Cells were then lysed in 50 mM Tris 7.5, 150 mM NaCl, 0.5% NP-40 and intracellular p24 levels measured by ELISA.

Resulting activities were normalized by the activity of the reference sequence with the native alanines serving as positive controls and empty vector serving as a negative control (Figure 2). These relative activities were grouped in a manner similar to the entropy analysis (Figure 3, Figure 4).

Results and Discussion

Alanine Scanning Using Cell-Based Reporters Provides a Residue Resolution Functional Map of Each Protein

In order to determine how accurately the entropy analysis from Chapter 2 recapitulated the contribution of each residue to protein function, we generated complete and separate sets of alanine point mutants for both Tat and Rev in standard mammalian expression vectors. This approach had the advantage of removing the overlap as each protein was tested, independently of the other, for its known function in a cell-based reporter. The activity of each Tat mutant was determined by its ability to transactivate a cell line containing an integrated firefly luciferase gene under the control of the HIV promoter¹⁴. For the Rev mutants, activity was tested by cotransfection of the mutant and a RRE containing gag-pol reporter construct that produced p24 only in the presence of Rev (Figure 1)¹⁵. In both cases activities were normalized to the activity of the reference sequence to produce relative activities. Loss of function indicated a functional contribution of the mutated side chain. This dataset is the first mutagenesis screen that provides complete and systematic coverage in consistent cell-based reporter assays for both Tat and Rev function (Figure 2).

Alanine Scan Accurately Recapitulates the Current Structural Understanding of Tat and Rev

Mapping the functional data onto the existing partial crystal structures of Tat and Rev clearly illustrates the known interfaces for both proteins (Figures 3 & 4). The activation domain (AD) of Tat was solved in complex with the host transcriptional complex PtefB (consisting of Cyclin T1 and CDK9)¹⁶. Tat mutations that abrogate function include important structural cysteines, and residues buried in the Tat-cyclin interface (Figure 3BC). Unfortunately the Tat ARM is not present in the structure and, therefore, cannot be visualized structurally, although mutation of the ARM appears to impair function.

The Rev crystal structure consists of an N-terminal region containing the ARM flanked by the OD^{17,18}. Both domains are clearly illustrated via the reporter data with mutations in the ARM diminishing but not abolishing function. Although the structures of the Rev C-terminus remains unknown, a peptide structure of the Rev NES bound to its export factor Crm1 has been solved¹⁹. Analysis of this structure reveals that mutation of each binding residue severely inhibits Rev function (Figure 4B).

Analysis of Residue Classifications based on Reporter Activity

Plots of the entropy data versus the functional data for each of the proteins (Figure 3) revealed several expected correlations. As our mutations resulted in the loss of a side chain and our entropy analysis was most sensitive to signatures of conservation, we emphasized loss of function mutations. In order to validate our relative entropy approach from Chapter 2, we examined Tat's large single frame region which

corresponds to its activation domain. As the only known selection pressure in this region is Tat function, we expected there to be a strong correlation between relative entropy and relative activity (i.e. sites which are strongly conserved are essential to protein structure and function). Majority of sites (81%) with a relative entropy less than 1 had a detrimental effect on activity, and those with high entropy generally appeared amenable to mutation (Figure 3A). Only one site, F32, had a significant loss in activity yet a high entropy; analysis of the patient data sets indicates that this site has tolerance for several bulky amino acids (F/W/Y) but not alanine.

We then examined the effectiveness of the entropy analysis matched in the multiple reading frame regions where selection pressures are less clear. Mutations of S_{none} residues did not generally affect either protein's function, while S_{Tat} affected only Tat and S_{Rev} affected only Rev. This behavior can clearly be seen in Figure 3B & Figure 3C. S_{Rev} sites act equivalently to S_{none} sites when considering only Tat entropy and activity; likewise S_{Tat} sites act like S_{none} when considering only Rev entropy and activity. Notably, some S_{none} sites show a loss of activity perhaps reflecting a moderately mutable site but with fragility to alanine, similar to the F32 case in the Tat single frame region.

Curiously, some of the S_{both} residues had no noticeable effect on the function in at least one of the assays, despite the fact that these residues are highly conserved in viruses. To see if any single nucleotide contributed to both known functions of Tat and Rev we graphed the relative activity of overlapped residues against one another (Figure 4). Plotting the residues and performing classifications solely on relative activities, we find

no sites remain in the S_{both} grouping. This implies that, for these roles of Tat and Rev, no single nucleotide appears to be selected on a functional level for both proteins.

The Tat/Rev Overlap Has No Compromising Residues

Despite the entropic classification of multiple residues in the S_{both} category, functional data suggests the source of high conservation in these pairings always came solely from the functional requirement of a single protein and the subsequent restrictions imposed by the genetic code. This startling result implies that the individual nucleotides in the overlap have individual been parceled out for a single protein and that there exists no nucleotide that encodes a functionally important residue for both proteins.

The reclassification of residues based on activities reflects several strengths and weaknesses of our previous entropy-based classifications (Table 1). Most importantly, no residue moves from S_{Tat} to S_{Rev} indicating that the relative entropy metric is able to clearly distinguish the difference between strong conservation signals of two separate proteins' function. Moreover, the elimination of the S_{both} category results in many residues moving to classifications in line with known domain organization of Tat and Rev. Majority of the residues moving from S_{both} to S_{Tat} make up the Tat ARM (Tat residues K50, K51, R52, R53, R55, and R56)⁵ while those moving from S_{both} to S_{Rev} make up the Rev OD (Rev I19) and ARM (Rev N40)^{7,11,20}. However several limitations of our entropy metric are highlighted in other reclassifications. The transitions from S_{none} to either S_{Rev} or S_{Tat} provide several examples of these limits. Despite having a relatively low absolute entropy, Tat residue R53 just misses correct classification, indicating that

threshold cutoff, based on the global one-frame entropy conservation, maybe be too conservative. Similarly this method is insensitive to mutations between amino acids of similar chemical properties and functionally critical positions such as Rev L18 and L21 which easily transition between I and L miss the significance threshold. Using a reduced amino acid alphabet by grouping similar amino acids overcomes this problem but results in a loss of sensitivity at sites across the protein and depends on the manner in which amino acids are grouped (Appendix C2).

Surprisingly several residues move into the S_{none} category indicating that, despite a strong signal of conservation, these residues do not appear to make any contribution to protein function. One possible explanation for these transitions is that the entropy analysis is correct, but a single mutation to alanine may simply be insufficient to disrupt a functional interface and thus escapes detection in our assay. A trivial case displaying this inability of alanine mutations to fully capture the mutational space of each protein is seen in site TatG48/Rev A2 which has a low entropy for both sites; as the reference sequence has a potential critical alanine we are not able to accurately probe the functional significance of that position. Another explanation for the behavior of these sites involves selection pressures outside our assay such as requirements of Envelope (in the 3 frame region) or unknown Tat and Rev functions. A simple example of an outside pressure missing from our reporter system is seen in the sites Tat K71/Rev S25 and Tat P73/Rev P27. These sites overlap contain essential parts of splice donor and acceptor sequence at the exon/intron boundaries of the *tat/rev* coding region. In a viral context

conservation of this site is essential; however our reporter system uses a mammalian expression construct based upon cDNA where this splicing is irrelevant.

Although the combination of our entropy and functional analysis establishes a clearer picture which indicates that tat and rev are organized in a selfish manner with no compromising sites, the question remains why certain selfish sites appear to be conserved in both frames. If the high conservation of the functionally selfish S_{both} residues is driven solely by functional requirements in one frame (with sequence constraints imposed by the genetic code conserving the alternate frame), then, in an uncoupled context, these reclassified residues should be free to mutate. If, however, these sites are conserved for reasons such as alternative function, then they should remain conserved even when the constraint of the overlap is removed. In Chapter 4 we establish an experimental methodology to directly address this question.

Summary

We determined the functional contribution of each residue in both Tat and Rev and compared these values to our relative entropy values. We find that, based on activity data, no site fits the S_{both} classification. It appears that the entropy signatures identified in these residues may be due to requirements of the genetic code, or selection pressures absent in our functional assays.

References

1. D'Orso, I. & Frankel, A. D. HIV-1 Tat: Its Dependence on Host Factors is Crystal Clear. *Viruses* **2**, 2226–2234 (2010).
2. Ott, M., Geyer, M. & Zhou, Q. The Control of HIV Transcription: Keeping RNA Polymerase II on Track. *Cell host microbe* **10**, 426–35 (2011).
3. Madore, S., Tiley, L., Malim, M. H. & Cullen, B. R. Sequence Requirements for Rev Multimerization in Vivo. *Virology* **202**, 186–194 (1994).
4. Madore, S. J. & Cullen, B. R. Genetic analysis of the cofactor requirement for human immunodeficiency virus type 1 Tat Genetic Analysis of the Cofactor Requirement for Human Immunodeficiency Virus Type 1 Tat Function. **67**, (1993).
5. Kuppaswamy, M., Subramanian, T., Srinivasan, A. & Chinnadurai, G. Multiple functional domains of Tat, the trans-activator of HIV-1, defined by mutational analysis. *Nucleic Acids Research* **17**, 3551–3561 (1989).
6. Zapp, M. L., Hope, T. J., Parslow, T. G. & Green, M. R. Oligomerization and RNA binding domains of the type 1 human immunodeficiency virus Rev protein: a dual function for an arginine-rich binding motif. *Proceedings of the National Academy of Sciences of the United States of America* **88**, 7734–8 (1991).
7. Hope, T. J., McDonald, D., Huang, X., Low, J. & Parslow, T. G. Mutational Analysis of the Human Immunodeficiency Virus Type 1 Rev Transactivator: Essential Residues Near the Amino Terminus. *Journal of virology* **64**, 5360–5366 (1990).
8. Fernandes, J., Jayaraman, B. & Frankel, A. The HIV-1 Rev response element: an RNA scaffold that directs the cooperative assembly of a homo-oligomeric ribonucleoprotein complex. *RNABiol* **9**, 6–11 (2012).
9. Pollard, V. W. & Malim, M. H. The HIV-1 Rev protein. *Annual review of microbiology* **52**, 491–532 (1998).
10. Malim, M. H., Böhnlein, S., Hauber, J. & Cullen, B. R. Functional dissection of the HIV-1 Rev trans-activator--derivation of a trans-dominant repressor of Rev function. *Cell* **58**, 205–214 (1989).
11. Jain, C. & Belasco, J. G. Structural model for the cooperative assembly of HIV-1 Rev multimers on the RRE as deduced from analysis of assembly-defective mutants. *Molecular cell* **7**, 603–14 (2001).

12. D'Orso, I. *et al.* Transition step during assembly of HIV Tat:P-TEFb transcription complexes and transfer to TAR RNA. *Mol Cell Biol* **32**, 4780–4793 (2012).
13. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081–1085 (1989).
14. D'Orso, I. & Frankel, A. D. RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. *Nature Structural & Molecular Biology* **17**, 815–821 (2010).
15. Smith, A. J., Cho, M. I., Hammarskjöld, M. L. & Rekosh, D. Human immunodeficiency virus type 1 Pr55gag and Pr160gag-pol expressed from a simian virus 40 late replacement vector are efficiently processed and assembled into viruslike particles. *Journal of Virology* **64**, 2743–2750 (1990).
16. Tahirov, T. H. *et al.* Crystal structure of HIV-1 Tat complexed with human P-TEFb. *Nature* **465**, 747–751 (2010).
17. Daugherty, M. D., Liu, B. & Frankel, A. D. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. *Nature Structural & Molecular Biology* **17**, 1337–1342 (2010).
18. DiMattia, M. A. *et al.* Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 5810–5814 (2010).
19. Güttler, T. *et al.* NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1. *Nature Structural & Molecular Biology* **17**, 1367–1376 (2010).
20. Battiste, J. L. *et al.* Alpha Helix-RNA Major Groove Recognition in an HIV-1 Rev Peptide-RRE RNA Complex. *Science* **273**, 1547–1551 (1996).

Figure Descriptions

Figure 1: Schematic of the Reporter Assays. A) Tat function is probed using a cell line containing the firefly luciferase gene under the control of the HIV-1 promoter. In the absence of Tat, transcriptional elongation is defective and no luciferase is produced. In the presence of reference Tat or a point mutant luciferase production is indicative of Tat-mediated activation. B) Rev function is used by cotransfection of a reporter consisting of a truncated HIV-1 genome. In the absence of Rev, the gag mRNA is spliced and no gag is expressed. Gag expression is indicative of Rev-mediated export.

Figure 2: Comprehensive Alanine Scanning of Tat and Rev. Relative activities of each alanine mutant. Activities are normalized to an average reference value performed in parallel with the mutants. Error bars represent the standard deviation from 5 biological replicates. Activities of Tat mutants are shown in A) and Rev in B).

Figure 3: Structural Heat Map of Tat A) Crystal structure of residues 1-49 of Tat with reporter data overlaid. Several interfaces, including a cluster of structural B) cysteines and the C) cyclin contact surface are clearly visible. CDK9 is shown in yellow and Cyclin T1 in green.

Figure 4: Structural Heat Map of Rev A) Crystal structure of Rev residues 9-63 with reporter data overlaid. The ARM and OD are clearly visible. B) Structure of a peptide consisting of the Rev NES (LQLPPLERLTL) in complex with Crm1 and with reporter data overlaid. All binding residues show a loss of function.

Figure 5: Comparison of Relative Entropies and Activities A) Correlation between relative entropy and activity in the single frame region of Tat is relatively strong as most residues with low entropy also exhibit low activity. B) Scatter plots of relative entropies from the patient data vs. relative function for both Tat and Rev C). Colors from the classifications made solely on relative entropy are also shown.

Figure 6: Comparison of Relative Activities of Shared Residues. Scatter plot of Tat and Rev activities with residues that share 2 nt once again paired. Coloring represents class assignments made on relative entropies. No residues fit the S_{both} classification based on activity data.

Reclassified to S_{tat}

Tat			Rev			Previous Class
Residue	Activity	Entropy	Residue	Activity	Entropy	
Y47	0.96	0.45	M1	1.00	0.14	S_{both}
R49	0.78	0.17	G3	1.56	0.19	S_{both}
K50	0.31	0.17	R4	1.50	0.17	S_{both}
K51	0.47	0.15	S5	1.60	0.50	S_{both}
R52	0.44	0.38	G6	2.00	0.48	S_{both}
R55	0.28	0.13	D9	1.52	0.34	S_{both}
R53	0.35	1.13	D7	1.74	1.30	S_{none}
R57	0.77	3.09	E11	0.99	3.42	S_{none}

Reclassified to S_{rev}

Tat			Rev			Previous Class
Residue	Activity	Entropy	Residue	Activity	Entropy	
Q60	1.04	3.41	R14	0.49	3.76	S_{none}
T64	1.39	3.57	L18	0.76	3.19	S_{none}
A67	1.00	4.15	L21	0.93	3.64	S_{none}
K85	1.54	2.10	R39	0.33	1.84	S_{none}
Q72	1.52	0.06	N26	0.05	0.08	S_{both}
H65	1.45	0.99	I19	0.32	0.21	S_{both}
T82	1.44	0.61	Q36	0.31	0.35	S_{both}
E86	1.75	0.57	N40	0.71	0.17	S_{both}

Reclassified to S_{none}

Tat			Rev			Previous Class
Residue	Activity	Entropy	Residue	Activity	Entropy	
G48	1.42	0.13	A2	1.08	0.13	S_{both}
Q66	1.34	0.11	K20	1.62	0.79	S_{both}
K71	1.30	0.97	S25	1.84	0.50	S_{both}
P73	1.81	0.49	P27	2.04	0.09	S_{both}
S75	1.94	2.16	P29	1.87	0.08	S_{both}
G79	1.44	0.55	G33	1.95	0.17	S_{both}
P73	1.81	0.49	P27	2.04	0.09	S_{both}
G83	1.47	0.22	A37	1.23	0.12	S_{both}
D80	1.19	2.09	T34	1.05	0.83	S_{rev}

Table 1: Reclassification of Residues Based Upon Reporter Data. Residues whose classification changes upon consideration of the reporter data is shown. Residues originally in S_{both} are colored purple, S_{Tat} blue, S_{Rev} red, and S_{none} black.

Figures

Figure 1

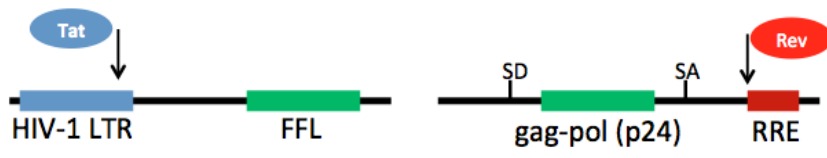
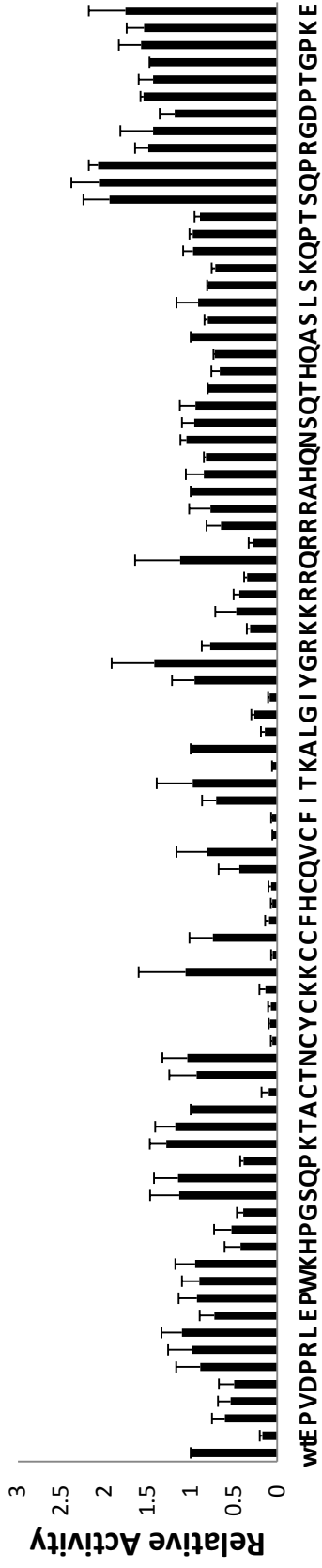


Figure 2

Tat Relative Activities Across Entire Protein



Rev Relative Activities Across Entire Protein

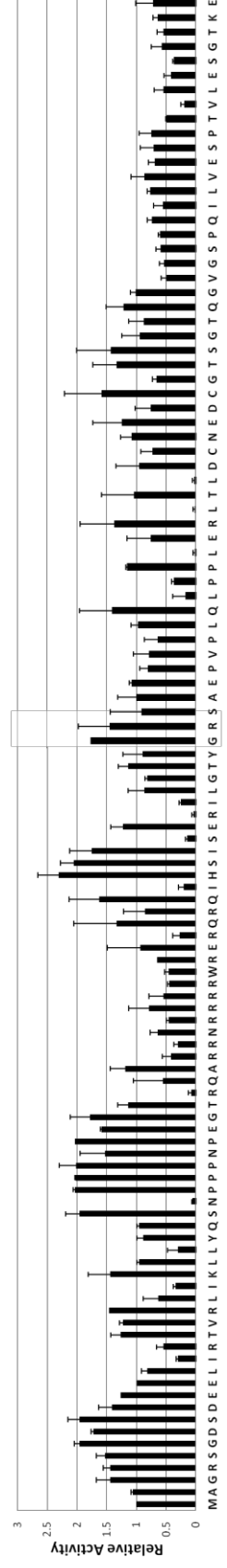
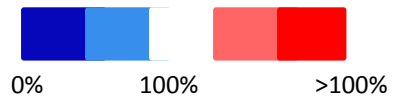
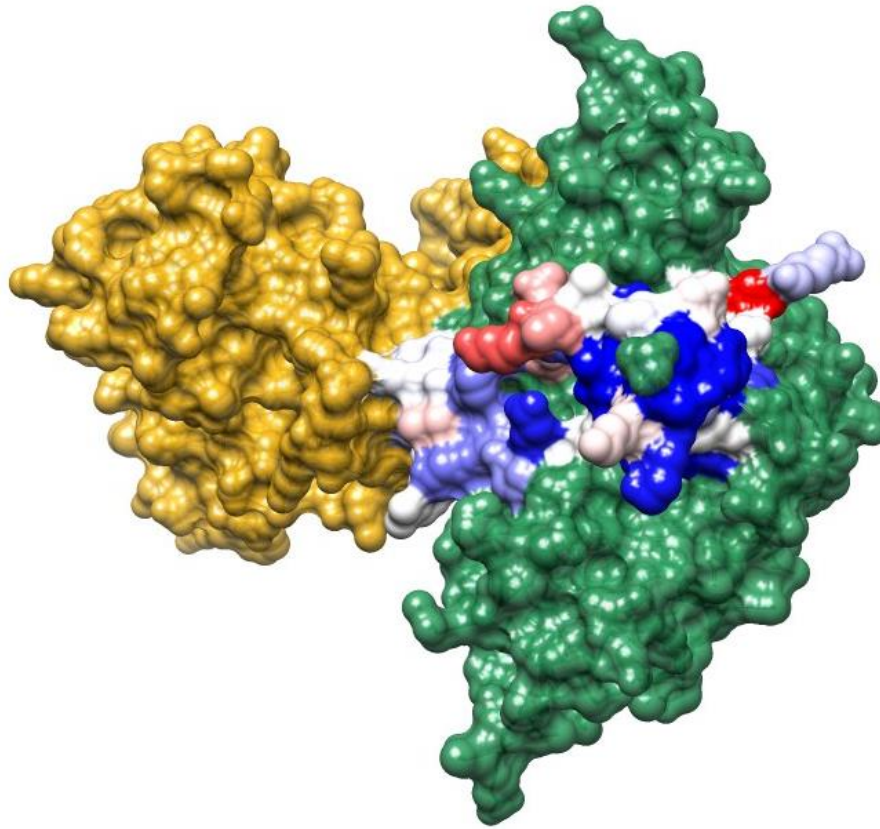
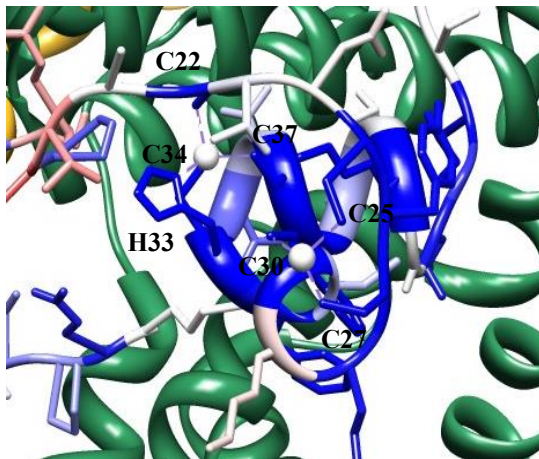


Figure 3

A



B



C

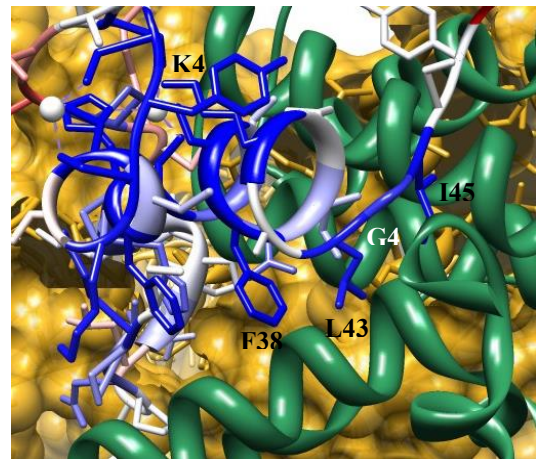
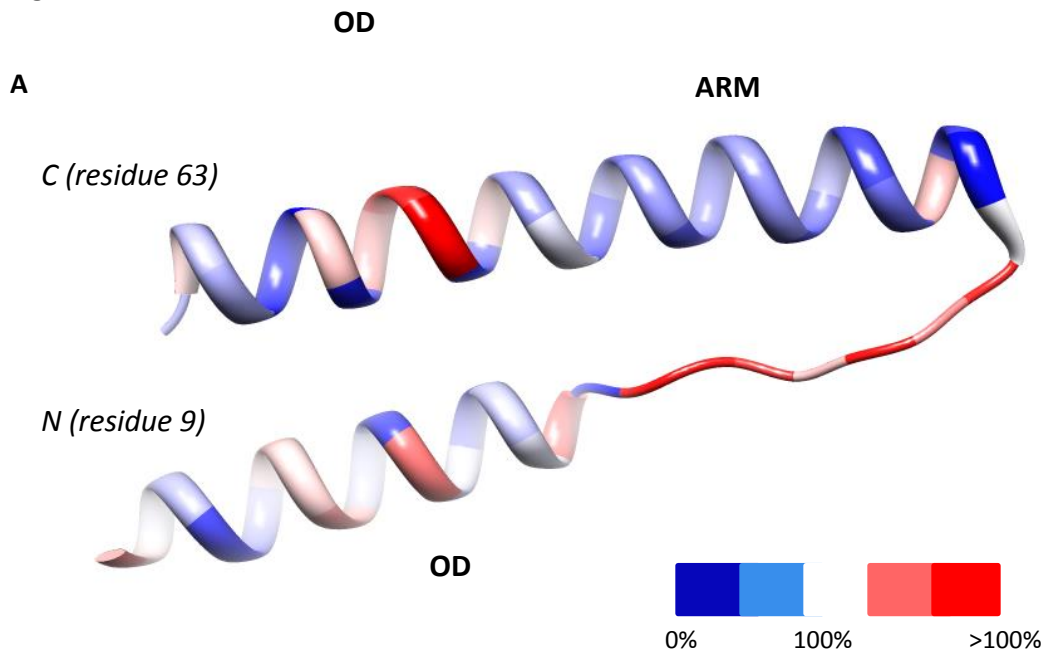


Figure 4



B

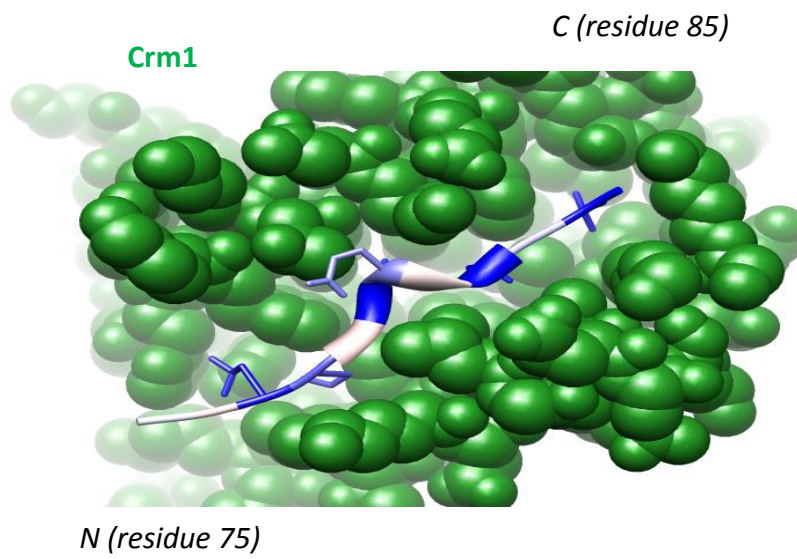
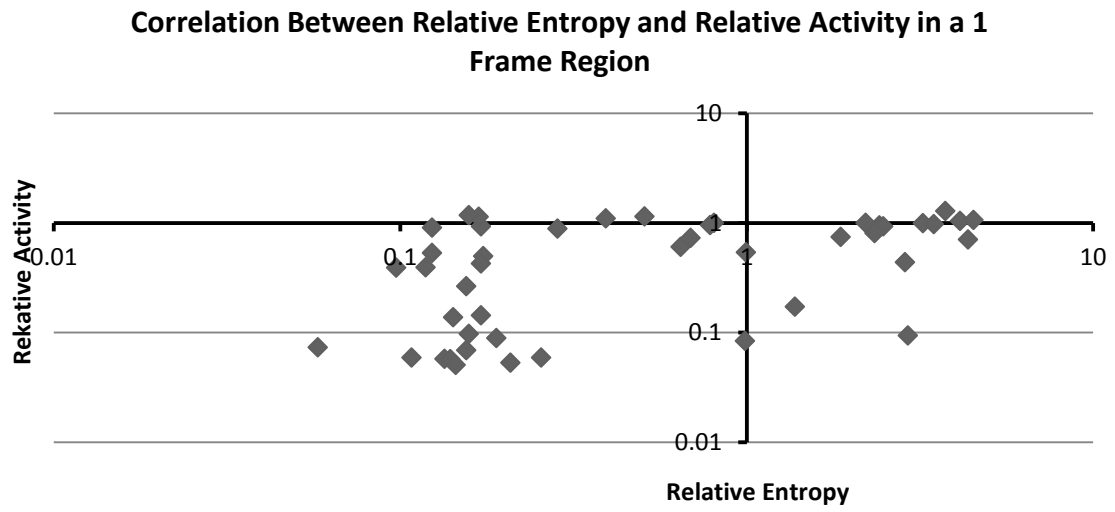
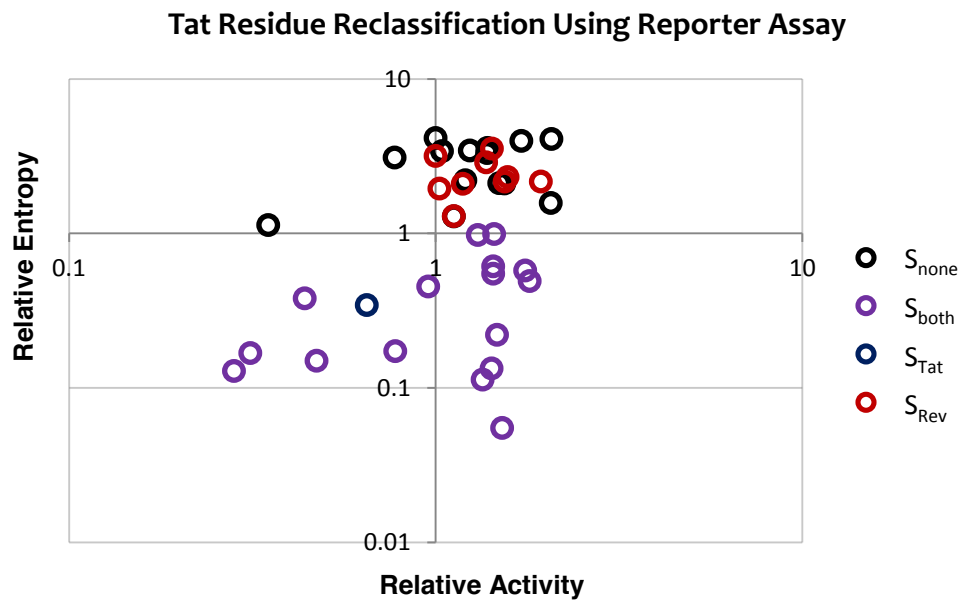


Figure 5

A



B



c

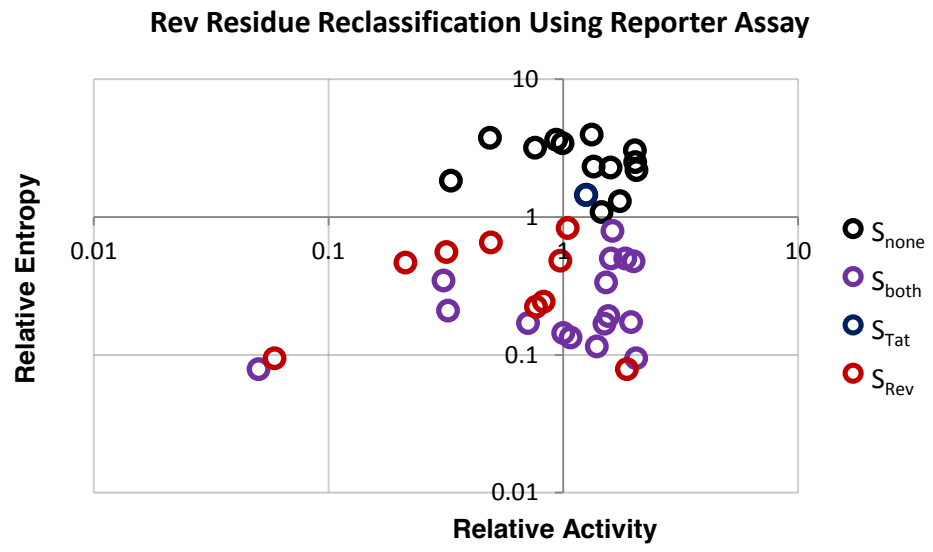
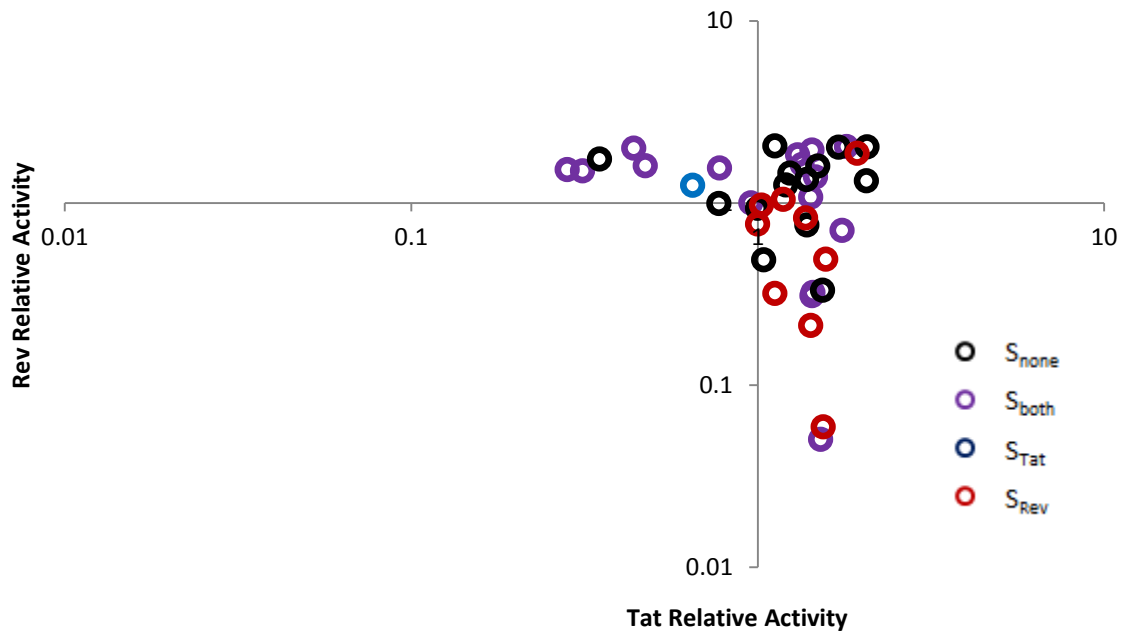


Figure 6



Chapter 4: Development of a Decoupled, Directed-Evolution Viral Platform

Introduction

In order to fully test our model of the *tat/rev* overlap, we required an experimental system that could allow the decoupled evolution of both proteins. Furthermore, this system would ideally measure fitness in the biologically meaningful context of viral replication. Unfortunately, HIV-1's complex splicing and low tolerance for direct repeats making "refactoring" strategies, in which one gene is inactivated in its endogenous position and reinserted immediately downstream of the first stop codon, impractical¹. In order to overcome these obstacles we engineered decoupled viruses by inactivating either *rev* or *tat* and moving a synonymous copy of the inactivated gene into the only non-overlapping gene in the HIV-1 genome, *nef*.

To test the robustness of this system, we then performed competition assays between viruses with known fitness in a variety of starting ratios. Lastly we demonstrated the ability of this system, in combination with directed evolution, to find "plastic" and fixed positions that are otherwise masked by the overlap.

Methods

Creation of Uncoupled Viruses

To facilitate repurposing *nef* to express Rev or Tat, HIV_{NL4-3}, an infection molecular clone was modified to introduce Sac II and Xba I restriction sites in place of the *nef* start ATG codon via site-directed mutagenesis. The native *rev* locus was deleted via mutagenesis from ATG to ACG (non-coding in the overlapping Tat frame) and a stop codon was substituted at Rev Y23 (also non coding in *tat*)². The endogenous *tat* locus was deleted

via mutagenesis of the start codon to ACG (silent in VpR) followed by the introduction of tandem stop codons at K12 and S16. To introduce *rev* or *tat* cDNA into *nef* it was necessary to synonymously substitute codons (tat-CS/rev-CS) to void direct repeat deletion during reverse transcription³ (Figure 1).

Creation of Proviral Libraries

Pools of 33 nt primers were synthesized (IDT) containing one completely degenerate codon (NNN) in the middle of the primer. Overlap extension PCR was then performed on plasmids containing either the rev-CS or tat-CS as template. These PCR products were then cloned into the appropriate knockout molecular clone in the SacII/XbaI cloning site. Transformation efficiency was judged by plating a portion of the ligation and a minimum of 250 colonies was used to generate a single proviral library representing randomization of a single residue. Although this methodology was applied to each site in tat-CS and rev-CS, only 3 libraries, Rev N40X, Rev L81X and Rev E47X, were selected for establishing the feasibility of this system to detect uncoupled fitness values. Additionally 4 simpler pools consisting of mutants (L12 (ref), L12A, L12D) with established replication kinetics were generated to test the effect of initial population ratios on fitness. These pools consisted of 9:1 and 1:9 ratios of L12:L12A and 9:1 and 1:9 ratios of L12:L12D.

Viral Selection and Generation of cDNA

Each pool of virus was raised separately and complemented with reference Rev and Tat *in trans* to generate virus in HEK293T cells. Briefly, 60K cells were transiently

transfected with 200 ng of proviral plasmid, and 2 ng each of pcDNA mammalian expression vectors containing *tat* or *rev* using polyethylenimine (PEI) at a 1:5 ratio of the DNA:PEI. Virus laden supernatant was collected 48 hours following transfection, cleared of cells via low speed centrifugation, DNase I treated to remove input plasmids, and stored at -80°C. Virus titers were determined by ELISA and virion-associated RNA was extracted from 100 uL using Qiagen RNeasy kits to determine the diversity of the viral population at inoculation.. Selection assays were initiated via centrifugal inoculation of 1 million SupT1 T-cells with 5ng p24 and 8 ug/mL polybrene for 2 hours at 1200 x g at 32°C in 96-well microtiter plates. Following inoculation, cells were washed twice with PBS to remove input virus and 250 uL media (RPMI-HEPES + 10% fetal calf serum) replaced per well. The infections were then monitored by immunofluorescence assay for cellular HIV antigen synthesis and by supernatant p24 ELISA to quantify progeny virion release. At 7 and 14 days post inoculation, time points correlating with peak infection for reference or defective viral pools, respectively (data not shown), 100 uL of virus-laden supernatant was collected, cleared of cells, and the virion-associated RNA content extracted via RNeasy (Qiagen) and DNase I-treated (Roche). Once all RNA sample were collected, cDNA were generated from 10uL of extracted RNA (correlating roughly to 100 ng of p24) using the BioRad (Hercules, CA) iScript Advanced cDNA synthesis kit primed with oligo dT and random hexamers.

Next Generation Sequencing

Amplicon specific primers were used to extract 175-nt fragments containing the randomized codon and add Illumina adapter sequence. Each amplicon specific primer consisted of a pool of 1-4 random nucleotides appended immediately 5' of the region of interest in order to increase basecall diversity during the NGS runs. A second step PCR added barcode sequences (courtesy J. DeRisi) and the remainder of the adapter. DNA from each selection experiment was quantified using qPCR (Kapa Biosciences) and then pooled in equal amounts and gel extracted to create the NGS library. Library quantification was then performed using qPCR (Kapa Biosciences) and then sequenced using PE150s on either MiSeq (Illumina). On average each run produced 10 million reads. Samples representing proviral DNA, pre-selection viral RNA, and post-selection RNA from two different time points were all sequenced.

Analysis of NGS Data

Barcodes were used to identify the random codon under selection and frequencies of each codon were calculated pre and post selection. Simple ratios of enrichment of post-selection: pre-selection were directly computed. Additionally a logarithmic relative frequency score relative to reference were calculated in a manner similar to McLaughlin et al.⁴:

$$\Delta E_i^x = \log \left[\frac{f_i^{x,sel}}{f_i^{x,unsel}} \right] - \log \left[\frac{f_i^{ref,sel}}{f_i^{ref,unsel}} \right]$$

where the relative frequency of an amino acid x at position i is greater than 0 if that amino acid is more fit than reference and less than 0 if it is less fit. The average value of a position i was calculated across all amino acids x .

Results and Discussion

Creation of Uncoupled Viruses

Engineering a virus that replicates with *tat* and *rev* unconstrained one another would allow us to convincingly demonstrate our conclusion that there are no compromising positions in the *tat/rev* overlap. To create this virus we decided to move one of the genes to the *nef* locus. This position in the genome has several advantages. First, the 5' end of *nef* has no coding or non-coding overlap (although the 3' UTR which overlaps with the 3' end of *nef* should not be disturbed). Second, *nef* is dispensable in cell culture, as Δ *nef* viruses replicate at rate near equivalent to reference (Figure 3). Indeed reporter genes have a long history of insertion into *nef*⁵. Finally, *nef* is positioned in the genome such that splicing patterns that give rise to *tat* and *rev* also produce *nef*; therefore the timing and expression of our engineered genes should be reasonably similar to reference. Indeed this approach has already been utilized in mutational studies of Tat⁶. However, simple insertion of reference sequence into *nef* produces direct repeats that lead to non-productive template switching during reverse transcription³. Therefore we performed synonymous substitutions to maximize nucleotide divergence from the reference sequence, thus creating create codon-swapped versions of *tat* (*tat*-CS) and *rev* (*rev*-CS). The engineering of these genes into

nef resulted in the creation of decoupled viruses we termed TxR(rev-CS in *nef*) and (tat-CS in *nef*) xRT (reference = TRx).

Establishment of an Experimental Workflow to Measure the Fitness of Codons in a Non-overlapped Context

As our goal was to create an experimental workflow (Figure 3) that could be used to measure codon fitness using libraries consisting of a mixture of point mutations of Tat and Rev we performed careful measurements of the replication kinetics of our decoupled reference viruses in high throughput formats. We found that replication continued steadily before peaking approximately at 10 days (Figure 4). We decided to harvest viruses at an early time point where our directed evolution should be the dominant force of evolution (as opposed to RT-mediated mutation and recombination) as well as a later time point in case some of the viruses had very few viable codons.

Creation of Directed Evolution Libraries

We then sought create randomized libraries that could be used to measure the experimental fitness of individual residues of Tat and Rev in an uncoupled context. Therefore, we created three separate libraries (Rev N40X, Rev E47X, Rev L81X), based upon the TxR virus, with the indicated codon mutated to NNN. Each library contained consisted of mixture of all 64 codons at these positions. These residues were chosen for their known functional requirements, activity in the reporter assay, and entropy signatures. N40X is a highly conserved residue which makes a specific contact to the high affinity binding site on the RRE and is therefore necessary in Rev-RRE recognition.

However despite this requirement, the alanine mutation has intermediate activity in our reporter assay. L81 was chosen for its low entropy signature and complete loss of activity in the reporter assay. Finally E47 was chosen for its relatively poor conservation and reference activity in the alanine scan.

Sequencing of Proviral Libraries

Each of our three directed evolution libraries (N40X, E47X, L81X) was sequenced on a MiSeq to determine the diversity and biases of the library. All libraries produced reads for every codon although N40X exhibited some sequence specific bias and all libraries appeared to have the reference rev-CS codon over-represented (most likely reflecting an artifact of the cloning procedure).

Determination of the Effect of Library Bias on Selection

In order to determine if the large variability in the number of sequences of each codon in our proviral libraries would affect our calculations of fitness, we performed a pilot experiment involving libraries composed of known ratios of known relative fitnesses. For these libraries we used TxR viruses harboring either reference rev-CS, L12A or L12D viruses. These viruses were harvested from individual clones and replicated in isolation. L12A has replication kinetics roughly equal to reference while L12D is much less fit (Figure 6). We performed four separate selection experiments; in two experiments L12:L12A was competed at 1:9 and 9:1 ratios and in the other two experiments the ratios were repeated but with L12:L12D. We reasoned that for the L:A comparisons the virus should maintain their input ratios, while the L:D comparisons should give us an

idea of the sensitivity of our assay. For each experiment, we computed a relative fitness score in the manner of McLaughlin et al. ⁴. As expected, the L:A experiments produced similar results (a score of 0 indicates fitness equivalent to wildtype) regardless of the time point or input ratio. The L:D experiments successfully recapitulated the L12D genotype as drastically less fit, with the time point and input ratio only effecting the final score by a maximum of 15%. Together these results gave us confidence that our assay would be sensitive to fitness disparities even in the face of an unequally distributed starting population.

Selection of Three Uncoupled Rev Sites

We then conducted the complete experimental workflow for the three previously described libraries (Rev N40X, Rev E47X, Rev L81X). Viruses were harvested at 7 and 12 days and sequenced according to the workflow (Figure 7). For all viruses there was negligible difference between the proviral plasmids and the viral input, indicating that no selection was occurring in the packaging line. Although unsurprising, this result was necessary to ensure that our results were not skewed by factors such as dominant negative phenotypes that could change our initial library composition. Furthermore, there was little difference between the t_7 and t_{12} timepoints indicating that the population had already begun to fix by day 7.

N40X Mirrors the Requirement of Specific RNA Binding

The N40X library displayed a strong selection for asparagine, with over 60% of the reads containing one of the two asparagine codons by the end of the experiment (up from an

input of 3.6%). A previously determined NMR structure of the Rev ARM in complex with the high affinity binding site of the RRE revealed that N40 is critical in determining specificity via a contact with a G:A base-pair⁷. Although this makes the selection for asparagine unsurprising, it is in stark contrast with the 70% activity of the N40A mutant in the reporter assay. The difference reflected between the two assays may reflect differences between levels of expression in each context; however even taking this difference into account, a 30% loss of activity is likely to be a major disadvantage in a competition assay. In fact, by the end of the experiment the four alanine codons make up roughly 8% of the population indicating that, though those viruses may be replication competent, they are at a significant selective disadvantage. Curiously, the replication data closely mirrors the patient data with a strong preference of asparagine and a slight tolerance of arginine. This is in keeping with our previous conclusion that this position in the genome is dedicated mostly to functional constraints from Rev.

L81X Demonstrates the Structure-Function Requirements of Rev

Rev residue L81 is a critical binding residue in the Rev NES which directly contacts the host export factor Crm1⁸. Disruption of this interface prevents nuclear export as evidenced by our own reporter data as well as thorough mutational and structural studies⁹⁻¹¹. From a structure-function perspective, Rev requires a bulky hydrophobic amino acid to correctly form an interface with Crm1. Our selection experiment is consistent with this structure-based understanding of L81's function, as all 6 leucine, and 2 of the 3 isoleucine codons, show signs of positive selection. In patient sequences,

Rev strongly favors the CTN leucine codons, perhaps reflecting the difficulty of the virus to take multiple steps through sequence space to reach the other codons which display fitness in our randomized library.

E47 is a Plastic Spot on Rev

While N40 and L81 demonstrate strong purifying selection due to Rev function, E47 appears to a relatively plastic position. In patient isolates, it commonly manifests as one of two amino acids with very different chemical properties: glutamic acid and alanine. Additionally, the E47A mutant has essentially reference activity in our reporter assay. Despite the seeming plasticity of this position, the residue rarely samples other amino acids, possibly because of constraints imposed by overlapping proteins. We expected that our library should freely sample a large number of residues once removed from this constraint. Indeed, the results of our experiment indicate that E47X does not display strong positive selection for any particular codon. However, several codons, such as the four proline codons and the three stop codons, display negative selection. These detrimental mutations are consistent with our understanding of Rev: a premature stop would eliminate Rev's essential NES, and a proline would likely disrupt the helical structure of Rev's ARM in which E47 is nestled. Thus, E47 demonstrates the existence of extremely plastic sites on Rev, yet whose degree of plasticity is potentially masked by alternative frames.

Summary

This chapter established a platform to examine the fitness of each codon at a given position in either tat or rev in an uncoupled context. We demonstrated with three sample libraries that, once removed from constraints from another protein, certain positions still require a particular amino acid, others may sample a larger group of chemically similar amino acids that they would not otherwise, and finally some may sample a large variety of chemically dissimilar amino acids.

References

1. Chan, L. Y., Kosuri, S. & Endy, D. Refactoring bacteriophage T7. *Molecular Systems Biology* **1**, 2005.0018 (2005).
2. Zolotukhin, A. S., Valentin, A., Pavlakis, G. N. & Felber, B. K. Continuous propagation of RRE(-) and Rev(-)RRE(-) human immunodeficiency virus type 1 molecular clones containing a cis-acting element of simian retrovirus type 1 in human peripheral blood lymphocytes. *Journal of Virology* **68**, 7944–7952 (1994).
3. An, W. & Telesnitsky, A. Effects of Varying Sequence Similarity on the Frequency of Repeat Deletion during Reverse Transcription of a Human Immunodeficiency Virus Type 1 Vector. *Journal of Virology* **76**, 7897–7902 (2002).
4. McLaughlin, R. N., Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R. The spatial architecture of protein function and adaptation. *Nature* **491**, 138–42 (2012).
5. Page, K. A., Liegler, T. & Feinberg, M. B. Use of a green fluorescent protein as a marker for human immunodeficiency virus type 1 infection. *Biochemical and Biophysical Research Communications* **233**, 288–292 (1997).
6. Neuveut, C. & Jeang, K. T. Recombinant human immunodeficiency virus type 1 genomes with tat unconstrained by overlapping reading frames reveal residues in Tat important for replication in tissue culture. *Journal of Virology* **70**, 5572–5581 (1996).
7. Battiste, J. L. *et al.* Alpha Helix-RNA Major Groove Recognition in an HIV-1 Rev Peptide-RRE RNA Complex. *Science* **273**, 1547–1551 (1996).
8. Güttler, T. *et al.* NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1. *Nature Structural & Molecular Biology* **17**, 1367–1376 (2010).
9. Jackson, L. K. *et al.* Structural basis for leucine-rich nuclear export signal recognition by CRM1. *Nature* (2009). doi:10.1038/nature07975
10. Dong, X., Biswas, A. & Chook, Y. M. Structural basis for assembly and disassembly of the CRM1 nuclear export complex. *Nature structural & molecular biology* 1–3 (2009). doi:10.1038/nsmb.1586
11. Malim, M. H., Böhnlein, S., Hauber, J. & Cullen, B. R. Functional dissection of the HIV-1 Rev trans-activator--derivation of a trans-dominant repressor of Rev function. *Cell* **58**, 205–214 (1989).

Figure Descriptions

Figure 1: Scheme for the design of uncoupled viruses. A cloning site was introduced into the 5' end of *nef* (which was inactivated through the removal of its start codon) to create the reference virus TRx. Then codon-swapped (Tat-CS or Rev-CS) versions of either *tat* or *rev* were cloned into the locus to create xRT and TxR viruses. To create libraries versions of Rev CS randomized at a particular position of interest were cloned into the TxR backbone.

Figure 2: Flowchart for Uncoupled Evolution Experiments. Libraries harboring the codon swapped gene with a single randomized amino acid are cloned into the provirus. Viral particles are then generated in a packaging line. The viruses are then used to infect T cells and RNA is harvested at pre-determined time points. The RNA is reverse transcribed into cDNA and prepped for NGS via the addition of adapters and barcodes.

Figure 3: Replication Rates of Uncoupled Viruses. Elimination of *nef* and substitution of *rev*-CS confers minimal defects in viral replication.

Figure 4: Time Course for Viral Replication. Replication kinetics in a 96 well dish demonstrate that infection continues to increase before peaking around day 9.

Figure 5: Diversity of Proviral Libraries. Distributions of the three directed evolution libraries pre-selection. Each library appears to have a slight overrepresentation of the wild-type codon and a large degree of variability between the most and least frequent codon. N40 appears to have a particularly biased distribution with CNN codons

underrepresented. Despite this, sequence coverage is able to accurately estimate pre- and post-selection levels of each codon.

Figure 6: Selection of Viruses with Known Relative Fitnesses. Mutants of Rev L12 (A and D) were replicated in mixtures of 1:9 or 9:1 (mutant:reference). A) Replication growth curves of ref, L12A and L12D. L12 is as fit as A12 and much more fit than D. B) Results of the viral competition experiment at two different time points. Ratios correspond to the number of reads for each codon from harvested virus at either day 7 or day 12. L12A and reference maintain their input ratios while L12D experiences negative selection. Frequency scores relative to reference are also shown.

Figure 7: Heatmaps of Directed Evolution Libraries. Heat maps showing the fitness landscape for each codon at Rev positions 40, 47 and 81. Each column represents a codon (denoted by its translation product above it) and each row represent a ratio between the number of reads of that codon at different conditions. The first row shows the ratio of proviral DNA/viral input (t_0) while the other two represent ratios between 7 (t_7) and and 12 (t_{12}) days of selection and the viral input (t_0).

Figure 1

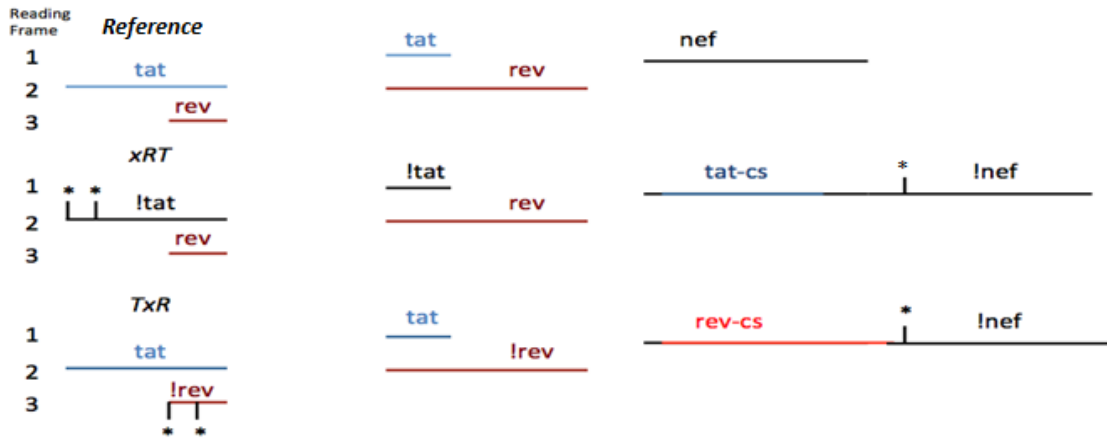


Figure 2

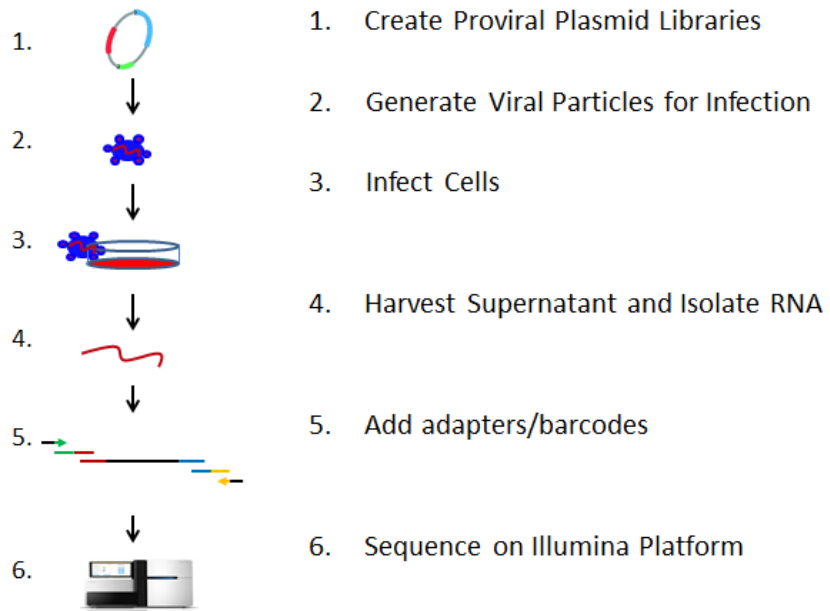


Figure 3

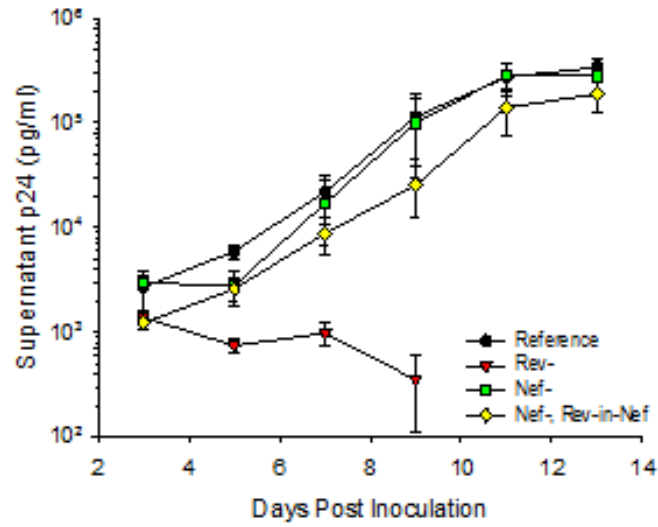


Figure 4

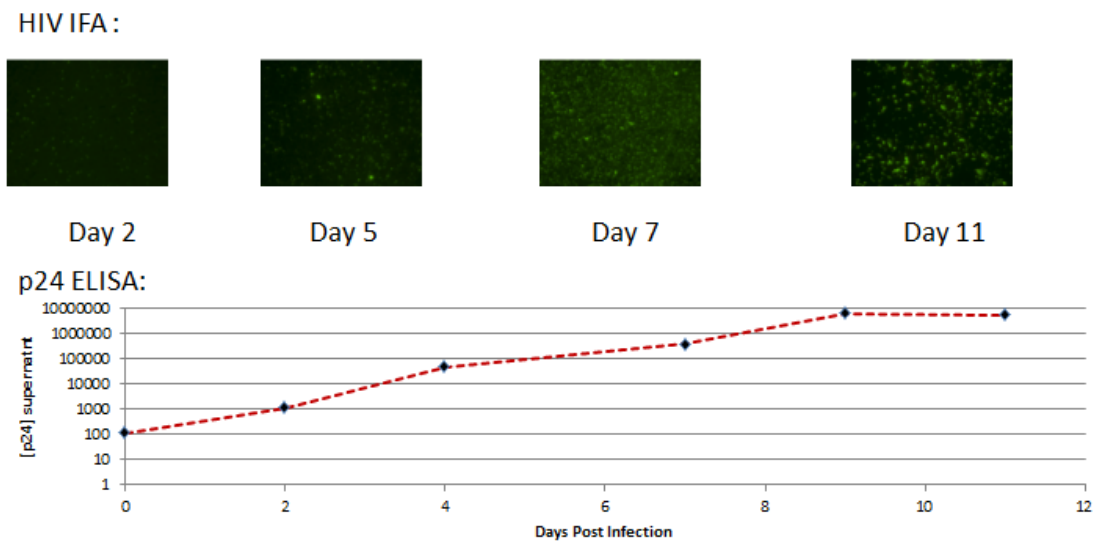
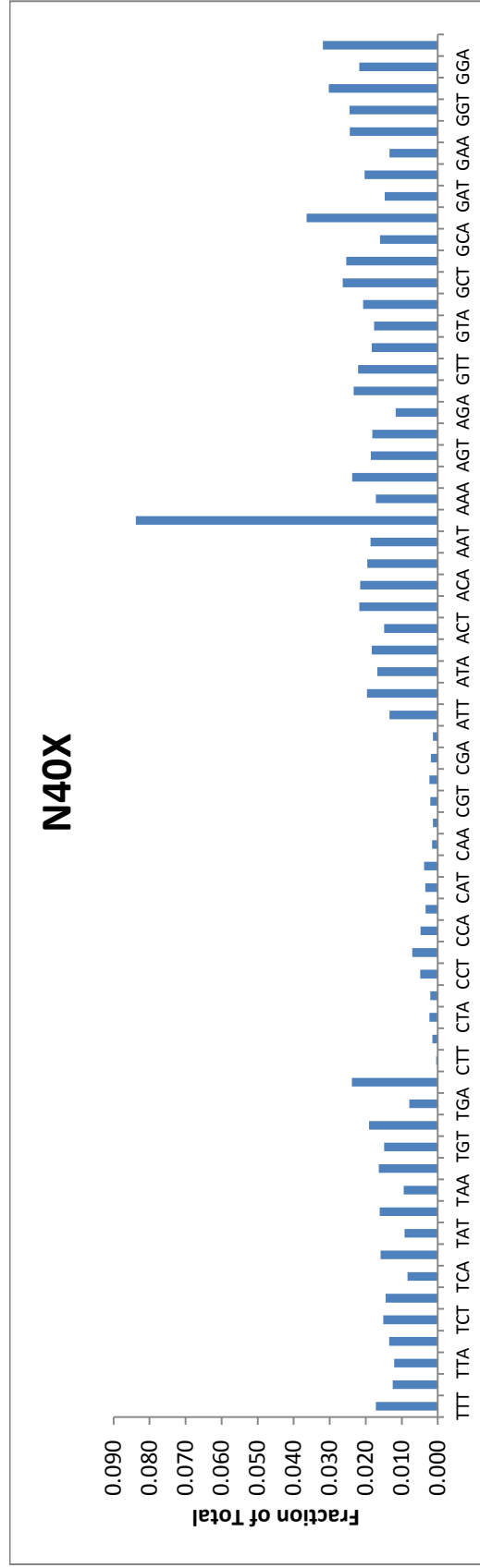
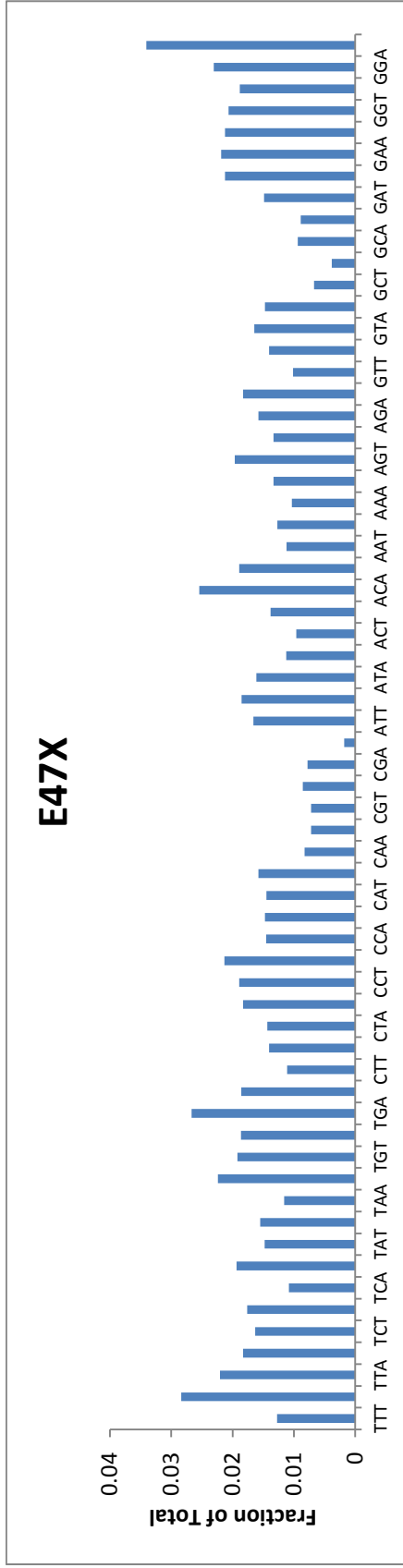


Figure 5



L81X

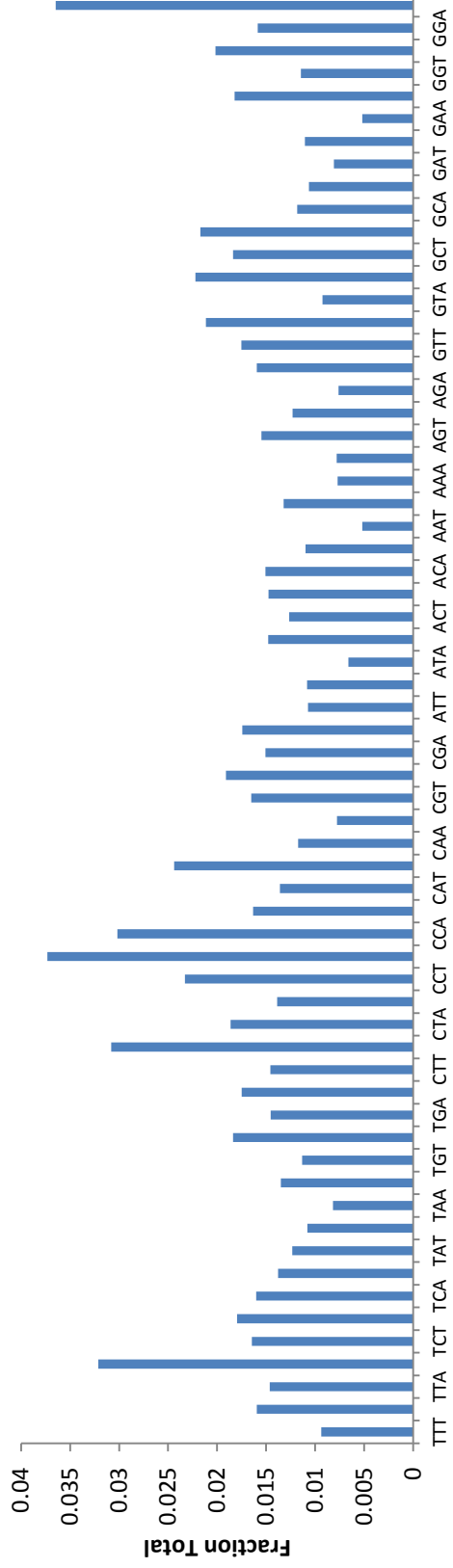
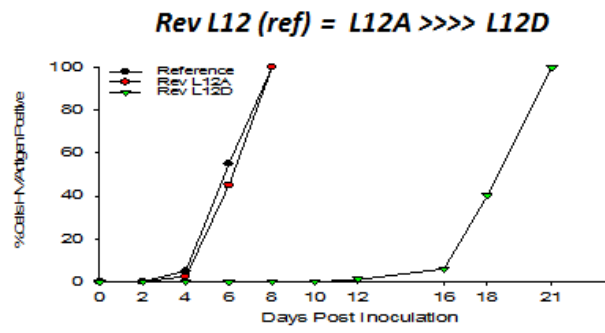


Figure 6

A



B

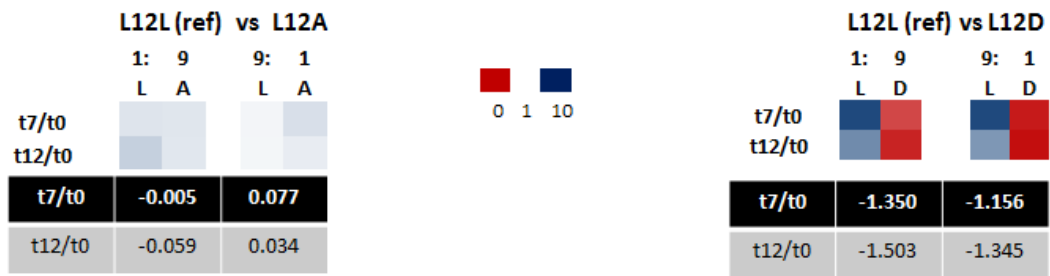


Figure 7

A



Alanine mutant has 70% activity in reporter assay

B

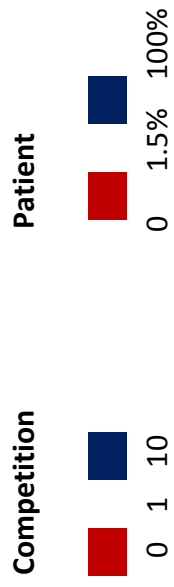


Alanine mutant has 0% activity in reporter assay

C



Alanine mutant has 100% activity in reporter assay



Chapter 5:

Conclusions, Impact and Future Directions

Introduction

We carried out analyses on two different datasets and established a platform to build a third to determine if the overlapping coding regions of *tat* and *rev* contained regions that were under selective constraints between both proteins. Together, these data are the strongest evidence to date that the *tat/rev* overlap evolves each protein almost independently. In this chapter we discuss the implications of this phenomena, its application to other systems, and future experiments.

Discussion

Comparison to Existing Models

In Chapter 1, we described several simple models of overlap evolution: the genome size model, the genomic novelty model, and the post-transcriptional regulation model. We will now revisit these models and discuss them in light of the results of this work.

*Genome Length Considerations do Not Adequately Explain the *tat/rev* Overlap*

As mentioned previously, the correlations between overlap proportion, polymerase fidelity, capsid size and genome length remain controversial¹⁻³. For HIV-1, capsid size seems to have a weak influence on genome size as the virus is capable of packaging genomes almost double its normal size⁴. Additionally, the decoupled viruses created in Chapter 4 increase the genome by the size of the overlap with no significant loss in fitness. Finally, the existence of overlaps in mammals, where the genomic savings of an overlap constitute an insignificant percentage of the entire genome, indicates that

genome length is not the sole reason for the existence of overlaps⁵. The relationship between genome size and the presence of overlap should not be ignored given the preponderance of overlapping genes in small viral genomes²; however it does not appear to be the dominant pressure appears in the *tat/rev* overlap.

Genomic Novelty and Intrinsically Disordered Proteins

Multiple groups have noted that overlapped proteins have unique structural properties^{6,7}. For instance, overlapped proteins have a propensity for intrinsic disorder which may make the proteins highly adaptable for multiple functions. Indeed Tat and Rev appear to be highly tolerant to mutation when compared to similar mutational scans of well-structured viral proteins such as capsid⁸. Not only does this structural arrangement allow robustness in the face of mutation, but it also promotes the use of short linear motifs which are highly adaptable^{5,9}. Furthermore the intrinsically disordered backbone can confer advantages, via aggregation and interactions with host proteins, in evading cellular surveillance mechanisms such as the immune system and the proteasome^{5,6,10-12}.

*Post-transcriptional Regulation Remains a Possible But Poorly Defined Justification for the *tat/rev* Overlap*

Beyond these inherent structural properties, overlaps also carry a potential benefit in post-transcriptional regulation. This behavior is seen in the influenza A virus, all retroviruses, and several human genes^{5,13-15}. Given that *tat* and *rev* have linked functions in the production and regulation of the viral RNA, it is entirely possible that

the genetic arrangement of these genes, which is largely conserved across the lentiviruses, helps encode important regulatory control mechanisms for viral replication^{16,17}. Indeed there is evidence that over-expression of Rev in a viral context can inhibit viral replication, potentially through feedback on Tat mRNA production (Figure 1). However the ability of the TxR virus to replicate indicates that this circuit is sufficiently robust to biologically relevant changes in timing and expression; it remains to be seen if Rev-mediated feedback on Tat is a significant selection pressure in either of these protein's evolution.

Overlaps May Protect Against Non-Productive Mutations

One way overlaps could confer a selective advantage is via protection from non-productive or lethal mutations. Surveys of point mutations in viruses have found that the average mutation has a negative fitness cost¹⁸. An overlap restricts the available mutational subspace, even in a selfish arrangement. For instance strict functional requirements from one frame may prevent the other frame from entering a non-productive state (Figure 2A). In this way, even though the overlap restricts the actual mutational subspace, it disproportionately decreases the occupancy of the non-productive subspace and thus mitigates any decrease in viral diversity.

Overlaps as an Alternative to Gene Duplication

In large viruses or higher order organisms, genomic novelty in response to a selection pressure can often evolve via gene duplication, adaptation and genome compression¹⁹. An alternative mechanism, possibly preferentially deployed in viruses

with small genomes where gene duplication may not be feasible, can occur through *de novo* gene creation via an overlapped reading frame. In the process of adaptation, the genes segregate their functional domains, thus avoiding a potential Achilles' heel and circumventing the need for genome compression (Figure 2B). In this model overlaps arise as adaptive changes, not as penalties in response to other pressures. Instead they are optimal solutions that take advantage of small, highly-adaptable functional domains templated onto an intrinsically disordered scaffold while working in the constraints imposed by the genome.

Therapeutic Implications of Overlaps

Although selfish organization is an elegant outcome to minimizing functional restraint, this organization appears to dampen the prospects for therapeutically targeting an overlap. However other studies demonstrate that, although evolution has sought to minimize the penalty by occupying a selfish portion of sequence space, mutational escape is still constrained. Rhesus monkeys infected with SIV can generate a T-cell epitope that targets the region of Tat that overlaps with Vpr. Production of this epitope results in nonsynonymous mutations to Tat, yet the virus mutates synonymously with respect to Vpr²⁰. In essence this creates an artificial "compromising" position via selection pressure by the host. A combination of epitopes might be successful in such cases as the overlap, by definition, decreases the viral mutational space. Indeed HBV, which has an astounding 50% gene overlap, appears to take pre-emptive steps to avoid

mutational trapping by the immune system by preferentially avoiding potential CTL epitopes in overlapped areas²¹.

Interestingly studies analyzing areas of overlap with non-coding elements show existing constrained variability and evolution^{22,23}. The discrepancy between coding/coding and coding/non-coding overlap may reflect different constraints between the offset genetic code and base-pairing considerations. Given that the aforementioned SIV experiments demonstrate that coding overlaps have some mutational constraint (even in a selfish arrangement), coding/non-coding overlaps may be additional ripe targets for therapy.

Mildly Selfish Sites

Although we have proposed two discrete models, a selfish model and a compromising model, these two models truly represent extremes upon a continuum. For instance, our analysis revealed several sites that were strongly conserved in a single characteristic rather than a single amino acid. This was most drastically seen in our mutation of Rev L81, but other residues such as Rev L21 transition frequently through both leucine & isoleucine codons. Even a seemingly plastic residue like Rev E47, which does not appear to possess a particular biochemical characteristic, has several codons which it actively selects against (i.e. proline and stop codons). Thus although it appears that each individual nucleotides in the overlap has been parceled out for a single protein, there are likely weak constraints that constitute weak compromises between the two proteins. Such behavior would most clearly be seen from a comprehensive survey of all sites in the overlap in the decoupled system.

Optimization of the Genetic Code

Although Francis Crick famously called the genetic code “a frozen accident”²⁴, several groups have argued that the genetic code is in fact optimal in several considerations, including the creation of new genes^{25,26}. Even amongst overlaps biases, certain biases are clear with +1 frameshifts being the most common manifestation of overlaps and reverse complement overlaps rarely seen². Although we attempted cursory investigation into this phenomena we were unable to discover any specific pattern except to note that arginine-rich proteins such as Rev and Tat are ideal for overlaps as these codons are six fold degenerate and allow preservation of particular characteristics in different frameshifts (Appendix C).

Independent Evolution of Reading Frames in Other Viruses

Although this study is the first to examine independent evolution of overlaps on a residue level, several other studies, performing gene wide dN/dS calculations have found observed the trend in which one gene displays a high mutation rate in response to conservation in another²⁷. This behavior has been seen in potato leafroll virus, HBV, SIV and some papillomaviruses^{20,28,29}. This suggests that selfish organization may be a general outcome of having two overlapped frames.

Complete Survey of All Residues

Chapter 4 clearly demonstrates the feasibility of directed evolution and selection for specific sites in a decoupled viruses. This method is generalizable to all sites in either

protein and comprehensive mutation of each site would convincingly show the presence or absence of any compromising sites. Furthermore analysis of these sites would demonstrate if, in avoiding an Achilles' heel, either protein has made functional sacrifices that can be optimized in a non-overlapped context. A complete dataset of all sites would further allow critical analysis of the "protective effect" model of overlaps as, with the fitness of all codons at every position, direct comparisons could be made between the fitnesses permitted and prevented by an overlapped architecture.

Summary

Taken together these data demonstrate a minimal negative impact on viral fitness imposed by the overlap. This is in stark contrast with the traditional dogma that overlaps are a large fitness cost driven by other considerations such as limitations on genome size¹⁻³. Indeed, the usual corollary of this model, that constraints from both frames should drive high conservation, appears to be untrue for HIV-1 as overlaps have a higher sequence entropy than non-overlapped areas. Rather it appears the existence of overlaps is the result of a confluence of both positive and negative selection pressures including the need for genomic novelty, restrictions on genome length, and possibly, though it lies outside the scope of this work, post-transcriptional control.

It is likely that the trends observed here are not an isolated case, but rather that this unilateral division of genetic information between overlapping genes is a common and clever evolutionary outcome that satisfies this mix of positive and negative pressures. Immune targeting and evolutionary analysis of overlaps suggests this modular

organization exists in several viruses such as HCV and *Paramyxovirinae* although their overlaps have yet to be completely and systematically dissected^{7,20,21,27}. The *tat/rev* overlap then is a logical evolutionary outcome, where each gene has selfishly claimed individual nucleotides for itself and thus incurred little to no functional penalty while still reaping the evolutionary rewards of an overlap. Indeed, overlaps in all viruses merit further analysis based on their impact in the context of viral evolution as a whole, rather than just dismissal as a disadvantageous consequence of other selection pressures.

References

1. Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral Mutation Rates. *Journal of Virology* **84**, 9733–9748 (2010).
2. Belshaw, R., Pybus, O. G. & Rambaut, A. The evolution of genome compression and genomic novelty in RNA viruses. *Genome Research* **17**, 1496–1504 (2007).
3. Chirico, N., Vianelli, A. & Belshaw, R. Why genes overlap in viruses. *Proceedings of the Royal Society B: Biological Sciences* **277**, 3809–3817 (2010).
4. Kumar, M., Keller, B., Makalou, N. & Sutton, R. E. Systematic determination of the packaging limit of lentiviral vectors. *Human Gene Therapy* **12**, 1893–1905 (2001).
5. Kovacs, E., Tompa, P., Liliom, K. & Kalmar, L. Dual coding in alternative reading frames correlates with intrinsic protein disorder. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 5429–5434 (2010).
6. Rancurel, C., Khosravi, M., Dunker, A. K., Romero, P. R. & Karlin, D. Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. *Journal of Virology* **83**, 10719–10736 (2009).
7. Karlin, D. Structural disorder and modular organization in Paramyxovirinae N and P. *Journal of General Virology* **84**, 3239–3252 (2003).
8. Rihn, S. J. *et al.* Extreme genetic fragility of the HIV-1 capsid. *PLoS pathogens* **9**, e1003461 (2013).
9. Neduva, V. & Russell, R. B. Linear motifs: evolutionary interaction switches. *FEBS Letters* **579**, 3342–3345 (2005).
10. Tompa, P., Szász, C. & Buday, L. Structural disorder throws new light on moonlighting. *Trends in Biochemical Sciences* **30**, 484–489 (2005).
11. Xue, B., Mizianty, M. J., Kurgan, L. & Uversky, V. N. Protein intrinsic disorder as a flexible armor and a weapon of HIV-1. *Cellular and molecular life sciences CMLS* **69**, 1211–59 (2011).
12. Tompa, P., Prilusky, J., Silman, I. & Sussman, J. L. Structural disorder serves as a weak signal for intracellular protein degradation. *Proteins* **71**, 903–909 (2008).

13. Chua, M. a, Schmid, S., Perez, J. T., Langlois, R. a & Tenover, B. R. Influenza a virus utilizes suboptimal splicing to coordinate the timing of infection. *Cell reports* **3**, 23–9 (2013).
14. Shehu-Xhilaga, M., Crowe, S. M. & Mak, J. Maintenance of the Gag/Gag-Pol Ratio Is Important for Human Immunodeficiency Virus Type 1 RNA Dimerization and Viral Infectivity. *Journal of Virology* **75**, 1834–1841 (2001).
15. Nekrutenko, A., Wadhawan, S., Goetting-Minesky, P. & Makova, K. D. Oscillating Evolution of a Mammalian Locus with Overlapping Reading Frames: An XLaS/ALEX Relay. *PLoS Genetics* **1**, e18 (2005).
16. Felber, B. K., Drysdale, C. M. & Pavlakis, G. N. Feedback regulation of human the Rev protein . Feedback Regulation of Human Immunodeficiency Virus Type 1 Expression by the Rev Protein. **64**, (1990).
17. Weinberger, L. S., Dar, R. D. & Simpson, M. L. Transient-mediated fate determination in a transcriptional circuit of HIV. *Nature Genetics* **40**, 466–470 (2008).
18. Sanjuán, R., Moya, A. & Elena, S. F. The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 8396–8401 (2004).
19. Elde, N. C. *et al.* Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. *Cell* **150**, 831–41 (2012).
20. Hughes, A. L., Westover, K., Da Silva, J., O'Connor, D. H. & Watkins, D. I. Simultaneous Positive and Purifying Selection on Overlapping Reading Frames of the tat and vpr Genes of Simian Immunodeficiency Virus. *Journal of Virology* **75**, 7966–7972 (2001).
21. Maman, Y. *et al.* Immune-induced evolutionary selection focused on a single reading frame in overlapping hepatitis B virus proteins. *Journal of Virology* **85**, 4558–4566 (2011).
22. Snoeck, J., Fellay, J., Bartha, I., Douek, D. C. & Telenti, A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. *Retrovirology* **8**, 87 (2011).
23. Sanjuán, R. & Bordería, A. V. Interplay between RNA structure and protein evolution in HIV-1. *Molecular Biology and Evolution* **28**, 1333–1338 (2011).

24. Crick, F. H. The origin of the genetic code. *Journal of Molecular Biology* **38**, 367–379 (1968).
25. Sella, G. & Ardell, D. H. The coevolution of genes and genetic codes: Crick's frozen accident revisited. *Journal of Molecular Evolution* **63**, 297–313 (2006).
26. Itzkovitz, S. & Alon, U. The genetic code is nearly optimal for allowing additional information within protein-coding sequences. *Genome Research* **17**, 405–12 (2007).
27. Zaaijer, H. L., Van Hemert, F. J., Koppelman, M. H. & Lukashov, V. V. Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. *The Journal of general virology* **88**, 2137–2143 (2007).
28. Hughes, A. L., Piontkivska, H., Krebs, K. C., O'Connor, D. H. & Watkins, D. I. Within-host evolution of CD8+TL epitopes encoded by overlapping and non-overlapping reading frames of simian immunodeficiency virus. *Bioinformatics* **21 Suppl 3**, iii39–i44 (2005).
29. Guyader, S. & Ducray, D. G. Sequence analysis of Potato leafroll virus isolates reveals genetic stability, major evolutionary events and differential selection pressure between overlapping reading frame products. *The Journal of general virology* **83**, 1799–1807 (2002).

Figure Descriptions

Figure 1: Inhibition of Viral Replication by Rev Overexpression

A cell line overexpressing Rev inhibits viral replication, possibly by repressing the production of Tat mRNA. This may be part of a larger role for the tat/rev overlap in the post-transcriptional control of a biological network. However this behavior has yet to be shown in a biologically relevant context.

Figure 2: Model for Fitness Advantages of an Overlap

A) “Pipe” network of codon fitness network. Nodes represent codons (with reference codon in the center) and connections represent evolutionary pathways (i.e. point mutations). Dashed connections indicate pathways that are no longer possible in the context of an overlap. Nodes are colored to represent fitness relative to reference: red nodes are less fit, neutral are white, and more fit residues are blue.

B) Potential mechanisms of adaptation in viruses. Large viruses with low mutation rates can accommodate multiple gene duplication events and then evolve functions from that scaffold (red and black boxes) in the face of selection pressure. Eventually only the needed functions are kept (blue, red) and the rest of the duplications are discarded. For small genomes, gene duplication may be infeasible and genomic novelty may arise simply by creating an alternative frame within an existing gene (black box). That gene is adapted and its functional domains isolated in nucleotide sequence space (the “driving apart” of the genes). No subsequent compression is required.

Figure 1

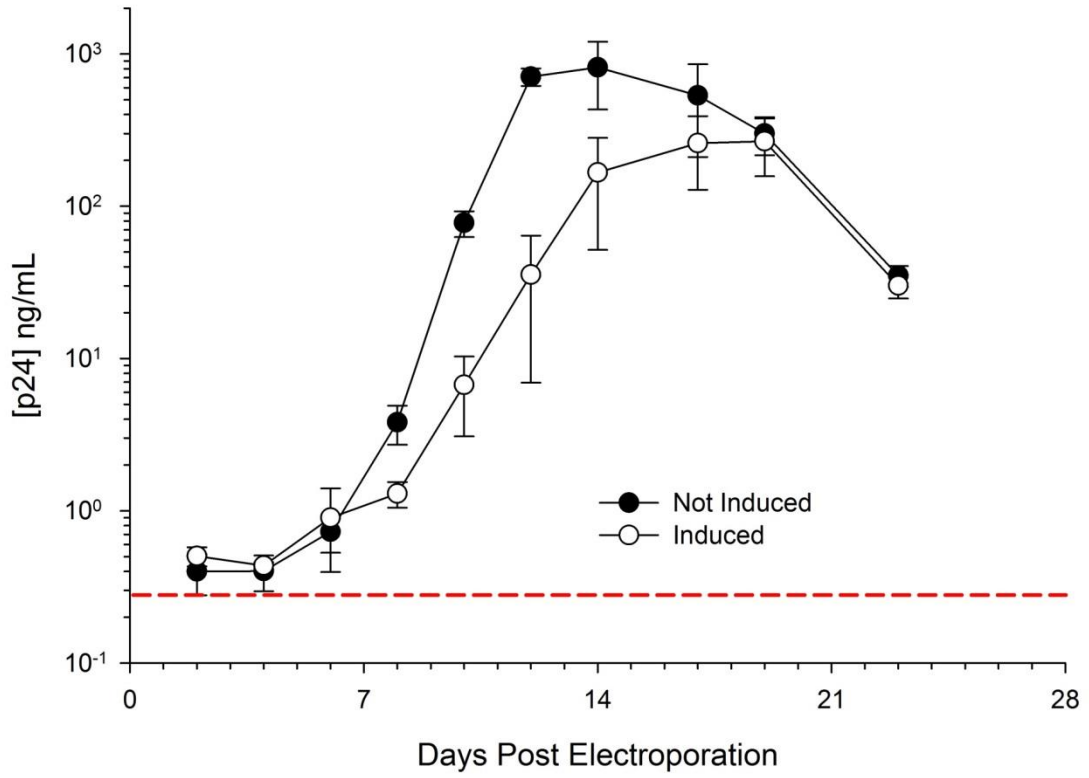
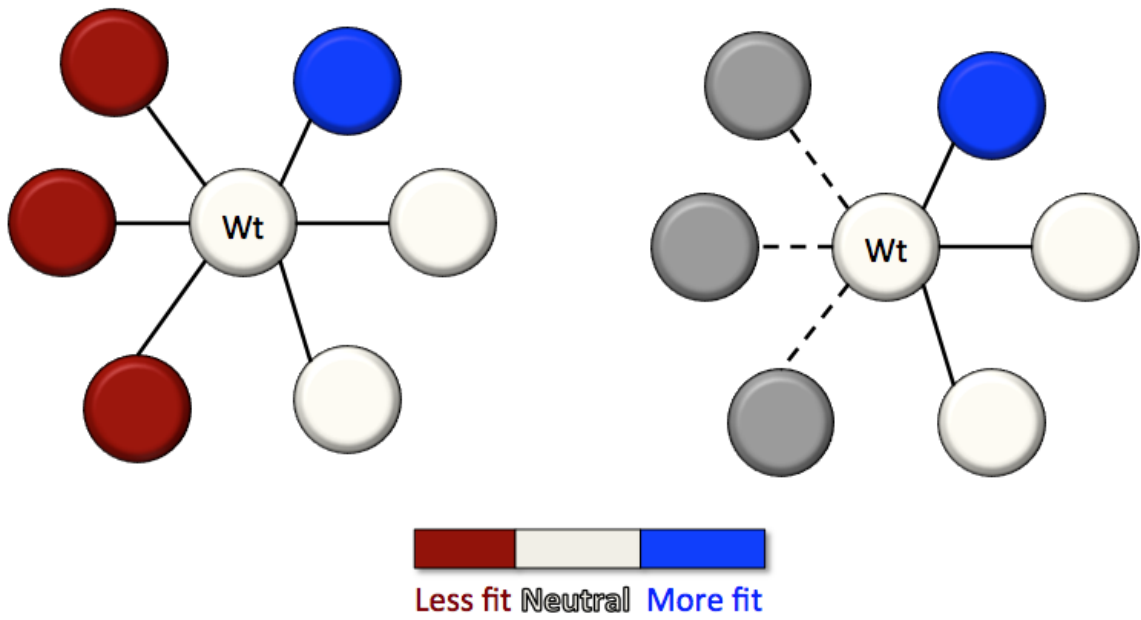
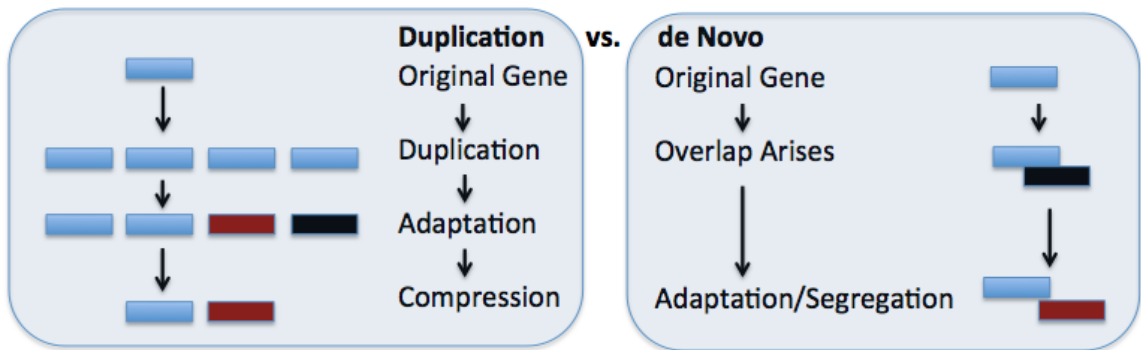


Figure 2

A



B



Chapter 6

Proteomic Characterization of Rev

Introduction

Although evolutionary studies of Tat and Rev provide an interesting case study of the selective forces that must be balanced within an overlapping reading frame, we are also interested in the specific mechanisms of viral protein function. In order to focus our effort, we concentrated on studying the details of Rev function by generating physical interactions maps between Rev and human host proteins¹.

Protein-Protein Interaction Maps Provide a Deeper Understanding of Protein Function

Within the cell, most proteins organize into larger functional complexes through protein-protein interactions (PPI). As most complex biological processes involve at least one multiprotein complexes, discovering and defining PPIs has become essential in achieving a better understanding of basic biological mechanisms²⁻⁵. Affinity purification coupled to mass spectrometry (AP-MS) is one of the most commonly used techniques to generate PPI network data. The AP-MS technique allows identification of complexes in their native context, produces a much lower false positive rate compared to other approaches, and is more amenable to complex identification⁶. AP-MS has also been employed to generate host-virus interactomes with success, and allows purification of protein complexes within the context of infection⁷. Global analyses of AP-MS maps have helped shape our view of protein complexes as the molecular machinery of the cell.

HIV-1 Requires Host Factors for Viral Replication and RNA export

The HIV-1 genome contains only nine genes encoding approximately fifteen proteins⁸. These proteins participate in a myriad of biological processes including, but not limited

to, receptor trafficking, RNA export, transcription, and ubiquitination⁹. Such a wealth of processes and minimal genomic information implies that virus is especially reliant on host factors in order to successfully replicate and spread.

As discussed in previous chapters, Rev has several well characterized interactions with host proteins. Rev contains a hydrophobic NES that recruits the host export factor Crm1 (XPO1) and forms a ternary complex with Ran. Several other proteins have been implicated in Rev function, but their roles remains less well characterized. The dead box helicase DDX3 is thought to attach to Rev-Crm1 complexes and remodel RNA cargoes in a manner that facilitates transit across the nuclear pore¹⁰. Other interactions of note include the recognition of Rev's NLS by B23 (Nucleophosmin) and importin-beta which allow nuclear localization of Rev^{11,12}. In addition to RNA export, alternative Rev functions in translation and splicing have been suggested yet remain controversial^{13,14}. More recent work has shown that RNA nuclear export and encapsidation are coupled processes, raising tantalizing post-transcriptional roles for Rev¹⁵⁻¹⁷.

Packaging and Cellular Trafficking of RNA is Linked to Rev-RRE Nuclear Export

It is possible to alter HIV-1 RNA export to proceed in a Rev independent manner. Two common approaches are the removal of instability elements via codon optimization, or the substitution of RRE for a constitutive transport element (CTE). Either of these modifications allows RNA export to proceed through the default cellular pathway mediated by Nxf1. Interestingly, simply changing the RNA export pathway from Crm1 to Nxf1 allows HIV-1 to overcome a virion assembly defect encountered in murine cells^{17,18}.

The linkage between RNA export and virion assembly is further strengthened by studies which show that viral RNAs exported by Crm1 pathways do not efficiently dimerize or co-package with viral RNAs exported by Nxf1¹⁶. These results suggest that the cytoplasmic fate of an RNA is at least partially influenced by the manner of its exit from the nucleus. A key difference between these two modes of export is the presence of Rev, suggesting that Rev may be a key determinant in downstream functions such as packaging and assembly¹⁵. In order to examine this possibility, we characterized a Rev “interactome” to find candidate host interaction partners that might be involved in these cytoplasmic roles of Rev.

Methods

Creation of an Affinity Tag and Purification System

A dual affinity tag containing a 2xStrep-3xFLAG vector spaced by a glycine-serine linker was synthesized and cloned into pcDNA4-TO to create pcDNA4-TO SF. This vector also contains a TEV cleavage site in between the Strep II (Novagen) and 3xFlag (Sigma). A corresponding N-terminal vector pcDNA4-TO FS as well as single tag versions (pcDNA4-TO Nstrep, pcDNA4-TO NFlag, pcDNA4-TO CFlag and pcDNA4-TO CStrep) were also created. Sequences corresponding to NL43 and HXB2 reference strains of Rev were cloned into this vectors using Afl II and Not I. This affinity tag combination was selected based on tests compared to His-Flag versions as well as literature analysis of common AP tags^{6,19,20}. This system was also used to characterize interactions for all HIV-1 proteins described in a separate work^{3,21}.

Large Scale Affinity Purification of Rev

HEK293 cells were cultured and plated onto four 15 cm dishes per each set of experimental conditions. At 60-75% confluency cells were transfected with 5 ug of pcDNA4-TO- FS Rev per plate. Transfection mixtures were prepared at a 1:5 ug DNA:ul Polyjet (SigmaGen) according to manufacturer's instructions. Cells were grown for 24-48 hours post-transfection.

Cells were washed with PBS, pelleted and resuspended in 1 ml/plate cold lysis buffer (0.5% NP40, 50 mM Tris-HCl pH 7.4; 150 mM NaCl, 1 mM EDTA, protease and phosphatase inhibitors). In order to test the optimal conditions for Rev purification we tested several purification conditions. High salt (500 mM NaCl and low salt 50 mM NaCl conditions were also tested) lysates were prepared as were standard RNAsed lysates (RNase A/T1 was added to both the lysis and wash buffers)²². Finally one set of dishes was used to test a hypotonic fractionation protocol. Lysis of the pellet was first performed with hypotonic lysis buffer (10 mM HEPES 7.9, 1.5 mM MgCl₂, 10 mM KCl 0.5 mM DTT) and dounce homogenization to make the cytoplasmic fraction. Nuclei were then washed with a sucrose cushion before being lysed in standard lysis buffer. All lysis and purification steps took place at 4°C After 30 minutes of incubation, lysates were pelleted to remove the insoluble fraction.

For large scale purifications binding was done on column. Briefly 100 ul of Strep Tactin Sepharose (Novagen) were packed onto plastic 2 mL column and washed 5x with ice cold lysis buffer. Lysates were then bound to the column with the flow through re-

collected and passed over the column a second time. After the second binding step columns were washed 2x (volume equivalent to lysate) with lysis buffer and then 3x with wash buffer (lysis buffer without detergent). Columns were then eluted 5x with 100 ul 2mM desthiobiotin elution buffer. After elution binding was repeated on a fresh column with M2 3xFLAG agarose and eluted with 100 ug/ml 3xFLAG peptide and 0.05% RapiGest. Samples were then analyzed by Western and silver stain for recovery of the bait and samples submitted for in-solution mass spectrometry analysis as described previously^{3,21}. These experiments produced a large list of candidate host partners for Rev.

Purification of Rev and Identified Interactors

As overexpression of known interaction partners can force Rev into biologically relevant complexes we also performed experiments in which Strep tagged Rev was coexpressed with 3xFLAG-tagged versions of known host cofactors Ran (or RanQ69L, which cannot hydrolyze GTP²³), and Crm1. Additionally we tagged several candidate factors to validate their interaction with Rev and further characterize their proteomes. For these experiments single FLAG or Strep purifications were performed with GFP-FLAG acting as a negative control. For smaller scale experiments, purifications were performed in batch with 1 hour binding and 10 minute washes²¹.

SILAC Affinity Purification of Rev

Several of the interesting Rev alanine mutants (R35A, L81A, and V109A) identified in Chapter 3 were purified using stable isotope labeling by amino acids in cell culture

(SILAC). Cells were grown in labeled media using SILAC kits (Thermo) supplemented with either ^{13}C ^{15}N lysine/arginine (heavy) or standard lysine/arginine (light). Samples were performed in duplicate with labels switched for reference and mutant. Single step, large-scale strep purifications were performed in batch as described above. Prior to IP a Bradford assay was performed and heavy and light mutant and reference samples were mixed to equal concentrations. Samples were then analyzed on an Orbitrap and z-scores computed for all purified proteins.

Reporter Assays

Candidate interactors identified by AP-MS were cloned into pcDNA-TO FLAG vectors. We then tested the effects of overexpression (0- 100 ng) of these factors in our gag-pol-RRE assay.

Co-localization Assays

Rev was cloned into a vector containing a GFP tag while host interactors were tagged with mCherry. Plasmids were transfected into 293 cells seeded on chambered slides and analyzed on a Nikon Ti-E Microscope 24 hours later. ImageJ was used to merge channels.

Results and Discussion

Defining the Rev Interactome Presents Many Unique Challenges

Characterization of the Rev proteome remains a daunting but intriguing challenge.

Despite Rev's small size, its propensity to bind RNA results in a large amount of noise via

non-specific RNA-mediated interactions. Despite this, one previous study has attempted to characterize the Rev interactome through *in vitro* reconstitutions of Rev-RRE export complexes²⁴. The authors identified several RNA helicases that appear to modulate Rev function but also note a high degree of non-specific interactions and require a sophisticated scoring system to filter signal from noise. Our own studies focused on *in vivo* complexes which allowed us to examine interactions, both relevant and artifactual, outside of the Rev-RRE-Crm1 complex. Although we observed some of these same helicases in our purifications we were unable to confidently identify them as relevant interactors as a comparison of AP-MS datasets of other HIV proteins showed limited specificity for these proteins¹.

Purification of Rev Under Varied Conditions

In order to test how dependent the interactome of Rev was on the conditions of purification of Rev we varied several conditions including salt (high/standard/low) the affinity tag used (FLAG or Strep), subcellular fractionation (cytoplasmic/nuclear fractions) and RNAses (Figure 1A). For these experiments we performed single step purifications as concurrent studies showed that post-purification analysis of two parallel but separate purification conditions was more effective increasing signal:noise than a single two step purification (which resulted in a large loss of signal)¹. Surprisingly, no condition identified Crm1 or Ran consistently, most likely reflecting difficulties in detecting a transient and low abundance complex. In order to minimize these difficulties we performed Strep purifications of FS RanQ69L coexpressed with His-Flag-

Crm1, and Flag-Rev. This purification procedure, in combination with the addition of 0.2% deoxycholate to the lysis buffer produced detectable complexes by Western, although MS analysis only semi-consistently detected Ran and failed altogether to detect Crm1 (Figure 1BC). The varied sets of purification conditions provided no clear correlation between higher quality MS results. High salt conditions resulted in a more complete lysis of the nucleus and identification of more proteins, but were experimentally difficult to work with due to their high viscosity. Strep purifications yielded a higher recovery of the bait but had identifiable tag specific noise making FLAG purifications an important orthogonal approach. Fractionation resulted in an overall loss of signal indicating that significantly more material would be required for adequate MS identification. The results of these experiments were combined and scored based on repeatability, functional annotation, and literature evidence to assemble a list of candidate interactors (Table 1, Table 2)¹.

TMEDs: An Intriguing Family of Candidates

One class of proteins consistently found in Rev purifications are the transmembrane emp24 domain-containing proteins. This family of proteins is involved in the early secretory pathway and thought to sort cargo into vesicles²⁵. Our screen identified 5 different TMED proteins (Table 3). We cloned tested them for interactions with a Strep tagged Rev (Figure 2A/B). All TMEDs successfully recovered Rev. Domain mapping experiments of the Rev-TMED interactions further showed requirement of Rev's unstructured C-terminus but not the N-terminus (Figure 2C). We also sought to

characterize the TMEDs in higher order complexes. MS analysis of TMED pulldowns identified our flag-tagged versions in complex with endogenous TMEDs (Figure 2D), suggesting that these interactions may constitute a larger multi-protein complex.

A Role for TMEDs in RNA Trafficking

The TMED-Rev interaction is intriguing given Rev's potential to influence RNA fate in the cytoplasm. The steps from nuclear export to encapsidation still remain largely unknown. Studies of HIV-1 RNAs labeled using MS2 binding sites and MS2-GFP have shown that HIV-1 RNAs move along endosomes, sites where TMEDs may be present²⁶. Although this endosomal trafficking is not required for HIV-1 spread, it appears that this pool of RNAs may play a role in cell-to-cell transmission. Furthermore, expression of Rev *in trans* has been shown to enhance encapsidation¹⁵. Given their location in a cellular compartment where trafficking to the membrane, and encounters with Env or Gag are possible, the Rev-TMED interaction may provide a mechanism for the packaging and assembly differences between Rev and Nxf1 mediated export.

Overexpression of TMEDs has Negligible effect in Gag-Pol Reporter System

Despite the strong evidence for a physical interaction between TMEDs and Rev, overexpression of TMEDs appears to have minimal effect on Rev-mediated export. However this is not an entirely unexpected result as the most likely role for TMEDs is downstream of export and not likely to be captured in this assay. (Figure 3)

TMED9 Changes Rev Localization

Expression of GFP-Rev is primarily nuclear with a diffuse cytoplasmic signal. TMED9 and TMED4 appear to have purely cytoplasmic location with discrete puncta near the membrane. Co-expression of TMED9 and Rev appears to cause Rev to persist in the cytoplasm, suggesting that a direct TMED9-Rev interaction occurs *in vivo*. No such effect is seen with TMED4. (Figure 4)

SILAC Analysis of Alanine Mutants

Three mutants were selected for SILAC analysis based on their activity in the reporter assay as well as their known functions. R35A is a particularly deleterious mutation in the ARM and should function to reduce Rev activity. L81A is a completely non-functional mutation of the NES and should remain nuclear. V109A has particularly diminished activity but no known functional role. R35A and L81A produced expected results with L81A having low amount of cytoplasmic interactions and R35A displaying decreased RNA binding likely through the loss of an arginine. Our AP-MS study confidently purified our candidate interactors; however the only statistically significant changes in both replicates occurred between V109A and several SRP proteins. While these proteins were not candidates from the initial screen, it is intriguing that machinery related to the ER appears to map to an unknown domain on Rev. (Table 4)

Post-Translational Modifications of Rev

MS analysis of Rev identified definitive phosphorylation of serine 8 (E-value 1.7×10^{-7}) and possible phosphorylation of serine 67 (1.7×10^{-5}). Serine 8 forms a classic casein kinase II site, and its phosphorylation has been linked to the down-regulation of Rev

function²⁷. Gly-Gly modification of lysine 20 was also detected (4.1×10^{-8}). This is likely the reason for Rev to occasionally appear as a doublet. Excision and MS analysis of the upper band shows this modification; the lower band has peptide coverage of lysine 20, but shows no modification. This modification is traditionally thought to be a sign of ubiquitination via ISG15²⁸.

Summary

Biologically meaningful proteomic characterization of Rev remains a challenge. The transience of known functional interactions makes it difficult to determine standards for “optimal purifications” and Rev’s RNA binding properties cause many RNA binding proteins to consistently co-purify with it. However Rev appears to be able to influence events far downstream of RNA export and interacts with a large amount of ER-associated machinery. Further work is needed to fully understand the relevance of these interactions in the context of HIV-1 replication.

References

1. Jäger, S. *et al.* Global landscape of HIV-human protein complexes. *Nature* **481**, 365–70 (2012).
2. Gavin, A. *et al.* Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636 (2006).
3. Krogan, N. J. *et al.* Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–43 (2006).
4. Devos, D. & Russell, R. B. A more complete, complexed and structured interactome. *Current opinion in structural biology* **17**, 370–7 (2007).
5. Superti-furga, G. Mass spectrometry–based functional proteomics: from molecular machines to protein networks. *Nature Methods* **4**, 807–815 (2007).
6. Chang, I.-F. Mass spectrometry-based proteomic analysis of the epitope-tag affinity purified protein complexes in eukaryotes. *Proteomics* **6**, 6158–66 (2006).
7. Cristea, I. M. *et al.* Tracking and elucidating alphavirus-host protein interactions. *The Journal of biological chemistry* **281**, 30269–78 (2006).
8. Frankel, A. D. & Young, J. a. HIV-1: fifteen proteins and an RNA. *Annual review of biochemistry* **67**, 1–25 (1998).
9. Greene, W. C. & Peterlin, B. M. Charting HIV’s remarkable voyage through the cell: Basic science as a passport to future therapy. *Nature medicine* **8**, 673–80 (2002).
10. Yedavalli, V. S. R. K., Neuveut, C., Chi, Y.-H., Kleiman, L. & Jeang, K.-T. Requirement of DDX3 DEAD box RNA helicase for HIV-1 Rev-RRE export function. *Cell* **119**, 381–92 (2004).
11. Szebeni, a *et al.* Nucleolar protein B23 stimulates nuclear import of the HIV-1 Rev protein and NLS-conjugated albumin. *Biochemistry* **36**, 3941–9 (1997).
12. Cullen, B. R. Nuclear RNA export. *Journal of Cell Science* **116**, 587–597 (2003).
13. Powell, D. M., Amaral, M. C., Wu, J. Y., Maniatis, T. & Greene, W. C. HIV Rev-dependent binding of SF2/ASF to the Rev response element: possible role in Rev-mediated inhibition of HIV RNA splicing. *Proc. Natl. Acad. Sci. USA* **94**, 973–8 (1997).

14. D'Agostino, D. M., Felber, B. K., Harrison, J. E. & Pavlakis, G. N. The Rev protein of human immunodeficiency virus type 1 promotes polysomal association and translation of gag/pol and vpu/env mRNAs. *Molecular and cellular biology* **12**, 1375–86 (1992).
15. Blissenbach, M., Grewe, B., Hoffmann, B., Brandt, S. & Uberla, K. Nuclear RNA export and packaging functions of HIV-1 Rev revisited. *Journal of virology* **84**, 6598–604 (2010).
16. Moore, M. D. *et al.* Probing the HIV-1 genomic RNA trafficking pathway and dimerization by genetic recombination and single virion analyses. *PLoS pathogens* **5**, e1000627 (2009).
17. Swanson, C. M., Puffer, B. A., Ahmad, K. M., Doms, R. W. & Malim, M. H. Retroviral mRNA nuclear export elements regulate protein function and virion assembly. *The EMBO journal* **23**, 2632–40 (2004).
18. Marques, S. M. P., Veyrone, J. & Kumar, A. Restriction of Human Immunodeficiency Virus Type 1 Rev Function in Murine A9 Cells Involves the Rev C-Terminal Domain. *Journal of virology* **77**, 3084–3090 (2003).
19. Lichty, J. J., Malecki, J. L., Agnew, H. D., Michelson-Horowitz, D. J. & Tan, S. Comparison of affinity tags for protein purification. *Protein expression and purification* **41**, 98–105 (2005).
20. Gloeckner, C. J., Boldt, K., Schumacher, A., Roepman, R. & Ueffing, M. A novel tandem affinity purification strategy for the efficient isolation and characterisation of native protein complexes. *Proteomics* **7**, 4228–34 (2007).
21. Jäger, S. *et al.* Purification and characterization of HIV-human protein complexes. *Methods San Diego Calif* **53**, 13–19 (2011).
22. Vardabasso, C., Manganaro, L., Lusic, M., Marcello, A. & Giacca, M. The histone chaperone protein Nucleosome Assembly Protein-1 (hNAP-1) binds HIV-1 Tat and promotes viral transcription. *Retrovirology* **5**, (2008).
23. Klebe, C., Bischoff, F. R., Ponstingl, H. & Wittinghofer, A. Interaction of the nuclear GTP-binding protein Ran with its regulatory proteins RCC1 and RanGAP1. *Biochemistry* **34**, 639–647 (1995).
24. Naji, S. *et al.* Host cell interactome of HIV-1 Rev includes RNA helicases involved in multiple facets of virus production. *Molecular & cellular proteomics : MCP* **11**, M111.015313 (2012).

25. Strating, J. R. P. M. & Martens, G. J. M. The p24 family and selective transport processes at the ER-Golgi interface. *Biology of the cell under the auspices of the European Cell Biology Organization* **101**, 495–509 (2009).
26. Molle, D. *et al.* Endosomal Trafficking of HIV-1 Gag and Genomic RNAs Regulates Viral Egress. *The Journal of Biological Chemistry* **284**, 19727–19743 (2009).
27. Meggio, F., D’Agostino, D. M., Ciminale, V., Chieco-Bianchi, L. & Pinna, L. a. Phosphorylation of HIV-1 Rev protein: implication of protein kinase CK2 and pro-directed kinases. *Biochemical and biophysical research communications* **226**, 547–54 (1996).
28. Skaug, B. & Chen, Z. J. Emerging role of ISG15 in antiviral immunity. *Cell* **143**, 187–190 (2010).
29. Reddy, T. R. *et al.* Inhibition of HIV replication by dominant negative mutants of Sam68, a functional homolog of HIV-1 Rev. *Nature medicine* **5**, 635–42 (1999).

Figure Descriptions

Figure 1: Silver stain of a Variety of Rev Purifications. A) Purifications performed with a normal lysis (Strep and FLAG), high salt (500 mM), low salt, (50 mM), RNase or fractionation conditions (Nuclear and Cytoplasmic). All purifications are single step Strep Purifications except for the FLAG purification which is an analogous, single-step FLAG purification. A Marker (M) is shown to the far left. B) Purification of a Rev-Crm1-RanQ69L complex. Flag-Rev, Crm1 HF and RRE are coexpressed with RanQ69L SF. A Strep pulldown is performed on RanQ69L and all components are visualized with FLAG antibody. Alternatively nickel resin is used to purify Crm1 HF and FLAG visualization once again. C) Silver stain of Rev, Ran, and Crm1 baits submitted for MS analysis. Rev's RNA binding activity gives it a relatively high background.

Figure 2: A) Silver stain FLAG Purifications of TMEDs +/- Rev. Each TMED successfully co-purifies with Rev. GFP is shown as a negative control. B) Western blots of TMED pulldowns. GFP and Tat are shown as specificity controls. Numbers represent MIST scores from Jager et al. (figure is reproduced and modified from this work)¹ C) Domain mapping of Rev-TMED interaction. All TMEDs require residues 9-116 for the interaction. Crm1 is a positive control and GFP is a negative control. D) TMED pulldowns (in the absence of Rev) submitted for AP-MS analysis. Network diagram shows which TMED bait recovers another (bait points to partner) or is recovered by a different TMED.

Figure 3: Titration of interactors in the presence of 50 ng Rev and 200 ng RRE reporter. GFP shows a slight enhancement of reporter activity indicating the uncertainty of the

assay. SAM68, a previously established Rev interactor²⁹ acts as a control showing a clear excitation of the reporter while Tmed9 has no clear effect.

Figure 4: Expression of GFP-tagged Rev and mCherry tagged TMEDs. Rev shows a nuclear localization while Tmed9 and 4 show cytoplasmic localizations. An NLS-GFP functions as a control. Coexpression of Tmed9 and Rev results in cytoplasmic colocalization while no such phenotype is seen with Tmed4.

Protein Name	UniProt	Protein Name	UniProt	Protein Name	UniProt	Protein Name	UniProt
THOC4	Q86V81	SFPQ	P23246	LBR	Q14739	NUP214	P35658
DDX24	Q9GZR7	NONO	Q15233	DDX5	P17844	S23IP	Q9Y6Y8
GRWD1	Q9BQ67	MATR3	P43243	DDX17	Q92841	SEC13	P55735
COPB2	P35606	TPR	P12270	XPO1	O14980	TMEDA	P49755
COPG2	Q9UBF2	NU153	P49790	DDX3X	O00571	TMED9	Q98VK6
NOG2	Q13823	NUP93	Q8N1F7	RAN	P62826	SPATS2	Q86XZ4
LPPRC	P42704	EXOS7	Q15024	RANQ69L	P62826	TMED4	Q727H5
KHDR1	Q07666	NXF1	Q9UBU9	HUWE1	Q72627	TMED2	Q15363
XPO4	Q9C0E2	LMNB1	P20700	ILF2	Q12905	RED	Q13123
XPO2	P55060	LMNA	P02545	ILF3	Q12906	TMED1	Q13445

Table 2: Top 40 candidates filtered by raw scores from Table 1, functional annotation and literature searches.

Name	Rank	Subfamily
TMED1	12	gamma
TMED10	40	delta
TMED9	41	alpha
TMED4	128	alpha
TMED2	337	beta

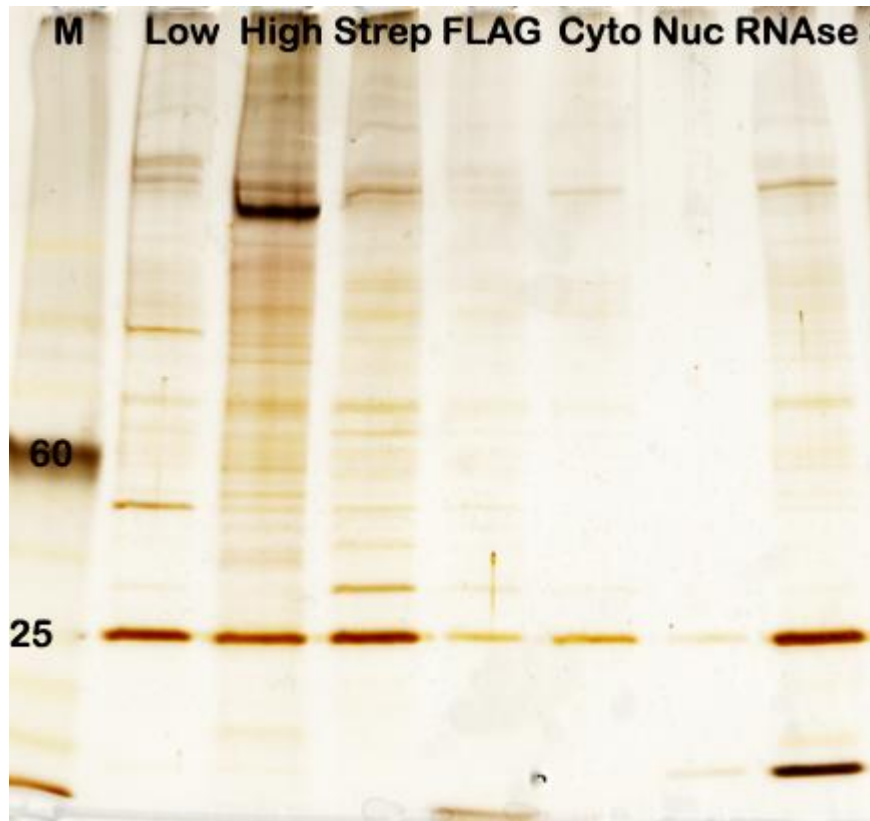
Table 3: TMED proteins identified in proteomic screen with raw rankings from Table 1 and subfamily membership.

	<i>V109A H</i>	<i>V109 L</i>
RPS7	1.2011	0.87967
SRP72	1.2921	0.86681
SRP19	1.5106	0.67425
SRP9	1.5163	0.87953

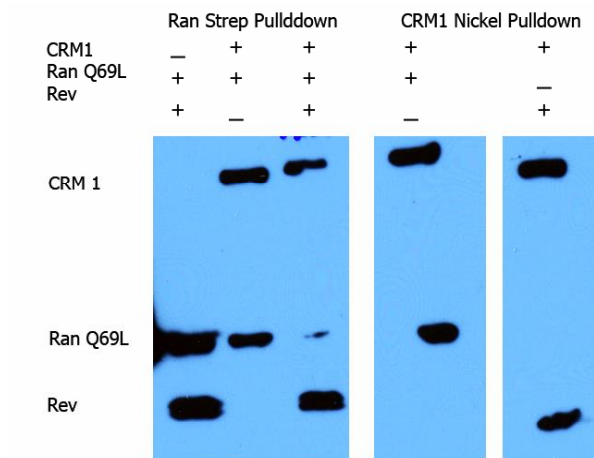
Table 4: Statistically Significant and Consistent changes in SILAC AP experiments for V109A vs reference. Both columns represent the heavy:light ratio suggesting that V109A preferentially associates with these proteins compared to reference.

Figure 1

A



B



C

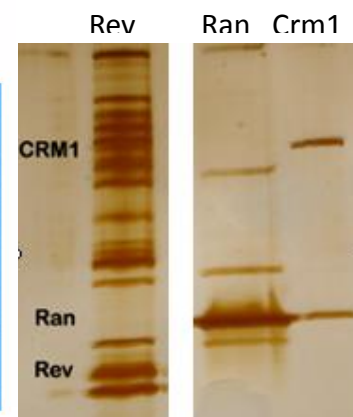
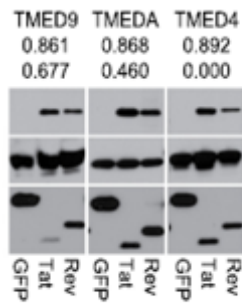


Figure 2

A



B



C



D

TMED
1 2 4 9

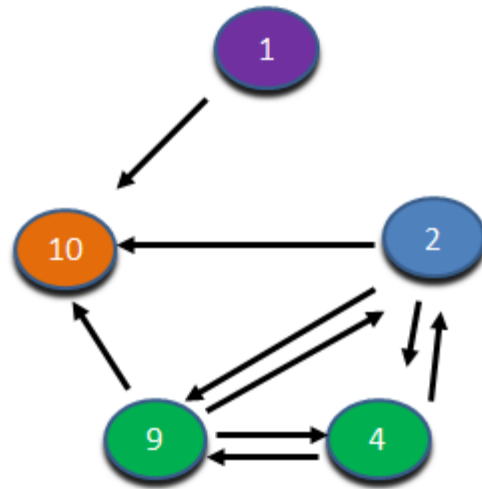
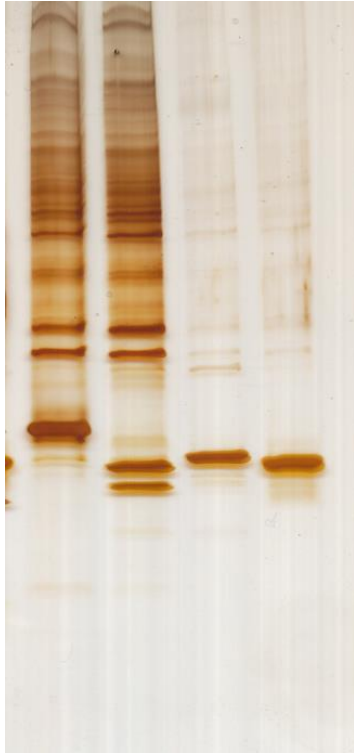


Figure 3

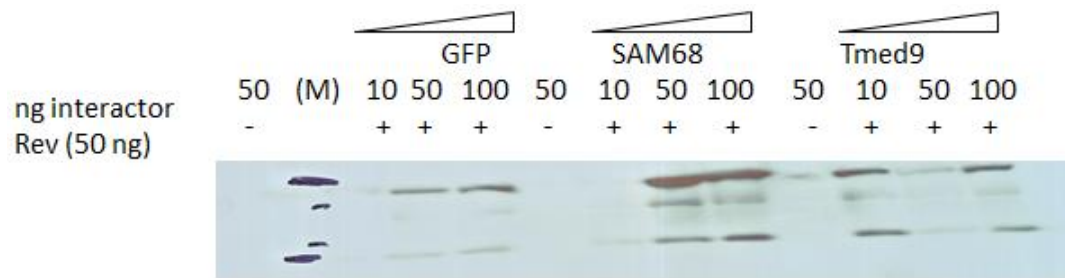
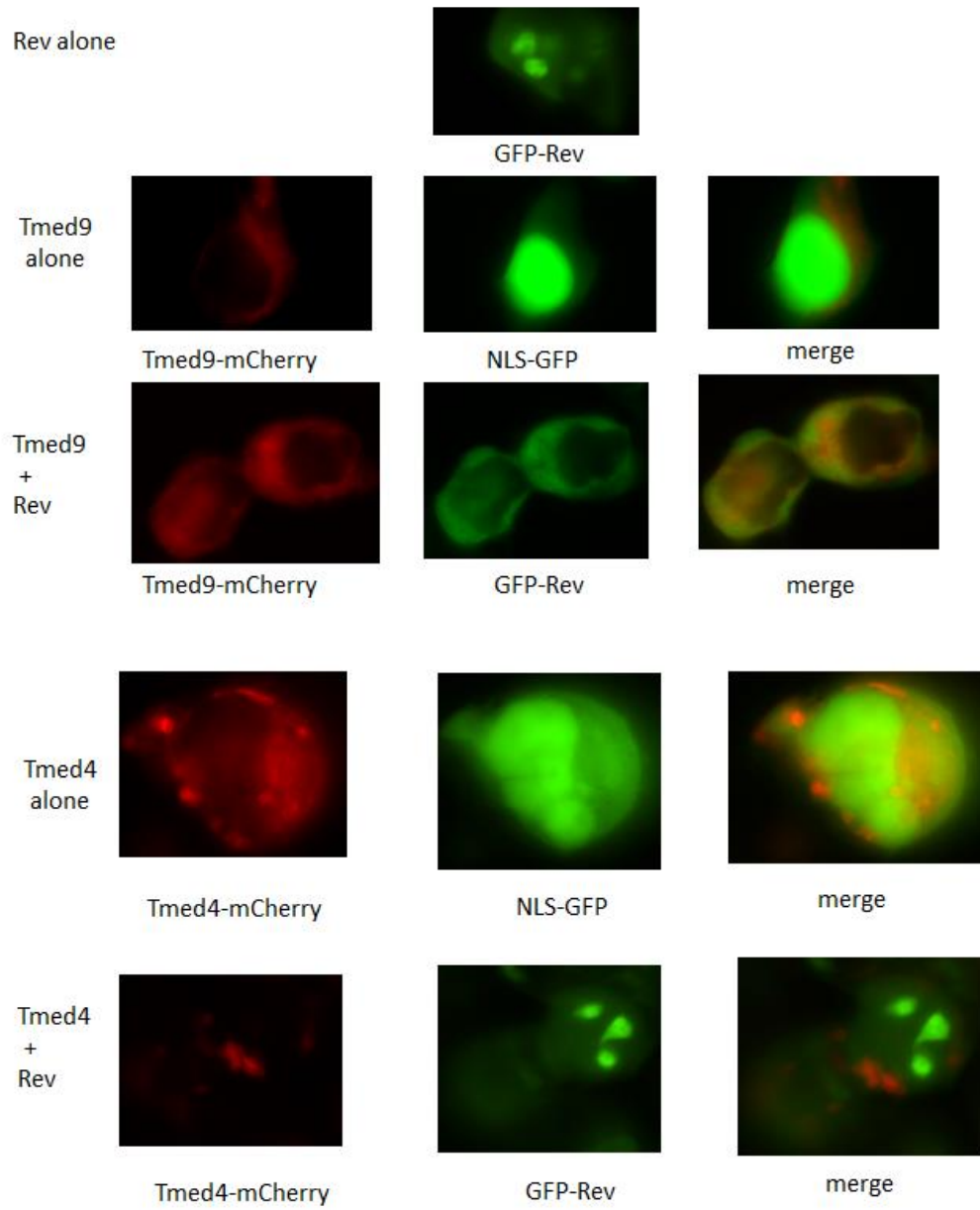


Figure 4



Concluding Remarks

The work outlined here represents the most in-depth dissection of a set of overlapping genes to date. The focus on residue level dissections of entropy, function, and fitness provide strong evidence for selfish organization of overlapped genes. It is our hope that these experiments further the rethinking of overlaps in the context of adaptive evolution instead of as artifacts of additional selection pressures.

Furthermore much work remains to be done in characterizing the molecular details of the functions that have evolved under this unique set of pressures. Our work on proteomic characterization of Rev interaction partners suggests that there still much to discover about the roles of these proteins in viral replication.

Appendix A

List of Common Abbreviations

AD:	Activation Domain
AP-MS:	Affinity Purification-Mass Spectrometry
ARM:	Arginine Rich Motif
dN:	Rate of non-synonymous substitution
dS:	Rate of synonymous substitution
FFL:	Firefly luciferase
HCV:	Hepatitis C Virus
HIV-1:	Human Immunodeficiency Virus type 1
LTR:	Long Terminal Repeat
NES:	Nuclear Export Sequence
NLS:	Nuclear Localization Sequence
OD:	Oligomerization Domain
p24:	HIV capsid
RBD:	RNA Binding Domain
RNP:	Ribonucleoprotein
SIV:	Simian Immunodeficiency Virus
TMED:	Transmembrane emp24 domain-containing protein

Appendix B:Supplemental Figures

Figure Descriptions

Figure B-1: The Gentic Code. Interpretation of codons in the standard genetic code.

Figure B-2: Groupings of Amino Acids by Similar Characteristics. Modified from:
Livingstone, C. D. and Barton, G. J. (1993), "Protein Sequence Alignments: A Strategy for the Hierarchical Analysis of Residue Conservation", *Comp. Appl. Bio. Sci.*, 9, 745-756.

Figure B-3: Unfiltered Protein Alignments. Raw protein alignments (not HXB2 numbering) of A) Tat and B) Rev. Note the common occurrence of Rev extensions including a fairly common QSQGTET insert in the C-terminus. Conservation and quality scores are included with the alignments.

Figure B-4: Mean Pairwise Distance Plots for Tat and Rev. Mean Pairwise Distance (MPD) is a useful metric to measure how frequently and drastically a codon mutates. MPD reflects the average number of nucleotide differences between all codons at a particular position in the alignment (0 = conserved codon, 3 = codons that differ by all 3 nt). MPD plots for Tat and Rev are provided here as they provide a deeper, nucleotide based analysis instead of simply looking at protein identity. MPD is plotted as nt from the Tat start (1 = A of Tat ATG) considering only the mature (i.e. introns ignored) mRNAs of Tat and Rev.

Figure B-5: Synonymous Rates of Evolution. Site-specific d_s values calculated according to the Nei-Gojobori method. Each site is represented by 3 data points so that all 3

curves can be properly shown and offset. Values were computed along the entirety of Tat, Rev, and the overlapping portion of Env.

Figure B-6: Non-synonymous Rates of Evolution. Figure B-5: Site-specific dn values calculated according to the Nei-Gojobori method. Each site is represented by 3 data points so that all 3 curves can be properly shown and offset. Values were computed along the entirety of Tat, Rev, and the overlapping portion of Env.

Figure B-7: Representative Flow Cell Image and Q-Score Distribution. Pictured is a flow cell image from a successful MiSeq run. Q score distributions show that ~98% of called bases had a quality score of 40 (1:10,000 error rate) or better.

Figure B-8: Biochemical Implications of Different Overlapped Architectures. The effect of overlapped reading frames can be represented by a single codon as a dipeptide (the codon is flanked with random nucleotides). Thus any single codon encodes a list of 64 possible dipeptides. By BLASTing this list against the list generated by the list produced by synonymous codons we can determine how constraining or variable the requirement of an amino acid is to the other frame. High scores indicate that that part of the dipeptide maintains a particular characteristic while a low score indicates a wide variety of disparate substitutions. For instance a requirement of W, TGG, fixes the +1 frame as G (GGN) and scores highly.

Figure B-9: Phylogenetic History of Retroviruses. Only a few families possess Rev and/or Tat orthologues (the lentiviruses, the delta retroviruses and the beta

retroviruses). Although it is unclear which gene is older, both genes appear to predate the HIV-1 accessory factors, vif, vpr, vpu and nef.

Figure B-10: Organization of the tat/rev Overlap in Other Viruses

Despite the large sequence and mechanistic diversity amongst different tat and rev orthologues, the positioning of rev +1 of tat seems to hold true. A notable exception is the oldest known lentivirus RELIK. RELIK's Rev and Tat remain poorly characterized and it is possible that they contain second, overlapping exons.

Figure B-1: The Standard Genetic Code

		Second Letter		T		C		A		G	
F i r s t L e t t e r	T	TTT	Phe F	TCT	Ser S	TAT	Tyr Y	TGT	Cys C	T	T
		TTC	Phe F	TCC	Ser S	TAC	Tyr Y	TGC	Cys C	C	h
		TTA	Leu L	TCA	Ser S	TAA	Stop *	TGA	Stop *	A	i
		TTG	Leu L	TCG	Ser S	TAG	Stop *	TGG	Trp W	G	r
	C	CTT	Leu L	CCT	Pro P	CAT	His H	CGT	Arg R	T	d
		CTC	Leu L	CCC	Pro P	CAC	His H	CGC	Arg R	C	
		CTA	Leu L	CCA	Pro P	CAA	Gln Q	CGA	Arg R	A	L
		CTG	Leu L	CCG	Pro P	CAG	Gln Q	CGG	Arg R	G	e
	A	ATT	Ile I	ACT	Thr T	AAT	Asn N	AGT	Ser S	T	t
		ATC	Ile I	ACC	Thr T	AAC	Asn N	AGC	Ser S	C	T
		ATA	Ile I	ACA	Thr T	AAA	Lys K	AGA	Arg R	A	E
		ATG	Met M	ACG	Thr T	AAG	Lys K	AGG	Arg R	G	R
	G	GTT	Val V	GCT	Ala A	GAT	Asp D	GGT	Gly G	T	
		GTC	Val V	GCC	Ala A	GAC	Asp D	GGC	Gly G	C	
		GTA	Val V	GCA	Ala A	GAA	Glu E	GGA	Gly G	A	
		GTG	Val V	GCG	Ala A	GAG	Glu E	GGG	Gly G	G	

Figure B-2: Amino Acid Groupings

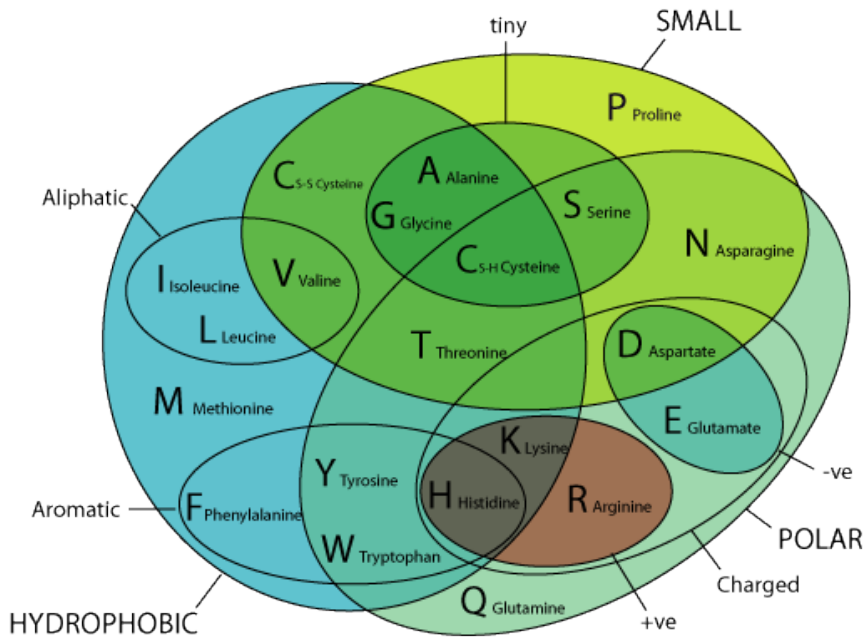
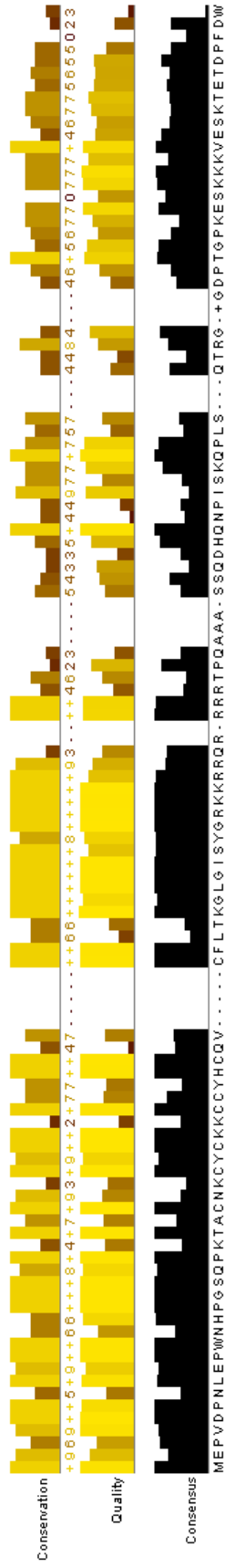
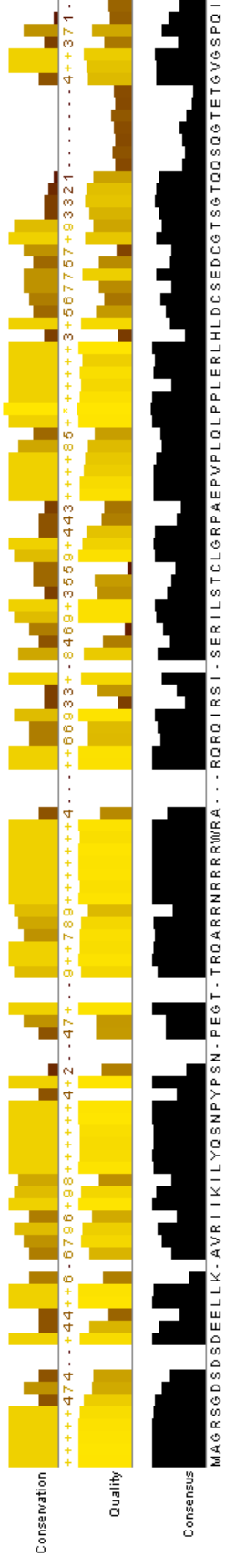


Figure B-3

A. Tat protein Alignment



B. Rev Protein Alignment



B. Rev alignment (continued)

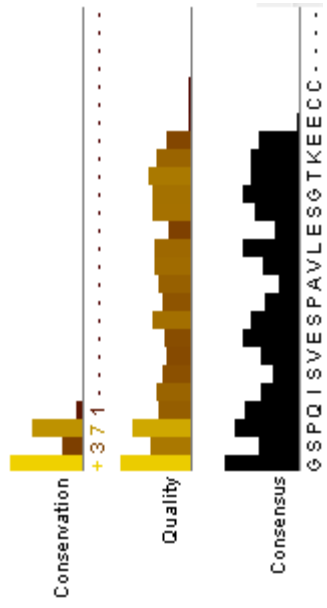


Figure B-4

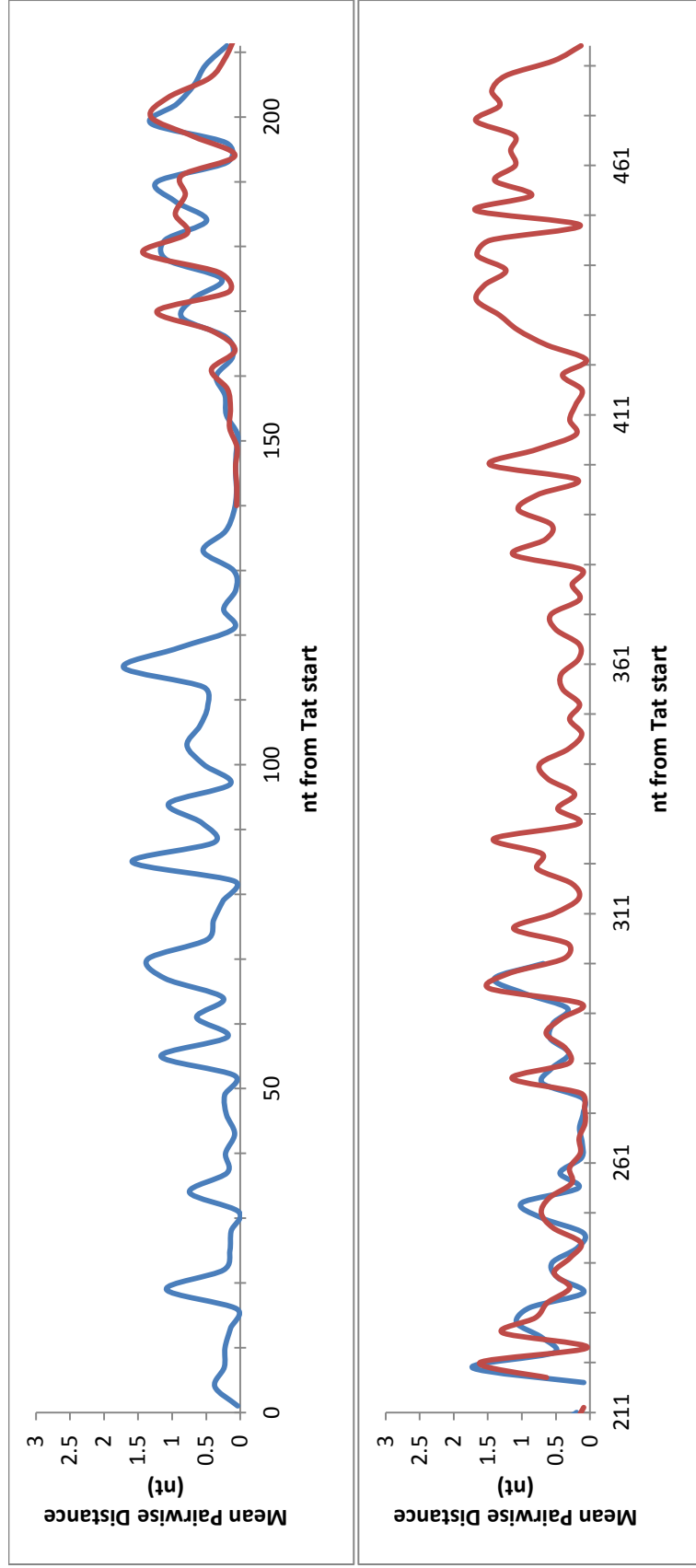


Figure B-5

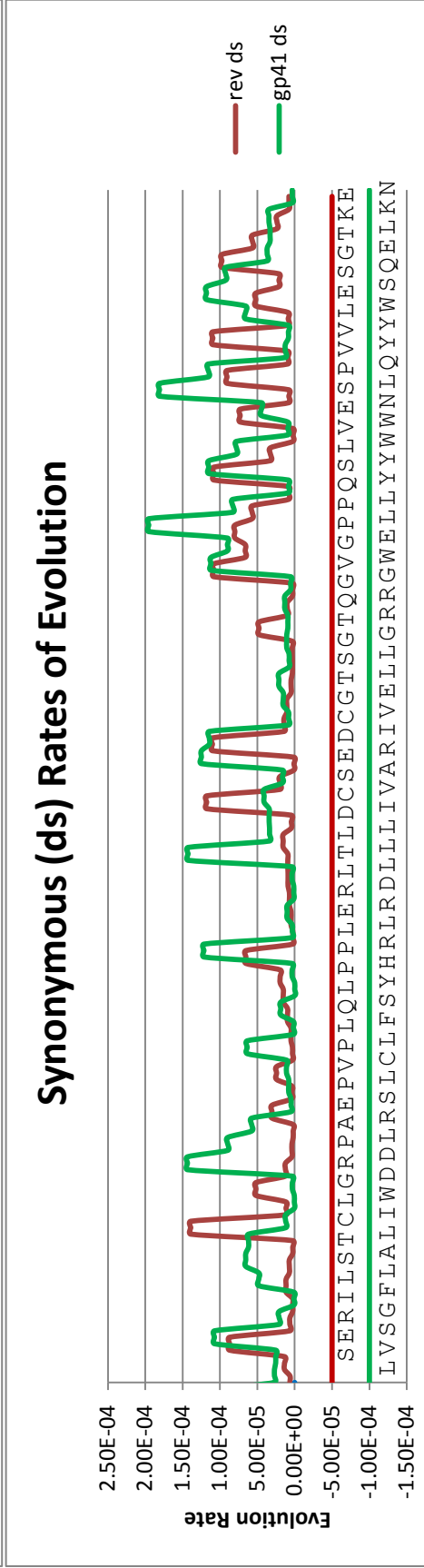
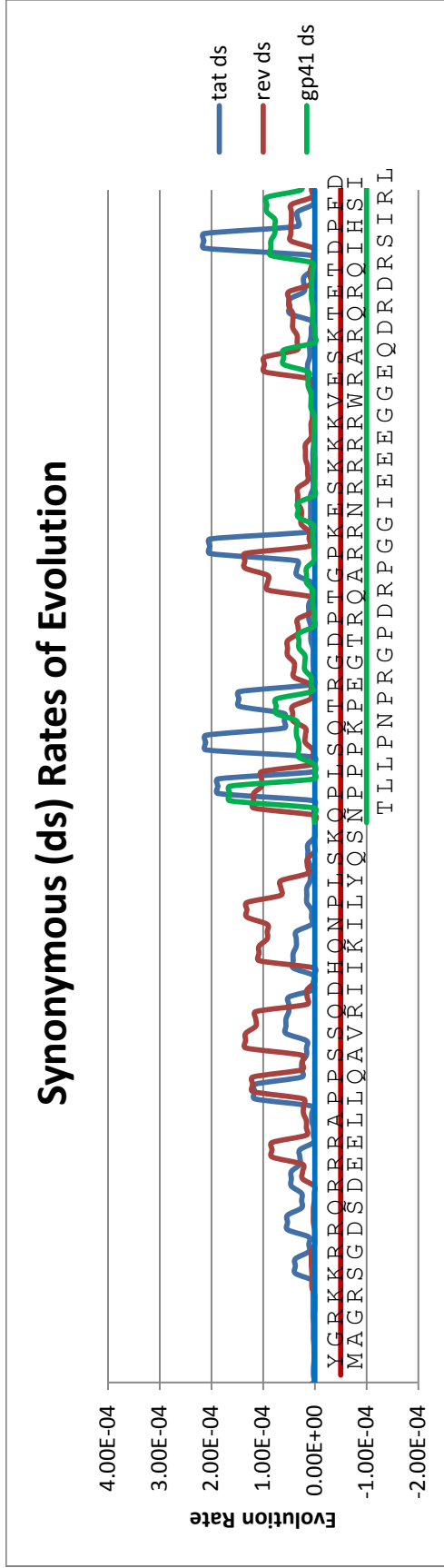
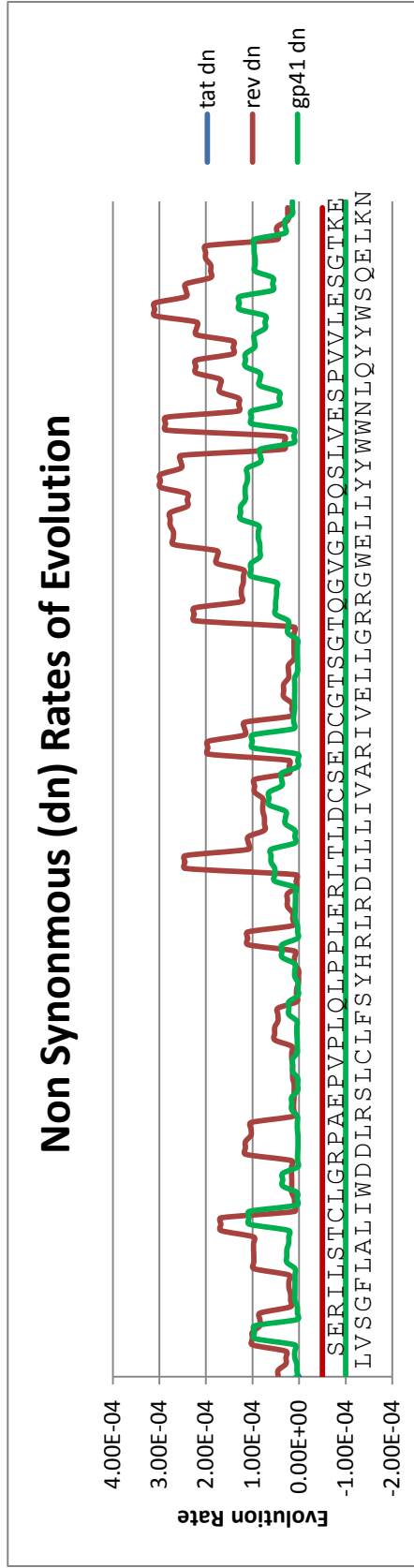
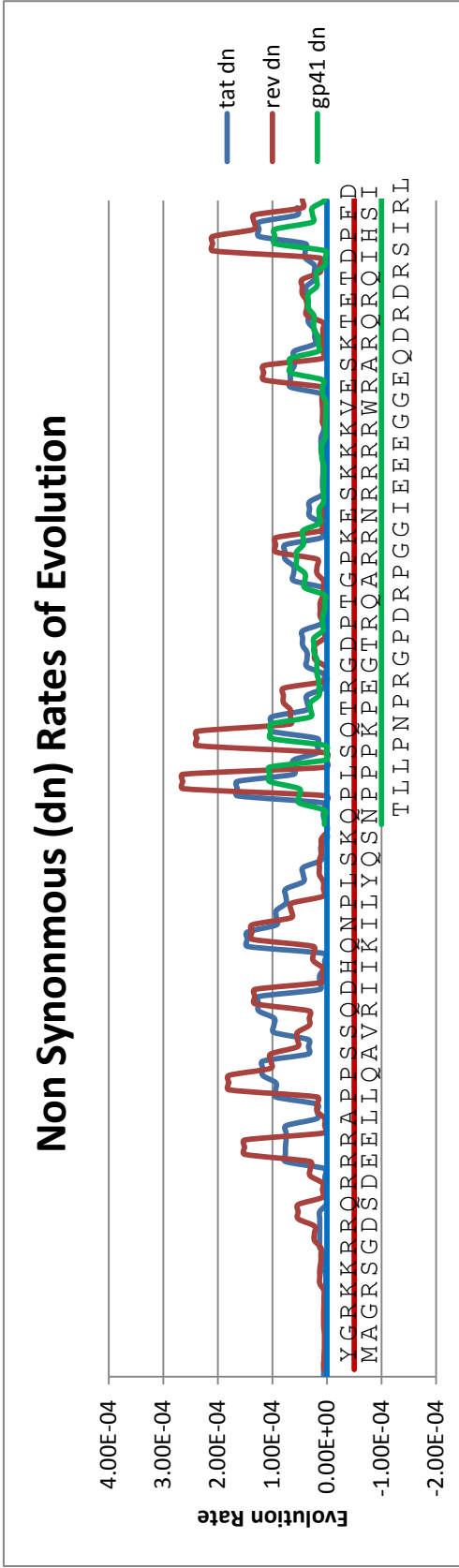
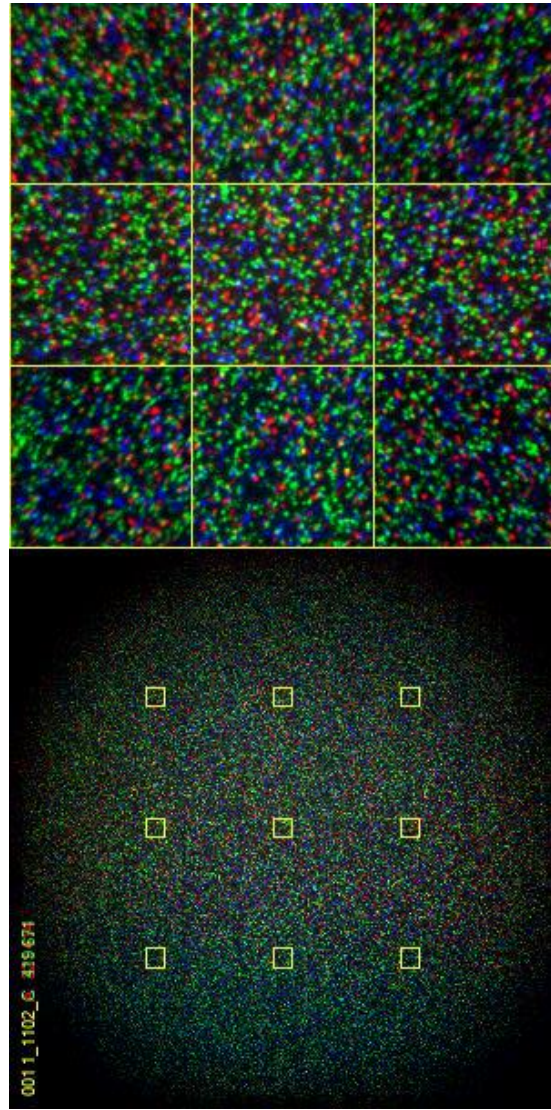


Figure B-6



Appendix B-7: Representative Flow Cell Image and Q-Score Distribution

A



B

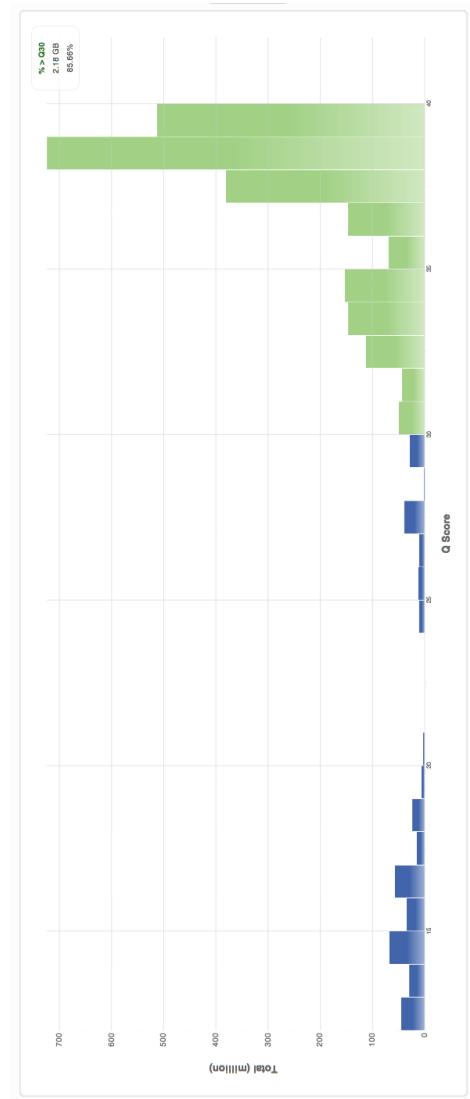
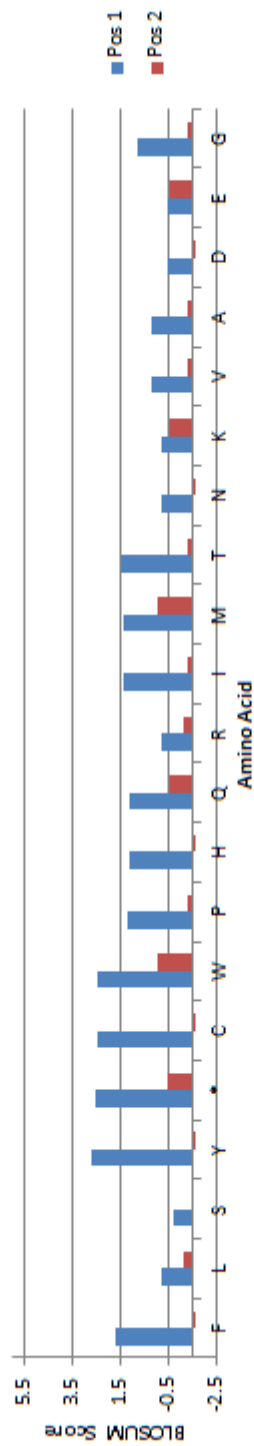


Figure B-8: Biochemical Implications of Different Overlapped Architectures

Dipeptide

+1: N N A T G N NNA TGN
 -1: N A T G N N NAT GNN

Mean BLOSUM62 Dipeptide Score (-1 Frame)



Mean BLOSUM62 Dipeptide Score (+1 Frame)

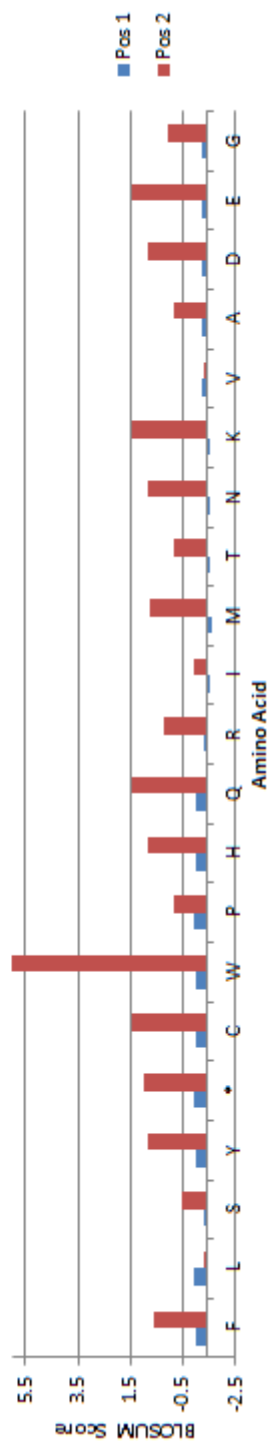


Figure B-9: Phylogenetic History of Retroviruses

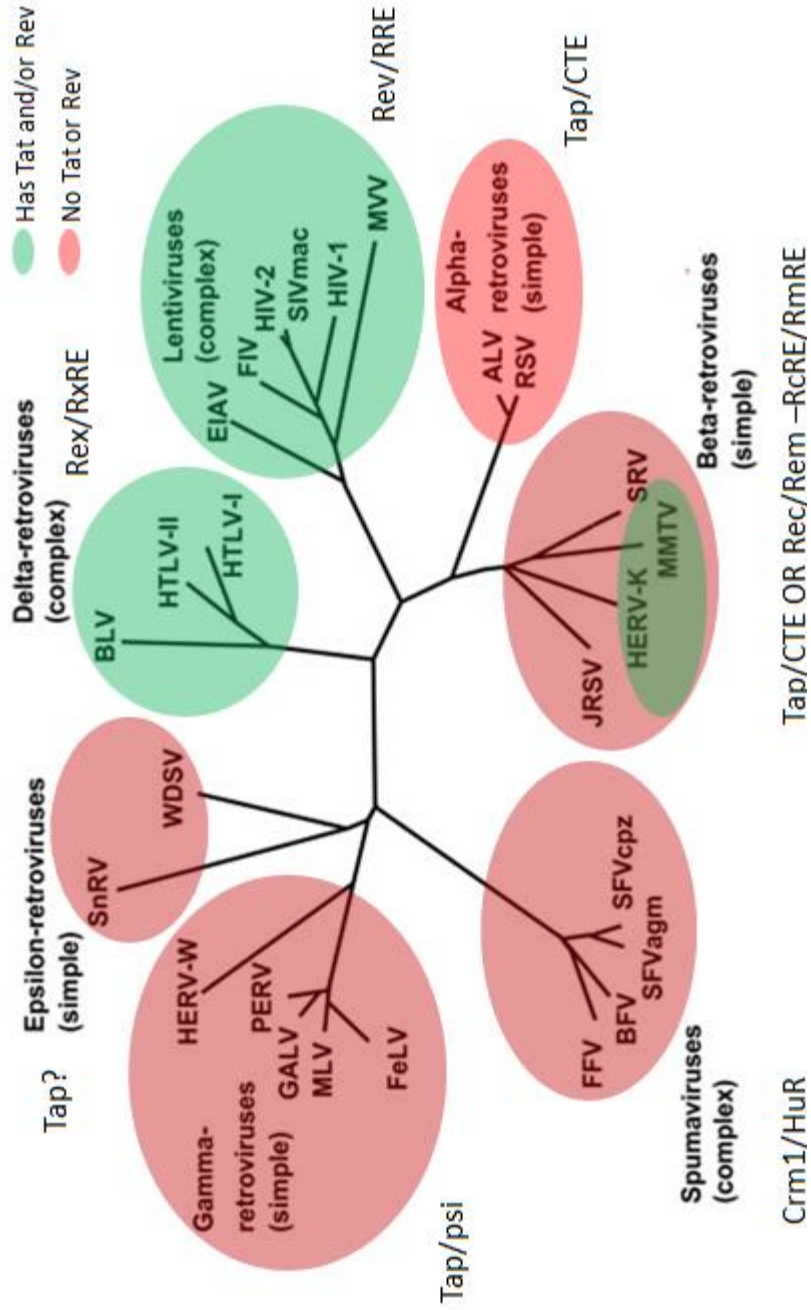
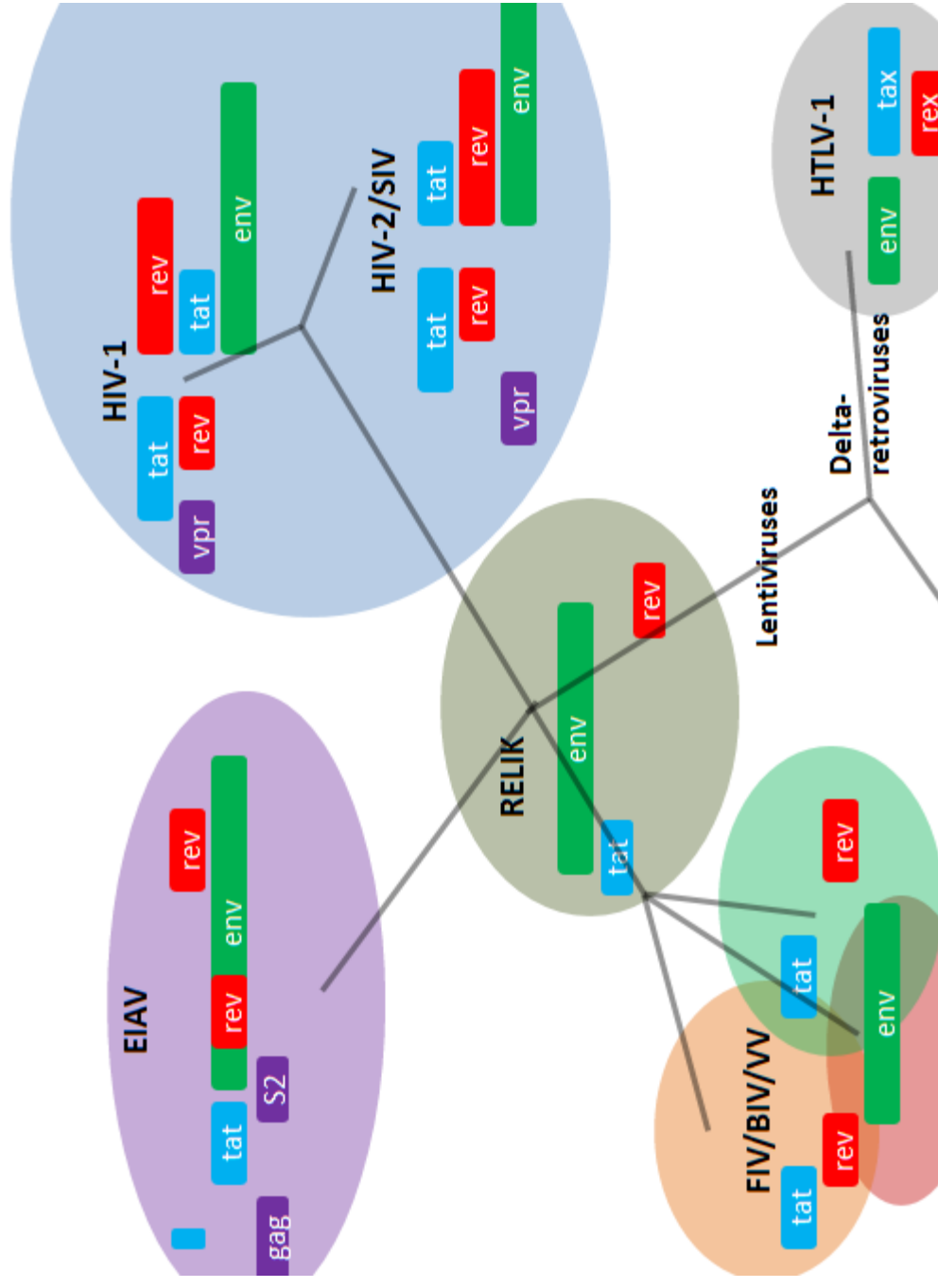


Figure B-10: Organization of the tat/rev Overlap in Other Viruses



Appendix C: Supplementary Tables

N40	F	F	L	L	S	S	S	S	S	S	S	S	Y	Y	Y	*	*	C	C	*	W	L	L	L	L
P	TTT	TTC	TTA	TTG	TCT	TCC	TCA	TCG	TAT	TAC	TAA	TAG	TGT	TGC	TGA	TGG	CTT	CTC	CTA	CTG	CTG	CTA	CTG	CTA	CTG
	0.017	0.012	0.012	0.013	0.015	0.014	0.008	0.016	0.009	0.016	0.009	0.016	0.015	0.015	0.019	0.008	0.024	0.000	0.002	0.002	0.002	0.002	0.002	0.002	
t0	0.017	0.011	0.012	0.013	0.017	0.013	0.009	0.017	0.011	0.015	0.011	0.017	0.015	0.015	0.018	0.009	0.030	0.001	0.001	0.003	0.002	0.002	0.003	0.002	
t7	0.001	0.002	0.001	0.009	0.016	0.007	0.010	0.009	0.005	0.003	0.003	0.008	0.001	0.002	0.007	0.001	0.016	0.000	0.001	0.000	0.000	0.000	0.000	0.000	
t12	0.001	0.002	0.001	0.005	0.016	0.017	0.007	0.006	0.003	0.003	0.005	0.001	0.001	0.001	0.005	0.001	0.017	0.000	0.001	0.000	0.000	0.000	0.000	0.000	

P	P	P	K	H	H	Q	R	R	R	R	R	V	V	V	V	V	A	A	A	A	A	A	M	T	T	T	
CCT	CCC	CCA	CCG	CAT	CAC	CAA	CAG	CGT	CGC	CGA	CGG	GTG	GTC	GTT	GTG	GTA	GCC	GCT	GCA	GCC	GCC	GCC	ATG	ACT	ACC	ACA	ACG
0.005	0.007	0.005	0.003	0.003	0.004	0.002	0.001	0.001	0.002	0.002	0.002	0.018	0.022	0.023	0.023	0.018	0.026	0.021	0.025	0.026	0.025	0.016	0.036	0.015	0.020	0.013	0.024
0.005	0.007	0.005	0.004	0.004	0.004	0.001	0.001	0.001	0.002	0.002	0.002	0.017	0.022	0.023	0.022	0.017	0.025	0.025	0.025	0.025	0.025	0.016	0.036	0.015	0.018	0.015	0.020
0.000	0.001	0.001	0.000	0.002	0.009	0.001	0.001	0.000	0.001	0.000	0.003	0.002	0.007	0.016	0.004	0.006	0.006	0.047	0.047	0.047	0.047	0.025	0.017	0.001	0.003	0.003	0.005
0.000	0.001	0.001	0.000	0.002	0.007	0.001	0.001	0.000	0.001	0.000	0.003	0.001	0.007	0.015	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.002	0.003	0.003	0.008

N	N	K	K	S	S	R	R	R	R	R	R	V	V	V	V	V	A	A	A	A	A	A	A	D	D	E	E
AAT	AAC	AAA	AAG	AGT	AGC	AGA	AGG	GTT	GTC	GTA	GTG	GTC	GTC	GTT	GTT	GTA	GCC	GCT	GCA	GCC	GCC	GCC	GAT	GAC	GAA	GAG	GAG
0.019	0.084	0.017	0.024	0.019	0.018	0.012	0.012	0.023	0.022	0.018	0.018	0.018	0.022	0.023	0.023	0.018	0.026	0.021	0.025	0.026	0.025	0.016	0.036	0.015	0.020	0.013	0.024
0.017	0.082	0.017	0.024	0.020	0.016	0.011	0.011	0.023	0.022	0.017	0.018	0.017	0.022	0.023	0.022	0.017	0.025	0.025	0.025	0.025	0.025	0.016	0.036	0.015	0.018	0.015	0.023
0.232	0.354	0.015	0.019	0.004	0.005	0.010	0.010	0.016	0.007	0.002	0.004	0.006	0.007	0.016	0.004	0.006	0.006	0.047	0.047	0.047	0.047	0.025	0.017	0.001	0.003	0.011	0.005
0.205	0.402	0.008	0.017	0.005	0.004	0.009	0.009	0.013	0.007	0.002	0.003	0.005	0.007	0.009	0.004	0.002	0.002	0.058	0.058	0.058	0.058	0.022	0.019	0.001	0.002	0.011	0.003

G	G	G	G
GGT	GGC	GGA	GGG
0.024	0.030	0.022	0.032
0.023	0.031	0.024	0.033
0.004	0.019	0.007	0.012
0.003	0.021	0.006	0.012

Table C1. Raw Data for Library Selection: Read frequencies of each codon for proviral plasmid (P), viral input (t0) and samples 7 (t7) and 12 (t12) days post selection.

C1. Raw Data for Library Selection

E47	F	F	L	L	S	S	S	S	S	Y	Y	Y	*	C	C	*	C	*	W	L	L	L	L	L	L
P	TTT	TTC	TTA	TTG	TCT	TCC	TCA	TCG	TAC	TAT	TAC	TAA	TAA	TAG	TGT	TGC	TGA	TGG	CTT	CTC	CTA	CTG	CTG	CTG	
t0	0.010645	0.027022	0.021745	0.022382	0.013193	0.019198	0.009826	0.024929	0.014375	0.017469	0.010736	0.026112	0.022411	0.019205	0.018623	0.026699	0.018563	0.011093	0.014074	0.014359	0.018314	0.018314	0.018314		
t7	0.005346	0.019331	0.018647	0.023123	0.014111	0.039533	0.009262	0.029028	0.012059	0.013675	0.00404	0.007894	0.034063	0.017964	0.009386	0.009635	0.003543	0.008143	0.005967	0.018834	0.018834	0.018834	0.018834		
t12	0.005894	0.017747	0.020953	0.024678	0.01425	0.031317	0.007125	0.032677	0.013084	0.014833	0.004631	0.008874	0.035754	0.020111	0.012177	0.011141	0.004728	0.00787	0.006639	0.024516	0.024516	0.024516	0.024516		

P	P	P	P	H	H	H	H	Q	Q	R	R	R	R	R	R	I	I	I	M	T	T	T	T	T
CCT	CCC	CCA	CCG	CAT	CAC	CAA	CAG	CAG	CAG	CGT	CGC	CGA	CGG	CGG	ATT	ATC	ATA	ATG	ACT	ACC	ACA	ACG	ACG	
0.018896	0.021331	0.014513	0.014703	0.014466	0.015772	0.008254	0.007185	0.007221	0.008563	0.007791	0.001829	0.016616	0.01854	0.016141	0.011235	0.009632	0.013789	0.025428	0.018896	0.018896	0.018896	0.018896	0.018896	
0.018743	0.021199	0.01574	0.01483	0.013557	0.014284	0.008734	0.008189	0.004913	0.00928	0.007916	0.002366	0.01665	0.016741	0.016923	0.011555	0.009462	0.012374	0.022655	0.020107	0.020107	0.020107	0.020107	0.020107	
0.003605	0.008764	0.016037	0.001243	0.004848	0.021942	0.007583	0.002486	0.006029	0.009448	0.005283	0.004413	0.035182	0.019642	0.030955	0.013053	0.011251	0.011748	0.019331	0.016037	0.016037	0.016037	0.016037	0.016037	
0.003724	0.006671	0.007125	0.001166	0.008355	0.024289	0.005511	0.002591	0.007125	0.010201	0.00625	0.003983	0.023512	0.018168	0.027236	0.013408	0.010493	0.012728	0.020241	0.016096	0.016096	0.016096	0.016096	0.016096	

N	N	K	K	S	S	S	S	R	R	V	V	V	V	V	V	A	A	A	A	D	D	D	D	E
AAT	AAC	AAA	AAG	AGT	AGC	AGA	AGG	AGG	GTT	GTC	GTA	GTA	GTG	GTC	GTC	GCC	GCA	GCG	GAT	GAC	GAA	GAG	GAG	
0.0112	0.012684	0.010345	0.013349	0.01962	0.013338	0.015784	0.018278	0.010155	0.014074	0.016449	0.014703	0.00671	0.003824	0.009371	0.008896	0.014893	0.021248	0.021877	0.021236	0.021236	0.021236	0.021236	0.021236	
0.008916	0.008825	0.00837	0.012374	0.021199	0.011464	0.015285	0.013829	0.00928	0.014557	0.01574	0.013375	0.006096	0.003548	0.009462	0.010008	0.014921	0.018834	0.020926	0.023929	0.023929	0.023929	0.023929	0.023929	
0.006651	0.003605	0.007086	0.021196	0.037544	0.011188	0.045002	0.02275	0.005532	0.022377	0.013799	0.038662	0.004786	0.002176	0.006029	0.0023	0.009075	0.009821	0.018772	0.051218	0.051218	0.051218	0.051218	0.051218	
0.007352	0.00366	0.007254	0.019464	0.035721	0.012112	0.042943	0.024969	0.004469	0.022605	0.013343	0.037567	0.003336	0.001976	0.006477	0.003109	0.008971	0.009975	0.016517	0.054764	0.054764	0.054764	0.054764	0.054764	

G	G	G	G
GGT	GGC	GGA	GGG
0.020654	0.018801	0.023065	0.034063
0.022564	0.017833	0.024748	0.038759
0.042268	0.005781	0.015912	0.041211
0.048287	0.006736	0.014152	0.041777

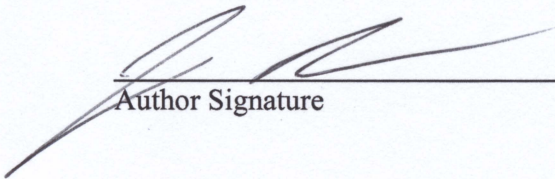
C1. Raw Data for Library Selection

l81	F	F	L	L	L	S	S	S	S	Y	Y	Y	Y	*	C	C	C	*	W	L	L	L	L	L
	TTT	TTC	TTA	TTG	TCT	TCC	TCA	TCG	TAT	TAC	TAA	TAG	TGT	TGC	TGA	TGG	CTT	CTC	CTA	CTG				
P	0.009388	0.01594	0.01463	0.032119	0.016465	0.017966	0.016012	0.013796	0.012342	0.010794	0.008173	0.013486	0.011342	0.018371	0.014535	0.017489	0.014582	0.030808	0.018633	0.013891				
t0	0.008936	0.016165	0.014759	0.036245	0.015361	0.017871	0.01255	0.012952	0.014659	0.011546	0.008333	0.012952	0.012952	0.018976	0.01496	0.021787	0.01506	0.029819	0.021185	0.016466				
t7	0.002198	0.002321	0.059652	0.200074	1.64E-04	0.001931	0.001089	5.75E-04	3.70E-04	0.00113	0.001212	0.005176	0.028807	0.014913	0.002403	8.22E-04	0.062939	0.069903	0.105953	0.140771				
t12	0.003019	0.002868	0.064075	0.224679	1.51E-04	0.002642	0.001208	3.77E-04	6.04E-04	6.04E-04	0.001434	0.006868	0.020981	0.011019	0.001434	4.53E-04	0.062113	0.072226	0.11117	0.135623				
P	P	P	P	P	H	H	S	Q	R	R	R	R	I	I	I	M	T	T	T	T	T	T	T	T
CCT	CCC	CCA	CCG	CCG	CAT	CAC	CAA	CAG	CGT	CGC	CGA	CGG	ATT	ATC	ATA	ATG	ACT	ACC	ACA	ACG				
0.023279	0.037313	0.030189	0.016322	0.013605	0.024375	0.011747	0.007768	0.016536	0.019086	0.015059	0.017418	0.010722	0.010818	0.006624	0.014797	0.012652	0.014773	0.015059	0.010984					
0.024096	0.03755	0.028313	0.01496	0.01255	0.028012	0.011546	0.007028	0.017771	0.019779	0.01506	0.014056	0.011044	0.010241	0.00743	0.013153	0.011345	0.013153	0.013956	0.010004					
5.34E-04	7.39E-04	8.83E-04	0.001089	3.70E-04	8.83E-04	8.83E-04	0.001725	5.55E-04	7.81E-04	2.05E-04	7.81E-04	4.52E-04	0.008443	0.102769	0.053018	0.027546	3.08E-04	8.22E-04	3.49E-04	6.37E-04				
3.77E-04	9.06E-04	6.79E-04	0.001358	7.55E-05	7.55E-05	7.55E-04	0.001585	3.02E-04	7.55E-04	2.26E-04	3.02E-04	8.30E-04	0.008604	0.091396	0.047623	0.035094	0	0.001283	3.77E-04	6.04E-04				
N	N	K	K	K	S	S	R	R	V	V	V	V	A	A	A	A	D	D	D	E	E	E	E	E
AAT	AAC	AAA	AAG	AAG	AGT	AGC	AGA	AGG	GTT	GTC	GTA	GTG	GCT	GCC	GCA	GCG	GAT	GAC	GAA	GAG				
0.005194	0.013224	0.00772	0.007815	0.015488	0.012319	0.007625	0.015964	0.017513	0.021135	0.009245	0.022207	0.018371	0.021706	0.011842	0.010651	0.008077	0.011056	0.005194	0.018204					
0.00502	0.012651	0.008434	0.008635	0.01757	0.012952	0.006627	0.017671	0.017871	0.020884	0.008032	0.024398	0.017169	0.020482	0.010241	0.011044	0.009438	0.00994	0.005924	0.012651					
2.88E-04	9.65E-04	0.058194	9.24E-04	6.16E-04	1.64E-04	3.90E-04	5.34E-04	5.34E-04	0.001582	0.008607	0.006142	0.002506	0.00417	4.72E-04	0.001274	4.52E-04	3.29E-04	4.31E-04	3.49E-04	0.001376				
1.51E-04	4.53E-04	0.047623	0.001208	3.02E-04	3.02E-04	3.02E-04	2.26E-04	2.26E-04	8.30E-04	0.005887	0.004981	0.00317	0.004075	0	0.001132	1.51E-04	7.55E-05	7.55E-04	1.51E-04	9.81E-04				
G	G	G	G	G	GGA	GGG	GGG	GGG																
GGT	GGC	GGA	GGG	GGG																				
0.011461	0.020182	0.015869	0.036455	0.012249	0.019478	0.016365	0.035643																	
3.49E-04	6.57E-04	5.55E-04	0.002054	3.77E-04	1.51E-04	4.53E-04	0.001434																	

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses, dissertations, and manuscripts. Copies of all UCSF theses, dissertations, and manuscripts will be routed to the library via the Graduate Division. The library will make all theses, dissertations, and manuscripts accessible to the public and will preserve these to the best of their abilities, in perpetuity.

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis, dissertation, or manuscript to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

 _____
Author Signature

9/9/13
_____ Date