Title

A Computational and Experimental Approach to Understanding HIV-1 Evolution and Latency for the Design of Improved Antiviral Therapies

Author

Dey, Siddharth Subhas

Publication Date

2012

Peer reviewed|Thesis/dissertation

A Computational and Experimental Approach to Understanding HIV-1 Evolution and Latency for the Design of Improved Antiviral Therapies

By

Siddharth Subhas Dey

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemical Engineering

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David V. Schaffer, Chair
Professor Harvey W. Blanch
Professor Adam P. Arkin

Spring 2012

Abstract

A Computational and Experimental Approach to Understanding HIV-1 Evolution and Latency
for the Design of Improved Antiviral Therapies

by

Siddharth Subhas Dey

Doctor of Philosophy in Chemical Engineering

University of California, Berkeley

Professor David V. Schaffer, Chair

With 33.3 million people presently infected with Human Immunodeficiency Virus-1 (HIV-1), combined with the 2.6 million new infections and 1.8 million AIDS related death in 2009 alone, HIV-1 continues to be one of the biggest global pandemics and medical challenges of the new millennium. Although the development of antiretroviral drugs was a major advance in the treatment of patients infected with HIV-1, complete eradication of HIV-1 has not been possible due to two major obstacles. First, the high mutation rate of the virus coupled with its rapid replication rate has given rise to drug resistant strains of HIV-1. Furthermore, latent viral reservoirs that are not directly targeted by anti-viral therapies or by the immune system can reactivate at a later time preventing complete viral clearance from a patient. Compounding these difficulties is the global diversification of viral strains or subtypes that have widely differing sequences, resulting in unique gene regulation and pathogenesis. Following integration into the host genome, activation of viral gene expression results in the production of new progeny whereas the inability to activate gene expression could initiate the establishment of viral latency. Thus, a better understanding of the mechanisms and factors that regulate viral transcription is critical towards eliminating latent viral populations. Therefore, the focus of this work has been to investigate the role of both cellular and viral factors in regulating HIV-1 gene expression and latency using a combination of computational and experimental techniques. This work may help develop novel therapy targets and better treatment regimens for different HIV-1 subtypes while concurrently providing new insights on mammalian gene regulation.

In studying viral factors that regulate gene expression in HIV-1, we focused attention on the HIV-1 promoter, a viral protein called Tat and a RNA hairpin called TAR. The error prone nature of HIV-1 replication has resulted in highly diverse viral sequences, and it is not clear how Tat, which plays a critical role in viral gene expression and replication, retains its complex functions. Although several important amino acid positions in Tat are conserved, we hypothesized that it may also harbor functionally important residues that may not be individually conserved yet appear as correlated pairs, and knowledge of such evolutionary information could help elucidate underlying mechanisms of Tat function. Using Information theory based approaches such as Mutual Information and protein engineering approaches, we found a pair of sites in Tat that are strongly coevolving and that provided insight into Tat-mediated viral transcription. In contrast to most coevolving protein residues that contribute to the same function, these studies showed that these two residues contribute to two mechanistically distinct

steps in gene expression: binding the cellular protein, positive transcription-elongation factor b (P-TEFb) and promoting P-TEFb phosphorylation of the C-terminal domain in RNA Polymerase II (RNAPII). Moreover, Tat variants that mimic HIV-1 subtype B or C at these sites have evolved orthogonal strengths of P-TEFb binding vs. RNAPII phosphorylation, suggesting that subtypes have evolved alternate transcriptional strategies that could differentially impact latency while achieving similar gene expression levels.

Interaction between Tat and the viral hairpin TAR is critical for efficient gene expression from the viral promoter and we therefore hypothesized that sequence diversity within these elements may dramatically alter the gene expression and latency properties of different subtype viruses. We found large differences in gene expression between subtypes using a variety of experimental models and showed that subtype TARs and Tats act independently to set the level of gene expression from the viral promoter. Further, using Mutual information and site-directed mutagenesis we showed that nucleotides in TAR are not coevolving with residues in Tat implying that HIV-1 has evolved a highly robust mechanism of activating gene expression in the face of rapid viral evolution.

Similarly, the promoters of different HIV-1 subtypes have evolved different architectures of transcription factor binding sites (TFBS) that result in widely varying levels of gene expression and viral replication. Within this large diversity of TFBS in the HIV-1 promoter, we used *in vitro* models of HIV-1 latency to identify the minimal set of TFBS that contribute to most of the observed differences in gene expression and latency at steady state. In contract, we found that the dynamics of gene expression is dependent on both the minimal set of TFBS and other sites in the viral promoter. Identifying other targets within the viral promoter will provide better mechanistic understanding of the establishment and reactivation of HIV-1 latency as well as potentially identify new molecular targets to counter latency.

While diversity in viral factors can contribute to differential regulation of viral gene expression, host factors can also play a significant role in this regulation. Since HIV-1 integrates semi-randomly within the human genome, another aspect of my thesis included studying the role of the cellular genomic location in regulating viral gene expression. We exploited the semi-random integration of HIV-1 to quantitatively study both how latent proviruses can be reactivated from different chromatin environments and to address a fundamental question in eukaryotic gene expression related to how the placement of a gene in the genome impacts its responsiveness to an input transcription factor signal. Using a tunable overexpression system for the transcription factor NF-κB RelA, we quantified HIV-1 expression as a function of RelA levels and chromatin features at a panel of viral integration sites. We demonstrated that chromatin environments at different genomic loci decouple transcription factor mediated gene expression induction thresholds from subsequent gene activation. We developed a functional relationship between gene expression, RelA levels, and chromatin accessibility that accurately predicted synergistic HIV-1 activation in response to combinatorial pharmacological perturbations. Thus, this quantitative study should help inform strategies for combinatorial therapies to combat latent HIV-1 and help unravel biological principles underlying selective gene expression in response to transcription factor inputs.

Finally, after HIV-1 integrates into the host genome, it can either activate gene expression that leads to viral replication or become transcriptionally silent that can result in viral latency. Since stochastic fluctuations in HIV-1 gene expression are one of several factors that have been implicated in influencing this decision and thus in the establishment of viral latency, we investigated the role of the local chromatin environment in regulating gene expression noise. We showed that for clones with similar mean gene expression levels, those integrated into more heterochromatic regions are associated with wider mRNA and protein distributions. Using a two-state stochastic model of gene expression, we showed that the repressed chromatin gives rise to noisier gene expression by lowering the burst frequency. In addition to more clearly defining the role of the chromatin environment in regulating the establishment of viral latency, this study has implications for the role of chromatin in modulating transcriptional noise in eukaryotes and its evolutionary consequences in the placement of genes within the genome.

Thus these studies of the role of sequence variation within the viral genome and its chromosomal integration site in regulating gene expression has resulted in better understanding of the mechanisms of gene expression and establishment of latency in HIV-1, while also helping to discern the role of chromatin in regulating mammalian gene expression.

To Ma, Baba and Dada

Table of Contents

Acknowledgements

I would like to express gratitude to several people with whom I have interacted and worked with and those who have assisted me in several ways during these years in graduate school.

I am deeply indebted to my parents for their support and encouragement through all the highs and lows of graduate studies. I am deeply grateful to by brother for the constant guidance and encouragement that he has always provided.

During the past six years I have had the chance to interact with several people in the Schaffer and Arkin laboratories who have helped me with various aspects of these projects. I would like to thank Morgan Price, Sharon Aviran, Marcin Joachimiak and Ron Skupsky from the Arkin laboratory with whom I have had several fruitful discussions about various computational aspects of these projects. I would like to thank John Burnett, Kathryn Miller-Jenson, Priya Shah and Jonathan Foley from the Schaffer laboratory with whom I have had helpful discussions on HIV biology. I would also like to thank Kwang Il-Lim, Shawdee Eshghi, Lukasz Bugaj and Randolph Ashton for other technical help. I would also like to thank other members of the Schaffer lab for their camaraderie and Wanichaya Ramey for administrative assistance.

Finally, I would like to thank Professor David Schaffer and Professor Adam Arkin for being wonderful mentors and role-models and guiding me though different stages of my graduate studies.

Chapter 1: Background and Motivation

1.1 Life cycle of Human Immunodeficiency Virus-1

Human Immunodeficiency Virus (HIV) is a member of the lentivirus subfamily of retroviruses. HIV carries its genetic information in the form of two single stranded (ss) RNA's that are slightly smaller than 10 kilobases in length. The viral genome contain 9 genes that encode for structural proteins (encoded by the Gag and Env genes), enzymes for important processes in its life cycle (like reverse transcriptase, integrase and protease encoded by the Pol gene), regulatory proteins (Tat and Rev), and auxiliary proteins (Nef, Vif, Vpr and Vpu). Both ends of the viral genome are flanked by Long Terminal Repeats (LTR) (Fig. 1.1*A*) (1, 2).

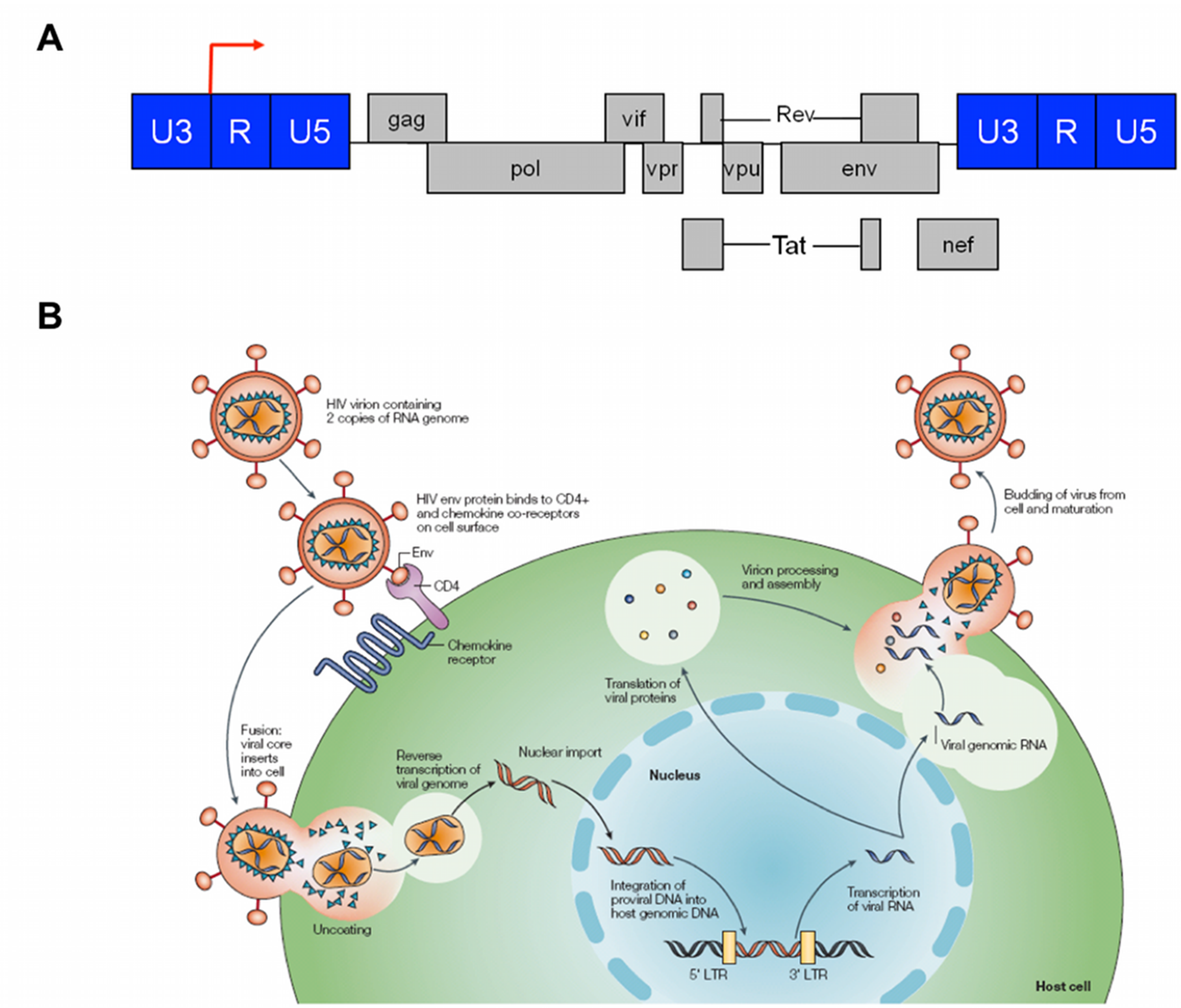

A
U3
R
U5
gag
vif
Rev
pol
vpr
vpu
env
U3
R
U5
Tat
nef
B
HIV virion containing 2 copies of RNA genome
HIV env protein binds to CD4+ and chemokine co-receptors on cell surface
Env
CD4
Chemokine receptor
Fusion: viral core inserts into cell
Uncoating
Reverse transcription of viral genome
Nuclear import
Nucleus
Integration of proviral DNA into host genomic DNA
5' LTR
3' LTR
Transcription of viral RNA
Translation of viral proteins
Virion processing and assembly
Viral genomic RNA
Budding of virus from cell and maturation
Host cell

Figure 1.1. Organization of the HIV-1 genome and viral life cycle. (A) HIV-1 genome showing the 9 viral genes flanked by the LTR (shown in blue) on both sides. (B) Life cycle of HIV-1 showing the major events such as viral entry, reverse transcription, integration into the host genome, transcription and translation followed by viral assembly and budding of progeny virus. Adapted from (1).

The first step in the life cycle of HIV-1 is the attachment of viral envelope glycoproteins to the CD4 receptor of T-cells, the primary cell target for this virus (Fig. 1.1*B*) (1). Viral entry into the cell also requires a chemokine co-receptor CXCR4 or CCR5 (3, 4). The viral envelope fuses with the cell membrane and releases the protein capsid into the cell. After uncoating, the viral enzyme, reverse transcriptase copies the viral genomic RNA into cDNA that is transported along with the viral protein integrase as part of the pre-integration complex to the nucleus where the integrase inserts the cDNA semi-randomly within the host genome (5-7) (Fig. 1.1*B*). Once integrated, the transcription factor binding sites on the 5'LTR recruit host cellular factors that mediate only a low, basal level of mRNA synthesis, or gene expression (8, 9). Initially, transcriptional elongation is inefficient and leads to the formation of a large number of abortive transcripts. However, this basal level of gene expression gives rise to a few rare full length genomic transcripts that are spliced by the host splicing machinery to produce the viral proteins, Tat (Transactivator of Transcription), Rev and Nef (10). Tat binds to a hairpin loop RNA structure called TAR (Transactivation Response Element) present at the 5' end of all HIV transcripts and greatly increases the processivity of RNA polymerase II (RNAP II), resulting in enhanced transcriptional elongation and the production of additional Tat and Rev protein (11-14). This strong positive feedback loop dramatically increases gene expression that drives the production of full length HIV transcripts. These full-length and partially spliced viral mRNA's all contain a structured RNA element called RRE (Rev Response Element) that binds to the viral protein Rev, which contains a nuclear export signal necessary to transport these intron-containing mRNAs to the cytoplasm (15-18). Once, in the cytoplasm, the partially spliced transcripts encode for the viral proteins Env, Vif, Vpr and Vpu, whereas the full length transcripts serve as templates for Gag and Gag-Pol as well as the viral genome for the progeny virus. Finally, the full length HIV transcripts are incorporated into new viral particles that bud off from the cell to infect the next generation of target cells (Fig. 1.1*B*).

1.2 HIV evolution and diversity

The reverse transcription process has low fidelity and unlike many cellular DNA polymerases, lacks the 3'→5' exonuclease proofreading activity, resulting in a high mutation rate of 0.2–0.3 mutations within its genome per replication cycle (5, 7). Since the virus carries two copies of ssRNA, the RT occasionally jumps from one strand to the other, giving rise to a high recombination rate of approximately 2.4 recombination events per replication cycle, along with insertions or deletions of the cDNA (7, 19, 20). When multiple virions infect a single cell, non-identical strands of HIV can get packaged into the progeny virus, and recombination between these non-identical strands during the next round of infection can give rise to additional diversity and new recombinant forms of the virus. The high error and recombination rates during reverse transcription, coupled with the large amounts of virus that are produced per day in a patient ($\sim 10^{10}$-10^{12}), allow the virus to effectively sample large parts of the sequence space, resulting in a huge sequence diversity of HIV-1 and allowing it to evolve rapidly (21, 22).

HIV was clinically first isolated in 1983 and the reconstruction of the evolutionary history of HIV-1 showed that its most common ancestor is a simian immunodeficiency virus (SIV) called SIVcpz, from which it diversified in the 1930s.The high error and recombination rate during reverse transcription, coupled with a very high replication rate has resulted in a huge global diversity of HIV. HIV-1 is phylogenetically categorized into three groups, 'M', 'N' and 'O' with group M being the most widespread form that is further classified into 9 subtypes and

15 recombinant forms (Fig. 1.2) (1). Depending on the genes being compared, the genetic distance between the different subtypes is between 10-30%, resulting in the different subtypes having widely differing gene regulation, pathogenesis, and transmission rates (23). Subtypes show differences in the elements involved in transactivation, such as differences in promoter architecture, diverse Tat protein sequences, and subtle differences in the TAR RNA that could result in different replication dynamics (24-27). Similarly, variation in Rev and RRE could impact the export of viral RNA to the cytoplasm and thus alter viral replication rates (28, 29). Within a single patient itself, there is a huge diversity of viral sequences that often exceeds the total global diversity of the influenza virus (30). This swarm of viral sequences is called a quasispecies (31, 32), which ensures that large parts of the sequence space is sampled by the virus, enabling the virus to rapidly evade the immune system and anti-retroviral therapies.

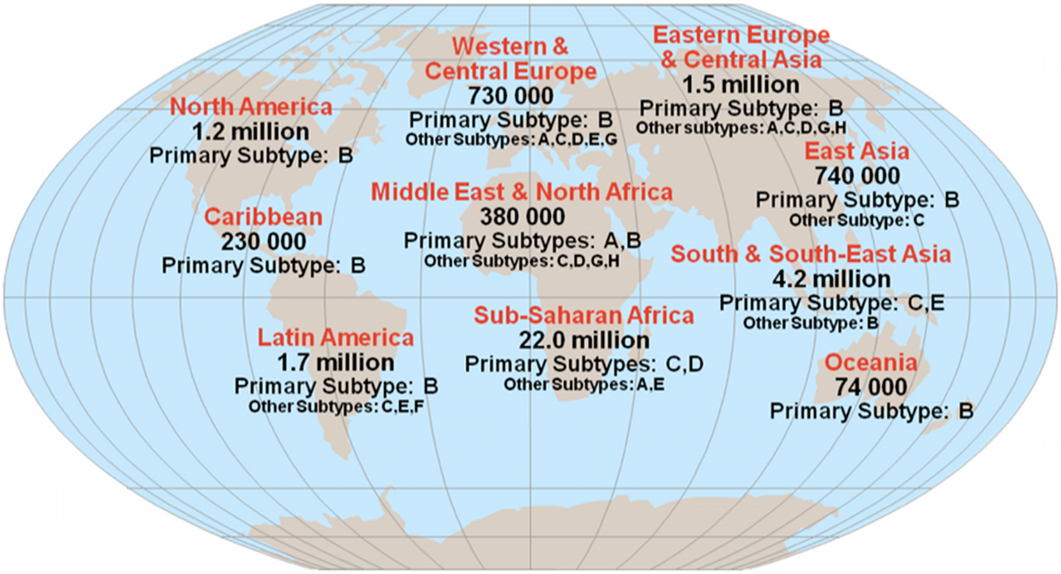
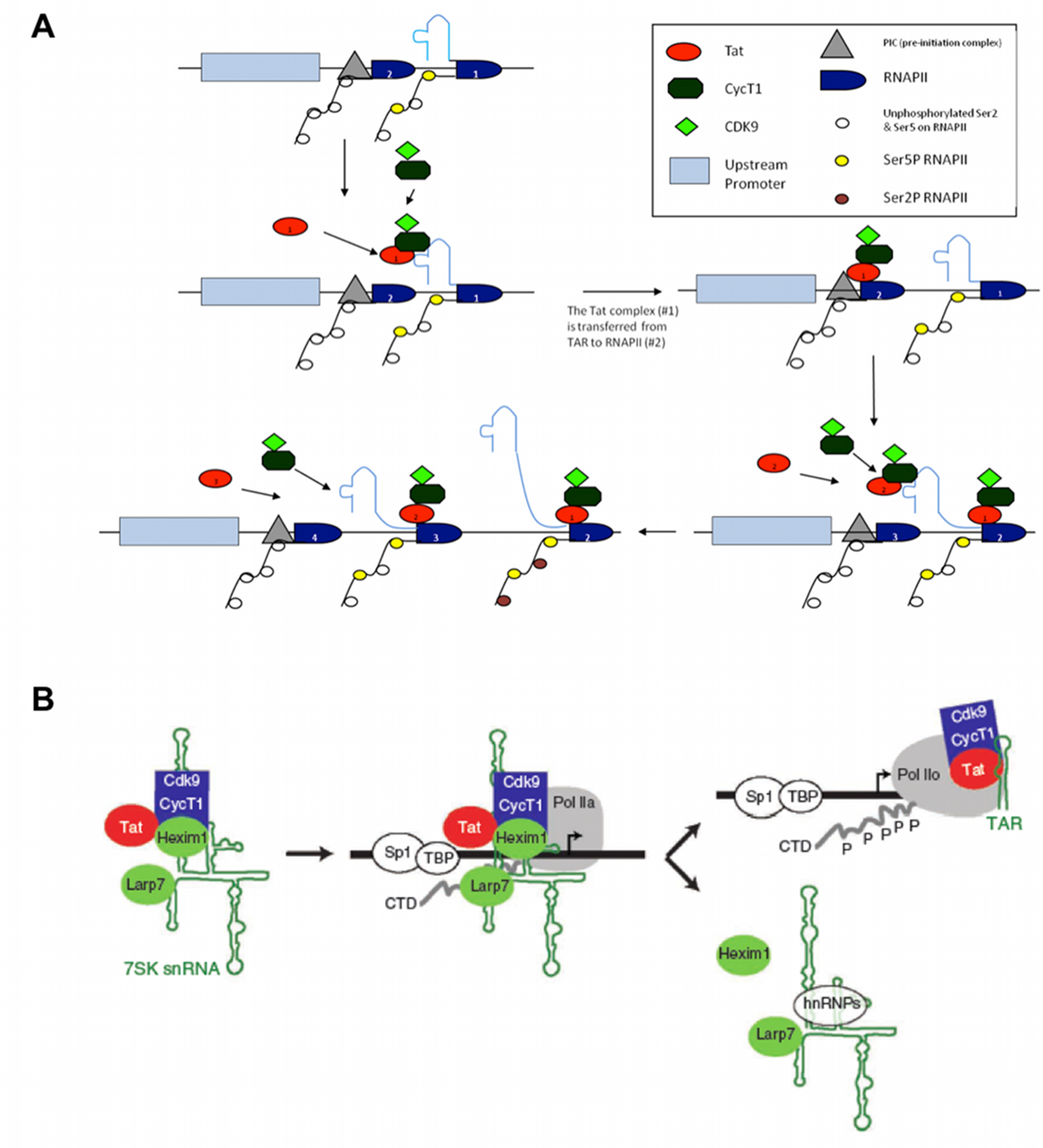


Figure 1.2. Distribution of HIV-1 subtypes in different parts of the world. Different subtypes and recombinant forms dominate various parts of the world. While subtype B remains the most studied subtype, it is not the most widespread form of the virus. Adapted from UNAIDS Report 2008.

Several small molecule drugs have been developed that interfere with different stages of the viral life cycle. There are four different classes of these drugs: those that inhibit viral fusion with the cell membrane, interfere with the process of reverse transcription, inhibit the enzyme integrase and prevent integration of the cDNA into the host genome, and those that inhibit the viral protease from cleaving viral precursor polypeptides into the proteins necessary to assemble a new virion (33). The current treatment for HIV-1 involves using a combination of these drugs, called as HAART (Highly Active Anti-Retroviral Therapy), to minimize the chances of viral escape. However, a number of drug-resistant strains have developed, so there is an urgent need for developing novel drugs. Identifying the underlying pattern of mutations within the viral genome could potentially be helpful in the rational design of novel therapies that minimize the chances of viral escape.

Although the viral proteins Tat and Rev show large sequence diversity, these different variants still function effectively to ensure efficient viral replication. The HIV-1 genome is constantly subject to mutations that allow it to rapidly adapt to selective pressures (34). Although certain mutations within a viral protein may help the virus escape the selective pressure, they may also reduce viral fitness. Purifying selection would purge such mutations that reduce the overall fitness of the virus; however compensatory mutations at other sites of the viral protein could compensate for such loss of fitness allowing correlated pairs of mutations to be fixed in the population (35). Identifying such correlated sites within different viral proteins will provide deeper understanding of the basic biology involving the structure and function of these proteins and their interaction with numerous cellular partners.



The Tat complex (#1) is transferred from TAR to RNAPII (#2)

Figure 1.3. Different proposed mechanisms of Tat-TAR mediated transactivation. (A) In this mechanism, Tat recruits P-TEFb to the viral promoter by binding to TAR. The Tat-P-TEFb complex is then transferred to the PIC, enabling P-TEFb to phosphorylate RNAPII and resulting in elongation. Newly synthesized viral RNA is shown in blue. (B) In this mechanism, P-TEFb is recruited as an inactive complex with 7SK. After recruitment of Tat to the promoter and transcription initiation, the TAR RNA displaces 7SK resulting in the activation of P-TEFb and efficient RNAPII phosphorylation leading to elongation. Adapted from (36, 37).

1.3 Gene regulation in HIV-1: Mechanism of Tat-TAR mediated transactivation

The exact molecular details of Tat-mediated viral transactivation are still being investigated. Briefly, the RNA hairpin TAR present at the 5' end of all viral transcripts, contains a two or three nucleotide bulge that is recognized by an arginine-rich motif (ARM, residues 49-57) in Tat (11, 12). This Tat-TAR binding is used to recruit other cellular proteins to the viral LTR that is necessary for transactivation (36). The few Tat molecules that are formed from basal transcription are acetylated at K28 by binding to a histone acetyl transferase (HAT) p300/CREB-binding protein associated factor (PCAF) (38). Ac28Tat has an increased affinity for the positive transcription-elongation factor b (P-TEFb), which consists of the cellular proteins Cyclin T1 (CycT1) and cyclin-dependent kinase 9 (Cdk9) (39-41). Binding to P-TEFb releases PCAF from Ac28Tat and the Tat-CycT1-Cdk9 complex then binds to TAR (Fig. 1.3*A*). Tat is then acetylated by another HAT p300/CREB binding protein (p300/CBP) at K50 which disrupts the Tat-pTEFb-TAR complex, and after dissociating with TAR, the Tat-pTEFb complex binds to PCAF (38, 42-44). This complex is then transferred to the pre-initiation complex (PIC) where Cdk9 phosphorylates the C-terminal domain (CTD) of RNAP II driving efficient transcription, resulting in transactivation (Fig. 1.3*A*) (45-47). More recently, in an alternate mechanism of Tat-mediated transactivation, it was shown that the P-TEFb complex is recruited to the promoter in an inactive form, bound to a small ribonucleoprotein (snRNP) called 7SK (Fig. 1.3*B*). Unlike the previous models, Tat assembles into the inhibitory 7SK snRNP along with P-TEFb and the formation of TAR after transcription initiation is used to displace to the 7SK snRNP which allows TAR to bind to the Tat-P-TEFb complex. Relieving P-TEFb from the inhibitory 7SK snRNP activates its kinase activity, resulting in the phosphorylation of the CTD of RNAPII and productive elongation (37) (Fig. 1.3*B*). Thus, gaining a better understanding of the structural constraints in Tat-TAR interactions in spite of their large sequence diversity and their interactions with various cellular factors is important for gaining better insights into Tat-mediated transactivation.

Tat is also post-translationally modified at specific sites by several cellular factors that impact transactivation. In addition to acetylation by PCAF and p300 as discussed above, there is evidence for methylation of Tat at Arg52 and Arg53 by the arginine methyltransferase PRMT6 that reduces Tat-mediated transcription (48, 49). In contrast, methylation of Tat at Lys51 by the lysine methyltransferase, Set 7/9 (KMT7) acts as an activator of gene expression (50). Similarly, other lysine methyltransferases like SETDB1 and SETDB2 have been shown to interact with Tat (51). Besides acetylation and methylation, Tat is also phosphorylated by Cdk2 at Ser16, Ser46 and by PKR at Ser62, Thr64, Ser68, enhancing viral gene expression (52, 53). Further evidence of the versatility of Tat can be seen in its interaction with other cellular proteins like SKIP/SNW1 and chromatin remodeling complexes like SWI/SNF that stimulate gene expression (54, 55). Besides these interactions with cellular factors that affect Tat-mediated transactivation, Tat has been implicated to interact with various factors involved in other functions. For instance, Tat has been shown to bind tubulin, enhance tubulin polymerization and induce mitochondrial apoptosis of cells that take up Tat (56, 57). Similarly, Tat has been shown to increase focal adhesion sites on brain microvascular endothelial cells (58).

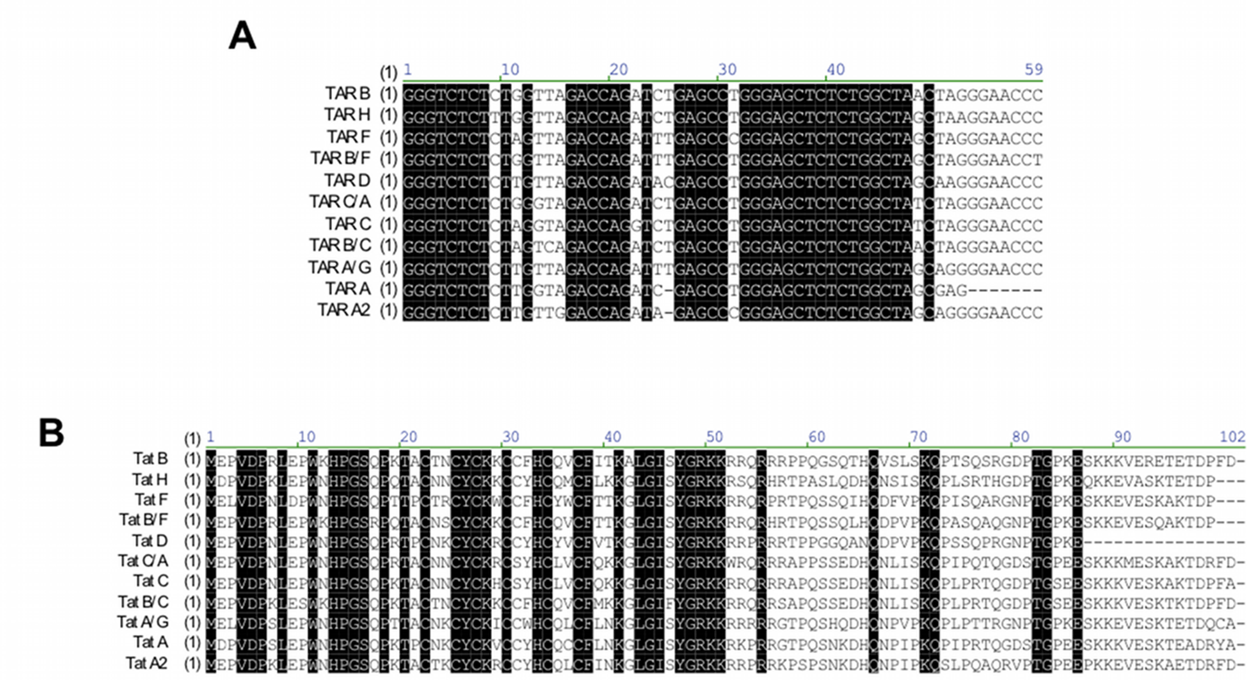

A
(1) 1 10 20 30 40 59
TAR B (1) GGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCC
TAR H (1) GGGTCTCTTTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAGCTAAGGAACCC
TAR F (1) GGGTCTCTCTAGTTAGACCAGATTTGAGCCCGGGAGCTCTCTGGCTAGCTAGGGAACCC
TAR B/F (1) GGGTCTCTCTGGTTAGACCAGATTTGAGCCTGGGAGCTCTCTGGCTAGCTAGGGAACCT
TAR D (1) GGGTCTCTCTTGTTAGACCAGATACGAGCCTGGGAGCTCTCTGGCTAGCAAGGGAACCC
TAR C/A (1) GGGTCTCTCTGGGTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTATCTAGGGAACCC
TAR C (1) GGGTCTCTCTAGGTAGACCAGGTCTGAGCCTGGGAGCTCTCTGGCTATCTAGGGAACCC
TAR B/C (1) GGGTCTCTCTAGTCAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCC
TAR A/G (1) GGGTCTCTCTTGTTAGACCAGATTTGAGCCTGGGAGCTCTCTGGCTAGCAGGGGAACCC
TAR A (1) GGGTCTCTCTTGGTAGACCAGATC-GAGCCTGGGAGCTCTCTGGCTAGCGAG-------
TAR A2 (1) GGGTCTCTCTTGTTGGACCAGATA-GAGCCCGGGAGCTCTCTGGCTAGCAGGGGAACCC
B
(1) 1 10 20 30 40 50 60 70 80 90 102
Tat B (1) MEPVDPRLEPWKHPGSQPKTACTNCYCKKCCFHCQVCFITKALGISYGRKKRRQRRRPPQGSQTHQVSLSKQPTSQSRGDPTGPKESKKKVERETETDPFD-
Tat H (1) MDPVDPKLEPWNHPGSQPQTACNNCYCKKCCYHCQMCFLKKGLGISYGRKKRSQRHRTPASLQDHQNSISKQPLSRTHGDPTGPKEQKKEVASKTETDP---
Tat F (1) MELVDPNLDPWNHPGSQPTTPCTRCYCKWCCFHCYWCFTTKGLGISYGRKKRRQRPRTPQSSQIHQDFVPKQPISQARGNPTGPKESKKEVESKAKTDP---
Tat B/F (1) MEPVDPRLEPWKHPGSRPQTACNSCYCKKCCFHCQVCFTTKGLGISYGRKKRRQRHRTPQSSQLHQDPVPKQPASQAQGNPTGPKESKKEVESQAKTDP---
Tat D (1) MEPVDPNLEPWNHPGSQPRTPCNKCYCKRCCYHCYVCFVTKGLGISYGRKKRRPRRRTPPGGQANQDPVPKQPSSQPRGNPTGPKE----------------
Tat C/A (1) MEPVDPNLEPWNHPGSQPRTACNNCYCKRCSYHCLVCFQKKGLGISYGRKKWRQRRRAPPSSEDHQNLISKQPIPQTQGDSTGPEESKKKMESKAKTDRFD-
Tat C (1) MEPVDPNLEPWNHPGSQPKTACNNCYCKHCSYHCLVCFQKKGLGISYGRKKRRQRRRAPQSSEDHQNLISKQPLPRTQGDPTGSEESKKKVESKAKTDPFA-
Tat B/C (1) MEPVDPKLESWKHPGSQPKTACTNCYCKKCCFHCQVCFMKKGLGIFYGRKKRRQRRSAPQSSEDHQNLISKQPLPRTQGDPTGSEESKKKVESKTKTDPFD-
Tat A/G (1) MELVDPSLEPWNHPGSQPTTACNKCYCKICCWHCQLCFLNKGLGISYGRKKRRRRRGTPQSHQDHQNPVPKQPLPTTRGNPTGPKESKKEVESKTETDQCA-
Tat A (1) MDPVDPSLEPWNHPGSQPKTPCNKCYCKVCCYHCQCCFLNKGLGISYGRKKRKPRRGTPQSNKDHQNPIPKQPIPRTQGDSTGPEESKKKVESKTEADRYA-
Tat A2 (1) MEPVDPKLEPWNHPGSQPKTACTKCYCKRCCYHCQLCFINKGLGISYGRKKRRPRRKPSPSNKDHQNPIPKQSLPQAQRVPTGPEEPKKEVESKAETDRFD-

Figure 1.4. Representative sequence alignment of TAR and Tat from different subtypes. Black columns represent completely conserved sites that are easily detected unlike coevolving sites. (A) Sequence alignment of representative TAR subtypes. (B) Sequence alignment of representative Tat subtypes.

Sequence alignment of a protein usually allows the easy identification of amino acids that are conserved and hence potentially allows detection of functionally important residues (Fig. 1.4). However, this approach misses cases where pairs of residues may coevolve, such that neither one is individually conserved, but specific pairs appear together and are important for maintaining the structural conformation of the protein or for its interaction with another protein. Given the large number of proteins with which the Tat-TAR complex and Tat interacts, and the various conserved sites that have been identified at which Tat is modified, we were interested in extending this analysis to identify functionally important coevolving positions between Tat-TAR and within Tat.

1.4 Thesis Goals and Organization

HIV-1 is a global pandemic and understanding how viral and cellular factors regulate gene expression and latency is central to identifying better treatment regimens or novel therapies to cure this disease. Although the development of antiretroviral drugs was a major advance in the treatment of patients infected with Human Immunodeficiency Virus-1 (HIV-1), complete eradication of HIV-1 has not been possible due to two major obstacles. First, the high mutation rate of the virus coupled with its rapid replication rate has given rise to drug resistant strains of HIV-1. Furthermore, latent viral reservoirs that are not directly targeted by anti-viral therapies or by the immune system can reactivate at a later time preventing complete viral clearance from a patient. Compounding these difficulties is the global diversification of viral strains or subtypes that have widely differing sequences, resulting in unique gene regulation and pathogenesis. Following integration into the host genome, activation of viral gene expression results in the production of new progeny whereas the inability to activate gene expression could initiate the establishment of viral latency. Thus, a better understanding of the mechanisms and factors that regulate viral transcription is critical towards eliminating latent viral populations. Therefore, the focus of my research projects has been to investigate the role of both cellular and viral factors in regulating HIV-1 gene expression and latency using a combination of computational and experimental techniques. This work may help develop novel therapy targets and better treatment regimens for different HIV-1 subtypes while concurrently providing new insights on mammalian gene regulation.

The first half of the thesis, chapters 2 through 5, focuses on various viral factors and mechanism by which these elements regulate gene expression and latency. In particular, we have focused on the role of sequence diversity within these different elements in differentially regulating gene expression and pathogenicity. In chapter 2, we have applied computational methods to mine large HIV-1 sequence databases to identify sites within the viral genome that may be coevolving and thereby gain better understanding of viral evolution and utilize this knowledge to piece together unknown mechanisms of gene regulation in HIV-1. In this chapter, we applied statistical methods to identify coevolving residues within Tat, between TAR and Tat, and within Rev. In chapter 3, we experimentally verified the sites in Tat that were predicted to be coevolving and showed that these sites may play a role in setting the rate of reactivation from latent viral populations. Further, we used the evolutionary information to gain deeper understanding of the mechanisms of Tat-mediated gene activation from the viral promoter. In contrast to previous studies, this study also provided new insights into protein evolution and showed that the coevolving sites we identified were unique in constraining two distinct transcriptional mechanisms critical in activating gene expression. In chapter 4, we studied how TAR and Tat sequences from different subtypes regulate gene expression and latency. While the interaction between TAR and Tat is critical in initiating strong gene expression, we found that the two viral elements act independently to set the level of gene expression. We also discovered that base pairs in TAR are not coevolving with resides in Tat. Since HIV-1 shows displays a high recombination rate, this study shows that the TAR-Tat axis in HIV-1 has evolved to be extremely robust such that a wide variety of TARs and Tats can interact with each other to efficiently activate viral gene expression. In chapter 5, we study another critical element, the viral genome in regulating viral gene expression, replication and latency. We discovered that promoters from different subtypes, in addition to having sequence diversity within transcription factors binding sites also have different architectures of binding sites that produces large differences in gene expression and propensities for latency between subtypes. We identified that the minimal set of transcription factor binding sites that contribute to most of the observed differences in gene expression at steady state. In contrast, we found that a combination of other transcription factor binding sites contribute to the dynamics of gene regulation, including the rates of reactivation from the latent state. Finally, we are currently using more clinically relevant primary cell culture models of HIV-1 latency, to probe differences in the propensity for latency between subtypes.

The latter half of the thesis, chapters 6 and 7, focuses on the role of cellular factors in regulating gene expression and latency in HIV-1. Since HIV-1 integrates semi-randomly within the human genome, we have studied the role of the cellular genomic location in regulating viral gene expression. In chapter 6, we exploited the semi-random integration of HIV-1 to quantitatively study both how latent proviruses can be reactivated from different chromatin environments and to address a fundamental question in eukaryotic gene expression related to how the placement of a gene in the genome impacts its responsiveness to an input transcription factor signal. We demonstrated that chromatin environments at different genomic loci decouple transcription factor mediated gene expression induction thresholds from subsequent gene activation. Using the functional relationship between gene expression, transcription factor levels, and chromatin accessibility, we accurately predicted synergistic HIV-1 activation in response to combinatorial pharmacological perturbations. Currently, we are using this system to identify drug regimens that maximize synergistic reactivation of latent HIV-1 populations such that these latent populations can be purged out most efficiently, enabling patients to completely eradicate the virus. Thus, this quantitative study should help inform strategies for combinatorial therapies to combat latent HIV-1 and help unravel biological principles underlying selective gene expression in response to transcription factor inputs. In chapter 7, we studied the role of the local chromatin environment around the HIV-1 promoter in regulating gene expression noise. Since it has previously been shown that gene expression noise could be one of many factors that contribute to the establishment and reactivation from latency, this chapter explores the origins and sources of this noise. We showed that increased levels of gene expression noise are associated with integrations into more heterochromatic regions. This increased gene expression noise, that could potentially be associated with greater propensity for latency, is associated with more infrequent transitions from the inactive to active promoter state. Thus, identifying mechanisms by which the frequency of transitions from the inactive to active promoter state could be increased could potentially reduce chances of establishment of viral latency or increase chances of reactivation thereby improving our chances of eliminating latent viral populations, the single greatest barrier to eradication of HIV-1 from an infected patient.

1.5 References

1. Rambaut A, Posada D, Crandall KA, & Holmes EC (2004) The causes and consequences of HIV evolution. *Nat Rev Genet* 5(1):52-61.
2. Kindt TJ, Goldsby RA, & Osborne BA (2007) *Kuby Immunology* (W.H. Freeman and Company) pp 504-521.
3. Feng Y, Broder CC, Kennedy PE, & Berger EA (1996) HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science* 272(5263):872-877.
4. Deng H*, et al.* (1996) Identification of a major co-receptor for primary isolates of HIV-1. *Nature* 381(6584):661-666.
5. Sarafianos SG*, et al.* (2009) Structure and function of HIV-1 reverse transcriptase: molecular mechanisms of polymerization and inhibition. *J Mol Biol* 385(3):693-713.
6. Delelis O, Carayon K, Saib A, Deprez E, & Mouscadet JF (2008) Integrase and integration: biochemical activities of HIV-1 integrase. *Retrovirology* 5:114.
7. Negroni M & Buc H (2001) Mechanisms of retroviral recombination. *Annu Rev Genet* 35:275-302.
8. Moses AV, Ibanez C, Gaynor R, Ghazal P, & Nelson JA (1994) Differential role of long terminal repeat control elements for the regulation of basal and Tat-mediated

transcription of the human immunodeficiency virus in stimulated and unstimulated primary human macrophages. *J Virol* 68(1):298-307.

9. Kao SY, Calman AF, Luciw PA, & Peterlin BM (1987) Anti-termination of transcription within the long terminal repeat of HIV-1 by tat gene product. *Nature* 330(6147):489-493.

10. Pollard VW & Malim MH (1998) The HIV-1 Rev protein. *Annu Rev Microbiol* 52:491-532.

11. Roy S, Delling U, Chen CH, Rosen CA, & Sonenberg N (1990) A bulge structure in HIV-1 TAR RNA is required for Tat binding and Tat-mediated trans-activation. *Genes Dev* 4(8):1365-1373.

12. Dingwall C*, et al.* (1990) HIV-1 tat protein stimulates transcription by binding to a U-rich bulge in the stem of the TAR RNA structure. *EMBO J* 9(12):4145-4153.

13. Brigati C, Giacca M, Noonan DM, & Albini A (2003) HIV Tat, its TARgets and the control of viral gene expression. *FEMS Microbiol Lett* 220(1):57-65.

14. Zhou Q & Yik JH (2006) The Yin and Yang of P-TEFb regulation: implications for human immunodeficiency virus gene expression and global control of cell growth and differentiation. *Microbiol Mol Biol Rev* 70(3):646-659.

15. Emerman M, Vazeux R, & Peden K (1989) The rev gene product of the human immunodeficiency virus affects envelope-specific RNA localization. *Cell* 57(7):1155-1165.

16. Felber BK, Hadzopoulou-Cladaras M, Cladaras C, Copeland T, & Pavlakis GN (1989) rev protein of human immunodeficiency virus type 1 affects the stability and transport of the viral mRNA. *Proc Natl Acad Sci U S A* 86(5):1495-1499.

17. Malim MH, Hauber J, Le SY, Maizel JV, & Cullen BR (1989) The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature* 338(6212):254-257.

18. Fischer U, Huber J, Boelens WC, Mattaj IW, & Luhrmann R (1995) The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell* 82(3):475-483.

19. Zhuang J*, et al.* (2002) Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots. *J Virol* 76(22):11273-11282.

20. Roda RH*, et al.* (2003) Role of the Reverse Transcriptase, Nucleocapsid Protein, and Template Structure in the Two-step Transfer Mechanism in Retroviral Recombination. *J Biol Chem* 278(34):31536-31546.

21. Perelson AS, Neumann AU, Markowitz M, Leonard JM, & Ho DD (1996) HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. *Science* 271(5255):1582-1586.

22. Najera R, Delgado E, Perez-Alvarez L, & Thomson MM (2002) Genetic recombination and its role in the development of the HIV-1 pandemic. *AIDS* 16 Suppl 4:S3-16.

23. Spira S, Wainberg MA, Loemba H, Turner D, & Brenner BG (2003) Impact of clade diversity on HIV-1 virulence, antiretroviral drug sensitivity and drug resistance. *J Antimicrob Chemother* 51(2):229-240.

24. van Opijnen T*, et al.* (2004) Human immunodeficiency virus type 1 subtypes have a distinct long terminal repeat that determines the replication rate in a host-cell-specific manner. *J Virol* 78(7):3675-3683.

25. Jeeninga RE*, et al.* (2000) Functional differences between the long terminal repeat transcriptional promoters of human immunodeficiency virus type 1 subtypes A through G. *J Virol* 74(8):3740-3751.

26. De Baar MP*, et al.* (2000) Subtype-specific sequence variation of the HIV type 1 long terminal repeat and primer-binding site. *AIDS Res Hum Retroviruses* 16(5):499-504.

27. Quivy V*, et al.* (2002) Synergistic activation of human immunodeficiency virus type 1 promoter activity by NF-kappaB and inhibitors of deacetylases: potential perspectives for the development of therapeutic strategies. *J Virol* 76(21):11091-11103.

28. Hua J, Caffrey JJ, & Cullen BR (1996) Functional consequences of natural sequence variation in the activation domain of HIV-1 Rev. *Virology* 222(2):423-429.

29. Phuphuakrat A & Auewarakul P (2003) Heterogeneity of HIV-1 Rev response element. *AIDS Res Hum Retroviruses* 19(7):569-574.

30. Walker BD & Burton DR (2008) Toward an AIDS vaccine. *Science* 320(5877):760-764.

31. Bull JJ, Meyers LA, & Lachmann M (2005) Quasispecies made simple. *PLoS Comput Biol* 1(6):e61.

32. Kamp C (2003) A quasispecies approach to viral evolution in the context of an adaptive immune system. *Microbes Infect* 5(15):1397-1405.

33. Pomerantz RJ & Horn DL (2003) Twenty years of therapy for HIV-1 infection. *Nat Med* 9(7):867-873.

34. Johnson VA*, et al.* (2009) Update of the drug resistance mutations in HIV-1: December 2009. *Top HIV Med* 17(5):138-145.

35. Camps M, Herman A, Loh E, & Loeb LA (2007) Genetic constraints on protein evolution. *Crit Rev Biochem Mol Biol* 42(5):313-326.

36. Bannwarth S & Gatignol A (2005) HIV-1 TAR RNA: the target of molecular interactions between the virus and its host. *Curr HIV Res* 3(1):61-71.

37. D'Orso I & Frankel AD (2010) RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. *Nat Struct Mol Biol* 17(7):815-821.

38. Kiernan RE*, et al.* (1999) HIV-1 tat transcriptional activity is regulated by acetylation. *EMBO J* 18(21):6106-6118.

39. Bieniasz PD, Grdina TA, Bogerd HP, & Cullen BR (1998) Recruitment of a protein complex containing Tat and cyclin T1 to TAR governs the species specificity of HIV-1 Tat. *EMBO J* 17(23):7056-7065.

40. Wei P, Garber ME, Fang SM, Fischer WH, & Jones KA (1998) A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA. *Cell* 92(4):451-462.

41. Bieniasz PD, Grdina TA, Bogerd HP, & Cullen BR (1999) Recruitment of cyclin T1/P-TEFb to an HIV type 1 long terminal repeat promoter proximal RNA target is both necessary and sufficient for full activation of transcription. *Proc Natl Acad Sci U S A* 96(14):7791-7796.

42. V. Bres HT, J. Peloponese, E. Loret, K. Jeang, Y. Nakatani, S. Emiliani, M. Benkirane, R. E. Kiernan (2002) Differential acetylation of Tat coordinates its interaction with the co-activators cyclin T1 and PCAF. *EMBO J* 21(24):6811-6819.

43. Benkirane M*, et al.* (1998) Activation of integrated provirus requires histone acetyltransferase. p300 and P/CAF are coactivators for HIV-1 Tat. *J Biol Chem* 273(38):24898-24905.

44. Marzio G, Tyagi M, Gutierrez MI, & Giacca M (1998) HIV-1 tat transactivator recruits p300 and CREB-binding protein histone acetyltransferases to the viral promoter. *Proc Natl Acad Sci U S A* 95(23):13519-13524.

45. Zhu Y*, et al.* (1997) Transcription elongation factor P-TEFb is required for HIV-1 tat transactivation in vitro. *Genes Dev* 11(20):2622-2632.

46. Yang X, Herrmann CH, & Rice AP (1996) The human immunodeficiency virus Tat proteins specifically associate with TAK in vivo and require the carboxyl-terminal domain of RNA polymerase II for function. *J Virol* 70(7):4576-4584.

47. Chun RF & Jeang KT (1996) Requirements for RNA polymerase II carboxyl-terminal domain for activated transcription of human retroviruses human T-cell lymphotropic virus I and HIV-1. *J Biol Chem* 271(44):27888-27894.

48. Boulanger MC*, et al.* (2005) Methylation of Tat by PRMT6 regulates human immunodeficiency virus type 1 gene expression. *J Virol* 79(1):124-131.

49. Xie B, Invernizzi CF, Richard S, & Wainberg MA (2007) Arginine methylation of the human immunodeficiency virus type 1 Tat protein by PRMT6 negatively affects Tat Interactions with both cyclin T1 and the Tat transactivation region. *J Virol* 81(8):4226-4234.

50. Pagans S*, et al.* (2010) The Cellular lysine methyltransferase Set7/9-KMT7 binds HIV-1 TAR RNA, monomethylates the viral transactivator Tat, and enhances HIV transcription. *Cell Host Microbe* 7(3):234-244.

51. Van Duyne R*, et al.* (2008) Lysine methylation of HIV-1 Tat regulates transcriptional activity of the viral LTR. *Retrovirology* 5:40.

52. Ammosova T*, et al.* (2006) Phosphorylation of HIV-1 Tat by CDK2 in HIV-1 transcription. *Retrovirology* 3:78.

53. Endo-Munoz L, Warby T, Harrich D, & McMillan NA (2005) Phosphorylation of HIV Tat by PKR increases interaction with TAR RNA and enhances transcription. *Virol J* 2:17.

54. Bres V, Yoshida T, Pickle L, & Jones KA (2009) SKIP interacts with c-Myc and Menin to promote HIV-1 Tat transactivation. *Mol Cell* 36(1):75-87.

55. Mahmoudi T*, et al.* (2006) The SWI/SNF chromatin-remodeling complex is a cofactor for Tat transactivation of the HIV promoter. *J Biol Chem* 281(29):19960-19968.

56. Chen D, Wang M, Zhou S, & Zhou Q (2002) HIV-1 Tat targets microtubules to induce apoptosis, a process promoted by the pro-apoptotic Bcl-2 relative Bim. *EMBO J* 21(24):6801-6810.

57. Campbell GR*, et al.* (2004) The glutamine-rich region of the HIV-1 Tat protein is involved in T-cell apoptosis. *J Biol Chem* 279(46):48197-48204.

58. Avraham HK, Jiang S, Lee TH, Prakash O, & Avraham S (2004) HIV-1 Tat-mediated effects on focal adhesion assembly and permeability in brain microvascular endothelial cells. *J Immunol* 173(10):6228-6233.

Chapter 2: Identifying Coevolving Sites within HIV-1 using Mutual Information

2.1 Introduction - Computational methods for identifying coevolving sites

Due to the increasing ease and reducing costs involved with sequencing, the database of HIV-1 sequences from infected patients is constantly growing. Such multiple sequence alignments can be used to predict and identify correlated sites within a protein that may be important for maintaining the function and/or structure of the protein. Several methods, including Mutual Information (MI), Observed Minus Expected Squared (OMES), Statistical Coupling Analysis (SCA) and McLachlan Based Substitution Correlation (McBASC) have been used to identify correlated position pairs from multiple sequence alignments of proteins (1-9). MI between two sites uses the well-known Shannon's entropy to estimate the reduction of uncertainty in the identity of amino acids at a particular site, given the identity of amino acids at another site, and is discussed in greater detail in the next section. OMES calculates correlation between two sites by using a measure that estimates the deviation between the observed and expected frequency of different pairs of amino acids at those sites (1). SCA identifies correlated sites by creating "perturbations" in the form of sub-alignments that consist of a particular amino acid at a site which is then compared to the original alignment to detect if there are any differences in the amino acid composition for different sites (3). McBASC computes a matrix of scores for each site in the alignment using a substitution rule, where scores of each element in the matrix corresponds to amino acids from two sequences at that site. Comparison of elements from two matrices, corresponding to two sites in the protein sequence alignment is then used to compute correlation scores in this method (2, 9). A difference between the McBASC and the other methods is that McBASC tends to give high correlation scores for a pair of sites even when both sites are highly conserved. Previously, such methods have been used to identify coevolving residues in HIV-1 for the proteins encoded by the *gag* gene and V3 loop of the *env* gene (6, 10). Further, these methods can easily be extended to DNA or RNA sequence alignments, as will be shown in this work later, and can be used to predict correlated sites between RNA bases and proteins amino acids that functionally interact with each other.

MI was first developed in the area of Information Theory, and has since then been used in computational biology to identify coevolving residues within the same protein or two interacting proteins (4, 11, 12). MI belongs to a family of methods that does not require structural or phylogenetic data for predicting coevolving sites. As compared to other methods briefly described above, MI initially detects a large number of pairs that are predicted to be coevolving (13). However, further application of methods to reduce background noise in the MI analysis allows for accurate estimation of the strongly correlated positions. In this work, we used the statistical measure MI to identify position pairs that were strongly correlated.

2.2 Predicting coevolving sites using Mutual Information

To predict the extent of correlation between two sites in a multiple sequence alignment using MI (4, 11), the degree to which each individual site is conserved needs to be initially evaluated using Shannon's entropy. Entropy gives the extent of uncertainty within a site and as is evaluated as follows:

$$H(X) = -\sum_{i=1}^{K} p(x_i)\log_b p(x_i)$$

where *H(X)* stands for the entropy of site *X*, and *X* ε (x_1, x_2……x_k) with K being 21 and 5 for a protein and RNA sequence alignment including gaps, respectively. $p(x_i)$ is the probability of observing residue x_i at site *X*. Thus, the entropy of a site that is completely conserved is zero. In contrast, a site for which the probability of observing any residue is the same, that is the amino acids are picked from a uniform random distribution, has maximum entropy of $\log_b K$. Choice of *b* is arbitrary and for all our calculations, *b = 2*. Similarly, the joint entropy, *H(X,Y)* between sites *X* and *Y* in a sequence alignment is given by:

$$H(X,Y) = -\sum_{i=1}^{K}\sum_{j=1}^{L} p(x_i, y_i)\log_b p(x_i, y_i)$$

where $p(x_i,y_i)$ is the joint probability of observing residues x_i and y_i at sites *X* and *Y,* respectively, for a given sequence. Based on these definitions, Mutual Information between sites *X* and *Y*, *MI(X,Y)*, a measure of the reduction in the entropy of site *X* given the identity of the residues at site *Y*, is given by:

$$\begin{aligned} MI(X,Y) &= H(X) - H(X \mid Y) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

Further, MI is commutative and thus, $MI(X,Y) = MI(Y,X)$.

MI gives the likelihood of observing a residue at a particular site in the multiple sequence alignment, given the identity of a residue at another site in the alignment. Thus, two sites that are constrained to certain pairs of amino acids due to structural or functional requirements of the protein will also result in higher MI scores between those sites. Thus, MI can be used to distinguish between functionally important non-conserved sites that are coevolving from other non-conserved sites. However, detection of sites that are actually coevolving is made harder due to background noise that arises from two sources; the finite size of the sequence databases and a phylogenetic contribution to noise (11). The finite size of the sequence database results in the probabilities used in the estimation of the entropies and joint entropies to be approximated by frequencies of amino acids at a particular site. Thus, alignments with larger number of sequences results in lower background noise due to the finite size of the sequence database. The phylogenetic contribution of noise arises due to the evolutionary history shared between different but closely-related sequences in the alignment. This shared ancestral history between sequences gives rise to apparent correlation between sites rather than a functional or structural context that constrains residues at those sites. Thus, to minimize the background noise that arises from these two sources, several methods have been proposed that help to filter out the background noise from the raw MI scores, some of which are explained below and applied later to the HIV-1 datasets:

1. Relatively conserved sites may have low MI scores even if it is coevolving with another site, due to the following constraint (4, 11):

$$0 \leq MI(X,Y) \leq \min\left\{H(X), H(Y)\right\}$$

Thus, a conserved site may have low MI scores even if it is coevolving with another site because the mutual information is constrained to lie below the smaller entropy of the lesser variable site. Since the MI between two sites is constrained by the smaller of the two entropies, sites that have high entropy also give rise to higher background MI. Normalizing the raw MI score with the smaller of the two entropies for these sites, could potentially help reduce background noise.

2. The joint entropy between two sites share the following relation with the entropy of the two sites:

   $$\max\left\{H(X), H(Y)\right\} \leq H(X,Y) \leq H(X) + H(Y)$$

   Thus, this relation also allows the raw MI scores to be normalized by the joint entropy between the two sites to improve the predictive ability of this method:

   $$MI(X,Y) \leq \min\left\{H(X), H(Y)\right\} \leq \max\left\{H(X), H(Y)\right\} \leq H(X,Y)$$

3. In this method, a random sequence alignment with the same number of sequences as the actual dataset is created in which the amino acid frequency at each site is based on a uniform distribution. Background MI scores computed for this simulated alignment is then subtracted from the raw MI scores of the actual dataset to correct for the background MI that arises due to random associations between sites and those due to the finite size of the sequence database.

4. Since the MI score is a function of the entropies of the individual sites, an improvement over the previous method is to create simulated alignments in which the amino acid frequency at each site is the same as that in the actual sequence database. This way, the entropy of each site is similar to that in the actual dataset and gives a better estimate of the background MI. Again, as in the previous method, subtracting the raw MI scores from the background MI is used to compute the corrected MI score.

5. To further minimize background noise arising from phylogenetic relations between sequences besides the contribution from the finite size of the sequence database, Dunn *et. al.* proposed a correction term, called average product correction (APC), which correlated well with background MI and subtracting the APC from the raw MI gave good estimate of strongly correlated sites by minimizing the influence arising due to shared ancestry and finite dataset sizes (14). APC is given by:

   $$APC(X,Y) = \frac{MI(X,\bar{i})MI(\bar{i},Y)}{\overline{MI}}$$

where, $MI(X,\bar{i}) = {}^{1}\!\!\diagup_{(n-1)} \sum_{i=1,i\neq X}^{n} MI(X,i)$ is the mean MI between site X and all other sites in the alignment and n is the number of sites in the alignment. $\overline{MI}$ is the mean MI between all sites in the alignment.

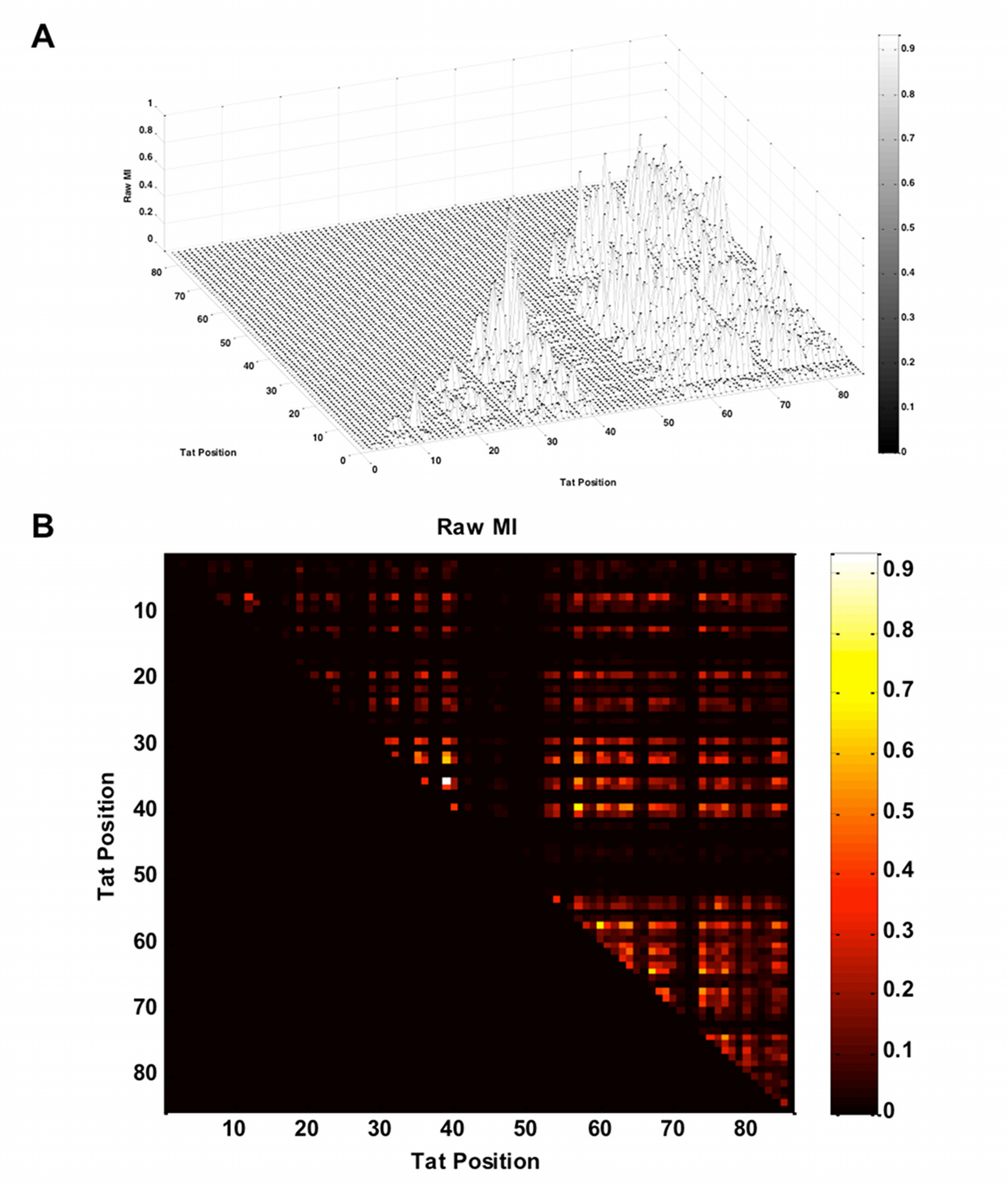


Figure 2.1. Raw MI plots for all positions within the first 86 amino acids of Tat (A) Mesh plot showing raw MI scores for position in Tat. (B) Heat Map of raw scores for position in Tat. Position pair 35-39 have the highest coevolution signal.

2.3 Using MI to identify coevolving residues within Tat

To identify correlated sites within Tat, we applied MI analysis to 917 pre-aligned Tat sequences obtained from the Los Alamos Sequence Database (http://www.hiv.lanl.gov/). The raw MI scores for all position pairs are shown as a 3-dimentional mesh and heat map (Fig. 2.1*A* and 2.3*B*). The raw MI scores show that position pairs 35 and 39, within the cysteine-rich region of Tat, have the highest MI score. The amino acids between residues 41-52, partly within the core and ARM domain of Tat, show very low MI scores with all other positions in Tat, possibly since this region is highly conserved across Tat sequences from different subtypes. Beyond amino acid 52, the MI landscape is rugged suggesting that there could be significant correlation between sites in the C-terminal end of Tat or that these could arise due to increased background noise since these sites also show higher amino acid variability or entropy.

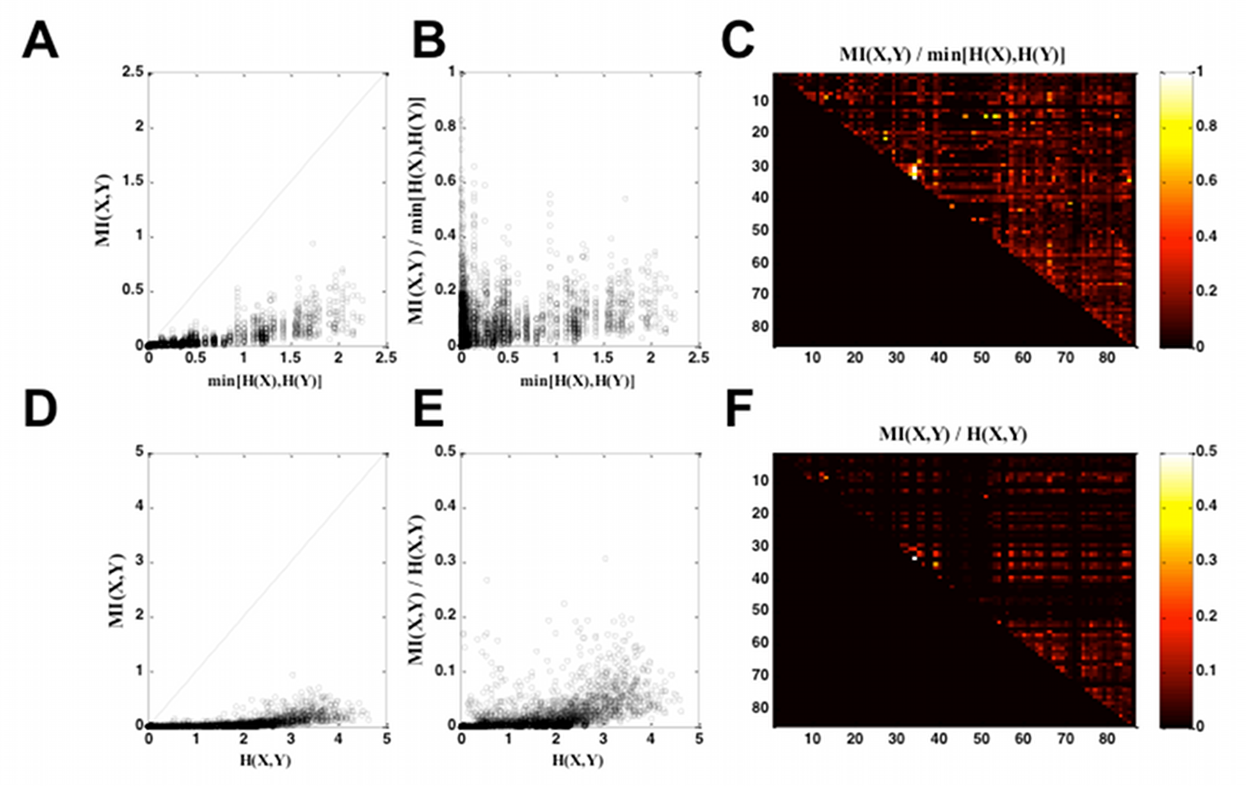


Figure 2.2. (A) Plot of MI(*X,Y*) vs.min{*H(X),H(Y)*} shows that the site pair 35-39 is closer to the theoretical limit corresponding to the diagonal line than other points. (B) Plot of MI(*X,Y*) /min{*H(X),H(Y)*} vs. min{*H(X),H(Y)*}. (C) Heat map of correlation between different sites in Tat after normalization with min{*H(X),H(Y)*}. (D) Plot of MI(*X,Y*) vs. H(*X,Y*) also shows that the site pair 35-39 lies above all other pairs, closer to the diagonal line. (E) Plot of MI(*X,Y*)/H(*X,Y*) vs. H(X,Y). (F) Heat map of correlation between different sites in Tat after normalization with H(*X,Y*).

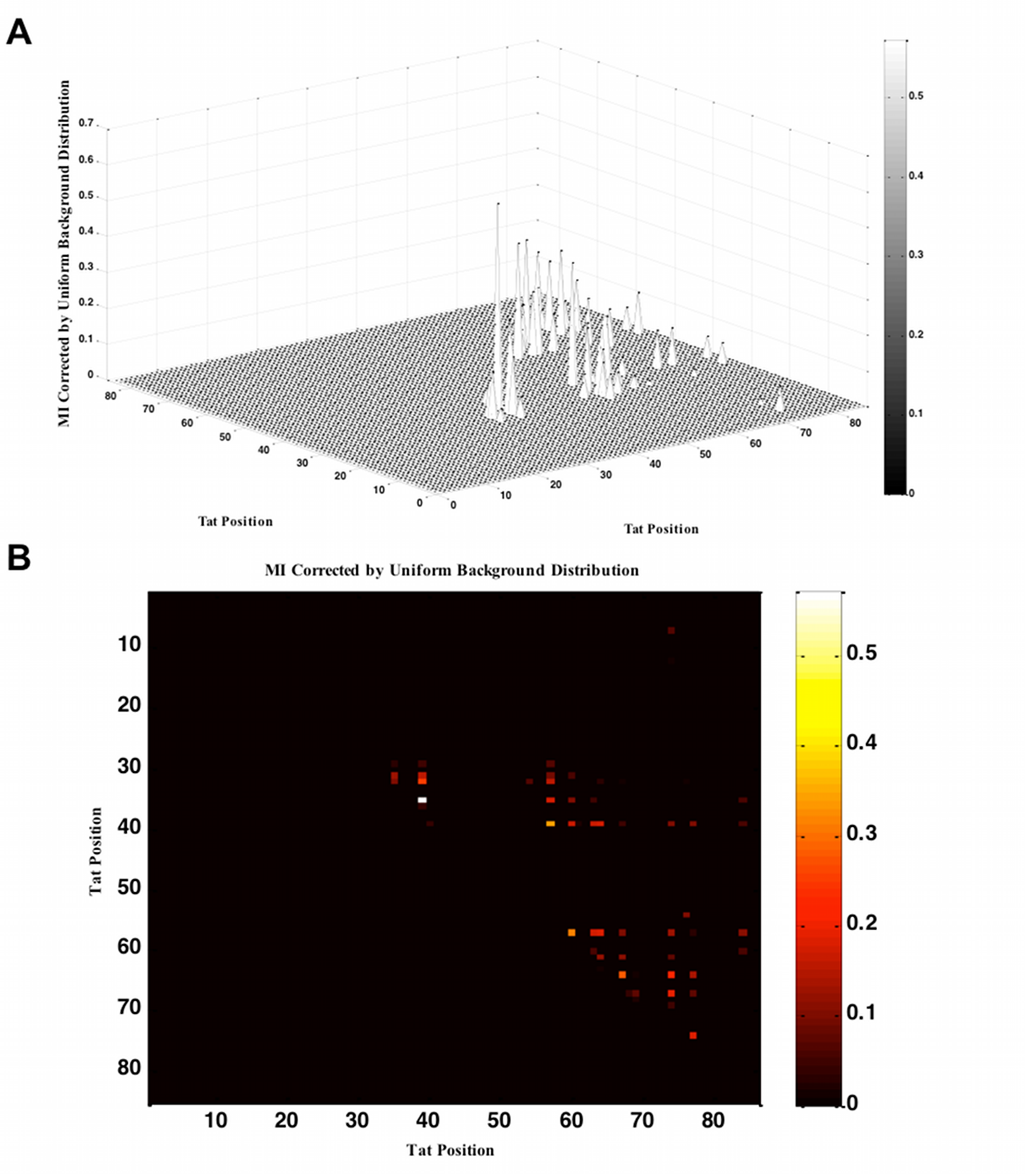


Figure 2.3. MI scores shown as (A) mesh plot and (B) heat map after substracting background MI generated from a simulated sequence alignment with equiprobable amino acid distribution.

To distinguish signal from noise, we initially normalized the raw MI scores with min$\{H(X),H(Y)\}$, as discussed previously. Since this quantity represents the maximum MI score that can be attained between a pair of sites, this normalization should help identify sites that are conserved but still coevolving. Thus, plotting MI(X,Y) vs. min$\{H(X),H(Y)\}$ should help identify position pairs that are coevolving since such pairs should be close to the diagonal line on this plot (Fig. 2.2*A*).

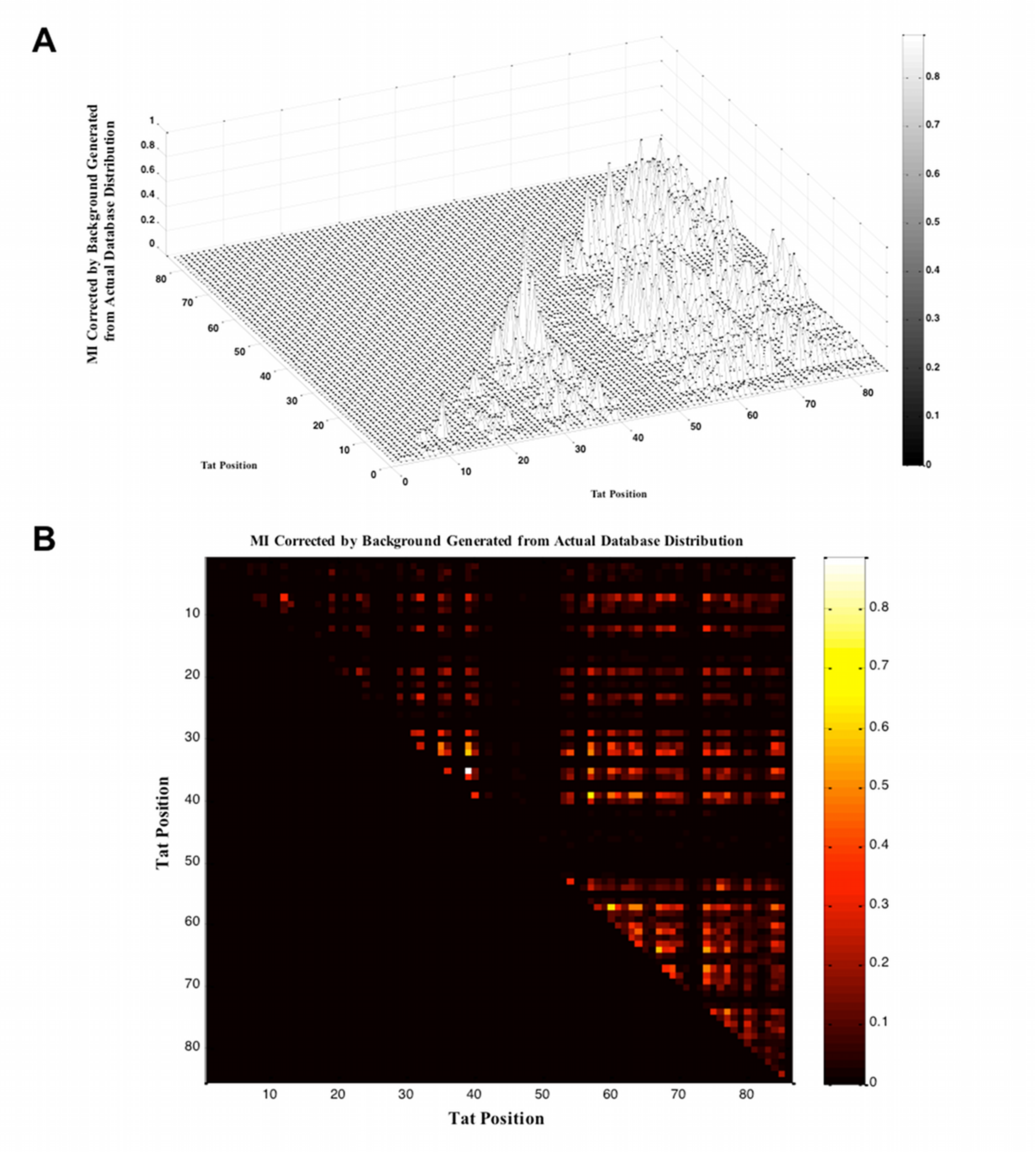


Figure 2.4. MI scores depicted as (A) mesh plot and (B) heat map after subtracting background MI generated from a simulated alignment with the same amino acid frequency at each site as the actual database.

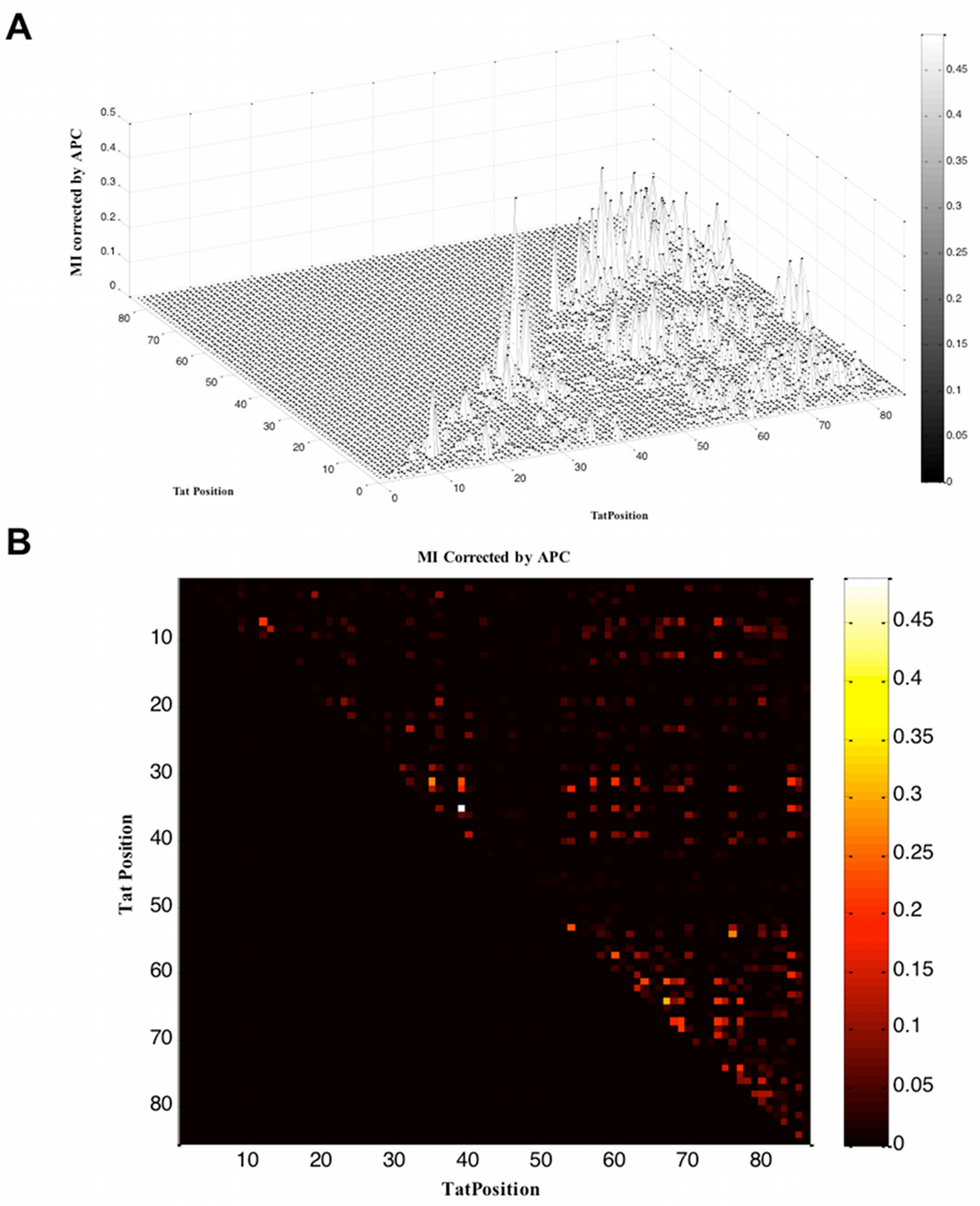


Figure 2.5. MI scores depicted as (A) mesh plot and (B) heat map after APC correction.

Again, we see that the position pair (35,39) is distinctly above all other site pairs, suggesting that this two positions may be coevolving and hence possibly critical for maintaining protein structure or function. However, plotting the normalized quantity, MI(*X,Y*)/min{*H(X),H(Y)*} vs. min{*H(X),H(Y)*} shows that this method may also be introducing false positives in the analysis since several sites that that have very low entropies, upon normalization with min{*H(X),H(Y)*} give rise to high normalized values although such sites are possibly not coevolving (Fig. 2.2*B*). Thus, it appears that although this normalization can help identify correlated positions within sites that are relatively conserved, this method does not work well for sites that are very highly conserved (and hence have very low entropy). The heat map further demonstrates this by showing that this normalization even results in some coevolution signal for sites within 41-52 with other sites in the protein suggesting that this normalization may not be suited for sites that are highly conserved (Fig 2.2*C*). Similarly, plotting MI(*X,Y*) vs. H(*X,Y*) shows that the position pair (35,39) is above all other pairs (Fig. 2.2*D*). Plotting the normalized quantity, MI(*X,Y*)/H(*X,Y*) again shows that this normalization may introduce false positives for positions pairs that are highly conserved with very small joint entropies (Fig. 2.2*E*). The heat map of MI(*X,Y*)/H(*X,Y*) shows that the mutual information landscape is qualitatively similar to the raw MI landscape (Fig. 2.2*F*).

Next, correcting the raw MI scores using a simulated alignment based on a random sequence alignment results in no correlation between most sites with the site pair (35,39) still having the highest corrected score (Fig. 2.3). A plausible reason for the reduction of correlation alignment have entropies that are close to the maximum entropy of 4.39 (= $\log_2 21$), that gives rise to a higher estimate of the background MI since most actual sites have much lower entropies and hence have smaller contributions to background MI. Thus, it appears that using a random sequence alignment for computing the background MI score results in its overestimation, thus eclipsing some functional/structural correlation that may exist between certain sites. To correct for the overestimation of the background MI, simulated sequence alignments were created that have the same amino acid distribution at each site (and therefore the same entropy) as that in the actual sequence database. The corrected MI appears to show a more realistic estimate of the correlation between all possible position pairs in Tat by accounting for the background noise that arises from the finite size of the sequence alignment (Fig. 2.4). Again sites 35 and 39 show the highest correlation with sites within 41-52 showing almost no correlation with any other site in Tat and sites beyond 52 showing significant correlation with other sites in the entire protein.

Finally, using the APC correction to minimize background noise arising from phylogenetic and finite sample size effects shows that the correlation between several sites is reduced dramatically while retaining high MI scores for certain other pairs. Within the activation domain of Tat, position pairs (35,39), (35,31), (31,39) and (7,12) show significant correlation; whereas correlation of several sites beyond amino acid 52 with other sites in the protein is reduced dramatically, suggesting that the high MI scores that were seen for such sites in the raw MI estimation possibly arose from background noise that was not taken into consideration (Fig. 2.5). Again, site pair (35,39) is the global maxima suggesting that this pair is coevolving and possibly critical for protein function (Fig. 2.5). Analyses of frequencies of amino acids at these sites show that a Leu at position 35 constrains position 39 primarily to a Gln and a Gln at position 35 results in a majority of Tat sequences having an Ile, Leu or Thr but not a Gln at position 39.

To identify a threshold corrected MI score to identify position pairs that are strongly correlated and hence possibly coevolving from pairs that do not interact with each other; we used a method described by Weigt *et. al.*, wherein a histogram of the corrected MI scores was constructed and fitted to an exponential function (15). Deviation from the exponential function was used to identify the threshold MI score. This method was applied to corrected MI scores obtained after normalization with background MI with the same amino acid frequency as the actual dataset, for which the threshold was 0.34 (Fig. 2.6*A*), or after APC correction, for which the threshold was 0.21 (Fig. 2.6*B*).

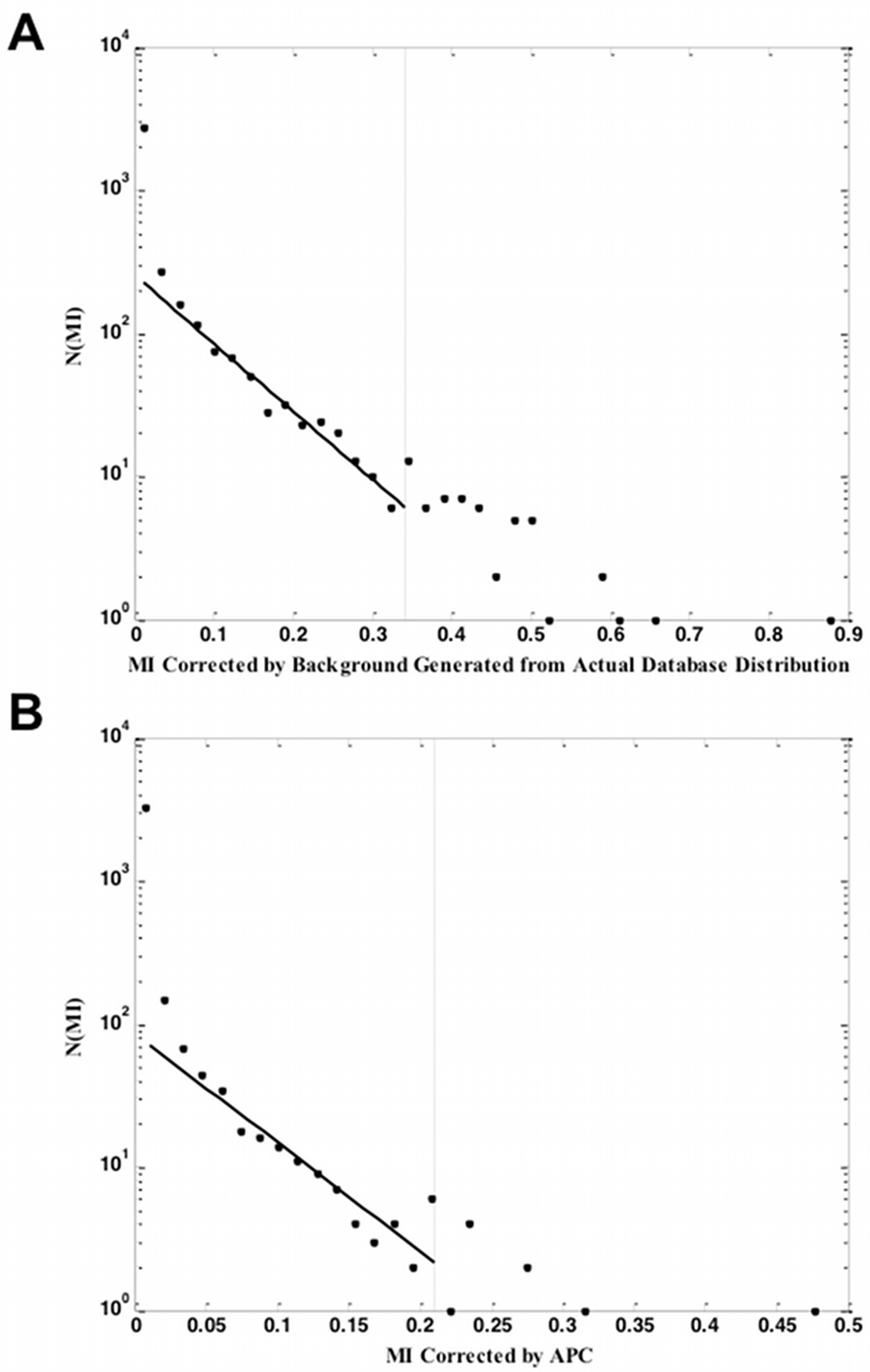


Figure 2.6. Identification of a threshold MI to distinguish correlated position pairs from other pairs. MI scores corrected by (A) actual database distribution and (B) APC for all position pairs in Tat are plotted against their frequency of occurrences. The threshold MI score is indicated by the dashed line. Points to the left of the dashed line indicate the exponential background. Deviation from the exponential function is used to identify the threshold MI.

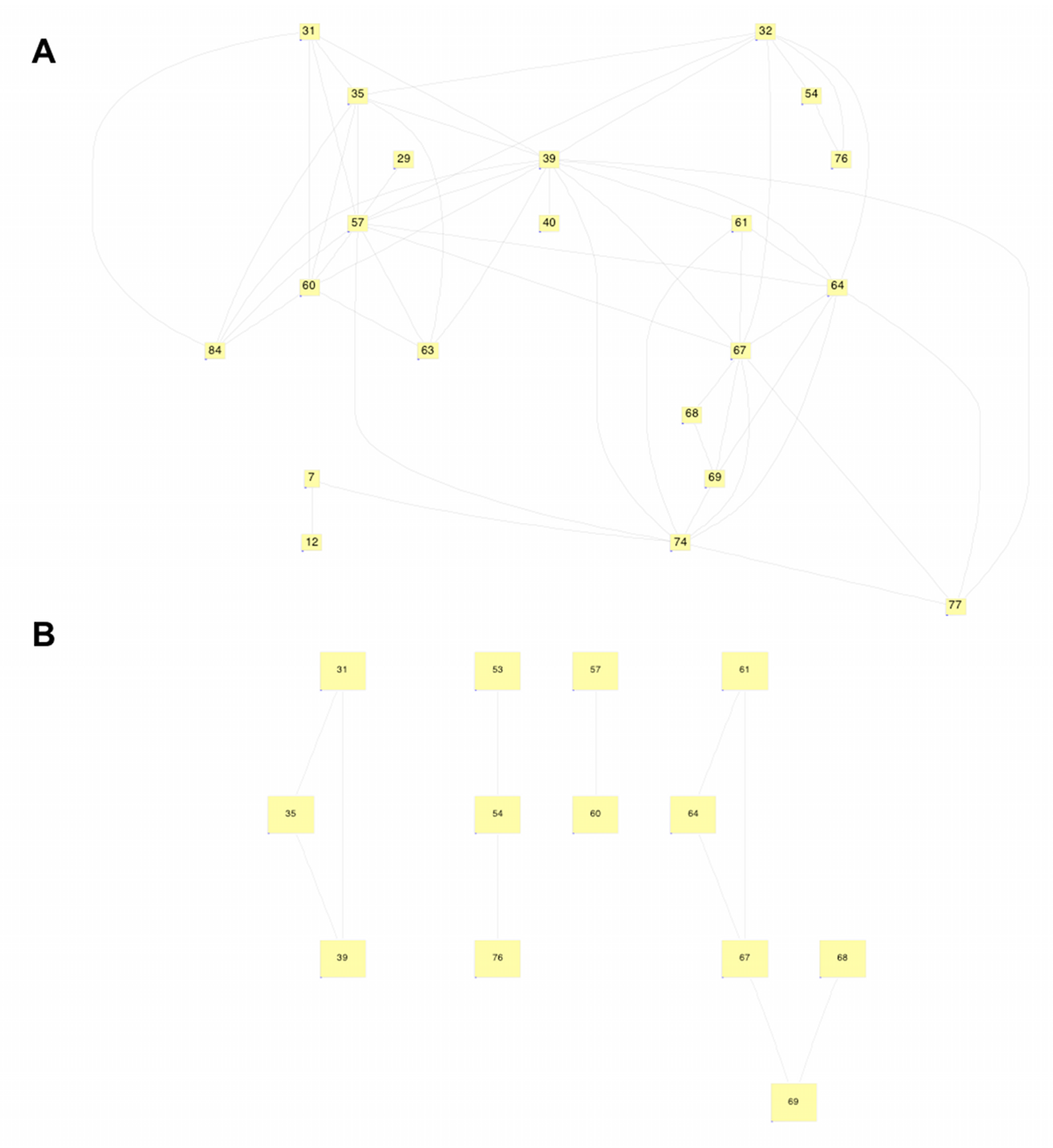


Figure 2.7. Network of position pairs above threshold MI. (A) Position pairs above threshold MI after background correction assuming the same amino acid distribution as the actual database distribution. (B) Position pairs above threshold MI after background correction by APC.

These threshold scores were then used to construct a network of interaction for all position pairs that scored above the threshold MI. For the corrected scores normalized by background MI with the same amino acid frequency as the actual dataset, 20 sites had scores above threshold with sites 39, 57, 67, 74 and 35 connected to several other sites, suggesting that these sites may be functionally or structurally important and critical for Tat function (Fig. 2.7*A*). For MI scores above threshold after APC correction, we see that there are fewer sites pairs that are above threshold with smaller independent cliques of sites that are connected to each other (Fig. 2.7*B*). 13 sites are above the threshold MI score. This normalization method suggests that the different independent cliques may be important for the distinct functions of Tat. For e.g., sites 31-35-39 that are correlated to each other, within the activation domain of Tat, which has been shown to be important for binding P-TEFb, may be critical for performing this function in Tat. Similarly, the correlation between sites 53-54-76 and 57-60, most of which are within the ARM motif of Tat, may play an important role in Tat nuclear localization or TAR binding. The correlation between sites 61-64-67-68-69 may be functionally important for interaction of Tat with PKR which induces phosphorylation of Tat at sites 62, 64 and 68. The presence of these correlated sites within the Gln-rich region of Tat may also imply that these sites could be important for Tat-mediated mitochondrial apoptosis of T-cells.

Thus, correction of raw MI scores with these two methods show one important distinction – the APC correction displays that the correlation between sites is localized and modular in nature, suggesting that they may be important for distinct functions of Tat whereas correction with background MI with the same amino acid frequency as the actual dataset shows greater correlations between different domains of Tat. It will be interesting to verify experimentally if such long range interactions exist in Tat or if they are modular in nature; this will be discussed in greater detail in the future directions chapter.

2.4 Using MI to identify coevolving sites between Tat-TAR

Besides the interaction of Tat with several cellular proteins, binding of Tat to TAR is critical for the recruitment of cellular factors to the viral promoter necessary for transactivation. Thus, we decided to study the interaction of Tat-TAR to identify residues in Tat that may be specifically coevolving with bases in the TAR RNA, using 182 sequences for the Los Alamos sequence database (http://www.hiv.lanl.gov/).

The landscape of raw MI scores between TAR and Tat show that the highly conserved sites in TAR display low scores with all sites in Tat and sites within the core and ARM domain of Tat (amino acids 41-52) show low scores with all bases in TAR, as would be expected due to the low entropy of these sites (Fig. 2.8). Further, the heat map shows that certain bases in TAR, such as 11, 13, 48 and 50 show high coevolution signal with almost the entire primary sequence of Tat. Since the MI score is related to the entropy of the two sites involved, the high coevolution signal seen for such sites in TAR possibly occurs due to their high entropy and thus appropriate corrections need to be introduced to filter out the noise amongst these sites, to identify site pairs that are actually coevolving (Fig. 2.8).

Plots of MI(X,Y) vs. min$\{H(X),H(Y)\}$ and MI(X,Y) vs. H(X,Y) show that no site pair is very close to the diagonal, suggesting that none of the site pairs shows significant correlation (Figs. 2.09*A* and 2.09*D*). Once again, the normalized quantities, MI(X,Y)/min$\{H(X),H(Y)\}$ and

MI(X,Y)/H(X,Y) give high values for some site pairs that are very highly conserved, which might not be a true indicator of coevolution between these sites and that, as in the case of Tat coevolution, these normalizations may not be appropriate for site pairs that are almost completely conserved (Figs. 2.09*B*,*C*,*E* and *F*).

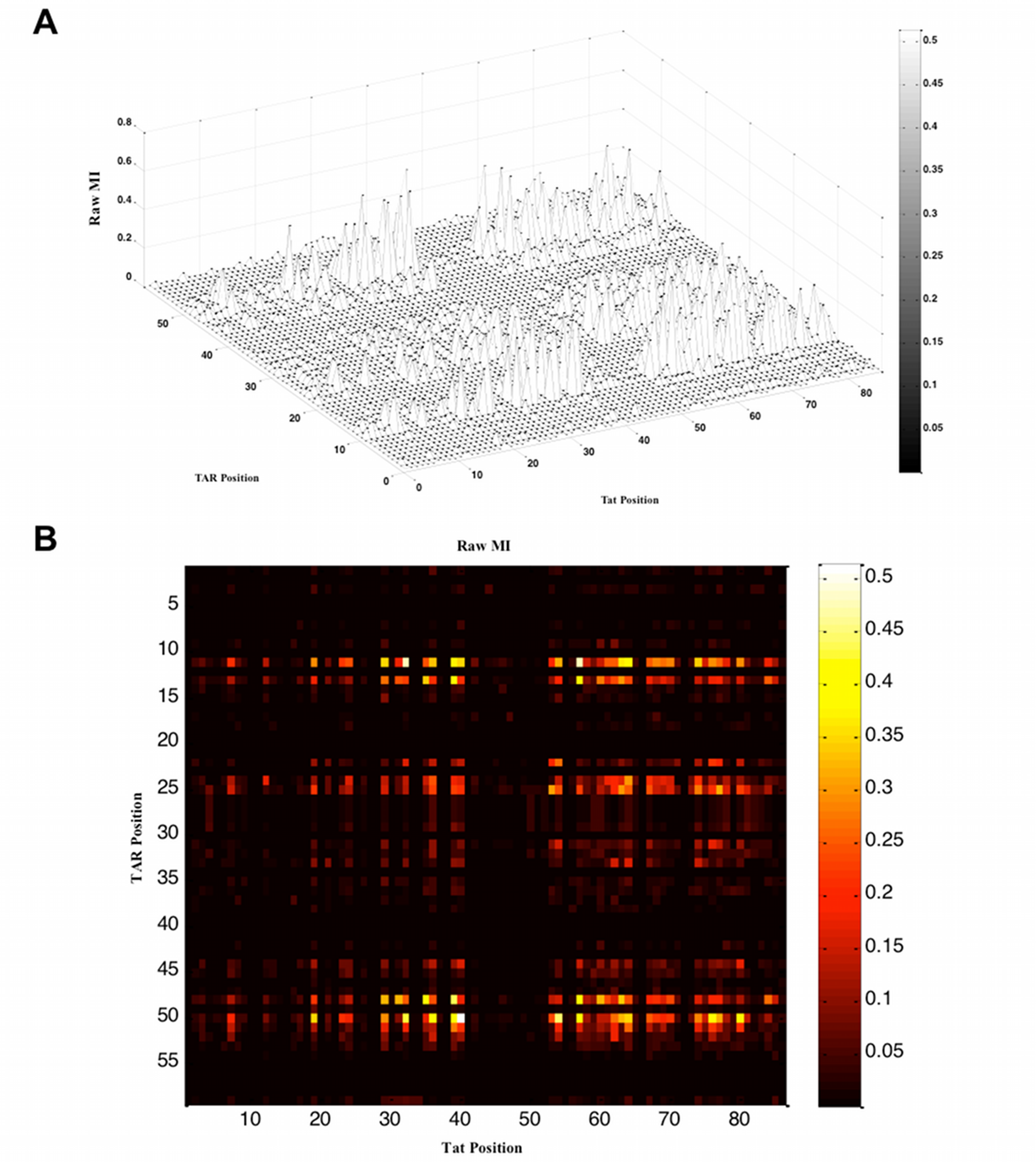


Figure 2.8. Raw MI scores between all 59 positions in TAR and 86 amino acids in Tat plotted as a (A) Mesh plot and (B) Heat map.

As in the case of intra-Tat coevolution, subtracting background MI scores, generated from a simulated sequence alignment where every base or amino acid is equiprobable, from the raw MI scores reduces the corrected MI scores to zero for most site pairs (Fig 2.10). This is due to the overestimation of background MI since sites in the simulated alignment have much higher entropies and thus to not accurately mimic the actual sequence alignment. Nevertheless, this correction shows that base 11 in TAR may be correlated with amino acid 32 within the activation domain, amino acid 57 within the ARM motif and amino acid 64 in Tat, and base 50 in TAR may be correlated with amino acid 40 in the activation domain and amino acid 54 within the ARM motif of Tat (Fig. 2.10). In the absence of a crystal structure of the entire TAR molecule in complex with the Tat protein, these correlations may be an indicator of the physical proximity between different sites in TAR and Tat that are important for maintaining the conformation of the RNA-protein complex.

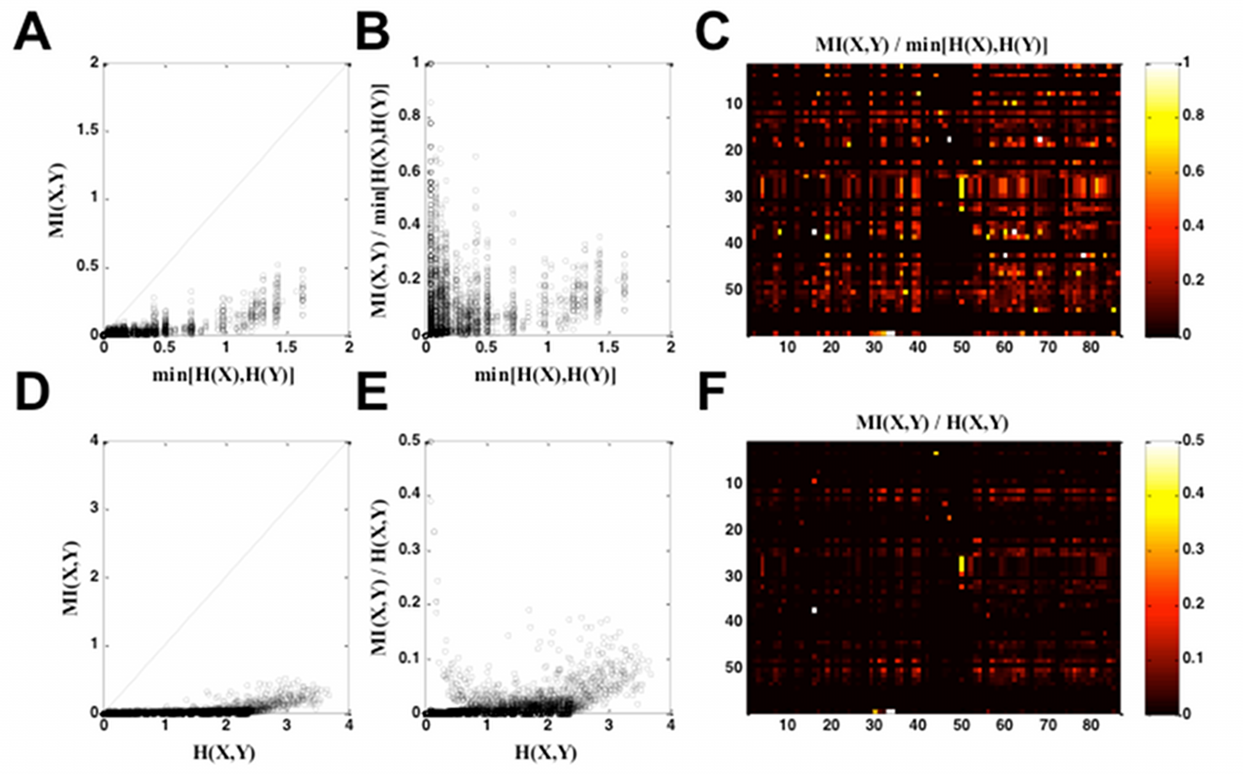


Figure 2.9. (A) Plot of MI(*X,Y*) vs. min{*H(X),H(Y)*} shows that no site pair is close to the dashed diagonal line. (B) Plot of MI(*X,Y*)/min{*H(X),H(Y)*} vs. min{*H(X),H(Y)*}. (C) Heat map of normalized MI after correction with min{*H(X),H(Y)*}. (D) Plot of MI(*X,Y*) vs. H(*X,Y*) shows that no site pair is close to the dashed diagonal line. (E) Plot of MI(*X,Y*)/H(*X,Y*) vs. H(*X,Y*). (F) Heat map of normalized MI after correction with H(*X,Y*).

To estimate the background MI more accurately, a simulated sequence alignment was constructed in which the base and amino acid frequency at each site was the same as the actual TAR and Tat database. The symmetry of MI scores seen in the heat map possibly occurs since site 11 in TAR base pairs with site 50 and site 13 in TAR base pairs with site 48, these base-pairings occurring in the lower stem of TAR (Fig. 2.11). The threshold MI score was determined as 0.38 which identified 3 Tat-TAR pairs as coevolving – (11,32), (11,57) and (50,40) (Figs.

2.12*A* and 2.12*B*). Interestingly, although site 11 base-pairs with site 50 in TAR, they display maximum coevolution signal with different residues in Tat, 32 and 40, respectively. This suggests that site 11 in TAR may be specifically interacting with residue 32 in Tat and site 50 in TAR may be specifically interacting with residue 40 in Tat, and that these coevolution signals do not arise as a consequence of the base-pairing constraint imposed by the stem structure of TAR.

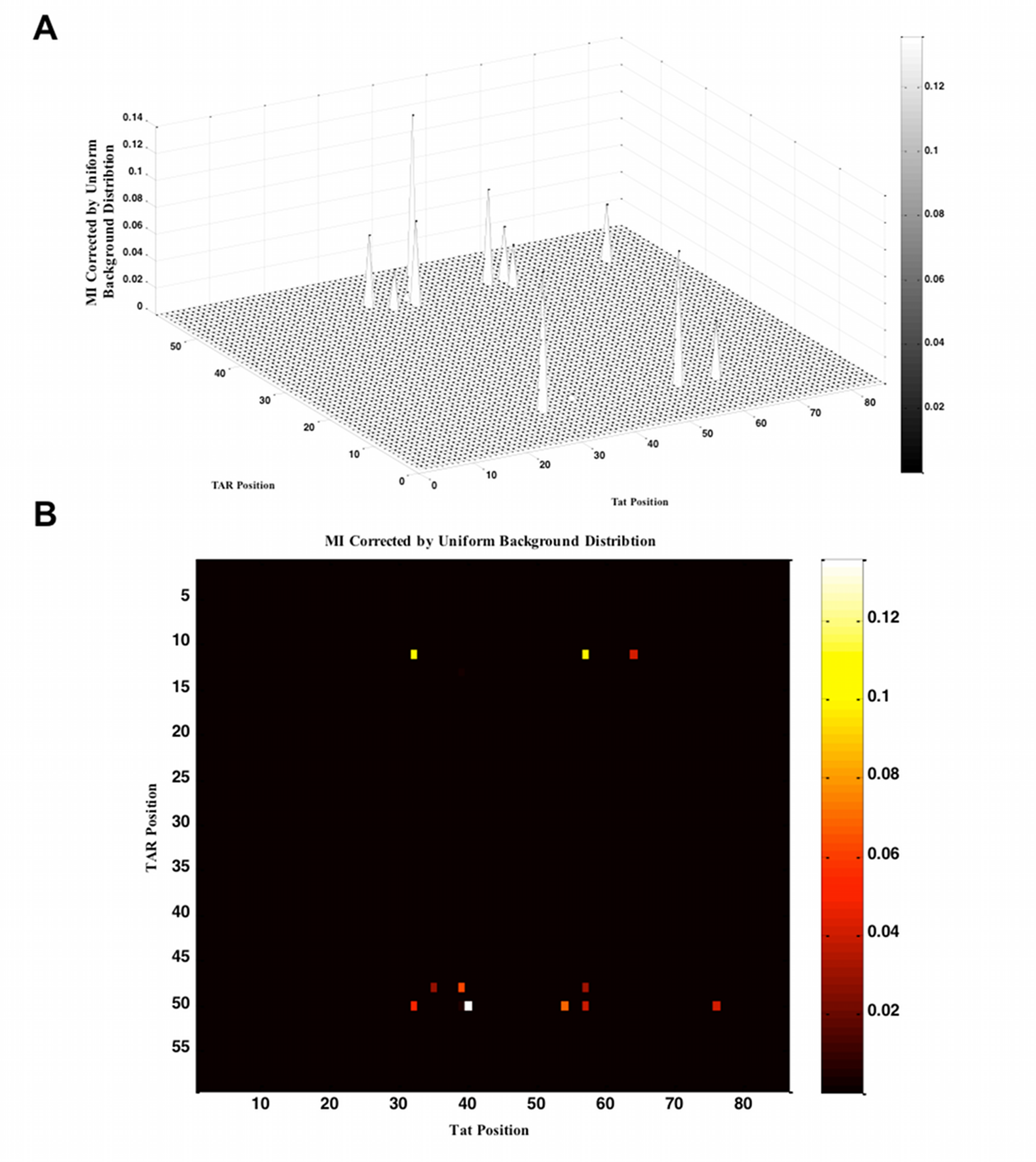


Figure 2.10. MI scores between positions in TAR and Tat plotted as a (A) mesh plot and (B) heat map after subtraction of background MI generated from a simulated sequence alignment with equiprobable base and amino acid distribution.

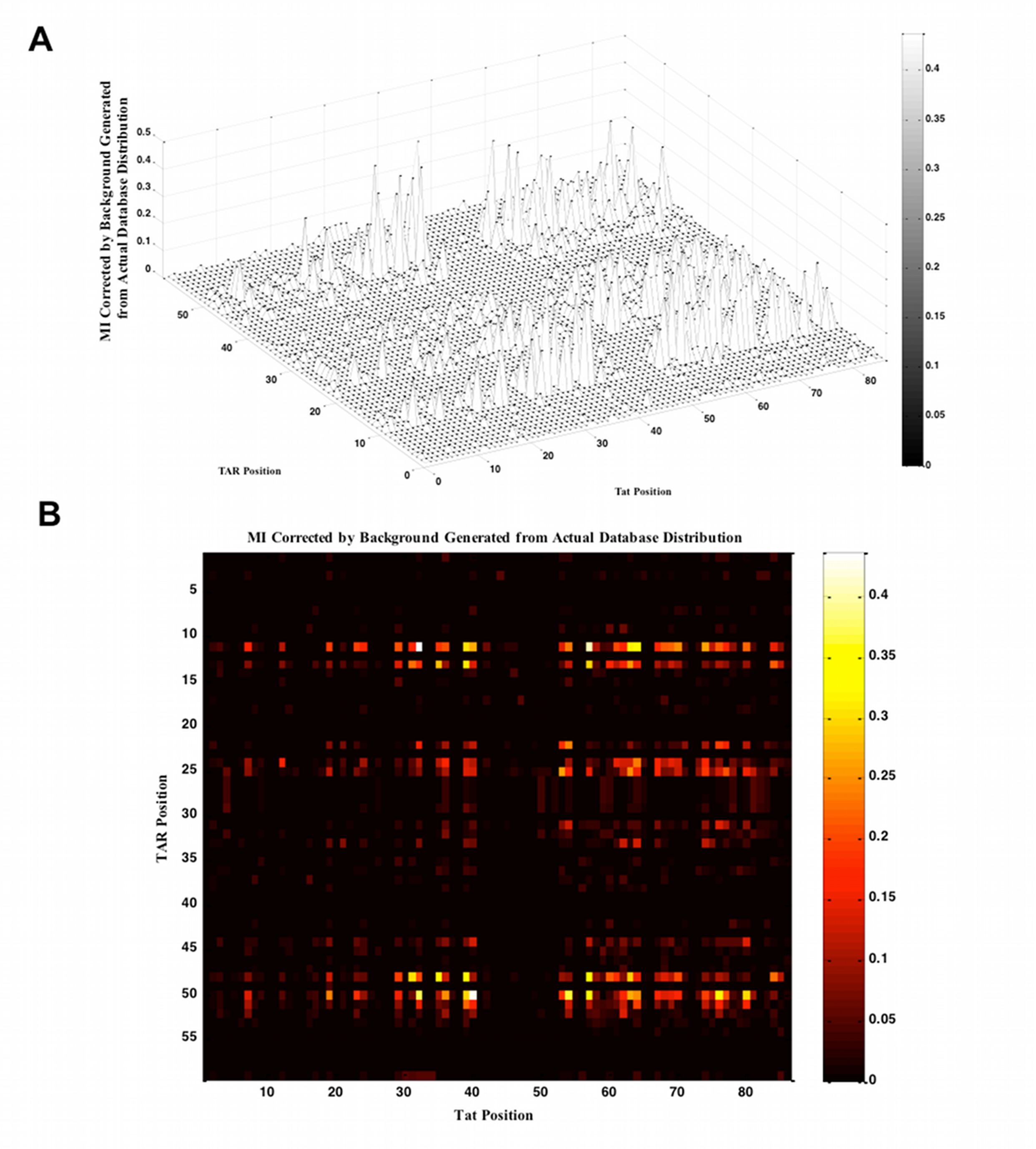


Figure 2.11: MI scores between positions in TAR and Tat plotted as a (A) mesh plot and (B) heat map after subtraction of background MI generated from a simulated sequence alignment with the same base and amino acid frequency as the actual database.

Correction with APC shows that the heat map of MI scores do not show the same symmetry in coevolution signals as seen in the other methods, suggesting that this correction, which minimizes the background signal arising from the phylogenetic relation shared between the sequences in the alignment, besides the structural constraints imposed by the secondary structure of TAR, was responsible for the symmetry of MI scores (Fig. 2.13). The threshold MI score was identified as 0.16, and construction of a network of sites in TAR and Tat above the threshold score reveals interesting trends (Fig. 2.14*A* and 2.14*B*). The presence of a one-to-one correlation in most cases shows that specific bases in TAR interact with specific residues in Tat, possibly due to their close physical proximity that is necessary for the formation of a stable Tat-TAR complex (Fig. 2.14*B*). The high coevolution signal between site 25 within the bulge of TAR and site 53 within the ARM motif of Tat confirm previously known data that the bulge in TAR interacts with the ARM motif of Tat (16, 17). Interestingly, the MI analysis also produces some new correlations between the lower stem of TAR (sites 11, 13 and 48) with the Cys-rich motif in Tat (sites 31, 32 and 35). The cellular protein PKR has been shown to bind to the lower stem of TAR as well as phosphorylate Tat. The correlation between sites in the lower stem of TAR and the Cys-rich region of Tat could arise from their interaction with PKR (18, 19). The MI analysis also reveals that a residue in the ARM motif (site 54) of Tat may also be interacting with a base (site 22) just below the bulge of TAR.

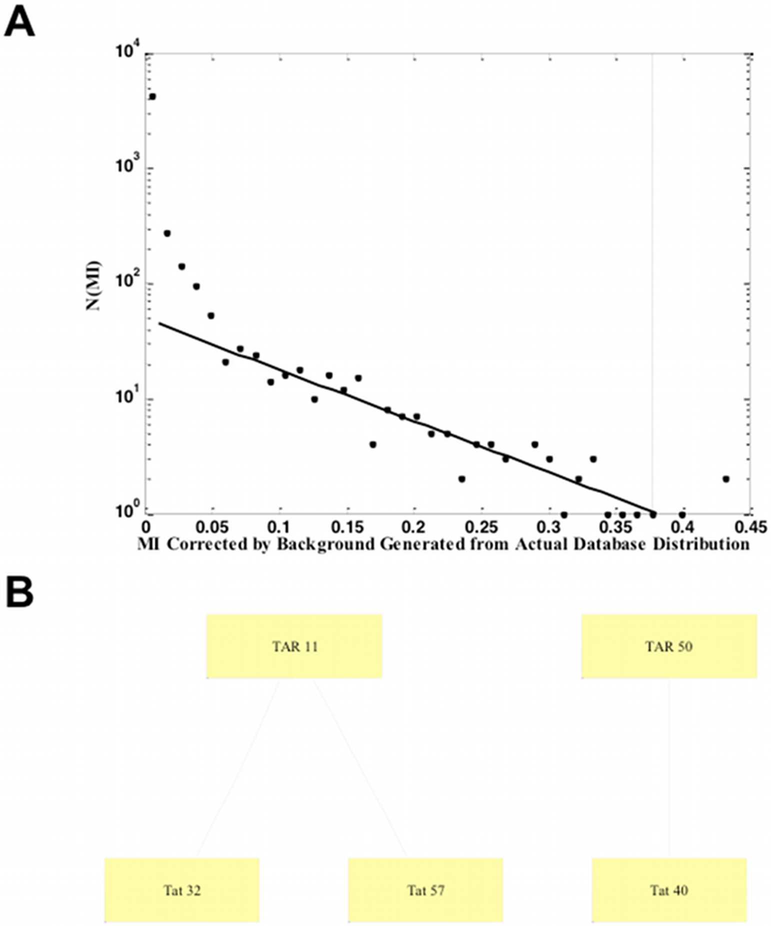


Figure 2.12. (A) Deviation from an exponential function is used to identify the threshold MI as 0.38. (B) Network of sites above threshold MI in TAR correlated with those in Tat.

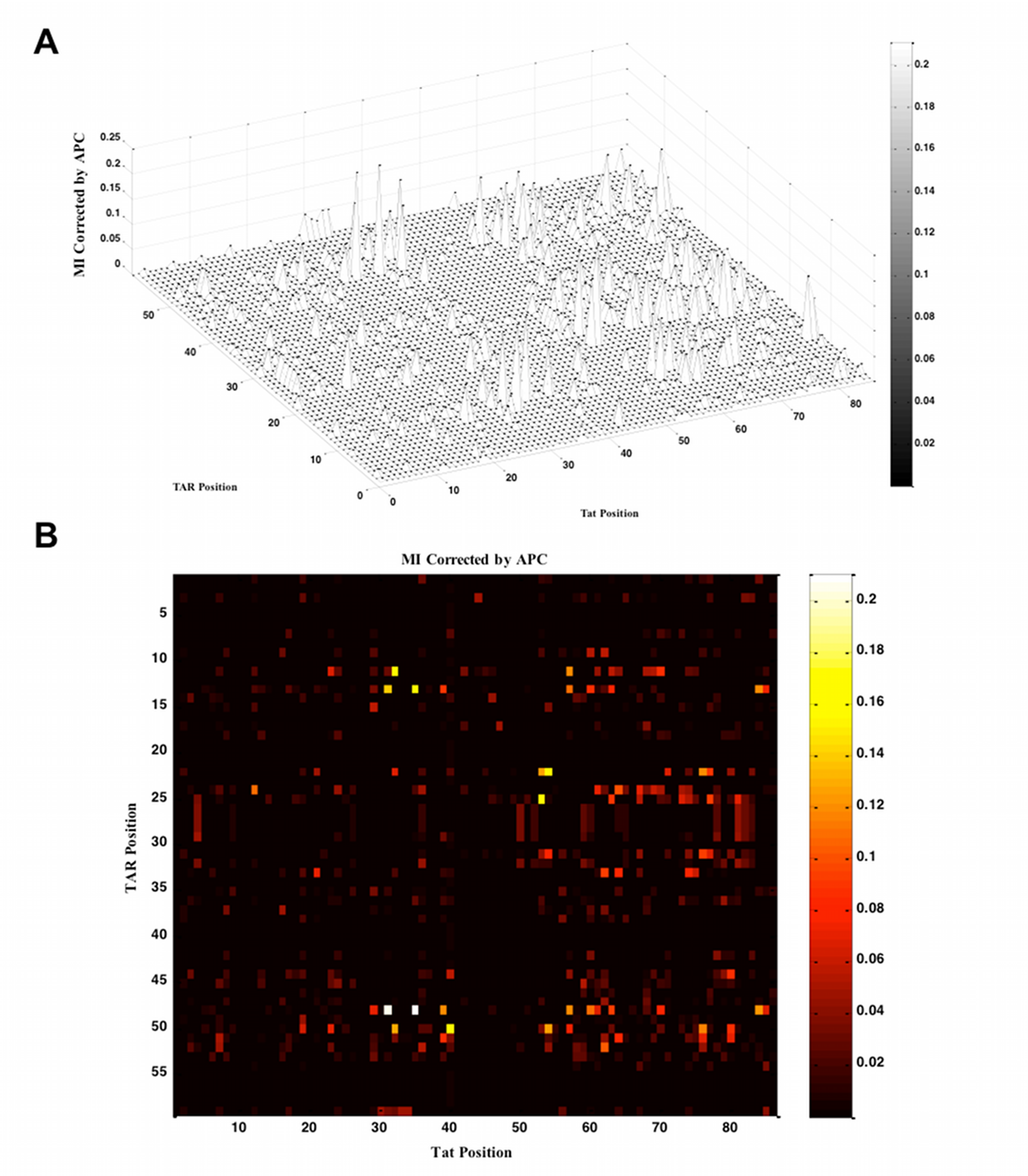


Figure 2.13. MI scores corrected by APC shown as a (A) mesh plot and (B) heat map.

The MI analysis between TAR and Tat, after correction of background MI using several methods, reveals linear correlation between sites in TAR and Tat. It would be interesting to test experimentally if these sites in TAR and Tat show this specificity in interaction as revealed by the MI analysis.

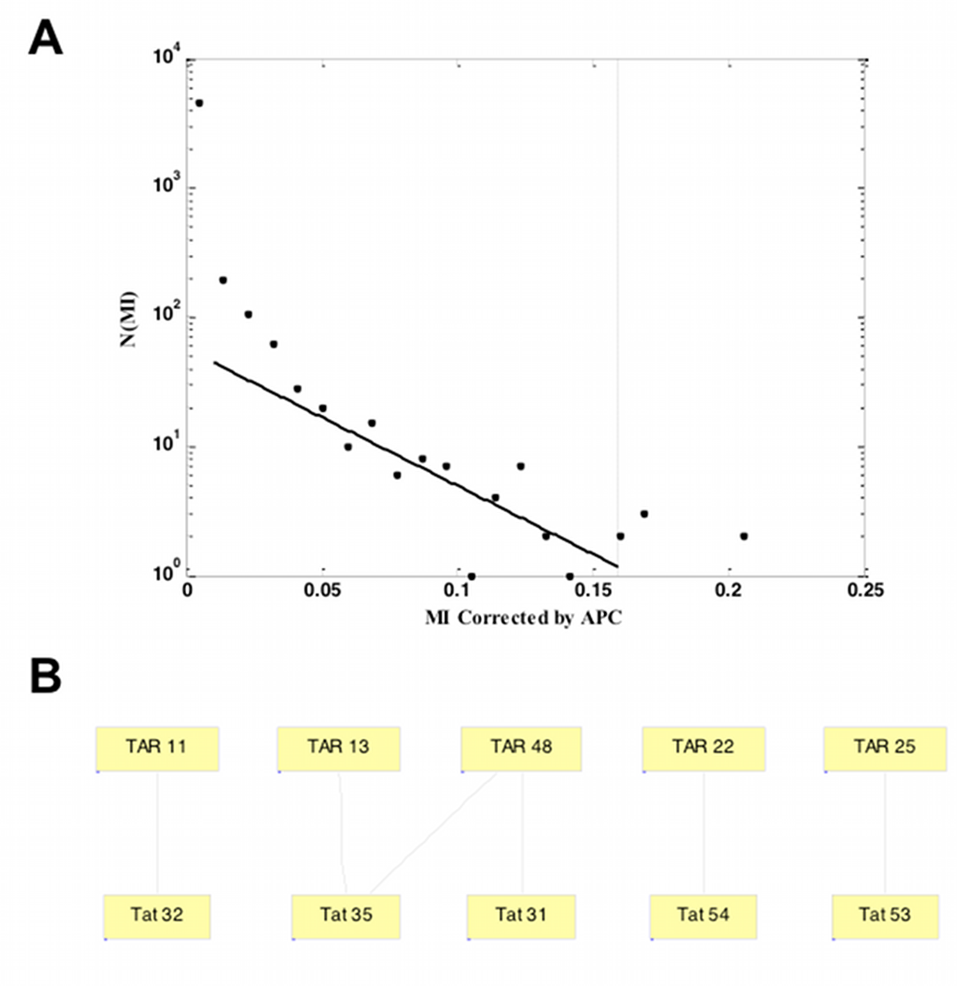


Figure 2.14. (A) Deviation from an exponential function is used to identify the threshold MI as 0.16. (B) Network of correlated sites in TAR and Tat that are above threshold MI.

2.5 Rev-mediated transport of incompletely spliced viral RNA to the cytoplasm

Rev (Regulator of Expression of Virion proteins) is a 116 amino acid regulatory protein that plays a critical role in transporting full-length or incompletely spliced, intron-containing viral RNA to the cytoplasm. During the early stages of viral replication, in the absence or under low levels of Rev, the full-length viral transcript in the nucleus of infected cells are spliced by the host splicing machinery to yield short (~2kb) transcripts that are constitutively exported to the cytoplasm (Fig. 2.15*A*). One of these short transcripts encodes for Rev which has a arginine-rich sequence (amino acids 34-50) that acts as nuclear localization signal allowing Rev to build up within the nucleus (20) (Fig. 2.15*A*). Rev then binds to a structured stem-loop RNA element called Rev Response Element (RRE) present within all full-length and partially spliced viral transcripts (21-23) (Fig. 2.15*B*). RRE contains a high-affinity site consisting of stems IIB and IID that binds to the same arginine-rich motif as the nuclear localization signal in Rev (23-27). After the initial RNA-protein interaction, additional Rev molecules bind to this complex through RNA-protein and protein-protein interactions (28-30). Sites such as 23, 25 and 26, and those on the other side of the nuclear localization signal have been shown to be important for Rev multimerization (31, 32) . Eight or more Rev monomers bound to the viral RNAs are then transported across the nuclear membrane with the help of a leucine-rich nuclear export signal present within Rev (amino acids 75-83) to rescue intron-containing viral RNAs from the nuclear splicing machinery (33) (Fig. 2.15*B*).

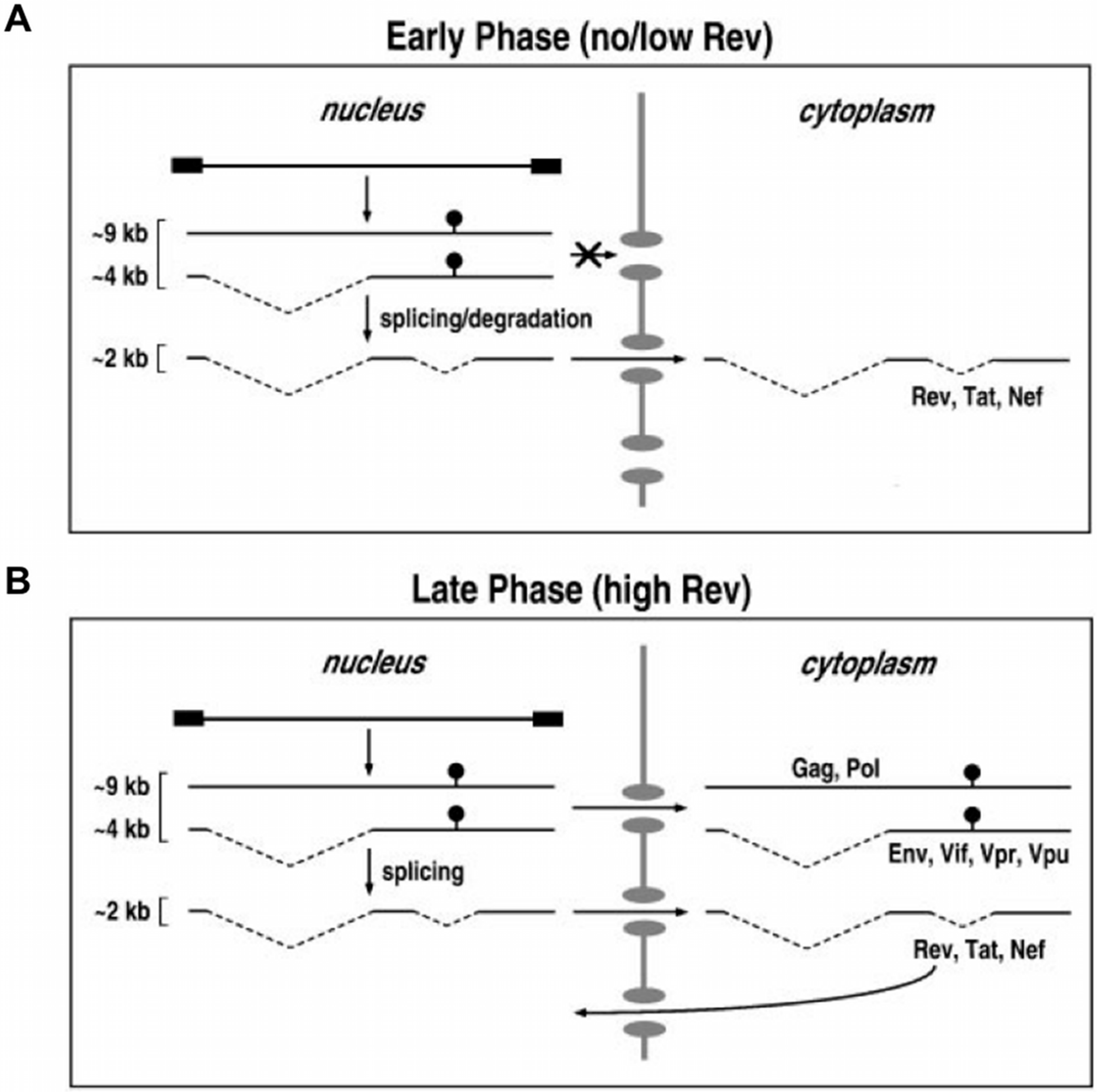
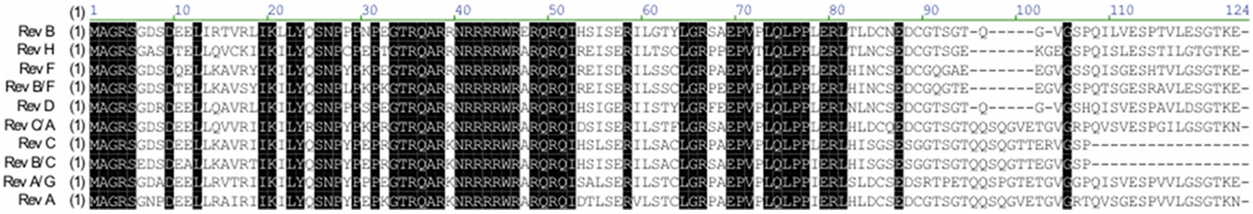


Figure 2.15. (A) During the early phase of viral replication, in the absence or under low levels of Rev, the full-length (~9kb) and partially spliced (~4kb) transcripts are retained in the nucleus and are either degraded or spliced to short (~2kb transcripts) that are exported to the cytoplasm. Translation of these short transcripts give rise to Rev, Tat and Nef. (B) Rev localizes to the nucleus and once enough Rev builds up, it binds to a structured RNA element called RRE (shown as a ball and stick in the figure) and exports the full and partially spliced transcripts to the cytoplasm. Adapted from (34).

Similar to the regulatory protein Tat discussed previously, Rev plays multiple functions in ensuring normal viral replication, despite having sequence variation within different sites (Fig 2.16). Besides binding to RRE and other Rev molecules, it also interacts with various cellular factors involved in nuclear export, such as Crm1 and RIP/Rab (35-37). Although, the presence of the N-terminal arginine-rich nuclear localization signal/RNA-binding domain and the C-terminal nuclear export signal makes Rev a relatively modular protein, with heterologous nuclear localization or nuclear export signal sequences ensuring normal Rev function, replacing both peptides with heterologous sequences result in a non-functional Rev (38-40). Thus, sites in the N- and C- terminus of the protein could be correlated that are necessary for normal protein function. We plan to identify such sites that may be important for the structural stability of a monomer, important for formation of Rev multimers or important for interactions with other cellular proteins to ensure normal Rev function.



```
           (1) 1        10        20        30        40        50        60        70        80        90        100       110           124
Rev B      (1) MAGRSGDSDEELIRTVRLIKLLYQSNPPPNPEGTRQARRNRRRRWRERQRQIHSISERILGTYLGRSAEPVPLQLPPLERLTLDCNEDCGTSGT-Q-----G-VGSPQILVESPTVLESGTKE-
Rev H      (1) MAGRSGASDTELLQVCKIIKILYQSNPCPEPTGTRQARRNRRRRWRARQRQIREISERILTSCLGRPPEPVTLQLPPLERLTLNCSEDCGTSGE-------KGEGSPQISLESSTILGTGTKE-
Rev F      (1) MAGRSGDSDQELLKAVRYIKILYQSNPYPKPEGTRQARRNRRRRWRARQRQIREISDRILSSCLGRPAEPVPLQLPPLERLHINCSEDCGQGAE-------EGVGSSQISGESHTVLGSGTKE-
Rev B/F    (1) MAGRSGDSDTELLKAVSYIKILYQSNPLPKPKGTRQARRNRRRRWRARQRQIREISERILSSCLGRPEEPVPLQLPPLERLHINCSEDCGQGTE-------EGVGSPQTSGESRAVLESGTKE-
Rev D      (1) MAGRSGDRDEELLQAVRLIKILYQSNPPPSPEGTRQARRNRRRRWRARQRQIHSIGERIISTYLGRFEEPVPLQLPPLERLNLNCSEDCGTSGT-Q-----G-VGSHQISVESPAVLDSGTKE-
Rev C/A    (1) MAGRSGDSDEELLQVVRIIKILYRSNPYPKPRGTRQARKNRRRRWRARQRQIDSISERILSTFLGRSAEPVPLQLPPLERLHLDCQEDCGTSGTQQSQGVETGVGRPQVSVESPGILGSGTKN-
Rev C      (1) MAGRSGDSDEELLKAVRIIKILYQSNPYPEPRGTRQARKNRRRRWRARQRQIHSLSERILSACLGRPAEPVPLQLPPLERLHISGSESGGTSGTQQSQGTTERVGSP-----------------
Rev B/C    (1) MAGRSEDSDEALLKAVRTIKILYQSNPYPEPRGTRQARKNRRRRWRARQRQIHSISERILSACLGRSAEPVPLQLPPIERLHISGSESGGTSGTQQSQGTTEGVGSP-----------------
Rev A/G    (1) MAGRSGDADEELLRVTRIIKILYQSNPYPPPEGTRQARKNRRRRWRARQRQISALSERILSTCLGRPAEPVPLQLPPIERLSLDCSEDSRTPETQQSPGTETGVGGPQISVESPVVLGSGTKE-
Rev A      (1) MAGRSGNPDEELLRAIRIIKILYQSNPYPEPKGTRQARKNRRRRWRARQRQIDTLSERVLSTCLGRPAEPVPLQLPPIERLHLDCSEDCGTSGTQQSQGVETGVGRTQVSGESPVVLGSGTKN-
```

Figure 2.16. Representative alignment of Rev sequences from different subtypes showing sequence diversity at different sites. Black columns represent completely conserved sites.

2.6 Using MI to identify coevolving residues within Rev

As briefly described in the previous section, the viral regulatory protein Rev interacts with a large number of cellular proteins similar to Tat, the other regulatory protein of HIV-1. We therefore decided to compute MI scores between different residues in Rev to identify non-conserved but functionally or structurally important sites in Rev using 1033 sequences from the Los Alamos Sequence Database (http://www.hiv.lanl.gov/). MI scores within Rev were calculated for the first 95 amino acids, instead of the entire 116 amino acids since the quality of sequence alignment beyond residue 95 was poor.

The raw MI landscape reveals that certain sites within amino acids 40-50 in the RRE-binding/nuclear localization signal, and sites within 75-80 in the nuclear export signal show low correlation with all other sites in Rev (Fig. 2.17*A*). This is expected since these residues are within functionally important motifs and hence highly conserved residues. Heat maps show that site pairs (88,89) close to the C-terminus end of the protein, and (11,14) close to the N-terminus end of the protein show high coevolution signal (Fig. 2.17*B*). Interestingly, site pairs (11,88) and

(11,89) also show high coevolution signal, suggesting that sites at the two ends of the protein may be functionally coupled and essential for protein function (Fig. 2.17*B*).

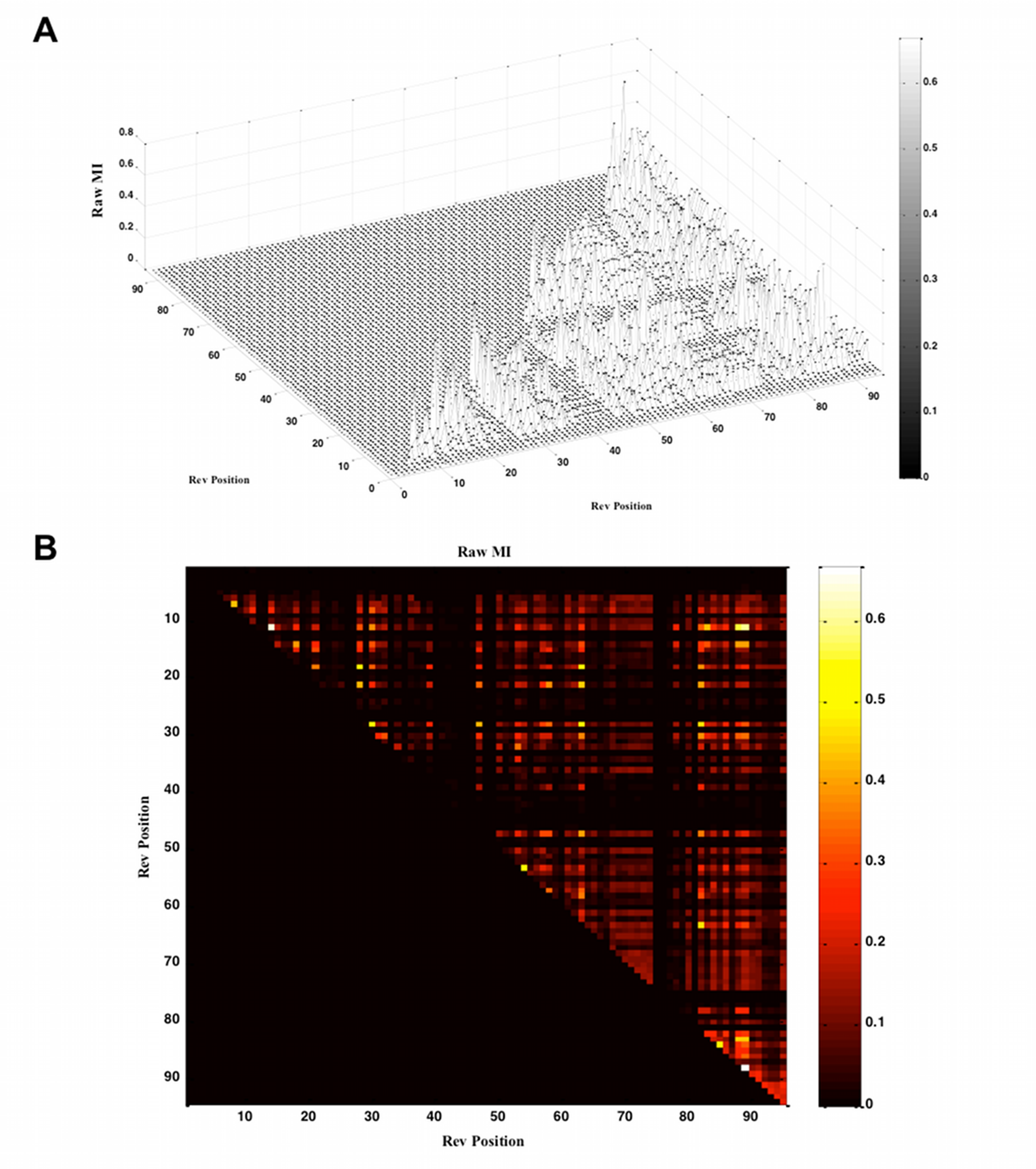


Figure 2.17. Raw MI scores between the first 95 sites in Rev shown as a (A) mesh plot and (B) heat map.

Plotting MI(*X,Y*) vs. min{*H(X),H(Y)*} again shows that a few site pairs, such as (11,14), (88,89), (11,88) and (11,89) appear to be closer to the dotted diagonal line, an indicator of significant correlation, than most other site pairs, suggesting these positions may be important for Rev function (Fig. 2.18*A*). However, a plot of MI(*X,Y*) vs. H(*X,Y*) shows that except for the site pair (88,89), most other pairs are within the background region (Fig. 2.18*D*). The normalized quantity MI(*X,Y*)/min{*H(X),H(Y)*} again gives high scores for positions that have very low entropies close to zero, and so this method may not be appropriate for positions that are almost completely conserved. However, this normalization reveals that sites 69-71 may be correlated and interacting with several other sites in the protein (Figs. 2.18*B* and 2.18*C*). Similarly, the MI(*X,Y*)/H(*X,Y*) normalization also suggests that sites 69-71 may be correlated with other sites in Rev (Figs. 2.18*E* and 2.18*F*).

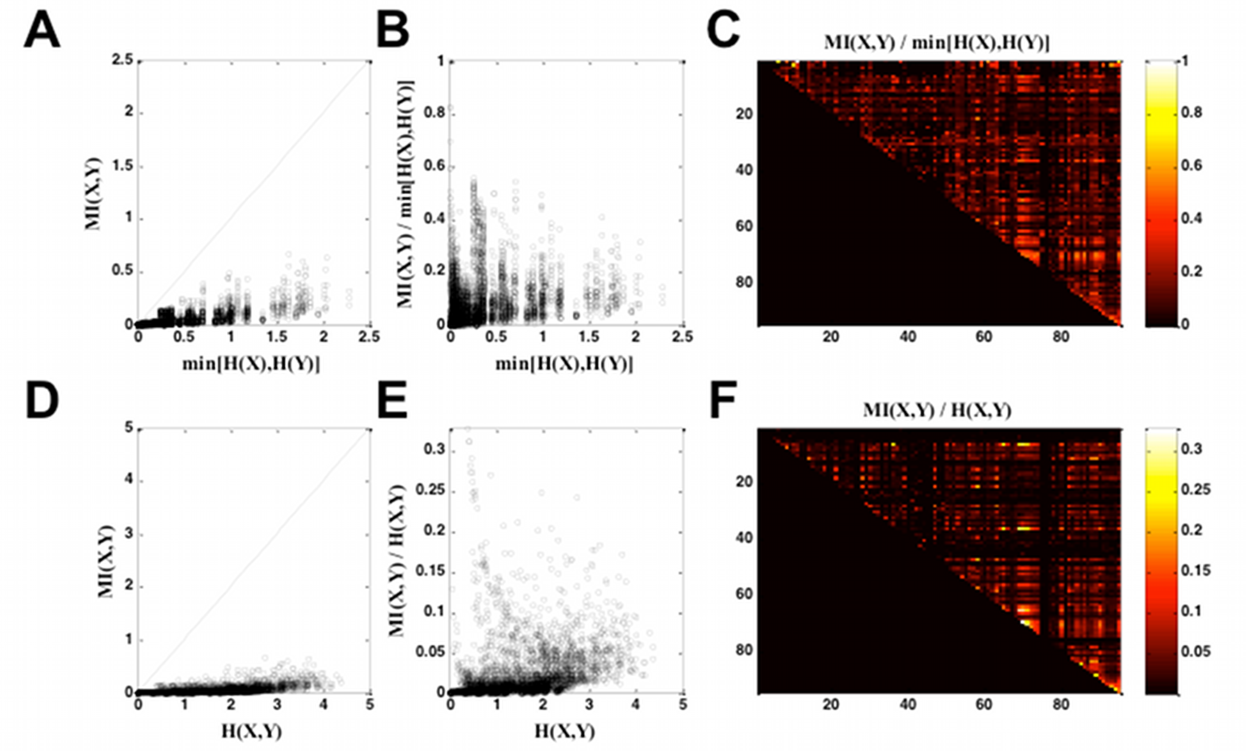


Figure 2.18. (A) Plot of MI(*X,Y*) vs. min{*H(X),H(Y)*} shows that a few site pairs such as (11,14), (88,89), (11,88) and (11,89) are closer to the diagonal line than most other site pairs. (B) Plot of MI(*X,Y*)/min{*H(X),H(Y)*} vs. min{*H(X),H(Y)*}. (C) Heat map of normalized MI after correction with min{*H(X),H(Y)*}. (D) Plot of MI(*X,Y*) vs. H(*X,Y*) shows that the site pair (88,89) is closer to the diagonal line than all other pairs. (E) Plot of MI(*X,Y*)/H(*X,Y*) vs. H(*X,Y*). (F) Heat map of normalized MI after correction with H(*X,Y*).

Correction of raw MI scores using a simulated sequence alignment where all residues are equiprobable at each site again results in the overestimation of background MI and reduces the score for most site pairs to zero. However, this correction also reveals position pairs that were previously identified to be potentially coevolving - (11,14), (88,89), (11,88) and (11,89) (Fig. 2.19). This correction reveals some interesting trends for the multimerization domains in Rev, which are present on either side of the RRE binding/nuclear localization signal (32). Site pair (53,54) appears to be correlated to each other (Fig. 2.19). Site 54 was initially believed to be important for Rev multimerization but was later shown to be important for RRE binding (41). If site 54 coevolves with site 53, it would be interesting to study how different amino acid combinations at these two positions impact RRE binding vs. Rev multimerization. Similarly, other sites – (28,30) and(18,28) - within or close to the Rev multimerization domains appear to be correlated to each other (Fig. 2.19).

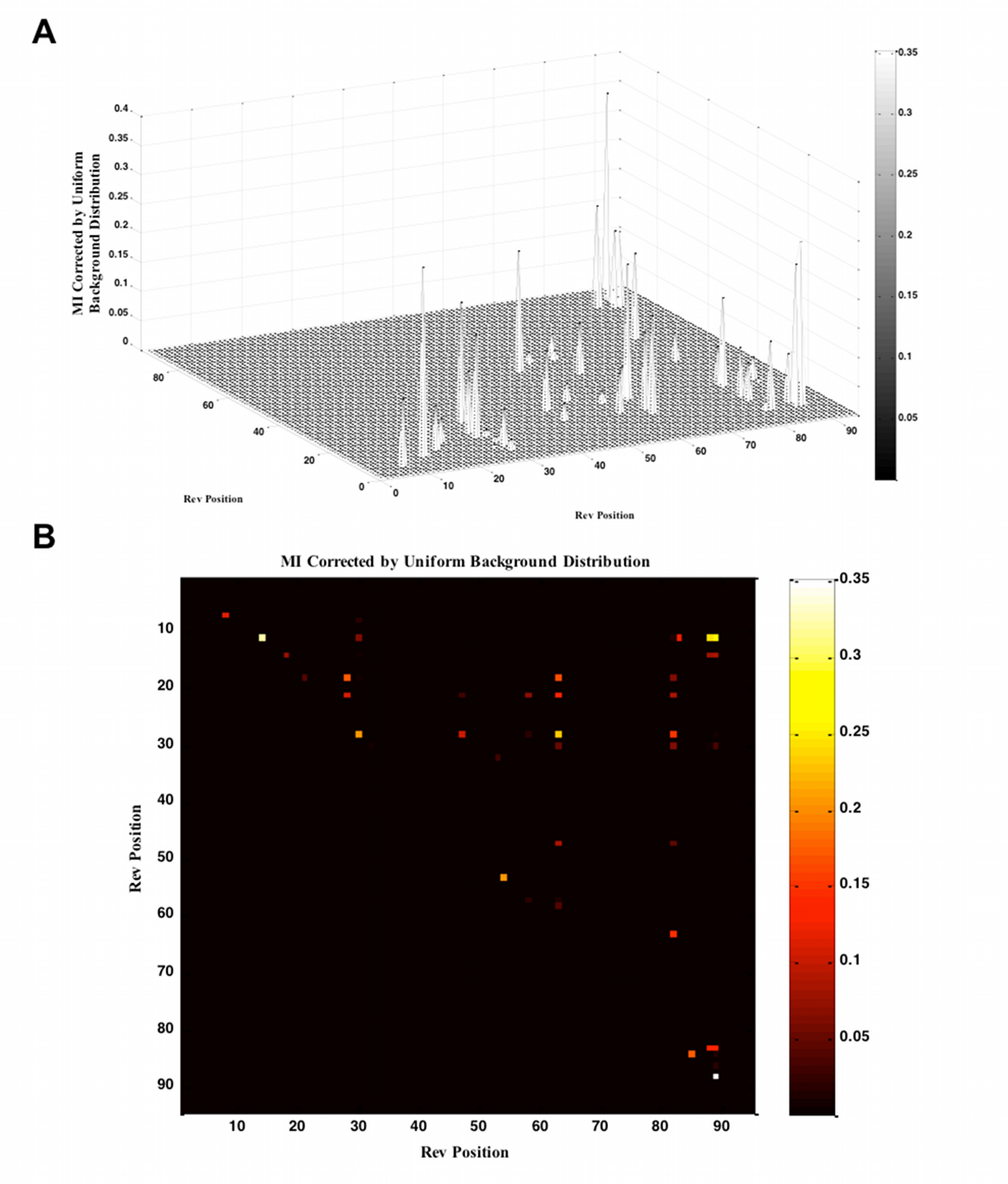


Figure 2.19. Corrected MI scores between positions in Rev shown as a (A) mesh plot and (B) heat map after subtraction of background MI generated from a simulated sequence alignment with equiprobable amino acid distribution.

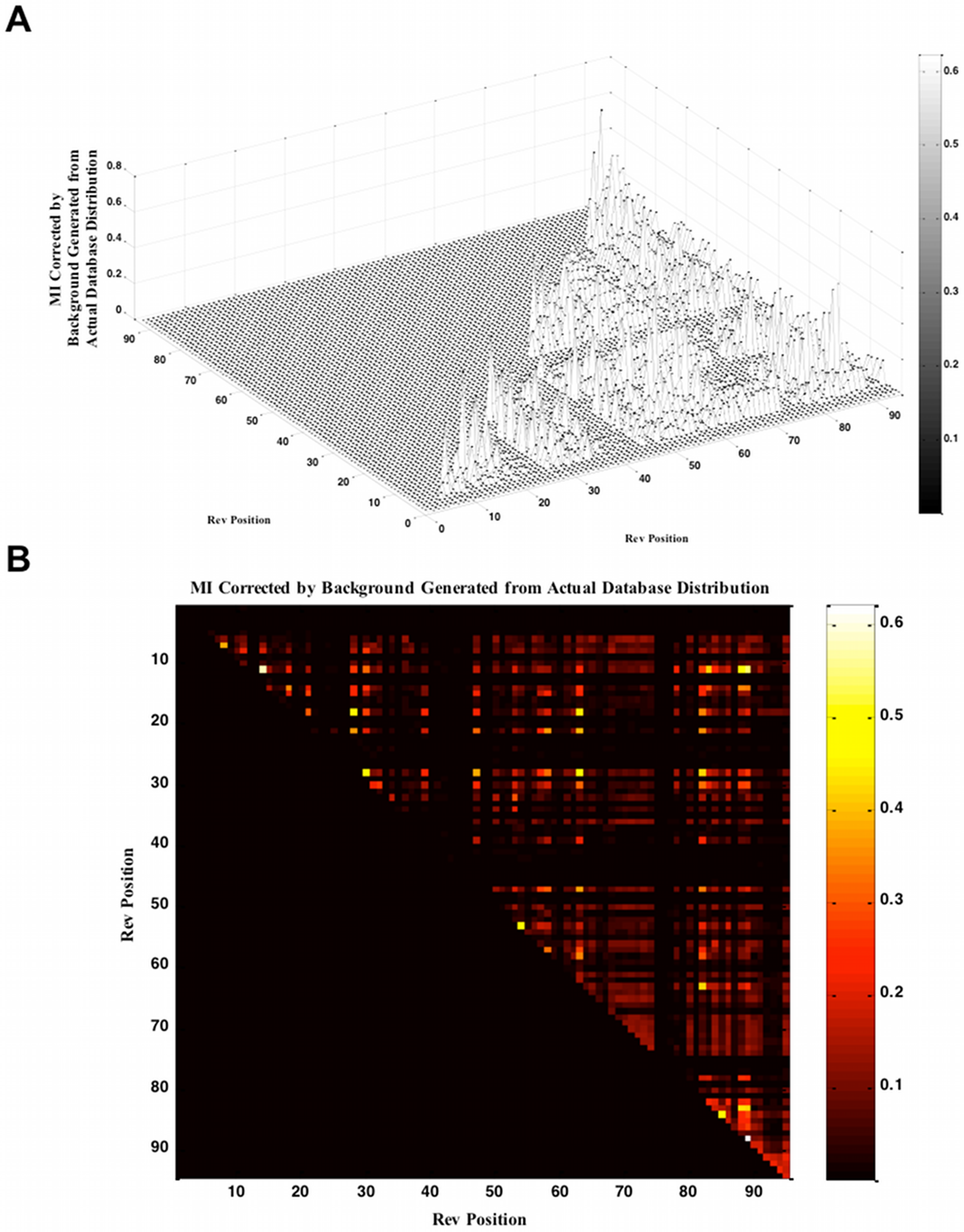


Figure 2.20. Corrected MI scores between positions in Rev shown as a (A) mesh plot and (B) heat map after subtraction of background MI generated from a simulated sequence alignment with the same amino acid distribution at each site as the actual database.

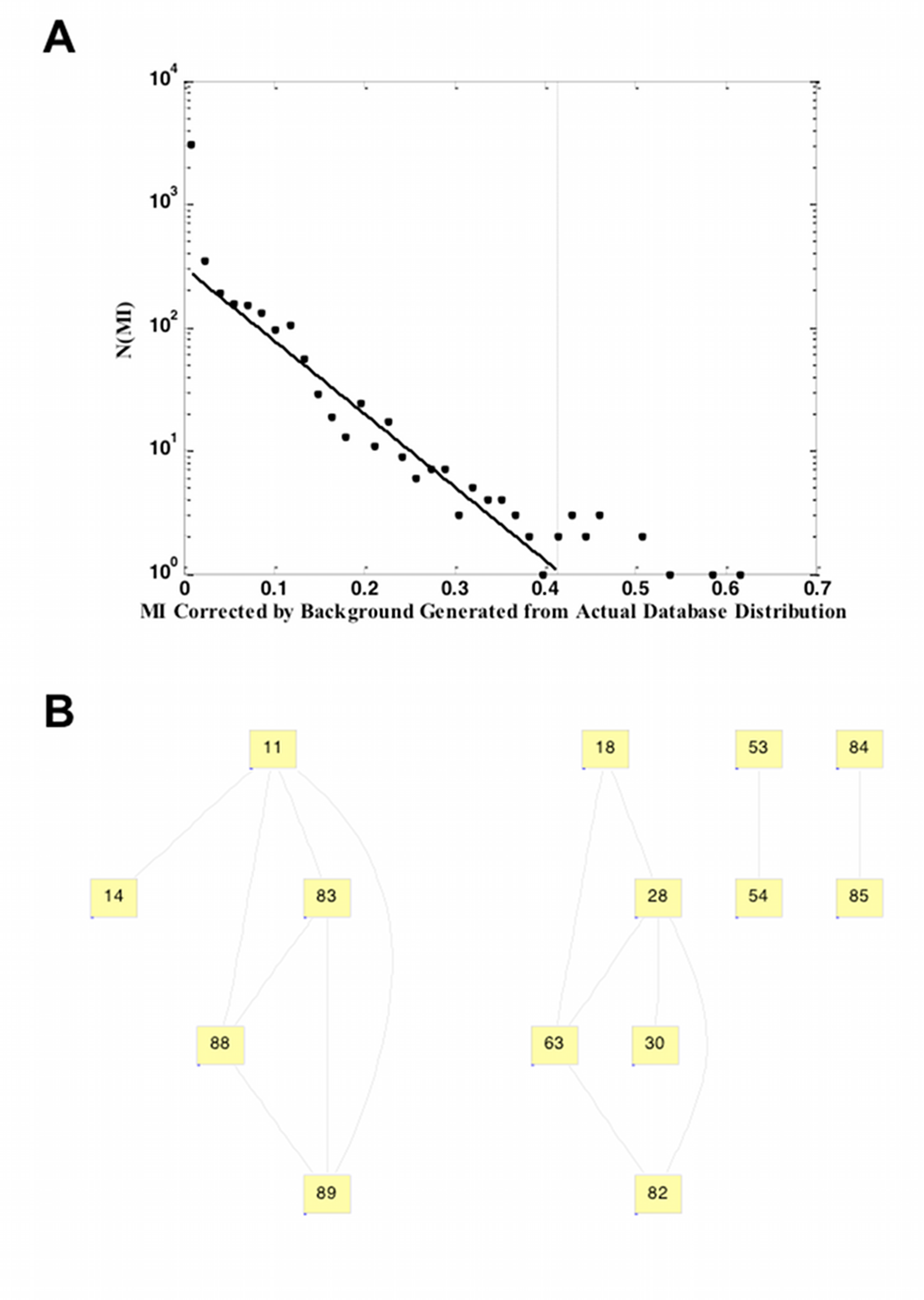


Figure 2.21. (A) Deviation from an exponential function is used to identify the threshold MI as 0.41. (B) Network of correlated sites in Rev that are above threshold MI.

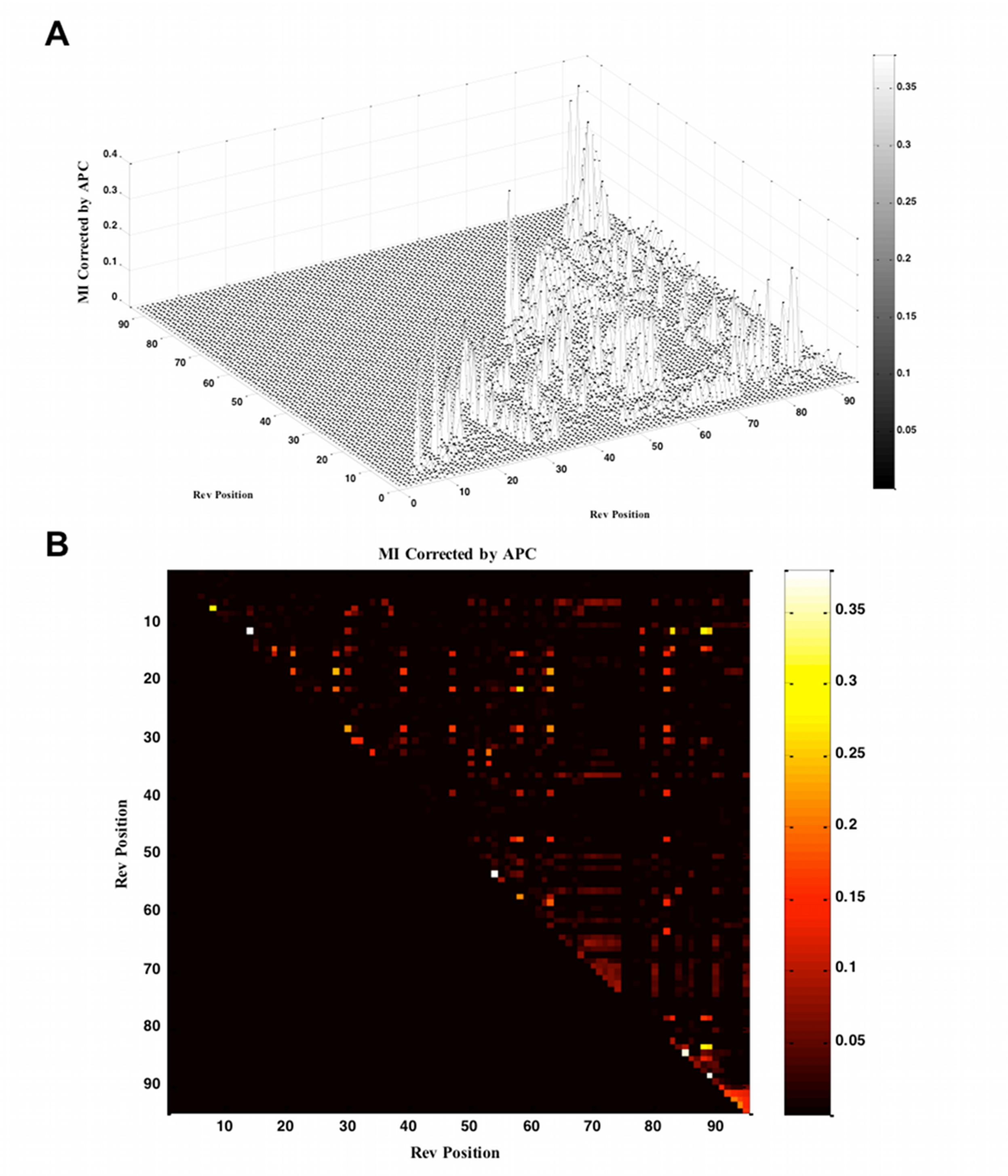


Figure 2.22. MI scores corrected by APC shown as a (A) mesh plot and (B) heat map.

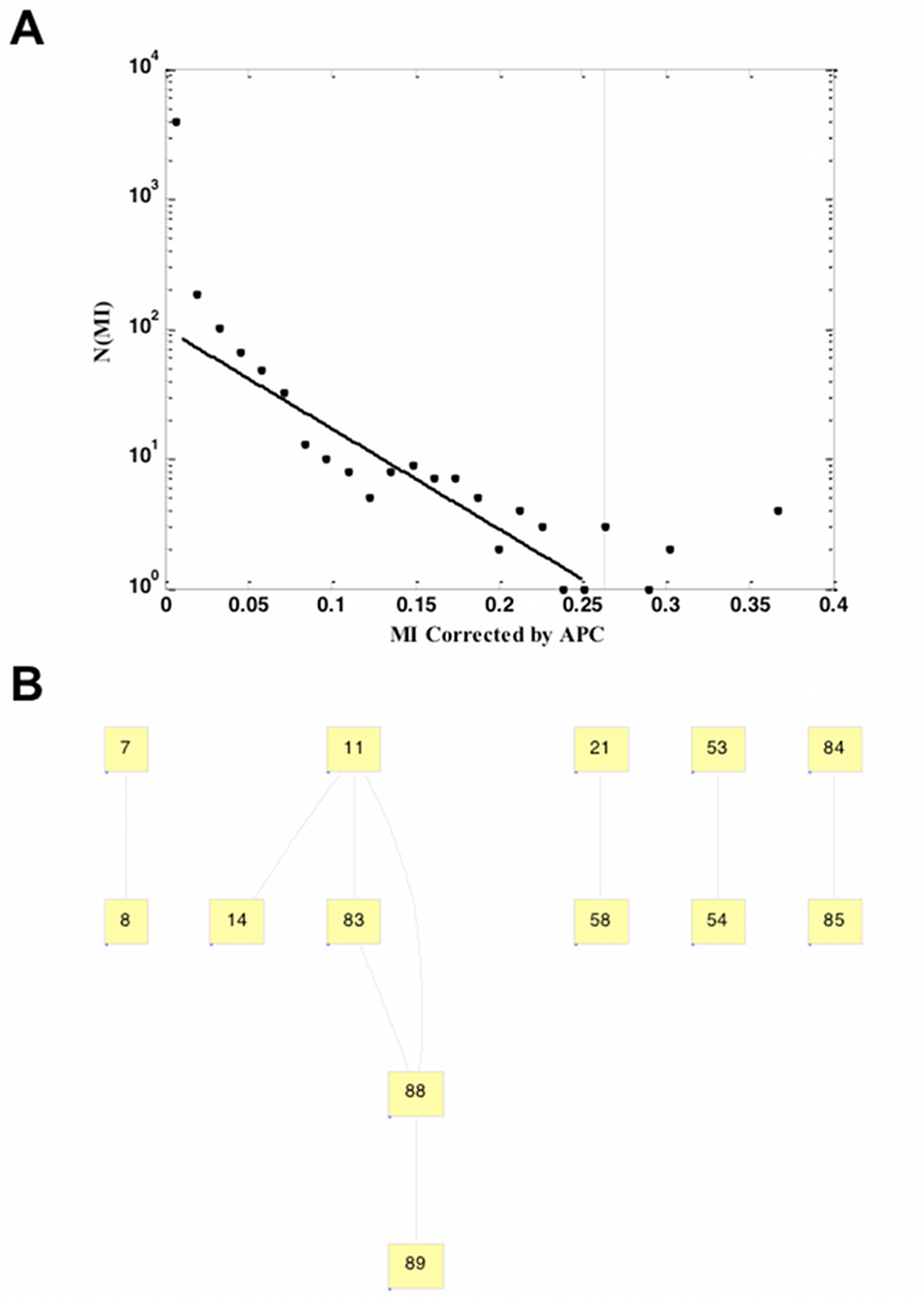


Figure 2.23. (A) Deviation from an exponential function is used to identify the threshold MI as 0.26. (B) Network of correlated sites in Rev that are above threshold MI.

Correcting raw MI scores using a simulated sequence alignment where the frequency of amino acids at each site is the same as the actual sequence database, identifies similar positions pairs to those identified by the previous methods (Fig. 2.20). The threshold MI score of 0.41 was used to identify the network of sites that are potentially coevolving (Fig. 2.21*A*). Sites 11, 83, 88 and 89 form a completely connected network with all sites correlated to the other (Fig. 2.21*B*). Site 11 also shows significant coevolution signal with site 14. Further, sites 18, 28, and 30 within the multimerization domain of Rev are correlated to each other and to sites 63 and 82 with site 28 appearing to be a particularly important site since it is linked to all the other four sites (Fig. 2.21*B*). As previously identified from the other methods, sites 53 and 54 appear to be correlated and may be functionally important for RRE binding or Rev multimerization.

Finally, correction with APC to minimize the phylogenetic and entropic contributions of noise to MI reveals site pairs – (53,54), (11,14), (88,89) and (84,85) - as the most strongly correlated pairs (Fig. 2.22). The threshold MI of 0.26 was used to construct a network of correlated sites above that cut-off score (Fig. 2.23*A*). The network of sites shows some interesting distinctions from the network generated by the previous method (Figs. 2.23*B* and 2.21*B*). Sites 11,14, 83, 88 and 89, are still correlated to each other though some of the edges have disappeared, suggesting that some of those correlations arose from phylogenetic contributions to noise. In agreement with the previous method, site pairs (53,54) and (84,85) are above the threshold MI score. Position pairs (7,8) and (21,58) that were previously not identified by any other method also show coevolution signals above threshold (Fig. 2.23*B*). Site pair (21,58) appears to be particularly interesting since it lies within the multimerization domain of Rev, on either side of the RRE binding/nuclear localization signal domain.

Thus, MI has allowed us to identify important correlated sites within Rev that are possibly linked to each other functionally or structurally. Sites close to the N-terminus, 11 and 14 appear to be correlated with sites 83, 88 and 89. Similarly, other sites 18, 28 and 30, within the Rev multimerization domain are linked to sites 63 and 82. Furthermore, site pairs (21,58) and (53,54), within either side of the Rev multimerization domain may be important for the interaction between Rev molecules. Interestingly, interactions between residues close to the N- and C-terminal end of the protein could possibly explain why replacing both the nuclear localization and export signal with homologous peptide signal sequences fail to function as wild-type Rev. A detailed study of the sites that were identified in the above analysis should help in gaining a better understanding of Rev function and the contribution of these non-conserved sites to protein function.

2.7 Materials and Methods

2.7.1 Mutual Information Analysis

Codes for Mutual Information to estimate raw and background scores were based on the mathematical equations presented in Section 2.2. They were written in Matlab® and will be made available upon request.

2.8 References

1. Kass I & Horovitz A (2002) Mapping pathways of allosteric communication in GroEL by analysis of correlated mutations. *Proteins* 48(4):611-617.
2. Gobel U, Sander C, Schneider R, & Valencia A (1994) Correlated mutations and residue contacts in proteins. *Proteins* 18(4):309-317.
3. Lockless SW & Ranganathan R (1999) Evolutionarily conserved pathways of energetic connectivity in protein families. *Science* 286(5438):295-299.
4. Cover TM & Thomas JA (1991) *Elements of information theory* (Wiley, New York) pp xxii, 542 p.
5. Tillier ER & Lui TW (2003) Using multiple interdependency to separate functional from phylogenetic correlations in protein alignments. *Bioinformatics* 19(6):750-755.
6. Fares MA & Travers SA (2006) A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics* 173(1):9-23.
7. Chiu DK & Kolodziejczak T (1991) Inferring consensus structure from nucleic acid sequences. *Comput Appl Biosci* 7(3):347-352.
8. Wollenberg KR & Atchley WR (2000) Separation of phylogenetic and functional associations in biological sequences by using the parametric bootstrap. *Proc Natl Acad Sci U S A* 97(7):3288-3291.
9. Olmea O, Rost B, & Valencia A (1999) Effective use of sequence correlation and conservation in fold recognition. *J Mol Biol* 293(5):1221-1239.
10. Korber BT, Farber RM, Wolpert DH, & Lapedes AS (1993) Covariation of mutations in the V3 loop of human immunodeficiency virus type 1 envelope protein: an information theoretic analysis. *Proc Natl Acad Sci U S A* 90(15):7176-7180.
11. Martin LC, Gloor GB, Dunn SD, & Wahl LM (2005) Using information theory to search for co-evolving residues in proteins. *Bioinformatics* 21(22):4116-4124.
12. Shannon CE & Weaver W (1949) *The Mathematical Theory of Communication*.
13. Fodor AA & Aldrich RW (2004) Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins* 56(2):211-221.
14. Dunn SD, Wahl LM, & Gloor GB (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics* 24(3):333-340.
15. Weigt M, White RA, Szurmant H, Hoch JA, & Hwa T (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* 106(1):67-72.
16. Dingwall C*, et al.* (1990) HIV-1 tat protein stimulates transcription by binding to a U-rich bulge in the stem of the TAR RNA structure. *EMBO J* 9(12):4145-4153.
17. Roy S, Delling U, Chen CH, Rosen CA, & Sonenberg N (1990) A bulge structure in HIV-1 TAR RNA is required for Tat binding and Tat-mediated trans-activation. *Genes Dev* 4(8):1365-1373.
18. Spanggord RJ, Vuyisich M, & Beal PA (2002) Identification of binding sites for both dsRBMs of PKR on kinase-activating and kinase-inhibiting RNA ligands. *Biochemistry* 41(14):4511-4520.

19. Carpick BW*, et al.* (1997) Characterization of the solution complex between the interferon-induced, double-stranded RNA-activated protein kinase and HIV-I trans-activating region RNA. *J Biol Chem* 272(14):9510-9516.

20. Kubota S*, et al.* (1989) Functional similarity of HIV-I rev and HTLV-I rex proteins: identification of a new nucleolar-targeting signal in rev protein. *Biochem Biophys Res Commun* 162(3):963-970.

21. Daly TJ, Cook KS, Gray GS, Maione TE, & Rusche JR (1989) Specific binding of HIV-1 recombinant Rev protein to the Rev-responsive element in vitro. *Nature* 342(6251):816-819.

22. Zapp ML & Green MR (1989) Sequence-specific RNA binding by the HIV-1 Rev protein. *Nature* 342(6250):714-716.

23. Bohnlein E, Berger J, & Hauber J (1991) Functional mapping of the human immunodeficiency virus type 1 Rev RNA binding domain: new insights into the domain structure of Rev and Rex. *J Virol* 65(12):7051-7055.

24. Holland SM, Ahmad N, Maitra RK, Wingfield P, & Venkatesan S (1990) Human immunodeficiency virus rev protein recognizes a target sequence in rev-responsive element RNA within the context of RNA secondary structure. *J Virol* 64(12):5966-5975.

25. Heaphy S*, et al.* (1990) HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the Rev response element region. *Cell* 60(4):685-693.

26. Tiley LS, Malim MH, Tewary HK, Stockley PG, & Cullen BR (1992) Identification of a high-affinity RNA-binding site for the human immunodeficiency virus type 1 Rev protein. *Proc Natl Acad Sci U S A* 89(2):758-762.

27. Cook KS*, et al.* (1991) Characterization of HIV-1 REV protein: binding stoichiometry and minimal RNA substrate. *Nucleic Acids Res* 19(7):1577-1583.

28. Daly TJ*, et al.* (1993) Biochemical characterization of binding of multiple HIV-1 Rev monomeric proteins to the Rev responsive element. *Biochemistry* 32(39):10497-10505.

29. Kjems J, Brown M, Chang DD, & Sharp PA (1991) Structural analysis of the interaction between the human immunodeficiency virus Rev protein and the Rev response element. *Proc Natl Acad Sci U S A* 88(3):683-687.

30. Zemmel RW, Kelley AC, Karn J, & Butler PJ (1996) Flexible regions of RNA structure facilitate co-operative Rev assembly on the Rev-response element. *J Mol Biol* 258(5):763-777.

31. Malim MH, Bohnlein S, Hauber J, & Cullen BR (1989) Functional dissection of the HIV-1 Rev trans-activator--derivation of a trans-dominant repressor of Rev function. *Cell* 58(1):205-214.

32. Malim MH & Cullen BR (1991) HIV-1 structural gene expression requires the binding of multiple Rev monomers to the viral RRE: implications for HIV-1 latency. *Cell* 65(2):241-248.

33. Fischer U, Huber J, Boelens WC, Mattaj IW, & Luhrmann R (1995) The HIV-1 Rev activation domain is a nuclear export signal that accesses an export pathway used by specific cellular RNAs. *Cell* 82(3):475-483.

34. Pollard VW & Malim MH (1998) The HIV-1 Rev protein. *Annual review of microbiology* 52:491-532.

35. Fornerod M, Ohno M, Yoshida M, & Mattaj IW (1997) CRM1 is an export receptor for leucine-rich nuclear export signals. *Cell* 90(6):1051-1060.

36. Fritz CC, Zapp ML, & Green MR (1995) A human nucleoporin-like protein that specifically interacts with HIV Rev. *Nature* 376(6540):530-533.

37. Bogerd HP, Fridell RA, Madore S, & Cullen BR (1995) Identification of a novel cellular cofactor for the Rev/Rex class of retroviral regulatory proteins. *Cell* 82(3):485-494.

38. McDonald D, Hope TJ, & Parslow TG (1992) Posttranscriptional regulation by the human immunodeficiency virus type 1 Rev and human T-cell leukemia virus type I Rex proteins through a heterologous RNA binding site. *J Virol* 66(12):7232-7238.

39. Venkatesan S, Gerstberger SM, Park H, Holland SM, & Nam Y (1992) Human immunodeficiency virus type 1 Rev activation can be achieved without Rev-responsive element RNA if Rev is directed to the target as a Rev/MS2 fusion protein which tethers the MS2 operator RNA. *J Virol* 66(12):7469-7480.

40. Tiley LS, Malim MH, & Cullen BR (1991) Conserved functional organization of the human immunodeficiency virus type 1 and visna virus Rev proteins. *J Virol* 65(7):3877-3881.

41. Brice PC, Kelley AC, & Butler PJ (1999) Sensitive in vitro analysis of HIV-1 Rev multimerization. *Nucleic Acids Res* 27(10):2080-2085.

Chapter 3: Mutual Information Analysis Reveals Coevolving Residues in Tat that Compensate for Two Distinct Functions in HIV-1 Gene Expression

3.1 Introduction

Genomes are continuously subjected to mutations that can in many cases undermine the structure and function of their encoded proteins. These processes may be especially important in rapidly evolving genomes, such as those of RNA viruses, which feature high rates of mutation and recombination during replication (1,2). For example, the retrovirus Human Immunodeficiency Virus-1 (HIV-1), exhibits enormous sequence diversity – both within individual patients and among numerous subtypes and recombinants circulating throughout the world – that in many cases reduces viral fitness but can also promote its ability to adapt to different selective pressures applied by the host immune system and anti-retroviral drugs (2-6). In many cases, purifying selection purges mutations that reduce overall fitness; however, deleterious mutations at one residue may also be compensated for by mutations at other sites to maintain protein structure and function (7). Such compensating positions within a protein may conceal significant evolutionary and functional information, yet are not readily apparent from an analysis of protein sequence. We use an approach whereby discovering coevolving sites within the HIV-1 protein Tat (Transactivator of Transcription) allows us to elucidate the underlying functional mechanism that constrains these sites to certain residue pairs for optimal function of this important viral protein.

Tat, which displays complex interactions with several cellular and viral factors that are critical to activating gene expression from the viral promoter, is an interesting substrate for analysis of the effects of protein evolution on complex, multifaceted protein functions. Briefly, once HIV-1 infects a host cell and integrates into its genome, transcription factor binding sites within the viral promoter recruit host cellular factors and mediate a low, basal level of gene expression in which transcriptional elongation is inefficient and yields primarily abortive transcripts (8). However, a small number of full length viral transcripts are produced and spliced to yield a mRNA species that encodes Tat. The few Tat molecules that are formed from this basal transcription bind to the positive transcription-elongation factor b (P-TEFb), which consists of the cellular proteins cyclin T1 (CycT1) and cyclin-dependent kinase 9 (Cdk9) (9). The resulting Tat-P-TEFb complex then binds to a RNA hairpin TAR (Transactivation Response Element) present at the 5' end of all viral transcripts and is subsequently transferred to the pre-initiation complex (PIC), wherein Cdk9 phosphorylates the C-terminal domain (CTD) of RNA polymerase II (RNAPII) and thereby greatly increases RNAPII processivity (10,11). In a recently proposed, alternate mechanism, P-TEFb and Tat may be recruited to the viral promoter in an inactive form, and the newly synthesized TAR may then bind to Tat and P-TEFb and displace the inhibitory 7SK snRNP to activate P-TEFb, which phosphorylates RNAPII (12). In either case, the increased RNAPII processivity greatly elevates viral gene expression and initiates a cascade of HIV-1 replication.

In addition to its interactions with RNAPII through P-TEFb, the multifunctional Tat interacts with numerous other host factors, and a balance among these various interactions and post-translational modifications is likely necessary for effective overall function (13-23).

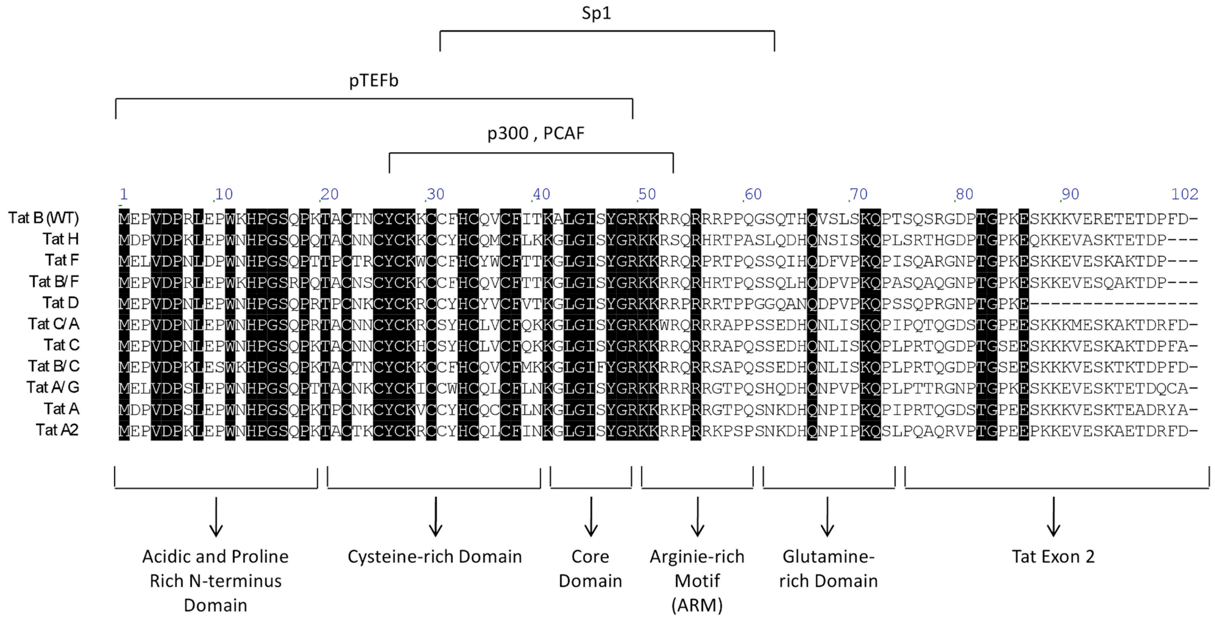

Sp1
pTEFb
p300 , PCAF
1
10
20
30
40
50
60
70
80
90
102
Tat B (WT) MEPVDPRLEPWKHPGSQPKTACTNCYCKKCCFHCQVCFITKALGISYGRKKRRQRRRPPQGSQTHQVSLSKQPTSQSRGDPTGPKESKKKVERETETDPFD-
Tat H MDPVDPKLEPWNHPGSQPQTACNNCYCKKCCYHCQMCFLKKGLGISYGRKKRSQRHRTPASLQDHQNSISKQPLSRTHGDPTGPKEQKKEVASKTETDP---
Tat F MELVDPNLDPWNHPGSQPTTPCTRCYCKWCCFHCYWCFTTKGLGISYGRKKRRQRPRTPQSSQIHQDFVPKQPISQARGNPTGPKESKKEVESKAKTDP---
Tat B/ F MEPVDPRLEPWKHPGSRPQTACNSCYCKKCCFHCQVCFTTKGLGISYGRKKRRQRHRTPQSSQLHQDPVPKQPASQAQGNPTGPKESKKEVESQAKTDP---
Tat D MEPVDPNLEPWNHPGSQPRTPCNKCYCKRCCYHCYVCFVTKGLGISYGRKKRRPRRRTPPGGQANQDPVPKQPSSQPRGNPTGPKE----------------
Tat C/ A MEPVDPNLEPWNHPGSQPRTACNNCYCKRCSYHCLVCFQKKGLGISYGRKKWRQRRRAPPSSEDHQNLISKQPIPQTQGDSTGPEESKKKMESKAKTDRFD-
Tat C MEPVDPNLEPWNHPGSQPKTACNNCYCKHCSYHCLVCFQKKGLGISYGRKKRRQRRRAPQSSEDHQNLISKQPLPRTQGDPTGSEESKKKVESKAKTDPFA-
Tat B/ C MEPVDPKLESWKHPGSQPKTACTNCYCKKCCFHCQVCFMKKGLGIFYGRKKRRQRRSAPQSSEDHQNLISKQPLPRTQGDPTGSEESKKKVESKTKTDPFD-
Tat A/ G MELVDPSLEPWNHPGSQPTTACNKCYCKICCWHCQLCFLNKGLGISYGRKKRRRRRGTPQSHQDHQNPVPKQPLPTTRGNPTGPKESKKEVESKTETDQCA-
Tat A MDPVDPSLEPWNHPGSQPKTPCNKCYCKVCCYHCQCCFLNKGLGISYGRKKRKPRRGTPQSNKDHQNPIPKQPIPRTQGDSTGPEESKKKVESKTEADRYA-
Tat A2 MEPVDPKLEPWNHPGSQPKTACTKCYCKRCCYHCQLCFINKGLGISYGRKKRRPRRKPSPSNKDHQNPIPKQSLPQAQRVPTGPEEPKKEVESKAETDRFD-
Acidic and Proline Rich N-terminus Domain
Cysteine-rich Domain
Core Domain
Arginie-rich Motif (ARM)
Glutamine-rich Domain
Tat Exon 2

Figure 3.1. Representative sequence alignment for HIV-1 Tat proteins of different subtypes. Sites in Tat that are conserved across these sequences are shaded in black. The domains of Tat that interact with a few well-known cellular proteins are shown in the upper half of the sequence alignment. The different domains of Tat are indicated in the lower half of the sequence alignment. The Tat B sequence is referred to as the wild-type (WT) sequence in the main text, into which mutations are introduced to identify coevolving residues.

However, it is unclear how Tat, despite its considerable sequence diversity among HIV-1 isolates globally, is able to mediate these critical processes of co-opting a series of host cellular mechanisms to orchestrate viral replication. Sequence alignment of a protein can enable the identification of single amino acids that are conserved and hence potentially important for specific functions, and a number of such sites have been identified within Tat (e.g. Fig. 3.1) (13,16). However, we hypothesized that correlated and coordinated amino acid changes may have played a role in diversifying Tat's sequence while preserving its interactions with many host proteins and thus its overall function (13-23). Furthermore, sequence alignments can readily miss situations where neither of two given residues is individually conserved, but instead where correlated pairs that make important contributions to protein structure and/or function appear together. Thus, bioinformatic and statistical approaches may help identify such sites and thereby gain greater molecular insights into a protein critical for HIV pathogenicity.

To address these hypotheses, we applied a statistical measure termed Mutual Information (MI) (24), one of several methods that can be used to identify correlated position pairs from multiple sequence alignments of proteins (25-27). Previously, such statistical measures have been applied to HIV-1 proteins, including the V3 loop of the *env* gene and the *gag* gene (28-30); however, these elegant computational analyses were not accompanied by experimental investigation. Similarly, such analysis has also been applied to other biological systems (31).

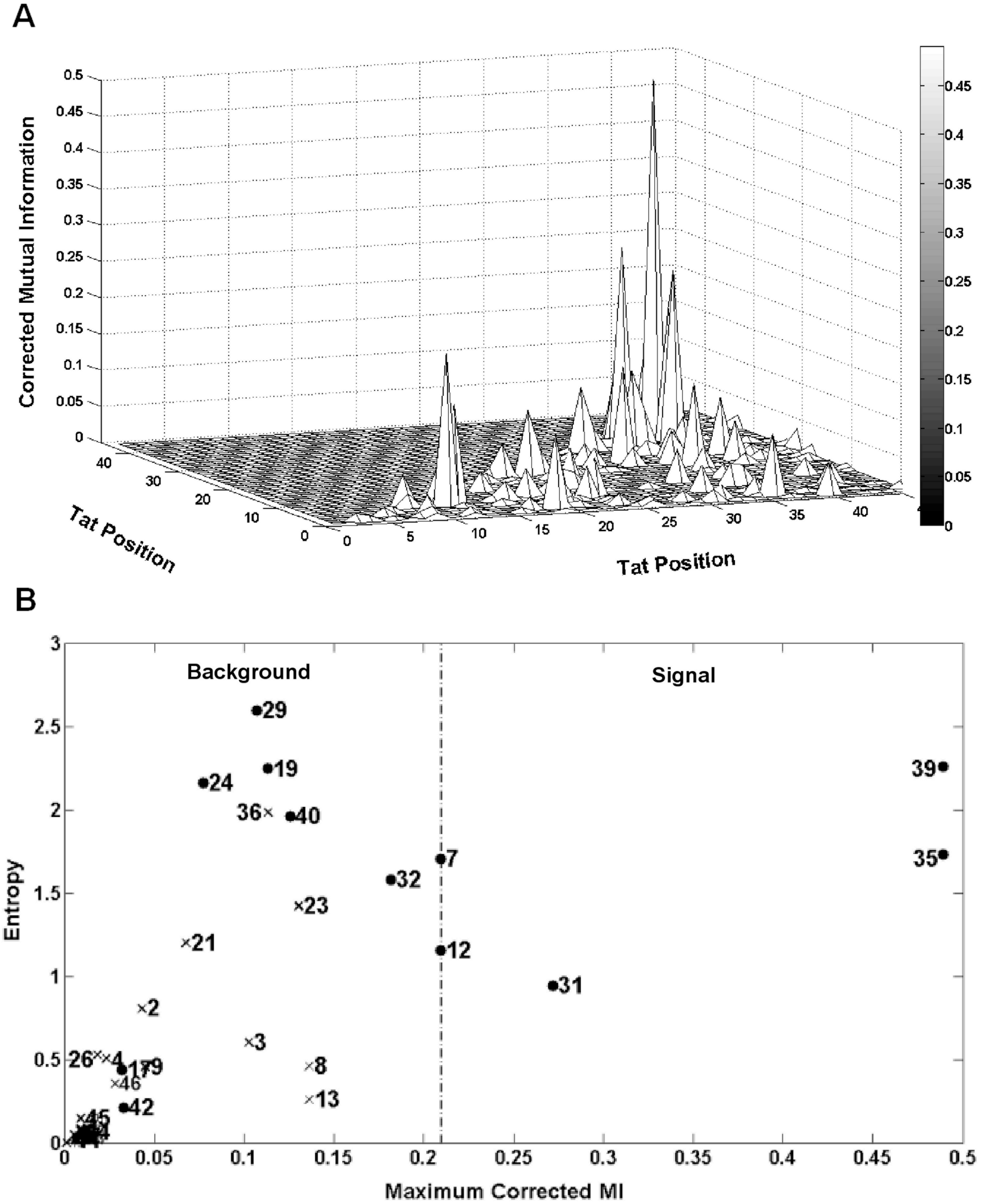


Figure 3.2. MI analysis applied to the Los Alamos Sequence Database reveals correlated position pairs in Tat. (A) MI scores between all possible position pairs within the first 46 residues in Tat. Sites (35,39), (31,35) and (31,39) have the highest scores. (B) Plot of Entropy of a site (a measure of amino acid conservation at a site) vs.

Maximum MI score for that site with any other site in Tat. The dotted line denotes the threshold MI score used to separate signal from background. Black solid circles represent experimentally tested positions in Fig. 3.4.

Because the background noise in multiple sequence alignments makes it difficult for most statistical measures to predict correlated residues accurately, accompanying experimental validation of these sites is critical to identify structurally or functionally constrained residue pairs.

Here, we present a combined computational and experimental approach to identify and investigate coevolving residues in Tat. Sites 35 and 39 emerged in this analysis, and the functional importance of these coevolving residues was verified experimentally by introducing single point mutations in Tat. While the single mutants proved non-functional, adding the second mutation restored viral gene expression. Surprisingly, despite their structural proximity, positions 35 and 39 appeared to be important for two distinct, underlying mechanisms – Tat binding to P-TEFb and Tat-mediated activation of P-TEFb to enable it to phosphorylate the CTD of RNAPII – and a combination of these two functions constrains the identities of these residues to certain pairs of amino acids.

Extending this analysis indicates that the Tat proteins of HIV-1 subtypes B and C appear to have evolved compensatory strengths for different steps of Tat-mediated transactivation to achieve similar overall viral gene expression levels.

3.2 Mutual Information Analysis Identifies Sites 35 and 39 in Tat as Coevolving

To identify correlated sites within Tat that are potentially important for maintaining Tat structure or function, we calculated MI between position pairs for 917 pre-aligned Tat sequences, from 9 viral subtypes and 14 recombinant forms, from the Los Alamos Sequence Database (http://www.hiv.lanl.gov/). To encode a large amount of information within a relatively small genome, HIV-1 uses overlapping reading frames and alternative splicing. Since Tat shares overlapping reading frame with another viral protein, Rev, beyond amino acid 47, analysis was restricted to the first 46 amino acids of Tat within its activation domain (amino acids 1-48) to ensure that the sequence conservation and structural constraints in Rev did not introduce false positives in the analysis.

Within a multiple sequence alignment, MI predicts the likelihood of observing an amino acid at a particular site *X* in an alignment, given the identity of the amino acid at another site *Y*, and is computed by:

$$I(X,Y) = \sum_{x,y=1}^{21} p(x,y)\log\frac{p(x,y)}{p(x)p(y)}$$

where $X,Y \in (1,2.....21)$ corresponds to one of the 20 amino acids or an alignment gap. $p(x)$ and $p(y)$ correspond to the probability of observing amino acid (or gap) x or y at sites X and Y, respectively, and $p(x,y)$ is the corresponding joint probability of observing amino acids (or gaps) x and y at sites X and Y. Higher MI values indicate stronger correlation between two sites.

The finite size of the sequence database and the phylogenetic relationship shared among these sequences result in non-zero background MI values.

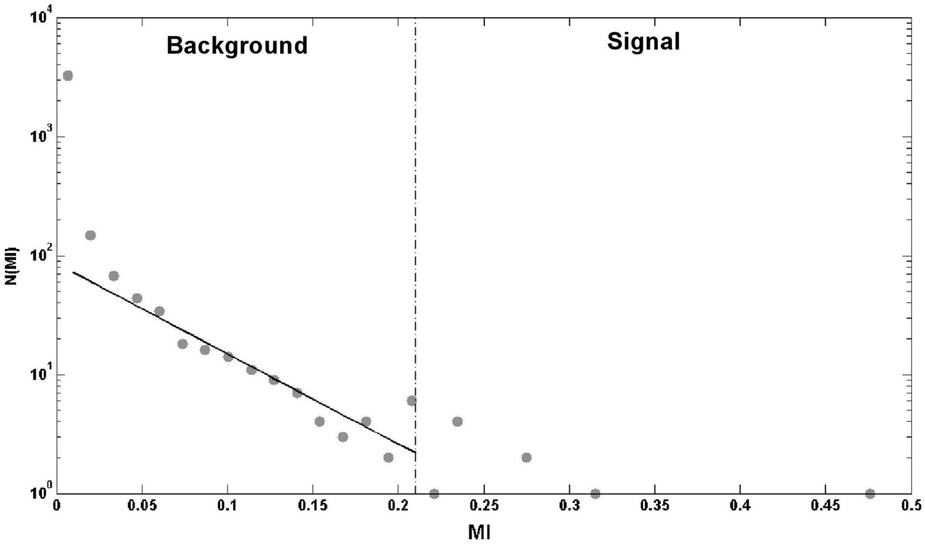


Figure 3.3. Identification of a threshold MI to distinguish correlated position pairs from other pairs. Corrected MI scores for all position pairs in Tat are plotted against their frequency of occurrences. The threshold MI score is indicated by the dashed line. Points to the left of the dashed line indicate the exponential background. Deviation from the exponential function is used to identify the threshold MI as 0.21.

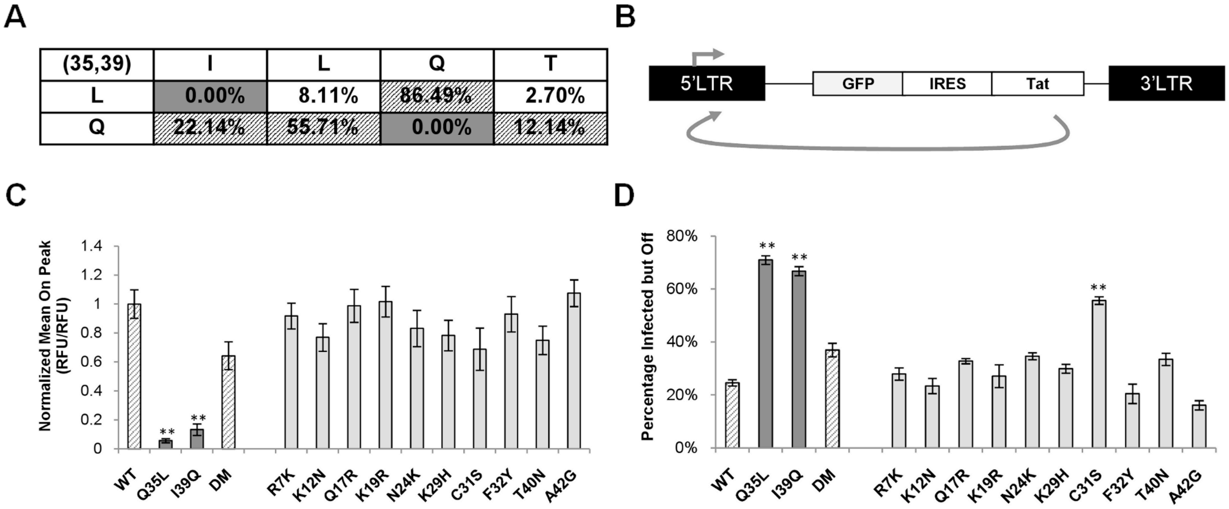
To minimize background noise, we subtracted background scores from raw MI scores using a method described by Dunn *et. al.* (32). After this normalization, unconstrained position pairs had values close to zero (Fig. 3.2*A*). To identify a threshold MI that distinguishes position pairs that are strongly statistically correlated and hence possibly coevolving from ones that do not interact with each other, a histogram of the corrected MI scores was constructed and fitted to an exponential function, as described by Weigt *et. al.*(Fig. 3.3) (33). Deviation from the exponential function was used to identify the threshold MI score as 0.21. Three position pairs – (35,39), (31,35), and (31,39) – had MI scores higher than the threshold (Fig. 3.2) (32,33).

The z-scores, defined as the number of standard deviations from the mean MI score for all position pairs within the activation domain of Tat, a measure of the strength of correlation between two sites, were 17.96, 9.94 and 8.62 for the position pairs (35,39), (31,35) and (31,39) respectively, against the mean z-score of 0.34 for all position pairs within the activation domain of Tat.

To ensure that these outcomes did not depend on the method used for background correction, an alternative method, in which the background was computed by creating a random sequence alignment in which the amino acid frequency at each site was the same as that in the

Tat sequence database, was used and yielded very similar corrected MI scores, with the highest pair (35,39) unchanged (24).

For positions 35 and 39, frequencies of the different amino acids, based on the Tat sequence database, show that a Leu at position 35 constrains position 39 primarily to a Gln.



(35,39)	I	L	Q	T
L	0.00%	8.11%	86.49%	2.70%
Q	22.14%	55.71%	0.00%	12.14%

Figure 3.4. Experimental identification of coevolving sites in Tat using gene expression studies. (A) Truncated 2x4 matrix showing amino acid residues usually observed at sites 35 and 39 from 917 sequences in the Los Alamos Sequence Database, rather than a sparse 21x21 matrix representing all the amino acids (and gaps) at sites 35 and 39 in Tat. In the complete matrix, each row sums to 100%. Shaded cells represent residues pairs commonly observed at sites 35 and 39. Gray cells represent amino acid pairs not observed in the database. (B) Schematic of the lentiviral vector (LGIT) used to study gene expression for different Tat variants. Jurkat cells were infected with the LGIT vector at low MOIs (0.05 – 0.1) to ensure single integration event per cell. (C) Gene expression levels based on GFP fluorescence for different Tat mutants normalized by WT Tat. (D) Percentage Infected but Off, a measure of the fraction of cells that are silenced and not expressing GFP. The shading of bars in (C) and (D) correlate with the matrix in (A) for easy visualization. Error bars represent S.D. for 3 independent infections. '**' denotes statistically significant differences (p<0.01) from WT Tat.

Similarly, a Gln at 35 results in a majority of Tat sequences having an Ile, Leu, or Thr but not a Gln at 39 (Fig. 3.4*A*). Based on the MI analysis, positions 31, 35, and 39 were chosen for experimental testing. In addition, a relatively conserved site that is coevolving with another site may in general have a low MI score due to the following mathematical constraints (24):

$$0 \leq I(X,Y) \leq \min\{H(X), H(Y)\}$$

where $H(X)$ stands for the entropy of site X and $I(X,Y)$ stands for the MI score between sites X and Y. Entropy of a site is a measure of the degree of conservation of a site, with lower Entropies representing more conserved sites and a value of zero implying a completely conserved site. Thus, a conserved site may have low MI scores even if it is coevolving with another site because the mutual information is constrained to lie below the smaller entropy of the lesser variable site. Therefore, we experimentally tested a few sites, within the background region that had low Entropies (Fig. 3.2*B*). By including such low entropy sites within the background region, we ensure that we do not fail to identify a coevolving pair that is constrained by one of the sites having low Entropy. Furthermore, we tested several non-conserved high entropy sites within background MI as controls to validate the low predicted functional relationship between such sites (Fig. 3.2*B*). Experimentally tested positions are shown by solid black dots (Fig. 3.2*B*).

3.3 Gene Expression Analysis of Coevolving Sites 35 and 39

If two sites functionally coevolve, then mutating either individually may impair biological activity, whereas mutating both together may rescue function. To test this hypothesis, amino acids in Tat from subtype B virus, broadly used in HIV-1 studies and referred to here as wild-type (WT) Tat (Fig. 3.1), were replaced with residues of other naturally occurring viral variants or subtypes (Fig. 3.4*A*). To test how sites with high MI scores predict Tat function, we studied the gene expression properties of different Tat mutants using a lentiviral vector model of HIV-1, in which the HIV-1 LTR drives expression of green fluorescent protein (GFP) and Tat, separated by an Internal Ribosome Entry Sequence (IRES) (LTR-GFP-IRES-Tat or LGIT) (Fig. 3.4*B*) (34). GFP expression from single integrations of LGIT in Jurkat cells was used to quantify LTR gene expression, and as previously observed the Tat positive feedback loop results in a bifurcated cell population with either very low or high levels of GFP expression, referred to as the Off and On populations, respectively (Fig. 3.5*A*) (34). Two metrics were used to quantify gene expression (35). First, the Mean On Peak indicates the average GFP level of cells above the background threshold of fluorescence (the On gate), a measure of the level of “closed-loop” transactivation attained by a particular Tat mutant. Second, the Percentage Infected but Off is the fraction of infected cells that are in the Off population, but that can be stimulated to express Tat and GFP via the addition of TNF-α (a strong activator of the NF-κB pathway, which directly activates the LTR) and TSA (an inhibitor of histone deacetylases that also activates HIV gene expression) (35). This metric is a measure of the inability of a particular Tat mutant to activate gene expression from the viral LTR.

When introduced into WT Tat, single point mutations Q35L or I39Q, chosen from the matrix in Fig. 3.4*A*, yielded non-expressing virus (Figs. 3.4*C* and 3.5*A*). The Percentage of Infected but Off cells for these single mutants was approximately 70%, nearly three times higher than WT Tat, again indicating that these Tat mutants fail to activate gene expression from the viral LTR (Fig. 3.4*D*). Strikingly, however, the introduction of both mutations into the same Tat sequence (henceforth referred to as the double-mutant, DM) rescued Mean On Peak levels close to that of WT Tat, with a similar fraction of silenced cells (Figs. 3.4*C*,*D* and 3.5*A*). Similarly, transfection of the WT, Q35L, I39Q, and DM Tat into a HeLa cell line containing a HIV-1 LTR Luciferase reporter showed analogous trends (Fig. 3.5*B*). From the 917 Tat sequences used in the MI analysis, 124 sequences shared the same Gln35-Ile39 residue pair as WT Tat, and 262 sequences shared the same Leu35-Gln39 residue pair as DM Tat, suggesting that both residue pairs are found in naturally occurring Tat sequences. Furthermore, to show that coevolution between sites 35 and 39 extend to another Tat subtype, we made single point mutants L35Q and Q39I of subtype C Tat. As previously observed for single mutations of Tat B, these mutants resulted in dramatic loss of gene expression and a three-fold increase in Percentage Infected but Off cells (Fig. 3.6*C,D*). However, gene expression was restored in the Tat C double-mutant, suggesting that sites 35 and 39 alone compensate for one other (Fig. 3.6*C,D*). Thus, evolutionarily only certain pairs of amino acids at sites 35 and 39 but not their intermediates are tolerated.

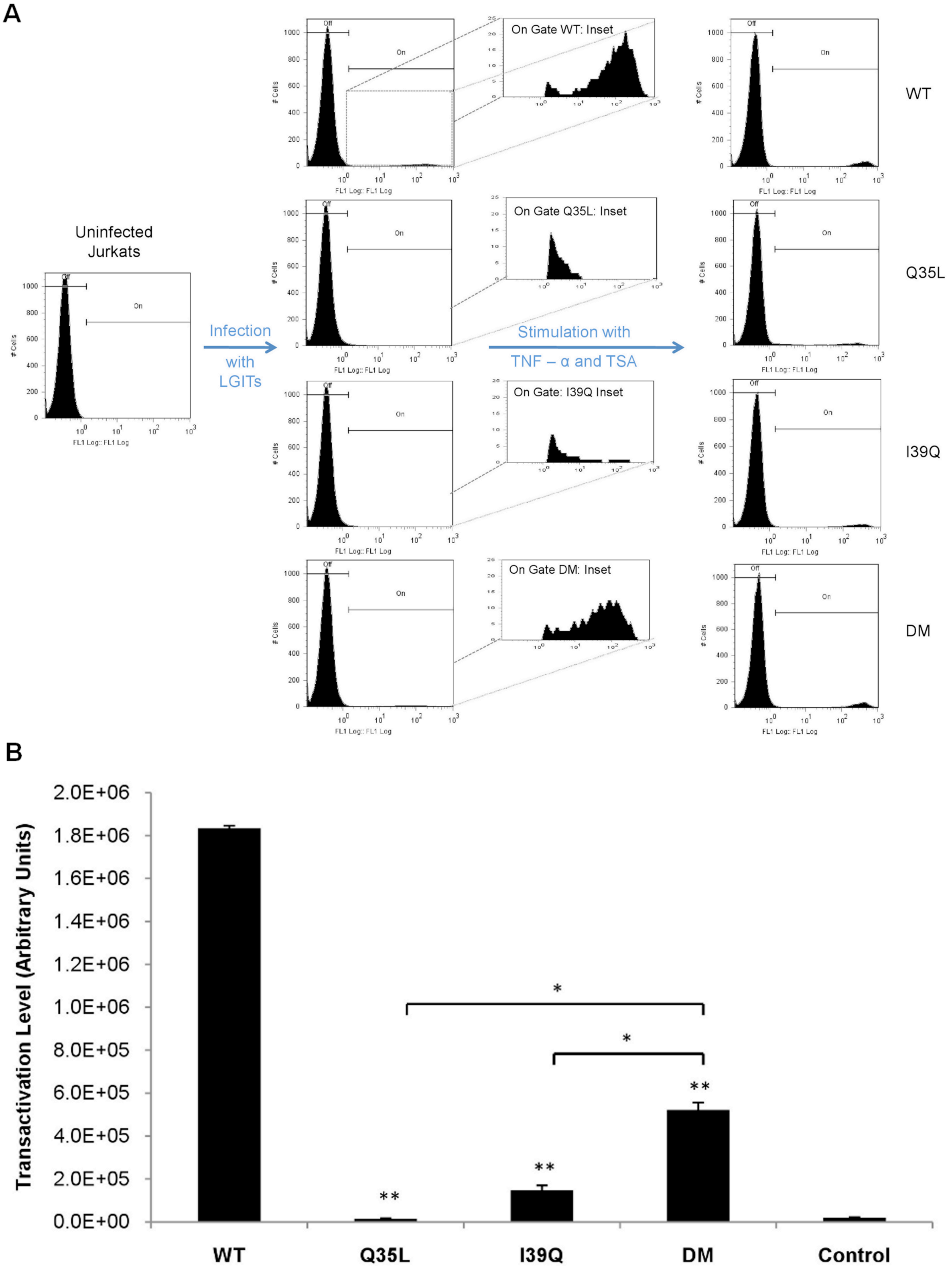


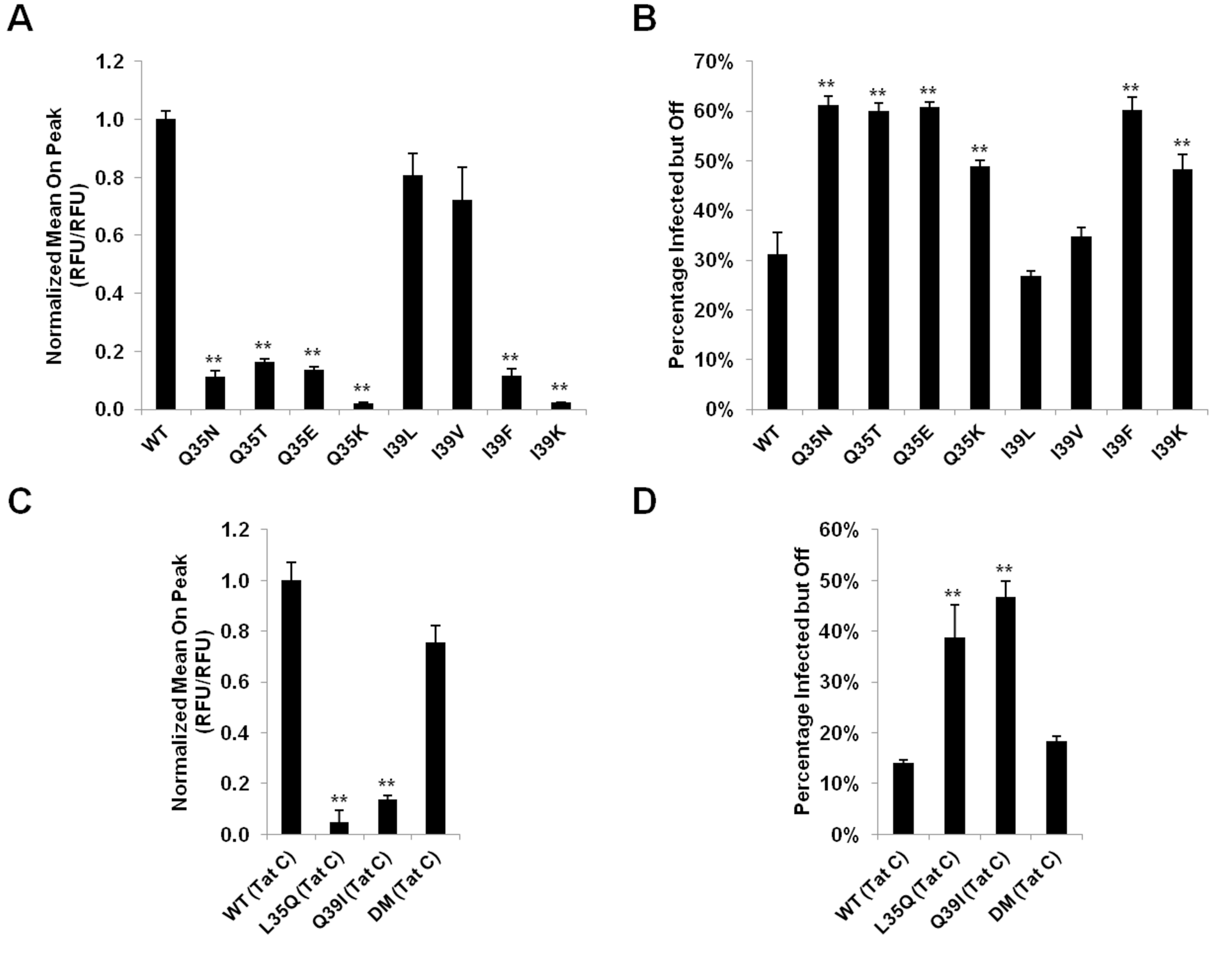
Figure 3.5. Additional experimental verification of coevolving sites 35 and 39. (A) The histogram to the left shows background fluorescence of uninfected Jurkat cells. Infection of Jurkat cells at low MOIs (0.05-0.1) with the LGIT vector containing WT, Q35L, I39Q or the DM Tat, to ensure a single integration event per cell, gave rise to a bimodal population of low and high GFP fluorescence, corresponding to cells in the Off and On gate, shown in the middle. Insets show the distribution of GFP fluorescence for the different Tat variants within the On gate. Histograms of the infected Jurkat cells after stimulation with TNF-α and TSA are shown to the right. The difference in the number of cells in the On gate before and after stimulation is used to calculate Percentage Infected but Off, the fraction of cells that are infected with the LGIT vector originally but not expressing GFP. (B) Experimental identification of coevolving residues in Tat using transfection of the UbChIT vector containing different Tat variants into a HeLa cell line containing a HIV-1 LTR Luciferase reporter (HL3T1 cell line). Gene expression is dramatically reduced from the single Tat mutants, Q35L and I39Q. The DM partially rescues gene expression from the single mutants. An empty vector is used as a control. '**' denotes statistically significant differences ($p<0.05$) from WT Tat. '*' denotes statistically significant differences ($p<0.05$) from the DM Tat.

In analysis of site 31 in Tat B, the mutation C31S also yielded a slight reduction in gene expression and a statistical increase in the Percentage of Infected but Off cells as compared to WT Tat ($p<0.01$); however, coevolution between site 31 and site 35 or 39 proved difficult to investigate, as mutation at either of the latter two exerted a dominant loss of gene expression. However, the interaction between sites 31, 35, and 39 can be observed statistically. A Gln at site 39 constrains sites 31 and 35 primarily to a Ser and Leu, respectively. Similarly, a Leu, Ile or Thr at site 39 primarily restricts sites 31 and 35 to a Cys and Gln, respectively (Fig. 3.7). These interactions between sites 31-35-39 are discussed from a structural perspective in additional detail in the discussion section.

We next replaced Gln35 and Ile39 in WT Tat with other amino acids based on residues that either appear or do not appear at sites 35 or 39 in the Los Alamos Sequence Database. As anticipated, replacing Gln35 with similar polar residues that are not found (Q35N, Q35E and Q35K) or rarely found (Q35T) in the database, and are thus not predicted to coevolve with residues at site 39, resulted in non-functional Tat proteins (Fig. 3.6*A,B*). Similarly, replacing Ile39 in WT Tat with non-polar residues (I39F), or polar residues that occupy similar side-chain volume (I39K) but are not found in the database, gave rise to non-expressing viruses (Fig. 3.6*A,B*). In contrast, naturally occurring Tat sequences with a Gln at site 35 have residues such as Leu and Val in addition to Ile at site 39 (Fig. 3.4*A*, Val is not shown in this matrix). I39L and I39V Tat mutants yielded similar Mean On Peak and Percentage Infected but Off levels as WT Tat (Fig. 3.6*A,B*), validating the predictions from MI analysis. Experimental data from these site-directed mutations and predictions from the MI analysis also corrected well with structural analysis at these positions.

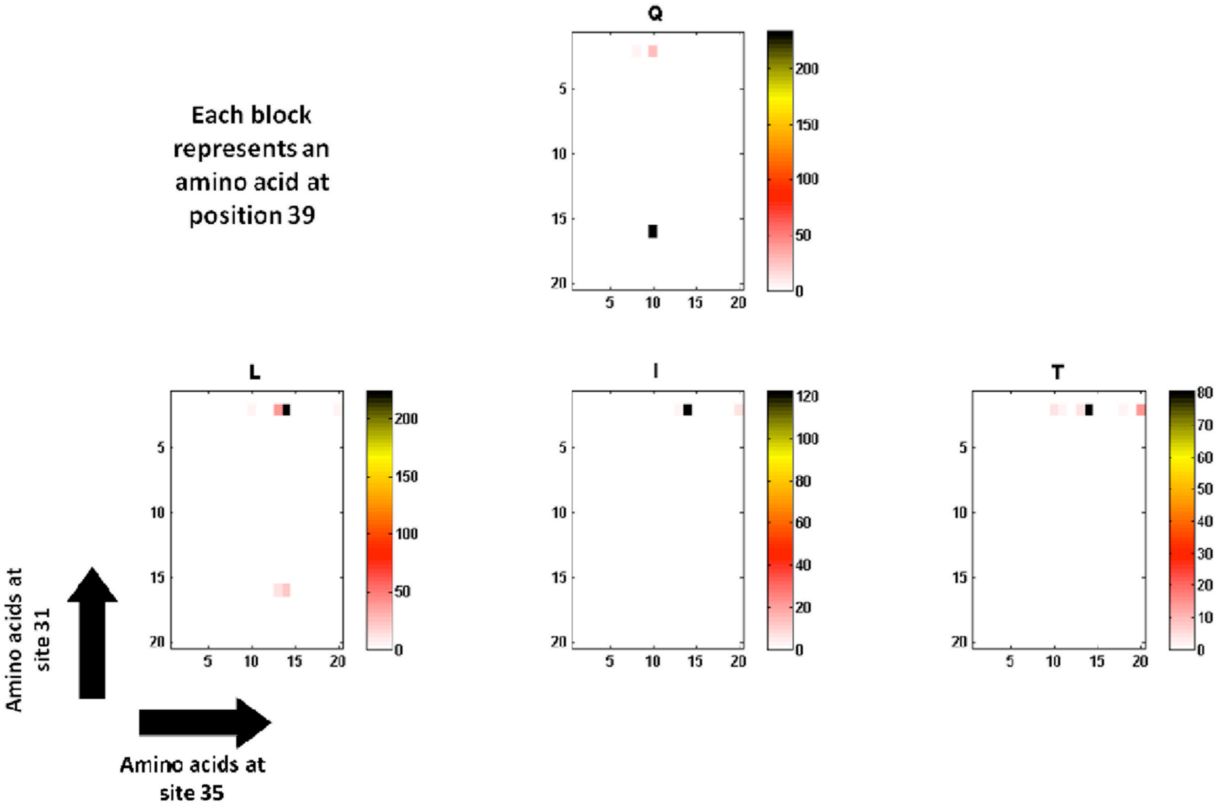
Structural analysis for the Q35N mutation shows that Asn35 of Tat fails to H-bond with Asn180 of CycT1. Compared to Gln at site 35 in Tat, the shorter side-chain length of Asn increases its distance from Asn180 of CycT1 that results in a loss of H-bonding. Other mutations at site 35 (Q35E, Q35T and Q35K) also result in a loss of H-bonding which may potentially be responsible for the loss of gene expression that is observed with these mutants (Fig. 3.6*A*).

Introducing mutations at site 39, such as I39F and I39K, results in a loss of gene expression (Fig. 3.6*A*). Replacing Ile39 in WT Tat with a non-polar residue Phe (I39F) shows that the structure of the Tat-P-TEFb complex (PDB: 3MI9) is destabilized by 3.7 kcal/mol*. Similarly, replacing Ile39 in WT Tat with the polar residue Lys (I39K) that occupies similar side-chain volume destabilizes the Tat-P-TEFb structure (PDB: 3MI9) by 5.92 kcal/mol*.



<u>Figure 3.6. Additional mutations at sites 35 and 39 further validate the MI analysis to establish these two sites as coevolving.</u> (A) and (B) Normalized Mean On Peak and Percentage Infected but Off for Tat variants with mutations at sites 35 and 39. As expected, mutants that are not found (Q35N, Q35E, Q35K, I39F and I39K) or rarely found (Q35T) in the Los Alamos Sequence Database fail to activate gene expression from the viral promoter with higher levels of Percentage Infected but Off cells than WT Tat. In contrast, the I39L and I39V mutants are also found in naturally occurring Tat sequences that have a Gln at site 35. Introducing these mutations in WT Tat does not alter the levels of the Mean On Peak or the Percentage of Infected but Off cells suggesting that only certain pairs of amino acids at sites 35 and 39 produce functional Tat protein. '**' denotes statistically significant differences ($p<0.05$) from WT Tat. (C) and (D) Normalized Mean On Peak and Percentage Infected but Off for Tat C mutants at sites 35 and 39. Single mutants, L35Q and Q39I give rise to non-functional Tat, whereas the presence of both mutations with the same Tat C sequence restores gene expression close to WT (Tat C) levels. Thus, sites 35 and 39 have coevolved with each other and this correlation between the two sites is not dependent on the subtype of Tat. '**' denotes statistically significant differences ($p<0.05$) from WT (Tat C) Tat.

The destabilization of the Tat-P-TEFb structure potentially explains the loss of gene expression associated with these mutations. In contrast, residues such as Leu and Val that are also found in naturally occurring Tat sequences at site 39, in addition to Ile, are associated with similar levels of gene expression as WT Tat (Fig. 3.6*A*). Analysis of the mutations I39L and I39V showed that the stability of Tat-P-TEFb structure (PDB: 3MI9) was almost unchanged (compared to WT Tat, the I39L and I39V mutations marginally stabilized the Tat-PTEFb structure by 0.02 kcal/mol* and 0.32 kcal/mol*, respectively), supporting the experimental evidence that these mutations result in similar levels of gene expression as WT Tat.



<u>Figure 3.7. Sites 31, 35 and 39 form a mini-network of coevolving residues.</u> The heat map shows the distribution of some commonly occurring amino acids at sites 31, 35 and 39 from Tat sequences obtained from the Los Alamos Sequence Database. Each block represents a 20x20 matrix of amino acids at sites 35 and 31 for a given amino acid at site 39, indicated at the top of that block. The x- and y-axis represent the 20 amino acids arranged alphabetically at sites 35 and 31, respectively. Thus, the upper block shows that a Gln at site 39 is primarily correlated with Leu at site 35 and Ser at site 31. Similarly, the three lower blocks show that a Leu, Ile or Thr at site 39 is primarily correlated with a Gln and Cys at sites 35 and 31, respectively.

Compared to sites above the threshold MI value, mutation of positions 7, 12, 17, 19, 24, 29, 32, 40, or 42, which were predicted to be within the background region from the MI analysis, did not show any statistical difference in the level of gene expression or the Percentage of Infected but Off cells compared to WT Tat (*p*> 0.01, Fig. 3.4*C*,*D*). These gene expression studies thus strongly support the *in silico* prediction that sites 35 and 39 strongly coevolve and are critical to ensure the primary function of Tat as a transactivator of the HIV-1 promoter (Fig. 3.4*C*,*D*).

Finally, WT and DM Tat have different pairs of amino acids at sites 35 and 39 yet induce similar levels of gene expression when present at high levels (Fig. 3.4*C*); however, we also wanted to explore their gene activation behavior at lower concentrations. Under these conditions, previous studies have shown that the Tat positive feedback loop is subject to stochastic fluctuations in Tat, one of several factors that may play a role in viral reactivation from latency (34,35). Since DM Tat has amino acids Leu and Gln at sites 35 and 39, respectively, residues shared by a majority of subtype C Tats at these positions, we included subtype C Tat in these studies to determine the contribution of sites 35 and 39 to this behavior. As described previously, cells were infected with LGIT variants at low MOI, and GFP+ cells were sorted by Fluorescence Activated Cell Sorting (FACS) 7 days post-infection after stimulation with TNF-α (35). The sorted cells were allowed to relax for 9 days, and GFP- cells infected with the LGIT vector, but not expressing GFP, were sorted (Fig. 3.8*A*). These silent proviruses were then monitored for GFP expression over the course of 12 days (Fig. 3.8*B*).

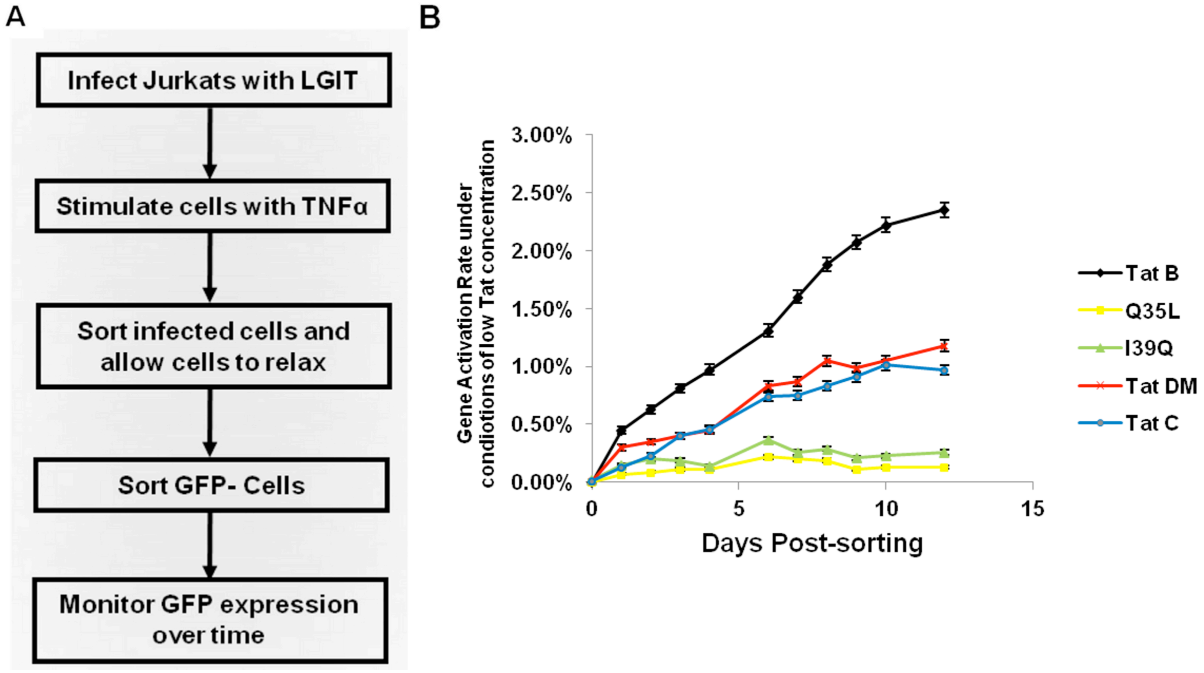


Figure 3.8. Gene activation under conditions of low Tat concentration. (A) Sorting scheme for isolating silent proviruses. Jurkat cells were infected with LGITs containing different Tat variants at a low MOI (to ensure single integration events per cell) and stimulated with TNF-α seven days post-infection, and GFP+ cells were isolated by FACS. The sorted cells were allowed to relax for 9 days and GFP- cells are then sorted from this population. GFP expression of this population was then tracked over time using flow cytometry. (B) Gene activation rates over time for different Tat variants. Q35L and I39Q Tat fail to initiate reactivation whereas Tat DM and C, with the same Leu-Gln residue pair at sites 35-39, have similar gene activation rates that partially restores gene activation to subtype B Tat levels. Error bars represent S.D. obtained by bootstrapping using a bootstrap sample size of 2000.

As anticipated, the functionally inactive Q35L and I39Q Tat variants showed very low levels of gene activation. However, the DM Tat partially restored gene expression to WT Tat levels, and very closely tracked the activation rate of Tat C (Fig. 3.8*B*), a result that suggests that residue pairs at sites 35 and 39 may be important determinants in setting the gene activation levels for different Tat variants.

3.4 Coevolving Sites 35 and 39 Impact both P-TEFb Binding and Phosphorylation at the CTD of RNAPII

To identify potential molecular mechanisms that restore gene expression for the DM Tat, we reasoned that the compromised transactivation of either Tat single mutant may be due to disruption in its binding to one of numerous cellular factors necessary for efficient transactivation. For example, the activation domain of Tat (amino acids 1-48) has previously been shown to interact with P-TEFb, which mediates the critical phosphorylation of the CTD of RNAPII and thus the production of full-length viral transcripts (9,36). HeLa cells were transfected with plasmids to express FLAG-tagged Tat under the control of the human ubiquitin promoter (Ubiquitin-mCherry-IRES-Tat or UbChIT), and immunoprecipitates of Tat were probed for Cdk9 and CycT1 (Fig. 3.9*A*,*B*) (18). WT Tat bound P-TEFb; however, the Q35L Tat mutant failed to efficiently bind either Cdk9 or CycT1, suggesting that site 35 is critical for binding P-TEFb (Fig. 3.9*A*,*B*), and the loss of this binding possibly underlies the defective gene expression for this mutant (Fig. 3.4*C*,*D*). Similarly, other factors that have recently been shown to interact with the Tat-P-TEFb complex and aid in transcriptional activation – such as ENL, AF9, AFF4 and ELL2 – failed to bind to the Q35L Tat mutant (Fig. 3.10) (37). Interestingly, the DM Tat partially restores binding with Cdk9 and CycT1, likely the mechanism by which the I39Q mutation rescues the Q35L mutant's loss of function (Fig. 3.9*A*,*B*).

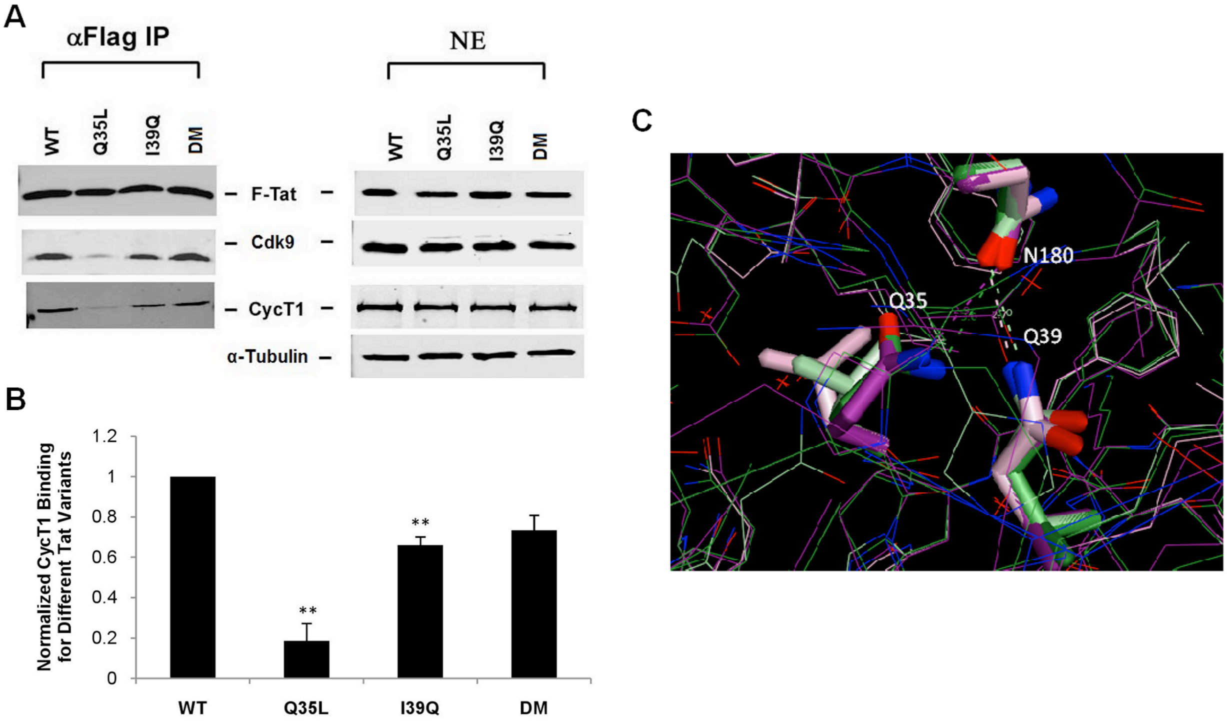


Figure 3.9. Co-immunoprecipitation and homology modeling shows that the Q35L Tat mutant fails to bind P-TEFb. (A) Immunoprecipitation (IP) of nuclear extracts (NE) with α-FLAG, obtained from HeLa cells transfected with the UbChIT vector, were followed by Western blots (WB) with α-Cdk9 and α-CycT1 antibodies. (B) Quantification of binding of different Tat mutants with CycT1. CycT1 is normalized to Tat, and its interaction with WT Tat is arbitrarily assigned the value 1. Error bars represent S.D. for two independent α-FLAG IPs and WBs. '**' denotes statistically significant differences ($p<0.05$) from WT Tat in CycT1 binding. (C) Interaction of P-TEFb with

WT and DM Tat are shown. Dark green and purple colors represent the ATP+ and ATP- structures for WT Tat, respectively, whereas light green and purple colors represent the ATP+ and ATP- structures for DM Tat. The structure shows hydrogen bonding between N180 in CycT1 with Q35 in WT Tat, as well as hydrogen bonding between N180 in CycT1 with Q39 in DM Tat. Hydrogen bonding is thus critical for Tat-P-TEFb binding.

To gain further insights into the loss of P-TEFb binding, we performed *in silico* modeling based on a recently solved structure of Tat-P-TEFb (36). These results indicated that Asn180 of CycT1 is positioned between and can form hydrogen bonds with a Gln at either Tat site 35 or 39.

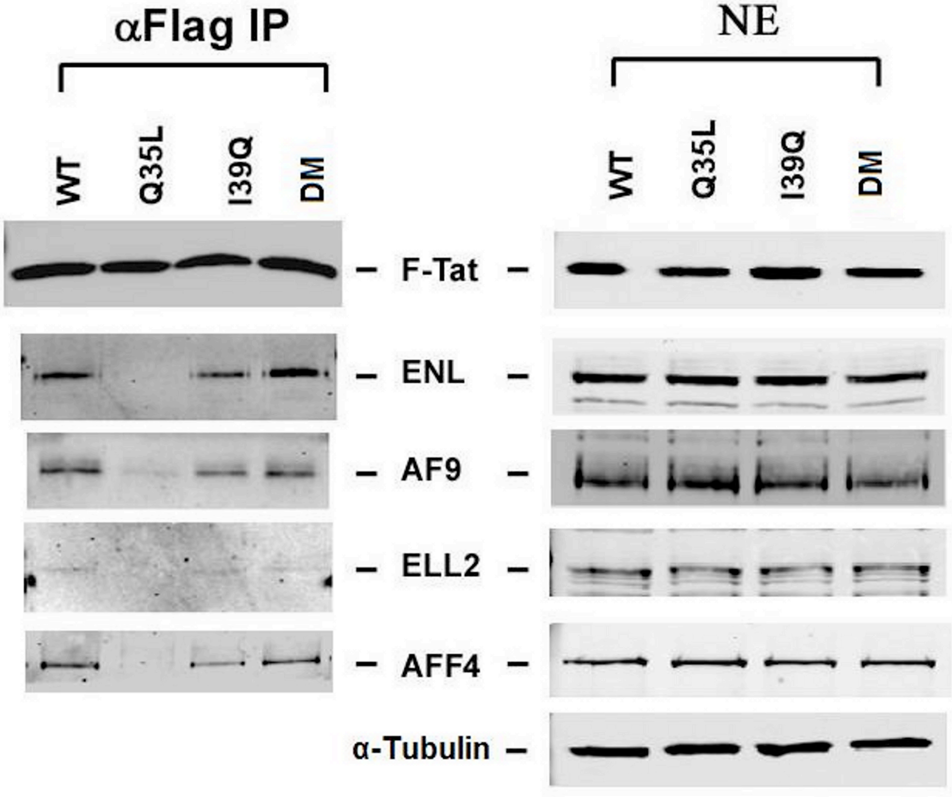


Figure 3.10. The Q35L Tat mutant fails to bind cellular factors that interact with the Tat-P-TEFb complex. The inability of the Q35L Tat mutant to bind P-TEFb also results in loss of binding with transcription factors AFF4, ENL, AF9 and elongation factor ELL2, that have been shown to interact with the Tat-P-TEFb complex. Immunoprecipitation of nuclear extracts with α-FLAG antibody, obtained from HeLa cells transfected with the UbChIT vector, were followed by western blots with α-AFF4, α-ENL, α-AF9 and α-ELL2 antibodies.

The Q35L mutation results in the loss of this hydrogen bonding, whereas the compensating mutation I39Q in the DM Tat enables Gln39 to replace this hydrogen bond with Asn180 in CycT1 (Fig. 3.9*C*). Although a previous study predicted that other naturally occurring mutations (except Tyr) could readily be accommodated at site 35 and maintain the structure of the protein complex (36), our analysis and accompanying experimental data suggest that the loss of hydrogen bonding may well be responsible for the drastic loss of function observed in the Q35L Tat mutant.

In contrast to the Q35L Tat mutant, however, the I39Q Tat mutant is able to bind CycT1 at levels close to the DM Tat (Fig. 3.9*B*), but lower than WT Tat, suggesting that its inability to activate gene expression (Fig. 3.4*C*) arises from reasons other than P-TEFb binding. To probe other transcriptional steps at which I39Q Tat may fail, we quantified viral transcripts. Cells infected with LGIT vectors containing one of the four Tat variants were stimulated with TNF-α seven days post-infection, and infected, GFP+ cells were isolated by FACS. The sorted cells were allowed to relax for 9 days (Fig. 3.11*A*), total cellular RNA was extracted, and the levels of viral transcripts were quantified using RT-qPCR (35). The I39Q and DM Tat both had similar percentages of elongated transcripts (Fig. 3.11*B*); however, the I39Q Tat has much lower levels of total transcripts compared to the DM Tat (Fig. 3.11*C*), suggesting that it fails to induce transcription at the same efficiency as the DM Tat.

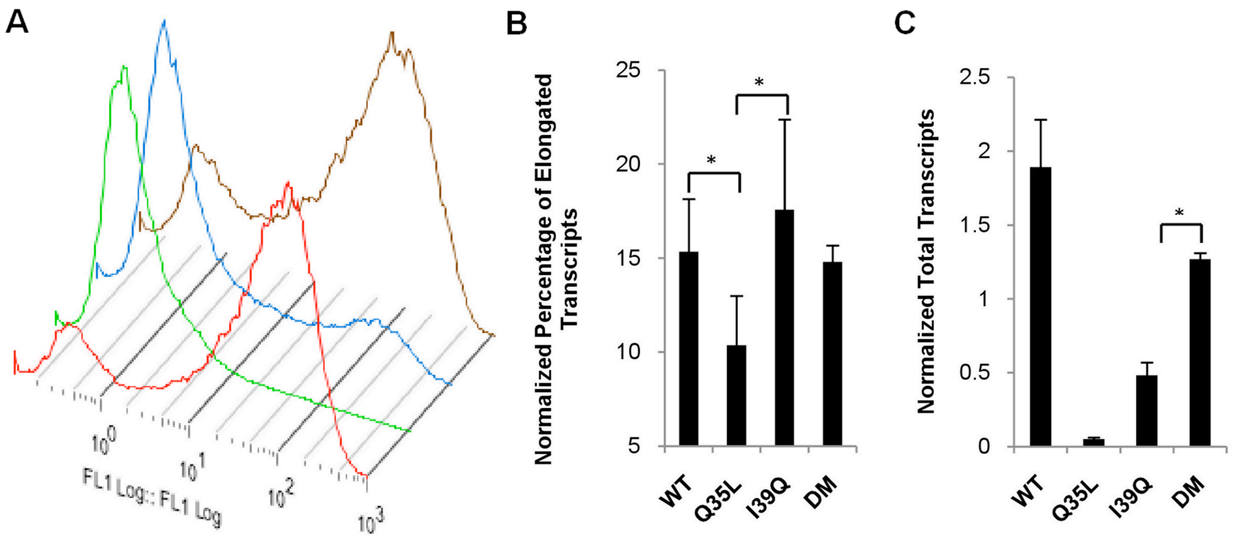


Figure 3.11. Viral transcript quantification reveals potential transcriptional step at which I39Q Tat may fail. (A) GFP histograms for Jurkat cells infected with wild-type (Red), Q35L (Green), I39Q (Blue), and DM (Brown) Tat 9 days post-sorting of TNF-α stimulated GFP+ cells. (B) and (C) Quantification of the percentage of elongated and total viral transcripts obtained from total cellular RNA of infected Jurkat cells. β-actin is used for normalization. The assay is able to detect and quantify transcripts containing the full TAR RNA but not very short aborted transcripts. All qPCR measurements are in triplicate and error bars represent S.D. '*' denotes statistically significant differences ($p<0.05$) between the indicated pairs of Tat variants.

We explored the possibility that loss of gene expression for I39Q Tat arises due to its inability to interact with an upstream transcription factor such as Sp1 or a chromatin remodeling complex such as SWI/SNF (21,38). However, co-immunoprecipitation showed no differences in Sp1 binding between the I39Q and DM Tat (Fig. 3.12*A*). A change in interaction with SWI/SNF could alter disruption of the nucleosome (Nuc-1) situated at the transcription start-site; however, nuclease sensitivity assays showed that both the I39Q and DM Tat had similar effects on Nuc-1 (Fig. 3.12*B*).

At the heart of viral gene expression is Tat's apparent ability to affect RNAPII phosphorylation. Sequential phosphorylation of serines at position 5 (Ser5) and 2 (Ser2) within an evolutionarily conserved but unstructured domain in mammalian RNAPII, consisting of 52 repeats of the heptapeptide $Y_1S_2P_3T_4S_5P_6S_7$ at its CTD, is critical for mRNA synthesis and processing (39). Normally, RNAPII recruited to the promoter of a gene is phosphorylated at Ser5 by Cdk7 within the transcription factor complex TFIIH (40). Shortly after transcription initiation, the polymerase briefly stalls ~30-40 bp downstream of the transcription start site to allow for pre-mRNA processing steps such as capping (41,42). Phosphorylation at Ser2 by the P-TEFb complex then promotes transcriptional elongation. In HIV-1 gene expression, however, Tat directly recruits and enables P-TEFb to phosphorylate both Ser5 and Ser2, and thereby greatly enhances transcriptional elongation (11,12,43).

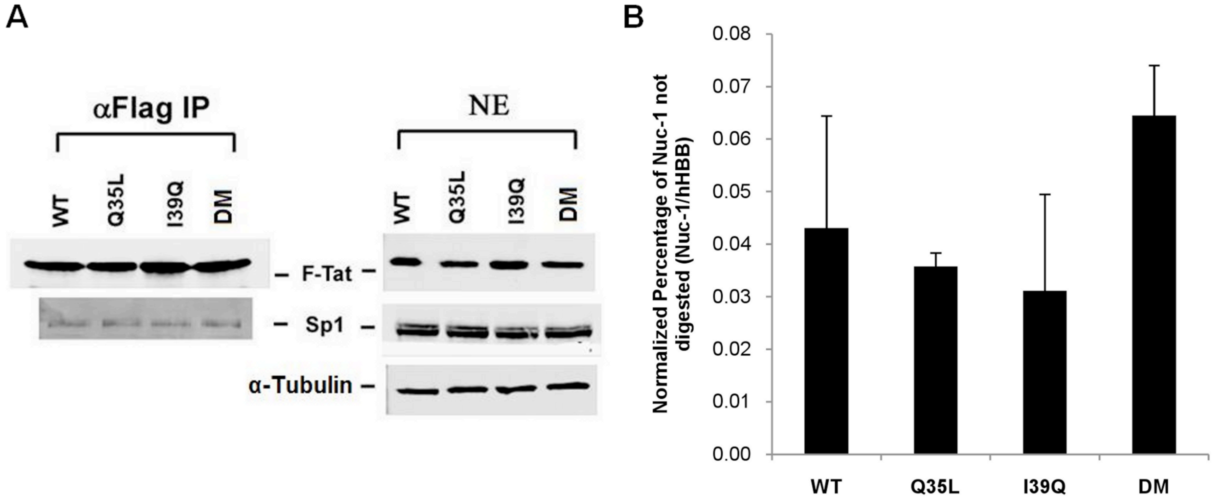


Figure 3.12. Mutations do not alter Tat binding with Sp1 or show differences in Nuc-1 disruption. (A) Immunoprecipitation of nuclear extracts with α-FLAG antibody, obtained from HeLa cells transfected with the UbChIT vector, were followed by western blots with α-Sp1 antibody. Both I39Q and DM Tat appear to bind Sp1 with similar affinity. (B) Jurkat cells infected with different Tat variants were incubated with or without a nuclease DNAse I and genomic DNA was extracted using the EpiQ Chromatin Analysis Kit. The human hemoglobin gene (hHBB) was used as an internal control. All qPCR measurements were made in triplicate and error bars represent S.D. None of the Tat variants showed any statistical difference in Nuc-1 disruption from each other (p>0.05).

Consistent with these results, the recently solved crystal structure of the P-TEFb-Tat complex shows that Tat binding induces P-TEFb conformational changes (36). To analyze the potential structural effects of mutations at sites 35 and 39, we performed additional *in silico* modeling based on this structure of Tat-P-TEFb, either in complex with or without an ATP analogue molecule (ATP+ or ATP-). Interestingly, both the ATP+ and ATP- structures of I39Q Tat are slightly energetically stabilized compared to WT Tat. In contrast, for the DM Tat the ATP+ structure was destabilized by 3.42 kcal/mol*, and the ATP- structure was stabilized by 2.38 kcal/mol* compared to WT Tat (Table 3.1). Based on the energetics of the ATP+ structures, these modeling results suggest that compared to I39Q Tat, P-TEFb associated with the DM Tat may have a higher propensity to transfer the phosphate group from ATP to a substrate and transit to the more stable ATP- state. Moreover, based on the collective evidence from literature, viral transcript data, and *in silico* modeling results, we hypothesized that the I39Q Tat mutant, unlike the DM, may fail to efficiently induce P-TEFb mediated phosphorylation of the CTD of RNAPII (Fig. 3.11*C* and Table 3.1) (36,43).

To explore this potential phosphorylation defect for I39Q Tat during transcriptional initiation and early elongation involved in efficient escape of RNAPII from the promoter, we performed chromatin immunoprecipitation (ChIP) with qPCR analysis to quantify the levels of total and Ser5 and Ser2 phosphorylated RNAPII associated with the HIV promoter in the presence of different Tat variants. Interestingly, even though similar levels of total RNAPII are recruited to the viral promoter (Fig. 3.13*C*), the level of Ser5P-CTD of RNAPII close to the transcription start site for the I39Q Tat mutant was dramatically lower than for the DM Tat, and slightly lower than for WT Tat (Fig. 3.13*A*). Similarly, the level of Ser2P-CTD of RNAPII for I39Q Tat during early elongation was significantly (p<0.05) lower than both WT and DM Tat (Fig. 3.13*B*).

Table 1. ***In silico*** **modeling results of the stability (*ΔΔG*) of the Tat-P-TEFb or Tat-P-TEFb-ATP complex after introducing mutations in Tat.**

ΔΔG **(kcal/mol*)**	**Tat-P-TEFb Complex (ATP -) (PDB: 3MI9)**	**Tat-P-TEFb-ATP Complex (ATP+) (PDB: 3MIA)**
Q35L	-3.00	4.95
I39Q	-1.01	-1.99
DM	-2.38	3.42

Negative values indicate greater stability of a complex as compared to the complex containing WT Tat. The asterisk over kcal/mol indicates that these values are computational determined. In contrast to the I39Q Tat, the Tat-P-TEFb-ATP complex for the DM Tat is destabilized and hence may have greater propensity to transfer the phosphate group to the CTD of RNAPII and transition into the more stable Tat-P-TEFb complex.

Thus, it appears that WT Tat's combination of weak Ser5P-CTD of RNAPII, strong P-TEFb binding affinity, and high Ser2P-CTD of RNAPII – or DM Tat's combination of high Ser5P-CTD of RNAPII, moderate P-TEFb binding affinity, and high Ser2P-CTD of RNAPII – mediates efficient escape of RNAPII from the HIV-1 promoter and activates gene expression for these two variants (Figs. 3.9*B* and 3.13*A*,*B*). Thus, it is possible that WT and DM Tat achieve similar levels of gene expression through orthogonal combinations of P-TEFb binding and Ser5P-CTD of RNAPII (Fig. 3.13*D*).

In contrast, although the I39Q Tat displays moderate P-TEFb binding affinity, the extremely low levels of Ser5P-CTD and Ser2P-CTD of RNAPII likely impairs its ability to activate gene expression (Figs. 3.9*B* and 3.13*A*,*B*). Therefore, it appears that a Gln at site 35 (as seen for the WT and I39Q Tat) reduces Tat's ability to induce P-TEFb-mediated phosphorylation at Ser5-CTD of RNAPII, but the presence of a Leu at site 35 (as in the Q35L and DM Tat) dramatically increases this function (Fig. 3.13*A*). However, Q35L Tat fails to bind P-TEFb and promote efficient escape of RNAPII from the promoter, as seen from the significantly lower levels (p<0.05) of Ser2P-CTD of RNAPII for this mutant compared to WT and DM Tat. Thus, although the I39Q and DM Tat both have similar, but lower, P-TEFb binding affinity than WT Tat, the high Ser5P-CTD and Ser2P-CTD of RNAPII observed with the DM but not I39Q Tat apparently rescues gene expression.

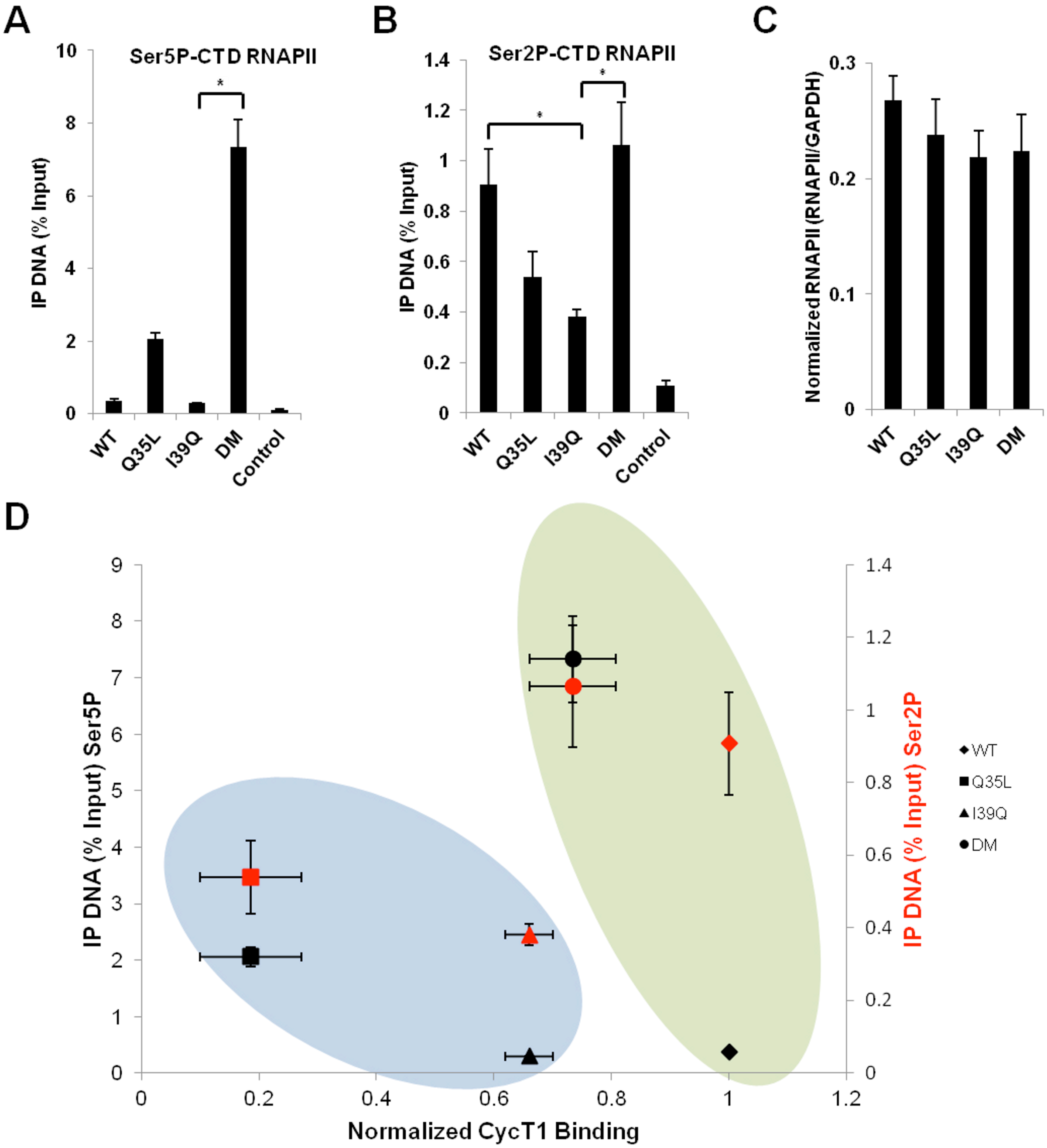


Figure 3.13. I39Q Tat fails to efficiently induce phosphorylation of the CTD of RNAPII, and subtype Tat's have potentially evolved alternate modes of inducing viral gene expression. (A), (B), and (C) ChIP for Ser5P-CTD of RNAPII, Ser2P-CTD of RNAPII and total RNAPII close to the transcription start site. Although similar levels of RNAPII are recruited to the viral promoter for all Tat variants, the I39Q Tat apparently fails to induce P-TEFb to efficiently phosphorylate the CTD of RNAPII, unlike the DM and WT Tat. Controls were performed without antibody. All qPCR measurements are in triplicate, and error bars represent S.D. '*' denotes statistically significant differences ($p<0.05$) between the indicated pairs of Tat variants. (D) Plot of Ser5P-CTD of RNAPII vs.

CycT1 binding, and Ser2P-CTD of RNAPII vs. CycT1 binding, for different Tat variants. The black and red symbols correspond to the levels of Ser5P-CTD of RNAPII and Ser2P-CTD of RNAPII for different Tat variants, respectively. The blue oval encompasses the Q35L and I39Q Tat variants that fail to activate gene expression, either due to its inability to bind P-TEFb or due to its failure to induce P-TEFb to efficiently phosphorylate the CTD of RNAPII. The green oval shows that the WT and DM Tat activate gene expression, though potentially through different mechanisms. Markers within the green oval show that subtype B Tat (WT Tat) displays high P-TEFb binding affinity and low Ser5P-CTD of RNAPII whereas the DM Tat, mimicking most subtype C Tats at sites 35 and 39, shows moderate P-TEFb binding and high Ser5P-CTD of RNAPII, with both Tat variants displaying comparable levels of Ser2P-CTD of RNAPII.

3.5 Discussion

For such a small protein, Tat shows a surprising diversity of function mediated by interaction with numerous cellular partners. It is post-translationally modified at specific sites by several cellular factors that impact transactivation. In addition to acetylation by PCAF and p300, there is evidence for methylation of Tat at Arg52 and Arg53 by the arginine methyltransferase PRMT6 (15) and methylation at Lys51 by the lysine methyltransferase Set 7/9 (KMT7). Similarly, other lysine methyltransferases have been shown to interact with Tat (16,17). Tat is also phosphorylated by Cdk2 (Ser16, Ser46) and PKR (Ser62, Thr64, Ser68) (18,19). Further evidence of the versatility of Tat can be seen in its interaction with other cellular proteins such as SKIP/SNW1 and SWI/SNF (20,21,23). Conserved and functionally important individual sites involved in these interactions can often be identified from multiple sequence alignments. However, the identification of mutually-dependent coevolving sites, which can readily be missed by simple site conservation, is enabled through the use of statistical measures such as MI.

To date, the experimental discovery of correlated sites within the HIV-1 proteome – such as in Tat, Reverse transcriptase, Nucleocapsid, and Rev – has involved creation of libraries of viral proteins or long-term culture of HIV-1 strains with single, site-directed mutations to reveal potential "suppressor mutations" (44-47). These approaches can sometimes yield either reversion of the introduced mutation or suppressor sites that do not naturally or specifically coevolve but act in a global, independent manner to increase fitness. By comparison, statistical analysis of viral sequence databases can identify positions whose evolution is correlated in a natural or clinical setting, as well as reveal correlations between new, unanticipated amino acid pairs. Here, we have harnessed MI to demonstrate functionally important correlations between pairs of sites in Tat.

In computationally guided experiments using a model lentiviral system mimicking the positive-feedback loop in HIV-1, we found that single point mutations Q35L and I39Q yielded Tat variants that failed to activate gene expression from the viral LTR, with a majority of the proviruses existing in a silenced state that was activated only upon stimulation with pharmacological agents. However, introduction of both mutations Q35L and I39Q into the same Tat protein restored gene expression (Fig. 3.4*C*,*D*). Thus, the Gln35-Ile39 and Leu35-Gln39 residue pairs both result in efficient gene expression from the viral promoter, for two Tat subtypes (Figs. 3.4*C* and 3.6*C*), confirming that sites 35 and 39 are coevolving. Furthermore, co-immunoprecipitation and ChIP studies revealed distinct, complementary mechanisms that constrain amino acid residues at these two sites: effective P-TEFb binding and alteration of P-TEFb substrate specificity to include the phosphorylation of Ser5 and Ser2 residues on the

RNAPII CTD. Specifically, we show that the Q35L single mutant fails to bind P-TEFb, whereas the DM partially rescues P-TEFb binding (Fig. 3.9). In contrast, the I39Q Tat binds P-TEFb at levels close to the DM Tat, yet still suffers from very low gene expression (Figs. 3.4*C* and 3.9*B*) potentially due to its inability to induce P-TEFb to phosphorylate the CTD of RNAPII (Fig. 3.13*A*,*B*). It is plausible that the inability of the I39Q Tat to induce efficient phosphorylation of the CTD of RNAPII involves loss of interaction with additional host factors that remain to be discovered. At any rate, unlike most coevolving or suppressor mutations that help restore a single biological function (48), the coevolving sites 35 and 39 each contribute to distinct Tat-mediated mechanisms that are integrated to yield an active protein. That is, mutationally-induced deficits in one mechanism can be compensated for by mutations in the coevolving site that affect the other.

Besides the strong coevolution signal observed between sites 35 and 39, site 31 was also correlated with sites 35 and 39, as seen in the MI analysis (Figs. 3.2 and 3.7), with decreased gene expression of a mutant at site 31 (Fig. 3.4*C*), and an accompanying statistical increase in the number of silenced cells observed experimentally (Fig. 3.4*D*). Extending our analysis beyond pairwise interactions, site 31, which is part of the 3_{10} helix, and sites 35 and 39, which are within the next α-helix, appear to be constrained to certain triplets of amino acids (Fig. 3.7). A triplet consisting of Gln at site 39, Ser at site 31 and Leu at site 35 can be observed in the Tat sequence database. Similarly, a Leu, Ile or Thr at site 39 is primarily correlated with a Cys at site 31 and a Gln at site 35. Interestingly, the residue with the smaller of the side-chain volumes at site 31 (Ser), is correlated with the larger of the side-chain volumes at site 35 (Leu), whereas the larger side-chain at site 31 (Cys) is correlated with the smaller side-chain (Gln) at site 35. These steric and volume effects possibly constrain site 31 to specific amino acid residues depending on the residues at sites 35 and 39 and help position residues at 35 and 39 within the hydrophobic groove in CycT1. Thus sites 31, 35, and 39 appear to form a mini-network of coevolving residues.

Most subtype B Tats pair Gln35 with Ile39/Leu39/Thr39, whereas a majority of subtype C Tats contain the Leu35-Gln39 residue pair, similar to the DM Tat, with a few having a Gln35-Leu39 residue pair. Based on the P-TEFb binding assay and levels of Ser5P-CTD of RNAPII, it appears that different subtypes could potentially have evolved alternate modes or "solutions" to inducing gene expression from the viral LTR. Subtype B Tats induce a low level of Ser5P-CTD in RNAPII (Fig. 3.13*A*), but the strong binding to P-TEFb could at least in part compensate for this deficit (Fig. 3.9*B*) and eventually produce efficient elongation, as can been seen from the high levels of Ser2P-CTD of RNAPII, a marker for P-TEFb-induced elongation (Fig. 3.14). In contrast, the DM Tat, which mimics the majority of subtype C Tats at sites 35 and 39, induces very high levels of Ser5P (Fig. 3.13*A*) coupled with relatively weaker P-TEFb binding (Fig. 3.9*B*) that in combination could ultimately drive comparable levels of Tat-mediated gene expression as measured by GFP expression (Fig. 3.4*C*), viral transcript analysis, and Ser2P-CTD of RNAPII (Fig. 3.14). Thus, the diversification of HIV-1 into different subtypes has apparently resulted in the evolution of compensatory mechanisms to trade off substrate binding and catalytic activity in inducing Tat-mediated gene expression from the viral LTR, such that the overall activity may be determined by the combination or "sum" of contributions from individual positions or functions (Fig. 3.13*D*). This novel finding has some parallels with other biological systems. For instance, it has been shown previously that autophosphorylation mutants of the epidermal growth factor (EGF) receptor stimulate similar levels of MAP kinase activation, gene expression, and mitogenesis as the WT EGF receptor though different compensatory mechanisms (49).

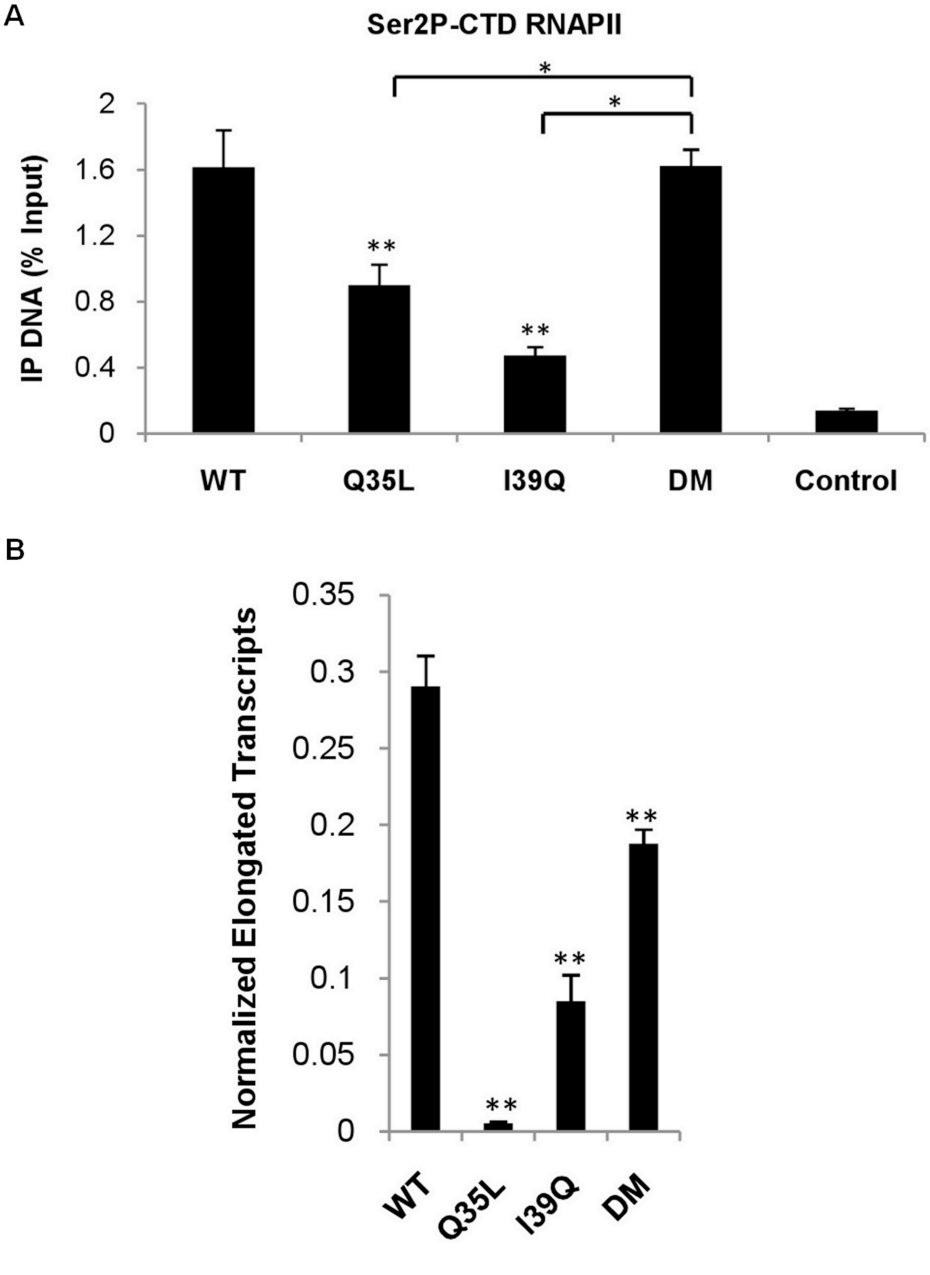


Figure 3.14 RNAPII elongation for different Tat variants using chromatin immunoprecipitation for Ser2P-CTD of RNAPII and viral transcript quantification. (A) ChIP for Ser2P-CTD of RNAPII +2215 bp downstream of the transcription start site for different Tat mutants. As expected, the WT and DM Tat have higher Ser2P-CTD signal than the two single-mutants, although the Q35L Tat mutant shows a higher signal than anticipated based on GFP expression (Fig. 2*C*). Controls were performed without antibody. All qPCR measurements are in triplicate and error bars represent S.D. '**' denotes statistically significant differences ($p<0.05$) from WT Tat. '*' denotes statistically significant differences ($p<0.05$) between the indicated pairs of Tat variants. (B) Elongated viral transcripts were quantified by RT-qPCR using β-actin for normalization. In agreement with the gene expression

results of Fig. 2*C*, the single mutants, Q35L and I39Q have much lower levels of elongated transcripts compared to the WT Tat. The DM Tat partially rescues genes expression to WT Tat levels and thus has higher levels of elongated transcripts as compared to the single mutant Tats. '**' denotes statistically significant differences ($p<0.05$) from WT Tat.

We have previously found that gene expression from the LTR is a stochastic process, with bursts of mRNA production separated by long intervals, a feature that could play an important role in the establishment of viral latency (34,35,50). Changes to Tat that distinctly affect transcriptional initiation or elongation could differentially impact the frequency and size of mRNA bursts. Gene expression data at low Tat levels indicates that Tat variants from different subtypes, with potentially alternate mechanisms for inducing gene expression, could impact probabilistic gene expression events (Fig. 3.8). Future work may explore whether these differences in Tat result in different propensities for viral latency.

In addition to its application to HIV-1, such an integrated computational and experimental approach could readily be extended to other pathogens to gain deeper insights into their function and evolution, as well as potentially aid in the rational development of novel therapeutic strategies.

3.6 Materials and Methods

3.6.1 Plasmids

The LGIT vector has been previously described (34). To construct the UbChIT vector, the internal CMV promoter from pCS-CG was replaced by the human ubiquitin promoter (from pFUGW) using Sac II/Eco RI sites (51,52). GFP was swapped for mCherry using Xba I/Eco RI sites. IRES-Tat was then inserted into this vector from LGIT using Eco RI/Xho I sites. Mutations in Tat in the LGIT and UbChIT vector were introduced using QuikChange PCR (Stratagene). Primers for the site-directed mutations will be made available upon request.

3.6.2 Cell Culture

Jurkat cells, used for infections, mRNA extraction and ChIP assays, were cultured in RPMI 1640 (Mediatech). HEK 293T cells, used for viral packaging, were cultured in Isocove's DMEM (Mediatech). HeLa cells, used for co-immunoprecipitation experiments, and the HL3T1 cell-line, used for the Luciferase assay, were cultured in DMEM. All cell media were supplemented with 10% fetal bovine serum (FBS) and 100U/mL Penicillin-Streptomycin (P-S). All cells were grown at 37^{o}C and 5% CO_2.

3.6.3 Viral Harvesting, Titering and Infections

To package the lentiviral vectors, 100 mm plates with HEK 293T cells were cotransfected with 10 μg of the plasmid of interest and the following helper plasmids: 5 μg pMDLg/pRRE, 3.5 μg pVSV-G and 1.5 μg pRSV-Rev (53). 36 hours post-transfection, virus was harvested by ultracentrifugation, and the viral pellets were resuspended in PBS and stored at -80^{o}C for future use. Viral titers were obtained by infecting $3x10^5$ cells with different viral volumes and measuring GFP expression of cells 8 days post-infection. On day 8 post-infection, cells were stimulated with TNF-α (20 ng/mL) and TSA (400 nM) for 18 hours prior to analysis of GFP expression by flow cytometry. Based on the resulting titering curves, Jurkat cells were infected at a MOI of 0.05-0.1 for experiments to ensure single integration events per cell.

3.6.4 Flow Cytometry and Cell Sorting

GFP fluorescence was monitored using the FC500 Flow Cytometer (Beckman Coulter) using the 488 nm laser and the 530 nm filter. Jurkat cells were stimulated with drugs 18 hours before sorting, and GFP^+ cells were sorted using a Cytopeia INFLUX Sorter or DAKO-Cytomation MoFlo High Speed Sorter to isolate Jurkat cells infected with the LGIT vector.

3.6.5 Transfections

For the co-immunoprecipitation experiments, 2 μg of the UbChIT vector was transfected into HeLa cells cultured in 150 mm plates using a PEI-based transfection method. For the Luciferase assay, the HeLa based cell-line HL3T1 was cultured in 6-well plates and transfected with 10 ng of UbChIT vector using the Lipofectamine Transfection Reagent and PLUS Reagent (Invitrogen).

3.6.6 mRNA Extraction and RT-qPCR

Total cellular RNA from $2x10^6$ sorted Jurkat cells infected with the LGIT lentivirus was extracted using Trizol (Invitrogen). Viral transcripts were quantified using the single step Quantitect SYBR Green RT-PCR kit (Qiagen) and a Bio-Rad iCycler (iQ5). Total viral transcripts were quantified using the primers LTR5 (5'-GTTAGACCAGATCTGAGCCT-3') and LTR3 (5'- GTGGGTTCCCTAGT TAGCCA-3'). Elongated viral transcripts were quantified using the primers GFP5 (5'- AGCAAAGACCCCAACGAGAA-3') and GFP3 (5'-CGTCCATGCCGAGAGTGAT-3'). β-Actin was used to normalize the samples using the primers β-Actin5 (5'- ACCTGACTGACTACCTCATGAAGATCCTCACCGA-3') and β-Actin3 (5'- GGAGCTGGAAGCAGCCGTGGCCATCTCTTGCTCGAA-3'). All RT-qPCR was performed in triplicate.

3.6.7 Co-Immunoprecipitation and Western Blots

HeLa cells transfected with UbChIT were lysed, and nuclear extracts (NE) prepared 48 hours post-transfection. Anti-FLAG M2 agarose beads (Sigma) were washed using the wash buffer (20 mM HEPES-KOH [pH 7.9], 15% glycerol, 0.2 mM EDTA, 0.2% NP-40, 1 mM dithiothreitol, and 1 mM phenylmethylsulfonyl fluoride) containing 0.3 M KCl. The beads were then incubated with the NE for 2 hours with rotation at 4°C. To minimize non-specific binding, the beads were then washed thrice with the wash buffer containing 0.3 M KCl and twice with the wash buffer containing 0.1 M KCl. The protein complexes bound to the beads were then eluted using 0.5 μg/mL of FLAG peptide in the wash buffer containing 0.1 M KCl. The protein complexes were then separated on a SDS-PAGE gel and blotted with the following antibodies: anti-Cdk9, anti-ENL (Abcam, research sample), anti-CycT1 (Santa Cruz Biotechnology, Calalog # sc-10750), anti-Sp1 (Millipore, Catalog # 07-645), anti-ELL2 (Bethyl Laboratories, Catalog # A302-505A), anti-AFF4 (Santa Cruz Biotechnology, Catalog # sc-101062) and anti-AF9 (Bethyl Laboritories, Catalog # A300-595A).

3.6.8 Chromatin Immunoprecipitation

Upstate EZ ChIP Kit reagents (Upstate) and protocol were used for the assay with variations. 1x10^7 sorted Jurkat cells were fixed at room temperature in 1% formaldehyde for 10 min followed by quenching of unreacted formaldehyde with 125 mM glycine for 5 minutes. After extensive washing with PBS, the cells were lysed with a SDS lysis buffer containing protease inhibitor complex. For the Ser5P and Ser2P ChIPs, a phosphatase inhibitor complex was also added during cell lysis. The cells were then sonicated using the Branson Sonifier 450 for 25 cycles at a power output of 2.5 and 25% duty cycle, with each cycle consisting of 15 pulses followed by incubation on ice for 1 min. Sheared DNA fragments from 0.2-1 kb were verified using DNA gel electrophoresis. For the Ser5P and Ser2P ChIPs, the following variations were introduced to the EZ ChIP Kit. Instead of using Protein G beads, Anti-mouse IgM agarose beads (Sigma) were used. These beads were washed with RIPA buffer and then blocked with salmon sperm DNA and yeast tRNA. In addition, these beads were incubated with the antibody-DNA/protein complex for 5 hours at 4°C with rotation.

Precipitated DNA was quantified using qPCR (Bio-Rad iCycler, iQ5) using the EpiQ Chromatin SYBR Supermix (Bio-Rad). All samples were run in triplicate, and melt curves were used to analyze the specificity of the PCR product. The following primers were used for the RNAPII, Ser5P-CTD RNAPII and Ser2P-CTD RNAPII ChIP close to the transcription start site: LTR5 (5'-GTTAGACCAGATCTGAGCCT-3') and LTR3 (5'- GTGGGTTCCCTAGT TAGCCA-3'). The RNAPII ChIP was normalized using GAPDH for which the following primers were used: GAPDH5 (5'-ACCTCCCATCGGGCCAATCTCAGTC-3') and GAPDH3 (5'-GGCTGACTGTCGAACAGGAGGAGCA-3'). The following primers were used for the Ser2P-CTD RNAPII ChIP +2215 bp downstream of the transcription start site: GFP5 (5'-AGCAAAGACCCCAACGAGAA-3') and GFP3 (5'-CGTCCATGCCGAGAGTGAT-3'). The following antibodies were used for ChIP: anti-RNAPII (Millipore, Catalog # 05-623), anti-Ser5P CTD RNAPII (Covance, Catalog # MMS-134R) and anti-Ser2P CTD RNAPII (Covance, Catalog # MMS-129R).

3.6.9 Nuclease Sensitivity Assay

The EpiQ Chromatin Analysis Kit (Bio-Rad) protocol and reagents were used for this assay. 2.5x10^5 sorted Jurkat cells were incubated with the chromatin buffer with or without 2 μL of nuclease DNAse I for 1 hour at 37°C. Cells were then incubated with 25 μL stop buffer to quench the reaction. After several wash steps, the digested and undigested genomic DNA were quantified using qPCR (Bio-Rad iCycler, iQ5) using the EpiQ Chromatin SYBR Supermix (Bio-Rad). The following primers were used to amplify the Nuc-1 region: Nuc5 (5'-GGACTTTCCGCTGGGGACTTTCCAGGG-3') and Nuc3 (5'-CTCGACGCAGGACTCGGCTTGCTGAAGCGCGC-3'). The hemoglobin gene was used as an internal control for the samples with the following primers: hHBB5 (5'-AAGCCAGTGCCAGAAGAGCCAAGGA-3') and hHBB3 (5'-CCCACAGGGCAGTAACGGCAGACTT-3'). All qPCR samples were run in triplicate and melt curves were used to ensure product specificity.

3.6.10 Luciferase Assay

The HL3T1 cell-line transfected with UbChIT were harvested 48 hours post-transfection using the Luciferase Assay System (Promega).

3.6.11 Mutual Information Analysis

Matlab codes for calculation of raw and background MI scores to estimate the corrected MI scores will be made available upon request.

3.6.12 Structure Modeling

Single and double mutations were modeled starting from the ATP-bound (PDB: 3MIA) and unbound (PDB: 3MI9) structures using RosettaDesign (54). Briefly, RosettaDesign uses a full-atom scoring function including Lennard-Jones, hydrogen-bonding, solvation, and torsional terms and models side-chain dihedral degrees of freedom by sampling from a backbone-dependent rotamer library (55). Here, the backbone was kept fixed, and side chains within 6Å of residues 35 and 39 were allowed to repack with extra subrotamers for the chi1 and chi2 dihedral angles. In addition, the slope of the Lennard-Jones repulsive term was reduced to be more forgiving of minor backbone differences between the two starting structures (56). Ligands and waters were not modeled and but do not occur within 6Å of residues 35 and 39 in the WT ATP-bound and unbound structures. Relative energies based on the RosettaDesign scoring function were computed as follows: $\Delta\Delta G$(mutation) = ΔG(mutant) - ΔG(WT). The structure graphic was generated with PyMol (The PyMOL Molecular Graphics System, Version 1.3, DeLano Scientific LLC.).

2.6.13 Statistical Analyses

All statistical significances were computed using one-way ANOVA followed by the Tukey-Kramer multiple comparison method to compare different pairs.

3.7 References

1. Negroni, M., and Buc, H. (2001) *Annu Rev Genet* **35**, 275-302
2. Johnson, V. A., Brun-Vezinet, F., Clotet, B., Gunthard, H. F., Kuritzkes, D. R., Pillay, D., Schapiro, J. M., and Richman, D. D. (2009) *Top HIV Med* **17**, 138-145
3. Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004) *Nat Rev Genet* **5**, 52-61
4. van Opijnen, T., Jeeninga, R. E., Boerlijst, M. C., Pollakis, G. P., Zetterberg, V., Salminen, M., and Berkhout, B. (2004) *J Virol* **78**, 3675-3683
5. Desfosses, Y., Solis, M., Sun, Q., Grandvaux, N., Van Lint, C., Burny, A., Gatignol, A., Wainberg, M. A., Lin, R., and Hiscott, J. (2005) *J Virol* **79**, 9180-9191
6. Kurosu, T., Mukai, T., Komoto, S., Ibrahim, M. S., Li, Y. G., Kobayashi, T., Tsuji, S., and Ikuta, K. (2002) *Microbiol Immunol* **46**, 787-799
7. Camps, M., Herman, A., Loh, E., and Loeb, L. A. (2007) *Crit Rev Biochem Mol Biol* **42**, 313-326
8. Kao, S. Y., Calman, A. F., Luciw, P. A., and Peterlin, B. M. (1987) *Nature* **330**, 489-493

9. Wei, P., Garber, M. E., Fang, S. M., Fischer, W. H., and Jones, K. A. (1998) *Cell* **92**, 451-462
10. Roy, S., Delling, U., Chen, C. H., Rosen, C. A., and Sonenberg, N. (1990) *Genes Dev* **4**, 1365-1373
11. Zhou, Q., and Yik, J. H. (2006) *Microbiol Mol Biol Rev* **70**, 646-659
12. D'Orso, I., and Frankel, A. D. (2010) *Nat Struct Mol Biol* **17**, 815-821
13. Kiernan, R. E., Vanhulle, C., Schiltz, L., Adam, E., Xiao, H., Maudoux, F., Calomme, C., Burny, A., Nakatani, Y., Jeang, K. T., Benkirane, M., and Van Lint, C. (1999) *EMBO J* **18**, 6106-6118
14. Marzio, G., Tyagi, M., Gutierrez, M. I., and Giacca, M. (1998) *Proc Natl Acad Sci U S A* **95**, 13519-13524
15. Xie, B., Invernizzi, C. F., Richard, S., and Wainberg, M. A. (2007) *J Virol* **81**, 4226-4234
16. Pagans, S., Kauder, S. E., Kaehlcke, K., Sakane, N., Schroeder, S., Dormeyer, W., Trievel, R. C., Verdin, E., Schnolzer, M., and Ott, M. (2010) *Cell Host Microbe* **7**, 234-244
17. Van Duyne, R., Easley, R., Wu, W., Berro, R., Pedati, C., Klase, Z., Kehn-Hall, K., Flynn, E. K., Symer, D. E., and Kashanchi, F. (2008) *Retrovirology* **5**, 40
18. Ammosova, T., Berro, R., Jerebtsova, M., Jackson, A., Charles, S., Klase, Z., Southerland, W., Gordeuk, V. R., Kashanchi, F., and Nekhai, S. (2006) *Retrovirology* **3**, 78
19. Endo-Munoz, L., Warby, T., Harrich, D., and McMillan, N. A. (2005) *Virol J* **2**, 17
20. Bres, V., Yoshida, T., Pickle, L., and Jones, K. A. (2009) *Mol Cell* **36**, 75-87
21. Mahmoudi, T., Parra, M., Vries, R. G., Kauder, S. E., Verrijzer, C. P., Ott, M., and Verdin, E. (2006) *J Biol Chem* **281**, 19960-19968
22. Deng, L., de la Fuente, C., Fu, P., Wang, L., Donnelly, R., Wade, J. D., Lambert, P., Li, H., Lee, C. G., and Kashanchi, F. (2000) *Virology* **277**, 278-295
23. Agbottah, E., Deng, L., Dannenberg, L. O., Pumfery, A., and Kashanchi, F. (2006) *Retrovirology* **3**, 48
24. Martin, L. C., Gloor, G. B., Dunn, S. D., and Wahl, L. M. (2005) *Bioinformatics* **21**, 4116-4124
25. Kass, I., and Horovitz, A. (2002) *Proteins* **48**, 611-617
26. Gobel, U., Sander, C., Schneider, R., and Valencia, A. (1994) *Proteins* **18**, 309-317
27. Lockless, S. W., and Ranganathan, R. (1999) *Science* **286**, 295-299
28. Korber, B. T., Farber, R. M., Wolpert, D. H., and Lapedes, A. S. (1993) *Proc Natl Acad Sci U S A* **90**, 7176-7180
29. Fares, M. A., and Travers, S. A. (2006) *Genetics* **173**, 9-23
30. Gilbert, P. B., Novitsky, V., and Essex, M. (2005) *AIDS Res Hum Retroviruses* **21**, 1016-1030
31. Skerker, J. M., Perchuk, B. S., Siryaporn, A., Lubin, E. A., Ashenberg, O., Goulian, M., and Laub, M. T. (2008) *Cell* **133**, 1043-1054
32. Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2008) *Bioinformatics* **24**, 333-340
33. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009) *Proc Natl Acad Sci U S A* **106**, 67-72
34. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., and Schaffer, D. V. (2005) *Cell* **122**, 169-182

35. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P., and Schaffer, D. V. (2009) *PLoS Pathog* **5**, e1000260
36. Tahirov, T. H., Babayeva, N. D., Varzavand, K., Cooper, J. J., Sedore, S. C., and Price, D. H. (2010) *Nature* **465**, 747-751
37. He, N., Liu, M., Hsu, J., Xue, Y., Chou, S., Burlingame, A., Krogan, N. J., Alber, T., and Zhou, Q. (2010) *Mol Cell* **38**, 428-438
38. Chun, R. F., Semmes, O. J., Neuveut, C., and Jeang, K. T. (1998) *J Virol* **72**, 2615-2629
39. Phatnani, H. P., and Greenleaf, A. L. (2006) *Genes Dev* **20**, 2922-2936
40. Serizawa, H., Makela, T. P., Conaway, J. W., Conaway, R. C., Weinberg, R. A., and Young, R. A. (1995) *Nature* **374**, 280-282
41. Zhang, Z., Klatt, A., Gilmour, D. S., and Henderson, A. J. (2007) *J Biol Chem* **282**, 16981-16988
42. Schroeder, S. C., Schwer, B., Shuman, S., and Bentley, D. (2000) *Genes Dev* **14**, 2435-2440
43. Zhou, M., Halanski, M. A., Radonovich, M. F., Kashanchi, F., Peng, J., Price, D. H., and Brady, J. N. (2000) *Mol Cell Biol* **20**, 5077-5086
44. Tachedjian, G., Aronson, H. E., and Goff, S. P. (2000) *Proc Natl Acad Sci U S A* **97**, 6334-6339
45. Verhoef, K., and Berkhout, B. (1999) *J Virol* **73**, 2781-2789
46. Cimarelli, A., Sandin, S., Hoglund, S., and Luban, J. (2000) *J Virol* **74**, 4273-4283
47. Jain, C., and Belasco, J. G. (1996) *Cell* **87**, 115-125
48. del Alamo, M., and Mateu, M. G. (2005) *J Mol Biol* **345**, 893-906
49. Li, N., Schlessinger, J., and Margolis, B. (1994) *Oncogene* **9**, 3457-3465
50. Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V., and Arkin, A. P. (2010) *PLoS Comput Biol* **6**
51. Miyoshi, H., Blomer, U., Takahashi, M., Gage, F. H., and Verma, I. M. (1998) *J Virol* **72**, 8150-8157
52. Greenberg, K. P., Geller, S. F., Schaffer, D. V., and Flannery, J. G. (2007) *Invest Ophthalmol Vis Sci* **48**, 1844-1852
53. Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., and Naldini, L. (1998) *J Virol* **72**, 8463-8471
54. Kuhlman, B., and Baker, D. (2000) *Proc Natl Acad Sci U S A* **97**, 10383-10388
55. Dunbrack, R. L., Jr., and Karplus, M. (1993) *J Mol Biol* **230**, 543-574
56. Friedland, G. D., Linares, A. J., Smith, C. A., and Kortemme, T. (2008) *J Mol Biol* **380**, 757-774

Chapter 4: TAR and Tat Independently Regulate the Strength of Gene Expression from the HIV-1 Promoter

4.1 Introduction

The high error and recombination rate of HIV-1 during reverse transcription has resulted in the rapid diversification of the most widespread group M into 9 subtypes and several recombinant forms (1-3). In addition to worldwide variations in host cultural norms and genetics, viral genetic diversity in HIV-1 has been shown to be an important factor in regulating gene expression, pathogenesis, and transmission rates (4-6). For instance, ex vivo models and patient data show that various viral subtypes can produce different transmission rates (5,7). At the molecular level, differences in the viral replication rate have been shown to arise from different reverse transcriptase and Tat sequences, as well as different architectures of transcription factor binding sites within the promoter. (6,8-11).

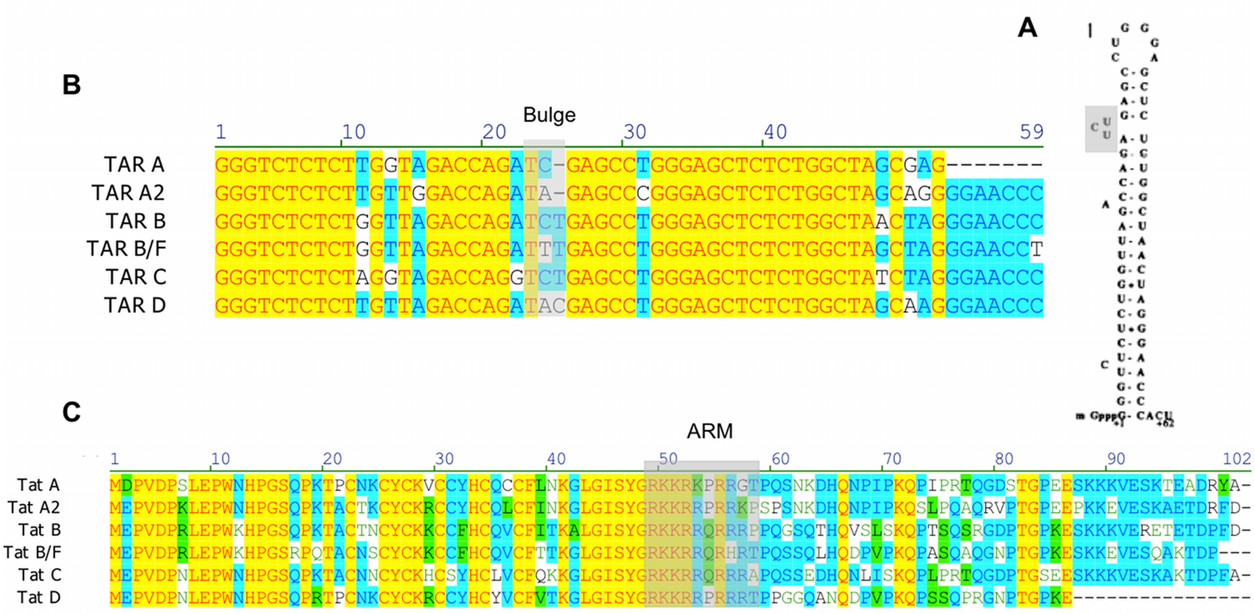

A
B
Bulge
1
10
20
30
40
59
TAR A GGGTCTCTCTTGGTAGACCAGATC-GAGCCTGGGAGCTCTCTGGCTAGCGAG-------
TAR A2 GGGTCTCTCTTGTTGGACCAGATA-GAGCCCGGGAGCTCTCTGGCTAGCAGGGAACCC
TAR B GGGTCTCTCTGGTTAGACCAGATCTGAGCCTGGGAGCTCTCTGGCTAACTAGGGAACCC
TAR B/F GGGTCTCTCTGGTTAGACCAGATTTGAGCCTGGGAGCTCTCTGGCTAGCTAGGGAACCT
TAR C GGGTCTCTCTAGGTAGACCAGGTCTGAGCCTGGGAGCTCTCTGGCTATCTAGGGAACCC
TAR D GGGTCTCTCTTGTTAGACCAGATACGAGCCTGGGAGCTCTCTGGCTAGCAAGGGAACCC
C
ARM
1
10
20
30
40
50
60
70
80
90
102
Tat A MDPVDPSLEPWNHPGSQPKTPCNKCYCKVCCYHCQCCFLNKGLGISYGRKKRKPRRGTPQSNKDHQNPIPKQPIPRTQGDSTGPEESKKKVESKTEADRYA-
Tat A2 MEPVDPKLEPWNHPGSQPKTACTKCYCKRCCYHCQLCFINKGLGISYGRKKRRPRRKPSPSNKDHQNPIPKQSLPQAQRVPTGPEEPKKEVESKAETDRFD-
Tat B MEPVDPRLEPWKHPGSQPKTACTNCYCKKCCFHCQVCFITKALGISYGRKKRRQRRRPPQGSQTHQVSLSKQPTSQSRGDPTGPKESKKKVERETETDPFD-
Tat B/F MEPVDPRLEPWKHPGSRPQTACNSCYCKKCCFHCQVCFTTKGLGISYGRKKRRQRHRTPQSSQLHQDPVPKQPASQAQGNPTGPKESKKEVESQAKTDP---
Tat C MEPVDPNLEPWNHPGSQPKTACNNCYCKHCSYHCLVCFQKKGLGISYGRKKRRQRRRAPQSSEDHQNLISKQPLPRTQGDPTGSEESKKKVESKAKTDPFA-
Tat D MEPVDPNLEPWNHPGSQPRTPCNKCYCKRCCYHCYVCFVTKGLGISYGRKKRRPRRRTPPGGQANQDPVPKQPSSQPRGNPTGPKE----------------
m GpppG- CACU

Figure 4.1. Representative sequence alignment for TAR and Tats from different HIV-1 subtypes. (A) Secondary structure of subtype B TAR RNA. TAR forms a hairpin structure with a 2-3 nucleotide bulge (shown in gray) that has been shown to interact with Tat. (B) Sequence alignment of TARs from different subtypes. Alignment shows that the TAR sequence is relatively highly conserved, possibly due to the secondary structure requirements of the TAR hairpin. The bulge region however shows considerable sequence diversity between subtypes. (C) Sequence alignment of Tat from different subtypes. Tat shows considerable sequence diversity between subtypes. The Arginine rich motif (ARM) interacts with the bulge of TAR and shows sequence diversity between Tat subtypes.

Furthermore, baseline polymorphisms and silent mutations within subtypes have been shown to alter susceptibility to anti-retroviral therapies and give rise to different resistance mutations (4,12,13). These studies indicate that different subtypes may produce significantly varying virulence and pathogenicity in patients. Furthermore, while the role of viral diversity in regulating HIV latency – the single greatest barrier to its elimination from a patient – and the rate of reactivation from latently infected cells is not entirely clear, recent evidence shows that genetic diversity in the promoters of different subtypes produce varied response to drugs that can be used to purge latent viral reservoirs (14).

To study how sequence diversity in the RNA hairpin TAR and the viral protein Tat regulate gene expression from the HIV-1 promoter, full length HIV-1 sequences from different subtypes were obtained from the NIH AIDS reagent program (15-18) and the TAR and Tat sequences were subsequently cloned into appropriate vectors. An alignment of TARs showed that stretches of its sequence was conserved across subtypes (Fig 4.1*B*). This is possibly due to the hairpin structure of TAR that constrains its bases (Fig 4.1*A*) (19). However, some sequence variation was observed at sites 11, 13, 48 and 50 within the lower stem of TAR. Interestingly, the sequence alignment also revealed diversity within the bulge of TAR, which consists of 2-3 unpaired bases that has been shown to interact with Tat and is critical for activating gene expression (19-21). In contrast, as discussed in Chapter 3, Tat shows considerable sequence diversity across subtypes, which was used to identify coevolving residues within the activation domain of Tat (1-48 amino acids). However, in this study, since we were particularly interested in the interaction of TAR and Tat, we focused our attention on the Arginine-rich motif (ARM) of Tat (Fig. 4.1*C*) (19,22-26). We found that Tat sequences show considerable diversity within this region, implying that subtype Tats may have differences in their interaction with TAR and their ability to activate gene expression from the viral promoter.

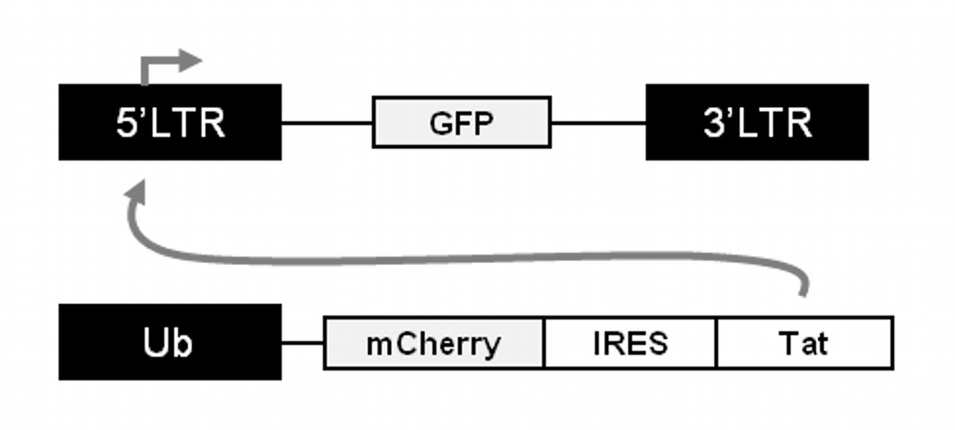


Figure 4.2. Open-loop construct used to study the impact of sequence diversity in TAR and Tat on viral gene expression. Schematic of the lentiviral LG vector used to create a stable polyclonal Jurkat cell line. Lower half shows the schematic for the UbChIT vector that is used to infect the LG Jurkat cell line. Different subtype Tats cloned into the UbChIT vector were used to study how they activate gene expression from the HIV-1 LTR.

4.2 Exploring the impact of sequence diversity in TAR and Tat on viral gene expression using a open-loop system

As a first step towards understanding how diversity within the sequences of TAR and Tat, in particular how the bulge of TAR and the ARM of Tat might alter the strength of gene expression (15-18), we created an open-loop circuit. The open-loop circuit allowed us to decouple the intrinsic properties of different subtype TAR and Tats from the role that the Tat-positive feedback loop plays in setting the level of gene expression. The open-loop circuit was created by placing green fluorescent protein (GFP) under the control of the HIV-1 LTR from subtype B (called as the LTR-GFP vector or LG) (Fig. 4.2) (27). Another vector was created in which Tat and mCherry were placed under the control of an Ubiquitin-C promoter, where the fluorescent protein mCherry serves as a quantitative indicator of the Tat expression level. Tat and mCherry were separated by the internal ribosome entry site (IRES), which allows for the translation of both mCherry and Tat from the same mRNA transcript. This vector was called Ubiquitin-mCherry-IRES-Tat or Ub-ChIT (Fig. 4.2).

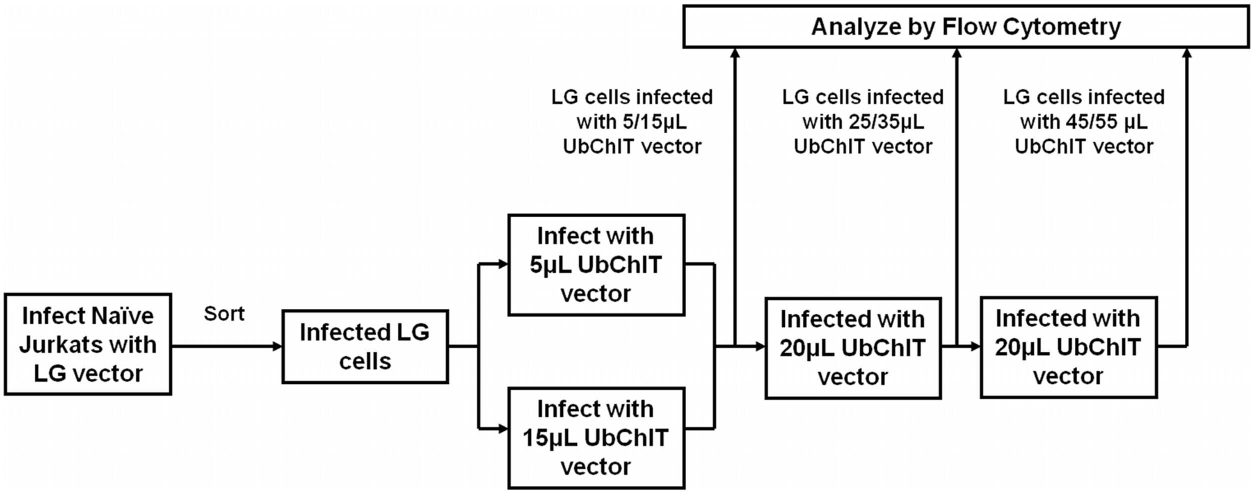

Analyze by Flow Cytometry
LG cells infected with 5/15µL UbChIT vector
LG cells infected with 25/35µL UbChIT vector
LG cells infected with 45/55 µL UbChIT vector
Infect Naïve Jurkats with LG vector
Sort
Infected LG cells
Infect with 5µL UbChIT vector
Infect with 15µL UbChIT vector
Infected with 20µL UbChIT vector
Infected with 20µL UbChIT vector

Figure 4.3. Experimental scheme for studying gene activation from TAR and Tats of different subtypes. Scheme shows Naïve Jurkats infected with the LG vector at low MOI (0.05 – 0.1) for single integration events per cell. The cells were stimulated with TNF-α 7 days post-infection and GFP+ cells were sorted. The cells were then allowed to relax and infected with 5µL and 15µL of UbChIT containing different subtype Tats. These two populations were further infected with increasing amounts (20µL in two further infections) of UbChIT and all the infections for UbChIT containing one subtype Tat were pooled and analyzed by flow cytometry. The LG Jurkat cell lines were infected with a wide range of UbChIT to allow quantification of gene expression for different Tat concentrations.

Cells of infected patients often carry multiple proviruses (an average of ~3-4 proviruses) per cells (28), and the high recombination rate of HIV-1 together imply that the large viral quasispecies within a patient could produce combinatorial Tat-TAR combinations *in vivo* (29). We wanted to explore how different combinations of Tat and TAR could produce variations in the level of gene expression resulting in these combinations having alternate propensities for latency. Since it has been reported that the ARM motif binds to the bulge of TAR (19,20), we wanted to study the interaction between different subtype TAR and Tats, that contain sequence differences within these important domains. To test the compatibility of different Tat-TAR pairs, we picked six different subtype pairs.

We infected Jurkat cells (a human T cell line) with the LG vector containing TAR sequences from different subtypes inserted into the LTR. LGs containing different subtype TARs were infected at low MOIs (0.05-0.1) to ensure single integrations of the vector in each cell. After sorting the infected cells using fluorescence activated cell sorting (FACS), they were infected with increasing levels of Ub-ChITs containing different subtypes Tat's to obtain cells expressing a wide range of Tat (Fig. 4.3). This open-loop construct was used to measure gene expression from the viral promoter as a function of Tat expression using flow cytometry. All 36 Tat-TAR pairs were explored and found to display a wide array of phenotypes that may possibly be mimicked within the >30 million HIV patient pool (Fig. 4.4).

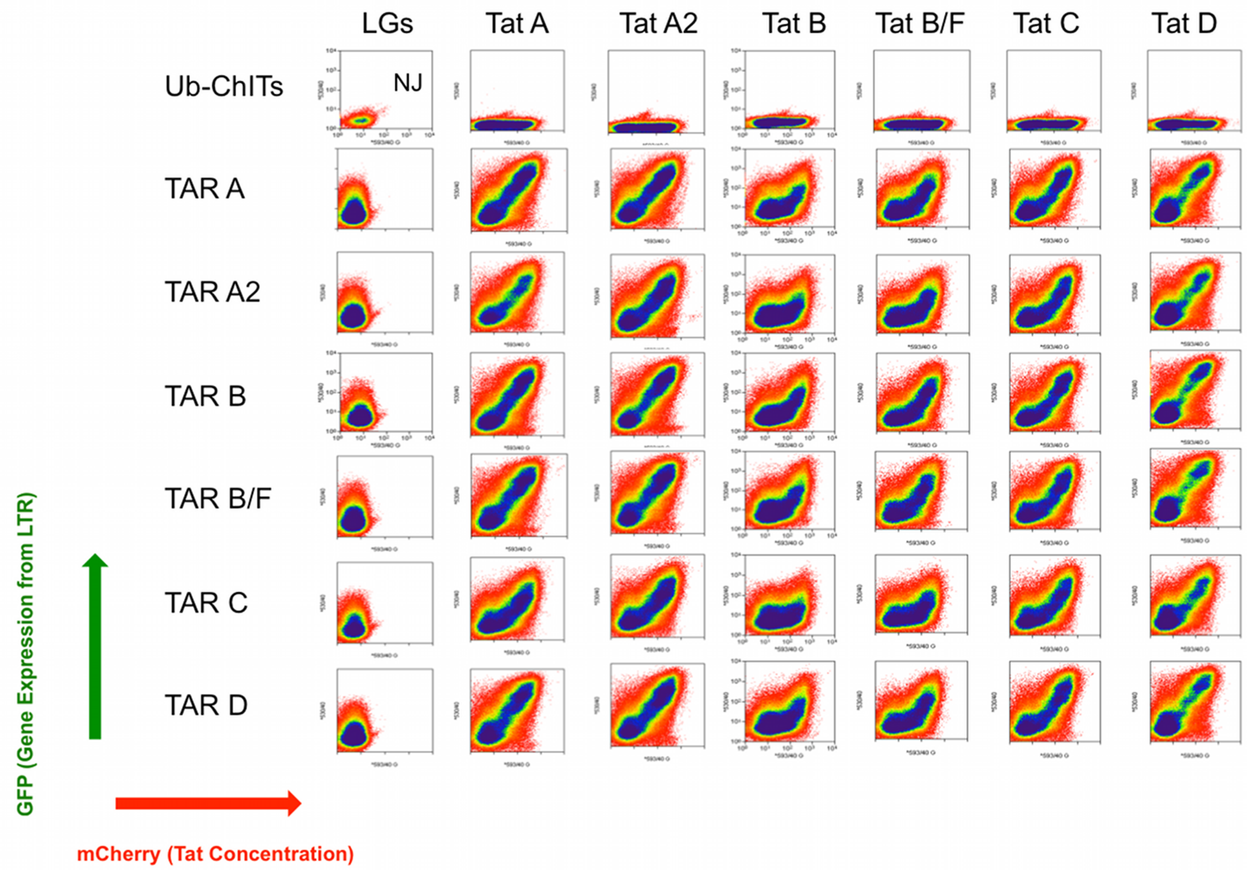
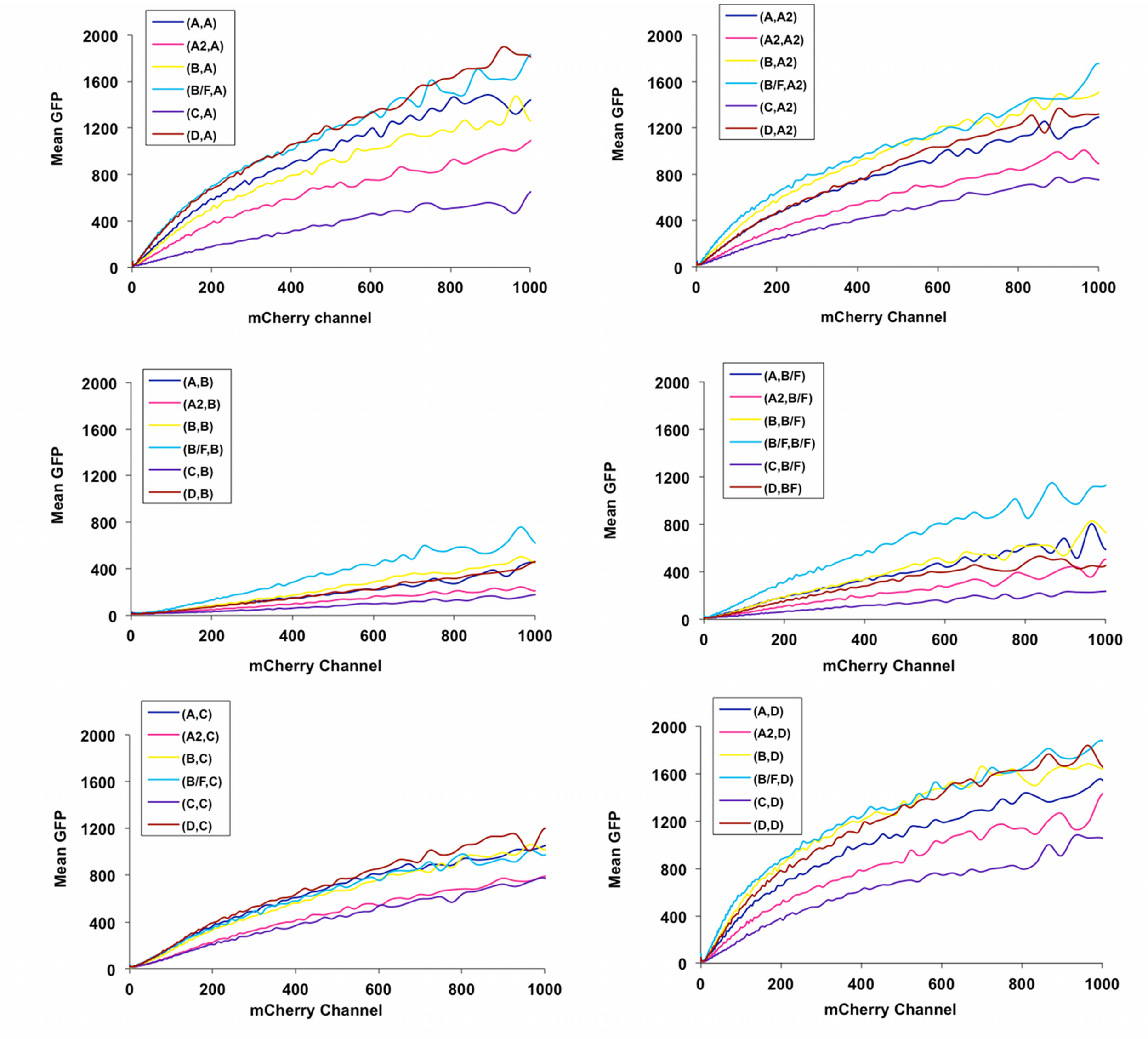
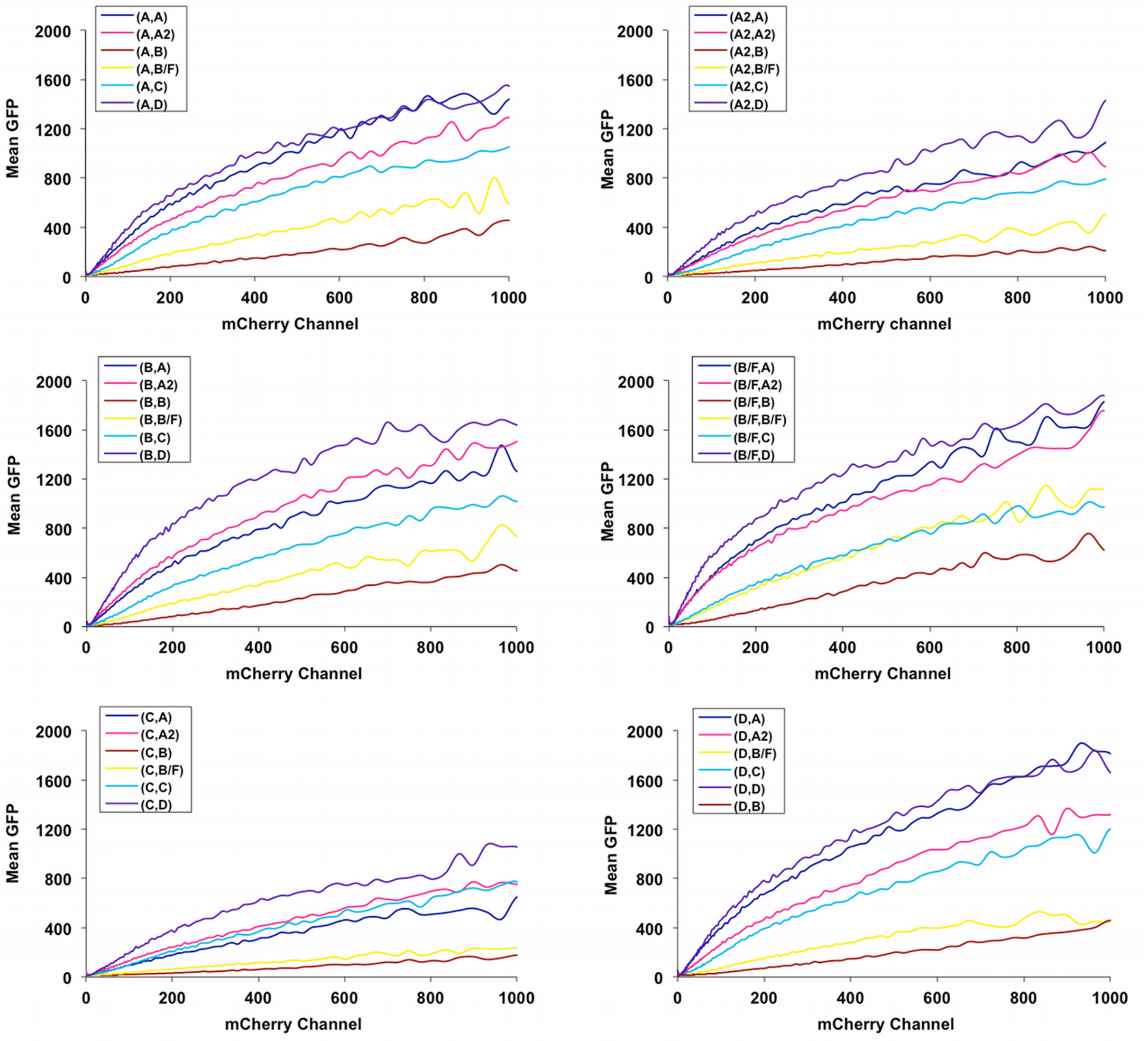


Figure 4.4. Matrix showing density plots for different Tat-TAR pairs in the LG-Ub-ChIT open-loop system. Within each density plot, each dot represents the density of cells. The density of cells increase in the order: red, yellow to blue. The x-axis represents the level of mCherry (or Tat) within the cell and y-axis represents the level of GFP (or gene expression from the viral promoter). The first column shows cells infected with only LGs containing different TARs. The first row represents cells infected with only Ub-ChIT. NJ stands for Naïve Jurkat. These density plots show that the gene expression levels show similar trends for the Tat subtype, independent of the TAR it interacts with.

We observed unique phenotypes for different pairs. From the density plots (Fig. 4.4), we estimated the mean level of gene expression from the HIV promoter at a particular Tat concentration (Fig. 4.5). We found that the initial activation of gene expression and the maximum level of expression attained varied substantially across subtypes (Fig. 4.5). This was used to estimate K_m, the mCherry relative fluorescence units (RFU) at half-maximum GFP RFU. A Chi-square test performed on K_m rejected the null hypothesis that the 36 $\mathrm{K_m}$'s were chosen from a random distribution implying that the different pairs had unique gene expression characteristics.

The density plots of Tat C, and in particular Tat B/F, paired with any of the TARs showed delayed activation of gene expression, such that a large concentration of these Tats is required to upregulate gene expression. This suggested that Tat B/F and C appear weak in their ability to activate gene expression, implying that stochastic effects could become important in directing these subtype viruses into either a transcriptionally repressed (OFF) state or a transactivated (ON) state. The delayed activation of gene expression associated with these subtype Tats suggest that there exists a larger window of Tat concentrations over which the Tat-mediated positive feedback loop may not be activated, allowing the provirus to become latent as the activated T-cell enters into a quiescent state. This result therefore has potential implications for the viral latency propensities of different subtypes and will be explored in greater detail in the next section.





<u>Figure 4.5. Gene expression levels for different TAR-Tat subtype combinations in the open-loop system.</u> (A) Each figure shows a particular Tat subtype paired with 6 other TARs. The figures show the mean level of GFP expression from cells expressing a particular mCherry (or Tat) level. As seen from all the 6 figures, independent of the Tat subtype, TARs B, B/F and D result in higher levels of gene expression compared to TARs A, A2 or C. (B) Each figure shows a particular Tat subtype paired with 6 other TARs. As seen in all the 6 figures, independent of the TAR subtype, Tats D, A and A2 produce higher levels of gene expression compared to Tats C, B/F or B. Further comparing (A) and (B) shows that different subtype Tats produce larger differences in gene expression than subtype TARs, implying that the hairpin TARs appear to act as an scaffold but differences in the interaction of the subtype Tats with TAR and other cellular primarily result in the large differences in gene expression.

Similarly, density plots of TAR C and A2 paired with different Tats also displayed slightly delayed gene activation, implying that certain nucleotide differences in these TARs, compared to those of other subtypes prevent them from interacting effectively with the Tats and suggesting that stochastic effects could play an important role in deciding the fate of such subtype viruses as well.

In contrast, Tat A, A2 and D coupled with the other TARs showed rapid activation of gene expression (Figs. 4.4 and 4.5). Further, the maximum level of gene expression attained with these Tat were dramatically higher than the weaker Tats (Figs. 4.4 and 4.5), suggesting that certain sequence features within these Tat sequences enabled them to be strong transactivators of gene expression that could possibly reduce their propensity to transition into a latent state.

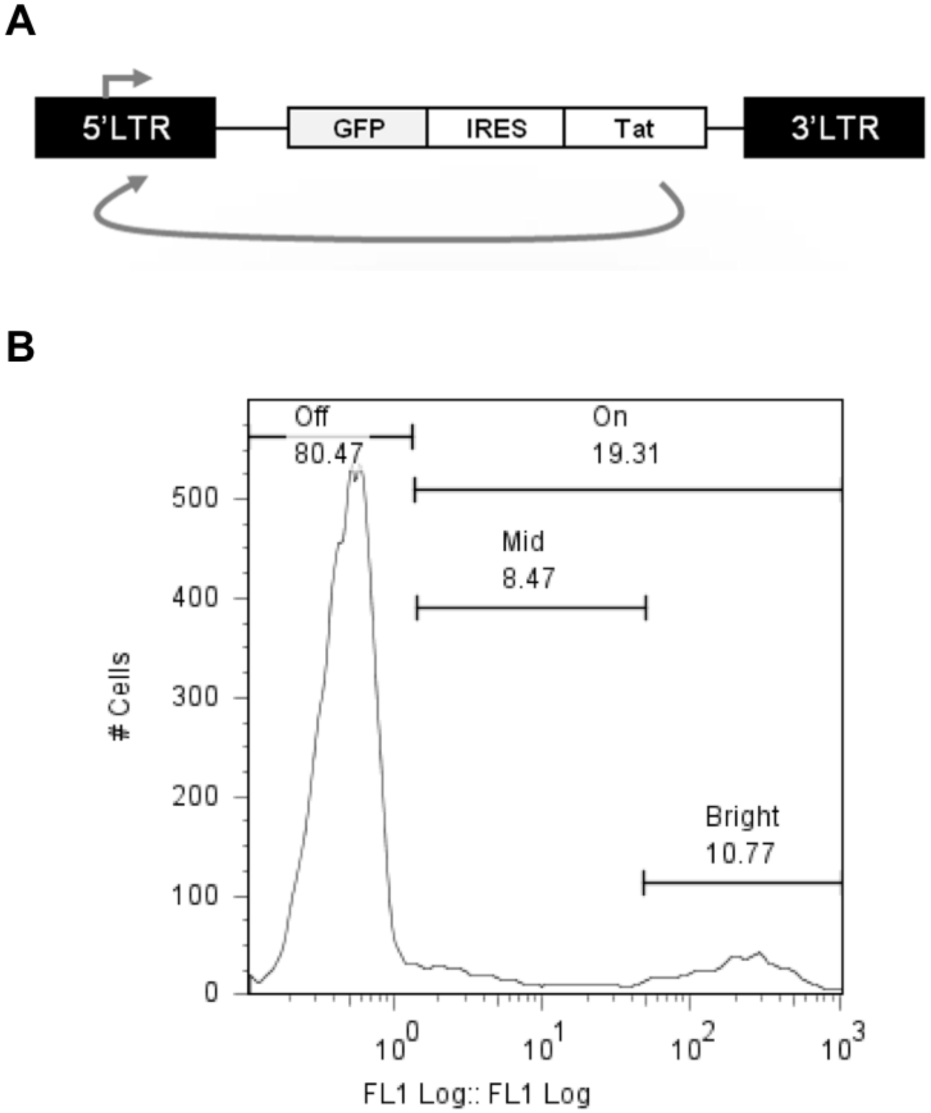


Figure 4.6. LGIT closed-loop system used to mimic the Tat positive-feedback loop in HIV-1. (A) Schematic of the LGIT lentiviral vector. The HIV-1 LTR drives expression of GFP and Tat, separated by IRES. The closed-loop system mimics the minimal Tat positive feedback loop in HIV-1. (B) Typical GFP histogram of Jurkat cells infected with the LGIT vector. The Tat positive-feedback loop results in bimodal gene expression, shown here as the Off and Bright gates. GFP+ cells are indicated as the ON population and cells with intermediate levels of gene expression indicated as Mid cells.

We also found that naturally occurring Tat-TAR pairs did not have the strongest transactivation. For example, the level of gene expression from TAR B/F paired to Tat B/F (denoted as (B/F,B/F)) was much lower than TAR B/F paired to Tat D (denoted as (B/F,D)). This suggests that different subtypes have not necessarily evolved to maximize gene expression. Interestingly, we found that the differences in gene expression for the Tat subtypes were fairly independent of the TAR subtype. For example, the relative differences in the levels of gene expression for the 6 Tat subtypes were similar when paired with either TAR A or B (Fig. 4.5). Similarly, we found that the differences in gene expression for the TAR variants were relatively independent of the Tat subtype used for the comparison. These results suggest that subtype TAR and Tats behave independently to program gene expression from the viral promoter. This may have significant evolutionary implications since it suggests that the Tat-TAR interaction that forms the central axis around which HIV-1 gene expression depends in robust to sequence variations and that different variants can effectively interact to activate gene expression from the viral promoter. The absence of particular sequence variants producing non-linear activation of gene expression suggests that TARs and Tats have evolved independently to maximize chances of activating gene expression and inducing viral replication. This lack of co-evolution between TAR and Tat will be explored systematically in section 4.4.

4.3 Exploring the impact of sequence diversity in TAR and Tat on viral gene expression using a closed-loop system

To test how different TAR-Tat pairs affect gene expression in the context of a closed-loop circuit that mimics the positive feedback loop in HIV-1, we designed a lentiviral vector in which Tat and GFP were placed under the control of HIV-1 LTR. Tat and GFP were separated by an IRES sequence and the vector was called LTR-GFP-IRES-Tat or LGIT (Fig. 4.6*A*) (27).

The open-loop LG-Ub-ChIT experiments described above provided basic information about the strength (or "gain") of different Tat-TAR pairs, allowing us to independently assess the characteristics of a particular subtype TAR or Tat. These experiments led us to test these different pairs in the context of the more biologically relevant closed-loop LGIT system. Along with the six naturally occurring Tat-TAR pairs, 13 other pairs were cloned into the LGIT vector based of their ability to activate gene expression in the open-loop system. A few weak pairs which would potentially have a large fraction of latently infected cells, and a sizeable number of cells with intermediate levels of gene expression exhibiting switching between OFF and ON states and a few strong pairs which would be anticipated to have a majority of the cells in an activated ON state were amongst those that were chosen for further experimental analysis. Based on the bimodal GFP expression observed for LGIT infections, we categorized cells into OFF or ON populations, with the ON population being further subdivided into Mid and Bright populations (Fig. 4.6*B*). We used three metrics to quantify gene expression (30):

1. Mean Bright Peak – The mean GFP level of cells in the Bright gate. We used it as a measure to quantify the level of closed-loop transactivation for a particular Tat-TAR pair.

2. Percentage Infected but Off – The percentage of cells in the OFF state out of the total number of infected cells. We used it to quantify the contribution of a particular Tat-TAR pair to the cell culture equivalent of a "latent infection" and the inability of the pair to activate gene expression. To obtain this parameter, the infected Jurkat cells were stimulated by the addition of the drugs TNFα (stimulates gene expression by activating

the NF-κB pathway) and TSA (an inhibitor of histone deacetylases) and the increase in the number of ON cells after stimulation was used to estimate this parameter.

3. Mid:On Ratio – The ratio of the number of cells in the Mid region to the total number of ON cells. It is a measure of the fraction of cells that display intermediate levels of gene expression which has previously been associated with stochastic gene expression allowing the provirus to flip between OFF and Bright states.

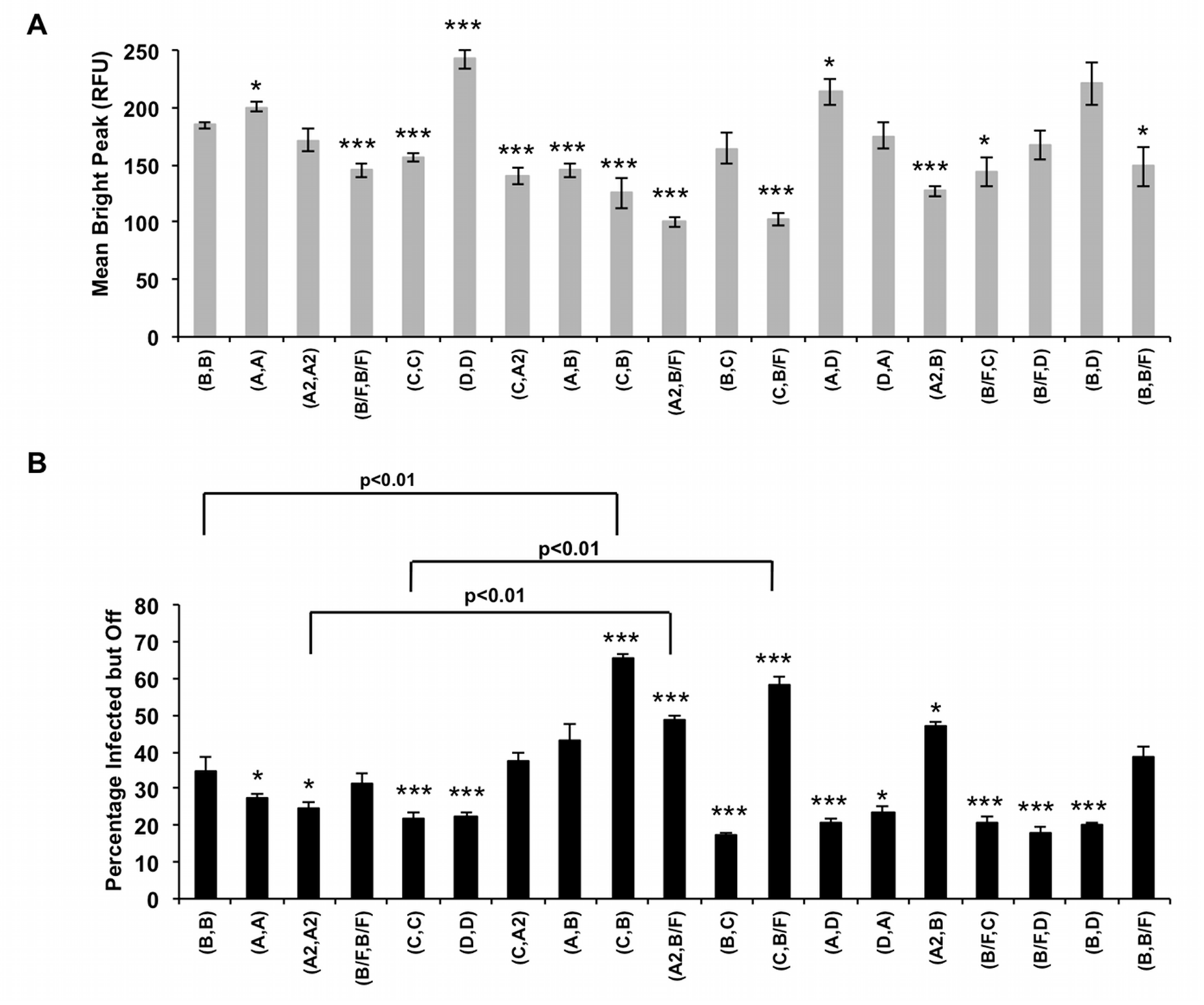


Figure 4.7. Subtype Tat-TAR combinations show dramatic differences in the level of gene expression and propensity for latency. (A) The Mean Bright Peak levels, a measure of the strength of gene expression from the viral promoter, shows that subtypes show large differences in gene expression. See text for details. (B) The Percentage Infected but Off, a measure of the propensity for latency shows that different Tat-TAR combinations can produce large differneces in latency. For example, changing the Tat subtype from A2 to B/F in (A2,A2) to (A2.B/F) doubles the number of latent cells. Similarly, changing the TAR subtype from B to C in (B,B) to (C,B) doubles the propensity for latency. Error bars represent S.D. '*' and '***' indicate statistically significant ($p<0.05$ and $p<0.01$, respectively) differences from (B,B). Statistically significant differences between a few other Tat-TAR pairs are shown within the figure.

Among the naturally occurring pairs, (D,D) had the highest Mean Bright Peak followed by (A,A) and (B,B) (Fig. 4.7*A*). In agreement with this data, the stronger Tat-TAR pairs, (D,D) and (A,A) had the lowest Mid:On Ratio, suggesting that these subtypes strongly activate gene expression from the viral promoter. Interestingly, despite having a large Mean Bright Peak value, (D,D) also had a substantial number of latent cells (Fig. 4.7). Similarly, (B,B) had largest fraction of latently infected cells amongst naturally occurring pairs although its Mean Bright Peak was not the lowest. This suggests that within naturally occurring pairs, the level of transactivation was not necessarily inversely related to the fraction of latent cells, suggesting that Tat-TAR pairs may be regulating multiple mechanisms to produce the observed levels of gene expression and latent cells. Finally, these parameters suggest that different naturally occurring subtypes have dramatically different gene expression levels at steady state, in agreement with the open-loop system that showed differences in the levels of gene expression at the highest Tat concentrations.

In analysis of the chimeric Tat-TAR pairs, as expected, some of the pairs that exhibited strong gene expression in the open-loop system, such as (A,D) and (B,D), also had the highest Mean Bright Peak values (Fig. 4.7*A*). However, (B,D) did not have the lowest number of latently infected cells. A comparison of (B,D) and (B/F,C) shows that although the Mean Bright Peak of (B,D) was significantly greater than (B/F,C); they had almost identical number of latently infected cells implying that strong viral gene expression and propensity for latency may be independent variables that are not correlated (Fig. 4.7). Amongst many of the weaker Tat-TAR pairs, the steady-state Mean Bright Peak was observed to be substantially lower than many of the naturally occurring pairs, in agreement with the open-loop LG-Ub-ChIT experiments. Further, the weaker Tat-TAR pairs had a larger population of latently infected cells in most cases. (C,B) had the most striking phenotype with ~65.3% latently infected cells, almost twice the number of latently infected cells in (B,B), the naturally occurring pair with the largest population of latent cells. This phenotype was particularly interesting since TAR B and TAR C differ at very few nucleotide positions suggesting that few sequence changes can dramatically alter latency characteristics. Similarly, in changing the Tat subtype from A2 to B/F in (A2,A2) and (A2,B/F), the fraction of latently infected cells increases two-fold. Comparing (C,C) and (C,B/F) also shows that the amino acid variations between subtype Tat C and B/F can greatly alter the propensity for latency (Fig. 4.7*B*). These results suggest that a few nucleotide or amino acid substitutions in TAR or Tat can dramatically alter gene expression characteristics implying that the huge viral quasispecies within a single patient may have very divergent properties that may help maximize viral fitness.

Finally, based on the open loop and closed loop experiments discussed above, we were able to assign strength of transactivation to different TARs and Tat's:

TAR: B>B/F>D>A>A2>C

Tat: D>A>A2>B>C>B/F

Since it appeared possible that a few nucleotide changes in TAR or amino acid changes in Tat could underlie the dramatic diversity in gene expression properties observed, we decided to rationally introduce point mutations into TAR and Tat to relate its sequence to the observed phenotype. Mutations were made in Tat B to mimic either weaker Tats such as B/F (R56H), and those in TAR B to mimic weaker TARs such as TAR C (A22G) with the hypothesis that this would result in weaker gene expression (Fig. 4.8). Since Tat B/F was the weakest Tat that had a mutation R56H with the ARM region, we decided to introduce this mutation in Tat B and test its effect on gene expression. We found that the R56H mutation in Tat B lowered its ability to activate gene expression and that it closely mimicked that of subtype B/F. Finally, to conclusively show that this site plays an important role in setting the level of gene expression in Tat B/F, we decided to make the reverse mutation H56R in Tat B/F with the hope of restoring gene expression close to the stronger Tat B. We found that the H56R mutation in Tat B/F rescued gene expression close to Tat B suggesting that this site within the ARM region independently plays an important role in setting the level of gene expression (Fig. 4.8*A*). Next, since TAR C was the weakest TAR, we decided to make the mutation A22G in TAR B, which is just below the bulge in TAR with the hypothesis that this mutation may alter the structure of the bulge resulting in a change in gene expression. As expected, the A22G mutation in TAR B reduced gene expression close to that of TAR C. As in the case of the Tat mutations, we decided to test the reverse mutation in TAR C to see if it restores gene expression. This was successfully validated, suggesting that site 22 in TAR primarily helps in differentiating the levels of gene expression and the propensity for latency between TAR B and C (Fig. 4.8*B*).

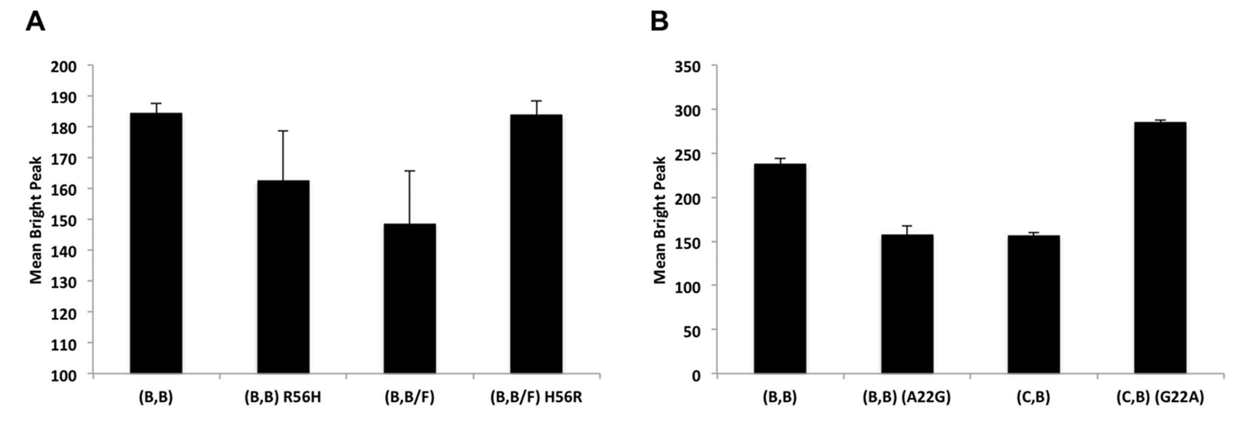


Figure 4.8. Single point mutations can alter gene expression phenotypes to mimic other subtypes. (A) Single point mutation R56H in subtype B Tat results in lowering of gene expression such that it mimics subtype B/F. Similarly, introducing the reverse mutation H56R in subtype B/F restores gene expression close to subtype B. (B) Single point mutation A22G in subtype B TAR results in loss of gene expression to mimic subtype C whereas the compensating mutation G22A in subtype C TAR increases gene expression to levels similar to subtype B.

Thus these point mutations help validate our observations from the open- and closed-loop systems that a few changes in the sequence of TAR and Tat can change gene expression dramatically. These point mutations also suggest that the amino acid or nucleotide at these positions independently alter gene expression suggesting the lack of cooperativity or coevolution between TAR and Tat. This hypothesis is also supported by the open-loop experiments that show that each Tat subtype appears to show specific gene expression characteristics independent of the TAR subtypes. Similarly, while the TAR subtypes show more subtle differences, their characteristics also appear to be independent of the Tat subtype. Thus, TARs and Tats from different subtypes seem to act additively in setting the level of gene expression pointing towards the lack of coevolution between this interacting viral protein and hairpin RNA structure. This lack of coevolution between TAR and Tat is analyzed quantitatively in the following section.

4.4 Subtype Tat and TARs appear to have evolved independently to produce a robust mechanism of activating gene expression from the viral promoter

As described in Chapter 2 (Fig. 2.16), we used Mutual Information (MI) to estimate coevolution between nucleotides in TAR and amino acids in Tat. As seen in Figure 2.16 and in Figure 4.9, most pairs of sites have MI below the threshold score. Off the 5 pairs of sites that have MI higher than the threshold score, the pair with the highest score is between site 48 in TAR and residue 35 in Tat. Since we have previously identified site 35 to be coevolving primarily with site 39 in Tat (Chapter 3, Figs. 3.2 and 3.4), the correlation between sites 48 in TAR and 35 in Tat is possibly due to these sites having relatively high entropy (Fig. 4.9) that gives rise to background MI that is not accurately corrected for. Further, the site pair 48-35 have a MI score that is only marginally higher than the threshold score as compared to sites 35-39 in Tat (Fig. 3.2*B*). Thus, we hypothesized that these 5 site pairs may not be functionally coevolving and that the observed statistical correlation is a consequence of phylogenetic history of these sequences. Thus from these 5 site pairs we decided to first experimentally study sites 22 in TAR and 54 in Tat since in addition to being marginally above the threshold score, site 22 in TAR is close to the bulge region whereas site 54 in Tat is within the ARM motif suggesting that these sites are within functionally important motifs and possibly within close physical proximity that would increase their chances of coevolving.

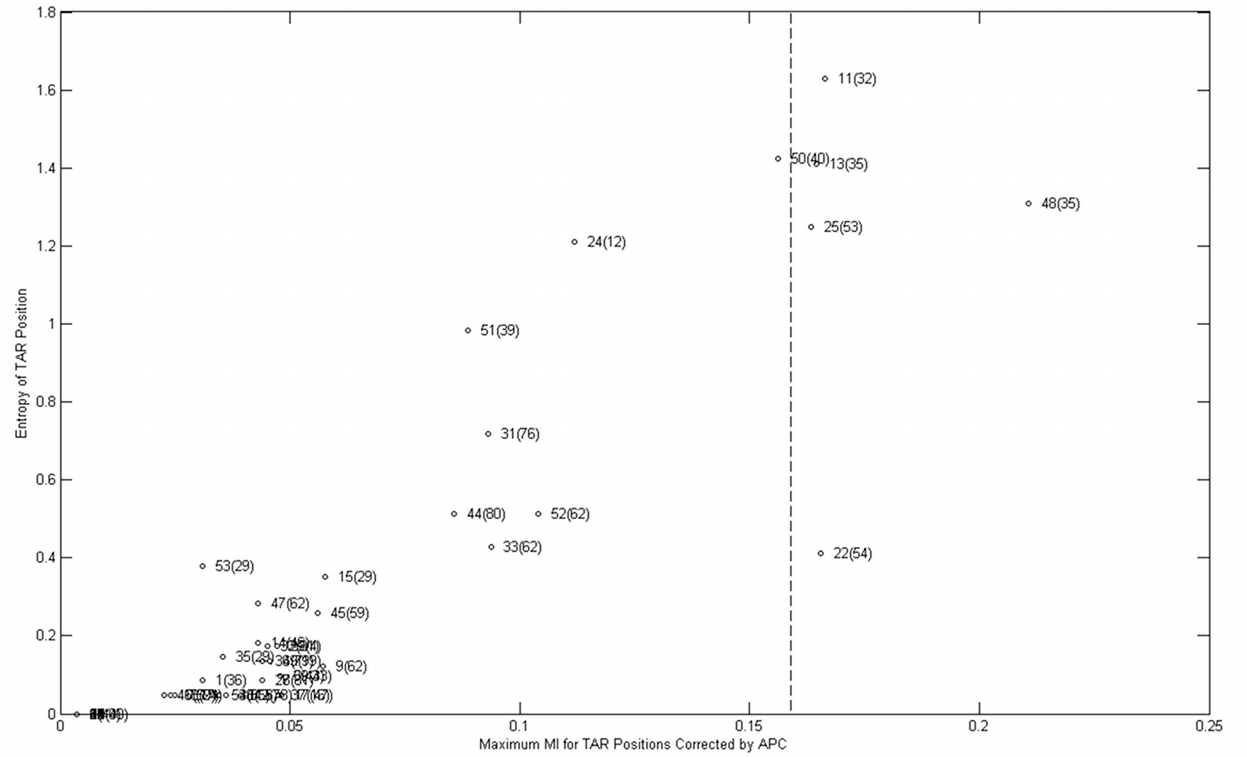
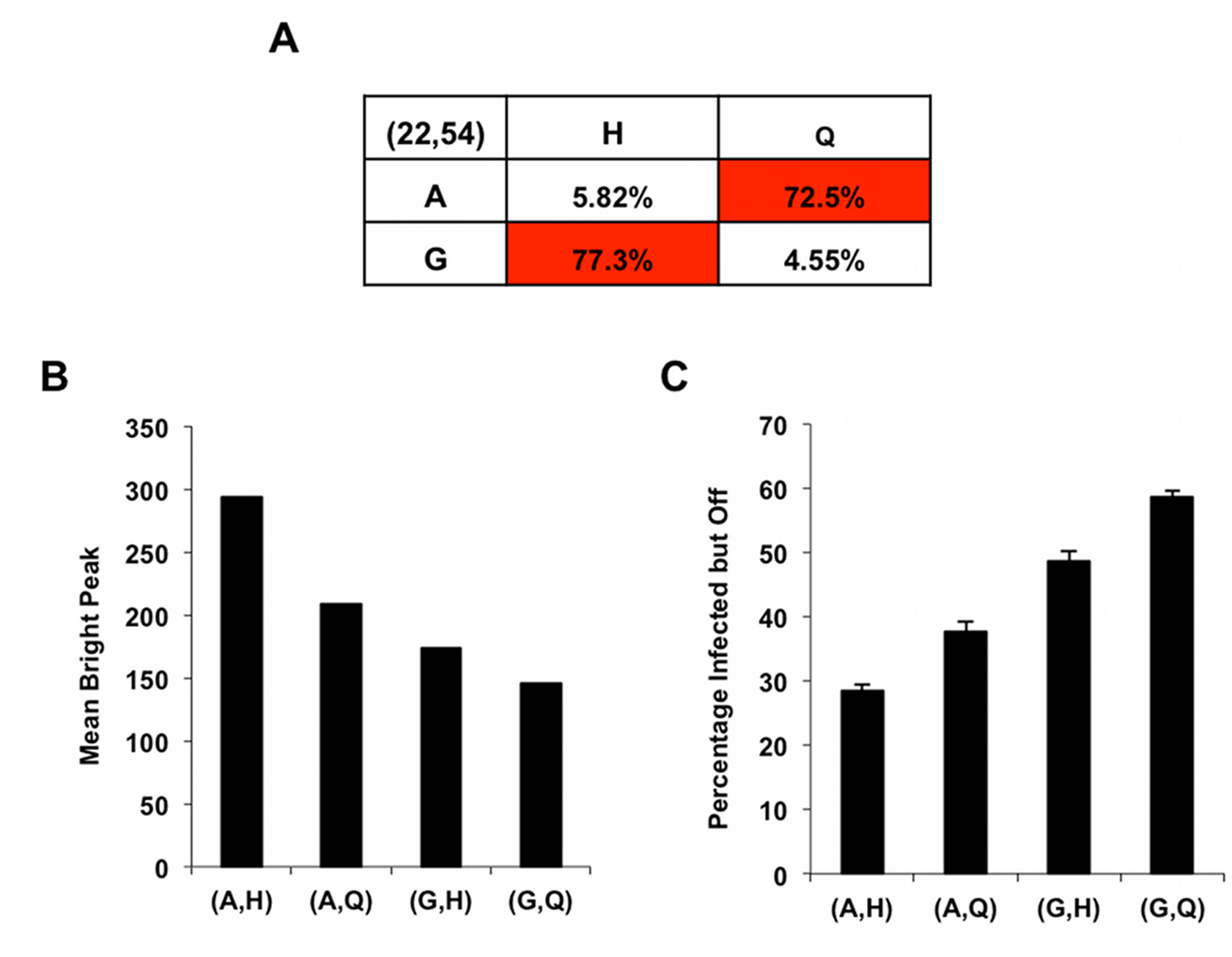


Figure 4.9. Mutual Information analysis shows weak coevolution signals between nucleotides in TAR and amino acids in Tat. Plot shows the entropy of a site in TAR and the maximum corrected MI of that TAR site with any other site in Tat. The site in TAR is indicated besides each dot and the site in Tat with which this position in TAR shares the highest MI signal is shown within parenthesis. 5 pairs of sites have MI scores marginally over the threshold score.

Based on the frequencies of nucleotides and amino acids at these two positions, we introduced the mutation Q54H in Tat B and A22G in TAR B (Fig. 4.10*A*). As in Chapter 3, we hypothesized that single point mutations at correlated sites should result in a loss of function that is rescued in the double mutant. We found no loss of gene expression upon introduction of these single mutations. Q54H in Tat resulted in stronger gene expression than WT Tat (Fig. 4.10*B*). While the single point mutations A22G in TAR resulted in weaker gene expression, the double mutant Q54H and A22G showed additive behavior, that is, the partial loss of gene expression in A22G is partly restored by the mutation Q54H that appears to increase gene expression (Fig. 4.10*B*,*C*). This suggests that these sites are not coevolving but instead display additive behavior implying that they act independently to regulate gene expression from the viral promoter.



(22,54)	H	Q
A	5.82%	72.5%
G	77.3%	4.55%

Figure 4.10. Experimental evidence to show that sites in Tat are not coevolving with those in TAR. (A) Frequency of sequences that have a particular combination of nucleotide and amino acid at positions 22 in TAR and 54 in Tat, respectively. A truncated matrix is shown here. Each row in the complete matrix sums to 100%. The red background indicates nucleotide-amino acid combinations that are most highly prevalent within naturally occurring sequences. Based on this matrix, we might expect to observe site pairs (A,H) and (G,Q) to have reduced gene expression levels. (B) and (C) In contract to the hypothesis outlined in (A), the mutation Q54H in Tat globally and independently activates gene expression (and thereby lowers the Percentage of Infected but Off cells). Similarly, the mutation A22G in TAR results in a loss of gene expression independent of the amino acid present at site 54 in Tat. These data indicate that sites in TAR and Tat do not interact with other and that each acts independently to activate or repress gene expression. Error bars indicate S.D.

4.5 Discussion

In this work, we have systematically studied the role of sequence diversity in the RNA hairpin TAR and viral protein Tat in regulating gene expression from the HIV promoter. We initially designed an open-loop system to independently assess the role of subtype TARs and Tats in activating gene expression. We found large differences in their ability to activate gene expression suggesting that certain subtype TARs and Tats may have increased propensity for latency. Importantly, we also discovered that in addition to these significant differences in gene expression, subtype TAR and Tats act independently to activate gene expression from the viral promoter. To study the differences in gene expression more systematically, we introduced these subtype TARs and Tats into a minimal lentiviral vector that retains the positive feedback loop of HIV-1. These studies in the closed-loop system confirmed the previous results by showing that subtypes display dramatically different levels of gene expression and latency and that a few specific mutations in TAR or Tat can dramatically alter the regulation of gene expression. To validate the latter observation from the open-loop system that TAR and Tat sequences act independently, we employed MI and site-directed mutagenesis to show that sites in TAR do not appear to be correlated with residues in Tat.

While other site pairs need to be tested to conclusively confirm the lack of coevolution between elements of TAR and Tat, these preliminary experiments in addition to the open loop system suggests that sites in TAR and Tat act in a mutually exclusive manner in setting the level of gene expression from the viral promoter. While the interaction of TAR and Tat is critical to recruit host cellular factors to the HIV promoter, the RNA hairpin and the viral protein are not constrained to particular sites to ensure the formation of the RNA-protein complex. Evolutionarily, this suggests that TAR-Tat mediated positive feedback loop in HIV is robust to sequence perturbations and not constrained to certain features in either TAR or Tat. As long as the TAR and Tat are individually functionally active, a wide variety of sequence diversity in TAR and Tat can be tolerated to activate the positive feedback loop. Thus the virus is able to effectively insulate the activation of gene expression from the high rates of mutation and recombination that potentially provide the virus with fitness advantages.

4.6 Materials and Methods

4.6.1 Plasmids

pcDNA 3.1 plasmid with the 5’ LTR and the pBS KSPS plasmid with the 3’LTR inserted in it, created by previous graduate students in the Schaffer group, were used to insert different subtype TARs using the restriction sites *Afl* II and *Bsa* I from the plasmids provided by NIH. The 3’LTR (with different subtype TARs) were then inserted into the pCLG plasmid (used to generate the LG infected cells) and pCLGIT plasmid (used to generate the LGIT infected cells) using restriction enzymes *Pme* I and *Xho* I. The 5’ LTR was then inserted in the next step into the pCLG and pCLGIT plasmids using *Mlu* I and *Not* I. Since Tat is encoded by two exons in wild-type HIV, we performed splice overlap PCR on the plasmids provided by NIH to combine the two exons to insert a single exon Tat into the pCLGIT and Ub-ChIT plasmids using the restriction enzymes *BstX* I and *Xho* I. Tat and TAR Mutants were created using QuikChange PCR (Stratagene). The TAR mutants were made in pcDNA 3.1 and pBS KSPS and then cloned into pCLGIT as described above.

4.6.2 Cell Culture

Jurkat cells were cultured in RPMI 1640 (MediaTech) with 10% fetal bovine serum (FBS) (Invitrogen). HEK 293T cells were cultured in Isocove's Modified Dulbecco's Medium (IMDM) (MediaTech) with 10% FBS.

4.6.3 Transfection and Virus Purification

Human embryonic kidney cells (HEK 293T) were cotransfected with the plasmid of interest (pCLG, pCLGIT or Ub-ChIT) along with three helper plasmids: (1) a pseudotyped plasmid encoding for the envelope from the vesicular stomatis virus (VSVG), (2) a plasmid encoding *rev* protein and (3) a plasmid that encodes for *gag* and *pol* proteins (31).

To harvest and purify virus, 10 mL viral supernatant from the HEK293T cells were centrifuged 36 hours post transfection at 2000 rpm for 2 minutes to remove cell debris. The supernatant was further filtered through a 0.45 μm filter and 1 mL of 20% sucrose was added to the bottom of the supernatant. The supernatant was ultracentrifuged at 25,000 rpm for 1.5 hours at 4^0C. The viral pellet obtained was then resuspended in 20 μL phosphate buffer saline (PBS) and stored at -80^0C for future infections.

4.6.4 Virus Titering, Infection and Stimulation by Drugs

Viral titers were obtained by infecting $3x10^5$ Jurkat cells with different volumes (0.1 μL to 5 μL) of virus and counting cells expressing GFP 8 days post-infection after stimulating the cells with TNFα and TSA and assuming a Poisson distribution for infection. Once the viral titers were obtained, $3x10^5$ Jurkat cells were infected with the appropriate volume of virus to obtain a multiplicity of infection (MOI) of 0.05 - 0.1 to ensure single viral integration per cell.

To calculate the number of latently infected cells, the cells were stimulated with TNF-α (20 ng/mL) and TSA (400 nM) once steady state GFP expression was attained (~8 days post infection) and analyzed by flow cytometry 18 hours post-stimulation.

4.6.5 Flow Cytometry and Cell Sorting

The GFP fluorescence in the closed-loop LGIT experiments was measured using a FC500 flow cytometer (Beckman-Coulter) with a 488 nm laser. GFP fluorescence was monitored through a 530 nm filter (FL1 channel). The GFP and mCherry fluorescence were measured in the open-loop LG-Ub-ChIT experiments using the Cytopeia Influx sorter with a 488nm and 561nm laser, respectively. GFP fluorescence was monitored through a 530nm filter (FL1 channel), and mCherry fluorescence was monitored through a 593nm filter (FL3 channel). Jurkat cells infected with the LG vector were sorted for GFP+ cells on a DAKO-Cytomation MoFlo High Speed Sorter.

4.6.6 Mutual Information Analysis

Code for Mutual Information to analyze sequence alignment data from the Los Alamos Sequence Data was written in Matlab®. These codes will be made available upon request.

4.7 References

1. Negroni, M., and Buc, H. (2001) *Annu Rev Genet* **35**, 275-302
2. Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004) *Nature reviews. Genetics* **5**, 52-61
3. Osmanov, S., Pattou, C., Walker, N., Schwardlander, B., and Esparza, J. (2002) *Journal of acquired immune deficiency syndromes (1999)* **29**, 184-190
4. Spira, S., Wainberg, M. A., Loemba, H., Turner, D., and Brenner, B. G. (2003) *J Antimicrob Chemother* **51**, 229-240
5. Abraha, A., Nankya, I. L., Gibson, R., Demers, K., Tebit, D. M., Johnston, E., Katzenstein, D., Siddiqui, A., Herrera, C., Fischetti, L., Shattock, R. J., and Arts, E. J. (2009) *J Virol* **83**, 5592-5605
6. Desfosses, Y., Solis, M., Sun, Q., Grandvaux, N., Van Lint, C., Burny, A., Gatignol, A., Wainberg, M. A., Lin, R., and Hiscott, J. (2005) *J Virol* **79**, 9180-9191
7. Koulinska, I. N., Villamor, E., Msamanga, G., Fawzi, W., Blackard, J., Renjifo, B., and Essex, M. (2006) *Virus Res* **120**, 191-198
8. Iordanskiy, S., Waltke, M., Feng, Y., and Wood, C. (2010) *Retrovirology* **7**, 85
9. Jeeninga, R. E., Hoogenkamp, M., Armand-Ugon, M., de Baar, M., Verhoef, K., and Berkhout, B. (2000) *J Virol* **74**, 3740-3751
10. van Opijnen, T., Jeeninga, R. E., Boerlijst, M. C., Pollakis, G. P., Zetterberg, V., Salminen, M., and Berkhout, B. (2004) *J Virol* **78**, 3675-3683
11. Roof, P., Ricci, M., Genin, P., Montano, M. A., Essex, M., Wainberg, M. A., Gatignol, A., and Hiscott, J. (2002) *Virology* **296**, 77-83
12. Martinez-Cajas, J. L., Pant-Pai, N., Klein, M. B., and Wainberg, M. A. (2008) *AIDS Rev* **10**, 212-223
13. Johnson, V. A., Brun-Vezinet, F., Clotet, B., Gunthard, H. F., Kuritzkes, D. R., Pillay, D., Schapiro, J. M., and Richman, D. D. (2010) *Topics in HIV medicine : a publication of the International AIDS Society, USA* **18**, 156-163
14. Burnett, J. C., Lim, K. I., Calafi, A., Rossi, J. J., Schaffer, D. V., and Arkin, A. P. (2010) *J Virol* **84**, 5958-5974
15. Gao, F., Robertson, D. L., Carruthers, C. D., Morrison, S. G., Jian, B., Chen, Y., Barre-Sinoussi, F., Girard, M., Srinivasan, A., Abimiku, A. G., Shaw, G. M., and Sharp, P. M. H., B.H. . (1998) *J Virol* **72**, 5690-5698
16. Lole, K. S., Bollinger, R. C., Paranjape, R. S., Gadkari, D., Kulkarni, S. S., Novak, N. G., Ingersoll, R., Sheppard, H. W., and Ray, S. C. (1999) *J Virol* **73**, 152-160
17. Gao, F., Vidal, N., Li, Y., Trask, S. A., Chen, Y., Kostrikis, L. G., Ho, D. D., Kim, J., Oh, M. D., Choe, K., Salminen, M., Robertson, D. L., Shaw, G. M., Hahn, B. H., and Peeters, M. (2001) *AIDS Res Hum Retroviruses* **17**, 675-688
18. Rodenburg, C. M., Li, Y., Trask, S. A., Chen, Y., Decker, J., Robertson, D. L., Kalish, M. L., Shaw, G. M., Allen, S., Hahn, B. H., and Gao, F. (2001) *AIDS Res Hum Retroviruses* **17**, 161-168
19. Roy, S., Delling, U., Chen, C., Rosen, C. A., and Sonenberg, N. (1990) *Genes & Dev* **1990**, 1365-1373
20. Dingwall, C., Emberg, I., Gait, M. J., Green, S. M., Heaphy, S., Karn, J., Lowe, A. D., Singh, M., and Skinner, M. A. (1990) *EMBO J* **9**, 4145-4153
21. Aboul-ela, F., Karn, J., and Varani, G. (1995) *Journal of molecular biology* **253**, 313-332

22. Mujtaba, S., He, Y., Zeng, L., Farooq, A., Carlson, J. E., Ott, M., Verdin, E., and Zhou, M. M. (2002) *Molecular cell* **9**, 575-586
23. Richter, S., Ping, Y. H., and Rana, T. M. (2002) *Proceedings of the National Academy of Sciences of the United States of America* **99**, 7928-7933
24. Dorr, A., Kiermer, V., Pedal, A., Rackwitz, H. R., Henklein, P., Schubert, U., Zhou, M. M., Verdin, E., and Ott, M. (2002) *EMBO J* **21**, 2715-2723
25. Kiernan, R. E., Vanhulle, C., Schiltz, L., Adam, E., Xiao, H., Maudoux, F., Calomme, C., Burny, A., Nakatani, Y., Jeang, K. T., Benkirane, M., and Van Lint, C. (1999) *EMBO J* **18**, 6106-6118
26. Pagans, S., Kauder, S. E., Kaehlcke, K., Sakane, N., Schroeder, S., Dormeyer, W., Trievel, R. C., Verdin, E., Schnolzer, M., and Ott, M. (2010) *Cell host & microbe* **7**, 234-244
27. Weinberger, L. S., Burnett, J. C. T., J. E. , Arkin, A. P., and Schaffer, D. V. (2005) *Cell* **122**, 169-182
28. Suryavanshi, G. W., and Dixit, N. M. (2007) *PLoS Computation Biology* **3**, 2003-2018
29. Jetzt, A. E., Yu, H., Klarmann, G. J., Ron, Y., Preston, B. D., and Dougherty, J. P. (2000) *J Virol* **74**, 1234-1240
30. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P., and Schaffer, D. V. (2009) *PLoS pathogens* **5**, e1000260
31. Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., and Naldini, L. (1999) *J Virol* **72**, 8463-8471

Chapter 5: Sequence and Architecture Variations within Transcription-Factor Binding Sites in the HIV-1 Promoter Differentially Regulate Viral Gene Expression, Replication and Latency

5.1 Introduction

Although highly active anti-retroviral therapy (HAART) has been extremely effective in reducing viral loads to undetectable levels and prolonging the lives of people infected with HIV-1, it was not been possible to completely eradicate the virus from a patient due to the presence of latent viral populations (1,2). These latent populations remain undetected by the surveillance mechanisms of the immune system and can therefore reactivate during late stages of the disease to initiate rapid viral replication and progress to acquired immunodeficiency syndrome (AIDS) (3,4).

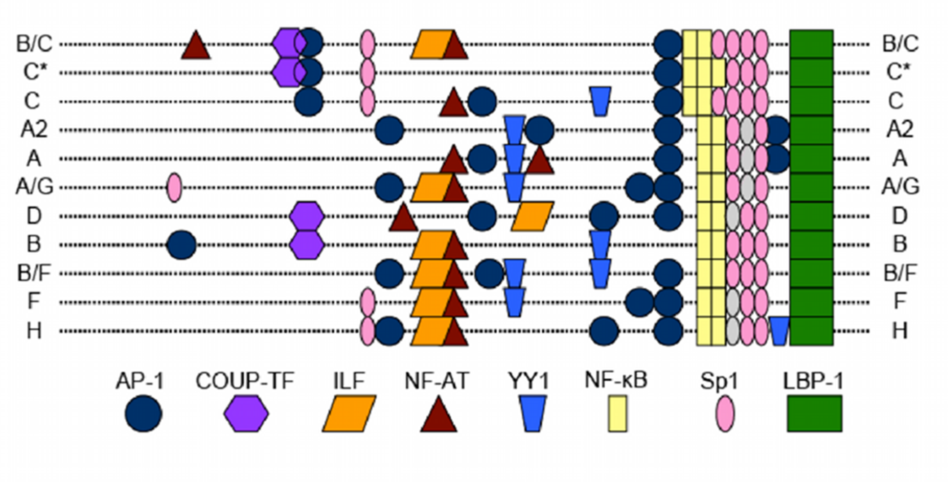


Figure 5.1. Schematic representation of transcription factor binding sites found in different HIV-1 subtype promoters. Pictorial representation shows that subtype promoters can have very different architectures of TFBS in addition to sequence variation within a single TFBS. Adapted from (5).

Understanding the origins of viral latency is therefore critical to identifying mechanisms by which dormant populations can be purged out of patients (6-9). Several underlying causes have been reported, including the role of the viral protein Tat in initiating gene activation (10,11), the TAR-Tat interaction (12), the site of viral integration (13-18) and cellular miRNAs (19) that suppress synthesis of new viral progeny. Once HIV-1 integrates semi-randomly within the host genome, recruitment of host factors to the viral promoter either initiates active transcription and production of new viral particles or transcriptional silence to enter a latent state. This decision between a lytic or latent state depends on the recruitment of host transcription factors to the viral promoter that initiates production of Tat, an early viral product. The viral protein Tat binds to TAR to initiate a positive-feedback loop that leads to robust viral gene expression (20,21). Since the viral promoter initiates transcription after viral integration or just before reactivation from latency, understanding how sequence diversity within the viral promoter may regulate gene expression is important for determining its role in the establishment and reactivation from latency (22,23).

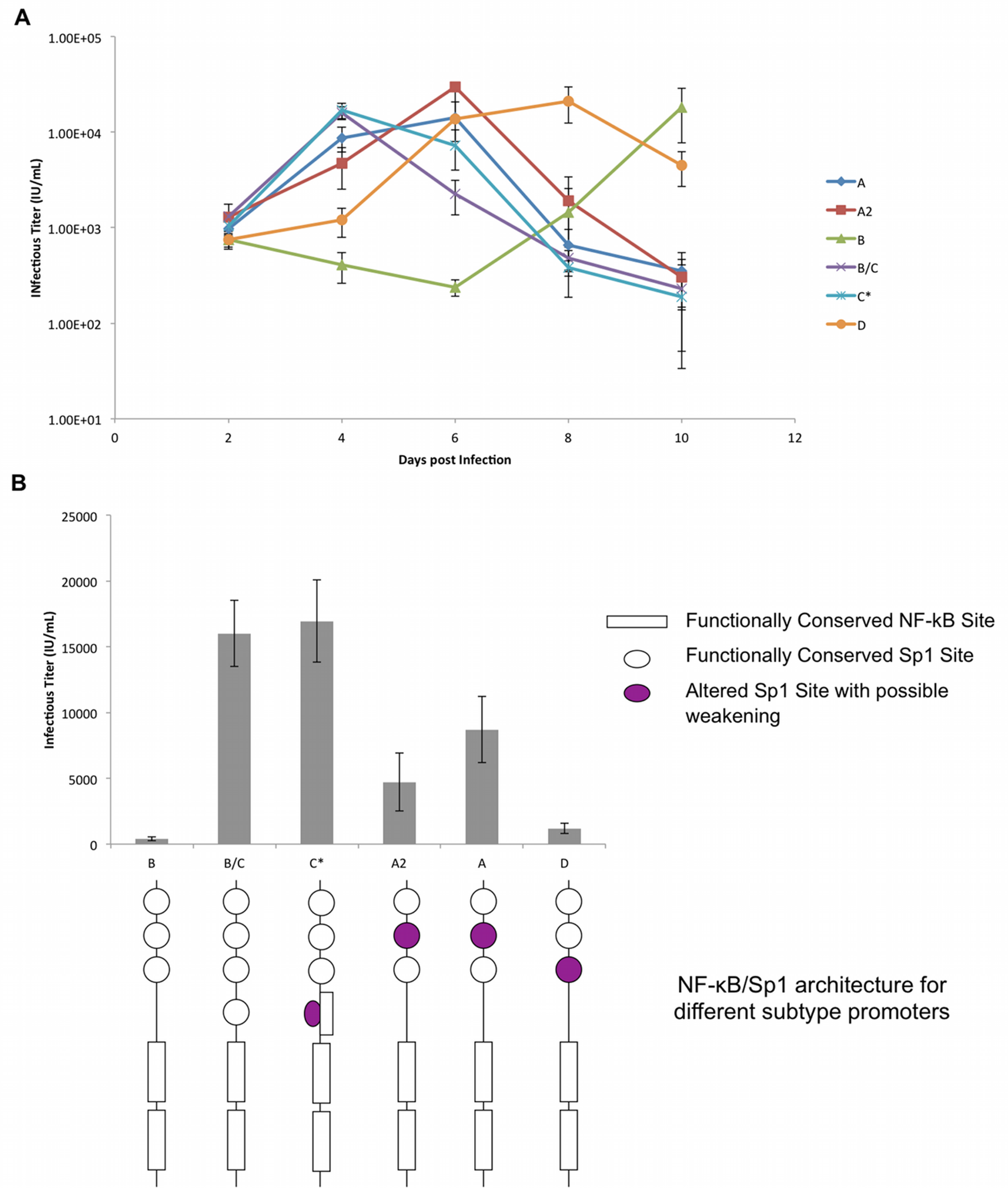


Figure 5.2. Subtype promoters produce large differences in viral replication rates. (A) Time course data of viral replication for full-length virus with promoters from different subtypes. The rest of the viral genome corresponds to subtype B. SupT1 cells were infected by the full-length virus and infectious titers were estimated using an indicator cell-line expressing GFP. Figure shows that subtypes such as C* and B/C show rapid initial replication, followed by subtypes such as A and A2. In comparison, subtypes such as B and D show very delayed replication. (B) The infectious titer during early stages of replication (day 4) are shown in the bar chart. Figure shows that subtypes such as B/C and C* that have an extra NF-κB/Sp1 site show rapid replication. In comparison,

other subtypes such as D that have mutations in Sp1 sites have reduced viral replication rates. Schematic representation of the architecture of the NF-κB/Sp1 sites for the different subtypes are shown at the bottom. Altered Sp1 sites that possibly result in weakening of viral replication rates are shown in purple. Experiments were performed in biological triplicate and error bars indicate standard deviations from the mean.

The HIV-1 promoter contains several transcription factor binding sites (TFBS) such as NF-κB, Sp1, YY1, NF-AT, LBP, COUP-TF, AP1, RBE III, Ets-1 including several others (24-26). These factors can recruit activating and repressive factors to the promoter to initiate gene activation or silencing. Binding of these factors to the promoter has been shown to recruit activating complexes such as histone acetyl transferases (HAT) and repressive factors such as histone deacetyl transfereases (HDAC) that may influence the fate of the virus by either driving it towards active replication or latency, respectively (13,18,27-33).

While the impact of point mutations within NF-κB and Sp1 sites have been extensively explored to show that these can dramatically alter gene expression and the stability of the active/inactive states of the virus (23,34-36), further work needs to be done to understand how other TFBS influence the decision between viral activation and silencing. In addition to sequence diversity within different TFBS, subtypes have also developed novel architectures of these TFBS (Fig. 5.1) (5,37). HIV-1 is phylogenetically classified into three major groups, M, N and O. The most prevalent group M is further classified into several subtypes and chimeric or recombinant forms of these subtypes (38). Most studies on HIV-1 gene expression, replication and latency have previously concentrated on subtype B, the clade most prevalent in the United States and Western Europe. Understanding how sequence diversity and unique TFBS architectures in the U3 region for different subtypes may alter recruitment of transcription factors is important to determining differences in viral pathogenesis and latency (39). Such knowledge will provide strategies to optimally combat HIV-1 latency for different subtypes (40-50).

In this work we show that subtypes with differences in TFBS sequence and architecture result in differences in steady-state gene expression levels and replication rates. These differences in gene expression were attributed to specific domains of the promoter. Further, we also studied the dynamics of gene expression and found that a combination of different TFBS may play a role in regulating these properties of the virus and thereby potentially alter the rates of reactivation from latent viruses. Thus based on the data presented in this chapter and experiments currently being pursued in primary cell culture models, we show that subtype promoters can lead to dramatically different levels of gene expression, viral replication and latency, thereby giving rise to subtypes with varying virulence and pathogenicity.

5.2 HIV-1 promoters from different subtypes produce virus with widely varying replication rates

To initially assess if subtype promoters, with differences in the sequence and arrangement of TFBS, result in phenotypic differences, we created full-length virus containing U3 regions from a limited number of different subtypes with the remaining viral genome corresponding to subtype B. This allowed us to test how subtype promoters may impact different properties of the virus. The U3 region from six subtypes, B, A, A2, C*, B/C and D were cloned into the sLTR vector (51) and virus packaged from these different viral genomes were used to infect a T-cell line, SupT1 at an MOI of 0.0005. Infectious viral titers were subsequently estimated using an indicator cell-line expressing GFP, as described previously. Viral titers were estimated every two days over a course of 10 days.

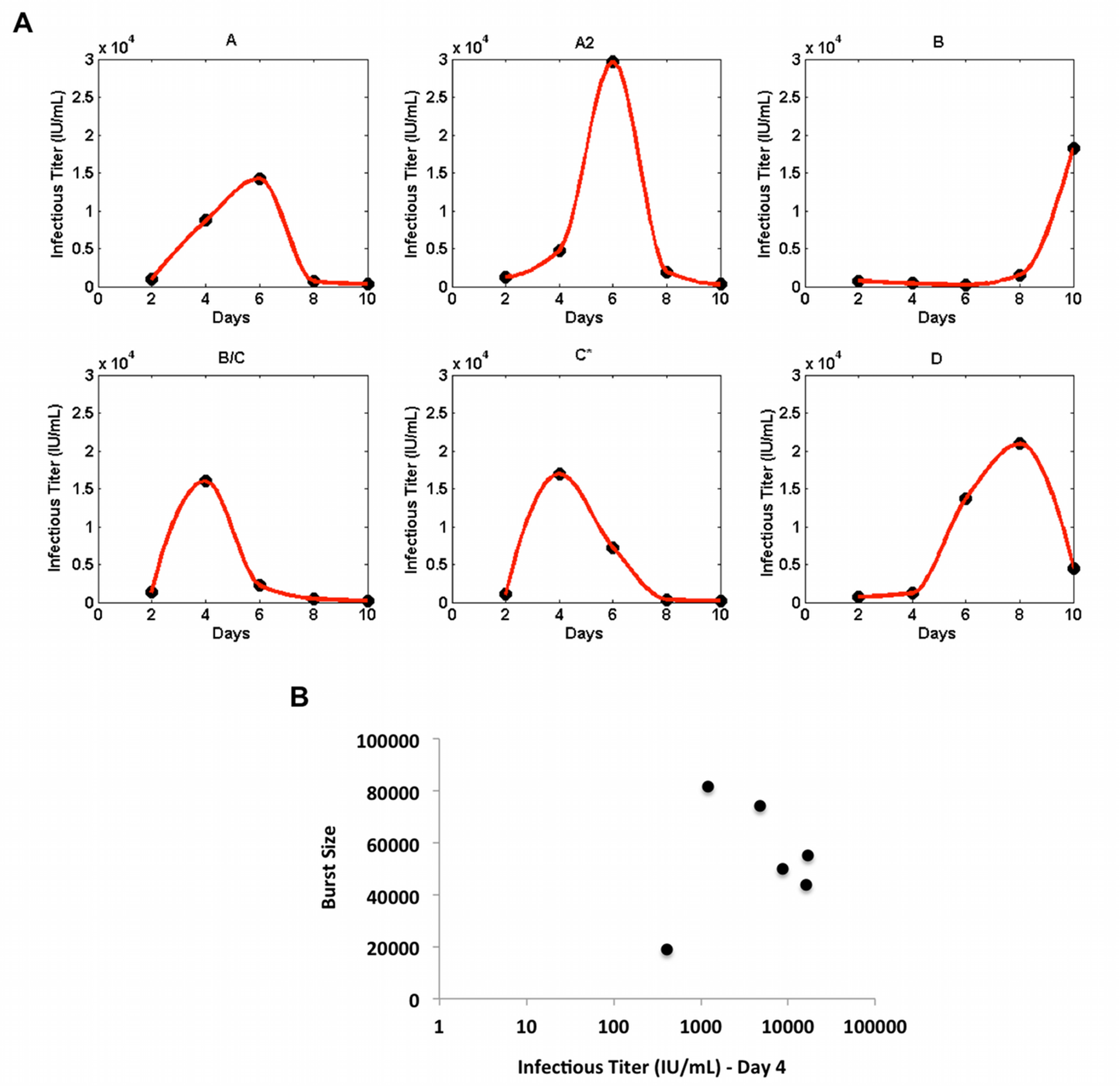


Figure 5.3. Burst sizes in virus production differ between subtype promoters. (A) The experimentally determined infectious titers of Figure 2A are shown with black dots. Burst sizes for each subtype promoter was estimated by calculating the area under each curve obtained by the piecewise cubic Hermite interpolation method. (B) Burst sizes vary over 4-fold between subtypes. Data shows that promoters that result in rapid viral replication produce smaller burst sizes, possibly as a result of increased T-cell death initially that leaves fewer cells for the virus to replicate in.

We found that different promoter stains produced widely variable replication rates. Certain subtypes, such as C* and B/C, show rapid viral production, peaking around day 4 followed by reduced viral titers over time, possibly due to cell death (Fig. 5.2*A*). Other subtypes, such as A and A2, displayed slightly slower replication with peak viral titers achieved around day 6 (Fig. 5.2*A*). At the other extreme, subtypes such as B and D showed much slower replication kinetics with peak viral titers achieved around day 8 or 10 (Fig. 5.2*A*).

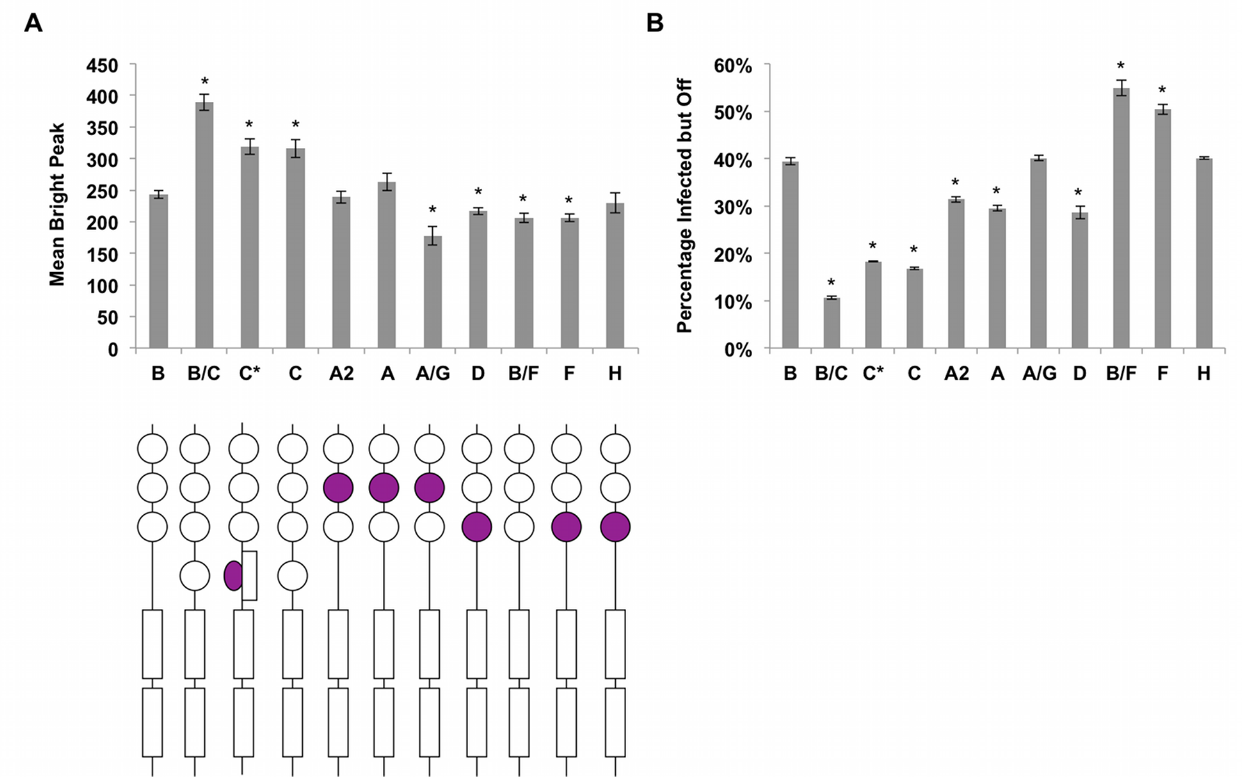


Figure 5.4. Gene expression levels at steady state vary dramatically between subtype promoters. (A) Figure shows the levels of the Mean Bright Peak for Jurkat cells infected with LGIT vectors containing different subtype promoters. The lower half of the figure shows the NF-κB/Sp1 sites for the different subtypes. As in Figure 2B, purple ovals indicate mutations within Sp1 sites with potential weakening. Subtypes such as B/C, C and C*, with an extra NF-κB/Sp1 site show increased gene expression levels compared to subtype B. In contrast, certain subtypes with mutations within Sp1 sites show lower gene expression levels compared to subtype B. (B) In agreement with data in Figure 4A, subtypes that show strong gene expression have fewer Percentage Infected but Off cells whereas those that show weaker expression result in more silenced cells. Experiments were performed in biological triplicate and error bars indicate standard deviations from the mean. '*' indicates statistically significant differences from subtype B (p<0.05).

Analyzing viral titers at day 4 post-infection provides a measure of the variation in the initial replication kinetics of the viruses before any substantial cell death. As observed with the time-course data, we found that subtypes C* and B/C show rapid replication, followed by subtypes A and A2 with subtypes B and D showing much slower viral replication rates (Fig. 5.2*B*). We found a 30-fold difference in the initial viral replication rate between subtypes, suggesting that variations in the sequence and architecture of TFBS could dramatically alter viral replication. To understand how variations in the replication rate when working with a fixed number of SupT1 cells in culture may affect the total number of viral particles produced over the course of the experiment (which we called the Burst Size), we fit the viral titering data using a piecewise cubic

Hermite interpolation method and calculated the area under each curve (Fig. 5.3*A*). We found that subtypes promoters that produce virus rapidly give rise to smaller Burst Sizes, possibly due to rapid T-cell death initially leaving fewer cells to replicate in over time (Fig. 5.3*B*). These initial experiments suggested that subtype promoters could produce widely varying viral replication rates that may result in variations in disease progression in patients infected with different subtypes. Further, we hypothesized that these differences in replication rate may arise from differences in the sequence of the promoter since subtypes showing rapid replication (C* and B/C) contain an extra NF-κB/Sp1 site within the core promoter that may result in stronger gene expression. In contract, some of the other subtypes, such as A, A2 or D contain a mutation within Sp1 site II or Site III that may be responsible for weaker replication rates. Thus, we decided to explore and characterize this potential link between the viral promoter genotype and phenotype more carefully.

5.3 HIV-1 subtype promoters show differences in gene expression

Since the 6 subtype promoters tested in the previous section showed differences in viral replication rates, we hypothesized that these differences may arise from variations in the level of gene expression that is a function of the intrinsic strength of the promoter resulting in differential recruitment of transcription factors and transcriptional initiation. To test this hypothesis, we decided to study a large set of 11 subtype promoters. To study if variations in the architecture and sequence of TFBS can result in differences in gene expression, we used the minimal positive-feedback LGIT (LTR-GFP-IRES-Tat) system previously described in Chapter 4. We used the metrics, Mean Bright Peak and Percentage Infected but Off, previously described in Section 4.3 to quantify gene expression from the viral promoter. U3 regions from different subtypes were cloned into LGIT (with the remainder of the lentiviral sequence corresponding to subtype B) and used to infect Jurkat T-cells at low MOI to ensure single integration events per cell.

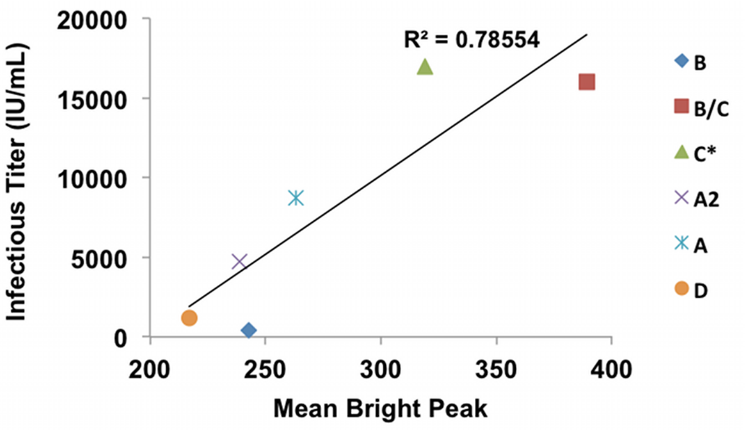


Figure 5.5. Viral replication rates for subtype promoters are strongly correlated to gene expression levels. The initial infectious titer obtained for different subtype promoters correlate well with the Mean Bright Peak levels indicating that differences in viral replication rates may be driven by corresponding variations in the levels of gene expression. These differences in gene expression may arise from differential recruitment of transcription factors to subtype promoters that result in characteristic strengths for these promoters.

Once steady state is attained, we found large differences in the level of gene expression (Fig. 5.4). When compared to subtype B, certain subtypes such as C, C* and B/C, each with an extra NF-κB/Sp1 site, showed statistically significant increases in the level of Mean Bright Peak. In comparison, other subtypes such as A/G, D, B/F and F, some with mutations within Sp1 sites that were predicted to weaken gene expression, showed lower Mean Bright Peak levels than subtype B (Fig. 5.4*A*). In support of this data, subtypes that showed strong gene expression, such as C, C* and B/C, had 2-3 fold fewer inactive cells than subtype B. Similarly, subtypes that showed weaker gene activation had higher fractions of Percentage Infected but Off cells, with subtypes such as B/F and F having half of the infected cells in the inactive state (Fig. 5.4*B*). Finally the five-fold difference in the Percentage Infected but Off between subtypes having the lowest and highest fraction of inactive cells showed that subtypes have large differences in gene expression that may give rise to different propensities for latency (Fig. 5.4*B*). Such differences in the establishment and reactivation from latency may imply that patients from different parts of the world would require therapy regimens that are tailored towards the infecting subtype.

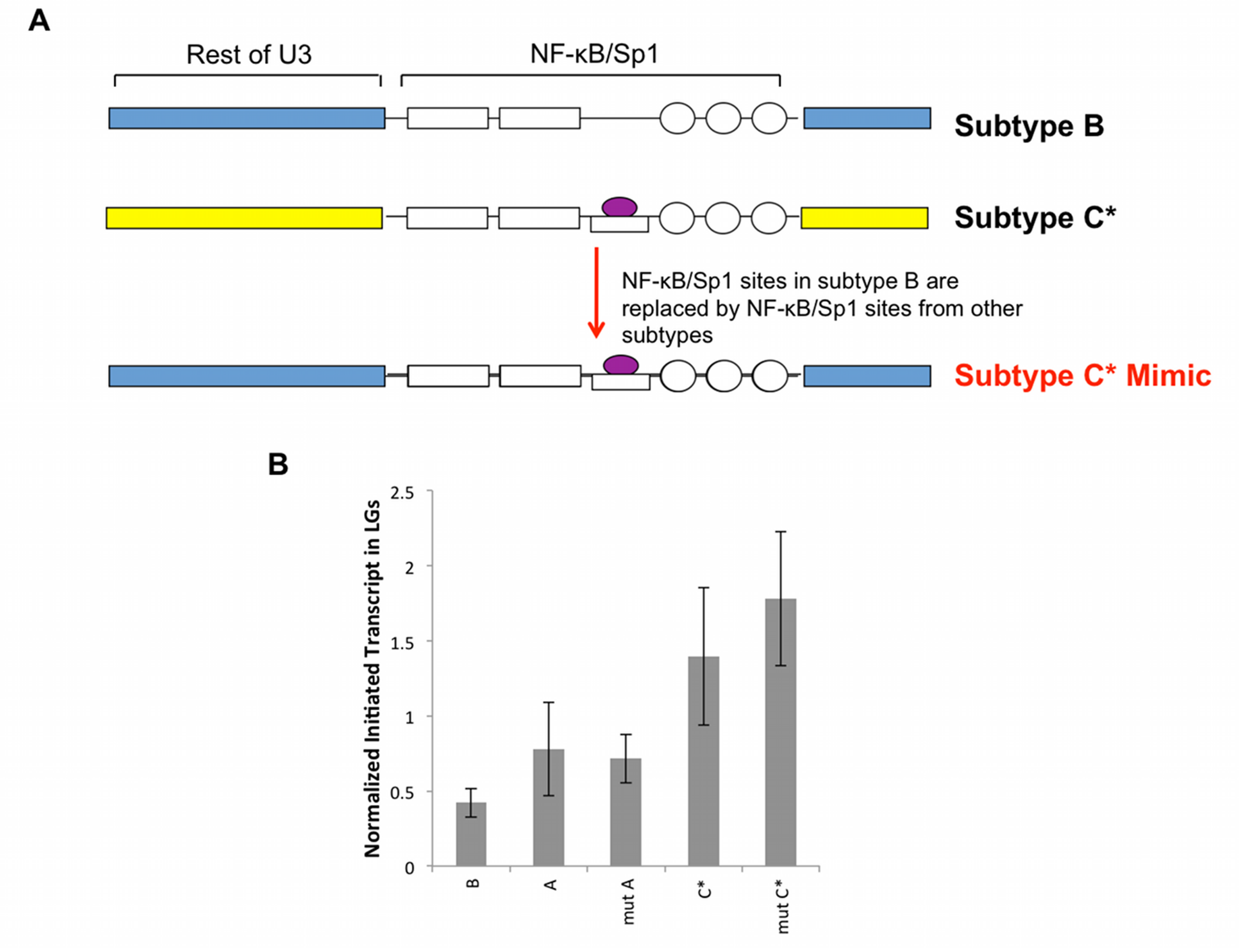


Figure 5.6. NF-κB/Sp1 sites determine the intrinsic strength of the promoter at steady state. (A) Schematic of the chimeric promoters created to study the impact of NF-κB/Sp1 sites on gene expression. The NF-κB/Sp1 sites in the subtype B promoter were swapped with the corresponding sites from another subtype promoter to generate a chimeric promoter that consisted of NF-κB/Sp1 sites from that subtype while the rest of the promoter consisted of subtype B. (B) To test the intrinsic strength (without the positive-feedback loop induced by Tat) of naturally

occurring and mimicking promoters, Jurkat cells were infected with LGs containing different promoters and total cellular RNA was extracted to quantify initiated viral transcripts. Data shows that the mimicking promoters track the naturally occurring promoters closely, suggesting that the NF-κB/Sp1 sites play an important role in setting the level of gene expression at steady state. qPCR was performed in triplicate on the BioRad iQ5 machine. Error bars indicate standard deviation from the mean.

Finally, we found a strong correlation between the infectious titers in the full-length virus containing variable U3 regions with the Mean Bright Peak levels of LGITs (Fig. 5.5). This strong correlation suggests that the differences in the replication rate of different subtype promoters may arise from variations in the level of gene expression due to the differential recruitment of transcription factors to the viral promoter (Fig. 5.5). Finally, this correlation also shows that the minimal LGIT system that mimics the positive-feedback loop in HIV-1 is a good model to study viral gene expression.

5.4 Identifying the minimal set of TFBS in the HIV-1 promoter that contribute to most of the observed differences in viral gene expression

Previous studies on the HIV-1 promoter have shown that the NF-κB/Sp1 sites play an important role in regulating gene expression (23,34-36). Further, our empirical observation that subtypes with an extra NF-κB/Sp1 site within the promoter showed stronger gene expression and faster viral replication while those with weakening mutations within the Sp1 sites showed weaker gene expression and slower replication kinetics, suggesting that the NF-κB/Sp1 sites may play a critical role in setting the steady state level of gene expression. To explore this hypothesis further, we replaced the NF-κB/Sp1 domain in subtype B with that of other subtypes (Fig. 5.6*A*). Thus, when we replaced the NF-κB/Sp1 domain in subtype B with subtype C*, we called the new promoter architecture as subtype C* Mimic (Fig. 5.6*A*). Thus, if our hypothesis that the NF-κB/Sp1 sites make the most important contribution to steady state gene expression were true, then the levels of gene expression for each mimicking subtype promoter should resemble that of the actual subtype promoter.

In another experiment, to test if the NF-κB/Sp1 domain is critical in setting the level of gene expression, we studied the extent to which the basal expression level at steady state, in the absence of Tat, depends on the architecture of the NF-κB/Sp1 sites. We cloned in subtype A, C* and their corresponding mimics into the LG lentiviral vector (Section 4.2). Jurkat cells were then infected with these LG constructs at low MOIs (~0.05-0.1) and stimulated with TNFα 7 days post-infection to sort GFP+ cells. These cells were allowed to relax which was followed by total RNA extraction. The levels of initiated viral transcripts were then quantified using RT-qPCR to access the contribution of the NF-κB/Sp1 domain to basal gene expression. Interestingly, while both subtype A and C* had different levels of initiated viral transcripts, their corresponding mimicking variants showed similar levels of initiated viral transcripts as the original subtype promoter that were significantly different from subtype B (Fig. 5.6*B*). This suggested that the NF-κB/Sp1 domain plays the most important role in setting the basal level of transcription from the HIV-1 promoter (Fig. 5.6*B*).

To access more generally if the architecture and sequence of the NF-κB/Sp1 sites make the most important contribution to viral gene expression, we made mimicking variants for 6 subtypes in the LGIT lentiviral system and infected Jurkat cells with the original and mimicking promoter variants. Interestingly, we found that the Mean Bright Peak levels for the mimicking variants tracked the original subtype very closely, suggesting that the NF-κB/Sp1 domain plays an critical role in setting the level of gene expression at steady state (Fig. 5.7*A*). Similarly, the Percentage Infected but Offs were also similar for the original and mimicking subtypes (Fig. 5.7*B*). These experiments showed that the NF-κB/Sp1 sites play the most important role in regulating steady state levels of gene expression and allowed us to identify the minimal set of TFBS that contribute to most of the observed differences in viral gene expression between subtypes.

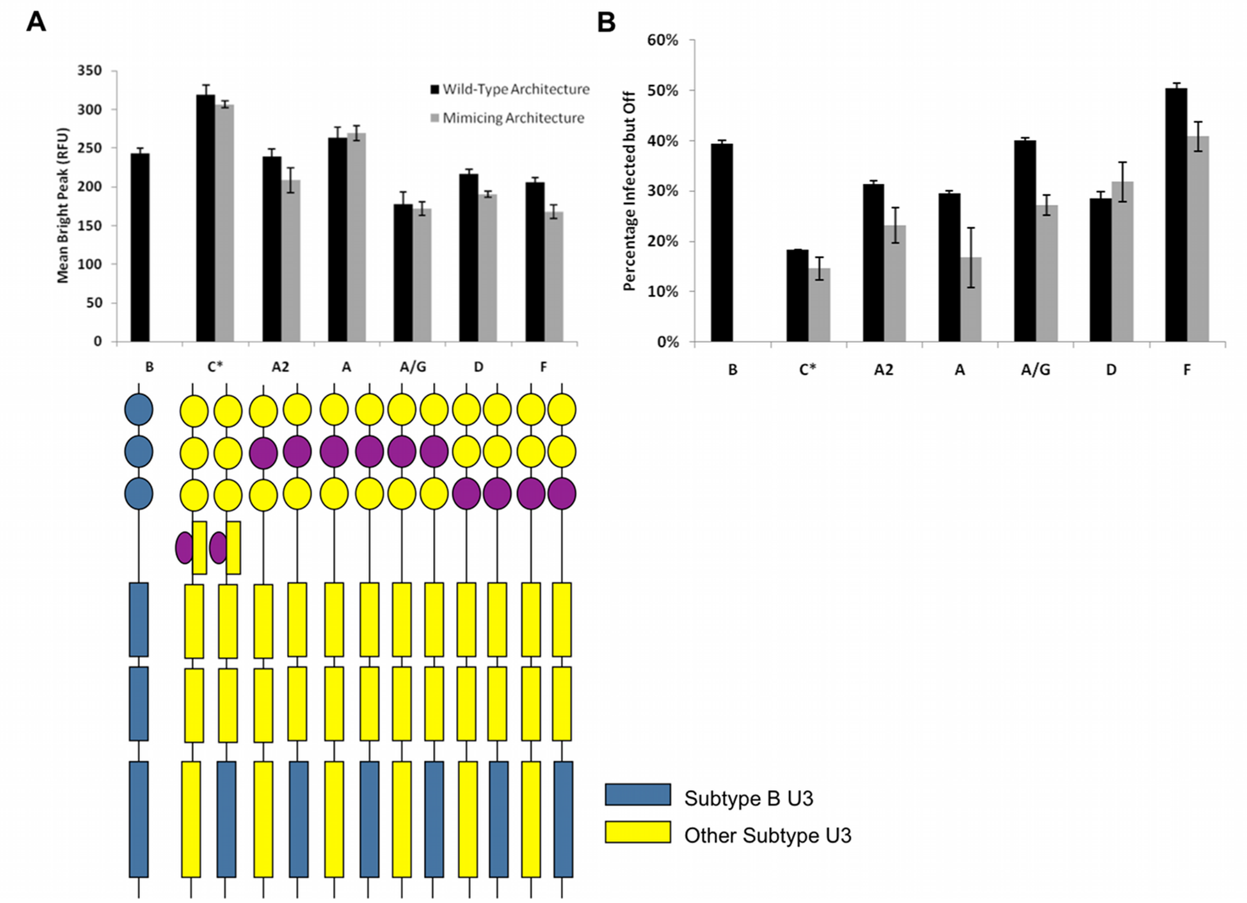


Figure 5.7. NF-κB/Sp1 sites determine the level of gene expression from the HIV-1 promoter at steady state. (A) Mean Bright Peak levels and (B) Percentage Infected but Off were determined for Jurkat cells infected with LGITs containing promoters from naturally occurring subtypes or their corresponding NF-κB/Sp1 mimicking variants. Lower half of the figure shows schematically the naturally occurring and chimeric promoters. Blue boxes and circles indicate sequence and TFBS corresponding to subtype B whereas yellow boxes and circles stand for sequence and TFBS corresponding to other subtypes. Purple circles stand for Sp1 sites with potentially weakening effects. Bar charts show that the gene expression levels of the mimicking promoters track the wild-type promoters very closely. Similarly, the fraction of silenced cells are similar for wild-type and their corresponding mimicking promoters. Experiments were performed in triplicate and error bars indicate standard deviations from the mean.

5.5 NF-κB/Sp1 sites and other TFBS in the HIV-1 promoter regulate viral gene expression dynamics

From the previous section, we established that the sequence and architecture of the NF-κB/Sp1 sites determine the steady state gene expression levels from the viral promoter. We next wanted to investigate how sequence diversity within the viral promoter may regulate gene expression dynamics from inactive (Off) and active (Bright) LGIT infected Jurkat populations.

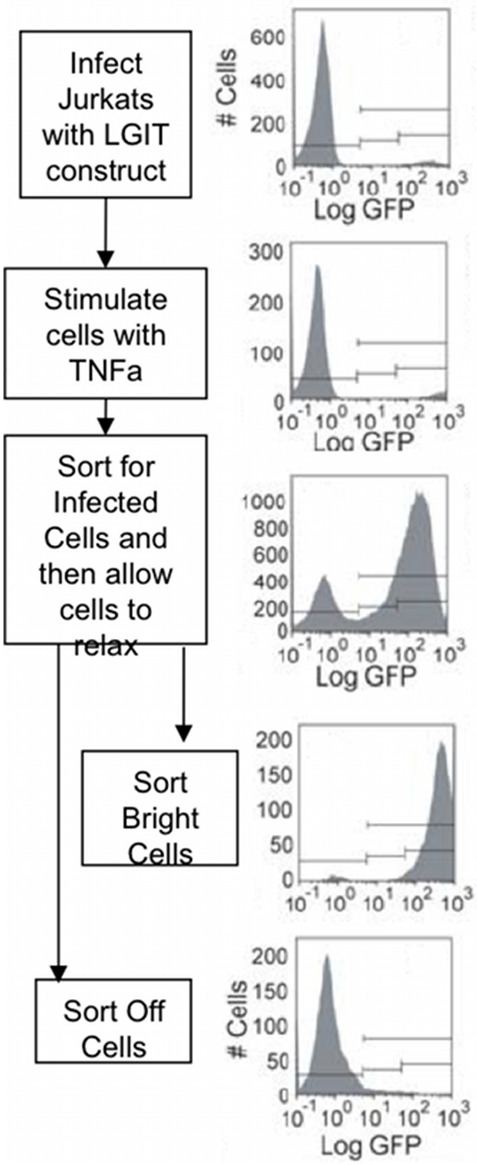


Figure 5.8. Sorting scheme to isolate inactive and active cell populations to study viral gene expression dynamics. Jurkat cells were infected with LGITs containing different subtype promoters at low MOIs (0.05-0.1) to ensure single integration events per cell. 7 days post infection the cells are stimulated with TNFα and sorted to isolate infected GFP+ cells. These sorted cells are allowed to relax for 1-2 weeks. Inactive (GFP-) and Active (stongly GFP+) cells are sorted from this population and the dynamics of gene activation/inactivation for these two population of cells are tracked over time. Adapted from (23).

The dynamics of gene activation from an Off population or gene inactivation from the Bright population allow us to estimate and predict the rate of establishment and reactivation of latent populations in patients. To study if promoter diversity alters gene activation or inactivation rates, we infected Jurkat cells with LGITs containing subtype B promoter or one of the six other subtype promoters and their corresponding mimics (used in the previous section). The sorting scheme to isolate inactive (Off) and active (Bright) cell populations in each case are shown in Figure 5.8. Infected Jurkat cells were stimulated with TNFα 7 days post infection and GFP+ cells were sorted by Fluorescent Activated Cell Sorting (FACS). These sorted populations were allowed to relax and GFP- (Off) and strongly activated cells (Bright) were sorted for each U3 subtype (Fig. 5.8). The dynamics of gene activation from the inactive (Off) population and gene inactivation from the active (Bright) population were then monitored over time by measuring GFP expression using flow cytometry.

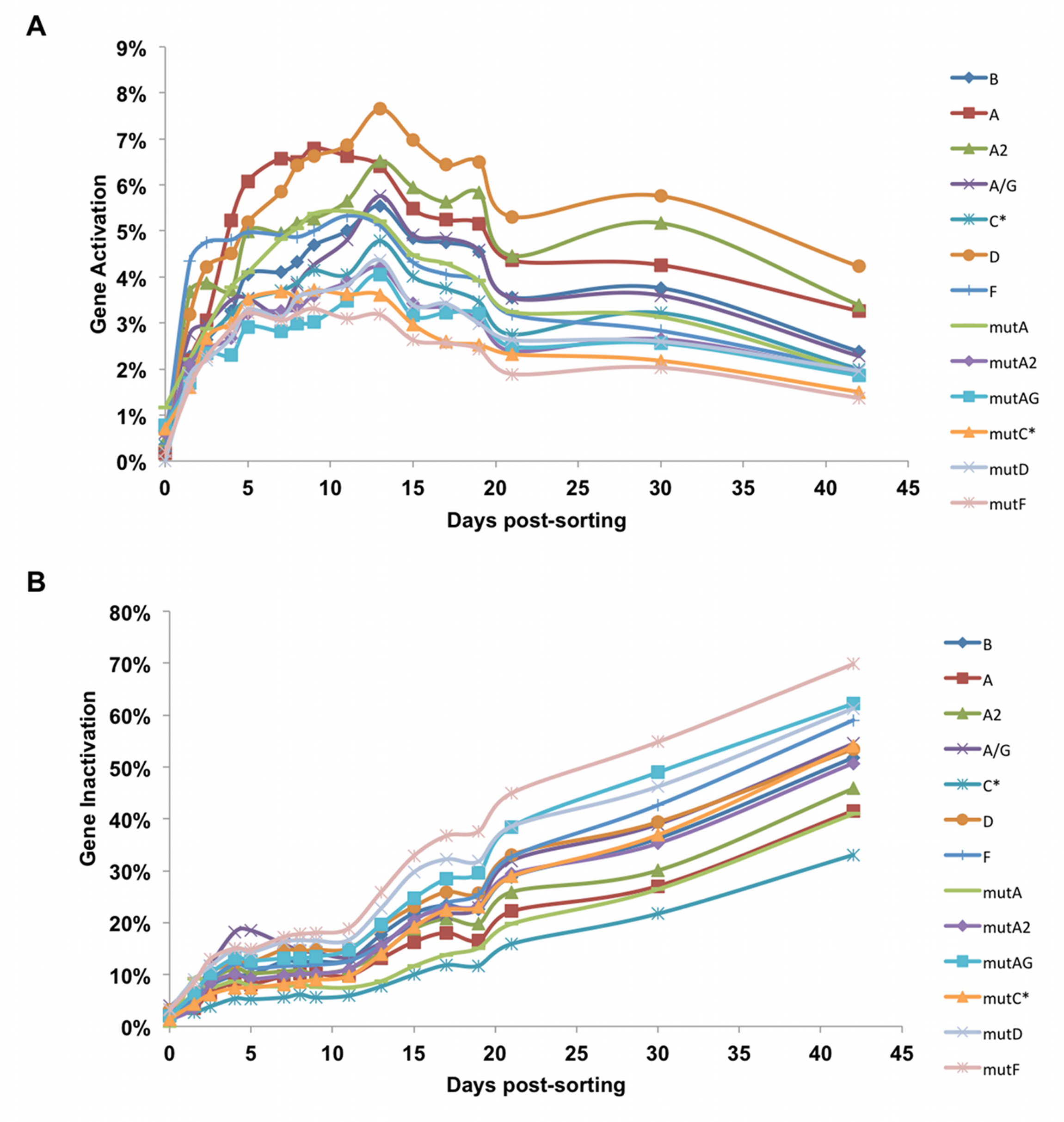


Figure 5.9. Subtype promoters show differential rates of gene activation and inactivation. (A) Time-course data of the percentage of cells that activate from the Off state over time. The rate of gene activation is estimated by quantifying GFP+ cells using flow cytometry. Subtypes show approximately 2-fold variations in the rate of gene activation. These differences in gene activation suggests that subtype promoters may lead to differences in the rate of reactivation from latent viral populations in patients. (B) Time-course data of the percentage of cells that relax from the Bright state over time. The rate of gene inactivation is estimated by quantifying GFP+ cells using flow cytometry. Similar to gene activation, the rate of gene inactivation from the Bright state also shows 2-3 fold differences between subtype promoters.

Gene activation and inactivation were measured for all 13 subtype variants over 42 days (Fig. 5.9). We found large differences in both the rate of activation and inactivation. Subtypes showed a 2 fold-variation in the rate of activation from the inactive state and a 2-3 fold variation in the levels of gene inactivation from the active state (Fig. 5.9). We analyzed the gene activation levels around day 13 post-sorting when the gene activation levels reached a maxima (Figs. 5.9*A* and 5.10*A*). Similarly, data for gene inactivation from the active state is plotted in Figure 5.10*B* around the mid time point of our analysis, day 17 post-sorting. As expected based on the time-course data, U3 subtypes showed large variations in the level of gene activation and inactivation.

In studying the levels of gene inactivation, we found that subtype C* had the lowest levels of gene inactivation compared to the other subtypes (Fig. 5.10*B*), suggesting that the extra NF-κB/Sp1 sites increase the stability of the active state. In comparison, we found that the mimicking variants for all subtypes, expect subtype A, showed higher rates of gene inactivation suggesting that the chimeric promoters tend to inactive the active state and that regions outside NF-κB/Sp1 may play a role in regulating the dynamics of gene expression (Fig. 5.10*B*).

Within naturally occurring subtypes, C* showed the lowest whereas D showed the highest levels of gene activation from the inactive state, with all the other subtypes showing gene activation levels between this 2-fold variation between the two extreme subtypes (Fig. 5.10*A*). This 2-fold variation in gene activation levels between subtypes may be clinically important since it may imply that subtype promoters may have different propensities for reactivation from latency. Under these conditions of latency, where levels of the viral protein Tat are extremely low or absent, sequence diversity and TFBS variations between subtype promoters may lead to differential recruitment of transcription factors to the promoter resulting in variations in the rate and level of reactivation from latency. Surprisingly, we also found that all the mimicking architectures displayed gene activation levels lower than their corresponding natural variants (Fig. 5.10*A*). Since a natural and its corresponding mimicking promoter variant shares the same NF-κB/Sp1 sites, this suggested that the region of the promoter outside the NF-κB/Sp1 sites (henceforth referred to as the Upstream promoter elements) for subtype B resulted in lower gene activation levels than upstream promoter elements for all other subtypes (such as A, A2, A/G, C*, D and F). Importantly, this also suggested that promoter elements outside the NF-κB/Sp1 sites may play an important in regulating gene expression dynamics from the viral promoter.

To better understand how upstream promoter elements may be influencing the dynamics of gene activation/inactivation, we plotted the activation rate of mimicking promoter variants vs. the activation rate of the naturally occurring variant (and similarly for the gene inactivation rates) (Fig. 5.11). In both cases, we found positive correlations between the mimicking and naturally occurring subtypes. Since a particular pair of mimicking and naturally occurring subtypes share the same NF-κB/Sp1 sites, this suggested that a particular NF-κB/Sp1 configuration made a contribution towards the dynamics of gene expression (Fig. 5.11). However, when these correlations were compared to the hypothetical diagonal line constructed in these figures, since the level of gene activation for the mimics were lower than the corresponding naturally occurring variants, this suggested that these differences might arise from differences in the upstream promoter elements (Fig. 5.11*A*). Similarly, deviations from the diagonal for the gene inactivation levels suggested that upstream promoter elements play a role in regulating this phenotype (Fig. 5.11*B*).

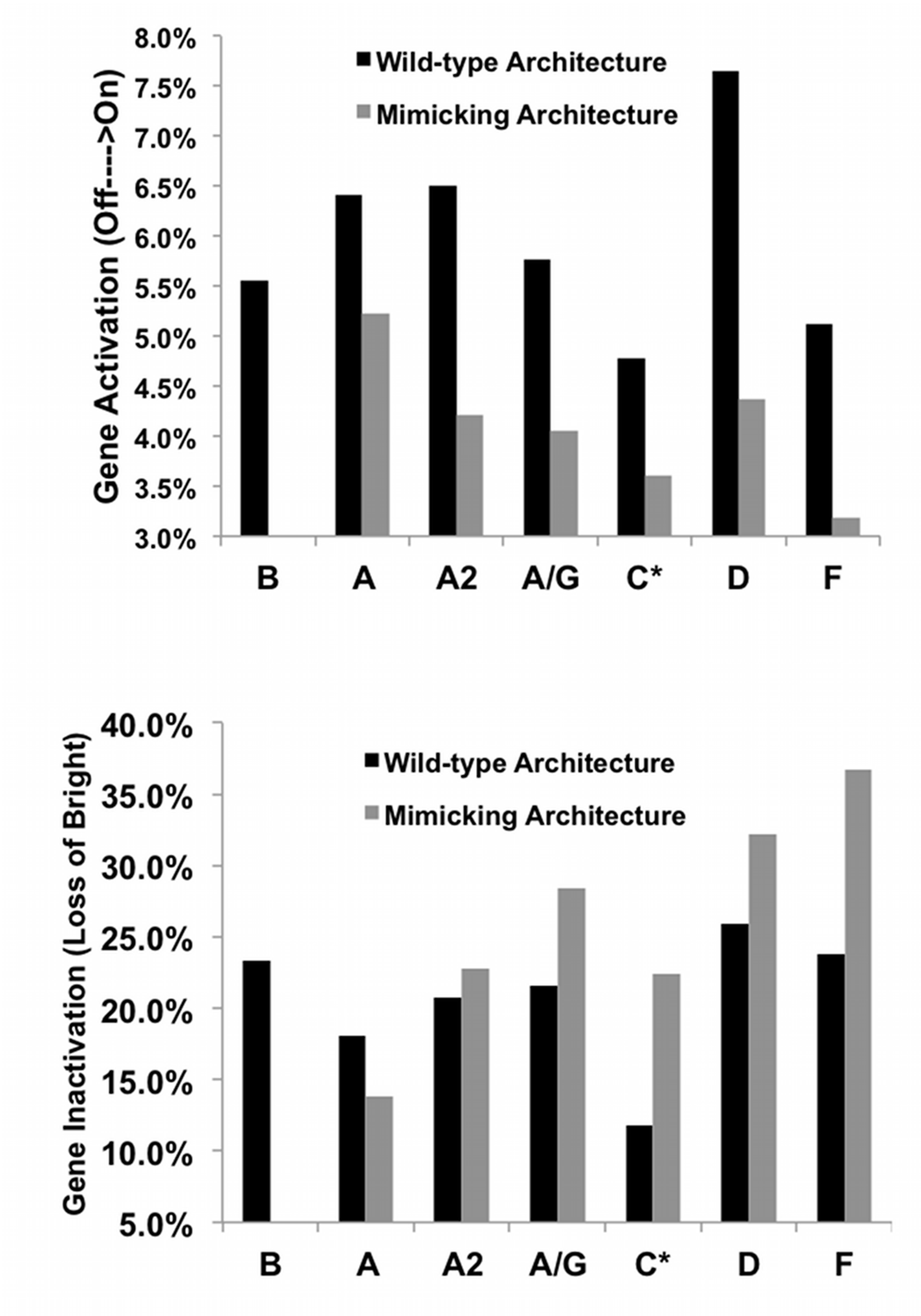


Figure 5.10. Levels of gene activation and inactivation vary with the subtype promoter. (A) Bar charts quantify the level of gene activation at day 13 post-sorting. Data shows that gene activation varies 2-fold over the naturally occurring subtypes. The large differences in the levels of gene activation between the naturally occurring and mimicking promoters suggest that regions outside the NF-κB/Sp1 sites may play an important role in regulating gene expression dynamics from the HIV-1 promoter. (B) Bar charts quantify the levels of gene inactivation from the active state at day 17 post-sorting. Gene inactivation rates vary 2-fold over wild-type promoters and as observed in Figure 10A, differences in gene inactivation between the wild-type and mimicking architecture point to the role of other elements besides NF-κB/Sp1 sites in regulating gene expression dynamics.

To further understand how subtype promoters influence gene expression dynamics, we analyzed the frequency of clones that show phenotypic bifurcation (PheB). It has previously been shown that the integration of LGIT into certain genomic locations give rise to clonal populations that have heterogeneous expression of GFP, with a fraction of cells having active expression whereas others showing no gene expression (10). Within the bulk sort, these PheB clones tend to be enriched within the Mid GFP expression range. It has been proposed that such PheB clones could stochastically alter between active and inactive states and thereby give rise to latent viral populations within patients that could reactivate at later time points to repopulate the actively replicating pool of virus. Therefore, analyzing if subtypes result in different frequencies of PheB clones could imply differences in the establishment and reactivation from latency.

To analyze potential differences in PheB, we infected Jurkats with LGIT from different subtypes at low MOIs and sorted single cells into 96-well plates from the Mid GFP region 7 days post-infection. These single cells were expanded into clonal populations and the GFP distribution of 80-100 clones were analyzed by flow cytometry for each subtype variant 20-30 days post-sorting. The fraction of clones for each subtype that displayed bimodal or PheB expression were then estimated using a statistical test called as the Hartigan's Dip Statistic (52). Hartigan's dip test allows us to estimate the statistical significance of the extent of unimodality of a histogram. Briefly, the dip statistic is the maximum difference over all data points between the distribution being tested and the unimodal distribution function that minimizes that maximum difference. The null distribution from which the dip statistic is determined is the normal distribution. Therefore, clones that showed statistically significant deviation from the unimodal distribution were scored as PheB clones.

We found a 4-fold variation in the frequency of PheB clones between subtypes (Fig. 5.12*A* and Table 5.1). Subtype promoters that previously showed strong gene expression, such as subtype C*, had low frequencies of PheB clones, possibly since the strong promoter biases such clones towards activation. In contrast, subtypes such as D and F, with weaker gene expression have higher frequencies of PheB clones, possibly due to their inability to activate the Tat mediated positive-feedback loop effectively, resulting in greater stochasticity (Fig. 5.12*A*). Finally, when we plotted the frequency of PheB clones for the mimicking variants vs. the naturally occurring subtypes, we found a positive correlation between these quantities suggesting that the NF-κB/Sp1 sites play a role in regulating gene expression dynamics (Fig. 5.12*B*). Further, as previously noted for the gene activation/inactivation levels, the data points deviate from the diagonal, suggesting that elements outside the NF-κB/Sp1 domain play an important role in regulating gene expression dynamics (Fig. 5.12).

As a variety of metrics to quantify gene expression dynamics suggested that upstream promoter elements together with NF-κB/Sp1 sites regulate this property of the virus, we computed the contribution of each of these elements to expression dynamics for the 6 subtype variants studied.

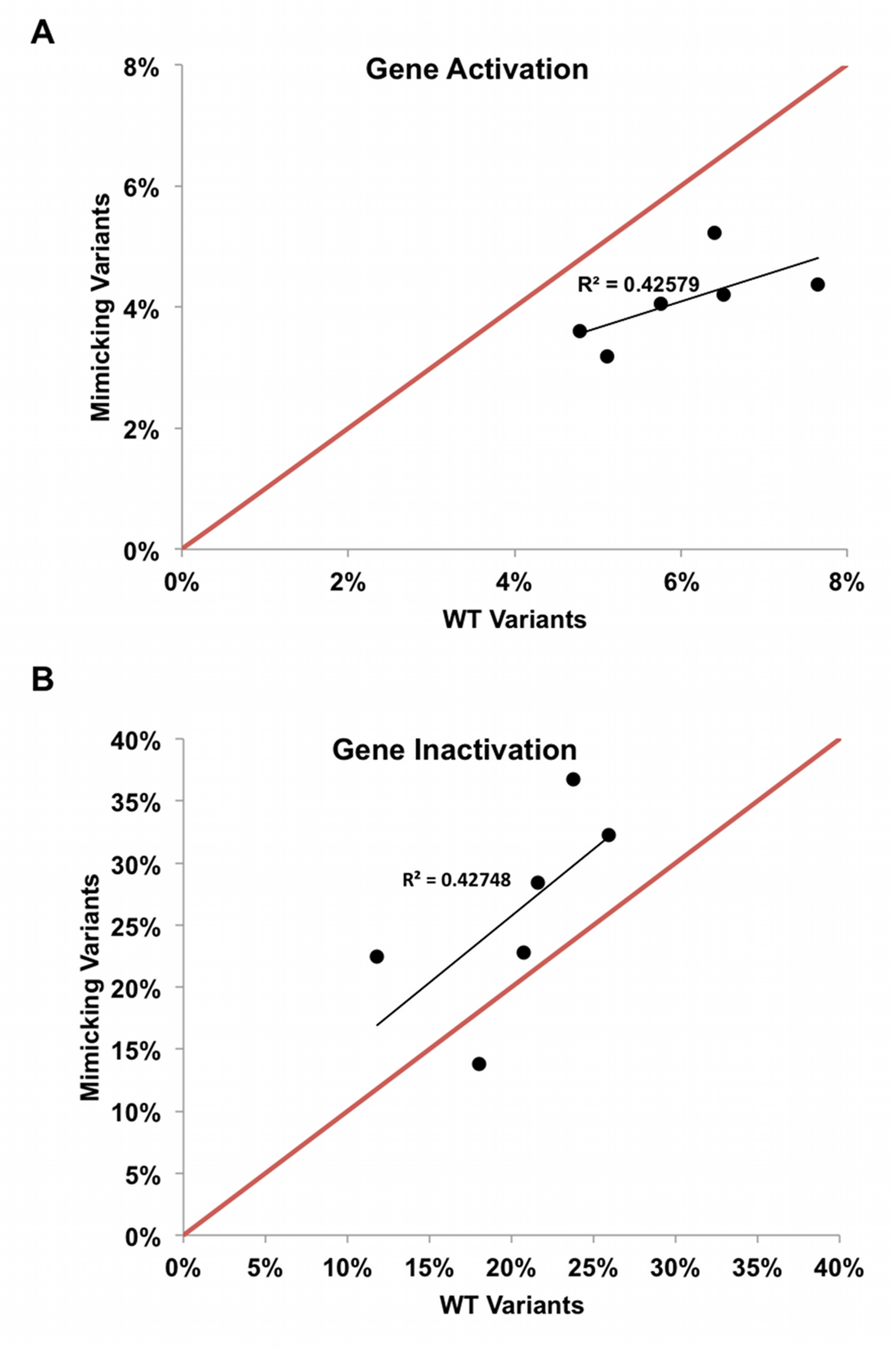


Figure 5.11. Upstream promoter elements together with NF-κB/Sp1 sites regulate gene expression dynamics from the HIV-1 promoter. Plots of (A) Gene activation and (B) gene inactivation for mimicking promoters vs.

wild-type promoters show moderate positive correlation. This suggests that changing the NF-κB/Sp1 sites produces similar changes in the gene activation/inactivation rate and therefore these sites contribute to gene expression dynamics. The red line shows a hypothetical diagonal line. If the NF-κB/Sp1 sites alone contributed to gene expression dynamics, then the data points should all lie close to the hypothetical diagonal. Since the data points either lie below (for gene activation) or above (for gene inactivation) the diagonal, this suggests that upstream promoter elements also contribute to gene expression dynamics from the HIV-1 promoter.

In estimating the contribution of the upstream promoter elements and NF-κB/Sp1 sites to gene activation levels, we made the simplifying assumption that these two modules within the promoter do not interact and thereby regulate gene activation independently. For all subtypes, except C*, we found that the upstream promoter elements contribute close to or greater than 50% of the gene activation levels (Fig. 5.13). The extra NF-κB/Sp1 site in subtype C* possibly skews the contribution towards the core promoter.

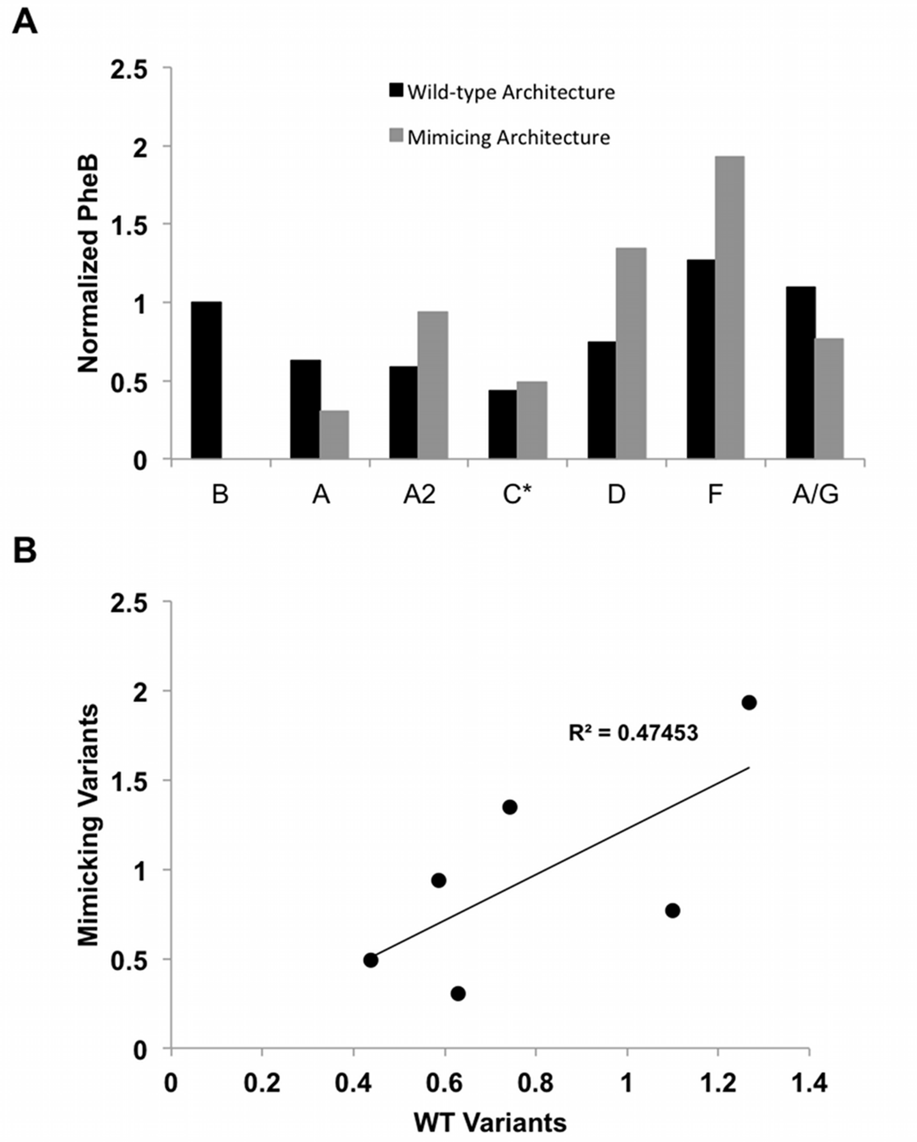


Figure 5.12. Upstream promoter elements and NF-κB/Sp1 sites both contribute to the fraction of clones that show stochastic switching between inactive and active states. (A) The fraction of clones that show stochastic switching between inactive and active states or phenotypic bifurcation (PheB) are shown for naturally occurring and mimicking promoter variants. Normalized PheB denotes the ratio of the fraction of clones that shows PheB with a particular promoter variant to that with subtype B. As with other dynamic properties of gene expression, promoter variants lead to differences in PheB. Differences in PheB between the wild-type and corresponding mimicking variants suggest that elements outside NF-κB/Sp1 sites contribute to PheB. (B) The modest positive correlation in the levels of normalized PheB between mimicking vs. wild-type promoters suggest that NF-κB/Sp1 sites also partly regulate this property. Thus, both upstream promoter elements and NF-κB/Sp1 sites contribute to PheB.

Table 5.1. The Hartigan's Dip Statistic is used to estimate the fraction of clones exhibiting PheB. For all the promoter variants tested, the fraction of clones exhibiting bimodal distribution within the Mid region are shown. This is used to estimate the fraction of clones exhibiting PheB in the entire population.

Subtype	Bimodal Clones	% of Bimodal Clones	% Mid Population	% PheB
B	9/121	7.438	1.36	0.101
D	3/40	7.5	1.00	0.075
mut D	7/79	8.861	1.54	0.136
C*	6/88	6.818	0.65	0.0443
mutC*	5/81	6.173	0.8	0.0494
A2	4/96	4.167	1.42	0.0592
mutA2	5/78	6.410	1.48	0.0949
A	4/85	4.706	1.33	0.0636
mutA	3/84	3.571	0.86	0.0307
A/G	7/88	7.955	1.4	0.111
mutA/G	2/42	4.762	1.63	0.0776
F	9/89	10.112	1.27	0.128
mutF	10/84	11.905	1.64	0.195
RBE	12/93	12.903	1.67	0.215

Thus, in contrast to the steady-state levels of gene expression that was primarily determined by the sequence and architecture of the NF-κB/Sp1 sites, gene expression dynamics was a function of both the NF-κB/Sp1 sites and upstream promoter elements.

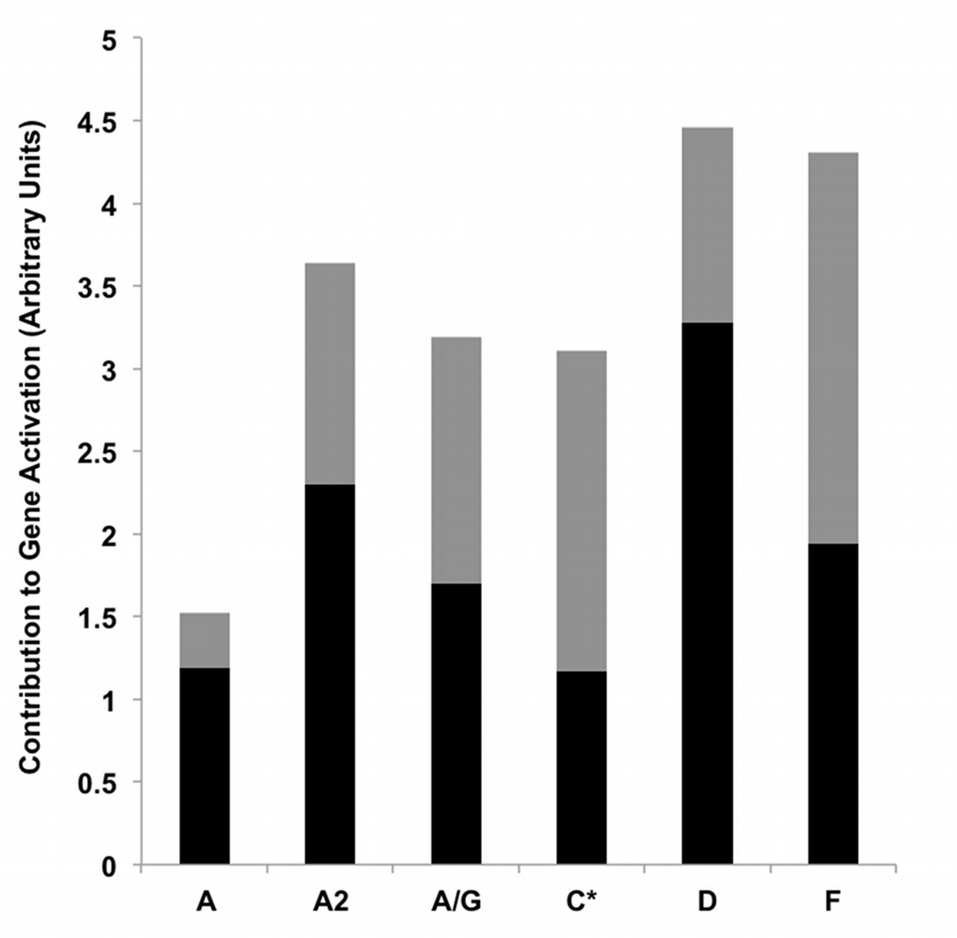
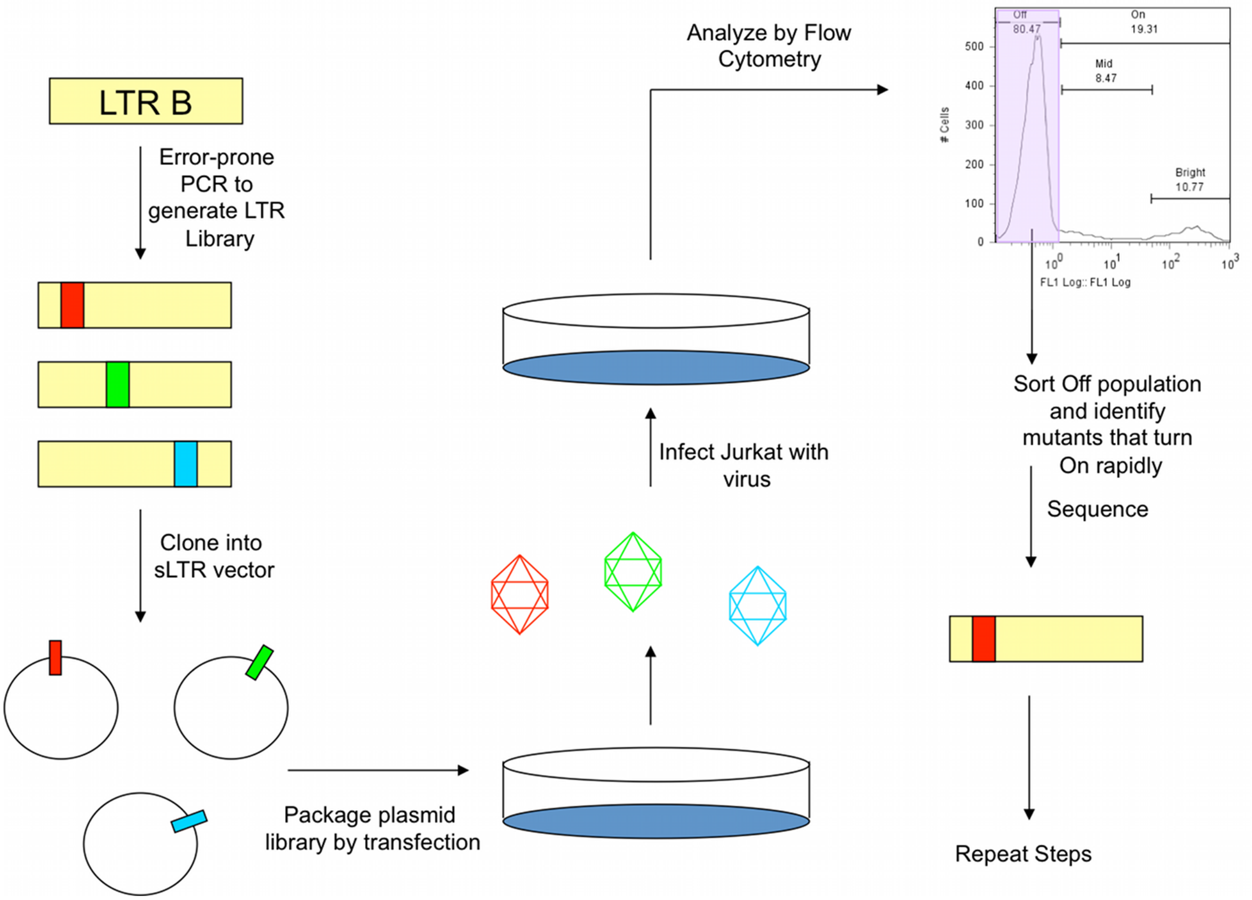


Figure 5.13. Contribution of upstream promoter elements and NF-κB/Sp1 sites to gene activation for different subtypes. Black bars correspond to the contribution of upstream promoter elements and gray bars correspond to the contribution of NF-κB/Sp1 sites to gene activation. Bar chart shows that for all subtypes, except C*, the upstream promoter elements contribute to more than 50% of the observed level of gene activation for a subtype. The increased contribution of NF-κB/Sp1 sites to gene activation in the case of subtype C* may arise due to the presence of an extra NF-κB/Sp1 site. The bar charts are estimated using the assumption that the upstream promoter elements and NF-κB/Sp1 sites act independently to regulate gene activation from the inactive state.

5.6 Identifying TFBS within upstream promoter elements that regulate gene expression dynamics

HIV-1 contains several TFBS within the upstream promoter region and therefore rational site-directed mutagenesis within a single or combination of these TFBS to identify sites that contribute to gene expression dynamics is not feasible. We therefore mined data collected by other members of the Schaffer group (unpublished data) to predict sites that may be important in regulating viral gene expression dynamics. In these experiments, a directed evolution approach was used to identify mutations within the viral promoter that increase the rate of gene activation from the inactive state (Fig. 5.14). Error-prone PCR of the subtype B promoter was initially used to create a large library of promoter variants. These were cloned into the sLTR lentiviral vector and packaged to create a library of viral particles. Jurkat cells were infected with this viral library and promoter variants that activated rapidly from the inactive state were isolated, sequenced and the process was repeated. After 3 enrichment cycles, clones that were found to increase the rate of gene activation from the inactive state were sequenced. As a negative control, clones that were not selected for were also sequenced.



<u>Figure 5.14. Directed evolution scheme to identify sites within the HIV-1 promoter that increase the rate of gene activation from the inactive state.</u> Random point mutations are introduced into the LTR of subtype B. This promoter library is cloned into the sLTR vector and virus packaged from this library is used to infect Jurkat cells. After cell sorting, the infected cells in the inactive state that active rapidly are isolated and the viral genome sequenced and subjected to further rounds of selection. This selection scheme should help enrich for sites within the promoter that increase the rate of gene activation.

The Fisher Exact test was used to estimate sites within the LTR that appeared to be have been selected to increase the rate of gene activation from the viral promoter. 18 such sites appeared to be selected for (p<0.05), off which 11 sites were within upstream promoter elements (Fig. 5.15). It has previously been shown that mutations within Sp1 site III makes the viral promoter more stochastic and decreases the stability of the inactive state thereby increasing the rate of gene activation. The selected library identified 4 such sites within Sp1 site III validating previously obtained results as well as providing guidance to identify potentially interesting TFBS within the upstream promoter region that may increase the rate of gene activation.

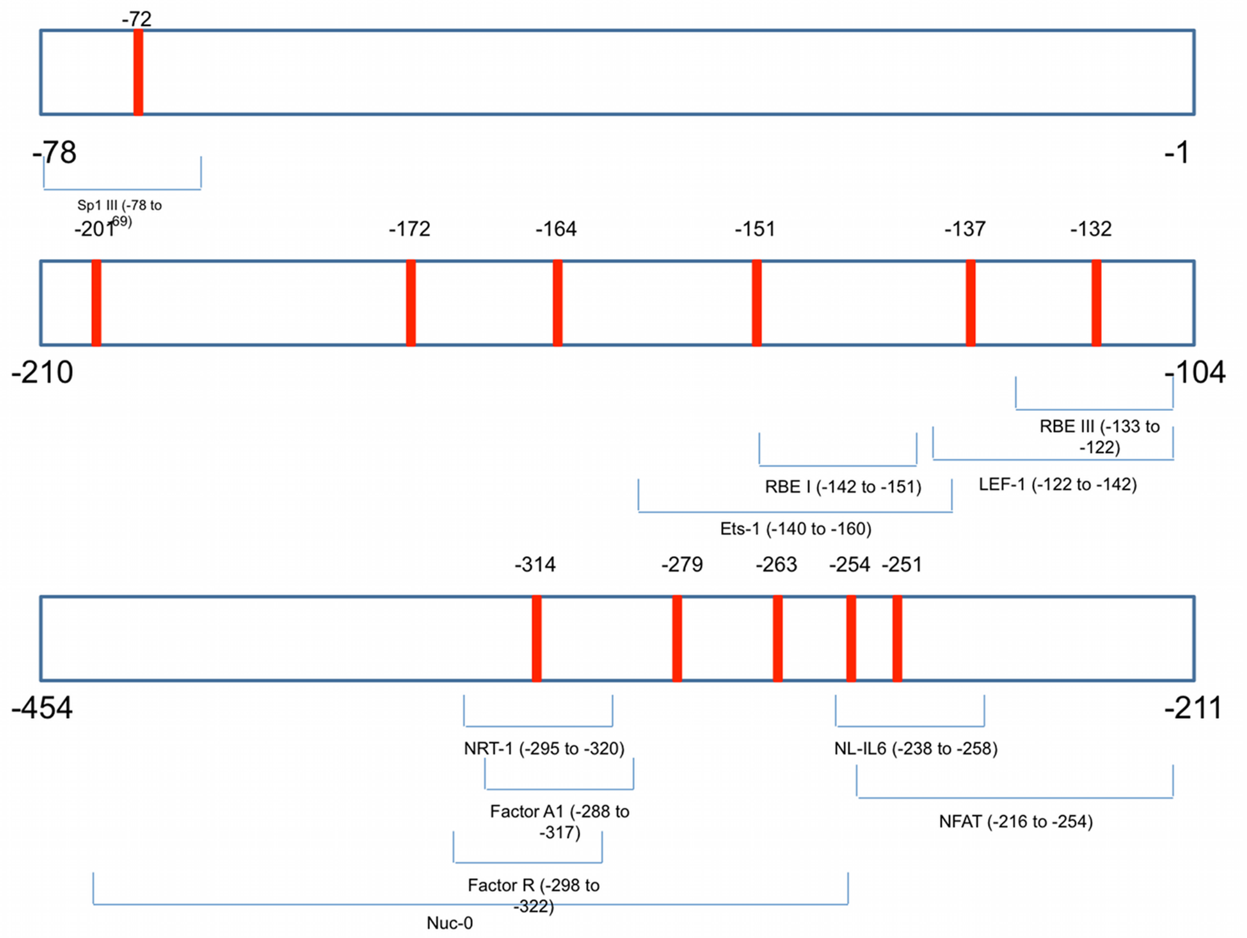


Figure 5.15. Directed evolution identified sites within the promoter that increase the rate of gene activation. Schematic shows sites in red that appear to statistically increase the rate of gene activation. The position of these sites in the promoter is shown above the red bars. Transcription factor binding sites in the vicinity of these sites are shown below the bar bars. 11 sites within upstream promoter elements were identified that may increase the rate of gene activation. The directed evolution approach also identified sites within Sp1 that had previously been shown to increase gene activation rates.

Off the 11 sites within the upstream promoter region that appeared to increase the rate of promoter activation, one of the sites (position -132) was within the TFBS RBE III. We decided to test if mutation at this site makes the promoter more stochastic and therefore potentially alter the latency properties of the virus. We made a single point mutation at position -132 in the subtype B promoter. We found that this increased the initial rate of gene activation from the inactive state as well as doubled the fraction of PheB clones (Fig. 5.16). Thus, we validated the library data for one of the positions that was predicted to change the viral phenotype as well as identify new TFBS within the HIV promoter that may be functionally important in regulating viral gene expression dynamics. Further investigation of the other mutations identified by the error-prone library will help identify important TFBS, and biochemical characterization using chromatin immunoprecipitation will reveal novel functions for these sites.

5.7 Studying HIV-1 latency for different subtype promoters using primary cell-culture models

The data presented above used full-length replication competent or lentiviral vectors to show that subtype promoters produce differential gene expression, replication and latency. To study how sequence differences may influence HIV-1 latency, we decided to use a clinically relevant primary cell culture model.

To use primary CD4+ T-cells, peripheral blood mononuclear cells (PBMCs) were isolated from 4 healthy human donors. Naïve CD4+ cells are then isolated from PBMCs using antibody-mediated magnetic associated cell sorting (MACS) and stimulated using α-CD3/CD28 beads (5).

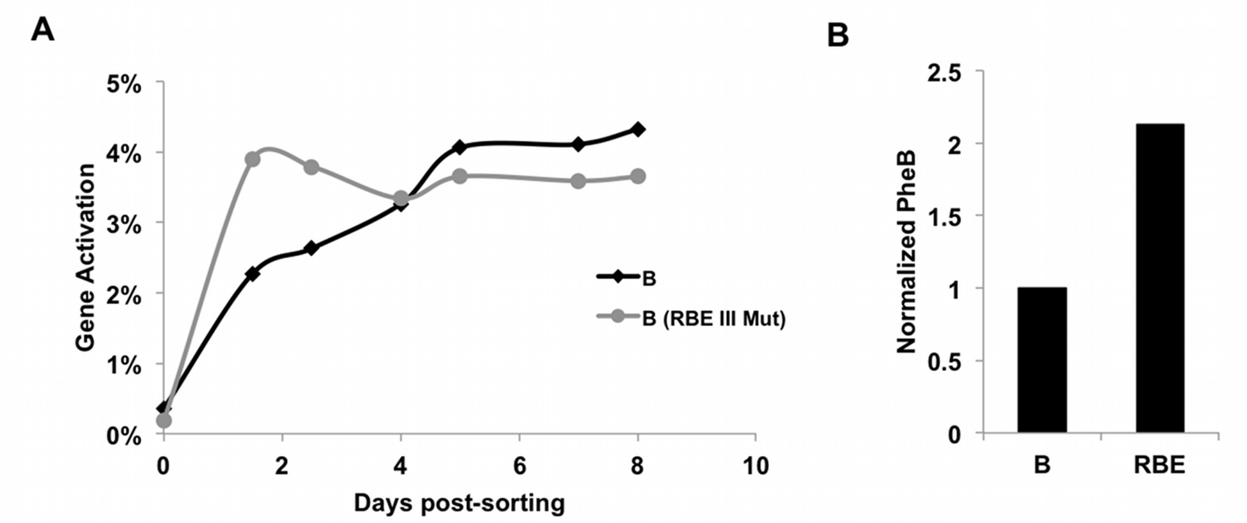


Figure 5.16. Mutation within the RBE III site increases the rate of gene activation from the inactive state. To validate the sites identified from the directed evolution experiment, we introduced mutation at site -132, within the RBE III TFBS. (A) As described in Figure 8, the rate of gene activation was monitored for subtype B and the RBE III mutant. We found that the RBE III mutation increased the initial rate of gene activation. (B) Analysis also showed that the RBE III mutant increases the fraction of PheB clones 2-fold, suggesting that this site and possibly the RBE III TFBS may be playing a critical role in regulating gene expression dynamics from the HIV-1 promoter.

Latently infected cells are rare compared to actively infected T-cells, and as we were specifically interested in latent viral populations and differences that are induced by subtype promoters, we cloned subtype B and subtypes A, A2, D, F, C* and their corresponding mimics into the sLTR vector described in Section 5.2. GFP in this sLTR vector was replaced by the Herpes Simplex Virus – Thymidine Kinase (HSV-TK) gene. Thus T-cells actively replicating would produce HSV-TK that can be killed by addition of the drug ganciclovir (GCV) that specifically targets HSV-TK, thus negatively selecting for latently infected cells. Differences in the fraction of latently infected cells for each subtype promoter will then be quantified using small-molecules (such as prostatin and SAHA) that active latent populations or by quantifying integrated proviruses using qPCR. Thus, the use of this clinically relevant primary cell culture model of latency will allow us to systematically probe differences in the propensity of latency for different HIV-1 subtype promoters.

5.8 Discussion

In this work we studied how sequence diversity and variations in the architecture of TFBS regulate viral gene expression and latency. Immediately post viral integration and during reactivation from latency, the viral protein Tat is present at low levels or absent, and recruitment of cellular factors to the viral promoter plays a critical role in initiating and upregulating viral transcription. Relating the viral promoter genotype to HIV-1 pathogenicity is important for understanding differences between subtypes. To accomplish this, we employed a combination of lentiviral vectors and full-length HIV to explore how variations in TFBS between subtypes differentially regulate gene expression, viral replication and latency. We found that subtypes have widely varying replication kinetics, with subtypes such as C* and B/C, containing an extra NF-κB/Sp1 site replicating rapidly. In contrast, subtypes such as D with mutations within Sp1 sites have lower replication rates.

In studying how the replication rate may be influenced by differential rates of gene expression, we found that subtypes with extra NF-κB/Sp1 sites (such as C*, C and B/C) have strong gene expression whereas those with mutations within Sp1 sites (such as F) have reduced gene expression at steady state with a large fraction of cells in the inactive or latent state. By replacing the NF-κB/Sp1 sites in the subtype B promoter with that of other subtypes, we were able to show that this domain determines the strength of gene expression at steady state. Further, to study the rate of reactivation from latency and integration into certain genomic locations that drive bimodal gene expression and increase chances of latency, we explored the dynamics of gene expression and found that both the NF-κB/Sp1 sites and upstream promoter elements play an important role in determining the rate and levels of gene activation/inactivation from an inactive/active state and the frequency of obtaining PheB clones. Finally, we are currently employing full-length virus in primary cell-culture models with different subtype U3 regions to understand differences in the propensity for latency between subtypes.

Thus we found that subtype promoters, with large sequence diversity produce differences in gene expression, replication rates and latency. Specifically, we were able to identify regions of the promoters that regulate different properties of viral gene regulation. We showed that the NF-κB/Sp1 sites constitute the minimal set of TFBS that regulate steady state gene expression whereas specific TFBS within upstream regions of the promoter and NF-κB/Sp1 sites regulate gene expression dynamics.

A number of studies have explored how subtypes can alter virulence, transmission and gene expression. However, several of these studies use transfection-based assays under basal and Tat-transactivated conditions. However, these studies do not take into account how polyclonal integration of the HIV provirus could impact gene expression dynamics and thereby affect viral latency. In contract to the data described above, a recent study found no differences in the levels of latency between most subtypes promoter and that of subtype B (53). However, discrepancy between this and our work possibly arises since they do not consider the entire subtype U3 region in their studies but instead only clone in the subtype-specific region corresponding to -177 to +68. Further, the use of p24 as a reporter of gene expression, which is a late product is possibly not ideally suited to quantify viral gene expression. In contrast, this study systematically explores how different elements within the viral promoter regulate gene expression and latency.

While the U3 region plays a critical role in regulating gene expression by recruiting several transcription factors to the promoters, other regions of the viral genome have been shown to be important in setting the strength of gene expression. The viral protein Tat, the TAR-Tat interaction, other structural and regulatory proteins of HIV-1 and the site of integration of the provirus have all been shown to influence viral gene expression and HIV-1 pathogenicity. Thus while the overall gene expression properties of any subtype will depend on a combination of all these factors, the promoter plays a critical role in the initial recruitment of factors to the promoter immediately after viral integration or during reactivation from latency. Thus, this study provides a better understanding of how sequence diversity in the HIV-1 promoter could differentially impact viral replication and the establishment and reactivation from latency.

5.9 Materials and Methods

5.9.1 Plasmids

U3 regions from different subtypes were cloned into pLG and pCLGIT as described previously (John Burnett, PhD Thesis). The NF-κB/Sp1 mimicking variants were created by introducing point mutations using QuikChange PCR (Stratagene) within NF-κB/Sp1 sites in the subtype B promoter. Primers will be made available upon request.

sLTR vectors containing different subtype promoters were made by PCR amplifying the promoters from the pLGIT vectors and cloning them into sLTR using the restriction enzymes Kas I and Pme I.

For selecting latent viral infections in the primary cell culture experiments, GFP in the sLTR vector was replaced by HSV-TK. HSV-TK was PCR amplified from the plasmid pHIV-TK obtained from the NIH AIDS Research and Reference Reagent Program (54) and cloned into sLTR using the restriction enzymes Pme I and Not I.

5.9.2 Cell culture

The Jurkat, SupT1 and CEM GFP cell lines were cultured in RPMI 1640 (Mediatech) with 10% fetal bovine serum (FBS) and 100U/mL Penicillin-Streptomycin (P-S). HEK 293T cell line was cultured in Isocove's DMEM (Mediatech) with 10% FBS and 100U/mL P-S. The cells were propagated at 37^0C and 5% CO_2.

5.9.3 Viral packaging and infections

To package the LG and LGIT vectors, HEK 293T cells were cotransfected with 10 μg of the pLG or pLGIT plasmids containing various subtype promoters and the following helper plasmids: pMDLg/pRRE, pVSV-G and pRSV-Rev (55). Virus was harvested by ultracentrifugation 36 hours post-transfection, and viral pellets were resuspended in PBS and stored at -80^0C. Viral titers were obtained by infecting $3x10^5$ cells with different viral volumes and measuring GFP expression of cells on day 8 post-infection after stimulating them with TNFα (20 ng/mL) and TSA (400 nM) for 18 hours prior to GFP measurements using flow cytometry. The tittering curves were used to infect Jurkat cells at a MOI of 0.05-0.1 to ensure single integration events per cell.

To package full-length virus, HEK 293T cells were cotransfected with the sLTR vectors containing different subtype promoters and the following helper plasmids to increase packaging efficiency: pMDLg/pRRE, pVSV-G, pRSV-Rev and pCLPIT-Tat mCherry (51). Prior to transfection, the sLTR vectors were digested using Eco RI and Pvu I for 1 hour and after heat inactivation of the restriction enzymes and extraction of DNA, the plasmid was ligated using DNA ligase. The virus was packaged as described above and amplified using SupT1 cells. Viral titering was performed by infecting the CEM GFP cell line with various viral volumes. The CEM GFP cells were then fixed in paraformaldehyde and GFP expression was monitored using flow cytometry.

5.9.4 Replication competent HIV propagation

$4x10^5$ SupT1 cells were infected at a MOI of 0.0005 in 12-well plates. Over a 10-day time-course experiment, 700 μL of media was extracted from the culture media every 2 days and replaced with fresh media. This culture media removed was used to estimate infectious viral units using the CEM GFP cell line.

5.9.5 Flow cytometry and cell sorting

GFP expression was monitored using the FC500 Flow Cytometer (Beckman Coulter). For the bulk studies, Jurkat cells infected with various LGITs were stimulated with TNFα 18 hours before sorting, and GFP+ cells were sorted using a Cytopeia INFLUX Sorter or DAKO-Cytomation MoFlo High Speed Sorter. These sorted cells were allowed to relax and GFP- and strongly GFP+ cells were sorted from this population. Relaxation of these two bulk populations was then monitored over time using the FC500 Flow Cytometer.

For identifying PheB clones from different subtype promoters, LGIT infected single Jurkats cells were sorted into 96-well plates. These single cells were expanded for 14-21 days and GFP expression of these clonal populations were measured using the FC500 Flow Cytometer.

5.9.6. Isolating primary CD4+ T-cells and primary cell culture experiments

Isolation of primary CD4+ T-cells was performed as described in Section 5.7 and in reference (5). Latently infected cells were selected by treating the primary CD4+ T-cells with ganciclovir (GCV) as described in Section 5.7.

5.10 References

1. Chun, T. W., Davey, R. T., Jr., Engel, D., Lane, H. C., and Fauci, A. S. (1999) *Nature* **401**, 874-875
2. Wong, J. K., Hezareh, M., Gunthard, H. F., Havlir, D. V., Ignacio, C. C., Spina, C. A., and Richman, D. D. (1997) *Science (New York, N.Y.)* **278**, 1291-1295
3. Finzi, D., Blankson, J., Siliciano, J. D., Margolick, J. B., Chadwick, K., Pierson, T., Smith, K., Lisziewicz, J., Lori, F., Flexner, C., Quinn, T. C., Chaisson, R. E., Rosenberg, E., Walker, B., Gange, S., Gallant, J., and Siliciano, R. F. (1999) *Nature medicine* **5**, 512-517

4. Siliciano, J. D., Kajdas, J., Finzi, D., Quinn, T. C., Chadwick, K., Margolick, J. B., Kovacs, C., Gange, S. J., and Siliciano, R. F. (2003) *Nature medicine* **9**, 727-728
5. Burnett, J. C., Lim, K. I., Calafi, A., Rossi, J. J., Schaffer, D. V., and Arkin, A. P. (2010) *Journal of virology* **84**, 5958-5974
6. Lassen, K., Han, Y., Zhou, Y., Siliciano, J., and Siliciano, R. F. (2004) *Trends in molecular medicine* **10**, 525-531
7. Geeraert, L., Kraus, G., and Pomerantz, R. J. (2008) *Annual review of medicine* **59**, 487-501
8. Colin, L., and Van Lint, C. (2009) *Retrovirology* **6**, 111
9. Coiras, M., Lopez-Huertas, M. R., Perez-Olmeda, M., and Alcami, J. (2009) *Nature reviews. Microbiology* **7**, 798-812
10. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., and Schaffer, D. V. (2005) *Cell* **122**, 169-182
11. Yukl, S., Pillai, S., Li, P., Chang, K., Pasutti, W., Ahlgren, C., Havlir, D., Strain, M., Gunthard, H., Richman, D., Rice, A. P., Daar, E., Little, S., and Wong, J. K. (2009) *Virology* **387**, 98-108
12. Bannwarth, S., and Gatignol, A. (2005) *Current HIV research* **3**, 61-71
13. Lewinski, M. K., Bisgrove, D., Shinn, P., Chen, H., Hoffmann, C., Hannenhalli, S., Verdin, E., Berry, C. C., Ecker, J. R., and Bushman, F. D. (2005) *Journal of virology* **79**, 6610-6619
14. Han, Y., Lin, Y. B., An, W., Xu, J., Yang, H. C., O'Connell, K., Dordai, D., Boeke, J. D., Siliciano, J. D., and Siliciano, R. F. (2008) *Cell host & microbe* **4**, 134-146
15. Mok, H. P., and Lever, A. M. (2007) *Genome biology* **8**, 228
16. Blazkova, J., Trejbalova, K., Gondois-Rey, F., Halfon, P., Philibert, P., Guiguen, A., Verdin, E., Olive, D., Van Lint, C., Hejnar, J., and Hirsch, I. (2009) *PLoS pathogens* **5**, e1000554
17. Kauder, S. E., Bosque, A., Lindqvist, A., Planelles, V., and Verdin, E. (2009) *PLoS pathogens* **5**, e1000495
18. Williams, S. A., Chen, L. F., Kwon, H., Ruiz-Jarabo, C. M., Verdin, E., and Greene, W. C. (2006) *The EMBO journal* **25**, 139-149
19. Huang, J., Wang, F., Argyris, E., Chen, K., Liang, Z., Tian, H., Huang, W., Squires, K., Verlinghieri, G., and Zhang, H. (2007) *Nature medicine* **13**, 1241-1247
20. Roy, S., Delling, U., Chen, C. H., Rosen, C. A., and Sonenberg, N. (1990) *Genes & development* **4**, 1365-1373
21. Dingwall, C., Ernberg, I., Gait, M. J., Green, S. M., Heaphy, S., Karn, J., Lowe, A. D., Singh, M., and Skinner, M. A. (1990) *The EMBO journal* **9**, 4145-4153
22. Bosque, A., and Planelles, V. (2009) *Blood* **113**, 58-65
23. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P., and Schaffer, D. V. (2009) *PLoS pathogens* **5**, e1000260
24. Pereira, L. A., Bentley, K., Peeters, A., Churchill, M. J., and Deacon, N. J. (2000) *Nucleic acids research* **28**, 663-668
25. He, G., and Margolis, D. M. (2002) *Molecular and cellular biology* **22**, 2965-2973
26. Li, Y., Mak, G., and Franza, B. R., Jr. (1994) *The Journal of biological chemistry* **269**, 30616-30619
27. Kaehlcke, K., Dorr, A., Hetzer-Egger, C., Kiermer, V., Henklein, P., Schnoelzer, M., Loret, E., Cole, P. A., Verdin, E., and Ott, M. (2003) *Molecular cell* **12**, 167-176

28. Mahmoudi, T., Parra, M., Vries, R. G., Kauder, S. E., Verrijzer, C. P., Ott, M., and Verdin, E. (2006) *The Journal of biological chemistry* **281**, 19960-19968
29. Hoberg, J. E., Popko, A. E., Ramsey, C. S., and Mayo, M. W. (2006) *Molecular and cellular biology* **26**, 457-471
30. Baeuerle, P. A., and Baltimore, D. (1989) *Genes & development* **3**, 1689-1698
31. Gerritsen, M. E., Williams, A. J., Neish, A. S., Moore, S., Shi, Y., and Collins, T. (1997) *Proceedings of the National Academy of Sciences of the United States of America* **94**, 2927-2932
32. Doetzlhofer, A., Rotheneder, H., Lagger, G., Koranda, M., Kurtev, V., Brosch, G., Wintersberger, E., and Seiser, C. (1999) *Molecular and cellular biology* **19**, 5504-5511
33. Suzuki, T., Kimura, A., Nagai, R., and Horikoshi, M. (2000) *Genes to cells : devoted to molecular & cellular mechanisms* **5**, 29-41
34. Gomez-Gonzalo, M., Carretero, M., Rullas, J., Lara-Pezzi, E., Aramburu, J., Berkhout, B., Alcami, J., and Lopez-Cabrera, M. (2001) *The Journal of biological chemistry* **276**, 35435-35443
35. Ross, E. K., Buckler-White, A. J., Rabson, A. B., Englund, G., and Martin, M. A. (1991) *Journal of virology* **65**, 4350-4358
36. McAllister, J. J., Phillips, D., Millhouse, S., Conner, J., Hogan, T., Ross, H. L., and Wigdahl, B. (2000) *Virology* **274**, 262-277
37. De Baar, M. P., De Ronde, A., Berkhout, B., Cornelissen, M., Van Der Horn, K. H., Van Der Schoot, A. M., De Wolf, F., Lukashov, V. V., and Goudsmit, J. (2000) *AIDS research and human retroviruses* **16**, 499-504
38. Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004) *Nature reviews. Genetics* **5**, 52-61
39. van Opijnen, T., Jeeninga, R. E., Boerlijst, M. C., Pollakis, G. P., Zetterberg, V., Salminen, M., and Berkhout, B. (2004) *Journal of virology* **78**, 3675-3683
40. Desfosses, Y., Solis, M., Sun, Q., Grandvaux, N., Van Lint, C., Burny, A., Gatignol, A., Wainberg, M. A., Lin, R., and Hiscott, J. (2005) *Journal of virology* **79**, 9180-9191
41. Jeeninga, R. E., Hoogenkamp, M., Armand-Ugon, M., de Baar, M., Verhoef, K., and Berkhout, B. (2000) *Journal of virology* **74**, 3740-3751
42. Siddappa, N. B., Venkatramanan, M., Venkatesh, P., Janki, M. V., Jayasuryan, N., Desai, A., Ravi, V., and Ranga, U. (2006) *Retrovirology* **3**, 53
43. Kurosu, T., Mukai, T., Komoto, S., Ibrahim, M. S., Li, Y. G., Kobayashi, T., Tsuji, S., and Ikuta, K. (2002) *Microbiology and immunology* **46**, 787-799
44. Aulicino, P. C., Holmes, E. C., Rocco, C., Mangano, A., and Sen, L. (2007) *Journal of virology* **81**, 427-429
45. Rousseau, C. M., Daniels, M. G., Carlson, J. M., Kadie, C., Crawford, H., Prendergast, A., Matthews, P., Payne, R., Rolland, M., Raugi, D. N., Maust, B. S., Learn, G. H., Nickle, D. C., Coovadia, H., Ndung'u, T., Frahm, N., Brander, C., Walker, B. D., Goulder, P. J., Bhattacharya, T., Heckerman, D. E., Korber, B. T., and Mullins, J. I. (2008) *Journal of virology* **82**, 6434-6446
46. Novitsky, V., Smith, U. R., Gilbert, P., McLane, M. F., Chigwedere, P., Williamson, C., Ndung'u, T., Klein, I., Chang, S. Y., Peter, T., Thior, I., Foley, B. T., Gaolekwe, S., Rybak, N., Gaseitsiwe, S., Vannberg, F., Marlink, R., Lee, T. H., and Essex, M. (2002) *Journal of virology* **76**, 5435-5451

47. Roof, P., Ricci, M., Genin, P., Montano, M. A., Essex, M., Wainberg, M. A., Gatignol, A., and Hiscott, J. (2002) *Virology* **296**, 77-83

48. De Arellano, E. R., Soriano, V., and Holguin, A. (2005) *AIDS research and human retroviruses* **21**, 949-954

49. Turk, G., Carobene, M., Monczor, A., Rubio, A. E., Gomez-Carrillo, M., and Salomon, H. (2006) *Retrovirology* **3**, 14

50. Montano, M. A., Novitsky, V. A., Blackard, J. T., Cho, N. L., Katzenstein, D. A., and Essex, M. (1997) *Journal of virology* **71**, 8657-8665

51. Shah, P. S., Pham, N. P., and Schaffer, D. V. (2012) *Molecular therapy : the journal of the American Society of Gene Therapy* **20**, 840-848

52. Das, J., Ho, M., Zikherman, J., Govern, C., Yang, M., Weiss, A., Chakraborty, A. K., and Roose, J. P. (2009) *Cell* **136**, 337-351

53. van der Sluis, R. M., Pollakis, G., van Gerven, M. L., Berkhout, B., and Jeeninga, R. E. (2011) *Retrovirology* **8**, 73

54. Smith, S. M., Markham, R. B., and Jeang, K. T. (1996) *Proceedings of the National Academy of Sciences of the United States of America* **93**, 7955-7960

55. Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., and Naldini, L. (1998) *Journal of virology* **72**, 8463-8471

Chapter 6: Chromatin accessibility at the HIV LTR promoter sets a threshold for NF-κB mediated viral gene expression

6.1 Introduction

A central question in eukaryotic gene expression is how the activation of gene expression depends simultaneously on transcription factor availability and quantitative features of the chromatin environment at different genomic locations (1) (Fig. 6.1*A*). Eukaryotic transcription factors commonly regulate multiple genes, yet extracellular stimuli that activate transcription factors result in selective expression of only a subset of these genes. The sequence and arrangement of transcription factor binding sites in different promoters cannot fully explain differential responses to the same transcription factor (2). Another important input, chromatin features of the genomic locus, can also provide regulatory selectivity in response to transcription factor activation, including in complex processes such as inflammation (3,4) and development (5). It would therefore be informative to quantify how the placement of a particular gene in the genome impacts its responsiveness to an input transcription factor signal and features of the local chromatin environment. Such a quantitative understanding of how chromatin environment impacts gene regulation may also improve rational design of therapies to reverse gene expression dysregulation induced by chromatin changes (6).

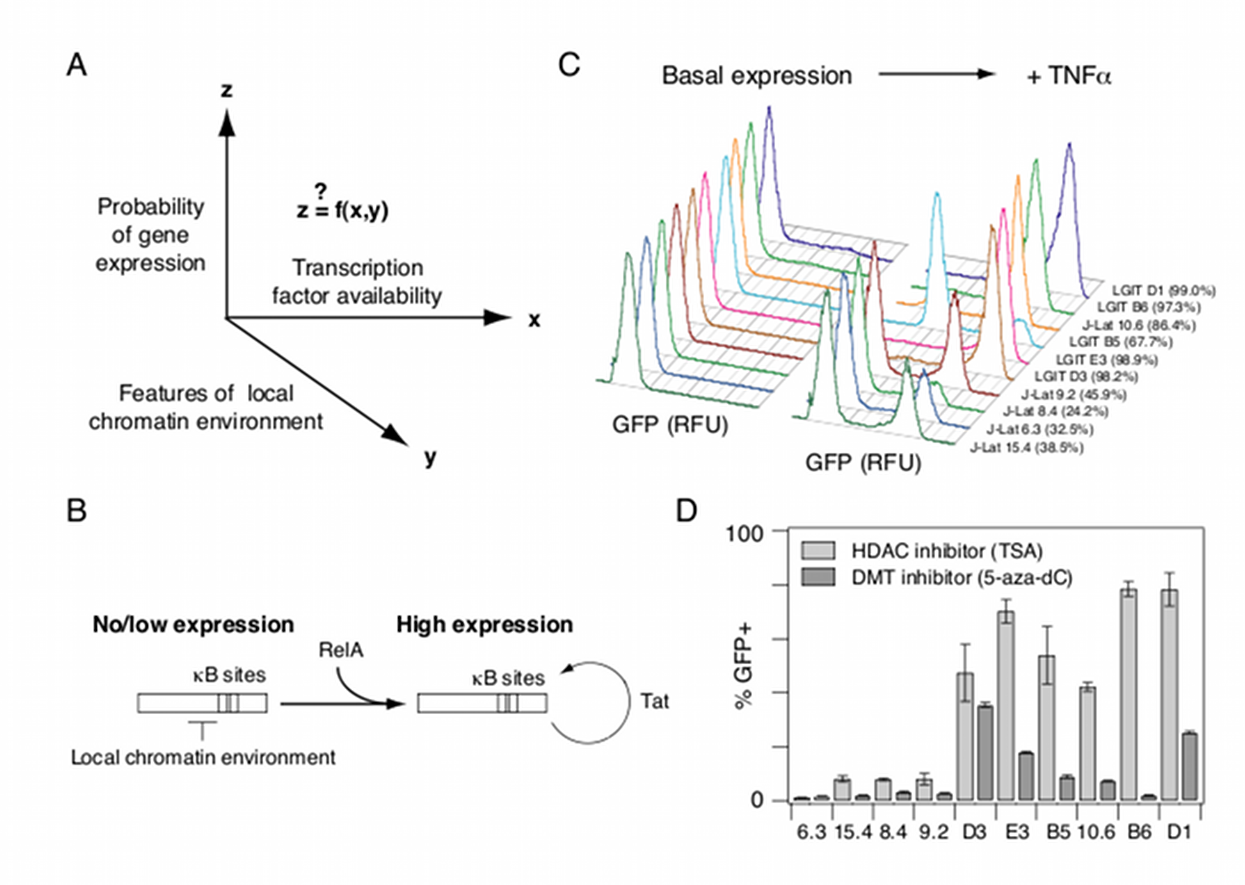


Figure 6.1. *In vitro* models of HIV gene expression provide an experimental system to study RelA-mediated gene expression in a range of chromatin environments. (A) There is general interest in how gene expression probability varies as a function of transcription factor availability and quantitative features of the local chromatin environment. (B) Schematic describing RelA-mediated gene expression in the HIV vectors before and after the Tat-mediated positive feedback loop is activated. (C) Representative flow cytometry histograms of GFP expression for the panel of clones each infected with a single integration of an inactive HIV provirus under basal conditions (left) and after stimulation with TNFα (20 ng/ml) for 48 hours (right). Percentage of TNFα-activated cells is indicated in parentheses. Clones are ordered according to increasing basal gene expression. (D) Infected clonal populations were stimulated with 400 nM TSA for 24 hours (light gray bars) or 5 μM 5-aza-dC for 48 hours (dark gray bars). Experiments were performed in biological triplicate. Data are presented as the mean ± standard deviation.

Studies in *S. cerevisiae* recently demonstrated that chromatin provides a mechanism for tuning gene expression in response to transcription factors by setting a gene induction threshold that is decoupled from gene expression range (7,8). However, it is unclear if a similar relationship holds for genes in multicellular organisms, in which gene expression attenuation and silencing are mediated by more complex repressive chromatin modifications (9). To address this question, we studied activation of the human retrovirus human immunodeficiency virus-1 (HIV). Because HIV integrates into the genome of its host cell in a semi-random fashion and responds to host transcription factors, it provides a unique opportunity to study activation of the same gene by the same transcription factor in different chromatin environments without altering promoter architecture (10,11).

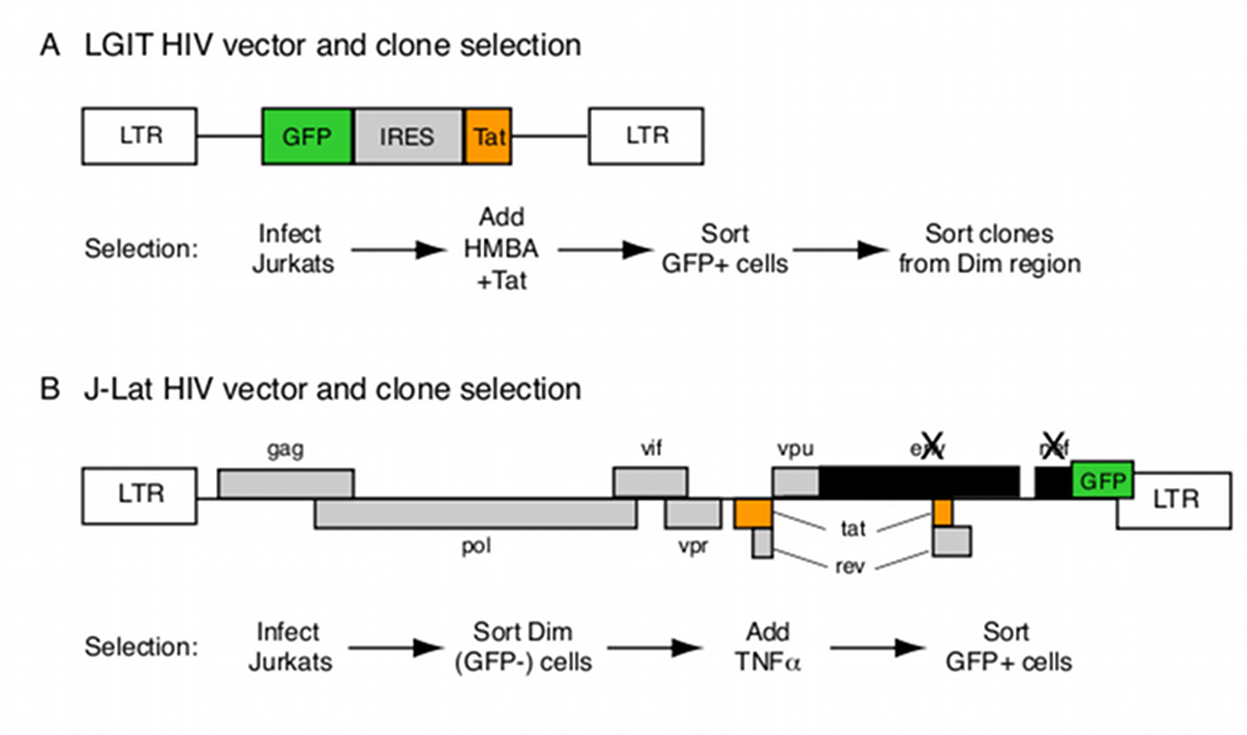
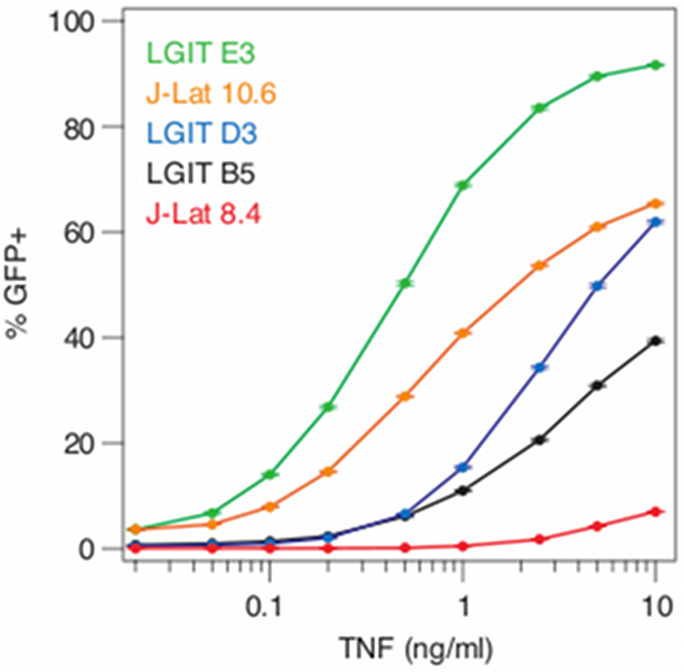


Figure 6.2. Schematic of the HIV vector model and a brief description of the selection procedure for clones infected with a single copy of the virus. (A) LTR-GFP-IRES-Tat (LGIT) clones (12) and (B) J-Lat clones as established by Jordan et al. (13). Please refer to the published references for full details of the selection procedures.

Following infection and integration into the host chromosome, initial expression from the HIV long terminal repeat (LTR) promoter is inefficient and subject to the availability of the host cell transcriptional machinery and to local factors operating at the integration site (Fig. 6.1*B*) (14-16). In some cases, chromatin-mediated repression of HIV gene expression – including histone deacetylation, histone methylation, and DNA methylation – results in inactive viral gene expression that may be related to viral latency, in which the virus adopts a quiescent phenotype but can be reactivated when stimulated with the appropriate transcriptional cues (13,17-20). Within inactive HIV-1 promoters, a nucleosome is precisely positioned immediately downstream of the transcription start site (TSS), and transcriptional activation of silent proviruses is strongly correlated with its removal via chromatin remodeling complexes (21,22). Upon such LTR activation, a virally-encoded transcriptional activator (Tat) feeds back on the LTR to amplify gene expression nearly 100-fold (Fig. 6.1*B*) (23,24), and stochastic effects in this process may also contribute to viral latency (12,25,26). Thus, inactive HIV integrated at different genomic locations offers a biomedically relevant system to study the probability of gene activation from the same mammalian promoter in a spectrum of repressive chromatin environments.



<u>Figure 6.3. Clones display differential sensitivity to TNFα dose.</u> Clones were treated with the indicated dose of TNFα for 24 hours and HIV gene expression was evaluated by flow cytometry. Data are presented as mean ± standard deviation as estimated by bootstrapping.

Like most cellular promoters, the HIV LTR is also strongly regulated by global host factors, prominently including the transcription factor nuclear factor-κB (NF-κB) p65/RelA. Transcription factors of the NF-κB/Rel family control complex transcriptional patterns in both the innate and adaptive immune responses, and these diverse patterns in part result from differences in the chromatin structure of target genes (3,27). Upon stimulation with a NF-κB pathway activator, such as the inflammatory cytokine tumor necrosis factor-α (TNFα), RelA translocates to the nucleus and binds to the HIV LTR to stimulate gene expression (21,28). Specifically, NF-κB RelA promotes elongation by RNA polymerase II (RNAPII) in the absence of Tat (29) and is thought to be important in mediating the activation of silent proviruses (30). Thus, as a model system, the HIV LTR provides a common promoter architecture to quantitatively explore how NF-κB RelA mediates gene expression in different chromatin environments.

Table 6.1: Genomic Locations of the integrated provirus for the clones used in this study

					Repeats		Proximity to CpG Island		
Clone ID	Chromosome	Location	Origin	Gene	Integrated in	Within 200 bp	Left	Right	Expression[1]
LGIT B5	12	51,557,735	–	TFCP2	SINE/Alu	LINE/L1	80,183	8,945	Below
LGIT B6	3	185,636,060	–	TRA2B	None	LTR /ERV1	91,914	18,997	Above
LGIT D1[2]	--	--	--	--	--	--	--	--	--
LGIT D3	17	5,254,712	+	RABEP1	None	SINE/Alu	68,839	67,795	Median
LGIT E3	8	61,524,041	+	RAB2A	None	LINE/L2	94,041	39,979	Median
J-Lat 6.3	19	46,884,266	+	PPP5C	LINE/L1	LINE/L1	33,588	31,536	Median
J-Lat 8.4	1	78,412,065	+	FUBP1	None	AT-rich	57,392	32,239	Above
J-Lat 9.2	19	46,884,266	+	PPP5C	LINE/L1	LINE/L1	33,588	31,536	Median
J-Lat 10.6	9	139,362,925	+	SEC16A	None	None	1,782	4,261	Median
J-Lat 15.4	19	34,932,169	–	UBA2	SINE/Alu	LINE/L1	12,199	39,881	Median

1. Refers to the expression level of the gene in the Jurkat cell line relative to the level of gene expression for all other cell lines in the NCI-60 cancer cell panel.
2. Integration position for clone D1 was not successfully characterized.

Here we quantified viral gene expression as a function of NF-κB RelA level and quantitative features of the chromatin environment at the viral integration site. In cell populations containing different clonal integrations of the LTR promoter, we found that the threshold level of RelA necessary to initiate gene expression in the cell population varied monotonically with the degree of chromatin accessibility at the LTR promoter. Furthermore, upon onset, gene expression increased as a function of additional RelA increases in a non-linear manner similar for all clones. Moreover, increasing chromatin accessibility via small molecule inhibition of either histone deacetylation or DNA methylation reduced the RelA threshold without otherwise changing this gene activation function. Finally, an empirical gene activation function describing the dependence of HIV gene expression on RelA level and chromatin accessibility accurately predicted synergistic activation in response to combinatorial treatment with chromatin- or DNA-modifying enzyme inhibitors and TNFα. Thus, our results demonstrate that chromatin accessibility at LTR promoters, mediated by complex epigenetic modifications acting at the integration site, sets a threshold level of RelA required for promoter activation, after which the activation profile is conserved across genomic locations. These findings point to a general mechanism by which genomic location may establish differential gene expression in response to the same transcription factor. These results may also aid efforts to develop combinatorial therapies to reverse chromatin repression and purge latent HIV reservoirs (31,32).

6.2 Inactive HIV infections in Jurkat T cells show varying degrees of repression and differential response to NF-κB pathway activation

Inactive HIV infections of Jurkat leukemic T cells provide an opportunity to study gene expression in response to the same transcription factor from a single promoter located in different genomic environments. Here, we studied two *in vitro* models previously used to study HIV latency, in which clonal populations of Jurkat cells harbor a single viral integration at different genomic locations (12,13). LGIT-infected clones contain a minimal, replication-incompetent HIV-based lentiviral vector with Tat and GFP under the control of the LTR promoter (Fig. 6.2*A*) (12), whereas J-Lat clones contain a full-length, replication-incompetent HIV virus with GFP in place of the Nef gene (Fig. 6.2*B*) (13). In the early stages of viral gene expression, Tat and GFP are the primary proteins expressed from the full-length virus, and the mechanism of transcriptional activation is thus similar for both models (19,25). Also, both J-Lat and LGIT exhibit bimodal gene expression, where the virus can exist in a non-expressing state, or where Tat basal expression is amplified by a positive feedback loop to yield transactivated expression (Fig. 6.1*B*).

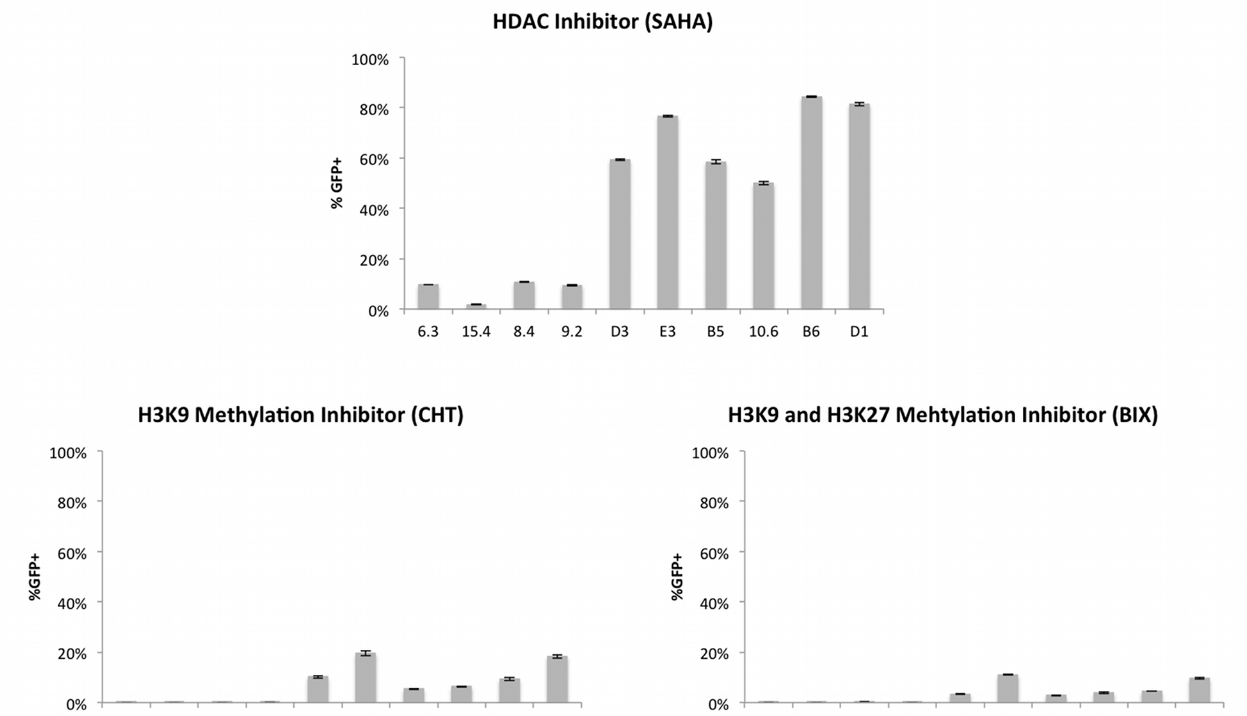


Figure 6.4. Clones show differential activation to small molecules that derepress different epigenetic mechanisms. Activation of gene expression was monitored by flow cytometry in clones after stimulating them with SAHA (4μM) for 24 hours and CHT (15nM) and BIX (3μM) for 48 hours. Experiments were preformed in biological triplicate and error bars indicate standard deviations from the mean.

To explore a range of behaviors, we selected complementary sets of LGIT clones – in which a small fraction of the cells exhibit active transcription and the rest remain inactive – or J-Lat clones – which are generally more silent since they were originally selected to have no basal gene expression unless stimulated with TNFα (Fig. 6.2*B*) (13). We compared five LGIT clones (B5, B6, D1, D3, and E3) and five J-Lat clones (6.3, 8.4, 9.2, 10.6, and 15.4) that showed low or no GFP expression from the LTR promoter in the absence of stimulation, as measured by flow cytometry (Fig. 6.1*C*). All clones were activated to some extent by NF-κB RelA via TNFα stimulation, indicating that all integrated promoters could support viral transcription (Fig. 6.1*C*); however, activation occurred to varying degrees. In general, TNFα stimulation activated a smaller fraction of J-Lat clonal populations compared to LGIT clonal populations, except for J-Lat 10.6, which was activated to a greater extent than LGIT B5. Moreover, the TNFα dose required to activate gene expression across clonal populations varied more than 10-fold (Fig. 6.3). Thus, the J-Lat and LGIT *in vitro* latency models display a range of gene expression activation in response to the transcription factor RelA in different genomic environments.

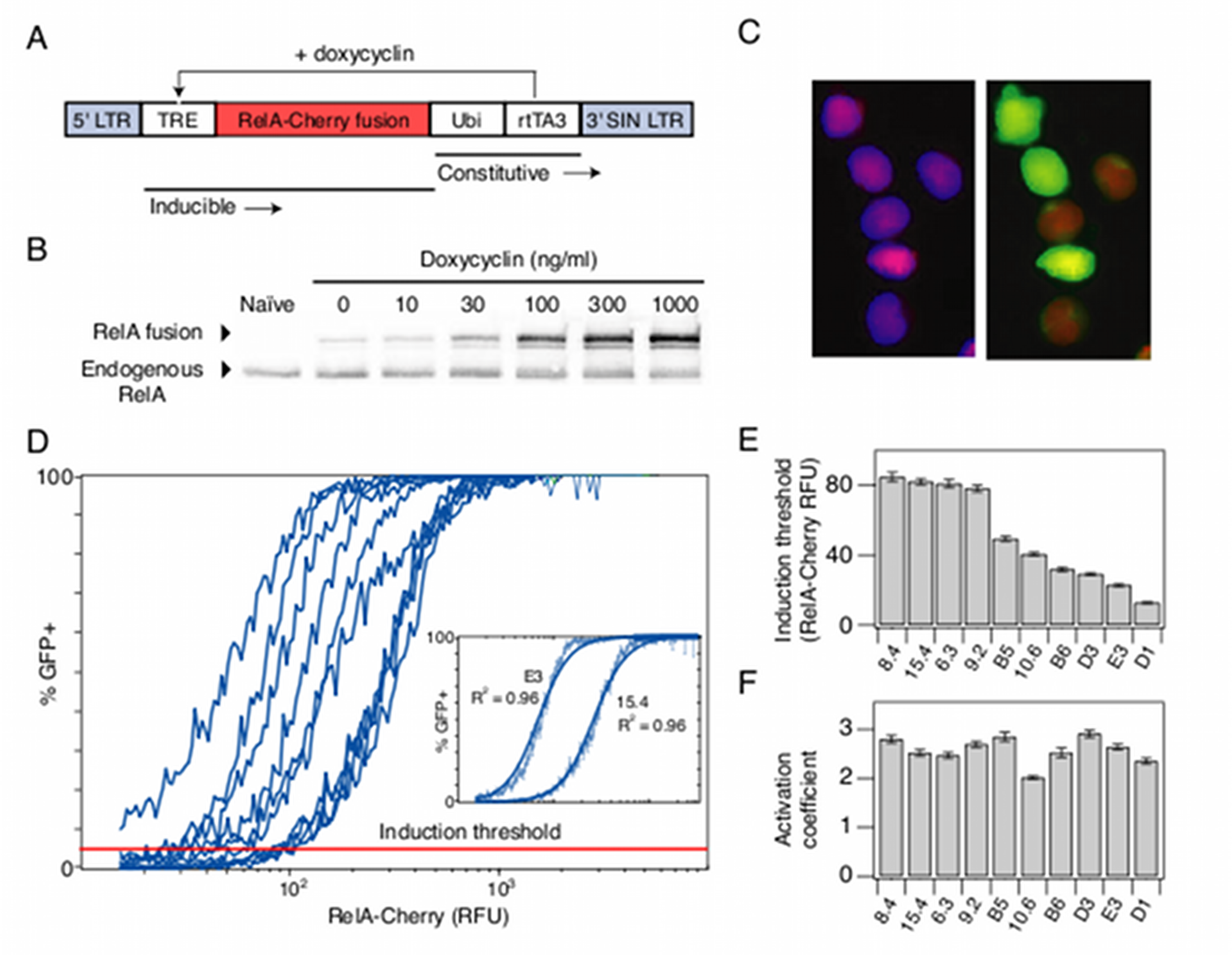


Figure 6.5. Inducing HIV gene expression by overexpression of RelA reveals an induction threshold of gene activation. (A) Schematic of the inducible RelA (iRelA) vector. (B) Immunoblot of total RelA-Cherry fusion protein and endogenous protein levels in clone 6.3 infected with iRelA 4 days after DOX induction. (C) Microscopy picture of clone 6.3 infected with iRelA 4 days after induction with 30 ng/ml DOX. (Left) DAPI and mCherry overlay. (Right) GFP and mCherry overlay. (D) Combined flow cytometry data for HIV-infected clones expressing iRelA in response to a range of DOX concentrations. More than 50,000 single cell events were divided into 256 bins of mCherry fluorescence, and the fraction of GFP+ cells was calculated and plotted for each bin. (Inset) Least squares fit line for clone 15.4 and E3. (E) Induction threshold (defined as the mCherry-RelA level at which 5% of the population expressed GFP) and (F) activation coefficient (defined as the Hill coefficient calculated from fitting

Hill functions to the curves in (D)) for each clone. Error bars in (D-F) represent standard deviations and were calculated by bootstrapping.

We hypothesized that differences in TNFα-mediated activation may be due to epigenetic modifications at the LTR promoter that result in higher order chromatin structure, as suggested in previous studies (18,19). Local genomic features of the integration site did not reveal any systematic differences among the clones (Table 6.1). To chemically probe the nature of repression at the site of integration in each clone, we added trichostatin A (TSA), an inhibitor of class I and II mammalian HDACs, or 5-aza-2'-deoxycytidine (5-aza-dC), which inhibits DNA methyltransferase (DMT) activity. Similar to TNFα treatment, TSA or 5-aza-dC stimulation activated gene expression to varying extents across the clonal populations (Fig. 6.1*D*). To probe other epigenetic mechanisms of gene expression repression, we also treated the cell lines with another HDAC inhibitor, suberoylanilide hydroxamic acid (SAHA), chaetocin (CHT), an inhibitor of SUV39H1 that methylates histone H3 at lysine 9 (H3K9) or the small molecule BIX-01294 (BIX) that has been shown to inhibit G9a, a methyltransferase that methylates H3K9 and H3K27. All clones showed strong reactivation after treatment with SAHA and weaker activation in general in response to treatment with CHT and BIX (Fig. 6.4). Therefore, these chemical perturbations showed that the panel of inactive integrated proviruses is subject to varying degrees of chromatin repression by multiple epigenetic mechanisms across integration sites and exhibit differential responses to TNFα-mediated RelA activation.

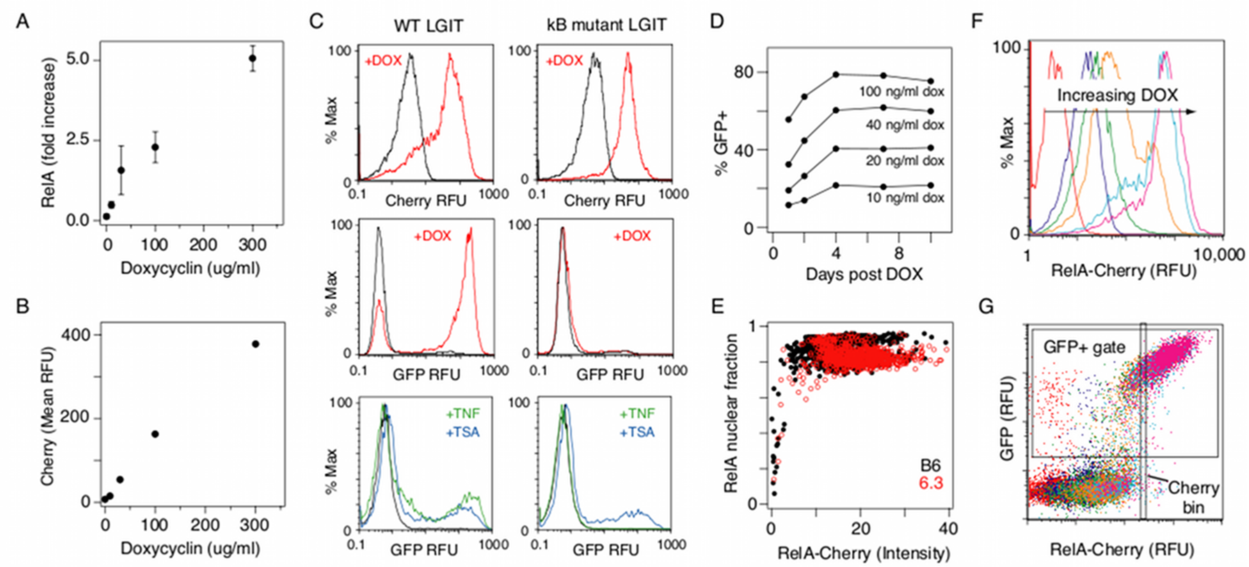
6.3 LTR activation by tunable overexpression of the transcription factor NF-κB RelA revealed an activation threshold that varied significantly across clones

To more quantitatively and directly analyze how RelA activates HIV gene expression in different chromatin environments, we modified a tetracycline inducible expression system (33) for variable expression of a mCherry-RelA fusion protein (iRelA, Fig. 6.5*A*). Treatment with increasing doses of doxycycline (DOX) induced a steady increase in total RelA expression relative to endogenous levels, ranging from approximately a 0.2-fold increase in RelA fusion protein relative to endogenous RelA in the absence of DOX (due to basal expression from the Tet promoter) to a 5-fold increase at high DOX concentrations (Figs. 6.5*B* and 6.6*A*). Total mCherry fluorescence varied with DOX dosage in a similar manner as protein level (Fig. 6.6*B*), confirming that the two measurements are monotonically related. Deletion of the κB sites from the HIV LTR promoter abolished activation by the inducible (iRelA) vector and TNFα, but retained activation by TSA (Fig. 6.6*C*), indicating that RelA overexpression activated the LTR via specifically binding to the LTR κB sites.

We introduced iRelA into the panel of clones and stimulated them across the full range of RelA expression until GFP expression and RelA levels reached steady-state 4 days post DOX addition (Fig. 6.6*D*). Within individual cells, mCherry-RelA predominantly localized to the nucleus for all but the lowest RelA levels (Figs. 6.5*C* and 6.6*E*), suggesting that RelA expression had largely overcome cytoplasmic sequestration by I-κB. Stimulation of an i-RelA-infected population of cells at a particular DOX concentration produced a wide distribution of mCherry-RelA expression (Fig. 6.6*F*). Therefore, to quantify gene activation in the population directly as a function of mCherry-RelA across this full range, we pooled flow cytometry measurements across all DOX levels and subdivided the single cell data into 256 mCherry-RelA bins (Fig. 6.6*G*). Gene expression for each clone varied from minimal activation with low mCherry-RelA to fully activated (i.e. 100% of the population expressing GFP) at maximal RelA levels (Fig. 6.5*D*). The resulting gene activation curves were fit to the Hill equation after log transforming it into a linear equation:

$$(\%GFP+) = \frac{(mCherry)^n}{K^n + (mCherry)^n}$$

$$\log\left[\frac{1}{(\%GFP+)} - 1\right] = n\log K - n\log(mCherry)$$



<u>Figure 6.6. Characterization of the iRelA vector.</u> (A) Quantification of RelA overexpression in clone 6.3 in response to increasing DOX dosage as measured by immunoblot. Western blot was performed in triplicate and mCherry-RelA was normalized to the endogenous RelA level. Data are presented as the mean ± standard deviation. (B) For the same conditions in (A), mean mCherry fluorescence was quantified by flow cytometry and plotted against DOX dosage. Data points represent mCherry RFU mean of ~10,000 cells. (C) Histograms of mCherry fluorescence (top panels) and GFP fluorescence (middle panels) in Jurkat cells infected with WT LGIT (left) or LGIT with κB-deletions in the LTR promoter (right) and co-infected with the RelA vector in the absence (black) and presence (red) of DOX. κB-deleted LGIT vector shows negligible increase in GFP when stimulated with DOX (middle left). (Bottom panels) κB-deletion mutants also show loss of response to TNFα (green) but retain activation by TSA (blue) as compare to WT. (D) Time course of % GFP positive cells at the indicated DOX treatment concentration for LGIT B6. (E) Nuclear mCherry fluorescence as a fraction of total mCherry quantified for 650 J-Lat 6.3 cells and 1250 LGIT B6 cells from single cell microscopy data. (F) Flow cytometry histograms of mCherry fluorescence for clone B6 infected with iRelA in response to increasing DOX dosages: red (uninfected), blue (0 μg/ml), green (10 μg/ml), yellow (30 μg/ml), turquoise (100 μg/ml), and pink (300 μg/ml). (G) Density plots of GFP versus mCherry for all dosages of DOX combined for clone B6. Measurements were divided into 256 mCherry bins and % GFP positive fraction was calculated for each bin.

The experimental gene activation curves were well described using the fit parameters, *K* and *n* (Fig. 6.5*D*, inset and Fig. 6.7) and the quality of the fits was independent of the total number of subdivisions (bins). Strikingly, we observed that gene expression in each clonal population is induced at a different level of RelA (mCherry), but after induction the increase in the GFP+ fraction as a function of RelA is similar (Fig. 6.5*D*). The mCherry-RelA level at which 5% of the population expressed GFP was defined as the induction threshold (Fig. 6.5*D*, red line), calculated using the Hill "gene activation" functions. Note that the relative difference in threshold of activation among clones was independent of the GFP level chosen for computing this metric (Fig. 6.8*A*).

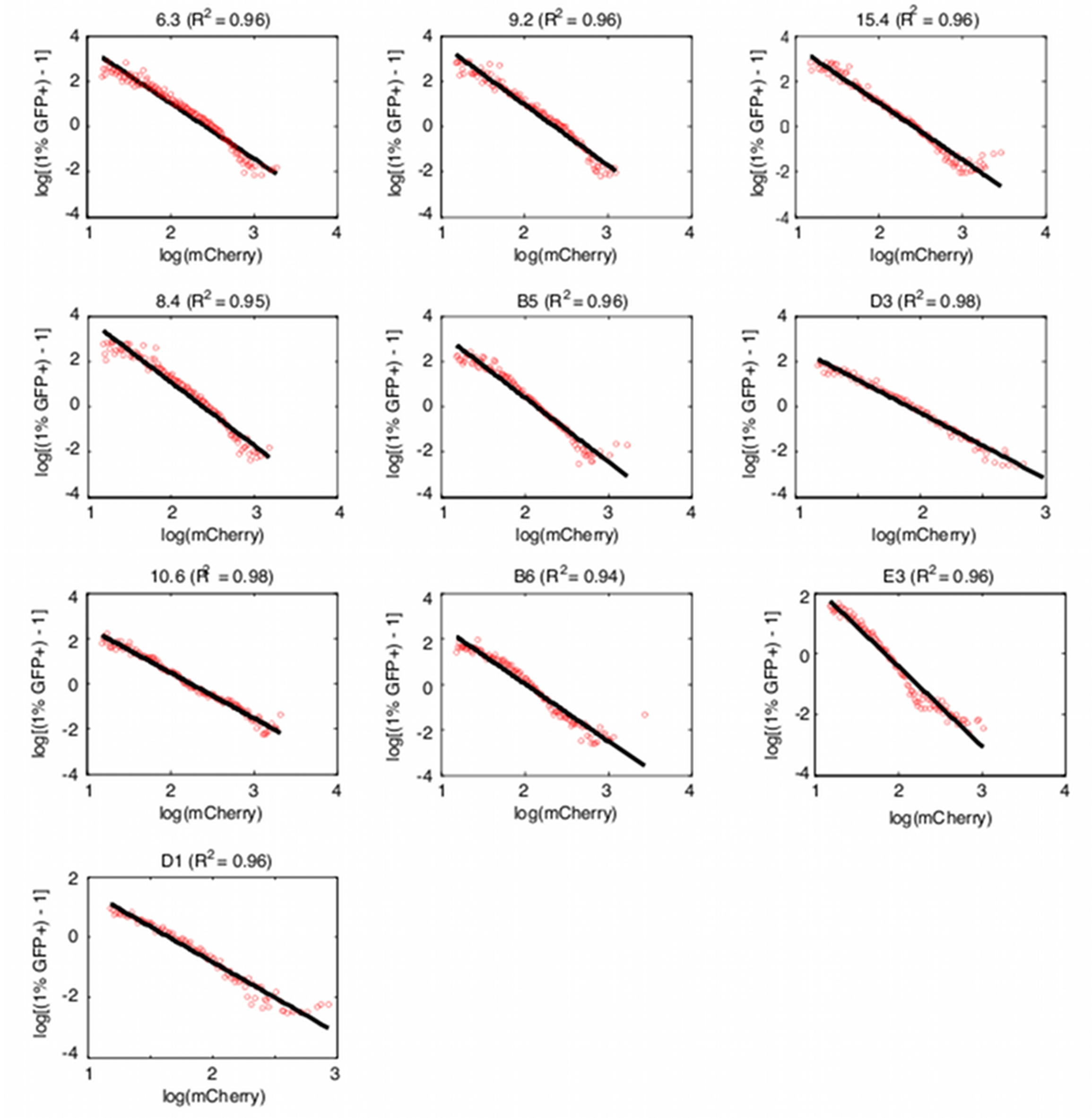


Figure 6.7. Least-squares fits of gene activation functions for the panel of clones studied. Flow cytometry data obtained from stimulating iRelA clones at different DOX levels were pooled together and binned into 256 GFP and mCherry channels. Each blue circle quantifies the metrics shown on the *x*- and *y*-axis, obtained from the mean %GFP+ and mCherry expression within each channel. The black line shows the best fit obtained from least-squares fitting after log transforming the Hill equation in to a linear equation. The clone IDs and the goodness of fits for each clone are indicated. Parameter estimates from the best-fit line were used to compute the induction threshold and activation coefficient shown in Fig. 6.5*E-F*.

The induction threshold exhibited a considerable 6-fold range of variation in mCherry fluorescence units (Fig. 6.5*E*), which was also reflected by variation in the fit parameter *K*, i.e. the mCherry-RelA level at half maximal GFP induction (Fig. 6.8*B-C*). In contrast, the apparent Hill coefficient *n*, which describes the steepness in the rise of the gene activation function, did not vary more than 1.5-fold among clones (Fig. 6.5*F*). The Hill coefficients, which we will refer to as the activation coefficients, were greater than 2, suggesting possible cooperativity in RelA- and Tat-mediated LTR activation (Fig. 6.1*B*).

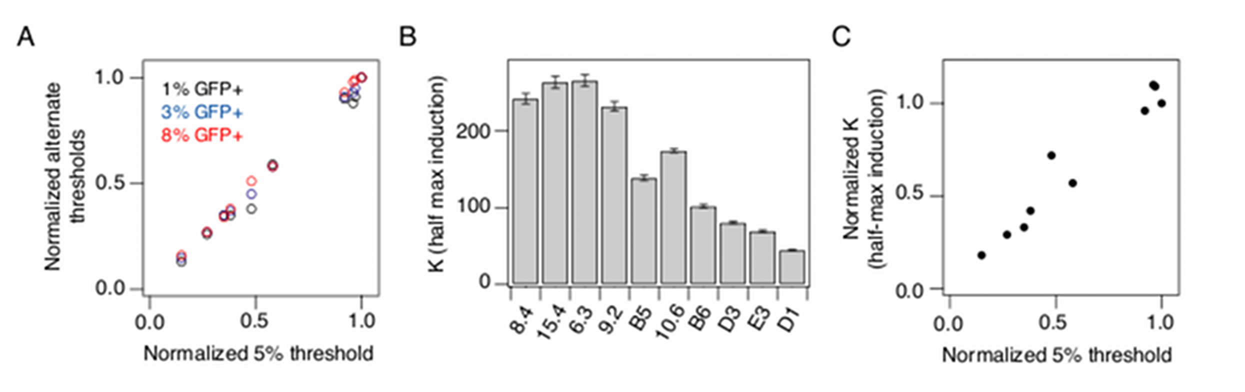


Figure 6.8. Analysis of the variation in induction threshold and K values. (A) Values of RelA induction threshold were calculated for different fractions of GFP+ population and compared to the chosen induction fraction of 5%. GFP+ fractions: 1% (black), 3% (blue), and 8% (red). (B) Values for the fit parameter K (half-max induction). Error bars represent standard deviations and were calculated by bootstrapping. (C) Comparison of K and 5% induction threshold.

Notably, clones that responded more strongly to drug treatments (Fig. 6.1*C-D*) also exhibited lower induction thresholds. Tat transcripts were undetectable below the induction threshold for both LGIT- and J-Lat-infected clones and Tat did not increase significantly until after the induction threshold was reached, indicating that any difference in transcription and Tat production between the two vectors did not affect the threshold (Fig. 6.9). Taken together, these data suggest that the genomic environment at the integration site affected the induction threshold of the LTR in response to RelA, but did not significantly affect progressive RelA-mediated increases in gene expression within the population once the gene had been induced.

6.4 Chromatin accessibility at the LTR across clones is strongly correlated with the RelA induction threshold

We reasoned that the local chromatin environment may affect the induction threshold by modulating chromatin accessibility at the promoter (7,34). To quantitatively compare general chromatin accessibility, we measured the extent to which chromatin limited the sensitivity to DNAse I digestion near the transcription start site (TSS) of the LTR in each clonal population (35,36). Nuclease sensitivity assay measurements of the LTR were normalized to the same measurement made on the highly repressed hemoglobin-β (HBB) reference gene (37) for each clone, and we refer to this normalized metric as the heterochromatin fraction (see Materials and Methods).

The panel of inactive clones harbored proviruses in a wide range of chromatin environments, with heterochromatin fractions varying 100-fold (from clone 6.3 down to clone B6) (Fig. 6.10*A*). The differences in heterochromatin fraction could be resolved into three groups ($p < 0.05$ by one-way ANOVA): strong repression (> 0.5), intermediate repression (0.05-0.5), and weak repression (< 0.05). Importantly, the induction threshold (Fig. 6.5*E*) showed a strong positive correlation with heterochromatin fraction (Fig. 6.10*B*; Pearson R = 0.82, $p < 0.01$), suggesting that chromatin accessibility at the promoter may be a determinant of RelA levels required to initiate gene expression. In contrast, the activation coefficient *n* did not show a significant correlation with nuclease sensitivity (Fig. 6.10*C*), consistent with the observation that this coefficient does not vary across clones (Fig. 6.5*F*). These results suggest that activation following initial gene expression may be an intrinsic property of the promoter, whereas initiation of gene expression is strongly correlated to the local chromatin environment at the site of integration.

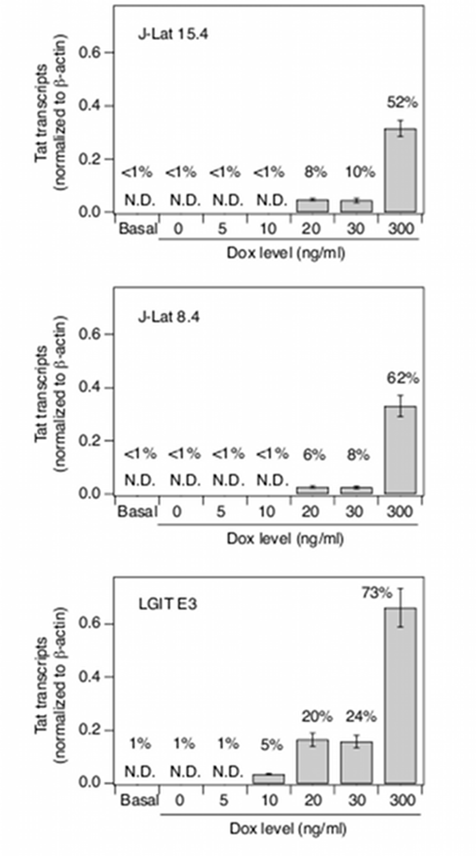


Figure 6.9. Tat transcript levels are undetectable until threshold is reached. RNA was isolated from clones 15.4, 8.4 and E3 in the basal state or in iRelA-infected cell lines after treatment with DOX for 24 hours at the indicated concentrations and Tat transcripts were measured by RT-PCR. β-actin transcription was used as a normalization control. Data are presented as mean ± standard deviation. N.D. indicates non detectable. Corresponding fraction of GFP+ cells for each condition is indicated above each bar. Basal Tat transcription was also measured in LGIT B5, LGIT D3, and J-Lat 10.6 and found to be N.D.

We next measured if the nuclease sensitivity assay was consistent with known molecular determinants of heterochromatin, and how these determinants correlated with the induction threshold and activation coefficient induced by RelA overexpression. Using chromatin immunoprecipitation, we measured the total amount of histone 3 (H3), presumably higher with increased nucleosome occupancy near the promoter; the level of H3 tri-methylation at lysine 9 (H3K9me3), associated with repressed promoters; and the level of H3 acetylation (AcH3), associated with active promoters.

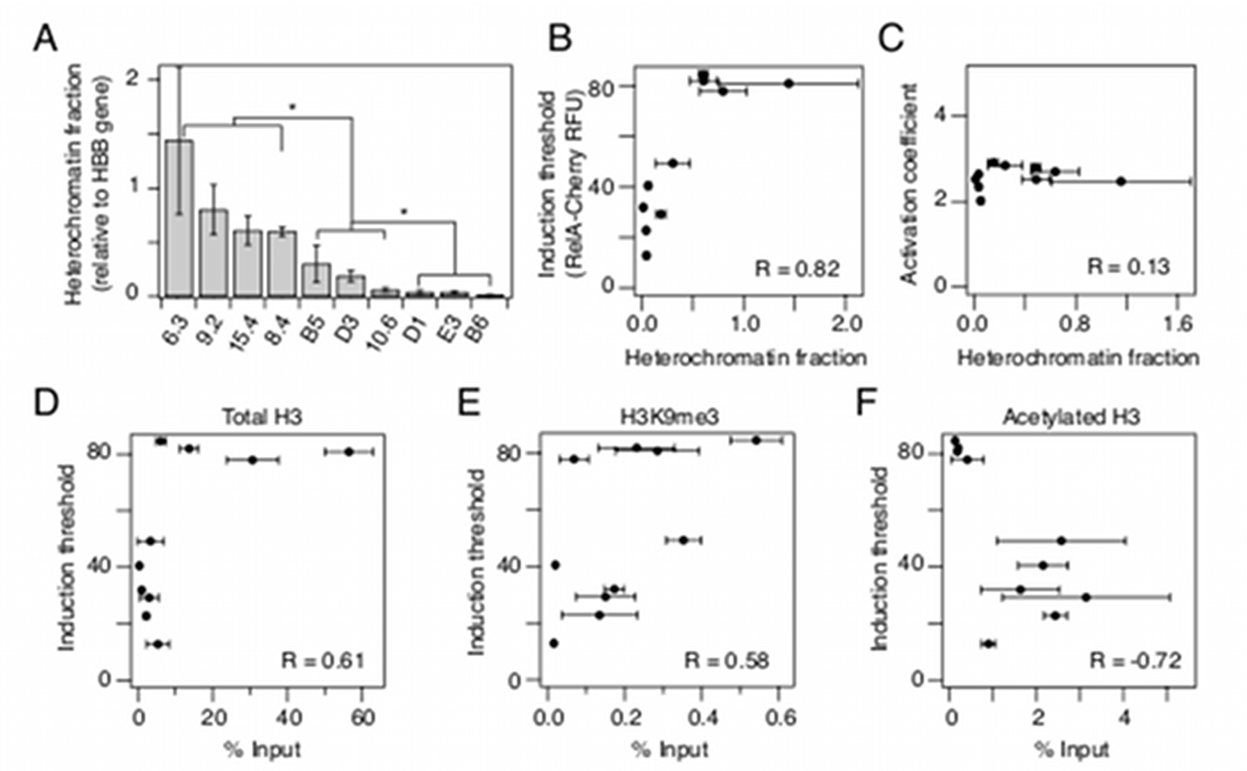


Figure 6.10. Chromatin accessibility is correlated with RelA induction threshold. (A) Heterochromatin fraction was quantified with a DNAse I sensitivity assay. Quantitative PCR was performed in triplicate and normalized to a hemoglobin-β (HBB) reference gene. (B-C) Correlation of heterochromatin fraction with (B) induction threshold and (C) activation coefficient extracted from the fits in Fig. 6.5*D-F*. (D-F) Chromatin immunoprecipitation for (D) total H3, (E) H3K9me3 and (F) acetylated H3 bound to the HIV promoter in unstimulated clones was correlated to the induction threshold. Quantitative PCR was performed in triplicate and normalized to an input control. Data are presented as the mean ± standard deviation. Differences are labeled as significant (*) if $p < 0.05$. Pearson correlation coefficient R is indicated on plot.

As anticipated, total H3 increased with increasing heterochromatin fraction and was positively correlated with the induction threshold (Fig. 6.10*D*; R = 0.61, p = 0.06). H3K9me3 levels were also generally higher for clones with higher heterochromatin fractions and also showed a positive correlation with the induction threshold (Fig. 6.10*E*; R = 0.58, p = 0.08). In contrast, total AcH3 was generally lower for increased heterochromatin fraction and negatively correlated with the induction threshold (Fig. 6.10*F*; Pearson R = -0.72, p = 0.02). The activation coefficient was not significantly correlated with total histone levels or histone modifications (Fig. 6.11). Therefore, the threshold level of RelA necessary to activate gene expression is significantly correlated with chromatin accessibility and molecular determinants of heterochromatin across loci.

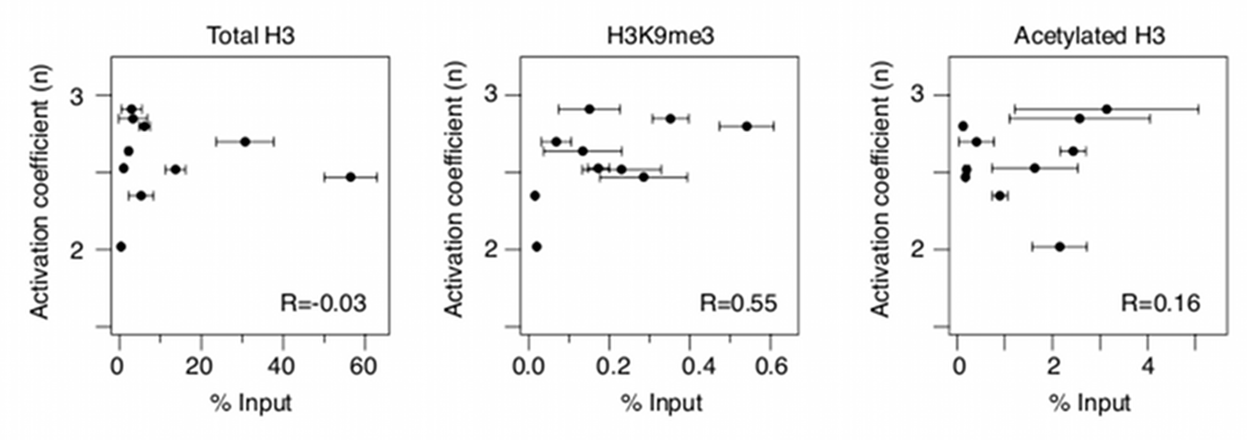
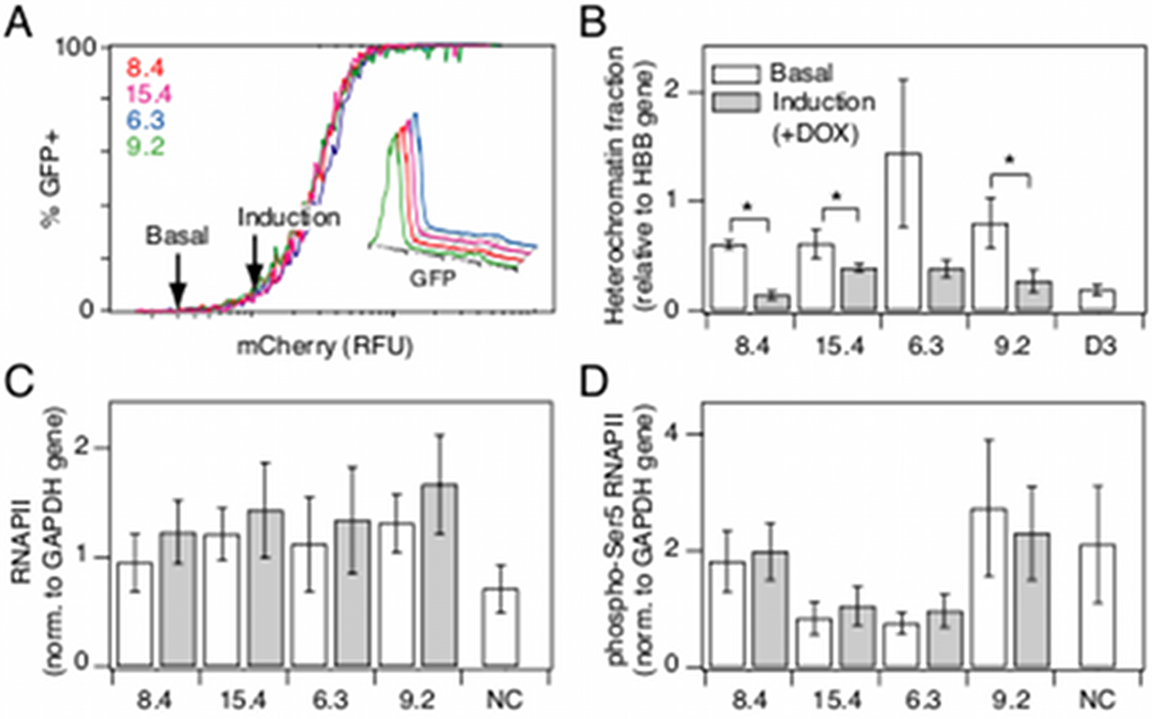


Figure 6.11. The induction threshold does not correlate with known repressive and activating epigenetic marks. Total H3, H3K9me3 and acetylated H3 bound to the HIV promoter in unstimulated clones was measured by chromatin immunoprecipitation and correlated to the activation coefficient that describes the gene activation function. Quantitative PCR was performed in triplicate and normalized to an input control. Data are presented as mean ± standard deviation.

6.5 Activation of gene expression is more strongly associated with a decrease in heterochromatin rather than an increase in RNAPII binding or phosphorylation

For strongly repressed clones (6.3, 9.2, 15.4, and 8.4), significant increases in RelA levels are necessary to reach an induction threshold (Fig. 6.5*D*). Therefore, we used these clones to test what quantitative features at the promoter change between the basal state and the point at which gene expression has just been initiated. Based on the measured correlations between the induction threshold and chromatin structure (Fig. 6.10), we hypothesized that at the point of gene expression onset, the heterochromatin fraction at the promoter may be reduced to that of clones that have induction thresholds close to the basal RelA level. We thus measured the heterochromatin fraction for each clone at a low DOX concentration (20 ng/ml) that approximately increased RelA to the induction threshold, at which point a small fraction of cells expressed GFP (5-8%; Fig. 6.12*A*, inset). The heterochromatin fraction at the induction threshold was compared to heterochromatin at the basal level for each clone, at which point <1% of cells express GFP. The level of heterochromatin at the induction threshold was reduced for all four clones, and three exhibited statistically significant decreases relative to the basal state (Fig. 6.12*B*, $p < 0.05$). Moreover, at the induction threshold, the measured heterochromatin fraction was not significantly different from that of clones displaying intermediate levels of repression (clones B5, 10.6, and D3; $p = 0.09$ by ANOVA), consistent with the hypothesis that chromatin accessibility becomes equalized at the induction threshold.

An alternative to alleviating promoter repression at the induction threshold would be increased recruitment of positive regulators of transcription, including RNAPII and the associated factors required for transcription initiation. We used chromatin immunoprecipitation to measure the level of total RNAPII and RNAPII phosphorylated at serine 5 (pSer5-RNAPII associated with transcription initiation) at the promoter. No significant differences in LTR-bound RNAPII or pSer5-RNAPII were measured in the basal state across the entire panel of clones (Fig. 6.13*A*-*B*) and both were low relative to an actively expressing GFP+ HIV-infected population (Fig. 6.13*C*-*D*).



<u>Figure 6.12. Induction of gene expression is associated with a decrease in heterochromatin fraction.</u> (A) Selected clones were treated with 20 ng/ml DOX to hold the clonal populations at the point at which gene expression in the population is just induced (arrow). (Inset) Flow histograms showing a low fraction of cells expressing GFP for each clone at the point of induction. (B) Heterochromatin fraction as quantified by nuclease sensitivity for clones at basal (white) and induction (gray) level of RelA. Quantitative PCR was performed in triplicate and normalized to a HBB reference gene. (C-D) Chromatin immunoprecipitation comparing (C) RNA polymerase II and (D) phospho-Ser5 RNAPII bound to the LTR promoter at basal (white) and induction (gray) level of RelA. Quantitative PCR was performed in triplicate and normalized to a GAPDH control gene. Data are presented as the mean ± standard deviation. Changes are labeled as significant (*) if $p < 0.05$.

Moreover, no significant changes were measured between basal conditions and threshold conditions at induction for either RNAPII (Fig. 6.12*C*) or pSer5-RNAPII bound to the promoter (Fig. 6.12*D*), consistent with our measurements of Tat transcription (Fig. 6.9). Overall, we conclude that the heterochromatin fractions in different clones begin to converge as they reach a gene expression threshold, prior to significant increases in Tat, RNAPII binding and phosphorylation at the promoter.

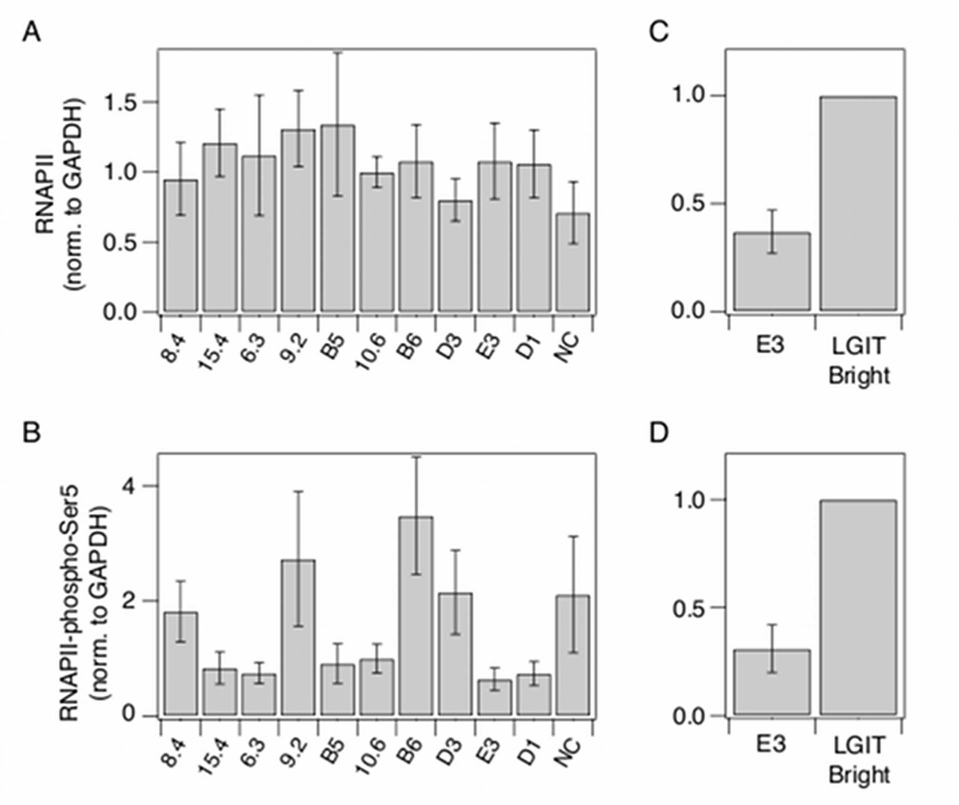


Figure 6.13. Total RNAPII and RNAPII phosphorylated at Ser5 (RNAPII-pSer5) are present at low levels at the HIV LTR promoter of inactive clones. Chromatin immunoprecipitation (ChIP) for (A) total RNAPII and (C) RNAPII-pSer5 bound to the HIV promoter in unstimulated clones. ChIP for (B) RNAPII and (D) RNAPII-phospho-Ser5 was compared between clone E3 and a polyclonal population of Jurkat cells singly-infected with an LGIT vector and sorted for GFP-expressing cells. Quantitative PCR was performed in triplicate and normalized to a GAPDH control. Data are presented as the mean ± standard deviation.

6.6 Increasing chromatin accessibility via small molecule inhibitors lowers the RelA induction threshold

If chromatin accessibility at the integration site is a determinant of the RelA induction threshold, then increasing chromatin accessibility at the LTR promoter of strongly repressed clones, which have relatively high induction thresholds, may shift the gene activation response curves to resemble more weakly repressed clones. While TSA or 5-aza-dC did not highly activate gene expression in clone 15.4 (approximately 1-2% for both drugs), these compounds may still modulate chromatin accessibility. We thus treated 15.4 with TSA (40 or 400 nM) or 5-aza-dC (5 μM) and analyzed nuclease sensitivity following incubation times previously demonstrated to be sufficient for producing measurable changes in H3 acetylation (4 hours for TSA) (17) or DNA methylation (48 hours for 5-aza-dC) (38). Nuclease sensitivity depended on TSA dosage (Fig. 6.14*A*). In addition, the higher 400 nM TSA dosage induced an approximately 3-fold decrease in the heterochromatin fraction, and 5 μM 5-aza-dC decreased the heterochromatin fraction by 2-fold, bringing these fractions into intermediate levels of basal heterochromatin (clones B5 and D3; Fig. 6.15*A*).

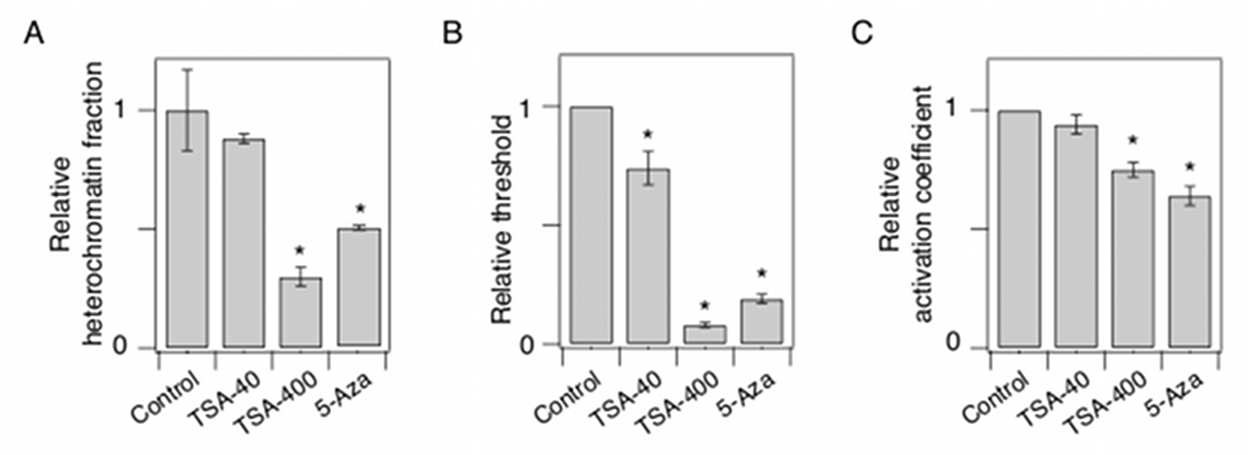


Figure 6.14. Increasing chromatin accessibility lowers the RelA induction threshold for clone 15.4. (A) Heterochromatin fraction for clone 15.4 was quantified with a DNAse I sensitivity assay following stimulation with TSA (40 or 400 nM) for 4 hours or with 5-aza-dC (5 µM) for 48 hours. Quantitative PCR was performed in triplicate and normalized to the HBB reference gene. Relative heterochromatin fraction was calculated by normalizing clone 15.4+drugs to the unstimulated 15.4 control. (B) Threshold and (C) activation coefficient extracted from fitting 15.4+iRelA gene activation functions in the presence of TSA and 5-aza-dC for the conditions in (A). Data are presented as the mean ± standard deviation. Standard deviation error bars for the threshold and activation coefficients were calculated by bootstrapping. Changes are labeled as significant (*) if $p < 0.05$.

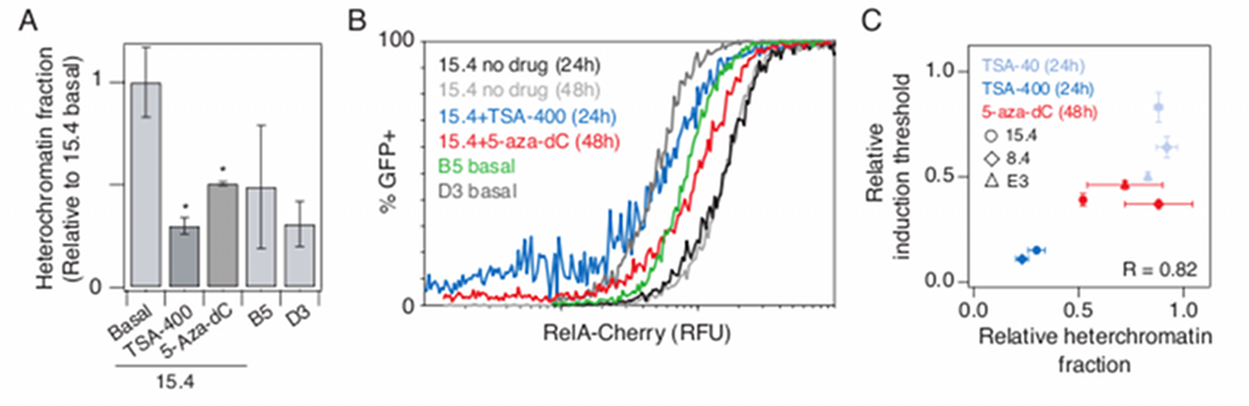


Figure 6.15. Increasing chromatin accessibility via drug treatment lowers the RelA induction threshold. (A) Heterochromatin fraction for clone 15.4 was quantified with a DNAse I sensitivity assay following stimulation with 400 nM TSA for 4 hours or with 5 µM 5-aza-dC for 48 hours. Quantitative PCR was performed in triplicate and normalized to the hemoglobin reference gene. Relative heterochromatin fraction was calculated by normalizing clone 15.4 with drugs, B5 and D3 to the unstimulated 15.4 control. (B) Combined flow cytometry data for 15.4 expressing iRelA in response to a range of DOX concentrations and simultaneous stimulation with 400 nM TSA for 24 hours (dark blue), 5 µM 5-aza-dC for 48 hours (red), and no drug treatment controls at 24 and 48 hours (black and light gray, respectively). iRelA dose response curves for clone B5 (green) and D3 (dark gray) without TSA or 5-aza-dC are included for comparison. (C) Relative change in induction threshold versus relative change in heterochromatin fraction for clones 15.4 (circles), 8.4 (diamonds) and E3 (triangles). Data for 15.4 are calculated from results presented in (A) and (B) and data for 8.4 and E3 are calculated from experiments presented in Fig. 6.16. All points are calculated by normalizing the value of heterochromatin fraction or threshold for the clone in the presence of drugs to the corresponding value for the unstimulated control clone. Data are presented as the mean ± standard deviation. Changes are labeled as significant (*) if $p < 0.05$. Pearson correlation coefficient R is indicated on plot.

To determine whether these shifts in chromatin accessibility lower the induction threshold for clone 15.4, we repeated the DOX induction of RelA-mediated gene activation in the presence of inhibitors at time points before these compounds affected cell viability (24 hours for TSA and 48 hours for 5-aza-dC). We then fit the resulting curves to the Hill equation (as in Fig. 6.5*D*) and extracted new values for the threshold and activation coefficient that define a new gene activation function. As anticipated, the induction threshold in the presence of either TSA or 5-aza-dC was significantly decreased compared to the control (7-fold and 2.5-fold, respectively; Fig. 6.14*B*) and importantly resulted in gene activation curves that resembled those of clones that had intermediate heterochromatin fractions (Figs. 6.15*B* and 6.10*A*). By comparison, the activation coefficient was modestly lower following drug treatment (approximately 25% and 40%, respectively; Fig. 6.14*C*).

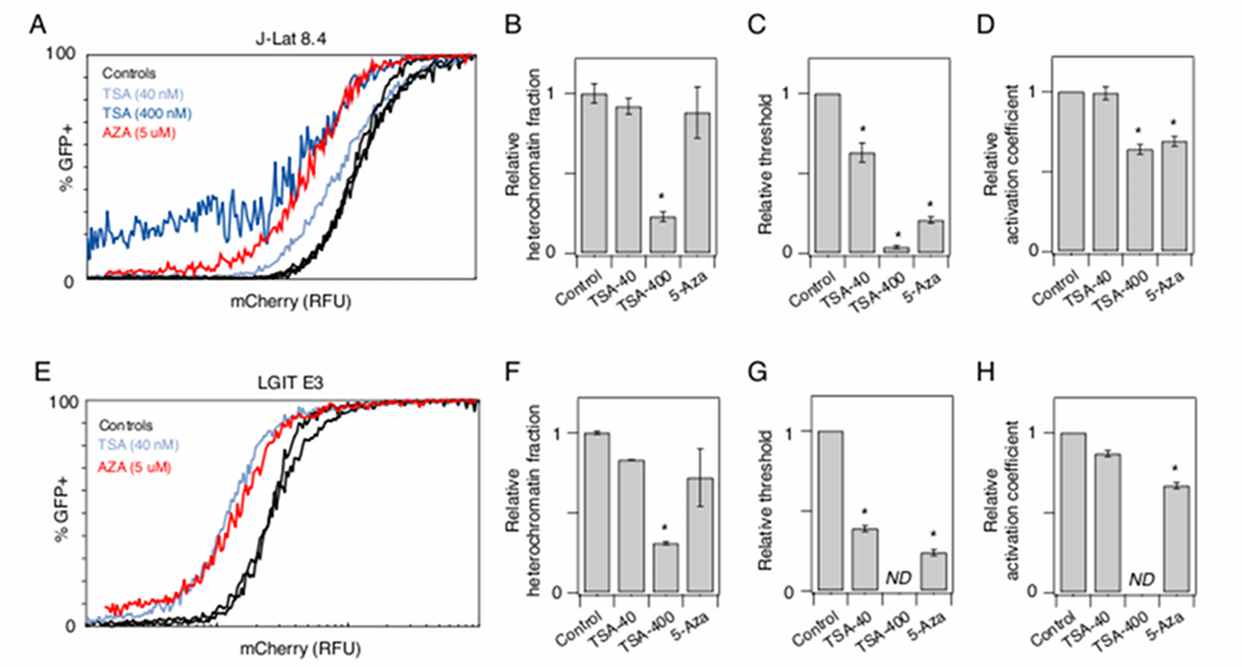


Figure 6.16. Increasing chromatin accessibility is associated with lowering of the RelA induction threshold for clones 8.4 and E3. (A) Combined flow cytometry data for clone 8.4 expressing iRelA in response to a range of DOX concentrations and simultaneous stimulation with 40 nM (light blue), 400 nM TSA (dark blue) for 24 hours, or 5 µM AZA (red) for 48 hours. Corresponding iRelA dose curves without drug treatment are in black. (B) Heterochromatin fraction was quantified with a DNAse I sensitivity assay following stimulation with TSA (40 or 400 nM) for 4 hours or with 5-aza-dC (5 µM) for 48 hours. Quantitative PCR was performed in triplicate and normalized to the HBB reference gene. Relative heterochromatin fraction was calculated as described in Fig. 6.14. (C) Threshold and (D) activation coefficient extracted from fitting 8.4+iRelA gene activation functions in the presence of TSA and 5-aza-dC for the conditions in (A). (E-H) Same conditions as described in (A-D) but repeated in clone E3. Note that 400 nM TSA induced a significant activation of clone E3 at basal RelA levels and so it was not possible to accurately fit a gene activation function. Data are presented as the mean ± standard deviation. Error bars for threshold and activation coefficient are calculated by bootstrapping. Changes are labeled as significant (*) if *$p < 0.05$*.

Furthermore, TSA and 5-aza-dC had similar effects on another repressed clone, 8.4, again inducing increased nuclease sensitivity and a lower induction threshold (Fig. 6.16*A-D*). Finally, we investigated whether increasing the chromatin accessibility could further reduce the RelA induction threshold of even a weakly repressed clone. Consistent with results for the two highly repressed clones, reducing the heterochromatin fraction for the weakly repressed clone E3 with TSA or 5-aza-dC caused a decrease in induction threshold (Fig. 6.16*E-H*).

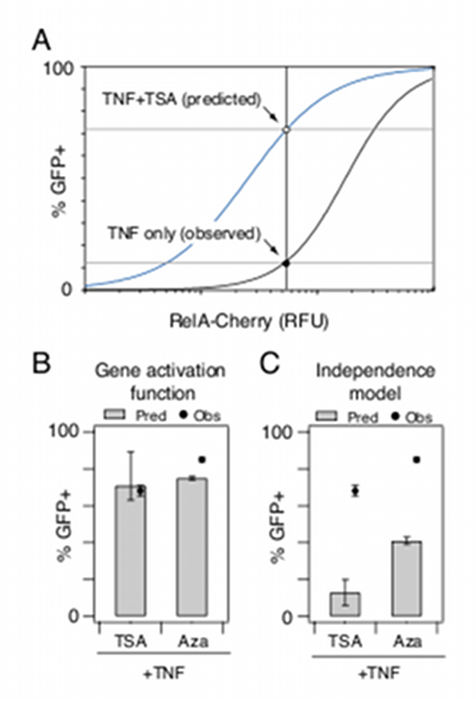


Figure 6.17. Gene activation function accurately predicts synergistic activation of HIV gene expression by simultaneous treatment with TNFα and HDAC or DMT inhibitors. (A) The empirically-derived gene activation function for clone 15.4+TSA was used to predict its response to combinatorial perturbation with TSA and TNFα. Approximate mCherry-RelA increases associated with TNFα treatment alone were estimated by locating the point on the gene activation curve for basal clone 15.4 that corresponded to the percentage of GFP+ cells that responded to TNFα treatment (~12%) (black line). This estimated TNFα-induced value of mCherry-RelA was used to predict the fraction of GFP+ cells expected for a combination of TNFα and TSA by solving the gene activation function for 15.4 treated with TSA (blue line). (B-C) Predicted (bars) and observed (dots) percentage of GFP+ cells following stimulation with TSA+TNFα or 5-aza-dC+TNFα based on (B) gene activation functions or (C) a Bliss independence model of drug response. Experiments were performed in biological triplicate and are presented as the mean ± standard deviation. Error bars for prediction were calculated as described in Materials and Methods.

When TSA and 5-aza-dC results were combined for all clones tested (15.4, 8.4, and E3), we observed that the change in heterochromatin fraction induced by inhibitor treatment showed a strong positive correlation with the resulting change in RelA induction threshold (Fig. 6.15*C*; R = 0.78, p = 0.03). This observation further supports the correlative relationship between chromatin accessibility and the RelA level required for induction observed for clones across different integration positions (Fig. 6.10*B*). Taken together, these data demonstrate that chromatin accessibility at the HIV promoter sets a threshold for transcription factor-induced activation and that altering chromatin accessibility via multiple epigenetic pathways shifts this induction threshold.

6.7 Gene activation functions account for synergistic increases in HIV gene activation following treatment with epigenetic modifiers and TNFα

Our small molecule perturbation data demonstrated that when the heterochromatin fraction for a repressed clone (e.g. clone 15.4) is decreased via small molecule inhibitors, chromatin accessibility is increased and the gene activation curve (or function) shifts such that it responds at lower RelA levels, similar to more weakly repressed clones. Since more weakly repressed clones also respond more robustly to TNFα stimulation (Fig. 6.1*C*), we considered whether the empirically measured gene activation function, i.e. gene expression as a function of RelA, could accurately predict gene expression in response to combined HDAC inhibition and NF-κB activation. Such predictions may be relevant to HIV latency therapy, as combinatorial treatment with a HDAC inhibitor and an activator of the TNF pathway has recently been observed to result in synergistic activation for *in vitro* HIV latency models (including J-Lat and LGIT) (19,31,32).

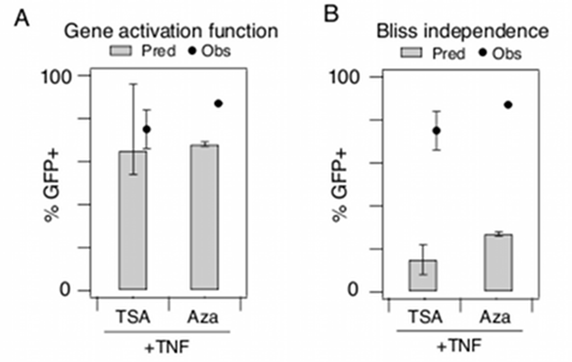


Figure 6.18. Threshold function predicts drug synergy for clone 8.4. (A-B) Predicted (bars) and observed (dots) GFP+ cells following stimulation with TSA+TNFα or 5-aza-dC+TNFα for a (A) gene activation function or (B) Bliss independence model of drug activation. Experiments were performed in biological triplicate. Data are presented as the mean ± standard deviation. Methods for prediction and error analysis are described in Materials and Methods.

To predict potential synergistic effects, we first inferred the approximate mCherry-RelA level associated with TNFα stimulation of clone 15.4 from earlier data (Fig. 6.1*C*). We then used this mCherry level and the 15.4+TSA activation function (Fig. 6.15*B*) to predict the population fraction activated in response to both TSA and TNFα (Fig. 6.17*A*). The 15.4+TSA gene activation function predicted a combined response of 71%, very close to the measured response of 68% (Fig. 6.17*B*). In contrast, when these data were used to predict combined responses under the assumption of Bliss independence (39), the expected activation in response to TNFα+TSA was 13% (i.e., 12% in response to TNFα only and 1% in response to TSA only) (Fig. 6.17*C*).

Also, the gene activation curve for 15.4+5-aza-dC accurately predicted synergistic gene activation in response to combined TNFα and 5-aza-dC stimulation (75% predicted activation versus 84% measured activation), while the Bliss independence model predicted only 41% activation (Fig. 6.17*B-C*). Gene activation functions derived for clone 8.4 treated with TSA or 5-aza-dC also predicted gene expression in response to a combination of TSA+TNFα or 5-aza-dC+TNFα more accurately than a Bliss independence model of drug response (Fig. 6.18). Our analysis collectively suggests that the predicted synergy occurred because treatment with TSA or 5-aza-dC lowered the RelA induction threshold significantly via increasing chromatin accessibility (Fig. 6.15*C*), such that TNFα-induced RelA activation resulted in a non-linear increase in population gene expression. Our prediction and observation that TSA and 5-aza-dC combine non-linearly with TNFα to stimulate gene expression is similar to the experimental synergy observed *in vitro* between activators of RelA and HDAC or DMT inhibitors in combinatorial anti-latency therapy strategies (19,31,32,38).

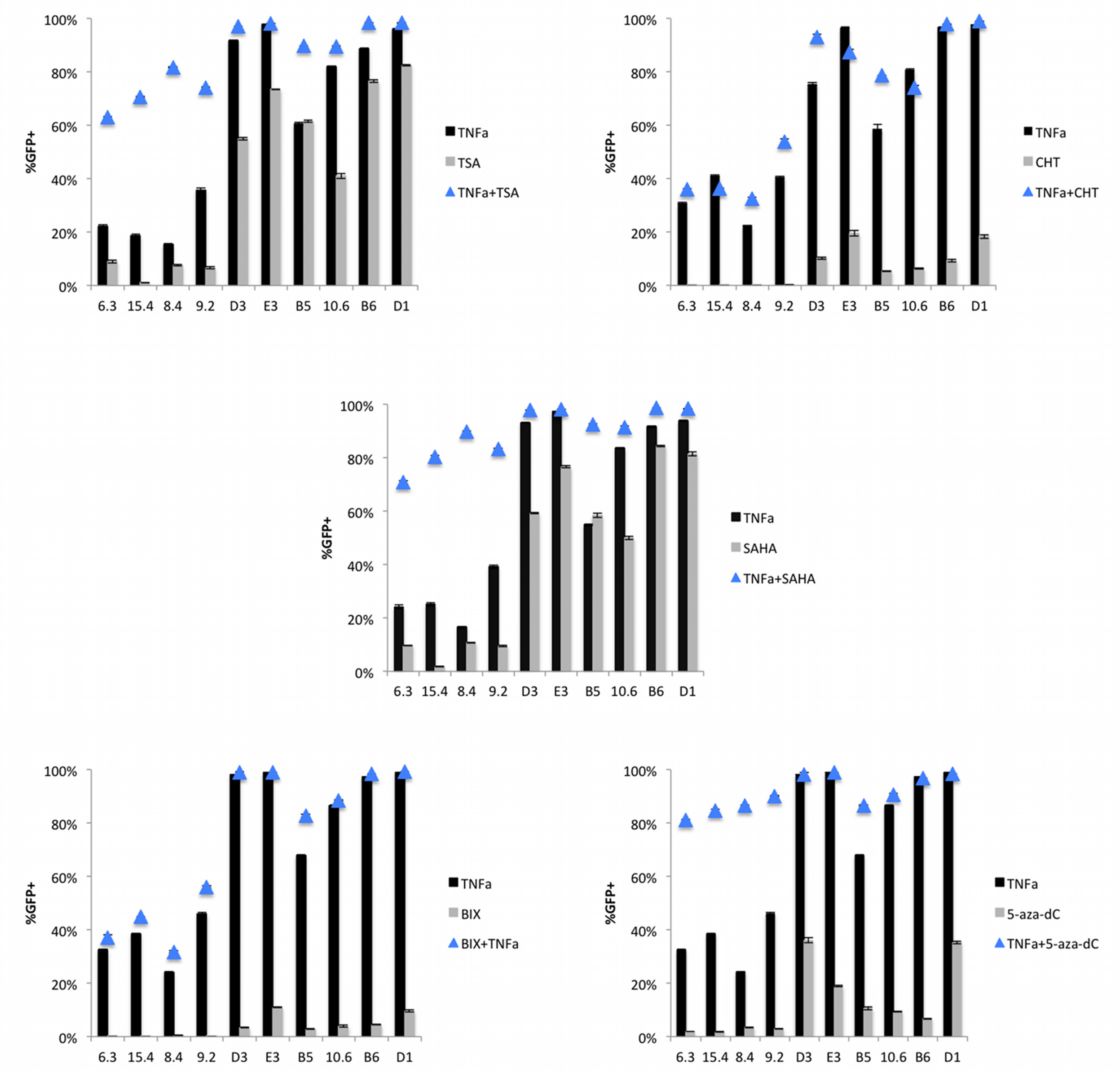


Figure 6.19. Latent HIV clones show synergistic reactivation when treated with TNFα and other small molecules. Within each panel, the level of reactivation for each clone after stimulation with either TNFα or a small molecule is indicated by bar charts. Synergistic reactivation of clones when treated with a combination of TNFα and a small molecule are shown by blue triangles. The fraction of the population that was GFP positive was monitored by flow cytometry. Flow cytometry measurements were made 24 hours after stimulation with TSA and SAHA and 48 hours after stimulation with 5-aza-dC, CHT and BIX. The TNFα and combined drug treatment measurements

were matched accordingly. Experiments were performed in biological triplicate. Data are presented as the mean ± standard deviation.

6.8 Gene activation functions show that synergistic reactivation of latent HIV clones is maximized at particular RelA concentrations

As shown in the previous section, the gene activation functions accurately predict synergistic reactivation of the latent clones, in contrast to the Bliss independence model. To estimate the existence and extent of synergy for all the clones, we treated them with a combination of TNFα and small molecules activators of gene expression. On average, we found stronger synergy when the clones were treated with a combination of TNFα and TSA (400nM), TNFα and SAHA (4μM) or TNFα and 5-aza-dC (5μM) as compared to the weaker synergy observed for stimulations with either TNFα and CHT (15nM) or TNFα and BIX (3μM) (Fig. 6.19 and Table 2).

Table 6.2: Synergy for clones used in this study between TNFα and the drugs listed in the table

Synergy	TSA	5-aza-dC	SAHA	CHT	BIX
6.3	2.17	2.39	2.25	1.16	1.13
15.4	3.61	2.14	3.04	0.88	1.16
8.4	3.76	3.23	3.53	1.45	1.30
9.2	1.85	1.90	1.85	1.32	1.22
D3	1.01	0.99	1.01	1.19	1.01
E3	0.99	1.00	0.99	0.90	1.00
B5	1.06	1.22	1.14	1.29	1.20
10.6	1.00	1.03	1.00	0.90	1.02
B6	1.01	0.99	1.00	1.01	1.01
D1	0.99	0.99	0.99	1.01	1.00

The synergies listed above are for the following drug concentrations: TSA (400nM), 5-aza-dC (5μM), SAHA (4μM), CHT (15nM) and BIX(3μM). Table shows that on average TSA, 5-aza-dC and SAHA show higher synergies than CHT and BIX for the clones used in this study.

Such synergistic reactivation of latent clones is potentially of great therapeutic value as it would permit purging out large fractions of latent viral pools from infected patients. Since the gene activation functions show non-linear increases in gene expression in response to increasing levels of RelA, we hypothesized that certain concentrations of RelA could potentially maximize the synergistic reactivation of these latent clones. To test this hypothesis we chose the drug combinations RelA and TSA (400nM) or RelA and 5-aza-dC (5μM) that showed higher synergies on average in Fig. 6.19. We determined the gene activation functions for clones 15.4, 8.4 and E3 for increasing levels of RelA in the presence or absence of the small molecules and estimated the corresponding extent of synergy at these different levels of RelA. For both small molecules TSA and 5-aza-dC, and for each of the three clones tested, we found that the synergy is maximized at a particular intermediate value of RelA (Fig 6.20). This implies that stimulating these clones at the maximal synergy inducing concentration of RelA could maximize the extent of latent viral reactivation at lower RelA concentrations.

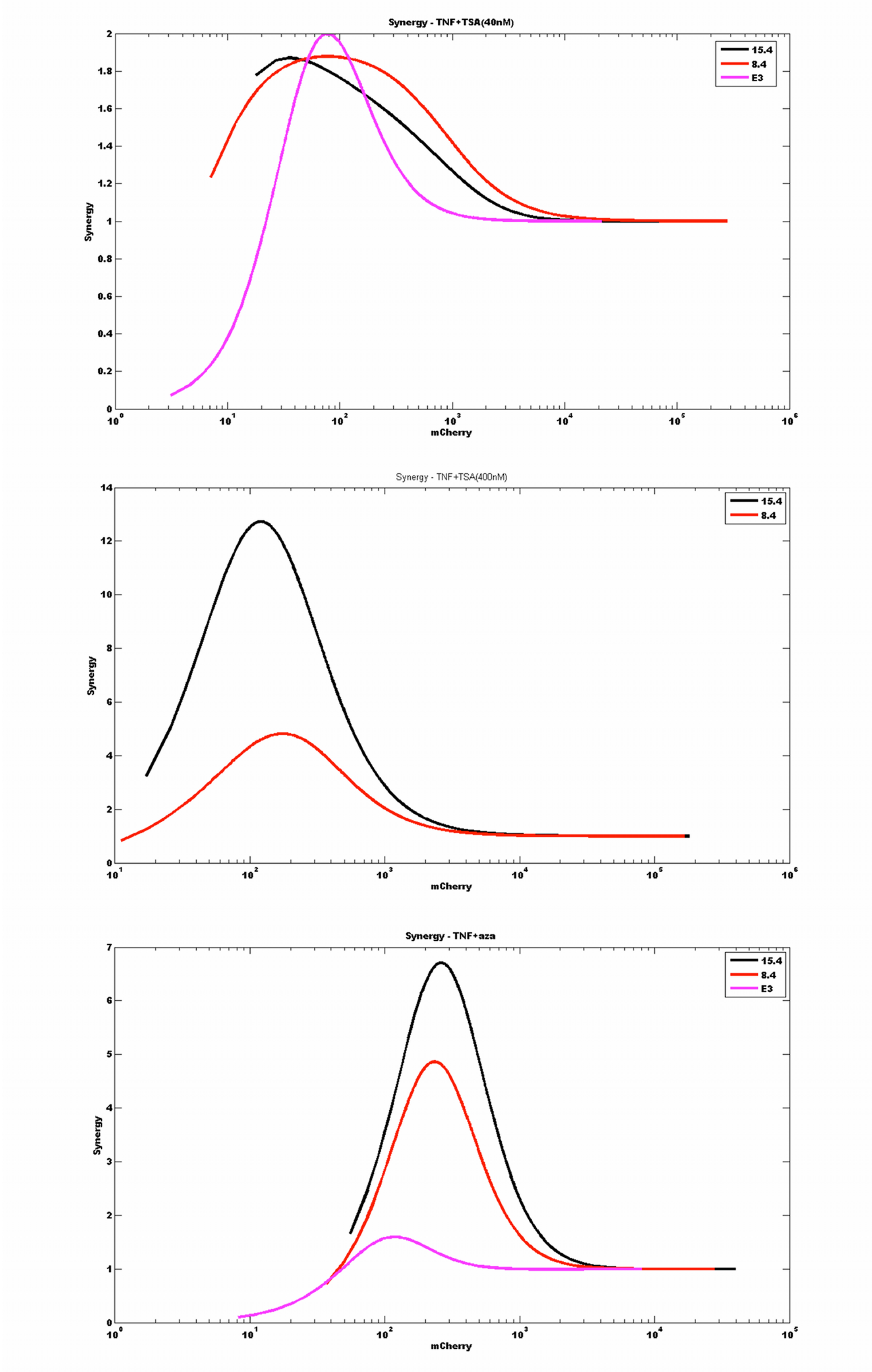


Figure 6.20. Synergistic reactivation of latent clones reveals a local optima at intermediate RelA concentrations. The iRelA clones 15.4, 8.4 and E3 were stimulated with TSA (at 40nM or 400nM) or 5-aza-dC (at 5µM) and increasing levels of RelA (by addition of DOX). The three panels show that for a given concentration of a small molecule, synergy is maximized at an intermediate concentration of RelA.

This could be of therapeutic interest as lower drug concentrations, resulting in better patient compliance could simultaneously be exploited to maximize the synergistic reactivation of latent clones. Further work will explore whether continuous reactivation of latent clones at this optimal RelA concentration is able to reactivate the same fraction of latent clones as stimulating the clones at the maximum dose of RelA.

6.9 Discussion

We have investigated how RelA level and features of the local chromatin environment quantitatively regulate the activation of HIV gene expression in a cell population. We demonstrated that gene expression is only induced when the cellular RelA level is brought above an induction threshold set by chromatin accessibility at the integration site, or conversely if chromatin accessibility is increased such that the induction threshold dips below the basal RelA levels (Figs. 6.5*D* and 6.15). A 3-D surface was constructed to incorporate and summarize data from Figs. 6.5 and 6.15 and to thereby show gene activation as a function of RelA for different genomic locations (Fig. 6.21). This functional surface – which offers the information discussed in Fig. 6.1*A* – indicates that the semi-random integration of HIV into the human genome causes it to sample a wide spectrum of chromatin environments that would lead integrated virus to respond differentially to global cellular activation, or to small molecule interventions designed to therapeutically activate gene expression.

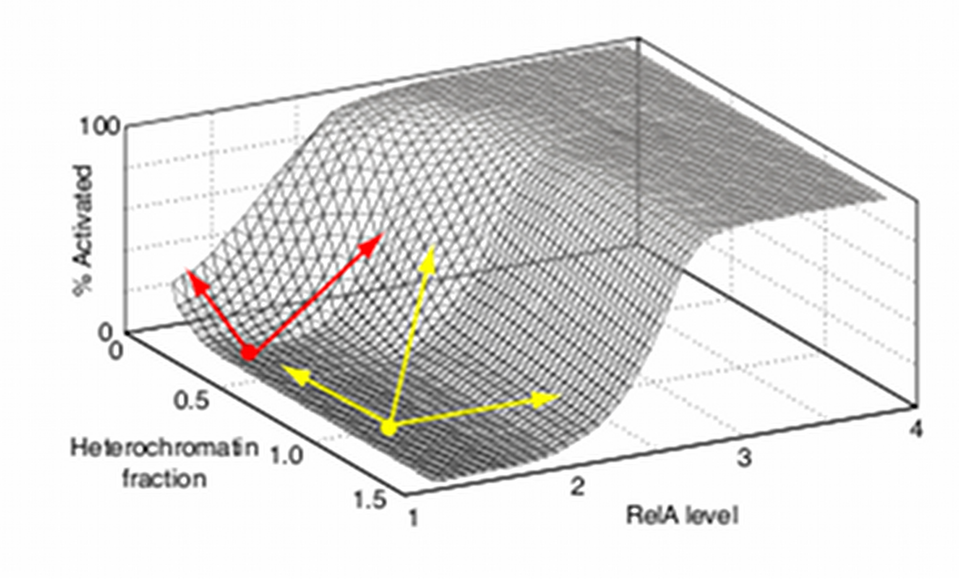


Figure 6.21. 3-D surface plot demonstrates gene activation as a function of RelA for different genomic locations. The plot was empirically derived by combining the gene activation functions for a subset of clones ranging from high to low repression. Surface plot provides a quantitative depiction of the function hypothesized in Fig. 6.1*A*. Yellow and red points and arrows describe behavior in different regimes of promoter repression. See text for discussion.

To qualitatively understand how the genomic environment of latent HIV infections may alter the response to small molecule activation, three regimes of gene expression "potential" may be considered (Fig. 6.21). In regime 1, proviruses are close to the induction threshold such that increasing either chromatin accessibility or RelA level will result in almost full activation of the population (Fig. 6.21, red). In regime 2, the level of RelA required to reach the induction threshold is sufficiently far from basal RelA such that increasing chromatin accessibility or raising RelA level alone will not be enough to activate the population, but moving along both axes will lead to activation (Fig. 6.21, yellow). Finally, it may be possible to have a promoter with sufficiently low chromatin accessibility (i.e. near the lower left corner of the functional surface) such that no combination of epigenetic modifiers and RelA activators will overcome the induction threshold and activate gene expression, though this scenario is outside the range of our experimental data. If other transcription factors that activate HIV display gene activation functions that are similar to RelA, then these infections may be difficult to activate therapeutically, but also may never result in a productive infection in activated T cells *in vivo*.

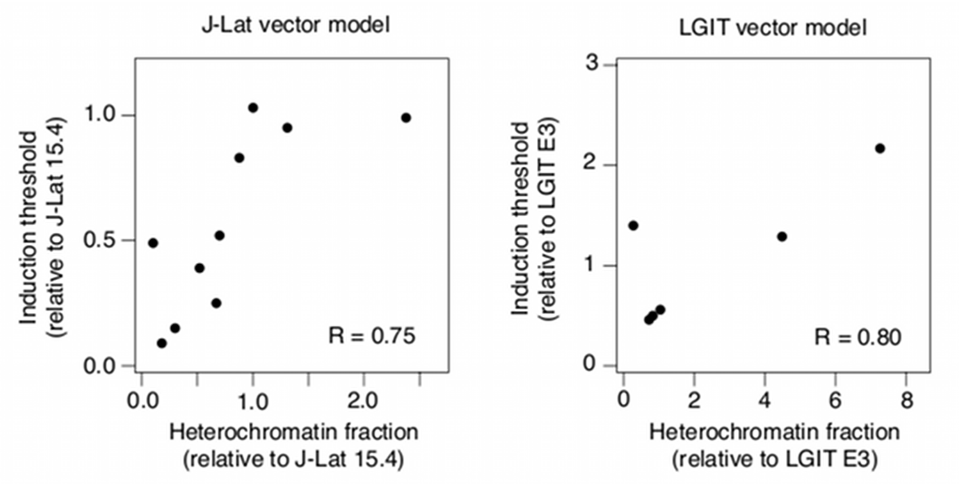


Figure 6.22. Correlation between heterochromatin fraction and RelA induction threshold is independent of HIV vector type. Data relating the induction threshold and measurements of heterochromatin fraction at basal level and in the presence of small molecule inhibitors were combined from Figs. 6.5*F*, 6.10*A*, 6.15*A-B*, and 6.16. Data for LGIT and J-Lat clones were considered separately and normalized to one clone (15.4 for J-Lat and E3 for LGIT). The correlation between heterochromatin fraction and threshold is significant for both vectors (J-Lat: Pearson R = 0.75; $p < 0.03$; LGIT: Pearson R = 0.8; $p < 0.05$).

The vectors compared in our study contain differences in sequence, Tat expression and splicing, and viral accessory proteins that could affect the threshold behavior. However, we demonstrated that Tat transcription is extremely low prior to reaching the induction threshold (Fig. 6.9). Furthermore, when measurements of chromatin accessibility and induction threshold under different conditions are separated by vector type, the strong correlation between heterochromatin fraction and induction threshold is maintained (Fig. 6.22). Although we think it is likely that each vector and selection strategy may optimally select for a particular range of chromatin environments, our data strongly support chromatin accessibility as the primary determinant of the induction threshold.

Induction thresholds set by chromatin have previously been shown in *S. cerevisiae* to be a mechanism for fine-tuning gene expression in response to transcription factors. Specifically, the affinity of the transcription factor Pho4 for its binding site in the PHO5 promoter sets a threshold for PHO5 activation by determining the level of Pho4 necessary to remodel a nucleosome positioned over the TSS (7,8). Interestingly, other transcription factor binding sites in the PHO5 promoter serve to scale expression after chromatin remodeling, suggesting that the two steps are independent. This is similar to our finding that the induction threshold for gene expression in the population is set by the local chromatin environment, but the increase in RelA-mediated gene activation and maximum fraction of activation achievable in the population is not. In the PHO5 study, the affinity of the Pho4 binding site was directly modified by introducing promoter mutations (7,8). In our study, the affinity of RelA for the κB sites on the LTR promoters is the same, and it is instead the chromatin accessibility at the site of integration that tunes the level of cellular RelA required for sufficient chromatin remodeling to activate gene expression. Further measurements are needed to determine if RelA binding to the promoter is directly or indirectly affected by changes in the affinity of nucleosomes for the LTR promoter.

The more general idea that chromosomal location modulates gene expression has been increasingly investigated since the study of position effect variegation (40). Our results explore how chromatin context quantitatively impacts activation by a single transcription factor (TF) input, and suggest that chromatin environment within the mammalian genome can threshold the activation of different genes to the same TF, without significantly affecting the TF-mediated expression after gene expression is induced in the population. Such a mechanism potentially contributes to observed differential activation of genes in response to proinflammatory stimuli (27), where stimulation by proinflammatory cytokines resulted in two waves of NF-κB recruitment to target genes – early and late – that are primarily differentiated by the chromatin configuration at the promoter and not the affinity of the binding site (3). Our analysis was performed at steady-state but could be extended to examine the role of a chromatin threshold in the dynamics of NF-κB recruitment and gene activation.

A recent genome-wide study of glucocorticoid receptor (GR) binding demonstrated that for a large majority of GR binding motifs, cell-specific differences in pre-existing patterns of chromatin accessibility at GR binding sites were a primary determinant of cell-selective GR occupancy, leading to cell-specific gene expression patterns (34). Our results also show that chromatin accessibility prior to stimulation plays a major role in determining NF-κB-mediated gene expression from the LTR, and thus appear to support an emerging general mechanism of how chromatin modulates TF–gene interaction specificity in diverse biological systems. Because TF binding in response to exogenous stimuli underlies all biological processes, a quantitative understanding of how these interactions are regulated by the local chromatin environment are important to decipher input-output responses of a cell.

6.10 Materials and Methods

6.10.1 Plasmids.

LGIT has been previously described (12). The inducible RelA (iRelA) vector was based on a single lentiviral vector platform for tetracycline-regulated expression of the product (33). The mCherry fluorescent protein was fused to the N-terminus of RelA by splice overlap PCR (41) and then cloned into the pEN-Tmcs (ATCC). The pEN-Cherry-RelA fusion plasmid was cloned into the pSLIK-Venus plasmid (ATCC) by LR recombination reaction (Invitrogen) as previously described (33), and the IRES-Venus sequence was removed. Cloning details and the final plasmid map is available upon request.

6.10.2 Cell culture.

Jurkat cells and HEK 293T cells (used for lentiviral packaging) were cultured as previously described (26). LGIT clones were sorted and cultured as previously described (26). J-Lat full length clones (13) were obtained from the laboratory of Dr. Eric Verdin via the NIH AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH.

6.10.3 Viral harvesting and infection of iRelA cell lines.

Lentiviral vectors were packaged as previously described (42). For infection with the iRelA vector, 3×10^5 LGIT and J-Lat clones were grown in 12-well plates and infected at a multiplicity of infection of 0.6. Four days later, infected cells were stimulated with 1 µg/ml doxycyclin for 48 hours and the top quartile of the mCherry-expressing population was sorted on a Cytopeia InFlux cell sorter (BD Biosciences). The sorted iRelA cell lines populations were expanded and frozen stocks were stored in liquid nitrogen.

6.10.4 Drug Stimulation.

The LGIT and J-Lat cell lines and the corresponding iRelA cell lines were treated with the following pharmacological agents for the indicated times and analyzed by flow cytometry: TNF-α at 20 ng/mL (24 or 48 hours post-stimulation), TSA at 40 nM or 400 nM (24 hours), and 5-aza-dC at 5 µM (48 hours). For the iRelA cell line stimulations, cells were treated with DOX at 0, 10, 30, 100 or 300 ng/mL or as indicated in the text.

6.10.5 Fitting the gene activation functions.

For each iRelA cell line, flow cytometry data collected from DOX stimulation at 0, 10, 30, 100 and 300 ng/mL were combined. The data were binned into 256 GFP and mCherry channels. For each mCherry channel, the percentage of GFP+ cells was computed and plotted, as in Fig. 6.5*D*. The Hill equation was log transformed into a linear equation and the curves in Fig. 6.5*D* were fit by least squares as shown in Fig. 6.7. The quality of the fits did not improve by changing the number of bins. The slope and intercept obtained from the least squares regression was used to compute the threshold and activation (Hill) coefficient (Fig. 6.5*E-F*). iRelA cell lines stimulated with chromatin modifying enzymes and DOX were analyzed similarly. Standard deviations for the threshold and activation coefficient were bootstrapped using 1000 bootstrapped data samples.

6.10.6 Western blotting.

J-Lat 6.3 cells were treated with DOX at the indicated concentrations for 4 days. Cells were pelleted and resuspended in lysis buffer containing IGEPAL (1%; Sigma), sodium dodecyl sulfate (SDS) (0.1%), phenylmethanesulfonylfluoride (0.1 mg/mL; Sigma), aprotinin (0.03 mg/mL; Sigma), and sodium orthovanadate (1mM; Sigma) in PBS. Lysate protein concentrations were quantified by BCA Protein Assay Kit (Pierce) according to manufacturer's instructions. 10 µg of protein from each lysate were electrophoretically separated by SDS-PAGE and transferred to nitrocellulose membranes (Bio-Rad Laboratories). Membranes were probed with anti-NF-κB p65 (C-20) primary antibody (Santa Cruz Biotechnology, sc-372) and horseradish peroxidaseconjugated goat anti-rabbit secondary antibody (Pierce, 31460), developed with ECL Plus (Pierce), and analyzed on the Versadoc 4000 imager (Bio-Rad).

6.10.7 Nuclease sensitivity assay.

The nuclease sensitivity assay was performed using the EpiQ™ Chromatin Analysis Kit (Bio-Rad) with minor modifications of the manufacturer's protocol. Briefly, 250,000 cells were incubated with DNAse I for 1 hour. Enzyme concentrations were adjusted to account for the range of nuclease sensitivities being compared (1X for measuring drug response in LGIT E3 and 3X for measuring basal chromatin across clones and drug response in J-Lat clones). Following extraction and purification of the genomic DNA, the level of HBB and LTR were quantified by qPCR. Primers were designed to prime within the DNase hypersensitive site located inside the core promoter and cover the binding site of nucleosome-1, a nucleosome whose remodeling is associated with activation of the latent promoter(21,22). See Supplementary Methods for sequences.

6.10.8 Chromatin immunoprecipitation.

Upstate EZ ChIP Kit Reagents (Upstate) and protocols were used with minor modifications. 10 million cells were fixed in 1% formaldehyde for 10 minutes, and the unreacted formaldehyde was quenched using 125 mM glycine for 10 minutes on ice. After extensive PBS washing, the cells were lysed with 1 mL of 1% SDS lysis buffer in the presence of a protease inhibitor cocktail. For the Ser5P-CTD of RNAPII ChIP, a phosphatase inhibitor cocktail was also added during the immunoprecipitation step. The cells were sonicated either using the Branson Sonifier 450 (Settings: 25 cycles at a power output of 2.5 and duty cycle of 25%. Each cycle consisted of 15 pulses followed by incubation on ice for at least 1 minute) or the Misonix Sonifier 3000 (Settings: 7 cycles at a power output of 4. Sonication was done for 30 sec in each cycle with 1 sec ON/OFF pulses, followed by incubation on ice for at least 1 min). DNA gel electrophoresis was used to verify that the sheared DNA fragments were within 0.1-1 kb. For the Ser5P-CTD of RNAPII ChIP, anti-mouse IgM agarose beads (Sigma) were used instead of Protein A or G beads that were used for the other ChIPs. The anti-mouse IgM agarose beads were washed extensively with RIPA buffer, then blocked with salmon sperm DNA and yeast tRNA. For the Ser5P-CTD of RNAPII ChIP, the beads were incubated with the antibody-chromatin complex for 5 hours at 4^{o}C. For all other ChIPs, the beads were incubated with the antibody-chromatin complex for 2 hours at 4^{o}C. The precipitated DNA was quantified using quantitative PCR (BioRad iCycler, iQ5) using the EpiQ Chromatin SYBR Supermix. qPCR was performed in triplicate and melt curves were run to ensure product specificity. The following antibodies were used in the immunoprecipitation step: anti-RNAPII (Millipore, Catalog # 05-623), anti-Ser5P CTD RNAPII (Covance, Catalog # MMS-134R), anti-histone H3 (Abcam, Catalog # ab1791), anti-acetyl histone H3 (Millipore, Catallog # 06-599), anti-histone H3K9me3 (Abcam, Catalog # ab8898). The following primers were used for the ChIP for AcH3, H3K9me3, and total H3 and the nuclease sensitivity assay: 5'-GGACTTTCCGCTGGGGACTTTCCAGGG-3' (forward) and 5'-GCGCGCTTCAGCAAGCCGAGTCCTGCGTCGAG-3' (reverse). Alternate primers were used for the ChIP for ChIP for RNAPII and phospho-Ser5-RNAPII: 5'-GACTTTCCGCTGGGGACTTTC-3' (forward) and 5′-GTGGGTTCCCTAGTTAGCCA-3′ (reverse).

6.10.9 Imaging protocol and analysis.

Clone 6.3 and B6 infected with iRelA were treated with 0, 30, and 300 ng/ml DOX for 4 days. One million cells per condition were washed twice with PBS, fixed with 4% Formaldehyde (F79-1, Fisher Scientific), and applied to the well of a glass bottom 6-well plate (P06G-1.0-20, MatTek Corp., Ashland,MA) treated with 0.01 mg/ml Poly-L-Lysine Solution (SDP8920A, Fisher Scientific) to promote cell adhesion. After 30 minutes, wells were washed twice with PBS and stored in 70% Ethanol at 4° C overnight before imaging. Wells were rehydrated twice for 15 minutes with PBS. Nuclei were stained with 0.0025 mg/ml DAPI (D1306, Invitrogen Corp. Carlsbad,CA) for 10 minutes and washed with PBS, and treated with an anti-bleach solution consisting of 10 mM Tris pH 8.0, 2xSSC, 0.4% glucose with 0.037 mg/ml glucose oxidase (G2133, Sigma-Aldrich Corp.) and 0.05 mg/ml catalase (C3515,Sigma-Aldrich Corp.) prior to applying the cover slip. Wells were imaged using an automated imaging system (ImageExpress Micro, Molecular Devices Inc.) with a 40X objective. Briefly, a 20x20 grid of independent fields was established in software per well and fields were imaged with hardware autofocus and a standard FITC, TexasRed, DAPI filter set. Exposure times were determined empirically to maximize signal to noise and prevent camera saturation. CellProfiler (Carpenter Genome Biology 2006) with a custom pipeline was used to segment cells and nuclei and to determine total and localized GFP and mCherry. The MeasureImageQuality module was used to reject significantly blurry fields using an empirically determined Focus Score of 0.004.

6.10.10 RNA Extraction and Quantification of Viral transcripts.

The indicated cell lines indicated were stimulated for 24 hours at different concentrations of DOX and treated with Trizol (Invitrogen) to extract total cellular RNA. Viral and cellular mRNA were quantified using the Quantitect SYBR Green RT-PCR kit (Qiagen) and a Bio-Rad iCycler (iQ5). The following primers were used to quantify Tat transcripts: Tat-F (5'-GCATCCAGGAAGTCAGCCT-3') and Tat-R (5'-CTCCGCTTCTTCCTGCCATAG-3'). B-Actin was used as a control and quantified using the primers, β-Actin-F (5'-ACCTGACTGACTACCTCATGAAGATCCTCACCGA-3') and B-Actin-R (5'-GGAGCTGGAAGCAGCCGTGGCCATCTCTTGCTCGAA-3'). qPCR was performed in triplicate and the error bars represent standard deviations from the mean.

6.10.11 Combinatorial drug predictions.

The GFP+ fraction activated by TNFα+TSA or TNFα+5-aza-dC according to the model of Bliss independence ($\mu_{\mathrm{TNF+inh,BLISS}}$) was calculated as follows: $\mu_{\mathrm{TNF+inh,BLISS}} = 1 - (1-\mu_{\mathrm{TNF}})*(1-\mu_{\mathrm{inh}})$ where μ_{TNF} and σ_{TNF} and μ_{inh} and σ_{inh} are the mean and standard deviation of the GFP+ fraction activated by TNFα and by TSA or 5-aza-dC, respectively. For predictions using the gene activation functions, for each clone of interest we located the point on the basal gene activation curve that corresponded to $\mu_{\mathrm{TNF}} \pm \sigma_{\mathrm{TNF}}$ and used this to estimate the approximate mCherry-RelA increase, $n_{\mathrm{TNF\text{-}RelA}} \pm e_{\mathrm{TNF\text{-}RelA}}$ associated with TNFα treatment alone (where $e_{\mathrm{TNF\text{-}RelA}}$ is the uncertainty in mCherry-RelA associated with σ_{TNF}). Finally, $\mu_{\mathrm{TNF+inh,GA}} \pm \sigma_{\mathrm{TNF,GA}}$ was calculated by solving the empirical gene activation function for the clone of interest in the presence of drug treatment (clone+TSA or clone+5-aza-dC) at the point $n_{\mathrm{TNF\text{-}RelA}} \pm e_{\mathrm{TNF\text{-}RelA}}$.

6.10.12 Statistical analysis.

We used Student's t-test to compare two means, and two-factor ANOVA to compare heterochromatin fraction across different clonal groups. Significance of Pearson correlation coefficients was calculated according to the following formula for the t statistic: $t = r*[(1-r^2)/(n-2)]^{-1/2}$ where r is the Pearson correlation coefficient and n is the sample size.

6.11 References

1. Segal, E., and Widom, J. (2009) *Nat Rev Genet* **10**, 443-456
2. MacQuarrie, K. L., Fong, A. P., Morse, R. H., and Tapscott, S. J. (2011) *Trends Genet* **27**, 141-148
3. Saccani, S., Pantano, S., and Natoli, G. (2001) *J Exp Med* **193**, 1351-1359
4. Smale, S. T. (2010) *Cell* **140**, 833-844
5. Chambeyron, S., and Bickmore, W. A. (2004) *Genes & Development* **18**, 1119-1130
6. Portela, A., and Esteller, M. (2010) *Nat Biotechnol* **28**, 1057-1068
7. Lam, F. H., Steger, D. J., and O'shea, E. K. (2008) *Nature* **453**, 246-250
8. Kim, H. D., and O'shea, E. K. (2008) *Nat Struct Mol Biol* **15**, 1192-1198
9. Kundu, S., and Peterson, C. L. (2009) *Biochim Biophys Acta* **1790**, 445-455
10. Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V., and Arkin, A. P. (2010) *PLoS Comput Biol* **6**
11. Miller-Jensen, K., Dey, S. S., Schaffer, D. V., and Arkin, A. P. (2011) *Trends in Biotechnology*
12. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., and Schaffer, D. V. (2005) *Cell* **122**, 169-182
13. Jordan, A., Bisgrove, D., and Verdin, E. (2003) *EMBO J* **22**, 1868-1877
14. Lassen, K., Han, Y., Zhou, Y., Siliciano, J., and Siliciano, R. F. (2004) *Trends Mol Med* **10**, 525-531
15. Kim, Y. K., Bourgeois, C. F., Pearson, R., Tyagi, M., West, M. J., Wong, J., Wu, S.-Y., Chiang, C.-M., and Karn, J. (2006) *EMBO J* **25**, 1-9
16. Barboric, M., Yik, J. H. N., Czudnochowski, N., Yang, Z., Chen, R., Contreras, X., Geyer, M., Matija Peterlin, B., and Zhou, Q. (2007) *Nucleic Acids Research* **35**, 2003-2012
17. Williams, S. A., Chen, L.-F., Kwon, H., Ruiz-Jarabo, C. M., Verdin, E., and Greene, W. C. (2006) *EMBO J* **25**, 139-149
18. Pearson, R., Kim, Y. K., Hokello, J., Lassen, K., Friedman, J., Tyagi, M., and Karn, J. (2008) *Journal of Virology* **82**, 12291-12303
19. Blazkova, J., Trejbalova, K., Gondois-Rey, F., Halfon, P., Philibert, P., Guiguen, A., Verdin, E., Olive, D., Van Lint, C., Hejnar, J., Hirsch, I., and Hope, T. J. (2009) *PLoS Pathog* **5**, e1000554
20. Hakre, S., Chavez, L., Shirakawa, K., and Verdin, E. (2011) *Current Opinion in HIV and AIDS* **6**, 19-24
21. Verdin, E., Paras, P., and Van Lint, C. (1993) *EMBO J* **12**, 3249-3259
22. el Kharroubi, A., and Verdin, E. (1994) *J Biol Chem* **269**, 19916-19924
23. Gatignol, A., Buckler-White, A., Berkhout, B., and Jeang, K. T. (1991) *Science* **251**, 1597-1600

24. Feinberg, M. B., Baltimore, D., and Frankel, A. D. (1991) *Proc Natl Acad Sci USA* **88**, 4045-4049

25. Weinberger, L. S., Dar, R. D., and Simpson, M. L. (2008) *Nat Genet* **40**, 466-470

26. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P., and Schaffer, D. V. (2009) *PLoS Pathog* **5**, e1000260

27. Natoli, G. (2009) *Cold Spring Harb Perspect Biol* **1**

28. Duh, E. J., Maury, W. J., Folks, T. M., Fauci, A. S., and Rabson, A. B. (1989) *Proc Natl Acad Sci U S A* **86**, 5974-5978

29. Barboric, M., Nissen, R. M., Kanazawa, S., Jabrane-Ferrat, N., and Peterlin, B. M. (2001) *Mol Cell* **8**, 327-337

30. Chan, J. K., and Greene, W. C. (2011) *Curr Opin HIV AIDS* **6**, 12-18

31. Reuse, S., Calao, M., Kabeya, K., Guiguen, A., Gatot, J.-S., Quivy, V., Vanhulle, C., Lamine, A., Vaira, D., Demonte, D., Martinelli, V., Veithen, E., Cherrier, T., Avettand, V., Poutrel, S., Piette, J., de Launoit, Y., Moutschen, M., Burny, A., Rouzioux, C., De Wit, S., Herbein, G., Rohr, O., Collette, Y., Lambotte, O., Clumeck, N., and Van Lint, C. (2009) *PLoS ONE* **4**, e6093

32. Burnett, J. C., Lim, K.-I., Calafi, A., Rossi, J. J., Schaffer, D. V., and Arkin, A. P. (2010) *Journal of Virology* **84**, 5958-5974

33. Shin, K.-J., Wall, E. A., Zavzavadjian, J. R., Santat, L. A., Liu, J., Hwang, J.-I., Rebres, R., Roach, T., Seaman, W., Simon, M. I., and Fraser, I. D. C. (2006) *Proc Natl Acad Sci USA* **103**, 13759-13764

34. John, S., Sabo, P. J., Thurman, R. E., Sung, M.-H., Biddie, S. C., Johnson, T. A., Hager, G. L., and Stamatoyannopoulos, J. A. (2011) *Nat Genet* **43**, 264-268

35. Wu, C., Bingham, P. M., Livak, K. J., Holmgren, R., and Elgin, S. C. (1979) *Cell* **16**, 797-806

36. Boyle, A. P., Davis, S., Shulha, H. P., Meltzer, P., Margulies, E. H., Weng, Z., Furey, T. S., and Crawford, G. E. (2008) *Cell* **132**, 311-322

37. Brown, K. E., Amoils, S., Horn, J. M., Buckle, V. J., Higgs, D. R., Merkenschlager, M., and Fisher, A. G. (2001) *Nat Cell Biol* **3**, 602-606

38. Kauder, S. E., Bosque, A., Lindqvist, A., Planelles, V., and Verdin, E. (2009) *PLoS Pathog* **5**, e1000495

39. BLISS, C. I. (1956) *Bacteriol Rev* **20**, 243-258

40. Henikoff, S. (2008) *Nat Rev Genet* **9**, 15-26

41. Heckman, K. L., and Pease, L. R. (2007) *Nat Protoc* **2**, 924-932

42. Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., and Naldini, L. (1998) *J Virol* **72**, 8463-8471

Chapter 7: Chromatin Features at the HIV-1 Promoter Regulate Transcriptional Noise

7.1 Introduction

Every cell population exhibits differences among individual cells, even when the cells are genetically identical and the environment is carefully controlled. While non-genetic heterogeneity might arise from varied sources, some fraction of it continuously arises from ‘biochemical noise’: random fluctuations in molecular concentrations and biochemical reactions that affect cellular mechanisms. Biochemical noise is especially apparent in gene expression within individual cells, because genes are usually present in very low numbers (typically 1-2 copies per cell). Gene expression is thus a fundamentally noisy process that can result in non-genetic heterogeneity in both prokaryotic and eukaryotic cell populations (1,2).

Stochastic fluctuations amplified by biological mechanisms such as regulatory circuits, can give rise to phenotypic heterogeneity (3). Such phenotypic heterogeneity arising from cell fate decisions driven by stochastic gene expression is emerging as a persistence mechanism in diverse mammalian diseases. For example, recent evidence suggests that biological noise may underlie probabilistic entry into and exit from mammalian viral latency (4), in which a subset of viruses establish “silent” infections that may permit viruses to evade the host immune system and reactivate later to produce more progeny (5). In a very different example, cell-to-cell variability in the proteome of cancer cells appears to permit a small population of “persister” cells to survive chemotherapy (6). Eukaryotic genes are subject to complex mechanisms of chromatin regulation mediated by transcription factors and chromatin modifying enzymes that modulate stochastic fluctuations in gene expression (7). Thus, chromatin may provide a mechanism for varying the probability of stochastic transitions between phenotypes, and possibly increase the stability of one phenotype versus another.

It is well documented that intrinsically noisy gene expression results, in large part, from bursts of transcript and protein production in a number of cellular systems (2). In prokaryotes, such noise is primarily attributed to translational bursts that occur when ribosomes generate many proteins from a single transcript (1,7,8). In contrast, noise in eukaryotic cells primarily arises from transcriptional bursts, which are compatible with a model in which the promoter infrequently transitions between an inactive and an active gene state (9-11). Transcriptional bursting has been studied most extensively in yeast, but there is also evidence of such bursting in mammalian cells (10,12). Furthermore, cell-to-cell variability in transcript and protein levels in human cells is consistent with a stochastic gene state transition model (12-14).

The gene state transition model is widely accepted; however, the source of transcriptional bursts is incompletely understood. One hypothesis that has gained considerable traction is that stochastic events in nucleosome remodeling cause the infrequent transitions between an inactive and active gene state, and thus underlie transcriptional bursting (7). Nucleosomes are the fundamental unit of chromatin, consisting of ~147 base pairs of DNA wrapped around an octamer of the four core histone proteins. Transcription factors must compete with nucleosomes for binding to the DNA, and therefore nucleosomes are considered to be general repressors of transcription (15,16). ATP-dependent chromatin remodeling enzymes periodically move or disassemble nucleosomes along the DNA, which "opens" the chromatin to favor transcription factor binding and gene activation (17).

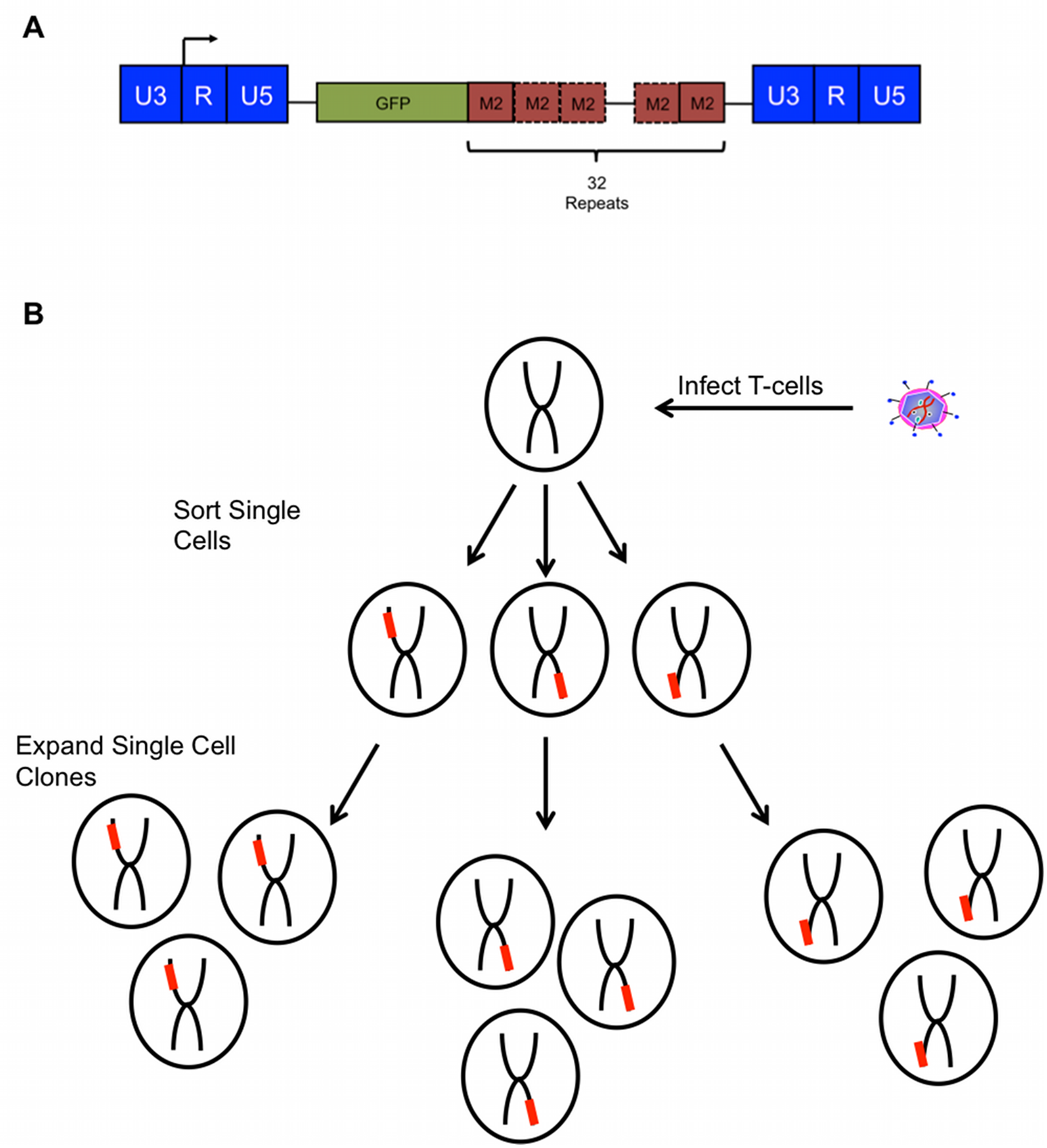


Figure 7.1. Experimental setup to study noise in gene expression from the HIV-1 promoter. (A) Lentiviral vector LGM2 used to quantify noise in protein and mRNA levels from the HIV-1 promoter. Green fluorescent protein (GFP) is used to quantify cell-to-cell variability in protein levels. 32 repeats of the M2 array at the 3' end of the transcript allows for hybridization of fluorescent probes allowing for the detection of single mRNA transcripts as diffraction limited spots. (B) Schematic representation for the isolation of Jurkat cell clones infected with a single copy of the LGM2 vector. Jurkat cells were infected with the LGM2 lentiviral construct at low MOIs of 0.05-0.1 to ensure single viral integrations per cell. The cells were stimulated with TNFα 7 days post-infection and GFP+ cells

were sorted. After allowing for the sorted cells to relax, single cells were sorted into 96-well plates. These single-cells were expanded to produce clonal populations, each clone containing a unique integration site for the virus.

Some of the first experimental studies to directly measure transcriptional bursts showed that the probability of gene activation varied with chromosomal position, which strongly suggested a link between chromatin remodeling and promoter state transitions (10,18). Recent genome-wide studies in yeast further demonstrated that increased variability in gene expression is positively correlated with nucleosome density close to the transcriptional start site (19-21). Yeast genes with higher expression noise were also more sensitive to perturbation of chromatin regulators, suggesting that noisy genes are subject to chromatin remodeling (19,20).

In the context of HIV-1, the choice between replication and latency, a decision with substantial consequences for human health, is an example of a heterogeneous fate decision that may result from stochastic gene expression (4). Following infection and integration in CD4+ T lymphocytes, HIV-1 usually actively replicates in the cell, but on rare occasions it fails to establish a productive infection and enters a latent state (22). Latent HIV-1 proviruses are highly stable and persist even in patients on highly active anti-retroviral therapy (23). Upon activation of the host T cell, latent virus can reactivate and re-seed viremia, and for this reason, patients must continuously take anti-viral therapy. Consequently, HIV-1 latency is the most significant barrier to curing viral infection (24).

The virally encoded transcriptional activator Tat is essential for establishing a productive infection and for reactivating latent virus. Tat is transcribed early during HIV-1 infection and significantly amplifies expression from the HIV-1 LTR promoter in a strong positive feedback loop. However, when Tat protein levels are low, such as just after infection or before reactivation, stochastic fluctuations in Tat gene expression can lead to delays before activation of the Tat-mediated positive feedback loop, resulting in subpopulations of low and high gene expression in a genetically-identical population of cells (25). In this case, stochastic gene expression noise coupled with a strong positive feedback loop operates as a genetic "switch" that regulates entry and exit from latency (26). If viral replication is sufficiently delayed, other cellular factors mediate chromatin changes that further suppress viral transcription and maintain (and further stabilize) the latent state (27,28).

While it has been shown that the HIV-1 promoter is noisy, it is unclear what the source of this noise is and ways in which it may influence the replication-versus-latency decision. Consistent with studies of stochastic fluctuations in eukaryotic gene expression discussed previously, our group and others recently demonstrated that a two-state model of transcriptional bursting could account for HIV-1 gene expression variance in the absence of Tat feedback (Fig. 2c) (12,29). A two-state bursting model of HIV-1 gene expression is consistent with the long-standing knowledge that nucleosomes are positioned at the HIV-1 transcriptional start site, and the observation that, in the absence of Tat, the HIV-1 promoter binds repressive factors that maintain an inactive chromatin configuration (30). Because the HIV-1 LTR promoter also contains binding sites for activating factors, binding competition with repressive factors could lead to an infrequent all-or-none binding of activating factors that directly remodel promoter-bound nucleosomes to establish a short-lived transcriptionally active chromatin configuration (31). Since features of the HIV-1 promoter that account for its heterogeneous expression pattern are likely present in other mammalian promoters, however, the development of new experimental methods to investigate its gene expression properties may yield general insights into mammalian transcriptional regulation.

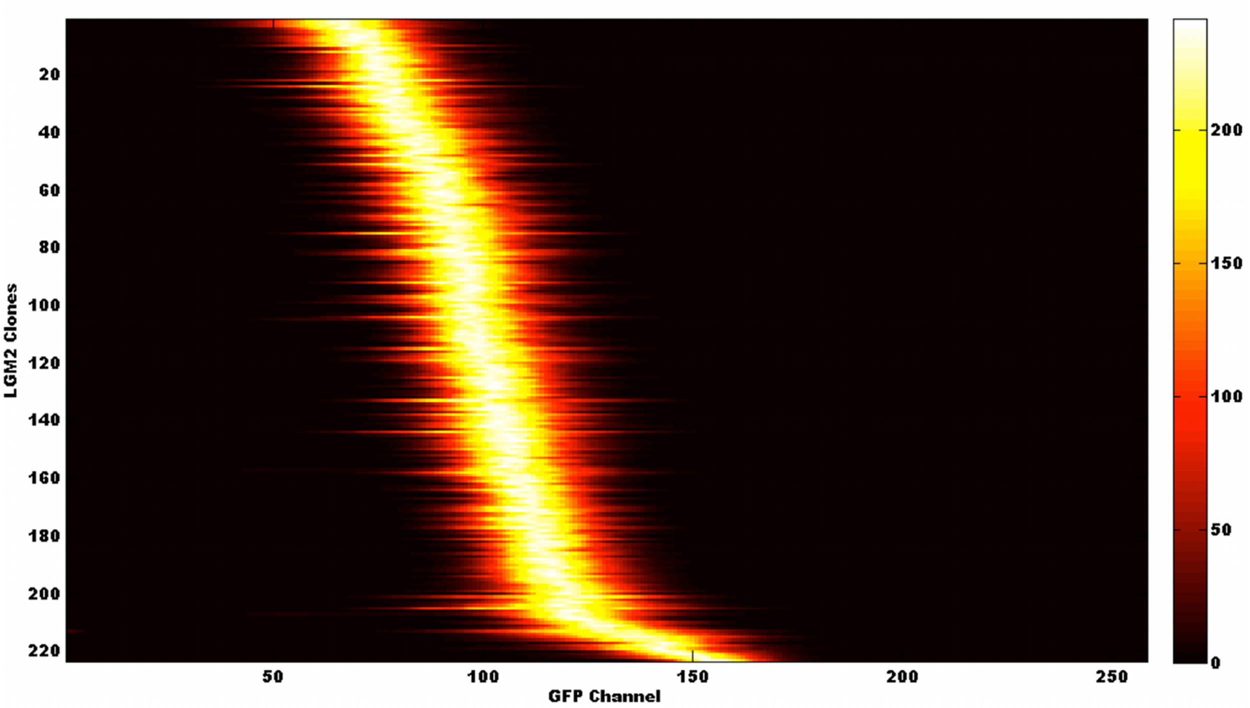


Figure 7.2. Heat map showing GFP distribution for 223 clones. The GFP distribution of each clones is represented in each row of the heat map. The clones are arranged in the order of increasing mean GFP expression. The color-coding indicates the number of cells in each GFP channel. The spikes in the heat map show that clones with different integration sites can have large variability in the variance of the GFP distribution.

In this work, we directly measure and quantify transcriptional bursting from the HIV-1 promoter, using Fluorescence *In Situ* Hybridization (FISH). RNA distributions for each clonal population allowed for the accurate estimation of the parameters of the two-state model. Further, we investigated the origins of transcriptional noise by studying clones that had similar mean levels of gene expression but different noise characteristics. We found systematic differences in the chromatin environment between high and low noise clones, thereby providing a mechanistic understanding of heterogeneity in a clonal population.

7.2 Quantifying cell-to-cell variability in the level of protein production from the HIV-1 promoter

To quantify variability in gene expression from the HIV-1 promoter, we created a vector in which GFP was used to quantify the level of protein expression from the full-length viral promoter. 32 tandem repeat oligonucleotides, denoted by M2, were added to the 3’ end of the transcribed mRNA for binding fluorescent probes to provide single mRNA detection resolution (10). This vector was denoted as LTR-GFP-M2 or LGM2 (Fig 7.1*A*). To obtain single integrations of the lentiviral construct, Jurkat cells were infected with LGM2 at low MOIs around 0.05-0.1. Cells were stimulated with TNFα 7 days post infection and GFP+ cells were sorted. The sorted cells were allowed to relax for a few days and single cells were thereafter sorted into 96-well plates. These single cells, each with a single unique integration site for LGM2 with the host genome, were expanded to produce 223 clonal populations (Fig 7.1*B*).

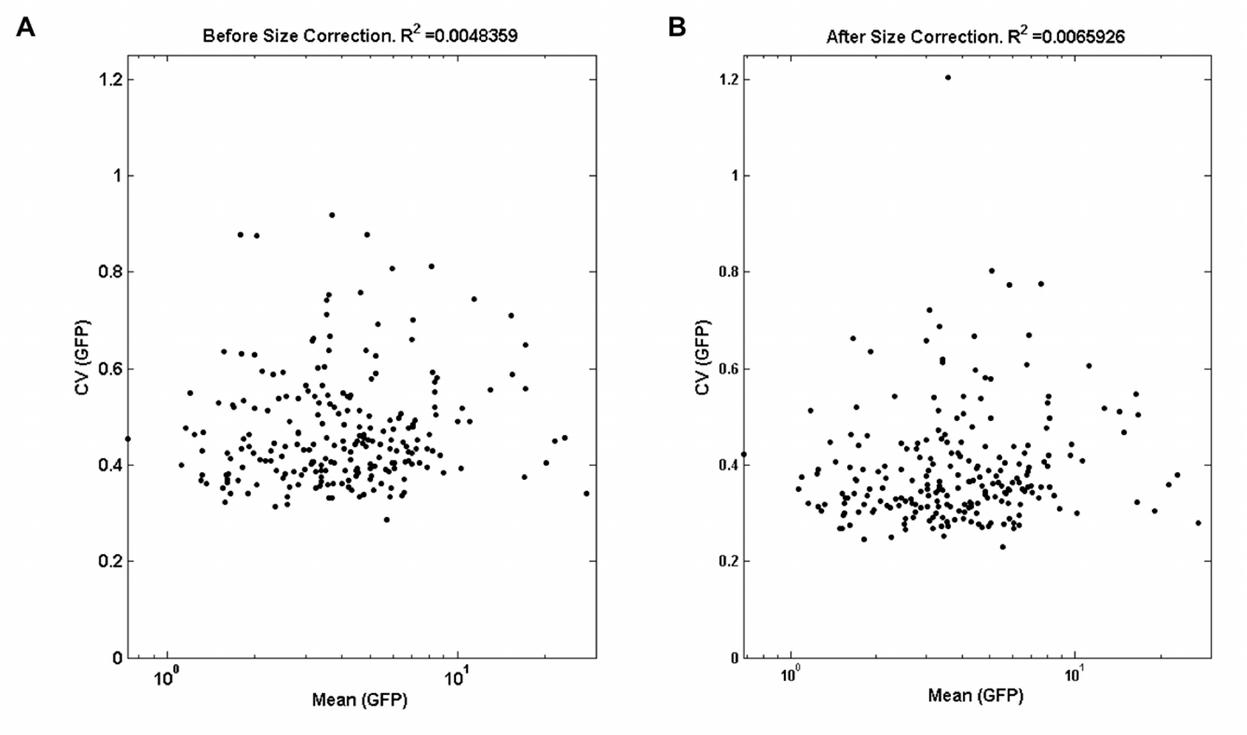


Figure 7.3. Clones show large variability in the level of noise in gene expression. (A) Coefficient of Variation (CV), a measure of noise in gene expression is plotted against the mean GFP expression for each clone. These moments of the GFP distribution are computed using 10,000 live cells. Data shows that the CV of the distribution is uncorrelated to the mean GFP expression level. (B) To minimize other sources of noise arising from cell size and shape, around 60% of the cells around the mean Forward and Side scatter were used to re-estimate the GFP distributions and the moments of the distribution. Figure shows that the CV is still uncorrelated to the mean.

Each clonal population had a distinct GFP distribution (Fig 7.2). In the heat map (Fig. 7.2), each row corresponds to the GFP distribution of a clone. The clones are arranged in the order of increasing mean GFP expression. These GFP distributions showed that clones had variable levels of gene expression with some clones exhibiting much wider distributions than others. Clones with wide distributions were skewed both to the left or right of the mean. These experiments suggested that the integration position may also impact the noise characteristics of the viral promoter.

We next quantified that how the mean level of GFP expression correlated with gene expression noise, quantified using the coefficient of variation (CV) of the distribution. In contrast to previous studies in eukaryotic systems (12,29,32,33), we found that the noise characteristics of the promoter were uncorrelated to the mean of the GFP distribution (Fig. 7.3*A*). To ensure that other sources of noise, such as cell size and granularity do not influence the relationship between the CV and mean of the distributions, we created a smaller gate around the mean of the Forward and Side Scatter to include ~50-60% of the cells analyzed by flow cytometry, as described in (12). We still found the CV and mean of the GFP distributions to be uncorrelated (Fig. 7.3*B*). Over the 10-fold variation in mean GFP expression across clones, the level of noise varied approximately two-fold. Therefore, we decided to study 18 clones across this entire range of mean GFP expression that have similar means but approximately two-fold variation in the level of gene expression noise. As an illustration, two such pairs of clones are shown in Figure 7.4. The clone pair C04 and A10 have low levels of mean GFP expression but display a 2.05-fold difference in CV. Similarly, the clone pair IC4 and IB4 have a 1.39-fold variation in CV though they have similar levels of mean GFP expression that is nearly 8 times higher than the other clone pair (Fig. 7.4).

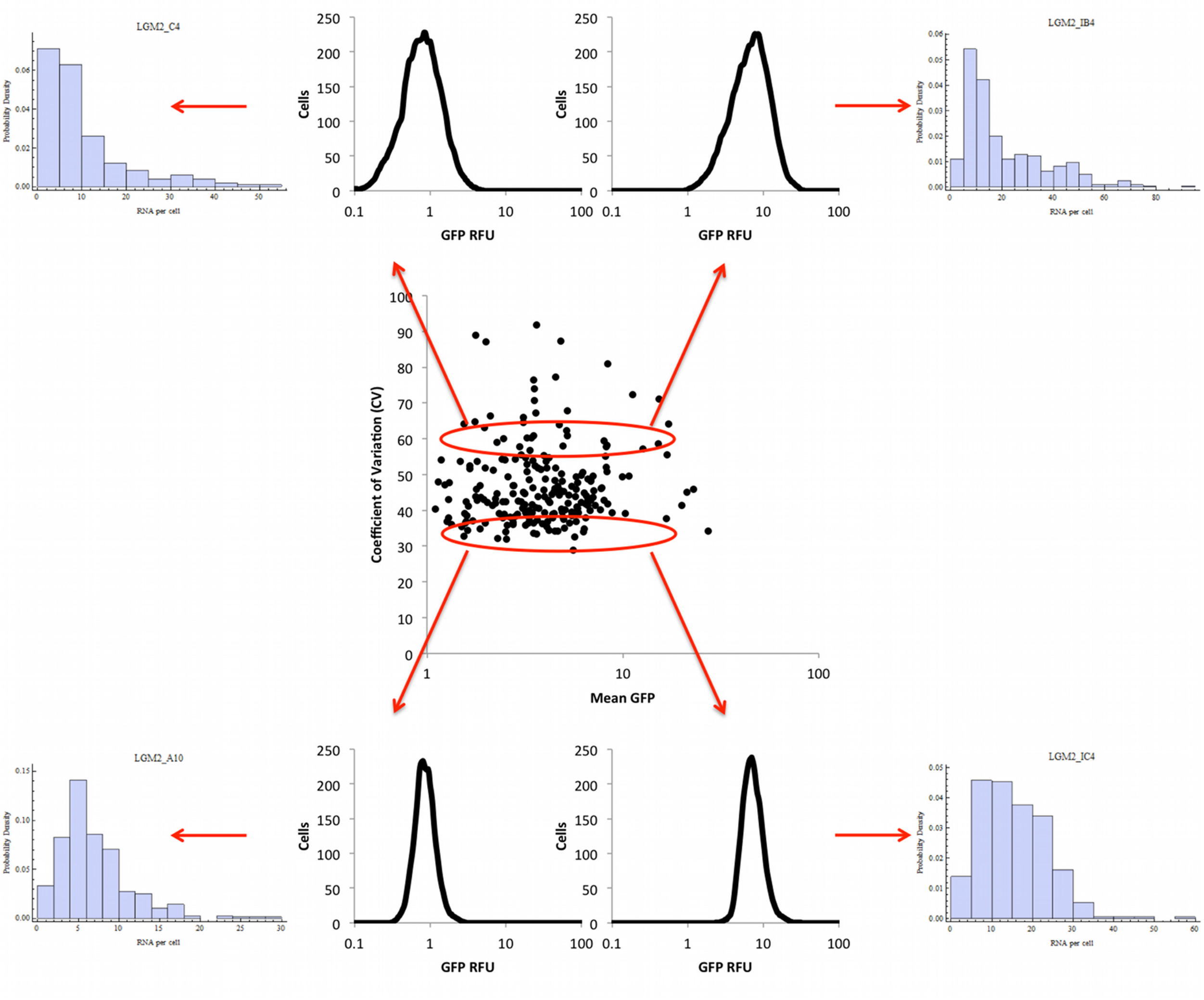


Figure 7.4. Representative example illustrating protein and mRNA distributions for high- and low-noise clones with different mean expression levels. To understand the biological mechanisms regulating gene expression noise, we selected pairs of clones with similar means but approximately two-fold variation in the level of CV. The clone pair A10/C04 have similar low mean expression levels but large differences in their protein and RNA distributions. Similarly, the clone pair IC4/IB4 show similar variations in their CV though they have higher mean expression levels. The cell-to-cell variability in protein expression is experimentally determined by measuring GFP expression using flow cytometry. The mRNA distributions are estimated using Fluorescence *in situ* hybridization and imaging 500 to 1000 cells using wild field fluorescence microscopy.

7.3 Quantifying cell-to-cell variability in the level of mRNA production from the HIV promoter

To accurately estimate the level of noise in transcription, we quantified mRNA distributions for 18 clones using FISH. The clones were fixed and hybridized with fluorescent probes complementary to the M2 array to allow detection of single transcripts as diffraction-limited spots in a wide field fluorescence microscope. To accurately distinguish real spots from background, the resulting z-stacks from each field were deconvolved. Further, algorithms were developed to automate cell identification, spot detection and counting. As expected, high-noise clones with wider GFP distributions showed greater cell-to-cell variation in the number of transcripts as compared to low-noise clones (Fig. 7.4). The mRNA distributions for the high-noise clones were much more skewed to the right with some cells having a very large number of transcripts (Fig. 7.5).

7.4 Fitting mRNA distributions to the two-state stochastic model of gene expression shows that noise in gene expression is correlated to the frequency of promoter transitions

To relate the mRNA distributions from the 18 clones to biologically interpretable phenomenon, we fit these distributions to the analytical solution for the mRNA distribution in the two-state model of gene expression. The previously solved analytical distribution for the mRNA distribution in the two-state model is given by (10,34):

$$\rho(m) = \frac{\Gamma\left(\frac{\lambda}{\delta}+m\right)}{\Gamma(1+m)\Gamma\left(\frac{\lambda}{\delta}+\frac{\gamma}{\delta}+m\right)} \frac{\Gamma\left(\frac{\lambda}{\delta}+\frac{\gamma}{\delta}\right)}{\Gamma\left(\frac{\lambda}{\delta}\right)} \left(\frac{\mu}{\delta}\right)^m \; {}_1F_1\left(\frac{\lambda}{\delta}+m, \frac{\lambda}{\delta}+\frac{\gamma}{\delta}+m, -\frac{\mu}{\delta}\right)$$

where $\rho(m)$ is the steady-state mRNA distribution. In the two-state model, the promoter is assumed to transition between an inactive and active state, with transcripts produced only from the active state. λ is the rate of promoter activation, γ is the rate of promoter inactivation, μ is the rate of transcription from the active state and δ is the rate of mRNA degradation.

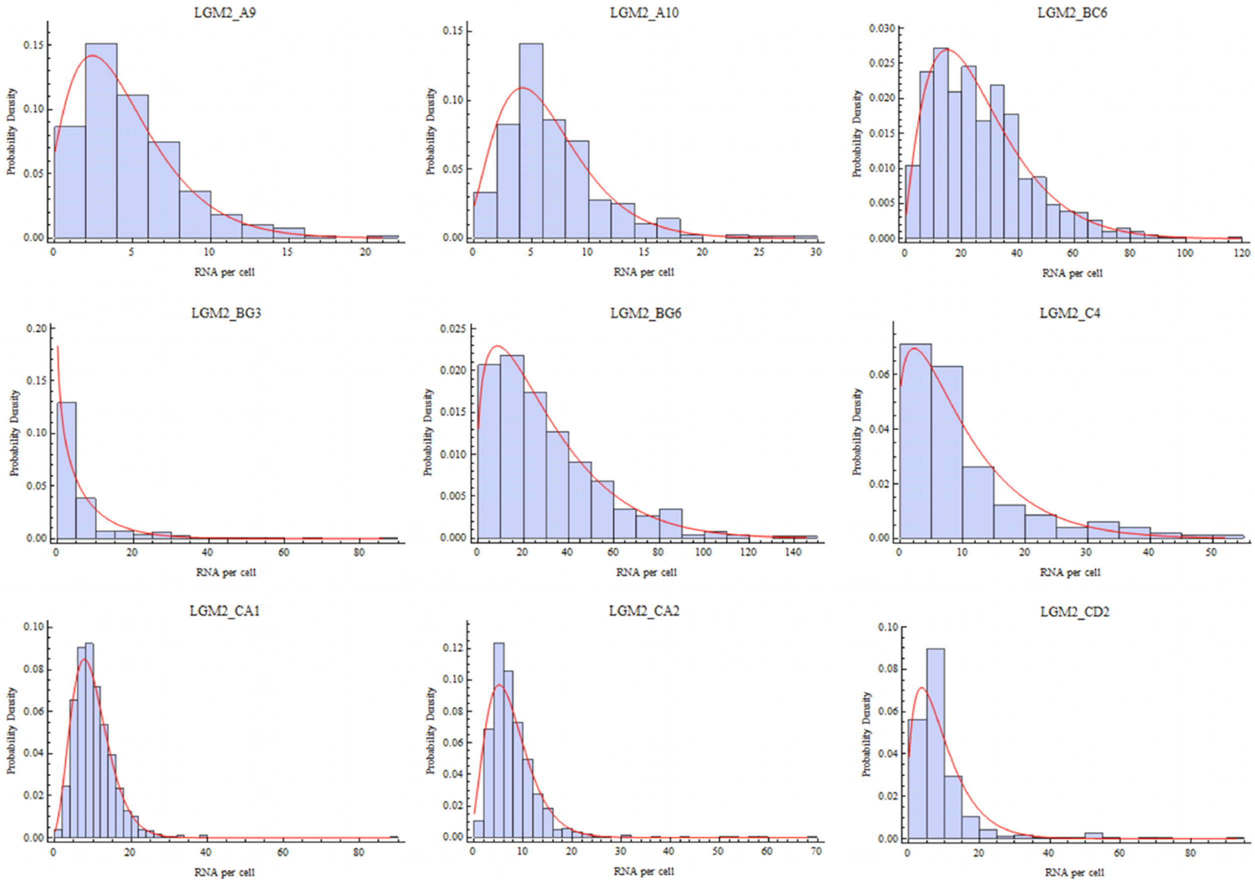
This simple stochastic model has been used to reproduce a range of single-gene expression profiles (10-12,35,36). Importantly, the relative values of the model rate constants, relative to the transcript degradation rate (δ), determine the regime of gene expression. If the rates of transition are very fast relative to transcript degradation ($\lambda, \gamma >> \delta$), then gene expression will follow a Poisson process. In contrast, if gene state transitions are extremely slow relative to transcript degradation ($\lambda, \gamma << \delta$), then each promoter state will be relatively stable, with transcripts produced in pulses that result in bimodal protein expression (35,36). Finally, if gene inactivation is much faster than activation ($\gamma >> \lambda$) and transcript degradation ($\gamma >> \delta$), then transcriptional 'bursting' results. In this regime, transcripts are produced in bursts during short-lived transitions to the active promoter state. The dynamics of bursting are often described using two parameters:

1. Burst Size: It is defined as μ / γ, with at least 1 transcript produced in the active state. It quantifies the number of transcripts produced every time the promoter transitions to the

active state.

2. Burst Frequency: It is defined as λ. It quantifies the frequency with which the promoter transitions into the active state.

Parameters in this model were estimated using Maximum Likelihood Estimation (MLE). In agreement with previous data, based on fitting of protein distributions (12), we found that the burst size for each clone correlates strongly with the mean of the RNA distribution but is not correlated to the noise in gene expression (Fig. 7.6*A,B*). Thus, clones that have higher mean expression levels produce larger transcriptional bursts. Interestingly, gene expression noise correlates strongly with the burst frequency, with noisier clones having more infrequent transitions into the active state. However, these promoter transitions do not influence the mean level of gene expression from the HIV promoter (Fig. 7.6*C,D*). Based on fitting the mRNA distributions to the two-state model, it appeared that the noise in gene expression is primarily influenced by the rate at which the promoter transitions into the active state. Since the site of integration of the viral promoter is the most distinguishing feature between the clones, we hypothesized that the chromatin environment may play a critical role in this transition between the promoter states resulting in clones with different noise characteristics.



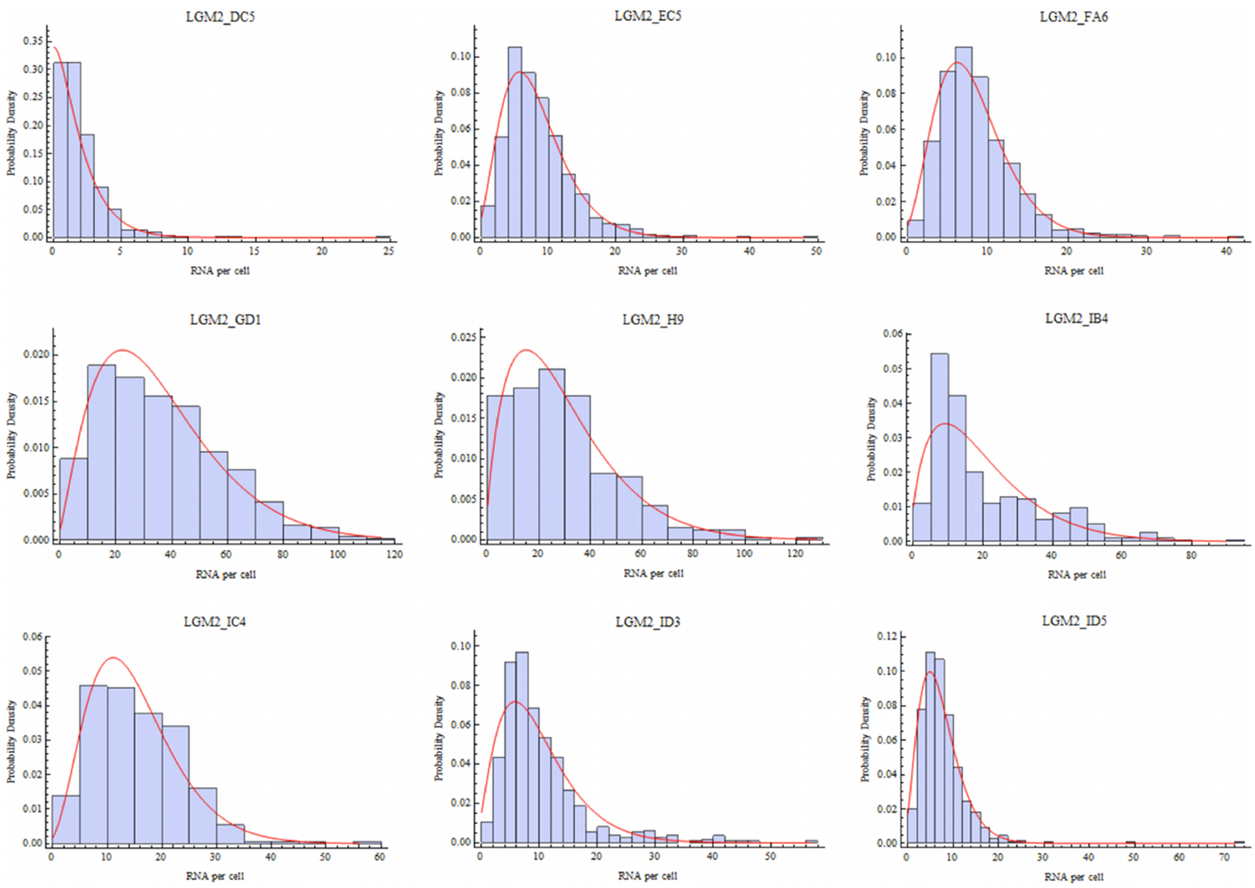


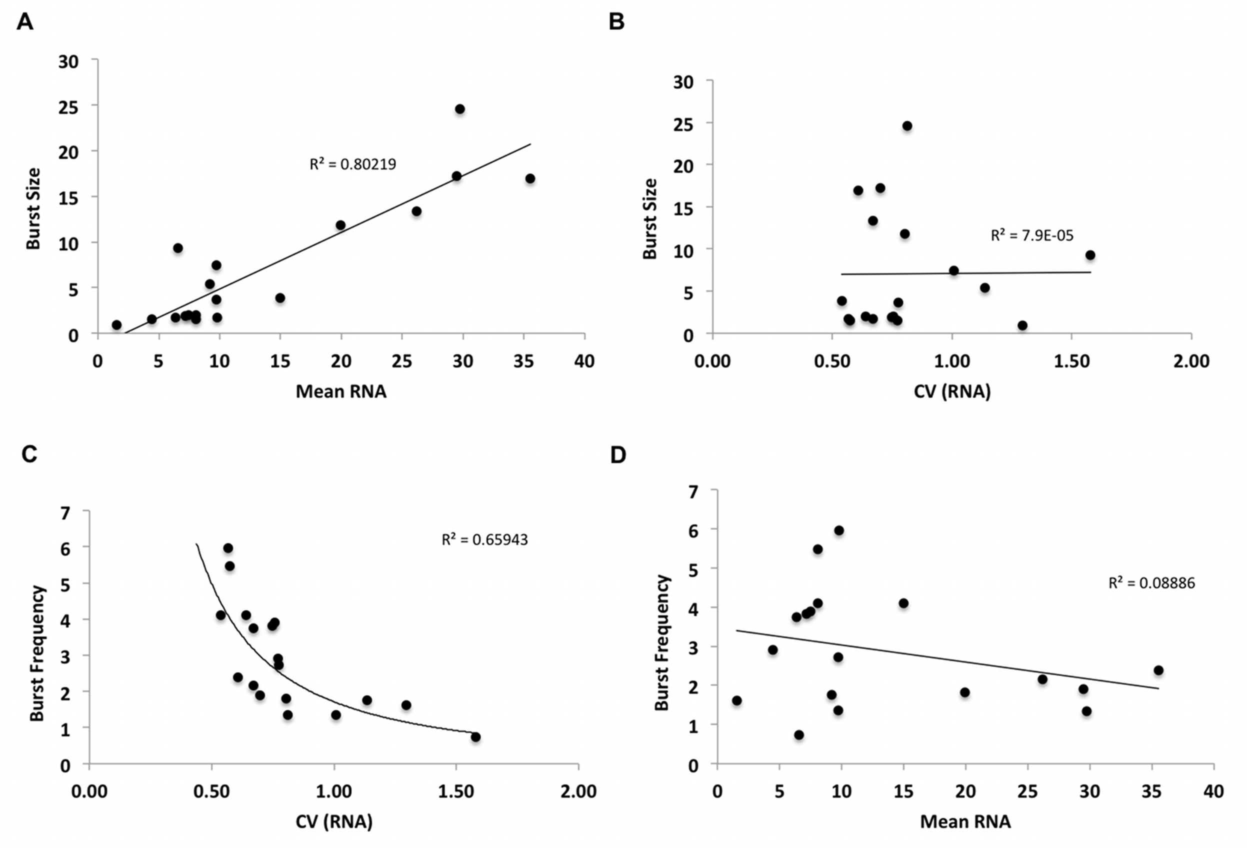
Figure 7.5. mRNA distributions for 18 clones obtained using FISH and corresponding fits to the two-state stochastic model of gene expression. FISH was used to image 500-1000 cells of each clone using wild field fluorescence microscopy. This is shown using the blue colored histograms. MLE fits to these distributions are shown using red curves.

7.5 DNase I sensitivity assay reveals systematic differences between high- and low-noise clones

The HIV-1 promoter has been well characterized (37-39), and it has been shown that the viral genome has precisely positioned nucleosomes along its entire length. Importantly, a nucleosome called Nuc-1 is positioned immediately beyond the transcription start site (TSS), which in the absence of activating factors remains bound to the viral promoter, preventing transcriptional initiation (Fig. 7.7*A*). The viral promoter has another upstream nucleosome called Nuc-0, with the nucleosome free region (NFR) between Nuc-0 and Nuc-1 containing several important transcription factor binding sites that have been shown to be important in recruiting both activating and repressive factors to the HIV-1 promoter (Fig. 7.7*A*). Further, it has recently been shown that nucleosomes are dynamically positioned within the LTR with chromatin remodeling complexes such as SWI/SNF playing an active role in nucleosome positioning.

To systematically access if differences in the chromatin environment around the site of integration regulates the noise characteristics of the viral promoter, we performed DNase I sensitivity assays. Clones were treated with DNase I and the digested chromatin was purified.

DNase I preferentially digests non-nucleosomal DNA and the ease with which a stretch of DNA is digested, a measure of the level of compacted DNA, was quantified using qPCR. The hemoglobin gene, inactive in Jurkat cells, served as an internal control to compare differences in the chromatin environment of the viral promoter between different clones.



<u>Figure 7.6. The burst frequency is strongly correlated to the level of noise in gene expression.</u> Figures show correlations between the fit parameters from the two-state model and the moments of the mRNA distribution. (A) and (B) The burst size is strongly correlated to the mean levels of gene expression but not correlated to noise in gene expression. (C) and (D) In contract, the burst frequency is strongly correlated to the level of noise in gene expression but uncorrelated to the mean expression level.

To identify differences between high- and low-noise clones, we initially accessed chromatin accessibility within a large part the HIV promoter, covering the NFR and Nuc-1 region. In support of our hypothesis that high-noise clones, with fewer promoter transitions into the active state, might be associated with more compacted chromatin, we found that chromatin inaccessibility for a clone increases with the CV of its distribution (Fig. 7.7*B*). Further, since we were specifically interested in comparing clones with similar levels of mean expression but different noise characteristics, we analyzed the ratio of chromatin inaccessibility between high- and low-noise clone pairs and found the ratio to be greater than one in all cases. This suggested that for a given mean level of gene expression, high-noise clones have more closed chromatin than low-noise clones for all pairs (Fig. 7.7*C*). Thus, it appeared that high-noise clones, with higher levels of heterochromatin, transition more infrequently into the active state resulting in greater cell-to-cell variability in gene expression.

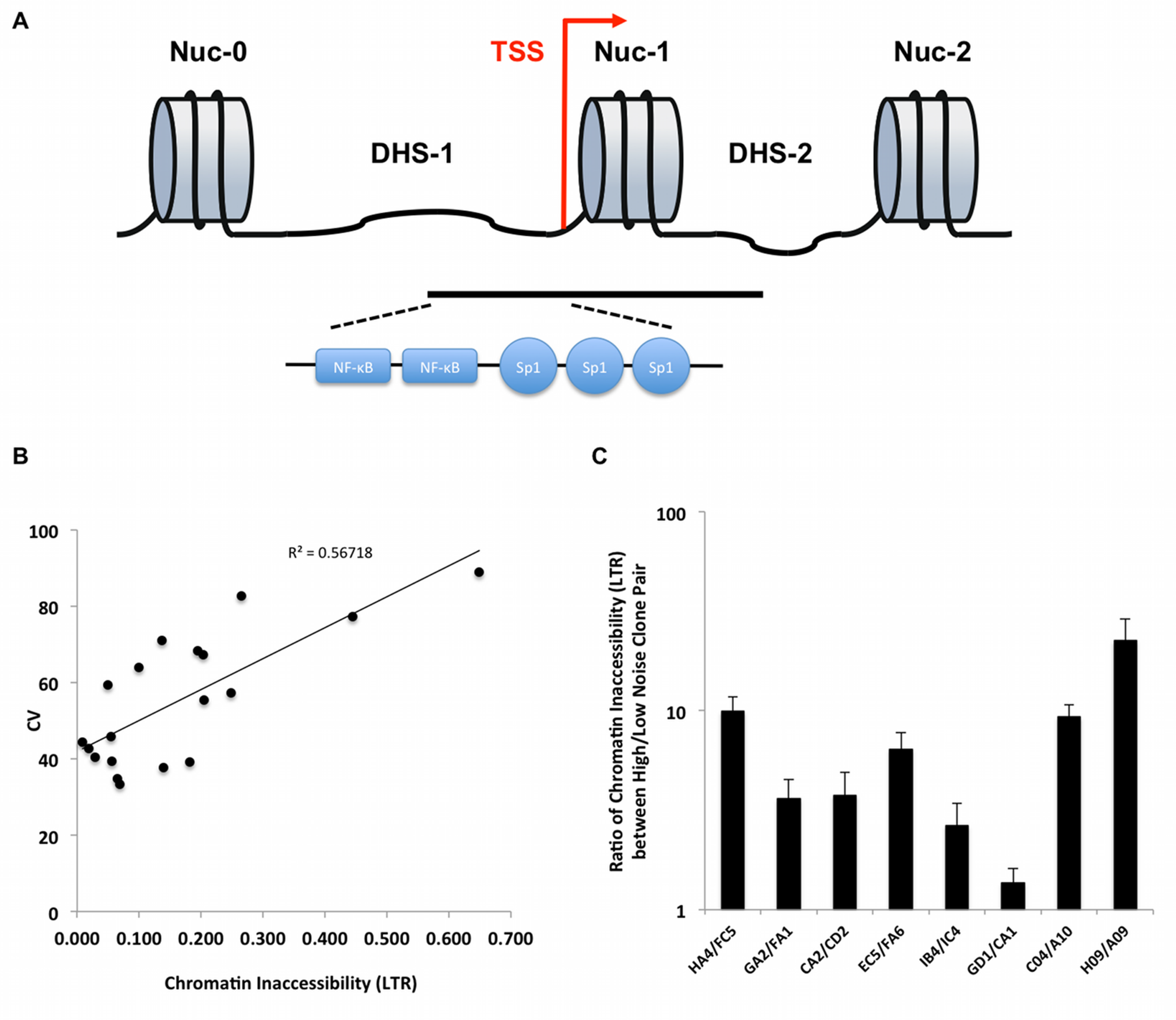


Figure 7.7. High-noise clones are integrated into more inaccessible chromatin. (A) Schematic of the HIV-1 LTR. The HIV-1 promoter has two well-positioned nucleosomes, Nuc-0 and another positioned just downstream of the transcription start-site (TSS). The DNase I hypersensitive site-1 (DHS-1) located between Nuc-0 and Nuc-1 contain several important transcription factor binding sites, such as NFκB and Sp1. The black bar shows the region of the promoter that was analyzed during the DNase I sensitivity assay. (B) Clones integrated into more inaccessible chromatin are associated with higher levels of gene expression noise. Chromatin inaccessibility was determined by designing primers that flank the black colored bar. The amplicon includes regions within DHS-1 and Nuc-1. (C) Comparison of chromatin inaccessibility between clone pairs that have similar mean expression level but show variation in the level of gene expression noise. Ratios greater than 1 for all clone pairs indicate that high-noise clones are integrated into more compacted chromatin than low-noise clones. Error bars indicate S.D. All qPCR was performed in triplicate and melt curves were run to ensure product specificity.

To study chromatin accessibility at the HIV-1 LTR more carefully, we looked at smaller regions of the promoter. We compared the Nuc-1 region, the NFR region and a region just downstream of the Nuc-0 region between high- and low-noise clones (Fig. 7.8*A*). In agreement with the previous data, we found that high-noise clones have more compacted chromatin across all 3 regions compared to low-noise clones, with ratios of chromatin inaccessibility being greater than 1 for all clone pairs (Fig. 7.8*A,B*). Surprisingly, we observed that for all clone pairs, the ratio of chromatin inaccessibility is highest within the NFR region. This might arise either due to high-noise clones being more compacted within the NFR or due to low-noise clones being more accessible within this region (Fig. 7.8*B*).

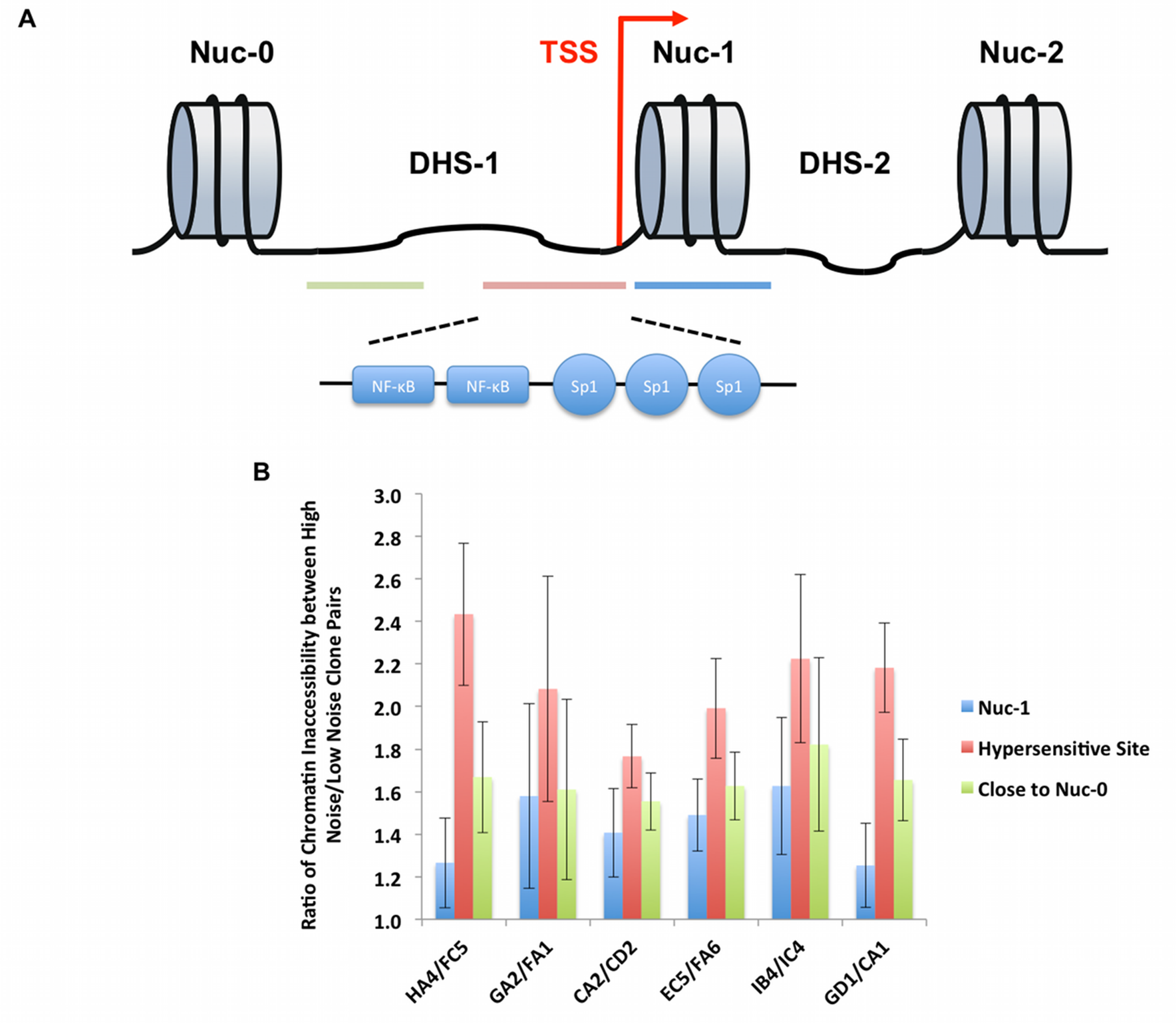
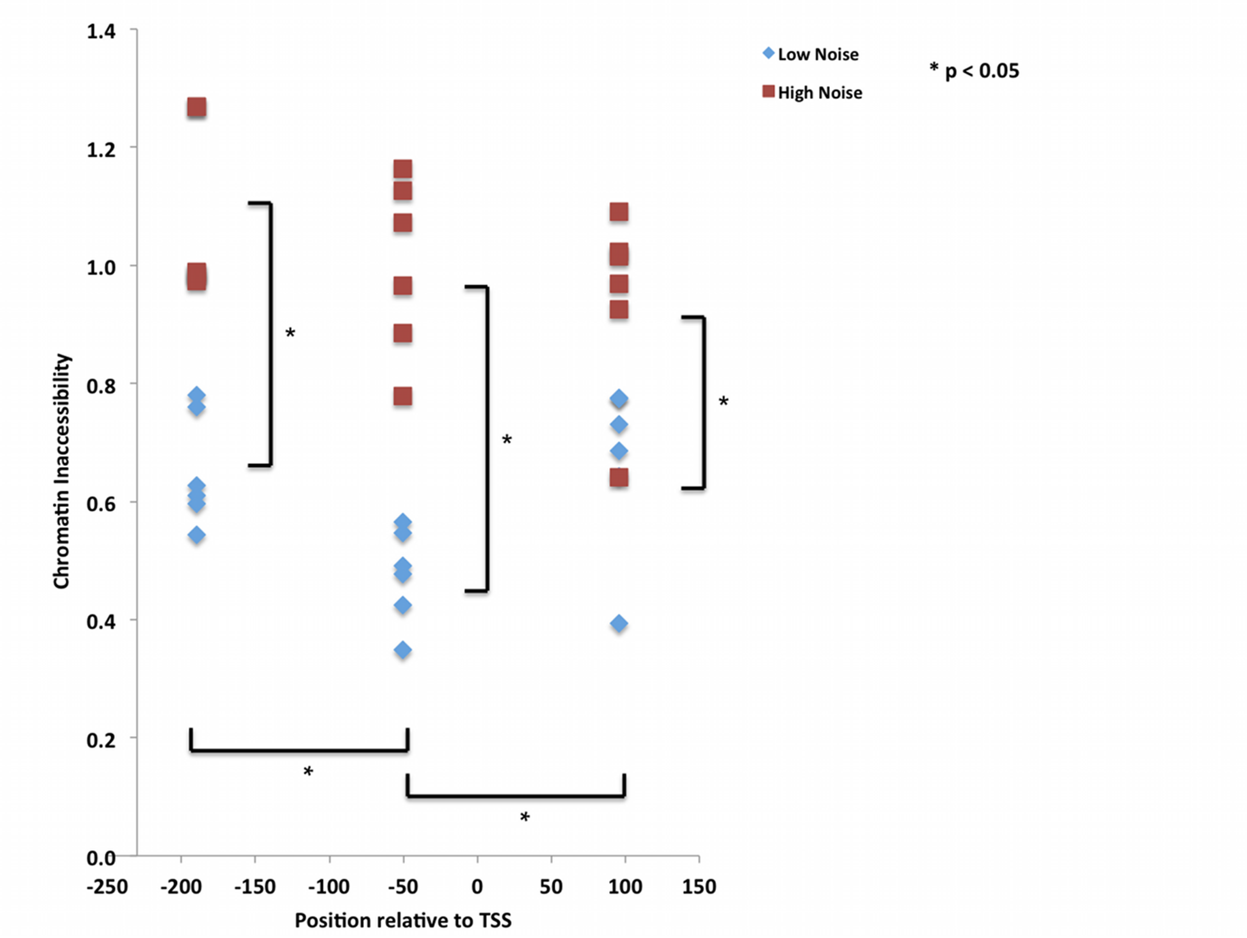


Figure 7.8. Detailed analysis of the HIV-1 LTR reveals systematic differences between high- and low- noise clones. (A) DNase I sensitivity assay was performed within three regions of the viral promoter, (1) Nuc-1 (blue bar), (2) region within DHS-1 that contains transcription factor binding sites for NFκB and Sp1 (red bar) and (3) downstream of Nuc-0 (green bar). (B) Comparison between high- and low-noise clones in these regions show ratios greater than 1 implying that high-noise clones have more dense chromatin across the entire promoter. For all clone pairs, the region with DHS-1 shows highest differences between high- and low- noise clones. Colors within the bar chart correspond to the amplicons shown in (A). Error bars indicate S.D. All qPCR was performed in triplicate and melt curves were run to ensure product specificity.

To identify the reason for higher ratios for clone pairs within the NFR, we analyzed the absolute values of chromatin inaccessibility. We found that while high-noise clones have similar levels of heterochromatin within the promoter, the low-noise clones have significantly lower levels of heterochromatin within the NFR when compared to other regions of its promoter (Fig. 7.9). This explains higher ratios for clone pairs within the NFR region (Fig. 7.8*B*). Since the NFR contains binding sites for several transcription factors that recruit both activating and repressive factors, these differences in chromatin accessibility within the NFR may have important consequences in the noise characteristics of the promoter (Fig. 7.9). Thus, it appears that there exists two systematic differences in the chromatin environment between high- and low- noise clones. First, high-clones clones in general have more compacted chromatin than low-noise clones. Second, while chromatin accessibility is relatively unchanged for high-noise clones across the entire promoter, low-noise clones appear to have especially open chromatin within the NFR. Thus, these systematic differences in the promoter chromatin environment possibly regulates the level of noise in gene expression.



<u>Figure 7.9. The chromatin around DHS-1 for low-noise clones is highly accessible.</u> Figure shows raw data from the DNase I sensitivity assay. Data for all clones, high- and low-noise, are plotted as a function of their position relative to the TSS. High-noise clones have similar chromatin inaccessibility across the entire promoter that is statistically higher than low-noise clones. For low-noise clones, the increase in chromatin accessibility at the DHS-1 site is statistically higher than at other sites in the promoter.

7.6 Discussion

This is the first study that measures noise in gene expression from the HIV-1 promoter at the single transcript level. Using the lentiviral vector LGM2, we were able to quantify noise in protein and mRNA expression from a large number of clones containing the vector integrated in different genomic locations. We fit the mRNA distributions of high- and low-noise clones to the two-state model of gene expression over the entire range of mean gene expression levels. As observed previously, we found that the burst size correlated well with the mean level of gene expression. Interesting, we also found that increasing levels of noise in gene expression was associated with more infrequent transitions into the active state.

To study how the chromatin environment might be regulating the frequency of transitions between promoter states and thereby regulating noise in gene expression, we performed DNase I sensitivity assays. We found two major differences in between high- and low-noise clones. High-noise clones appeared to be more compacted across the entire promoter supporting data from the two-state model that such clones have more infrequent transitions to the active state. Further, the NFR within the promoter is particularly accessible to transcription factors within the low-noise clones. The region of the NFR studied in our experiments contain two of the most important transcription factor binding sites, NFκB and Sp1. The NFκB sites can recruit both repressive factors HDAC1 and HDAC3 to the viral promoter and activating Histone Acetyltransferase (HATs) such as p300. Similarly, Sp1 can recruit histone modifying proteins HDACs and HATs to the viral promoter. Thus differences in accessibility and recruitment of these activating and repressive factors may play a critical role in regulating the noise characteristics from the promoter. Further chromatin immunoprecipitation (ChIP) experiments for transcription and chromatin modifying factors may help provide a molecular basis to understand the differences in the level of noise in gene expression as a function of the integration site.

Together these data suggest that by sampling different chromatin environments, HIV-1 establishes a range of noisy gene expression distributions, which may act to specify distinct infected cell fates when coupled to Tat positive feedback. In particular, we might speculate that, because productive viral replication depends on robust expression of the HIV-1 protein Tat, integrations with high basal gene expression (large burst size) and low noise (high burst frequency) will robustly generate sufficient viral Tat protein to replicate, whereas integrations with very low burst size result in unproductive infections. In contrast, those integrations with small or intermediate basal burst sizes with large noise in gene expression (resulting from infrequent transitions into the active state) may stochastically generate sufficient Tat for positive feedback activation, favoring latency. Therefore, nucleosome remodeling and features of the chromatin environment may lead to HIV-1 phenotypic diversity that may facilitate viral persistence through the establishment of latency.

7.7 Materials and Methods

7.7.1 Plasmids

The M2 repeat array was inserted into pLG by another graduate student in lab, Jonathan Foley, to obtain the plasmid pLGM2.

7.7.2 Cell Culture

Jurkat and HEK 293T cells used in these studies were cultured in RPMI 1640 (Mediatech) and Isocove's DMEM (Mediatech), respectively, at 37^0C and 5% CO_2. Both cell media were supplemented with 10% fetal bovine serum and 100U/mL Penicillin-Streptomycin.

7.7.3 Viral harvesting and infections

To package LGM2, 100 mm plates with HEK 293T cells were cotransfected with 10 μg of the plasmid pLGM2 and the following helper plasmids: 5 μg pMDLg/pRRE, 3.5 μg pVSV-G and 1.5 μg pRSV-Rev (40). Cell media was replaced 12 hours post-transfection and 24 hours after that virus was harvested by ultracentrifugation, and the viral pellets were resuspended in 100 μL PBS and stored at -80^oC for future use. $3x10^5$ cells were infected with different viral volumes and GFP expression from these cells were measured 8 days post-infection after treatment with TNFα (20 ng/mL) and TSA (400 nM) for 18 hours to obtain viral titers. To ensure single integration events per cell, the titering curves were used to infect Jurkat cells at a MOI of 0.05-0.1.

7.7.4 Cell sorting and flow cytometry

For bulk sorts, LGM2 infected Jurkat cells were stimulated with TNFα (20 ng/mL) 18 hours prior to sorting and infected GFP+ cells were sorted.

For single cell sorts, Jurkat cells each infected with a single copy of LGM2 at unique integration sites were sorted as single cells into 96-well plates. These single cells were cultured and expanded for 14-21 days and viable clones were transferred to 24-well plates. The GFP distribution of viable clones were measured using the FC500 Flow Cytometer (Beckman Coulter).

7.7.5 RNA fluorescence *in situ* hybridization

To image fixed Jurkat cells, poly-L-Lysine coated plates are use to adhere the cells to the plate. 2-3 million cells are added to each coated plate and allowed to stand for 15 minutes. The cells are then fixed by treatment with formaldehyde and stored in 70% ethanol. The fixed cells are rehydrated using 2X SSC (300 mM NaCl, 30 mM sodium citrate, pH 7.0) and 35% formamide and hybridized for 16-18 hours at 30^0C in 40 μL of a mixture containing the probe (10% dextran sulfate, 2 mM vanadyl-ribonucleoside complex, 0.02% RNAse-free BSA, 40 μg *E.coli* tRNA, 2x SSC, 35% formamide, 30 ng of probe). After overnight incubation, the slides are washed and treated with DAPI to identify cell nuclii. Since the slides are imaged for ~10-12 hours, they are treated with 100 μL of buffer containing Glucose Oxidase and Catalase to prevent photo bleaching and mounted with a coverslip to prepare it for imaging.

7.7.6 Stochastic model of gene expression

The RNA distributions acquired from RNA FISH were used to estimate the parameters in the stochastic two-state model of gene expression using the steady-state solution for the mRNA distribution shown in Section 7.4. Codes for Maximum Likelihood Estimation to obtain parameter values for the two-state model were written in Mathematica. Codes will be made available upon request.

7.7.7 DNase I sensitivity assay

The EpiQ™ Chromatin Analysis Kit (Bio-Rad) was used for the DNase I sensitivity assay. Briefly, 250,000 cells were incubated with 2 μL DNAse I for 1 hour and after quenching the digestion reaction, genomic DNA is extracted from the samples. The level of HBB and LTR are then quantified by qPCR (Bio-Rad iCycler, iQ5) using the EpiQ Chromatin SYBR Supermix (Bio-Rad). Primers were designed to prime the following regions: 1) Within a large region of the LTR – LTR-F (5'-GGACTTTCCGCTGGGGACTTTCCAGGG-3') and LTR-R (5'-GCGCGCTTCAGCAAGCCGAGTCCTGCGTCGAG-3'); 2) Within Nuc-1 – Nuc1-F (5'-AGCTCTCTGGCTAACTAGGG-3') and Nuc1-R (5'-AAAGGGTCTGAGGGATCTCTAG-3'); 3) Within DHS-1 – DHS1-F (5'- GGGACTTTCCGCTGGGGAC-3') and DHS1-R (5'-CCCAGTACAGGCAAAAAGCAGC-3'); and 4) Close to the 5' end of Nuc-0 – Nuc0-F (5'-GAGCCTGCATGGGATGG-3') and Nuc0-R (5'- CTCCGGATGCAGCTCTC-3'). Primers used to quantify HBB were: HBB-F (5'-AAGCCAGTGCCAGAAGAGCCAAGGA-3') and HBB-R (5'-CCCACAGGGCAGTAACGGCAGACTT-3').

7.8 References

1. McAdams, H. H., and Arkin, A. (1997) *Proc Natl Acad Sci U S A* 94, 814-819
2. Eldar, A., and Elowitz, M. B. (2010) *Nature* 467, 167-173
3. Losick, R., and Desplan, C. (2008) *Science* 320, 65-68
4. Singh, A., and Weinberger, L. S. (2009) *Curr Opin Microbiol* 12, 460-466
5. Stumpf, M. P., Laidlaw, Z., and Jansen, V. A. (2002) *Proc Natl Acad Sci U S A* 99, 15234-15237
6. Niepel, M., Spencer, S. L., and Sorger, P. K. (2009) *Curr Opin Chem Biol* 13, 556-561
7. Raj, A., and van Oudenaarden, A. (2008) *Cell* 135, 216-226
8. Thattai, M., and van Oudenaarden, A. (2001) *Proc Natl Acad Sci U S A* 98, 8614-8619
9. Newman, J. R., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S. (2006) *Nature* 441, 840-846
10. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006) *PLoS Biol* 4, e309
11. Zenklusen, D., Larson, D. R., and Singer, R. H. (2008) *Nat Struct Mol Biol* 15, 1263-1271
12. Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V., and Arkin, A. P. (2010) *PLoS Comput Biol* 6
13. Sigal, A., Milo, R., Cohen, A., Geva-Zatorsky, N., Klein, Y., Liron, Y., Rosenfeld, N., Danon, T., Perzov, N., and Alon, U. (2006) *Nature* 444, 643-646
14. Cohen, A. A., Kalisky, T., Mayo, A., Geva-Zatorsky, N., Danon, T., Issaeva, I., Kopito, R. B., Perzov, N., Milo, R., Sigal, A., and Alon, U. (2009) *PLoS One* 4, e4901
15. Segal, E., and Widom, J. (2009) *Nat Rev Genet* 10, 443-456
16. Mao, C., Brown, C. R., Griesenbeck, J., and Boeger, H. (2011) *PLoS One* 6, e17521
17. Kundu, S., and Peterson, C. L. (2009) *Biochim Biophys Acta* 1790, 445-455
18. Becskei, A., Kaufmann, B. B., and van Oudenaarden, A. (2005) *Nat Genet* 37, 937-944
19. Tirosh, I., and Barkai, N. (2008) *Genome Res* 18, 1084-1091

20. Choi, J. K., and Kim, Y. J. (2008) *Nat Genet* 40, 141-147
21. Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., Widom, J., and Segal, E. (2008) *PLoS Comput Biol* 4, e1000216
22. Han, Y., Wind-Rotolo, M., Yang, H. C., Siliciano, J. D., and Siliciano, R. F. (2007) *Nat Rev Microbiol* 5, 95-106
23. Joos, B., Fischer, M., Kuster, H., Pillai, S. K., Wong, J. K., Boni, J., Hirschel, B., Weber, R., Trkola, A., and Gunthard, H. F. (2008) *Proc Natl Acad Sci U S A* 105, 16725-16730
24. Richman, D. D., Margolis, D. M., Delaney, M., Greene, W. C., Hazuda, D., and Pomerantz, R. J. (2009) *Science* 323, 1304-1307
25. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., and Schaffer, D. V. (2005) *Cell* 122, 169-182
26. Weinberger, L. S., Dar, R. D., and Simpson, M. L. (2008) *Nat Genet* 40, 466-470
27. Blazkova, J., Trejbalova, K., Gondois-Rey, F., Halfon, P., Philibert, P., Guiguen, A., Verdin, E., Olive, D., Van Lint, C., Hejnar, J., and Hirsch, I. (2009) *PLoS Pathog* 5, e1000554
28. Margolis, D. M. (2010) *Curr HIV/AIDS Rep* 7, 37-43
29. Singh, A., Razooky, B., Cox, C. D., Simpson, M. L., and Weinberger, L. S. (2010) *Biophys J* 98, L32-34
30. Sadowski, I., Lourenco, P., and Malcolm, T. (2008) *Curr HIV Res* 6, 286-295
31. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P., and Schaffer, D. V. (2009) *PLoS Pathog* 5, e1000260
32. Kaern, M., Elston, T. C., Blake, W. J., and Collins, J. J. (2005) *Nat Rev Genet* 6, 451-464
33. Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O'Shea, E., Pilpel, Y., and Barkai, N. (2006) *Nat Genet* 38, 636-643
34. PECCOUD, J., and YCART, B. (1995) *THEORETICAL POPULATION BIOLOGY* 48, 222-234
35. Blake, W. J., M, K. A., Cantor, C. R., and Collins, J. J. (2003) *Nature* 422, 633-637
36. Raser, J. M., and O'Shea, E. K. (2004) *Science* 304, 1811-1814
37. Verdin, E., Paras, P., Jr., and Van Lint, C. (1993) *The EMBO journal* 12, 3249-3259
38. el Kharroubi, A., and Verdin, E. (1994) *The Journal of biological chemistry* 269, 19916-19924
39. Rafati, H., Parra, M., Hakre, S., Moshkin, Y., Verdin, E., and Mahmoudi, T. (2011) *PLoS Biol* 9, e1001206
40. Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., and Naldini, L. (1998) *Journal of virology* 72, 8463-8471