JBC Papers in Press. Published on January 17, 2012 as Manuscript M111.302653
The latest version is at http://www.jbc.org/cgi/doi/10.1074/jbc.M111.302653

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

Siddharth S. Dey[1], Yuhua Xue[3], Marcin P. Joachimiak[4,5], Gregory D. Friedland[8,9], John C. Burnett[1,7], Qiang Zhou[3], Adam P. Arkin[2,4,5,6,†], David V. Schaffer[1,2,4, †]

From the [1]Department of Chemical and Biomolecular Engineering and the Helen Wills Neuroscience Institute, the [2]Department of Bioengineering, and the [3]Department of Molecular and Cell Biology, University of California, Berkeley, CA 94720

[4]Physical Biosciences Division, the [5]Virtual Institute of Microbial Stress and Survival, and the [6]DOE, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

[7]Current Address: Division of Molecular and Cellular Biology, Beckman Research Institute of City of Hope, Duarte, CA 91010

[8]Technology Division, Joint BioEnergy Institute, Emeryville, CA 94608

[9]Biomass Science and Conversion Technology Department, Sandia National Laboratories, Livermore, CA 94551

[10]Current Address: School of Pharmaceutical Sciences, Xiamen University, Xiamen 361005, China

*Running Title: *Functional characterization of coevolving sites in HIV-1 Tat*

To whom correspondence may be addressed: David V. Schaffer (Department of Chemical and Biomolecular Engineering and the Helen Wills Neuroscience Institute, Department of Bioengineering, University of California, Berkeley, CA, USA 94720 and the Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA 94720, Tel.: (510) 643-5963; Fax: (510) 642-4778; Email: schaffer@berkeley.edu) or Adam P. Arkin (Department of Bioengineering, University of California, Berkeley, CA USA 94720, and the Physical Biosciences Division, Virtual Institute of Microbial Stress and Survival, and DOE, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA 94720, Tel: (510) 495-2366; Fax: (510) 486-6059; Email: aparkin@lbl.gov)

Keywords: HIV-1 Tat; protein coevolution; Tat-P-TEFb interaction; HIV-1 subtypes; Mutual Information

Background: It is not well understood how HIV-1 Tat performs complex functions despite its considerable sequence diversity.
Results: Individually non-conserved but functionally coevolving sites in Tat were identified.
Conclusion: In contrast to most coevolving sites, these sites are critical for two distinct functions regulating viral gene expression.
Significance: Balancing the two mechanisms is a strategy to maintain Tat function and may impact viral latency.

SUMMARY

Viral genomes are continually subjected to mutations, and functionally deleterious ones can be rescued by reversion or additional mutations that restore fitness. The error prone nature of HIV-1 replication has resulted in highly diverse viral sequences, and it is not clear how viral proteins such as Tat, which plays a critical role in viral gene expression and replication, retain their 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, whose analysis could yield new mechanistic insights into Tat function and evolution. To identify such sites, we combined Mutual Information analysis and experimentation to identify coevolving positions

and found that residues 35 and 39 are strongly correlated. Mutation of either residue of this pair into amino acids that appear in numerous viral isolates yields a defective virus; however, simultaneous introduction of both mutations into the heterologous Tat sequence restores gene expression close to wild-type Tat. Furthermore, in contrast to most coevolving protein residues that contribute to the same function, structural modeling and biochemical studies showed that these two residues contribute to two mechanistically distinct steps in gene expression: binding P-TEFb and promoting P-TEFb phosphorylation of CTD in RNAPII. Moreover, Tat variants that mimic HIV-1 subtype B or C at sites 35 and 39 have evolved orthogonal strengths of P-TEFb binding vs. RNAPII phosphorylation, suggesting that subtypes have evolved alternate transcriptional strategies to achieve similar gene expression levels.

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)[11], 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). 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. supplemental Fig. S1) (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). 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.

EXPERIMENTAL PROCEDURES

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.

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 (32). 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.

Co-Immunoprecipitation and Western Blots - Protocol details are provided in supplemental Experimental Procedures. The following antibodies were used for this assay: 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).

Chromatin Immunoprecipitation - Upstate EZ ChIP Kit reagents (Upstate) and protocol were used for the assay with variations. 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). See supplemental Experimental Procedures for details of protocol and primers.

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.

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

Details of *Plasmids; Flow Cytometry and Cell Sorting; Transfections; mRNA Extraction and RT-qPCR; Co-Immunoprecipitation and Western Blots; Chromatin Immunoprecipitation; Nuclease Sensitivity Assay; Luciferase Assay; and Structure Modeling;* can be found in supplemental Experimental Procedures.

RESULTS

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.

Using a method described by Dunn *et al.* to minimize background noise, three position pairs – (35,39), (31,35), and (31,39) – were identified with MI scores higher than a threshold score, computed using a method described by Weigt *et al.* (Fig. 1*A*,*B* and supplemental Fig. S2). See supplemental Results for details (33-34). 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. 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. 2*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 certain mathematical constraints (see supplemental Results for details) (24). We therefore also experimentally tested several sites whose MI did not reach threshold. Furthermore, we tested several non-conserved sites within background MI as controls to validate the low predicted functional relationship between such sites. Experimentally tested positions are shown by solid black dots (Fig. 1*B*).

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 (supplemental Fig. S1), were replaced with residues of other naturally occurring viral variants or subtypes (Fig. 2*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. 2*B*) (35). 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 (supplemental Fig. S3*A*) (35). Two metrics were used to quantify gene expression (36). 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) (36). 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. 2*A*, yielded non-expressing virus (Fig. 2*C* and supplemental Fig. S3*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. 2*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 (Fig. 2*C,D* and supplemental Fig. S3*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 (supplemental Fig. S3*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 (supplemental Fig. S4*C,D*). However, gene expression was restored in the Tat C double-mutant, suggesting that sites 35 and 39 alone compensate for one other (supplemental Fig. S4*C,D*). Thus, evolutionarily only certain pairs of amino acids at sites 35 and 39 but not their intermediates are tolerated.

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 (supplemental Fig. S5). These interactions between sites 31-35-39 are discussed from a structural perspective in additional detail in the supplemental Results 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 (supplemental Fig. S4*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 (supplemental Fig. S4*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. 2*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

(supplemental Fig. S4*A,B*), validating the predictions from MI analysis. These mutations are discussed in greater detail from a structural and energetic perspective in the supplemental Results section.

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. 2*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. 2*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. 2*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 (35-36). 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 (36). The sorted cells were allowed to relax for 9 days, and GFP- cells infected with the LGIT vector, but not expressing GFP, were sorted (supplemental Fig. S6*A*). These silent proviruses were then monitored for GFP expression over the course of 12 days (supplemental Fig. S6*B*).

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 (supplemental Fig. S6*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.

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 (9,37). 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*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 (Figs. 3*A*,*B*), and the loss of this binding possibly underlies the defective gene expression for this mutant (Fig. 2*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 (supplemental Fig. S7) (38). 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*A*,*B*).

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 (37). 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. 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*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 (37), 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*B*), but lower than WT Tat, suggesting that its inability to activate gene expression (Fig. 2*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. 4*A*), total cellular RNA was extracted, and the levels of viral transcripts were quantified using RT-qPCR (36). The I39Q and DM Tat both had similar percentages of elongated transcripts (Fig. 4*B*); however, the I39Q Tat has much lower levels of total transcripts compared to the DM Tat (Fig. 4*C*), suggesting that it fails to induce transcription at the same efficiency as the DM Tat.

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 (21,39). However, co-immunoprecipitation showed no differences in Sp1 binding between the I39Q and DM Tat (Fig. 5*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. 5*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 (40). Normally, RNAPII recruited to the promoter of a gene is phosphorylated at Ser5 by Cdk7 within the transcription factor complex TFIIH (41). 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 (42-43). 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 (11-12,44).

Consistent with these results, the recently solved crystal structure of the P-TEFb-Tat complex shows that Tat binding induces P-TEFb conformational changes (37). 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 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. 4*C* and Table 1) (37,44).

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. 6*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. 6*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. 6*B*).

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*B* and 6*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. 6*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*B* and 6*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. 6*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.

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" (45-48). 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. 2*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. 2*C* and supplemental Fig. S4*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). In contrast, the I39Q Tat binds P-TEFb at levels close to the DM Tat, yet still suffers from very low gene expression (Figs. 2*C* and 3*B*) potentially due to its inability to induce P-TEFb to phosphorylate the CTD of RNAPII (Figs. 6*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 (49), 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.

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. 6*A*), but the strong binding to P-TEFb could at least in part compensate for this deficit (Fig. 3*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 (supplemental Fig. S8). 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. 6*A*) coupled with relatively weaker P-TEFb binding (Fig. 3*B*) that in combination could ultimately drive comparable levels of Tat-mediated gene expression as measured by GFP expression (Fig. 2*C*), viral transcript analysis, and Ser2P-CTD of RNAPII (supplemental Fig. S8). 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. 6*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 (50).

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 (35-36,51). 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 (supplemental Fig. S6). 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.

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. Dull, T., Zufferey, R., Kelly, M., Mandel, R. J., Nguyen, M., Trono, D., and Naldini, L. (1998) *J Virol* **72**, 8463-8471
33. Dunn, S. D., Wahl, L. M., and Gloor, G. B. (2008) *Bioinformatics* **24**, 333-340
34. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A., and Hwa, T. (2009) *Proc Natl Acad Sci U S A* **106**, 67-72
35. Weinberger, L. S., Burnett, J. C., Toettcher, J. E., Arkin, A. P., and Schaffer, D. V. (2005) *Cell* **122**, 169-182
36. Burnett, J. C., Miller-Jensen, K., Shah, P. S., Arkin, A. P., and Schaffer, D. V. (2009) *PLoS Pathog* **5**, e1000260
37. Tahirov, T. H., Babayeva, N. D., Varzavand, K., Cooper, J. J., Sedore, S. C., and Price, D. H. (2010) *Nature* **465**, 747-751

38. 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
39. Chun, R. F., Semmes, O. J., Neuveut, C., and Jeang, K. T. (1998) *J Virol* **72**, 2615-2629
40. Phatnani, H. P., and Greenleaf, A. L. (2006) *Genes Dev* **20**, 2922-2936
41. Serizawa, H., Makela, T. P., Conaway, J. W., Conaway, R. C., Weinberg, R. A., and Young, R. A. (1995) *Nature* **374**, 280-282
42. Zhang, Z., Klatt, A., Gilmour, D. S., and Henderson, A. J. (2007) *J Biol Chem* **282**, 16981-16988
43. Schroeder, S. C., Schwer, B., Shuman, S., and Bentley, D. (2000) *Genes Dev* **14**, 2435-2440
44. 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
45. Tachedjian, G., Aronson, H. E., and Goff, S. P. (2000) *Proc Natl Acad Sci U S A* **97**, 6334-6339
46. Verhoef, K., and Berkhout, B. (1999) *J Virol* **73**, 2781-2789
47. Cimarelli, A., Sandin, S., Hoglund, S., and Luban, J. (2000) *J Virol* **74**, 4273-4283
48. Jain, C., and Belasco, J. G. (1996) *Cell* **87**, 115-125
49. del Alamo, M., and Mateu, M. G. (2005) *J Mol Biol* **345**, 893-906
50. Li, N., Schlessinger, J., and Margolis, B. (1994) *Oncogene* **9**, 3457-3465
51. Skupsky, R., Burnett, J. C., Foley, J. E., Schaffer, D. V., and Arkin, A. P. (2010) *PLoS Comput Biol* **6**

Acknowledgements - We thank Prof. Kathryn Miller-Jenson for careful perusal of the manuscript and helpful discussions. We thank Dr. Sharon Aviran and Morgan Price for technical assistance with the MI analysis, Jasper Rine's lab and Laura Lombardi for use of the Branson Sonifier 450 and Steve Okino (Bio-Rad) for the generous gifts of the EqiQ Chromatin SYBR Supermix and Analysis Kit. Cell sorting was performed at the UC Berkeley Flow Cytometry Core Facility with assistance from Hector Nolla and Alma Valeros.

FOOTNOTES

* This work was supported by the NIH grant R01GM073058 to A.P.A. and D.V.S., and NIH grants R01AI41757-11 and R01AI41757-11S1 to Q.Z.

† To whom correspondence may be addressed: D.V.S (Department of Chemical and Biomolecular Engineering and the Helen Wills Neuroscience Institute, Department of Bioengineering, University of California, Berkeley, CA, USA 94720 and the Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA 94720, Tel.: (510) 643-5963; Fax: (510) 642-4778; Email: schaffer@berkeley.edu) or A.P.A (Department of Bioengineering, University of California, Berkeley, CA, USA 94720, and the Physical Biosciences Division, Virtual Institute of Microbial Stress and Survival, and DOE, Joint BioEnergy Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA 94720, Tel: (510) 495-2366; Fax: (510) 486-6059; Email: aparkin@lbl.gov).

[11]The abbreviations used are: HIV-1, human immunodeficiency virus-1; P-TEFb, positive transcription-elongation factor b;CycT1, cyclin T1; cyclin-dependent kinase 9 (Cdk9); TAR, transactivation response element; PIC, pre-initiation complex; CTD, C-terminal domain; RNAPII, RNA polymerase II; snRNP, small nuclear ribonucleoprotein; MI, Mutual Information; WT, wild-type; DM, double-mutant; LTR, long terminal repeat; IRES, internal ribosome entry sequence; LGIT, LTR-GFP-IRES-Tat; UbChIT, Ubiquitin-mCherry-IRES-Tat; TNF-α, tumor cecrosis factor- α; TSA, trichostatin A.

FIGURE LEGENDS

FIGURE 1. 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. 2.

FIGURE 2. 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.

FIGURE 3. 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.

FIGURE 4. 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.

FIGURE 5. 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 ($p>0.05$).

FIGURE 6. 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.

TABLES

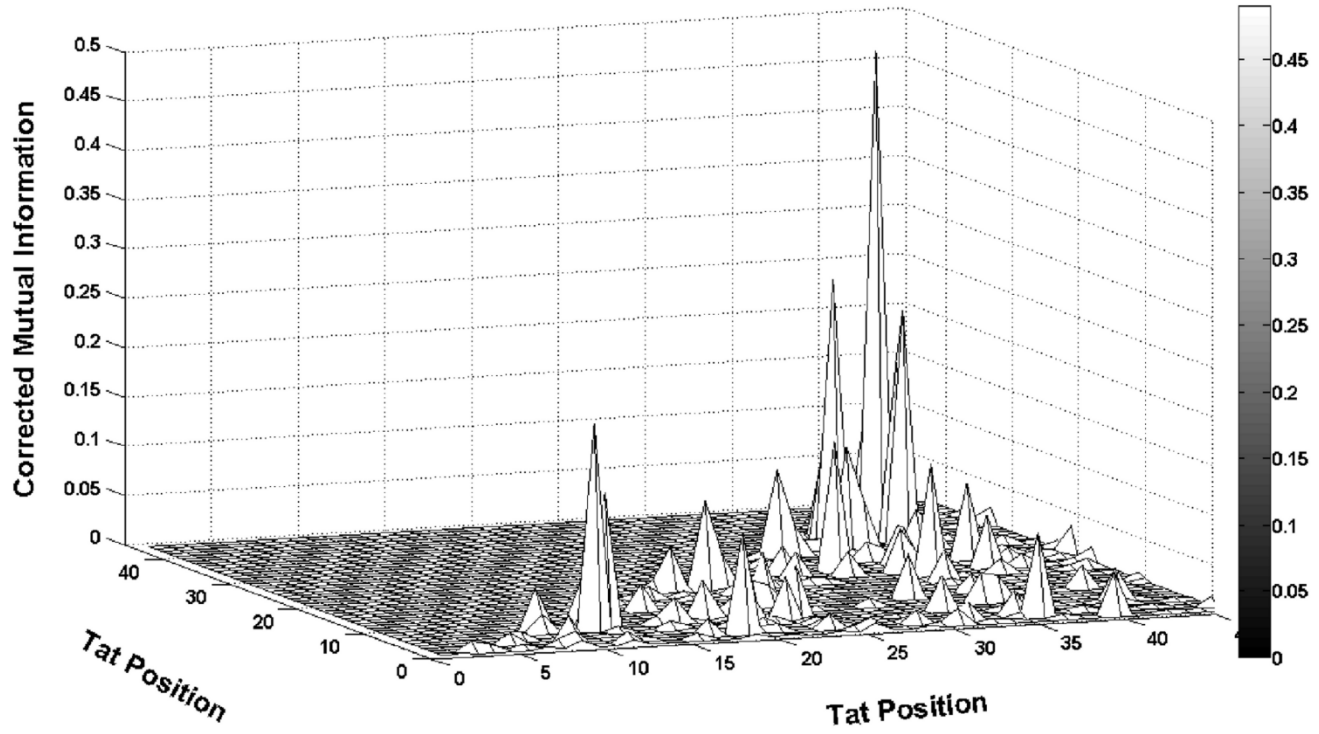
TABLE 1. *In silico* modeling results of the stability (*ΔΔG*) of the Tat-P-TEFb or Tat-P-TEFb-ATP complex after introducing mutations in Tat.

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.

Figure 1

A



B

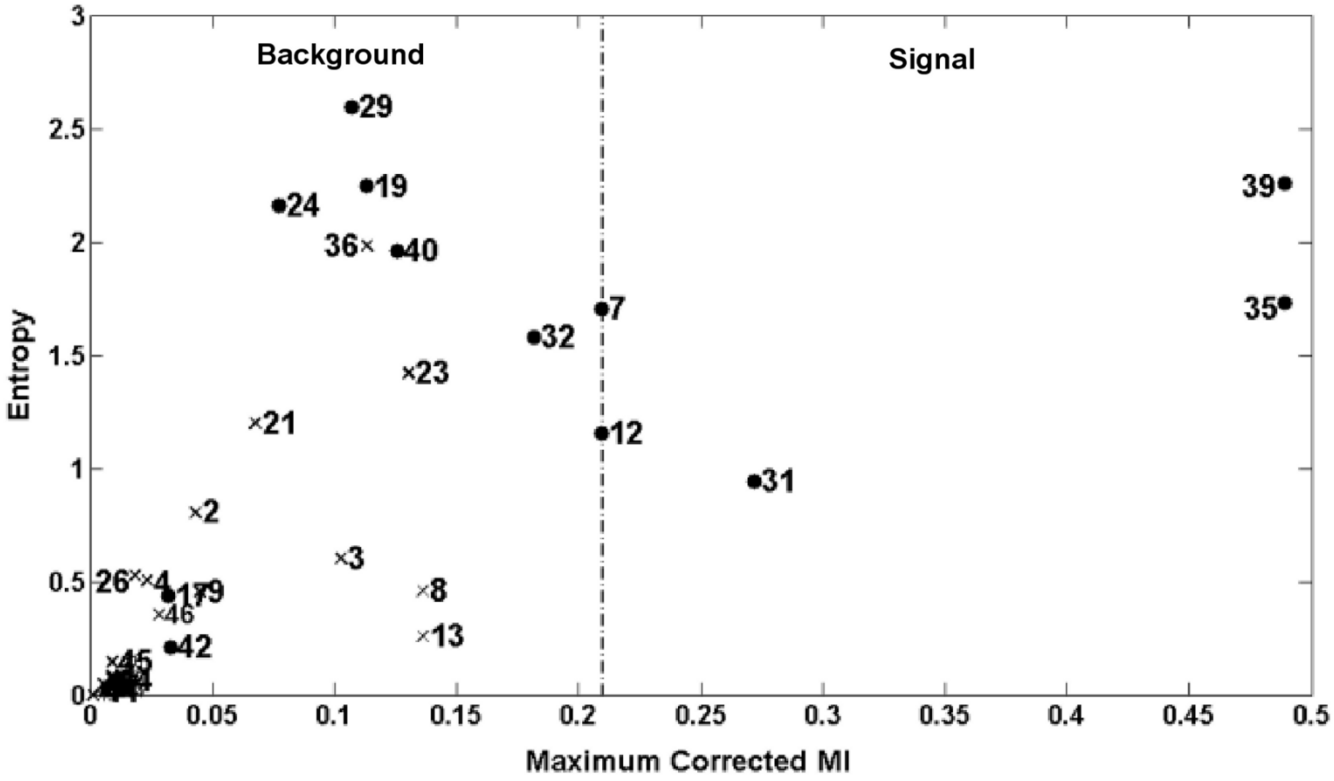


Figure 2

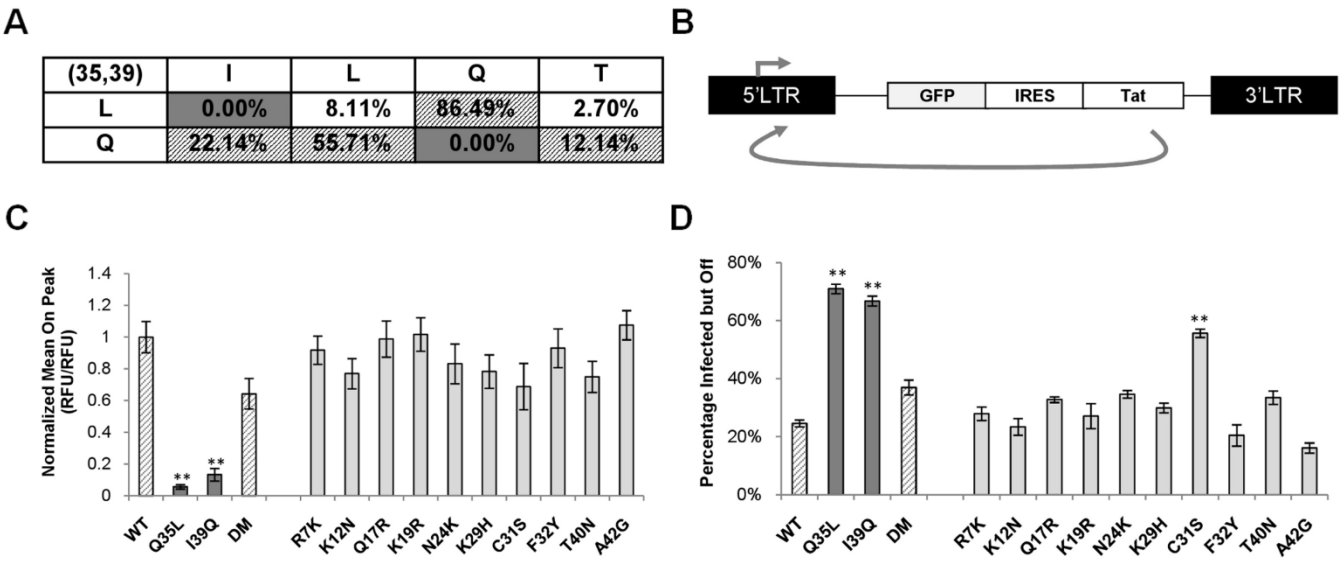


Figure 3

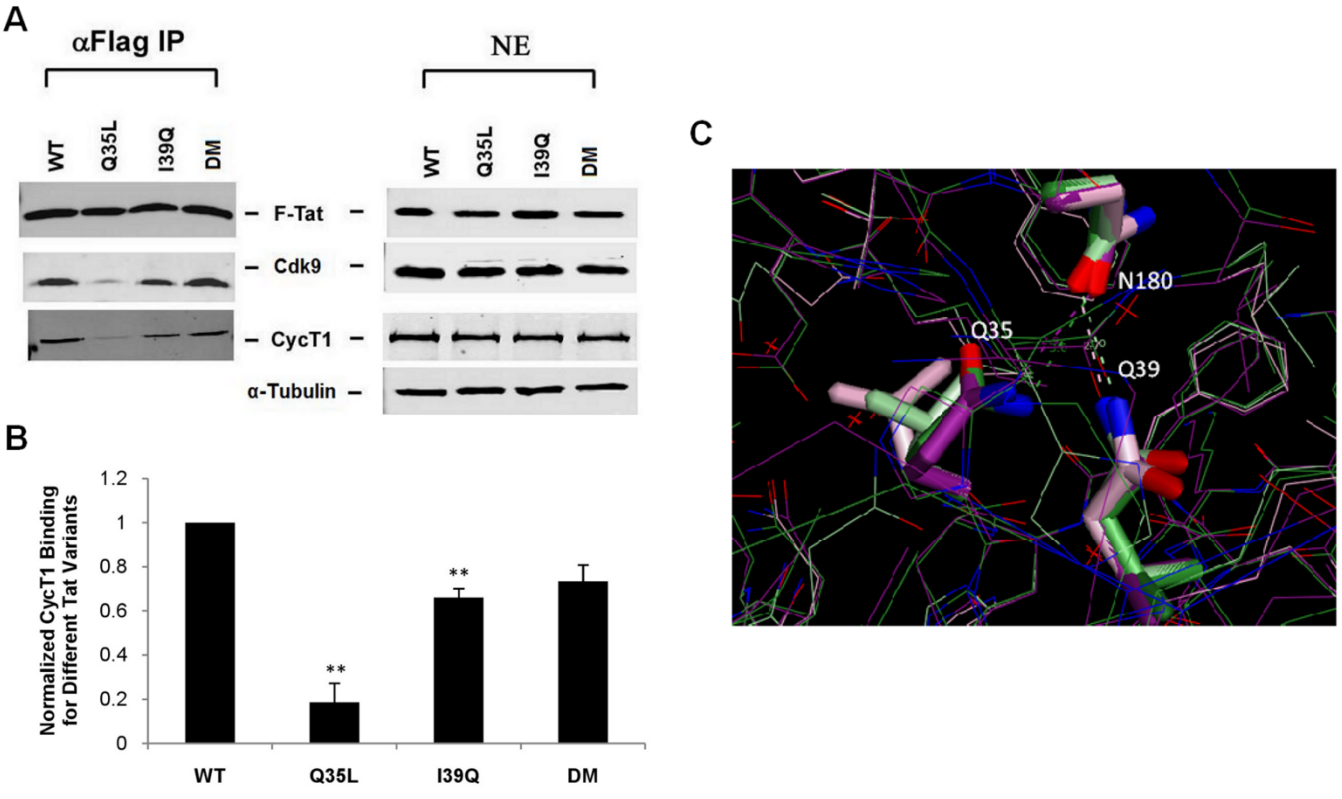


Figure 4

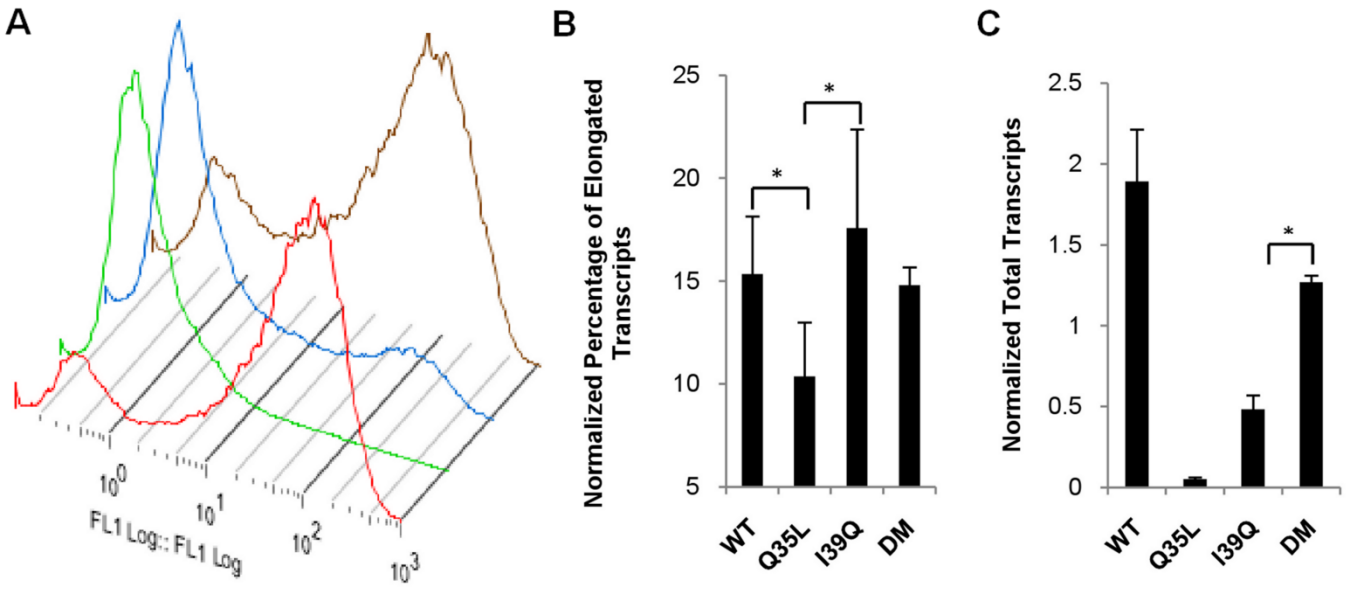


Figure 5

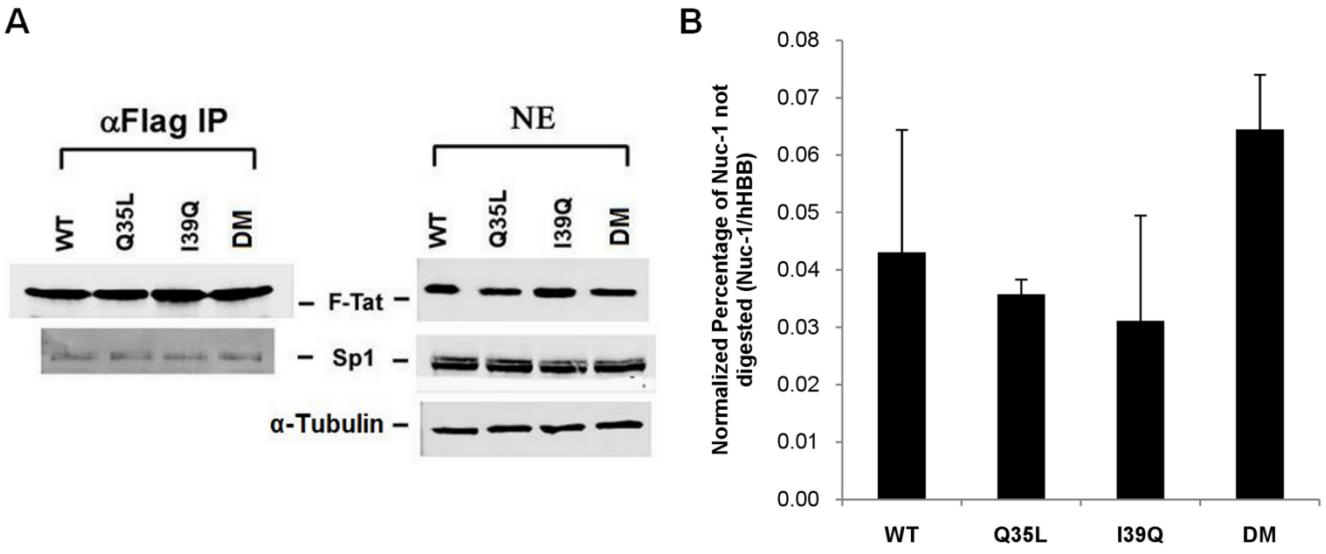


Figure 6