Title
Ubiquitin signaling regulates multiple steps in Tat-dependent HIV-1 transcription

Author
Faust, Tyler Bischoff

Publication Date
2016

Ubiquitin signaling regulates multiple steps in Tat-dependent HIV-1 transcription

by

Tyler B. Faust

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biochemistry and Molecular Biology

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Acknowledgements

When I entered graduate school, I was rather naïve. I expected a significant amount of direct guidance, given that is how I had experienced science to that point. As an undergraduate researcher for three years, I had become comfortable with day-to-day experimentation (e.g. running gels, mini-preps). However, I did not realize I was not yet a polished scientific thinker. All of this changed after joining the Frankel lab.

I want to start by thanking Alan Frankel who is responsible for my maturation as a scientific thinker. Alan's lab is dedicated to understanding how the HIV-1 proteins Tat and Rev regulate viral replication. However, he allows the members of his lab to venture out and explore broader questions that interest them. This mentoring style inherently drives everyone to become the expert on his or her project. It can also lead to times of feeling lost as you develop the project, yet Alan is always generous with his time to evaluate where things stand and where you should be going. Alan may not be an expert in the new fields of study brought to the lab by this process, but he always provides critical analysis of all projects. He has a keen ability to extract the underlying logic of every project in the lab, whether it is an X-ray crystal structure or next generation sequencing data. In this way, he can provide feedback on how sound and how well-controlled your conclusions are. Over many years of practicing science with Alan's mentorship, everyone in the lab becomes a great scientific thinker.

The Frankel laboratory was an incredible environment to work for seven years. Everyone in the lab had a great personality and was willing to offer advice on experiments and ideas. I also developed my scientific thinking from members of the lab, and I want to especially thank David Booth and Jason Fernandes in this regard. David

Booth combined a broad interest in many biological processes with an ability to distill these processes into their fundamental physical and chemical properties. I had not encountered anyone else at UCSF with this perspective, and it is one that I hope to develop more in myself. Jason Fernandes was whom I sought out when I wanted feedback on an idea or model. He was very well-read in many fields and had an incredible memory of the primary literature. Both David and Jason were a great resource in the lab, and I value the time I spent with them.

I want to thank my thesis committee of Raul Andino, J.J. Miranda, and Geeta Narlikar for providing valuable input as my project progressed. It is easy to lose perspective when you are swimming in the details of your project. I learned from my committee that it is essential to step back occasionally and ask yourself “what I am working on and is it interesting?”. Even more, the currency of science is publications and it is useful to periodically examine your work and map out how it will eventually become a paper.

My classmates were a wonderful group of people that I enjoyed spending time with outside of lab. I want to thank Lindsey Pack, Coral Zhou, Sam Liang, Isabel Nocedal, Alison Leaf, Trevor Sorrells, Doug Tischer, and Scott Coyle for being interesting people that made me laugh. We all made an effort to regularly get together, even as some of us entered our seventh year in graduate school. Obtaining a Ph.D. is difficult and discouraging at times, so developing these friendships with my classmates was key to persevering. In particular, I want to thank Lindsey Pack for her incredible support. Lindsey has several great qualities that make her a wonderful person, but I want to highlight that she is a great listener. It seems like no matter what she is in the

middle of, if you come to her with a question or you need advice she will immediately focus on you. Not only is she a great listener, but she always has great input, whether that input is personal or scientific.

As has been true the rest of my life, my mom and dad are always there for me. There are many times in science when you feel like you have gone down the wrong track and wasted months of time. In the midst of these doldrums, my parents would unknowingly calm me by telling me how proud they were just that I was pursuing my Ph.D. Their support was an essential comforting influence on me throughout my graduate work. My parents, along with my mother-and father-in-law, also provided immense help once my daughter, Adriana, was born. It is still quite difficult to pursue an advanced degree in science and start a family at the same time. Luckily for my wife and I, our parents went out of their way to help raise our daughter while we pursued our passion. Neither my wife nor I could have continued our education without this help.

Finally, I want to thank my daughter, Adriana, and wife, Silvi. I didn't enter graduate school expecting to meet my wife, but I guess you are never expecting it. However, meeting Silvi caught me off guard as we became very close very quickly. I have never been more connected to another person than I am with Silvi. We share an odd sense of humor, and that allows us to deflate the pressure of living the life of a scientist and raising a child. Silvi is an incredible scientist and a great resource for my own work. I didn't know what to expect when my daughter was born. During Silvi's pregnancy, I was excited but the whole experience was a bit of an abstraction given my focus on completing my work. However, seeing your child for the first time is not an abstraction, and I was overwhelmed by a sense of responsibility and awe. Raising

Adriana has been a joy and a challenge, and I am excited every day to see how she will surprise me.

Acknowledgement of previously published work in this thesis.

Chapter 2 was published in the following scientific journal:

Fernandes JD, Faust TB, Strauli NB, Smith C, Crosby DC, Nakamura RL, Hernandez RL, Frankel AD. Functional Segregation of Overlapping Genes in HIV. Cell. 2016; 167: 1-12.

Abstract

Ubiquitin signaling regulates multiple steps in Tat-dependent HIV-1 transcription

by

Tyler B. Faust

HIV-1 gene expression has provided an excellent model of transcription elongation control. Transcription complexes that assemble at the HIV-1 promoter efficiently initiate transcription but generate paused RNA polymerase II downstream from the start site. The virally encoded Tat protein hijacks P-TEFb to phosphorylate and activate this paused polymerase. Tat binds additional host proteins, but it is unclear if and how they regulate RNAP II elongation. Further, Tat is a particularly small, natively unfolded protein, so it is puzzling how this viral transcription factor is able to bind and redirect the activity of multiple, large host complexes. In this thesis we address these questions by first identifying all Tat amino acids that are essential for its transcriptional activity. The analysis of HIV-1 sequences from infected patients, the activities of an entire set of Tat alanine point mutants, and the replication efficiencies of a randomly mutagenized library of Tat alleles cloned into the virus reveal that Tat and Rev segregate their functional residues in their overlapping coding sequence. Next, a candidate RNAi screen of previously identified novel Tat host interactors reveals that multiple proteins involved in ubiquitin conjugation are essential for Tat activity. Using *in vivo* purifications and *in vitro*

reconstitution, we show that PJA2, a RING-H2 E3 ubiquitin ligase, directly ubiquitinates Tat. Extensive mutagenesis of Tat and ubiquitin expression constructs reveals both that multiple lysines in Tat can serve as ubiquitin acceptor sites and that multiple lysines in ubiquitin can be used to generate polyubiquitin chains. This immense plasticity of ubiquitin signaling through Tat is likely a way for the virus to encode a transcriptional activity that is robust to a high mutation rate. Using a similar approach, we then demonstrate that Tat hijacks the UBE2O ligase to robustly ubiquitinate specific proteins in the 7SK snRNP. Surprisingly, this ubiquitination occurs in the cytoplasm and functions to remodel the snRNP complex. Remarkably, Tat targets the cytoplasmic 7SK snRNP, releasing P-TEFb from its inhibitor, resulting in nuclear import of the free kinase to activate transcription. This re-localization provides a unique paradigm for P-TEFb control.

Table of Contents

List of Tables

Chapter 3: PJA2 ubiquitinates the HIV-1 Tat protein with atypical chain linkages to activate viral transcription

List of Figures

Chapter 2: Functional Segregation of Overlapping Genes in HIV

Chapter 3: PJA2 ubiquitinates the HIV-1 Tat protein with atypical chain linkages to activate viral transcription

Chapter 1: Introduction

Human immunodeficiency virus type 1 (HIV-1) is a retrovirus that infects an estimated two million people per year. Specifically, the viral capsid is surrounded by an envelope that interacts with the CD4 receptor of a subset of T cells resulting in fusion of the virus with the cell membrane. If left untreated, an infected individual can develop acquired immune deficiency syndrome (AIDS), which increases the risk of opportunistic infections. HIV-1 contains a small 9 kilobase, positive-sense RNA genome that encodes only nine genes. Because of its small size, the virus relies extensively on host protein complexes to complete infection. Many drugs have been developed that target different stages of the viral life cycle, and when used in combination can significantly decrease viral titers in the blood. However, this treatment, referred to as highly active antiretroviral therapy (HAART), is only a functional cure as the virus is not eradicated from the body, and treatment must be continued throughout the lifetime of the infected individual. Being a retrovirus, HIV-1 reverse-transcribes a DNA copy of its genome, which is subsequently integrated into the host cell genome. This process of integration permanently ligates the virus into the genetic material of the infected cell and is the major reason full eradication is not currently possible.

Once the HIV-1 provirus is integrated, the cellular RNA polymerase II (RNAP II) assembles at the viral promoter encoded in the 5' LTR and transcribes the viral RNA genome. The HIV-1 promoter contains *cis* elements for host transcription factors, including three Sp1 sites and two NF-κB sites, which provides an entirely cell-intrinsic mechanism to activate transcription, especially in activated T cells targeted by HIV-1 where the NF-κB pathway is induced (1–3). However, despite this appropriation of cellular mechanisms to increase gene expression, transcription complexes that

assemble at the HIV-1 promoter predominantly generate short transcripts of approximately 60 nucleotides (4). To bypass this block to transcription elongation, HIV-1 encodes it own transcription factor, Tat, which hijacks the cellular kinase positive transcription elongation factor b (P-TEFb) composed of Cyclin T1 (CCNT1) and Cyclin-dependent kinase 9 (CDK9) subunits, to phosphorylate and activate RNAP II that is paused at the viral promoter (5,6). The Tat-P-TEFb complex potently stimulates gene expression, initiating the post-integration steps of the life cycle, which eventually leads to viral budding and the infection of new cells. Despite both cellular and virally encoded means to activate transcription, individual T cells can nonetheless enter a state of latent infection in which no or minimal viral RNA is produced (7). This latent reservoir is responsible for persistent HIV-1 infection, despite aggressive HAART treatment.

The prevalent model by which latency is established is that activated $CD4^+$ T cells, which have been infected by HIV-1, revert to a resting, memory state where inducible transcription factors that also aid viral transcription are no longer activated (7). However, recent work has demonstrated that the return of activated T cells to a resting state is not sufficient to explain the transition to latent infection (8,9). Further, Tat increases its own expression by activating the LTR, and this positive feedback loop buffers HIV-1 transcription to changes in cellular state (8). Once latency is established, the integrated virus remains dormant in the infected host genome until future stimulation by antigen results in viral production. Therefore, in addition to vaccine development and the use of broadly neutralizing antibodies to cure infection (10), current efforts are also directed at developing drugs that reverse latency by inducing viral gene expression, which would purge the latent reservoir by viral antigen presentation and immune

clearance pathways (11). Critically, it has also been shown that Tat more potently induces viral gene expression from a dormant LTR compared to a suite of cell-state modifiers typically used as latency-reversing agents (8). Therefore, understanding the mechanisms by which Tat activates transcription would immensely aid the effort to develop highly active compounds that reverse latency.

In an effort to completely define the set of host interactions for all HIV-1 proteins, a collaborative effort at UCSF recently performed comprehensive affinity purification-mass spectrometry (AP-MS) of tagged viral proteins (12). Identifying host protein targets globally could eventually assist in the development of new HIV-1 therapeutics for any stage of the life cycle. Tat proteomics uncovered the well-established P-TEFb, 7SK snRNP, and super elongation complex interactions, however sixteen novel interactors were also identified. The 7SK snRNP is an RNA-based complex that sequesters P-TEFb in an inactive state (13,14) in which the HEXIM1 protein inhibits the kinase activity of CKD9 in an RNA-dependent manner (15,16). Previous models of Tat activity had proposed that Tat, which is an RNA-binding protein, directly recruits the P-TEFb heterodimer to a nascent, structured RNA (TAR) associated with RNAP II (17). However, we have recently shown that 7SKsnRNP-inhibited P-TEFb is recruited to the HIV-1 promoter (18), which provides a simple explanation of why RNAP II transcription is non-processive basally. When Tat protein is made, it can also bind the HIV-1 promoter with the 7SK snRNP, even in the absence of TAR, which demonstrates that the viral RNA is not necessary for recruitment of Tat or P-TEFb. Once the TAR RNA is transcribed by RNAP II, the proteins of the inhibitory 7SK snRNP are displaced from DNA-bound transcription complexes, which activates the kinase activity of P-TEFb to

phosphorylate RNAP II. It appears that the Tat-TAR RNA-P-TEFb complex functions as a viral mimic of the HEXIM1-7SK snRNA-P-TEFb complex. It has also been shown that Tat can displace HEXIM1 from the snRNP, forming a Tat-7SK snRNA-P-TEFb complex (19), which might itself be recruited to the viral promoter. TAR expression would then provide a handoff for Tat from the cellular RNA to the viral RNA, completing the displacement of the inhibitory complex.

The goal of this thesis was to interrogate the function of the novel Tat-interacting host proteins and determine how they might regulate distinct steps of the newly proposed activation model. To facilitate a new mechanistic understanding, it was also important to understand which residues in Tat were critical for its transcriptional activity. Importantly, the coding sequence for Tat is overlapping with the HIV-1 Rev sequence, making it difficult to determine essential Tat residues by conservation alone, as the overlapping genetic information could experience selection pressure from both reading frames. In Chapter 2 of this thesis, a combination of patient data analysis, alanine-scanning mutagenesis, and randomized mutational libraries in the context of viral replication reveal that the Tat and Rev proteins segregate their functional residues in the overlap, in which important sequence for one protein overlaps non-functional sequence for the other protein. This exhaustive analysis also provides a comprehensive examination of all functional Tat residues.

Chapters 3 and 4 explore the function of the novel Tat host interactors. The foundation for both chapters is a targeted RNAi screen of the sixteen new proteins from proteomics coupled to a functional assay for Tat activity. This screen revealed that nine host factors are required for full Tat transactivation, with an unexpected enrichment for

proteins involved in ubiquitin conjugation pathways. Interestingly, the ubiquitin modifications detailed in the thesis are non-degradative and instead serve positive signaling functions. Chapter 3 describes how a RING-H2 finger E3 ubiquitin ligase PJA2 ubiquitinates Tat and how this modification regulates transcription elongation at the HIV-1 promoter. Surprisingly, many lysines in Tat can function as acceptor sites for PJA2. This result is supported by the analysis of Chapter 2, which shows that the plastic ubiquitination sites in Tat are not highly conserved. Chapter 4 explores how the hybrid E2-E3 ubiquitin ligase UBE2O forms a complex with two other Tat interactors, ZFP91 and NAP1L4 (ZUN complex). UBE2O ubiquitinates the HEXIM1 and MEPCE proteins of the 7SK snRNP, which leads to a remodeling of the complex. Several highly conserved and functionally important Tat residues identified in Chapter 2 are surface-exposed in the Tat-P-TEFb crystal structure (20) and do not make contact with either CCNT1 or CDK9. Two of these Tat amino acids, Lys28 and Phe32, are required for the proper binding of 7SK snRNP and ZUN proteins. Chapter 4 also reveals a significant fraction of P-TEFb-7SK snRNP in the cytoplasm. Tat targets this cytoplasmic pool of the snRNP to release P-TEFb for nuclear import. In summary, the thesis provides a detailed analysis of the functional landscape of Tat sequence space and integrates these findings with a demonstration of how multiple novel Tat host factors support full transactivation.

References

1. Chen BK, Feinberg MB, Baltimore D. The kappaB sites in the human immunodeficiency virus type 1 long terminal repeat enhance virus replication yet are not absolutely required for viral growth. J Virol. 1997;71(7):5495–504.
2. Ross EK, Buckler-White a J, Rabson a B, Englund G, Martin M a. Contribution of NF-kappa B and Sp1 binding motifs to the replicative capacity of human immunodeficiency virus type 1: distinct patterns of viral growth are determined by T-cell types. J Virol [Internet]. 1991;65(8):4350–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=248874&tool=pmcentrez&rendertype=abstract
3. Kang S, Tran A, Grilli M, Lenardot MJ. NF-KB Subunit Regulation in Nontransformed CD41 T Lymphocytes. Science (80-). 1992;256(June):1452–6.
4. Lassen KG, Bailey JR, Siliciano RF. Analysis of Human Immunodeficiency Virus Type 1 Transcriptional Elongation in Resting CD4+ T Cells In Vivo. J Virol. 2004;78(17):9105–14.
5. Wei P, Garber ME, Fang SM, Fischer WH, Jones KA. 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. 1998;92(4):451–62.
6. Feinberg MB, Baltimore D, Frankel AD. The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation. Proc Natl Acad Sci U S A [Internet]. 1991;88(9):4045–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51590&tool=pmcentrez

&rendertype=abstract

7. Siliciano RF, Greene WC. HIV latency. Cold Spring Harb Perspect Med. 2011;1(1).

8. Razooky BS, Pai A, Aull K, Rouzine IM, Weinberger LS. A hardwired HIV latency program. Cell [Internet]. Elsevier Inc.; 2015;160(5):990–1001. Available from: http://dx.doi.org/10.1016/j.cell.2015.02.009

9. Chavez L, Calvanese V, Verdin E. HIV Latency Is Established Directly and Early in Both Resting and Activated Primary CD4 T Cells. PLoS Pathog [Internet]. 2015;11(6):e1004955. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4466167&tool=pmcentrez&rendertype=abstract

10. Kwong PD, Mascola JR, Nabel GJ. Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning. Nat Rev Immunol [Internet]. Nature Publishing Group; 2013;13(9):693–701. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23969737\nhttp://www.nature.com/doifinder/10.1038/nri3516

11. Margolis DM, Garcia JV, Hazuda DJ, Haynes BF. Latency reversal and viral clearance to cure HIV-1. Science [Internet]. 2016;353(6297):aaf6517. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27463679

12. Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global landscape of HIV-human protein complexes. Nature [Internet]. 2012;481(7381):365–70. Available from: http://dx.doi.org/10.1038/nature10719

13. Nguyen VT, Kiss T, Michels AA, Bensaude O. 7SK small nuclear RNA binds to

and inhibits the activity of CDK9/cyclin T complexes. Nature [Internet]. 2001;414(6861):322–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11713533\nhttp://dx.doi.org/10.1038/35104581

14. Yang Z, Zhu Q, Luo K, Zhou Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature [Internet]. 2001;414(6861):317–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11713532\nhttp://dx.doi.org/10.1038/35104575

15. Kobbi L, Demey-Thomas E, Braye F, Proux F, Kolesnikova O, Vinh J, et al. An evolutionary conserved Hexim1 peptide binds to the Cdk9 catalytic site to inhibit P-TEFb. Proc Natl Acad Sci [Internet]. 2016;113(45):201612331. Available from: http://www.pnas.org/lookup/doi/10.1073/pnas.1612331113

16. Michels A a, Fraldi A, Li Q, Adamson TE, Bonnet F, Nguyen VT, et al. Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. EMBO J. 2004;23(13):2608–19.

17. Peterlin BM, Price DH. Controlling the Elongation Phase of Transcription with P-TEFb. Mol Cell. 2006;23(3):297–305.

18. D'Orso I, Frankel AD. RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. Nat Struct Mol Biol. 2010;17(7):815–21.

19. Muniz L, Egloff S, Ughy B, Jády BE, Kiss T. Controlling cellular P-TEFb activity by the HIV-1 transcriptional transactivator tat. PLoS Pathog. 2010;6(10).

20. Tahirov TH, Babayeva ND, Varzavand K, Cooper JJ, Sedore SC, Price DH. Crystal structure of HIV-1 Tat complexed with human P-TEFb.pdf. Nature [Internet]. Nature Publishing Group; 2010;465(7299):747–51. Available from: http://dx.doi.org/10.1038/nature09131

Chapter 2: Functional Segregation of Overlapping Genes in HIV

Summary

Overlapping genes pose an evolutionary dilemma as one DNA sequence evolves under the selection pressures of multiple proteins. Here, we perform systematic statistical and mutational analyses of the overlapping HIV-1 genes *tat* and *rev,* and engineer exhaustive libraries of non-overlapped viruses to perform deep mutational scanning of each gene independently. We find a “segregated” organization in which overlapped sites encode functional residues of one gene or the other, but never both. Furthermore, this organization eliminates unfit genotypes, providing a fitness advantage to the population. Our comprehensive analysis reveals the extraordinary manner in which HIV minimizes the constraint of overlapping genes, and repurposes that constraint to its own advantage. Thus, overlaps are not just consequences of evolutionary constraints, but rather can provide population fitness advantages.

Introduction

The sequencing of the first complete DNA genome, ΦX174, revealed the startling discovery that genes can overlap with one another (1,2). Since this initial observation, overlapping reading frames have been observed in most viruses and across all domains of life (3–5). In viruses, these regions are traditionally thought to arise as consequences of error prone polymerases and constraints on the size of viral capsid proteins (3,6). For instance, high polymerase error rates favor short genomes thereby decreasing the probability of catastrophic mutations, while the viral capsid imposes a biophysical limit on genome size. Other models suggest that overlap formation is driven by selection pressures favoring evolutionary innovation (7–10), as overlaps are also found in large genomes. Regardless, once present in a genome, overlapping genes must balance

nucleotide usage so that the functions of each reading frame are satisfied. Several studies have used computational methods to estimate gene-wide selective forces (11–13) but only a few have generated experimental data (14). Computational analyses of protein structure have demonstrated that overlapped proteins in all viruses tend towards intrinsic disorder (10), but how structured and/or functional regions are divided at the amino acid level remains unknown. It is possible to envisage two extreme models for this simultaneous evolution: 1) a “segregated” model in which the amino acid/nucleotide preferences for one gene dominate and the other gene accommodates with no observable benefit to itself, or 2) a “shared” model in which both genes exert selective forces at the same site, enforcing strong conservation in both frames (Figure S1). It is unlikely that segregated or shared decisions are uniform over an entire overlap, so defining the selective forces on a per residue basis becomes critical to understanding how the functions of a pair of proteins can be properly balanced.

HIV-1 provides a compelling model as it contains eight distinct areas of coding overlap (Figure S2A) constituting ~8% of its entire genome, and extensive sequence information from many virus isolates is available (15) (www.hiv.lanl.gov). The *tat* and *rev* regulatory genes (Figure 1A) are a particularly interesting case as both are essential for virus replication and thus experience strong simultaneous selective pressure, both have well-established functions and assays, and both have partial structures available to help interpret the functional consequences of sequence variants (Figure S2B) (16–18). Tat activates transcriptional elongation at the HIV-1 promoter via its interactions with host transcription factors (most notably P-TEFb) and an RNA element at the 5’ end of viral transcripts known as TAR (trans-activation response element) (19). Rev facilitates the

nuclear export of partially spliced and unspliced viral RNAs that encode essential late-stage viral proteins and genomic RNA for packaging (20). Rev binds as an oligomer to an RNA element present in viral introns known as the RRE (Rev response element) and guides the RNAs to the cytoplasm via interactions with the Crm1 nuclear export machinery.

In order to understand the consequences of the *tat/rev* overlap for viral evolution, we compare sequence conservation in patient isolates to comprehensive residue-by-residue functional maps of Tat and Rev generated by alanine scanning. We further compare these datasets to replication experiments in viruses in which we removed the *tat/rev* overlap and subsequently measured the experimental fitness of every amino acid at each position in each protein independently. We find that HIV-1 has evolved in a segregated manner - to separate functionally important amino acids for Tat and Rev - and moreover, the arrangement decreases sampling of unfit genotypes, thereby turning an apparent genetic constraint into a fitness advantage. The combination of these orthogonal datasets provides the most complete picture of an overlap to date, and demonstrates another way in which HIV-1 has efficiently utilized the coding capacity of its small genome to optimally arrange its core regulatory machinery.

Results

Overlapped Regions in HIV-1 are Not More Conserved than Single Frame Regions

In order to explore the constraints imposed in the *tat/rev* overlap, we first examined the conservation of each protein residue in the HIV-1 proteome using patient data from the Los Alamos Database (hiv.lanl.gov). These datasets comprise ~2,000 HIV-1 sequences

in high-quality alignments. We then calculated the Shannon entropy for each residue of each HIV-1 protein. We used this relatively simple metric – in which a low value indicates high amino acid conservation - as it requires no assumptions about selection or substitution rates, which are challenging to approximate in overlapped regions. Furthermore this metric has previously been used to quantify diversity of viral sequences, to identify interaction surfaces on proteins, and analyze overlapping reading frames (21,22). We examined each residue in every HIV-1 gene and compared the single, double, and triple reading frame regions to one another (Figure S2C), and discovered that overlapped regions were, surprisingly, not more conserved than single frame regions. This behavior held for almost every HIV-1 gene (Figure 1B), though this may reflect inherent biases in the structural features of these proteins as opposed to the presence of dual coding. Regardless, these results demonstrate that regions of overlap do not experience high conservation, providing evidence against a shared organization (with the caveat that positive selection and variations in mutational tolerance can confound this analysis). Genome-wide selection analyses using sophisticated evolutionary rate calculations have produced results consistent with this data, although the inability to accurately estimate neutral substitution (or determine selective forces on a per-residue basis) have prevented these studies from definitively interpreting these findings (23,24).

Classification of Overlapped Residues as Segregating or Sharing Based on Entropy

To narrow our focus and analyze specific residues within Tat and Rev, we defined a statistic, normalized-to-mean entropy (NME), in which the entropy of each residue was

normalized to the average entropy of the one-frame regions of the viral genome. We reasoned that amino acid conservation within single-frame regions reflects a simple requirement for protein function without the confounding constraints of an overlapping gene. Thus a NME < 1 would be suggestive of functional importance even within overlapped regions. We calculated the NME values of all residues from Tat and Rev that share two nucleotides within a codon - for example, Y47 of Tat (tAT) shares two nucleotides with M1 of Rev (ATg) - and plotted Rev NME versus Tat NME (Figure 1C). By separating the plot into quadrants, we can speculatively classify residues according to their conservation for one protein alone (Tat or Rev), for both (Shared), or for neither. Thus, we can tentatively hypothesize that residues highly conserved in only either the Tat or Rev reading frame are segregated, devoted to a single protein while residues conserved in both frames may share functionally important positions. However, this naïve classification requires detailed experimental testing as it does not attempt to estimate a neutral rate of substitution or consider the extent to which the organization of the genetic code may constrain the codon possibilities at the overlapped positions.

Functionally Important Residues in Tat and Rev are Segregated

To evaluate which amino acids in Tat and Rev are conserved for function versus those restricted by codon usage in the alternative frame, we generated complete sets of alanine point mutants for both proteins and tested their activities in separate cell-based reporter systems, effectively removing the coding constraint of the overlap. Tat activity was measured by its ability to enhance transcription of an integrated firefly luciferase gene under the control of the HIV-1 promoter (25) (Figure S3A), and Rev activity was measured by co-transfecting each mutant with a CMV-driven, RRE-containing, *gag-pol*

reporter that produces viral capsid (p24) in a Rev-dependent manner (26) (Figure S3B). Heat maps (Figure 2A-B) and graphs (Figure S3C-D) show the relative activities of each mutant, normalized to a reference sequence (HXB2 & NL4-3 isolates), and highlight the functionally important residues, which correspond well to established motifs of these proteins. As expected, mutation of conserved residues among patient isolates in the single-frame region of Tat (Figure 1A) correlated well with loss of function (Figure 2C). The remainder of *tat* overlaps with *rev*, *env*, or *vpr*, whereas *rev* overlaps with *tat* or *env* over its entire length. Within the *tat/rev* overlapped region, mutation of residues conserved in either only Tat or only Rev correlated with their respective loss of function, whereas mutation of residues not conserved in either protein generally did not affect function (Figure 2C). Somewhat surprisingly, however, mutation of residues that show conservation in both frames generally correlated with loss of function of only one of the proteins (Figure 2C).

To examine more closely if residues in this overlapping region might be simultaneously important for both proteins, we plotted the activities of Tat and Rev mutants against one another and found, remarkably, that almost no sites contributed to both functions, with one exception (Figure 2D, “Shared” quadrant). This result stands in stark contrast to the naïve expectation from the entropy analysis (Figure 1D vs Figure 2D). This functional segregation at the single amino acid level, despite the overlap, is readily apparent when the important functional amino acids are mapped onto their primary structures (Figure 2E). The lone exception is the overlap between Tat Y47 and the Rev initiation codon M1, where Y47 is important for Tat activity and M1 is important only for Rev protein synthesis, as mutating the Rev start codon in the context of an N-terminally tagged Rev

has no effect on activity (Figure S3D). The striking segregation of functional residues suggests that Tat and Rev conform to a segregated model of evolution.

Deep Mutational Scanning and Selection of Viruses with Uncoupled *tat* and *rev* Genes

Our patient dataset contained several sites that were conserved in both the *tat* and *rev* frames. Our functional data suggests that this conservation is not a consequence of simultaneous selection for the function of both proteins, but rather reflects the requirements of only a single protein with subsequent restriction imposed by the genetic code in the alternative frame. If this is true, then we would expect the conserved but non-functional sites to mutate more freely absent the constraint of the overlap. To examine this, we created viruses (NL4-3 isolate) in which the *tat and rev* genes were engineered into different parts of the viral genome by first mutating either the *tat* or *rev* start codon and introducing downstream stop codons, synonymous in the alternative frame, to ablate the endogenous locus. We then generated exhaustive libraries of *tat* or *rev* variants in which every individual codon was randomized, and cloned the resulting codon library into the non-essential *nef* region of the genome (Figure S4A). This resulted in 202 virus pools (116 for *rev* and 86 for *tat*), each composed of 64 different alleles. We performed independent, competitive replication experiments with each pool and determined allele frequencies, pre- and post-selection, by deep sequencing (Figure S4B).

Population Genetics Modeling Allows Inference of Significant Selection

In order to infer selection from the changes in allele frequency, we developed a Wright-Fisher based model (27,28) using population numbers from our own replication data

(Figure S4B). This model allowed us to estimate the experimental fitness for each allele and statistically test if the allele was significantly under selection (given experimental error and the possibility of neutral variation) (Figure S5A). This method has an overall high correspondence (Figure S5B) with logarithmic fitness values often calculated in deep mutational scanning experiments (29), but has the advantage that statistical testing, rooted in population genetics principles, can be performed to assess confidence in selection values (Figure S5C). We used this model to calculate the mean selection coefficient (experimental fitness) for each amino acid allele (grouping synonymous codons) to produce non-overlapped mutational profiles of Tat and Rev (Figures 3 and 4). Amino acid experimental fitness profiles were clustered by chemical properties (30) to help visualize side chain preferences at every position, and the resulting landscapes were compared to the overlapped patient frequencies (Figures 3 and 4).

Additionally, we identified a subset of alleles from the selection data that displayed consistent and statistically significant selection (Figure S5D). The complete results of this analysis are provided (Supplemental Table 1) as a resource for further study of individual alleles and comparison to existing deep mutational scanning datasets (31).

Disordered Residues Mutate Freely in the Selection Dataset

Strikingly, the patient datasets show much stricter conservation than the selection datasets (Figures 3 and 4). This may in part reflect different selective forces in patients (i.e. immune system pressure, virus latency, or length of selection) but this phenomenon is commonly observed in deep mutational scanning experiments, where codon randomization is able to sample sequence space much more rapidly than natural evolution (29,32). Regardless, both proteins show strong selection within their known

functional motifs (Figures 3 and 4; gold is positive selection and blue is negative) whereas the linker regions show generally neutral variation (white). In order to map the data to the protein structures, we used the median experimental fitness for each residue to approximate the strength of selection for each residue (Figure S6; low median value = high selection), and plotted these values onto the structures of Tat and Rev (Figure 5). Interaction and folding surfaces between Tat and P-TEFb (Cysteine/Core) and TAR (ARM) are readily visible as are surfaces between Rev, RRE (ARM), Crm1 (NES) and other Rev molecules (OD). In sharp contrast, the disordered regions experience a large amount of non-synonymous, but neutral substitution. Notably these regions do not express strong signals of positive selection even absent the constraint of the overlap, suggesting that the presence of an overlapped gene does not prevent the formation of additional, extended interaction surfaces. Interestingly, the neutral selection values for stop codon alleles at residue 67 of Tat and residue 86 of Rev indicate that the C-termini of both proteins are dispensable for replication in this context. In some cases, we observed shifts from the NL4-3 reference sequence to the patient consensus (e.g. Rev P28Y), suggesting that NL4-3 sequences are not fully optimized for individual protein function. We also identified important alanine residues not captured by alanine mutagenesis, such as Tat A42, which showed strong conservation of small side chains (A/G) in both patient and selection datasets. Positions such as this, as well as the fact that experimental fitness of alanine variants does not always match the median experimental fitness, suggests that alanine scanning, while grounded in biophysical reasoning, may not always provide the best measure of functional importance of any particular site.

All Three Datasets Exhibit Signatures of Segregated Evolution

To assess the agreement between datasets, and compare them to our models of segregated and shared evolution, we grouped alleles into two categories: those present in the patient dataset (frequency >1%) and those absent (frequency = 0%). In a segregated model of evolution, a dominant gene (Figure 6A: grey gene), should act similarly regardless of whether it is overlapped or not, as its own requirements for protein function drive selection at these nucleotides. In a single-frame context, the dominant gene should continue to match the overlapped dataset with absent alleles remaining unfit, and present alleles remaining fit. In contrast, the accommodating gene (black) should have many alleles that are absent in the patient population (due to functional requirements of the grey gene) but are actually fit in a single frame context. In a shared model, the selective pressures for the functions of both proteins constrain the nucleotide sequence and thus removing the overlap provides only a minimal relaxation. At best, only a small number of chemically conservative amino acids, absent in the overlapped dataset due to the alternative frame, will be fit (Figure 6A: 0% long tail). The single frame region of Tat provides a good test for this behavior as Tat, the only gene present, should act in a segregated manner in both patients and our randomized selection dataset. As expected there is a strong correlation between patient allele frequency and selection (Figure 6B) with absent alleles unfit and present alleles fit. The correlation is much weaker in the multiple frame regions largely due to a subset of alleles that produce fit viruses but are absent in patients, presumably due to the constraint of the overlap (Figure 6A). These results are consistent with our segregated model, with Rev acting as the dominant gene over the majority of sites in the overlap

(Figure 6B; lower panel). However, the degree to which each gene dominates varies over the region of overlap.

To compare the selection dataset to the functional dataset, we once again used the median experimental fitness for each residue to approximate overall strength of selection at each position in each protein (Figure 6C). We then segregated sites into those identified as functionally important by the alanine scan (<50% relative activity) and those identified as unimportant (>90%). As expected those sites identified as important for both Tat and Rev, have strong signals of selection, while those identified by the alanine scan as unimportant for function have median experimental fitness near zero, suggesting no preference for any particular amino acid (Figure 6C).

Comparison of Single Frame and Overlapped Viral Mutational Profiles

Although the segregation of function for Tat and Rev helps minimize the evolutionary penalty of the overlap, we wondered whether there might be measurable benefits to overlapping genes as opposed to encoding each gene in a single frame. Simulation studies have hypothesized that simply decreasing the length of a genome by introducing overlaps can increase viral fitness by decreasing the probability of detrimental mutations introduced by polymerase (3). In addition, the actual coding requirements of each protein can provide additional buffering by decreasing the number of detrimental allelic combinations. Specifically, an overlapped virus has, by definition, a limited number of possible genotypes compared to its single frame equivalent. This restriction can result in an overlapped viral fitness landscape that is, proportionally, more, less, or equivalently fit when compared to a virus with both genes in single frames (Figure 7A). The manner in which the overlap restricts the landscape depends on each

gene's functional requirements and the genetic code: detrimental restriction results when fit alleles in one gene encode unfit alleles in the alternative frame thereby preventing the fit allele from existing in the population. Conversely, gain of fitness restriction occurs when fit alleles are unable to encode unfit alleles in the alternative frame. This arrangement can effectively purge unfit viral genotypes, in which the fit allele is sabotaged by an unfit allele in the other gene, from the genotypic landscape of the virus.

To examine how HIV-1 Tat and Rev affect each other, we used the selection data from the non-overlapped viruses to approximate the differences in the fitness distributions of both overlapped and non-overlapped viruses. For example, if both genes were encoded in a single frame context, viruses with the Tat Y47 allele would not affect the composition of the Rev allele at position 1, resulting in 21 possible Rev alleles (including stop codons). In the overlapped context, however, Tat Y47 restricts Rev codons to M1, I1, and T1, resulting in the loss of 18 viral genotypes (e.g. Tat Y47, Rev G1 cannot exist in an overlapped context). Importantly, the codon restrictions are not the same in both directions, as fixing the Rev M1 allele constrains Tat alleles (Y, H, D, and R) at position 47 in a different manner. Therefore we approximated the effect of each gene's constraint on the other in both directions and compared the unconstrained single-frame context to the overlapped case (Figure 7B-C). Because most sites in the overlap actually restrict two adjacent codons in the alternative frame (for example, Tat G48 influences both Rev M1 and A2), we made a simplifying assumption that the experimental fitnesses of the two amino acids were additive, and approximated the experimental fitness of the overlapped protein (Figure S7). Finally, as both Rev and Tat

are required for viral replication, we approximated the overall viral fitness as the lower value of either the Rev or Tat allele.

Tat Restricts Rev to Increase the Fitness of the Overlapped Viral Landscape

Examining the restriction Tat imposes on Rev (Figure 7B) reveals a large population of potential genotypes in the single-frame context, with only a small fraction of fit alleles. By overlapping the genes, the total number of possible phenotypes is drastically reduced; however a disproportionate number of unfit combinations are purged. Explicit calculation of the genotypic proportions shows that this restriction results in a doubling of the proportion of fit genotypes (Figure 7B). In essence, the overlap takes advantage of subtle preferences in the alternative frame to reduce the number of unfit genotypes. For example, the preference for R53 in Tat biases the overlapping but relatively tolerant residue 9 of Rev towards either D or A, but away from the highly unfit R, K, and stop codons. In contrast, Rev's constraint on Tat does not significantly alter the overall percentage of fit viruses (Figure 7C), most likely because the majority of the functional motifs in Rev overlap with the readily mutable C-terminus of Tat. Thus, the genomic arrangement and functional amino acid requirements of Tat and Rev appear to confer a distinct selective advantage to the virus.

Discussion

Overlapping genes were originally considered unfortunate consequences of other selection pressures such as high polymerase error rates and capsid size (3,6). However more recently, this dogma has been questioned (7) with gene delivery experiments demonstrating that HIV-1 is capable of packaging genomes of considerably larger sizes (33), and with polymerase error appearing to insufficiently explain overlap proportion

(6). Indeed, the creation of fit, uncoupled viruses in this work and others (34,35) makes explanations of overlaps as simply negative consequences of pressures on genome size highly unsatisfactory.

In our investigation of this problem, we initially hypothesized two simple models for how overlapped genes might constrain one another: a segregated model in which one protein drives evolution with little regard for the other, or a sharing model in which common nucleotides encode important amino acids in each protein. To test these models, we analyzed three datasets: 1) a patient dataset of HIV-1 isolates in which we examined frame-specific conservation on the amino acid level; 2) a functional dataset in which we tested the functional contribution of each side chain in Tat and Rev; and 3) a virus selection dataset in which we created non-overlapped Tat and Rev viruses and determined experimental fitness values for every single amino acid allele. Analysis of these datasets led to three key findings, all consistent with a segregated organization of the overlap: 1) overlapped regions in HIV-1 are not more conserved than non-overlapped ones; 2) functional motifs of Tat and Rev are segregated from each other at the level of individual nucleotides and amino acids; and 3) the *tat/rev* overlap reduces the probability of generating unfit combinations of these genes.

The relatively low sequence conservation observed within the overlapped regions of Tat and Rev can be rationalized in light of a segregated organization. Both proteins consist of flexible linkers and short motifs that acquire their proper three-dimensional structures using host protein and viral RNA surfaces as templates (19,20). By organizing their functional motifs such that critical functional residues in one protein overlap with the

highly mutable linker regions of the other protein, HIV-1 can experience a high rate of non-synonymous substitutions that remain functionally neutral.

This segregation also allows the overlap to increase the fitness of the viral population by taking advantage of the drastic reduction in the number of possible allelic combinations of *tat* and *rev*. While initial models of overlaps suggested that such a reduction might reduce overall fitness since fit alleles would be linked with unfit alleles (36), this line of reasoning does not take into account that the equivalent single frame virus can harbor every allelic combination, including all possibilities where just one of Tat or Rev is unfit and thus are replication incompetent. Whether an overlap is advantageous or disadvantageous to the viral population depends on which genotypes are lost when genes are overlapped: if most purged genotypes are fit, the overlapped viruses have a lower probability of producing fit progeny whereas if more purged viruses are unfit, the fitness of the population increases.

Segregation predisposes HIV-1 to exhibit beneficial purging of genotypes in a non-intuitive manner. Even though Tat and Rev are organized such that functionally important residues tend to overlap with flexible linker regions of the opposing protein, there is often selection against disruptive alleles (such as a proline or stop codon). Thus even in these regions, if a functionally important allele in one frame prevents a disruptive and unfit allele in the alternative frame, the overlapped virus can have a selective advantage over the single frame virus. In contrast, a shared (rather than segregated) organization is very likely to reduce the fit sequence space of the overlapped gene because synonymous (or chemically conservative) fit alleles of one

protein would almost always be detrimental in the alternative frame. In a segregated organization, such changes are likely to be neutral in the alternative frame.

Taken together, these data support a simple model for how an overlap might arise and evolve. An ancestral ORF can encode a new ORF in an alternative frame via alternative splicing (37) or atypical translation initiation (38), which can gradually evolve its own function. Functional motifs in the new ORF are more likely to arise in regions that correspond to highly mutable regions of the ancestral ORF to minimize disruptions to the ancestral function. As the new gene function becomes refined, mutations that reduce fitness of the ancestral gene continue to be avoided, and are therefore most likely to encode unfit or neutral alleles in the younger frame.

Such a model is consistent with our data as well as previous studies suggesting that Tat-like proteins are older than Rev-like proteins (39). In the context of the *tat/rev* overlap, the functionally important Y47 residue in Tat could have led to a start codon for Rev in the alternative reading frame. The nascent Rev ORF would then evolve only in the context of fit Tat alleles, avoiding simultaneously unfit mutations of both reading frames in the process. Most functional motifs of Rev overlap with the mutationally tolerant C-terminus of Tat and thus mutations of these positions would not confer any particular advantage or disadvantage to Tat.

Despite the appeal of this speculative model, several caveats should be noted. First, two ancient endogenous lentiviruses have no known *tat/rev* overlap (40,41), although this might result from gene loss events, independent evolutionary origins, or poor splicing annotation. The *tat/rev* overlap also may be more recent than the formation of the individual genes, with *rev* acquiring its first exon and *tat* acquiring its second exon at

later times. Second, our analyses have made rough estimations of viral fitness and examined only single point mutations in one viral background in an artificial genomic locus, ignoring the potential for epistatic effects at multiple positions. However, positive epistasis is rare while negative epistasis is common (42), suggesting that our simplification is conservative and likely underestimates the severity of purged deleterious genotypes (43), while missing only a few restorative double mutations. Regardless, further studies of overlaps in HIV-1 and other systems may shed more light on these aspects. Indeed, recent analyses of polyomaviruses have traced the clade-specific birth of an overlapped gene (44), and detailed functional studies in this system may provide good tests of the model.

It seems clear that overlaps are not necessarily constraints driven by other evolutionary forces, but may in fact reflect a common evolutionary strategy especially prevalent in small viruses. In large viruses, evolution of genomic novelty in response to a selection pressure can occur via gene duplication, adaptation and genome compression (45). An alternative mechanism - possibly preferentially deployed in viruses with small genomes and high recombination rates where gene duplication may not be feasible - is *de novo* gene creation via an overlapped reading frame. Indeed, there is ample evidence that the creation of genes via overlapped reading frames leads to a large degree of genetic novelty (3,7,8,10) in viral proteomes and across the tree of life (4).

The unilateral division of genetic information between Tat and Rev likely reflects a common and clever evolutionary outcome that satisfies a mix of positive and negative pressures and may therefore occur in other systems. Computational studies suggest that other overlapped proteins also segregate (22,46,47), although detailed

experimental validation is needed to determine the extent of segregated versus shared behavior on a per residue basis. At the same time, while functional segregation appears to be a favorable outcome, it is not an inevitable one. The intrinsically disordered nature of Tat and Rev allows considerable flexibility in the placement and ordering of functional motifs. Although overlapped proteins tend to be intrinsically disordered (10,48), other viral proteins, such as the HIV-1 capsid, are extremely genetically fragile (49) and would likely be unable to completely segregate from another gene in an alternative frame. Even the ability of the *tat/rev* overlap to preferentially purge unfit genotypes is only advantageous given a fixed set of selection pressures. Indeed, the constraint of an overlap prevents the exploration of large regions of sequence space and suboptimal alleles, and thus limits the extent of adaptation (50). In the case of Tat and Rev, the need to preserve their essential regulatory functions may simply outweigh any loss of adaptation (46). Such a tradeoff may be detrimental for viral proteins that evolve rapidly to antagonize host defenses. Furthermore, selective advantages can emerge from shared organizations, as has been suggested for the co-evolution of tRNA synthetases encoded on opposite strands (51).

Finally, it is interesting to consider that adding positive selection to an overlapped region via a targeted therapeutic could prove challenging for viral escape. Studies of cytotoxic T-cells in rhesus monkeys have demonstrated CTL targeting of a non-functional portion of SIV *tat* in the *vpr/tat* overlap, and viral escape from these pressures is limited to entirely predictable synonymous mutations (47). Combinations of epitope targeting in overlapped regions could create a viral “Achilles’ heel”, in which combinations of selection pressures from multiple frames drives the virus to unescapable fitness minima.

Given that our data suggests that the virus normally evolves to avoid such situations, the design of these kinds of therapeutic interventions will require systematic approaches such as the work presented here to reveal HIV's true mutational weak spots.

Experimental Methods

Cell lines: The HeLa HIV-1 LTR FFL line (D'Orso and Frankel, 2010) was used for the Tat alanine scan. Cells were cultured under standard conditions (DMEM supplemented w/ 10% FBS, 1% Penn-Strep). Reporter assays were carried out in a 48-well format described below. 293 (ATCC CRL-1573) lines were used for the Rev reporter assay (described below) in standard culture conditions (DMEM supplemented w/ 10% FBS, 1% Penn-Strep) in a 24-well format. 293FT cells were used to generate virus (described below) under standard culture conditions (DMEM supplemented w/ 10% FBS, 1% Penn-Strep). SupT1 (ATCC CRL-1942) human T-cell lines were used for HIV-1 spreading and competition assays in standard culturing (RPMI-1640 supplemented w/ 10% FBS, 1% Penn-Strep).

Patient Analysis: Curated web alignments of protein sequences were downloaded from the Los Alamos HIV-1 Sequence using the filtered web alignments from 2014. Database for each HIV-1 protein. Each region of each protein was then divided into areas of one, two, or three coding overlaps and entropy statistics generated for each region and gene. Only positions present in the HXB2 reference sequence were considered in this analysis.

Tat and Rev reporter assays: All 202 alanine mutants were created in independent site directed mutagenesis reactions in mammalian expression vectors. Reference alanines were not mutated. For the Tat reporter assays, 1ng of plasmid encoding wild

type HXB2 or mutant Tat proteins was co-transfected with 1ng CMV-Renilla control into a previously described HeLa cell line harboring an integrated copy of firefly luciferase under control of the HIV-1 LTR (25). HeLa cells were plated and transfected in a 48 well plate with PolyJet (SignaGen) lipofection reagent for 48 hours. After 48 hours, cells were lysed with 1x Promega passive lysis buffer and firefly and Renilla luciferase values were measured using a luminometer. Tat-dependent firefly luciferase values were normalized to Renilla luciferase values. The assay was performed in biological quintuplet. For the Rev reporter assays, 250 ng of pCMVgagpolRRE was cotransfected with 25 ng of a pcDNA4 NL4-3 Nstrep-Rev mutant into 293 cells in a 24 well format (26) using PolyJet (SignaGen) lipofection reagent. Cells were then lysed and intracellular p24 levels measured by ELISA. The assay was performed in biological triplicate. For all mutants, activity was normalized to the unmutated reference activity.

Creation of Uncoupled Viruses: To create viruses with *tat* and *rev* in the *nef* locus, we introduced mutations into the start codon of the *nef* locus of pNL4-3 as well as sets of mutations into either Tat or Rev that ablated the start codon and introduced two downstream stops. The ablated frame was then reintroduced into *nef*. To create *tat-in-nef* and *rev-in-nef* viruses we introduced Sac II/Xba I sites upstream of the *nef* coding region and ablated the nef start codon. We then mutated either the endogenous start codon for *tat* and *rev* and introduced two downstream stops to prevent reversion. The ablated gene was then reintroduced into the engineered Sac II/Xba I sites. The reintroduced gene was synonymously substituted throughout the gene body to create codon-swapped versions (tat-cs and rev-cs, Table S2) in order to prevent recombination during viral replication.

Creation of Proviral Libraries*:* 202 pools of 33 nt primers were synthesized containing NNN in the middle of the primer (see Table S2 for reference sequences). These primers were used in overlap extension PCRs then cloned into the *nef* locus of the appropriate knockout molecular clone using Gibson Assembly Master Mix (NEB) after digesting the backbone with Sac II and Xba I. A minimum of 1000 colonies was used to generate a single proviral library representing randomization of a single residue. The resulting colonies were scraped together and spin-column (EconoSpin) plasmid DNA purification was performed. Each randomized proviral library (202 total) was deep sequenced to ensure high codon diversity.

Virus Generation and Spread*:* Each proviral plasmid library was raised separately and complemented with reference *rev* and *tat in trans* to generate virus in 293T cells. Briefly, 60,000 293FT cells were transiently transfected with 200 ng of proviral plasmid, and 2 ng each of pcDNA mammalian expression vectors containing *tat* or *rev* using PolyJet (SignaGen) lipofection reagent at a 1:5 ratio of the DNA:Polyjet. Virus-laden supernatant was collected 48 hours following transfection, cleared of cells via low speed centrifugation at 500 x g for 5 minutes, and stored at -80°C. Virus titers were determined by p24 ELISA.

Viral Competition and Selection*:* Selection assays were initiated via centrifugal inoculation of 10^6 SupT1 T-cells with 5ng p24 and 8 ug/mL polybrene for 2 hours at 1200 x g at 32°C in 96-well microtiter plates. Following inoculation, cells were washed twice with PBS to remove input virus and 250 uL media (RPMI-HEPES + 10% fetal calf serum) replaced per well. The infections were then monitored by immunofluorescence assay for cellular HIV antigen synthesis and by supernatant p24 ELISA to quantify

progeny virion release. At peak infection for reference or defective viral pools 150 uL of virus-laden supernatant was collected, cleared of cells, and the virion-associated RNA content extracted via ZR-96 Viral RNA extraction kits (Zymo) and eluted in 15 ul water. Once all RNA samples were collected, cDNA was generated from 5uL of extracted RNA using the iScript Advanced (BioRad) cDNA synthesis. Selection experiments for each residue (202 total + reference wells) were performed in biological duplicate.

Amplicon Generation*:* Amplicon specific primers were used to extract 175-nt fragments containing the randomized codon and add Illumina adapter sequence (3 total primer sets for Tat and 5 for Rev; Table S2). Each amplicon specific primer consisted of a pool of 1-4 random nucleotides appended immediately 5' of the gene-specific region in order to increase basecall diversity during the NGS runs. A second step PCR added barcode sequences (courtesy of J. DeRisi) and the remainder of the adapter. DNA from each selection experiment was quantified using qPCR (Kapa Biosciences) and then pooled in equal amounts and gel extracted to create the NGS library. Library quantification was then performed using qPCR and sequenced using PE150s on a MiSeq (Illumina). Samples that produced low read counts, were either re-pooled in a smaller mixture or repeated starting at the initial infection stage.

Data Processing*:* Samples were de-multiplexed using MiSeqReporter to produce fastq files, each representing a single competition experiment. The resulting paired-end fastq files were then aligned to one another and the reference codon-swapped sequence (Table S2) and the frequencies of every codon were calculated. The randomized position was then extracted and a post-and pre-selection allele frequencies calculated. Synonymous codons were pooled as correlations between amino acids across

biological replicates was much stronger than individual correlations of codons (data not shown).

Approximation of Overlap Effect on Viral Fitness*:* In order to calculate the manner in which one gene restricts the other we took the endogenous NL4-3 sequence and fixed one allele in one gene. For instance, to examine the effect of Tat Y47 on Rev Position 1, we substituted both Y codons (TAT, TAC) into the rev sequence (ATg (M) or ACg (T)) and estimated the resulting fitnesses of the viral genotypes (i.e. Tat Y47, Rev M1) as the smaller of the two alleles (i.e. either Tat Y47 or Rev M1). The comparable single frame virus has 21 viral genotypes with Tat Y47 (i.e. Rev A1…Rev Y1) and the fitness of these genotypes can be approximated in a similar manner (Figure S7). The resulting single frame landscape of these 21 viruses can be compared to the restricted landscape of the two overlapped genotypes. This procedure was applied for every possible amino acid at every position in the overlap, first holding Tat fixed and calculating the resulting Rev fitness (and subsequently vice versa). As holding most positions fixed influences a dipeptide in the alternative frame, we added the selection coefficients of the resulting alleles in the alternative frame. The resulting single frame and overlapped distributions were then plotted in R using ggplot2's density function.

Experimental Replication*:* Alanine scanning was performed in biological quintuplet (Tat) or triplicate (Rev). For the deep mutational scanning dataset, each competition experiment (86 for Tat, 116 for Rev) was performed in biological duplicate.

QUANTIFICATION AND STATISTICAL ANALYSIS

Entropy Comparisons between Overlapped and Non-Overlapped Regions in Patient Data*:* To test for statistically significant differences between regions of overlap

(Figure S2B), we used two approaches: 1) We used a Wilcoxon test to compare the empirical distributions of entropy values within a bin to the distribution of entropy values in another bin, and 2) we used a permutation-based approach where we compared the mean entropy value from the positions that correspond to a given bin class to randomly sampled similarly sized, contiguous segments of the HIV exome. We then determined if the observed mean entropy value was significantly outside of the distribution of entropy values from these randomly sampled segments. The latter method is best illustrated by example: Consider two regions in the exome that code for 3 genes, where each of these hypothetical regions are 5 amino acids long. First, we pool the entropy values that correspond to these regions, and calculate their mean entropy level. We then randomly sample (with replacement) two contiguous segments of the HIV exome that are each 5 amino acids long, and calculate the mean entropy level for these sampled segments. We then repeat this process 10,000 times to arrive at an empirical null distribution of mean entropy values. We then compare our observed mean entropy value for regions of 3 coding reading frames to this empirical null distribution to see if it is significantly outside of what one would expect from chance.

Alanine Scanning*:* Tat reporter assay values are the mean of five independent transfections with error bars representing the standard deviation of these replicates. Rev reporter assays represent the mean of three independent transfections (error bars are the standard deviation).

Estimations of Experimental Error*:* Because wild-type (non-randomized) positions were also sequenced in our data, we were able to use this information to estimate the overall error rate for each amino acid type. For wild-type positions, the expectation is

that all sequenced reads will have the reference allele, and any reads that deviate from this expectation are the result of some type of error during the course of our protocol. These errors include random mutations during the experiment, PCR errors, or sequencing errors. We estimate the combined 'experimental error' for each amino acid type by using the non-randomized positions to count the number of instances that we observed a mutation 'away' from a given amino acid, and the number of times we observe a mutation 'to' that amino acid. We then normalized these counts by the total number of reads observed for the amino acid to determine mutation rates. Thus, each amino acid has an estimated 'away' mutation rate and an estimated 'to' mutation rate associated with it. We represent the rate at which amino acid Y mutates to any other amino acid as $R_{Y,\cdot}$, and the rate at which any other amino acid mutates to Y as $R_{\cdot,Y}$.

Estimating Population Growth*:* We used data on the growth of a wild-type population of HIV under the same experimental protocol as in our evolution experiments, to estimate the overall population growth in each experiment. Specifically, we measured the p24 concentration (in pg/uL) for several time points during the course of the experiment in triplicate. We set the overall population size of HIV to be equal to the mean [p24] value at each time-point. This is a good estimate of the HIV population size, as we used this value to empirically determine our MOI and thus this is a reasonable estimate of the infectious population. We then fit a logistic curve to these observed values to get a population growth function over the time-course of the experiments. The result of this fit produced the following equation:

$$N_t = f(t) = \frac{6604652.673}{1+e^{10.7974-2.646389t}},$$

where N_t is the population size at generation t.

Neutral Simulations*:* We sought to simulate, *in silico*, our evolution experiments in order estimate the range that a neutral allele's frequency could feasibly change during the experiments. This simulated range served as our null distribution, and we generated a unique null distribution for each allele in our observed data. The neutral simulations had five parameters: the overall population growth function, the number of generations, the starting allele frequency, the ending read depth for the experiment, and the amino acid identity of the allele. Each of these will be explained below.

A set of 100,000 neutral simulations were run for each allele in the data, and each simulation was run as follows. The starting neutral allele frequency was set to begin at the same starting frequency of the observed allele. We then ran a Wright-Fisher simulation of neutral drift over 6 generations (52). We model the neutral allele in question as allele A, with frequency p and collapsed all the other alleles in the population to be allele a, with frequency $q = 1 - p$. If the population size and frequency for allele A at generation t is N_t and p_t, respectively, then the count of A is $P_t = N_t p_t$, and the count of a is $Q_t = N_t - P_t$. The probability of an allele count in the next generation (P_{t+1}) follows the binomial distribution,

$$\Pr(P_{t+1}) = \binom{f(t+1)}{P_{t+1}} p_t^{P_{t+1}} q_t^{Q_{t+1}} .$$

Thus, to get a random value for P_{t+1} in the simulation we simply randomly sample from the above distribution. Once the Wright-Fisher simulation is complete, we again use the binomial distribution to simulate sub-sampling of the population. We do this because the true population size of HIV at the end of the experiment is quite large, and when we

sequence this population, we are only observing a subset. The size of this subsample is equal to the number of reads that were sequenced at the end of a given evolution experiment (one per randomized position). If $N_6$is the final population size, let the subsample size be N_6'. Similarly, if P_6 is the count of A at the final generation, then P_6' is the count of A resulting from subsampling. We again use the binomial to model the probability of P_6',

$$\Pr(P_6') = \binom{N_6'}{P_6'} p_6^{P_6'} q_6^{N_6'-P_6'} ,$$

and randomly sample from this distribution to simulate a value for P_6'. In order to simulate random additions to and subtractions from allele A that occur due mutations from the combined experimental error, we again use a binomial sampling approach. If the amino acid identity of A is Y then the rate at which Y is mutated to anything else is $R_{Y,\cdot}$, and the rate at which anything else mutates to Y is $R_{\cdot,Y}$. Let the number of units that is added to P_6' due to experimental error be represented as $X_{\cdot,Y}$, and the number of units subtracted be $X_{Y,\cdot}$. The probability of these two values are again found using the binomial,

$$\Pr(X_{\cdot,Y}) = \binom{N_6'-P_6'}{X_{\cdot,Y}} R_{\cdot,Y}^{X_{\cdot,Y}} (1-R_{\cdot,Y})^{N_6'-P_6'-X_{\cdot,Y}}; \text{ and } \Pr(X_{Y,\cdot}) = \binom{P_6'}{X_{Y,\cdot}} R_{Y,\cdot}^{X_{Y,\cdot}} (1-R_{Y,\cdot})^{P_6'-X_{Y,\cdot}} .$$

We randomly sample from these two distributions to get discrete values for $X_{\cdot,Y}$ and $X_{Y,\cdot}$. Finally, Let the count of allele A, at generation 6, after subsampling, and after introducing random experimental error be P_6''. P_6'' is then given by,

$$P_6'' = P_6' - X_{Y,\cdot} + X_{\cdot,Y} .$$

P_6'' is the final allele count that we use in the simulation. This process is then repeated 100,000 times to arrive at a simulated distribution for P_6''. This distribution represents the

range of ending allele frequencies that one could reasonably expect for a neutral allele, if this allele had the same starting frequency, read depth, amino acid identity as a given observed allele in our data.

Simulations with Selection*:* We ran the same simulations of our evolution experiments as described above, but included a selection parameter in order to test the accuracy of our fitness estimates of observed alleles. The only component of the framework described above that changed in this case was the Wright-Fisher simulations. Here, the expected frequency of allele A with a fitness of s at generation $t+1$ is given by,

$$E[p_{t+1}] = \frac{p_t(1+s)}{q_t + p_t(1+s)} ,$$

and the probability of P_{t+1} with selection is given by,

$$\Pr(P_{t+1}) = \binom{f(t+1)}{P_{t+1}} E[p_{t+1}]^{P_{t+1}} (1 - E[p_{t+1}])^{Q_{t+1}} .$$

Point Estimate of Selection Coefficient*:* We used population genetics theory to estimate the fitness, or selection coefficient (s) of an allele when given the starting frequency, ending frequency, and number of generations in-between (53). The equation for s is given by,

$$s = e^{\frac{\ln(p_{i+t}/q_{i+t}) - \ln(p_i/qi)}{t}} .$$

Where i indexes the starting generation and t is the number of generations that have elapsed.

Our simulation framework with selection (described above) allowed us to test the accuracy of our fitness estimates. Because selection can be incorporated into our

simulations, we can simulate an allele with a pre-specified ‘true’ fitness value. We can then compare this value with our estimate of fitness that is based only upon the results of our stochastic simulation (agnostic to the ‘true’ fitness value). To ensure that the parameters for these simulations with selection were relevant to our study, we gathered the parameters that pertain to each of the observed alleles in our data. These parameters are: starting allele frequency (to give the starting point of the simulation), the read depth at the end of the evolution experiment (to give the subsample size), and the amino acid identity of the allele (to give the experimental error rate). We then ran a single simulation pertaining to each of these alleles’ parameter sets, and importantly, included a fitness value with each of these simulations. To select a fitness value we randomly sampled from a uniform distribution between -2 and 2. This gives the true fitness. We then find the estimated fitness by incorporating the results of the stochastic simulation into the equation for the selection coefficient (given above). We found that the estimated fitness values track very well with the true fitness (Figure S5C), and thus our point estimate of fitness is accurate. However, we found that if the starting allele frequency is low (which by design is almost always the case for our data), one does not have the ability to infer extreme negative fitness values. This is because a moderately negatively selected allele will fall to a frequency of 0 just the same as an extremely negatively selected allele, if the starting frequency is low. This conceptual limit to one’s ability to infer negative selection with low starting frequency is well illustrated by the plateauing of points on the left-hand side of Figure S5C.

Calculations of Overlap Restriction*:* In order to calculate the manner in which one gene restricts the other we took the endogenous NL4-3 sequence and fixed one allele in one gene. For instance, to examine the effect of Tat Y47 on Rev Position 1, we substituted both Y codons (TAT, TAC) into the rev sequence (ATg (M) or ACg (T)) and estimated the resulting fitnesses of the viral genotypes (i.e. Tat Y47, Rev M1) as the smaller of the two alleles (i.e. either Tat Y47 or Rev M1). The comparable single frame virus has 21 viral genotypes with Tat Y47 (i.e. Rev A1…Rev Y1) and the fitness of these genotypes can be approximated in a similar manner. The resulting single frame landscape of these 21 viruses can be compared to the restricted landscape of the two overlapped genotypes. This procedure was applied for every possible amino acid at every position in the overlap, first holding Tat fixed and calculating the resulting Rev fitness (and subsequently vice versa). As holding most positions fixed influences a dipeptide in the alternative frame, we added the selection coefficients of the resulting alleles in the alternative frame. The resulting single frame and overlapped distributions were then plotted in R using ggplot2's density function.

DATA AND SOFTWARE AVAILABILITY

Deep mutational scanning data has been deposited in ArrayExpress, accession E-MTAB-5154.

Population Genetics Modeling and other code resources are available at https://github.com/nbstrauli/allele_frequency_trajectory_sim

AUTHOR CONTRIBUTIONS

Jason D Fernandes and Alan D Frankel designed the experiments with input from Tyler B Faust, David C Crosby and Robert L Nakamura. JDF and ADF wrote the manuscript

with input from all authors. JDF and Nicholas B Strauli analyzed the patient data. JDF and TBF performed the alanine scans. JDF, Cynthia Smith and DCC performed viral selection and sequencing experiments. NBS and Ryan D Hernandez developed the population genetics model and aided JDF in analysis of the selection dataset.

ACKNOWLEDGMENTS

We thank J. DeRisi and M. Stenglein for barcode sequences and advice for NGS experimental design; R. Andino and A. Acevedo for use of equipment and advice for performing selection experiments. We also thank P. Babbitt, J. Gross, J. Fraser, R. Andino, M. Daugherty, L. Gitlin and members of the Frankel and Andino labs for helpful conversations and review of the manuscript. JDF was supported in part by NIH training grant T32GM007175 and an Amgen Research Excellence Fellowship. This work was supported by NIH grant P50GM088250 to ADF.

References

1. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, et al. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 1977;265(5596):687–95.
2. Barrell BG, Air GM, Hutchison CA. Overlapping genes in bacteriophage φX174. Nature. 1976 Nov;264(5581):34–41.
3. Belshaw R, Pybus OG, Rambaut A. The evolution of genome compression and genomic novelty in RNA viruses. Genome Res. Cold Spring Harbor Laboratory Press; 2007;17(10):1496–504.
4. Makalowska I, Lin C-F, Makalowski W. Overlapping genes in vertebrate genomes. Comput Biol Chem. 2005 Feb;29(1):1–12.
5. Rogozin IB, Spiridonov AN, Sorokin A V, Wolf YI, Jordan IK, Tatusov RL, et al. Purifying and directional selection in overlapping prokaryotic genes. Trends Genet. Elsevier; 2002;18(5):228–32.
6. Chirico N, Vianelli A, Belshaw R. Why genes overlap in viruses. Proc R Soc B Biol Sci. The Royal Society; 2010;277(1701):3809–17.
7. Brandes N, Linial M. Gene overlapping and size constraints in the viral world. Biol Direct. BioMed Central; 2016 Dec;11(1):26.
8. Keese PK, Gibbs a. Origins of genes: “big bang” or continuous creation? Proc Natl Acad Sci U S A. 1992 Oct;89(20):9489–93.
9. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by overprinting. Mol Biol Evol. 2012;
10. Rancurel C, Khosravi M, Dunker AK, Romero PR, Karlin D. Overlapping Genes

Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation. J Virol. American Society for Microbiology (ASM); 2009;83(20):10719–36.

11. Sabath N, Landan G, Graur D. A Method for the Simultaneous Estimation of Selection Intensities in Overlapping Genes. Pybus OG, editor. PLoS One. Public Library of Science; 2008;3(12):7.
12. Hein J, Støvlbaek J. A maximum-likelihood approach to analyzing nonoverlapping and overlapping reading frames. J Mol Evol. 1995;40(2):181–9.
13. Wei X, Zhang J. A simple method for estimating the strength of natural selection on overlapping genes. Genome Biol Evol. 2015 Jan;7(1):381–90.
14. Kawano Y, Neeley S, Adachi K, Nakai H. An experimental and computational evolution-based method to study a mode of co-evolution of overlapping open reading frames in the AAV2 viral genome. PLoS One. 2013 Jan;8(6):e66211.
15. Foley B, Leitner T, Apetrei C, Hahn B, Mizrachi I, Mullins J, et al. HIV Sequence Compendium 2013. Los Alamos National Laboratory, NM: Theoretical Biology and Biophysics Group; 2013.
16. Tahirov TH, Babayeva ND, Varzavand K, Cooper JJ, Sedore SC, Price DH. Crystal structure of HIV-1 Tat complexed with human P-TEFb.pdf. Nature [Internet]. Nature Publishing Group; 2010;465(7299):747–51. Available from: http://dx.doi.org/10.1038/nature09131
17. DiMattia MA, Watts NR, Stahl SJ, Rader C, Wingfield PT, Stuart DI, et al. Implications of the HIV-1 Rev dimer structure at 3.2 Å resolution for multimeric binding to the Rev response element. Proc Natl Acad Sci U S A. National

Academy of Sciences; 2010;107(13):5810–4.

18. Daugherty MD, Liu B, Frankel AD. Structural basis for cooperative RNA binding and export complex assembly by HIV Rev. Nat Struct Mol Biol. Nature Publishing Group; 2010;17(11):1337–42.

19. Ott M, Geyer M, Zhou Q. The Control of HIV Transcription: Keeping RNA Polymerase II on Track. Cell host microbe. Elsevier Inc.; 2011;10(5):426–35.

20. Pollard VW, Malim MH. The HIV-1 Rev protein. Annu Rev Microbiol. 1998 Jan;52:491–532.

21. Pan K, Deem MW. Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. J R Soc Interface R Soc. 2011;8(64):1644–53.

22. Zaaijer HL, Van Hemert FJ, Koppelman MH, Lukashov V V. Independent evolution of overlapping polymerase and surface protein genes of hepatitis B virus. J Gen Virol. 2007;88(Pt 8):2137–43.

23. Ngandu NK, Scheffler K, Moore P, Woodman Z, Martin D, Seoighe C. Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. Virol J. BioMed Central; 2008;5:160.

24. Snoeck J, Fellay J, Bartha I, Douek DC, Telenti A. Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints. Retrovirology. BioMed Central Ltd; 2011;8(1):87.

25. D'Orso I, Frankel AD. RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. Nat Struct Mol Biol. 2010;17(7):815–21.

26. Smith AJ, Cho MI, Hammarskjöld ML, Rekosh D. Human immunodeficiency virus type 1 Pr55gag and Pr160gag-pol expressed from a simian virus 40 late replacement vector are efficiently processed and assembled into viruslike particles. J Virol. 1990;64(6):2743–50.
27. Wright S. Evolution in Mendelian Populations. Genetics. 1931 Mar;16(2):97–159.
28. Feder AF, Kryazhimskiy S, Plotkin JB. Identifying signatures of selection in genetic time series. Genetics. 2014 Feb;196(2):509–22.
29. McLaughlin RN, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. Nature Publishing Group; 2012 Nov;491(7422):138–42.
30. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution map of a gene's fitness landscape. Mol Biol Evol. 2014 Jun;31(6):1581–92.
31. Doud M, Bloom J. Accurate Measurement of the Effects of All Amino-Acid Mutations on Influenza Hemagglutinin. Viruses. Multidisciplinary Digital Publishing Institute; 2016 Jun;8(6):155.
32. Thyagarajan B, Bloom JD. The inherent mutational tolerance and antigenic evolvability of influenza hemagglutinin. Elife. eLife Sciences Publications Limited; 2014 Jan;3:e03300.
33. Kumar M, Keller B, Makalou N, Sutton RE. Systematic determination of the packaging limit of lentiviral vectors. Hum Gene Ther. Mary Ann Liebert, Inc.; 2001;12(15):1893–905.
34. Chan LY, Kosuri S, Endy D. Refactoring bacteriophage T7. Mol Syst Biol. EMBO

and Nature Publishing Group; 2005;1(1):2005.0018.

35. Neuveut C, Jeang KT. Recombinant human immunodeficiency virus type 1 genomes with tat unconstrained by overlapping reading frames reveal residues in Tat important for replication in tissue culture. J Virol. 1996;70(8):5572–81.

36. Miyata T, Yasunaga T. Evolution of overlapping genes. Nature. 1978 Apr;272(5653):532–5.

37. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science. 2014 Feb;343(6172):769–72.

38. Touriol C, Bornes S, Bonnal S, Audigier S, Prats H, Prats A-C, et al. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. Biol Cell. 2003 May;95(3–4):169–78.

39. Pavesi A, Magiorkinis G, Karlin DG. Viral proteins originated de novo by overprinting can be identified by codon usage: application to the “gene nursery” of Deltaretroviruses. PLoS Comput Biol. 2013 Jan;9(8):e1003162.

40. Keckesova Z, Ylinen LMJ, Towers GJ, Gifford RJ, Katzourakis A. Identification of a RELIK orthologue in the European hare (Lepus europaeus) reveals a minimum age of 12 million years for the lagomorph lentiviruses. Virology. 2009 Feb;384(1):7–11.

41. Katzourakis A, Tristem M, Pybus OG, Gifford RJ. Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A. National Academy of Sciences; 2007;104(15):6261–5.

42. Bank C, Hietpas RT, Jensen JD, Bolon DN. A systematic survey of an intragenic epistatic landscape. Mol Biol Evol. 2015 Jan;32(1):229–38.

43. Boucher JI, Bolon DNA, Tawfik DS. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature. Protein Sci. 2016 Jul;25(7):1219–26.

44. Carter JJ, Daugherty MD, Qi X, Bheda-Malge A, Wipf GC, Robinson K, et al. Identification of an overprinting gene in Merkel cell polyomavirus provides evolutionary insight into the birth of viral genes. Proc Natl Acad Sci U S A. 2013 Jul;110(31):12744–9.

45. Elde NC, Child SJ, Eickbush MT, Kitzman JO, Rogers KS, Shendure J, et al. Poxviruses deploy genomic accordions to adapt rapidly against host antiviral defenses. Cell. Elsevier; 2012 Aug;150(4):831–41.

46. Maman Y, Blancher A, Benichou J, Yablonka A, Efroni S, Louzoun Y. Immune-induced evolutionary selection focused on a single reading frame in overlapping hepatitis B virus proteins. J Virol. 2011;85(9):4558–66.

47. Hughes AL, Westover K, Da Silva J, O'Connor DH, Watkins DI. Simultaneous Positive and Purifying Selection on Overlapping Reading Frames of the tat and vpr Genes of Simian Immunodeficiency Virus. J Virol. American Society for Microbiology; 2001;75(17):7966–72.

48. Kovacs E, Tompa P, Liliom K, Kalmar L. Dual coding in alternative reading frames correlates with intrinsic protein disorder. Proc Natl Acad Sci U S A. National Academy of Sciences; 2010;107(12):5429–34.

49. Rihn SJ, Wilson SJ, Loman NJ, Alim M, Bakker SE, Bhella D, et al. Extreme genetic fragility of the HIV-1 capsid. PLoS Pathog. 2013 Jun;9(6):e1003461.

50. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying

Selection in TEM-1 β-Lactamase. Cell. 2015 Feb;160(5):882–92.

51. Rodin SN, Ohno S. Two types of aminoacyl-tRNA synthetases could be originally encoded by complementary strands of the same nucleic acid. Orig Life Evol Biosph. 1995 Dec;25(6):565–89.

52. Principles of Population Genetics. 4th ed. Sinauer Assosiates, Inc.; 2007. 565 p.

53. Felsenstein J. Theoretical evolutionary genetics. 2015th ed. 5th international workshop, WABI 2005, Mallorca, Spain, October 3-6, 2005 : proceedings. 2005. 436 p.

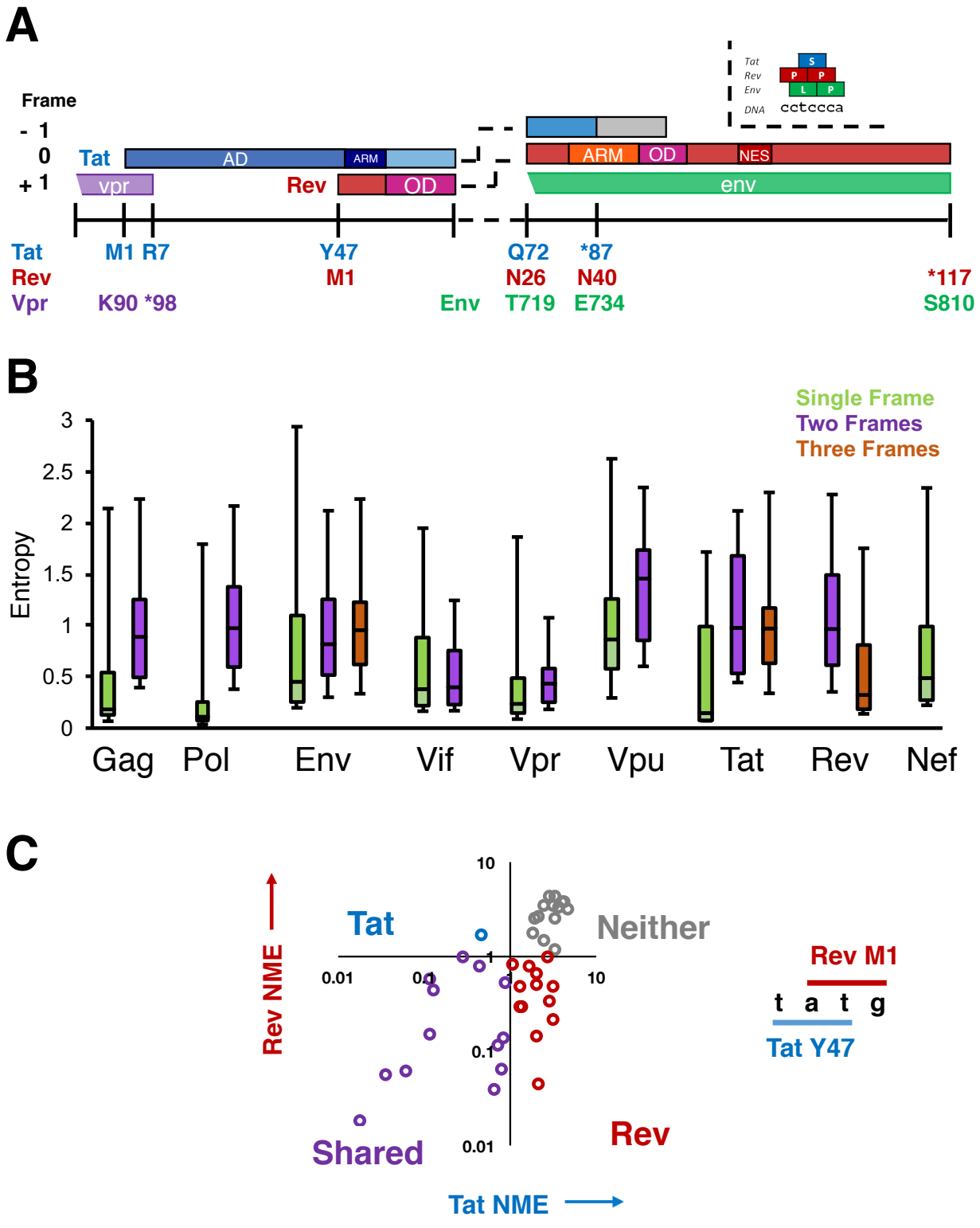


Figure 1: Organization and Conservation of HIV-1 Overlaps

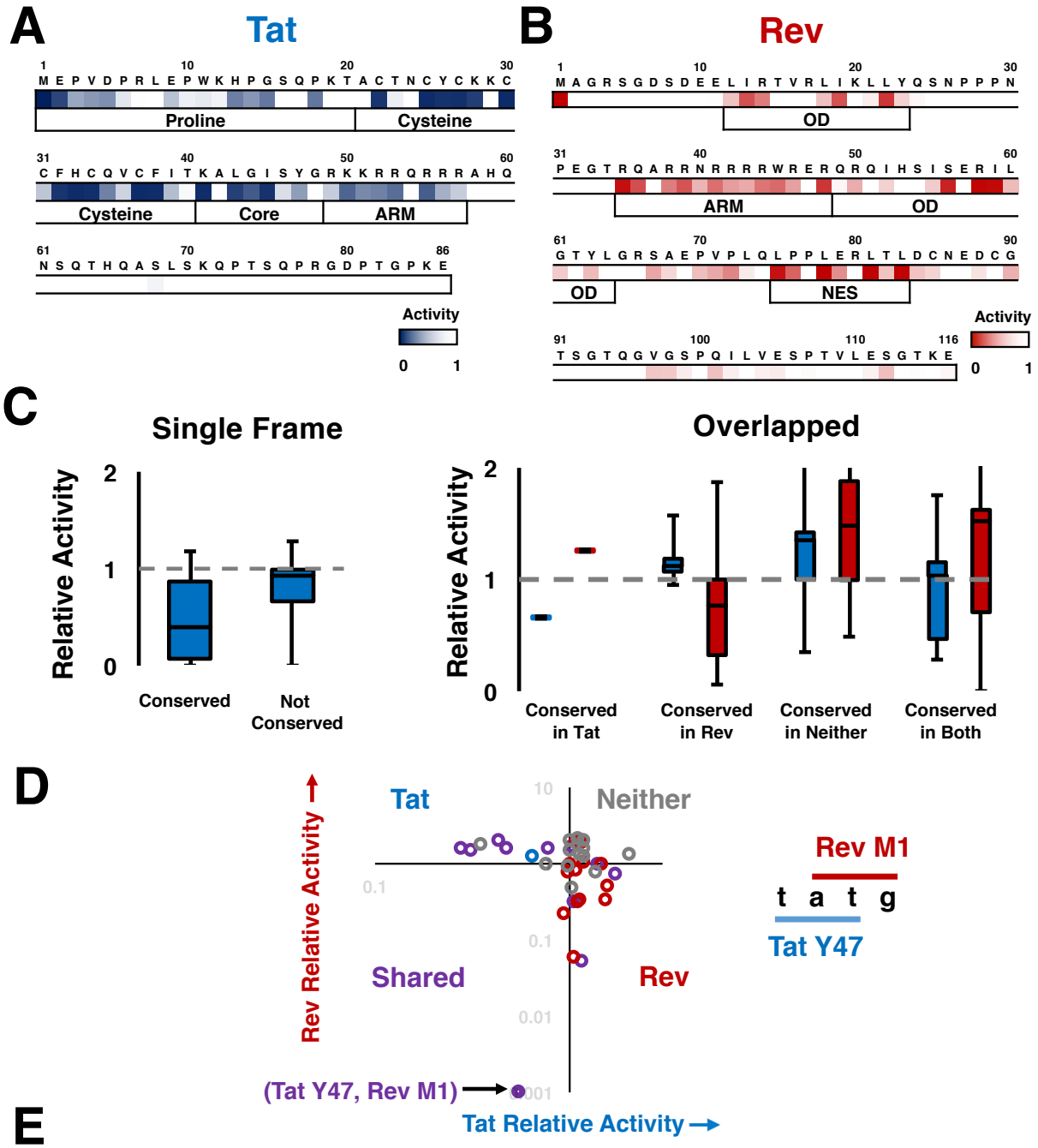


Figure 2: Functional Dissection of Tat and Rev

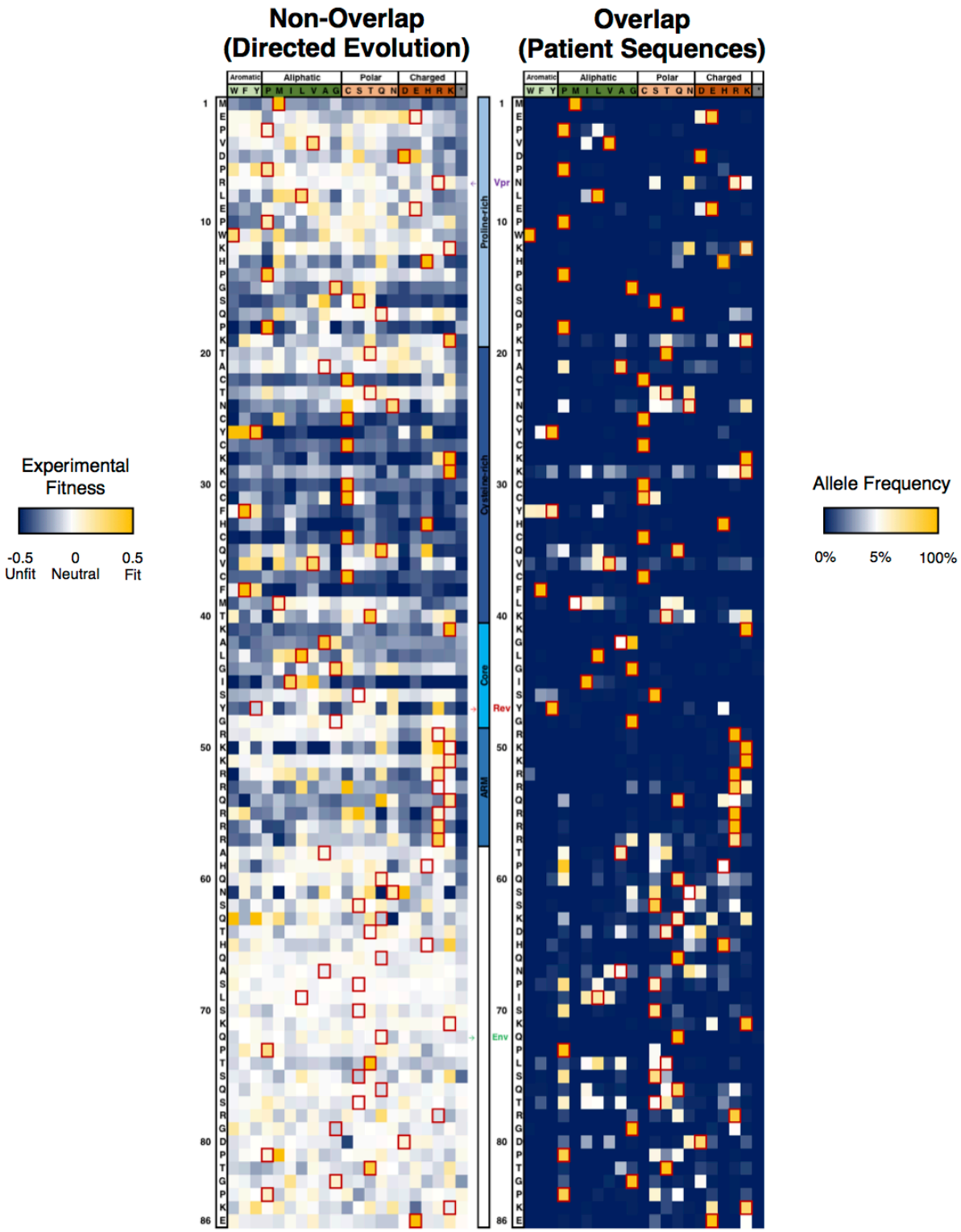


Figure 3: Mutational Profiling of HIV-1 Tat

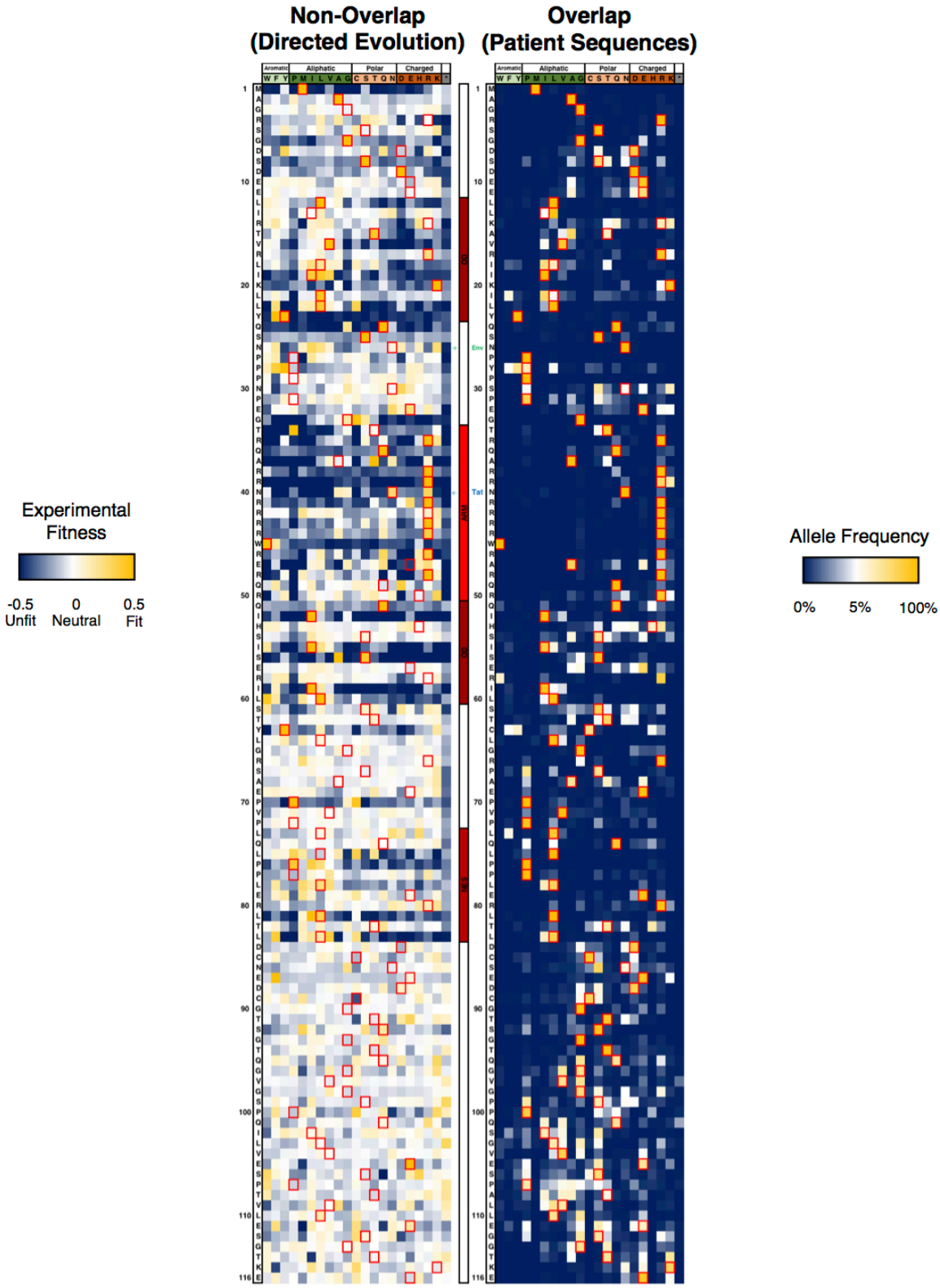


Figure 4: Mutational Profiling of HIV-1 Rev

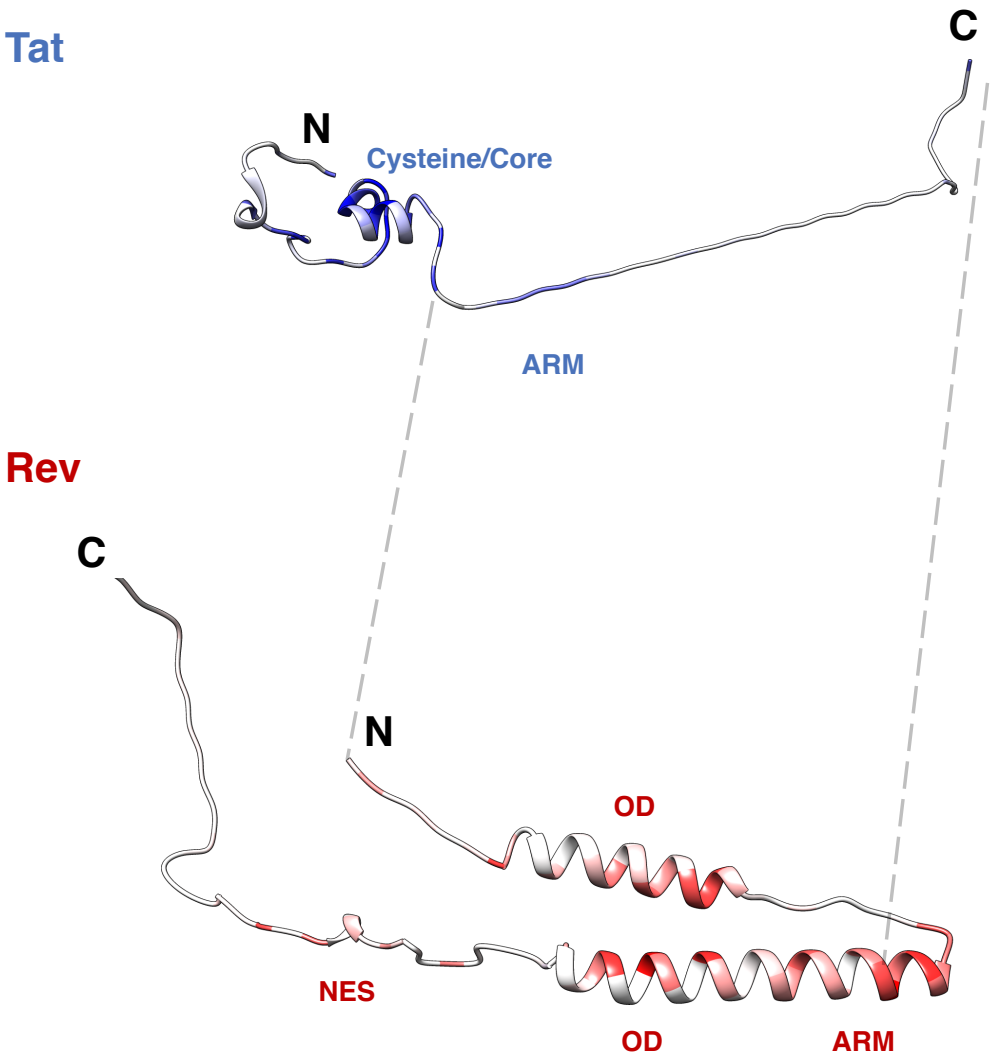


Figure 5: Structural Profiles of Tat and Rev

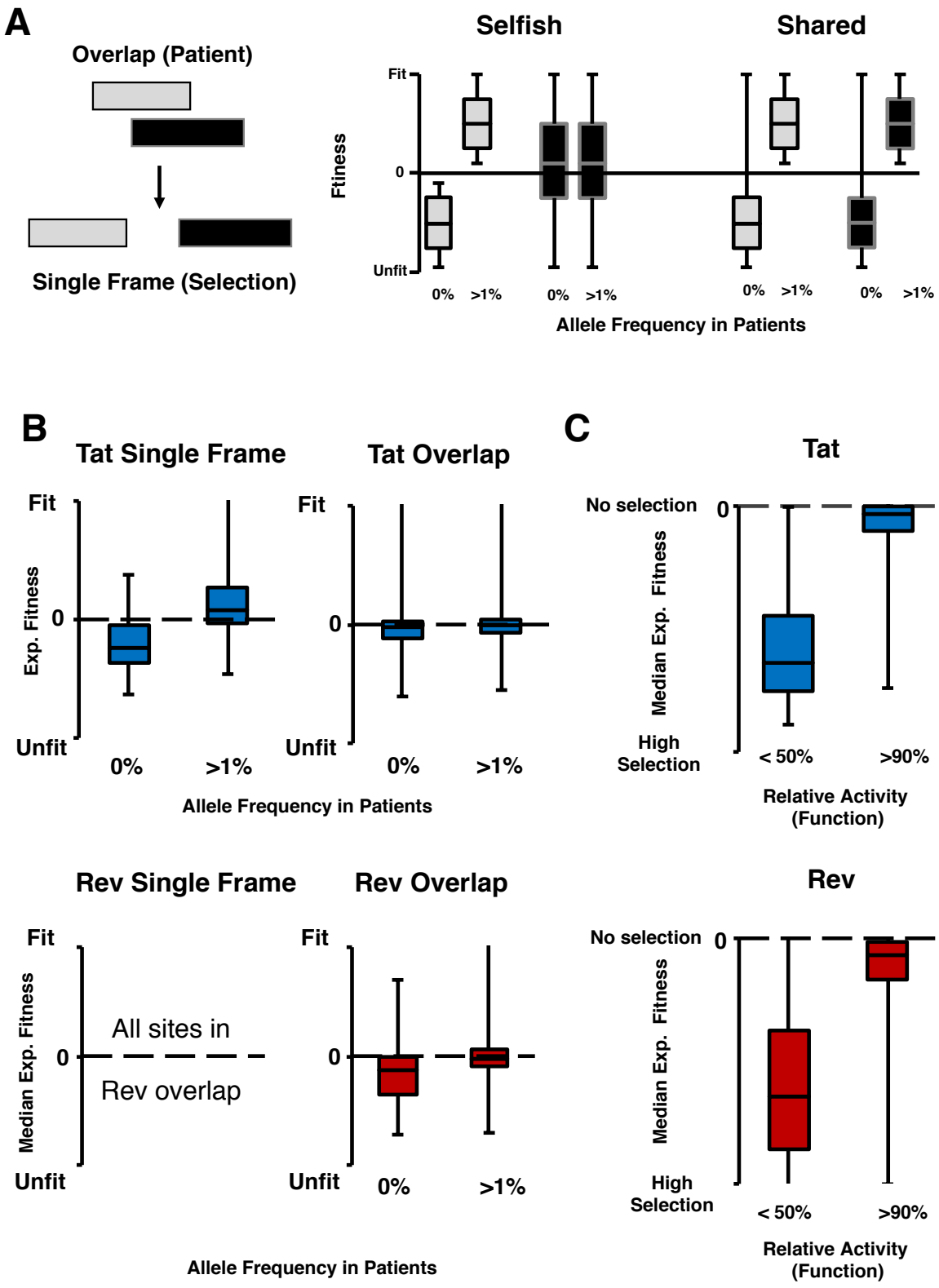


Figure 6: Comparisons between Datasets

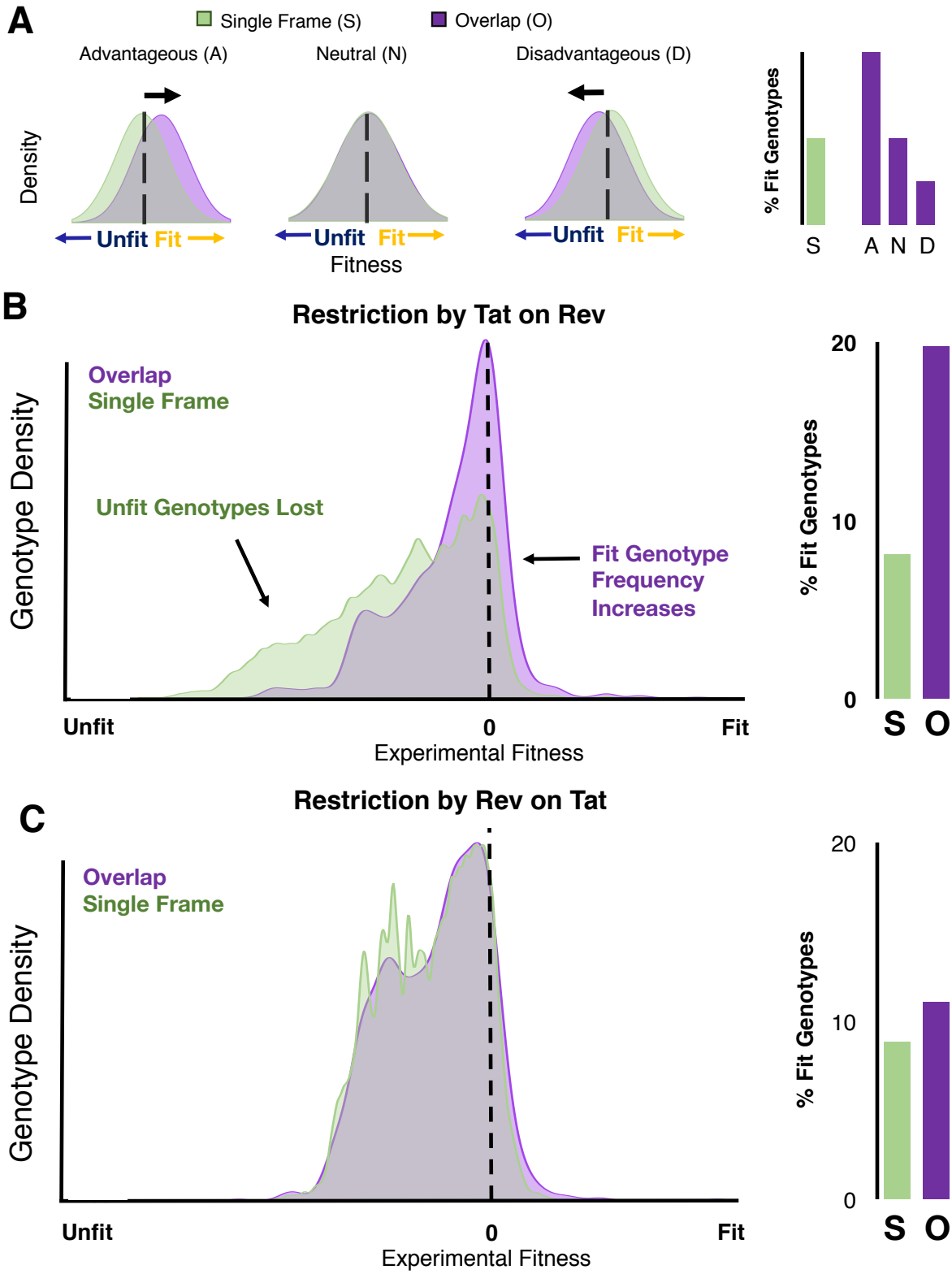


Figure 7: Advantageous Restriction by the Overlap

Figure Legends

Figure 1: Organization and Conservation of HIV-1 Overlaps

A) Layout of the Genetic Organization of the *tat/rev* overlap in HIV-1. ARM = arginine rich motif/nuclear localization sequence, OD = oligomerization domain NES = nuclear export sequence. In HIV-1 NL4-3 Tat is 86 residues, although many patient genes are 101 residues (grey box).

B) Individual gene entropy analysis for overlapped and single-frame regions in the HIV-1 genome (see Fig S2A). Entropy values were computed at the protein level for each frame and Shannon Entropy values for alignments of HIV-1 patient sequences are shown. Median, range, and IQR are shown in the box and whiskers plot. A score of 0 indicates absolute conservation and a score of 3 indicates near-absolute degeneracy

C) Categorization of sites by Normalized Mean Entropy (NME) in the *tat/rev* overlap. Residues are grouped into pairs that share two nucleotides, and their NME plotted accordingly (Tat RE, Rev RE). Quadrants are labeled to indicate which genes are conserved in that region.

Figure 2: Functional Dissection of Tat and Rev

A-B) Alanine scanning of Tat (A) and Rev (B) in functional reporter assays.

C) Relationship between entropy and function. (Left) Highly conserved (NME < 1) residues in the Tat single frame region generally result in a large loss of Tat activity when mutated to alanine, while residues that are not conserved (NME > 1) do not. A similar pattern is seen in the overlap (right) when only a single frame is conserved (blue = Tat activities, red = Rev activities). When neither gene is conserved there is generally

no loss of function. When both frames are conserved the relationship between entropy and function is less clear.

D) Categorization of residues in the *tat/rev* overlap by function. Residues are grouped into sites that share two nucleotides and colored as in Figure 1D. The only site that is required for the activity of both proteins is (Tat Y47, Rev M1) (labeled). Coloring of points represents classifications made by NME as in Fig 1D.

E) Residues determined to be functionally important by alanine scan in the context of the overlap. Dark blue residues are important for Tat function (<50% activity) while dark red residues are important for Rev (<50%activity). Only Tat Y47 and Rev M1 share nucleotides.

Figure 3: Mutational Profiling of HIV-1 Tat

(Left) Experimental fitness of every residue (all synonymous codons grouped) of Tat in non-overlapped tat-in-nef viruses after ~6 generations of selection in SupT1 cells. Red boxes and row headers denote the NL4-3 sequence. Dark blue indicates negative selection, white indicates neutral, and gold indicates positive selection. (Center) Motif and overlap organization of Tat. Labels of other indicate stop (vpr) and start codons (rev, gp41). (Right) Overlapped/Patient conservation of Tat residues. The neutral expectation (white) is approximated as a residue being equally represented by all amino acids (1/20). Red boxes denote the NL4-3 sequence while row headers denote the consensus sequence.

Figure 4: Mutational Profiling of HIV-1 Rev

(Left) Experimental fitness of every residue (all synonymous codons grouped) of Rev in non-overlapped rev-in-nef viruses and (right) overlapped, patient conservation of Rev

residues. Coloring and labeling are consistent with those shown in Figure 3. Tat stop codon and start of Env coding overlap are labeled.

Figure 5: Structural Profiles of Tat and Rev

Crystal Structures of Tat (3MI9) and Rev (3LPH, 3NBZ) colored by median experimental fitness values. Dark blue/red coloring indicates strong selection for a particular amino acid at that site (white indicates experimental fitness of 0). Unstructured regions in Tat (residues 50-86) and Rev (residues 1-9, 65-73, 86-116) were built in Pymol and added to the structures for visualization purposes. The grey dashed lines flank the overlap in both proteins.

Figure 6: Comparisons between Datasets

A) Model for how overlapped proteins might behave in a single frame context. For a dominant gene (grey) in a segregated organization, alleles that are absent (0% frequency) remain unfit while those that are present (>1%) remain fit. For an accommodating gene (black) many absent alleles are fit. In a shared organization, removal of the overlap does not change the fact that both proteins experience strong selection at that site. B) The single frame region of Tat acts in a segregated manner, with high correspondence between the patient and selection data. In the overlapped regions of Tat and Rev many absent alleles are fit in the single frame context, with Rev as the dominant gene (on average). C) Functional data matches selection data with functional residues (activity <50%) correlating well with selection (low median experimental fitness) while non-functional residues (activity >90%) have a median experimental fitness near 0 (neutral).

Figure 7: Advantageous Restriction by the Overlap

A) The fitness distribution of an overlapped virus (purple) is restricted compared to a equivalent single frame virus (S: green) restriction can be advantageous (A), neutral (N), or disadvantageous (D). B) Approximated fitness distribution of viral point mutations in a single frame virus compared to an overlapped virus in which Rev alleles are calculated for every possible Tat allele; (right) The resulting overlapped landscape (O) has twice the percentage of fit genotypes (fitness > 0) as the single frame (S) virus. C) Equivalent comparison between a single frame virus and overlapped one, in which Tat alleles are calculated for all Rev alleles; (right) The resulting overlapped landscape is approximately equivalent to the single frame with a slight increase in the percentage of fit genotypes.

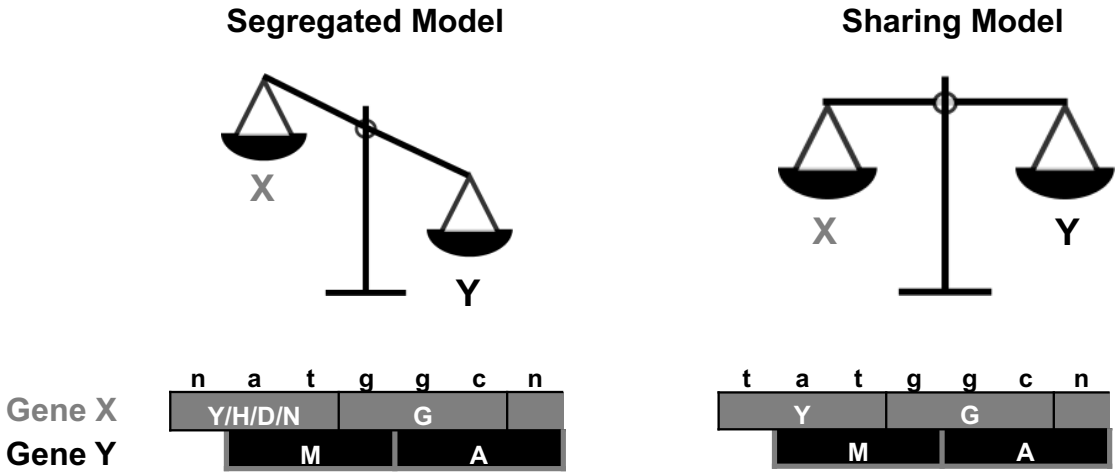


Figure S1: Models of Overlapped Evolution

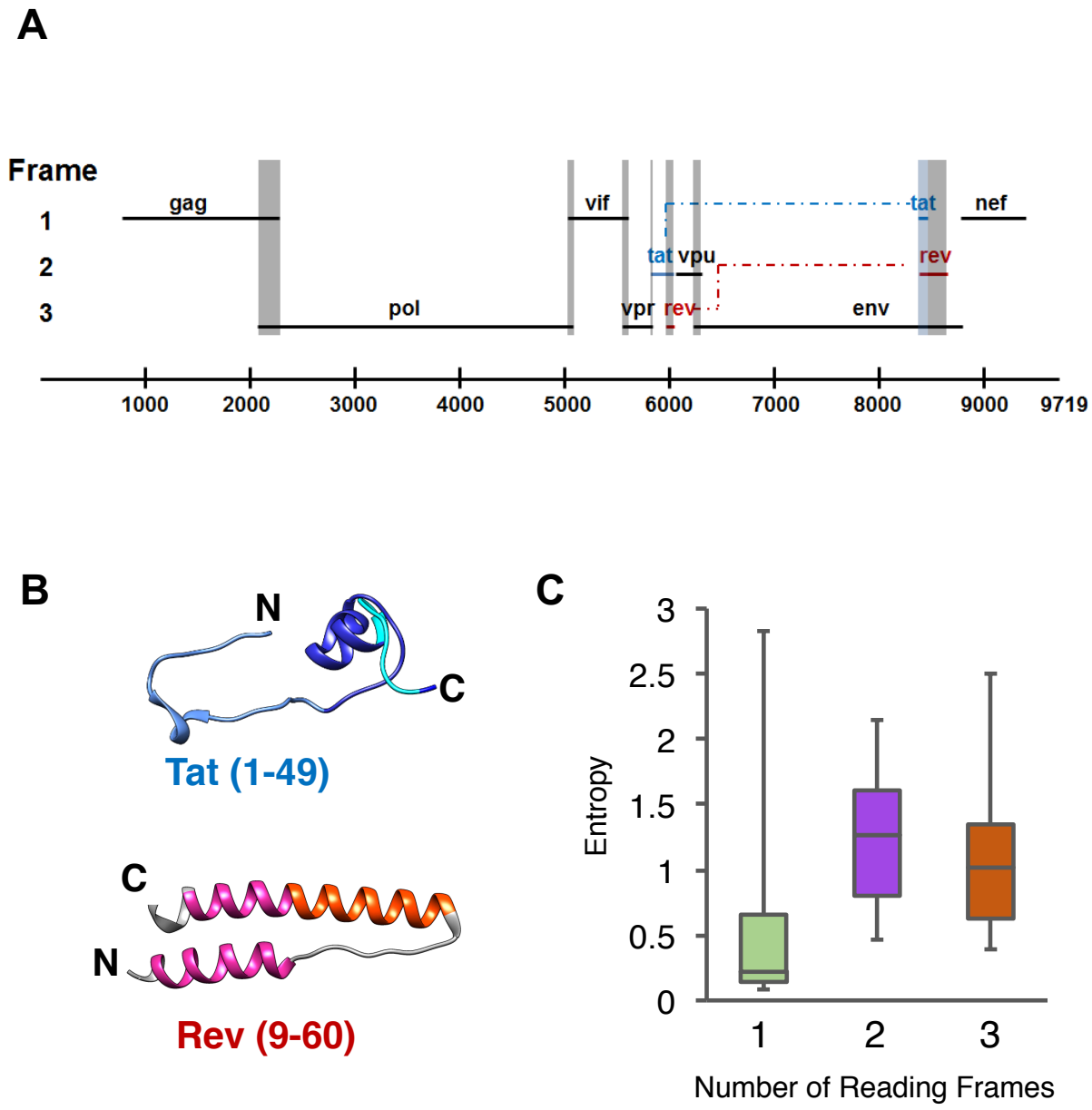


Figure S2: Conservation of Overlaps in the HIV-1 Genome

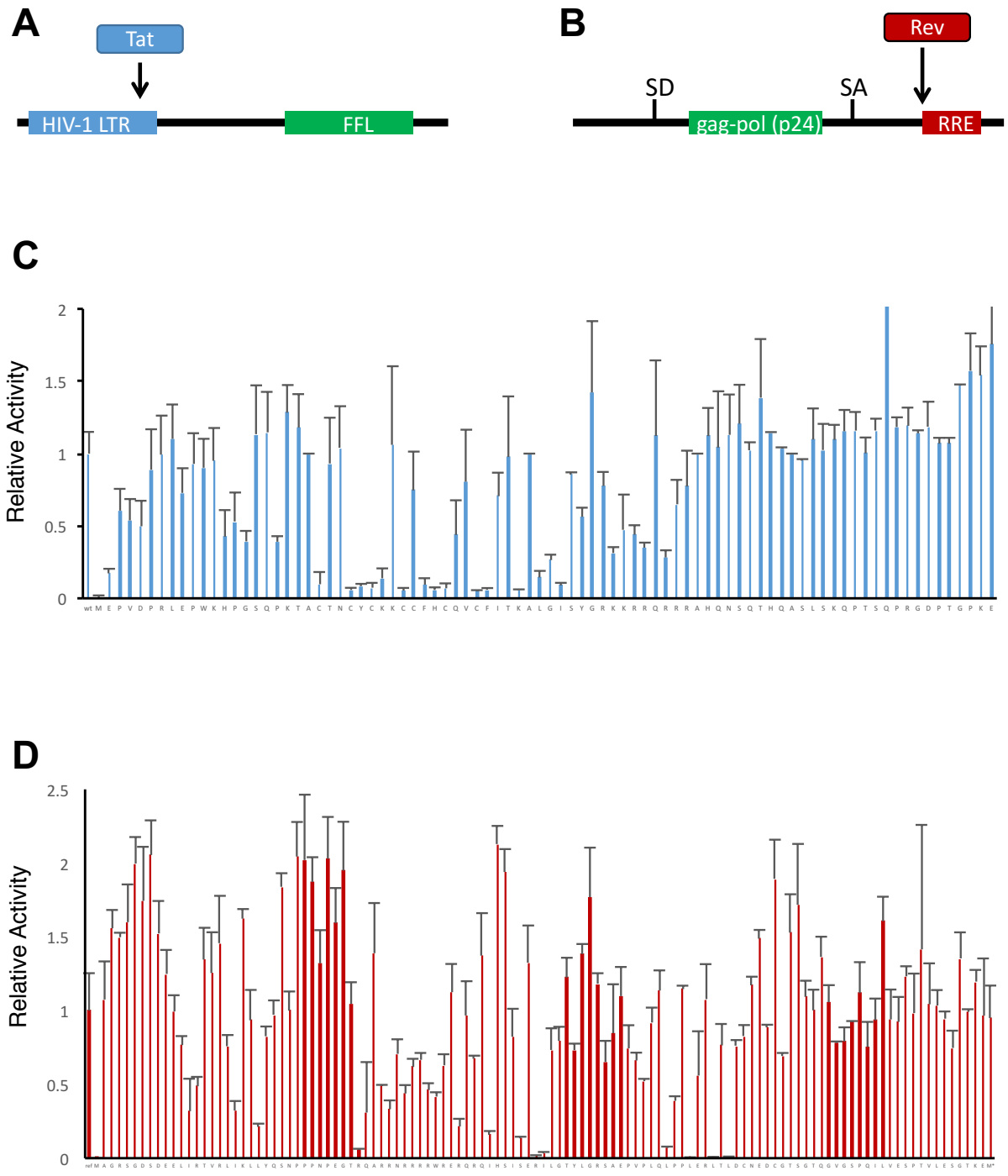


Figure S3: Alanine Scanning of Tat and Rev

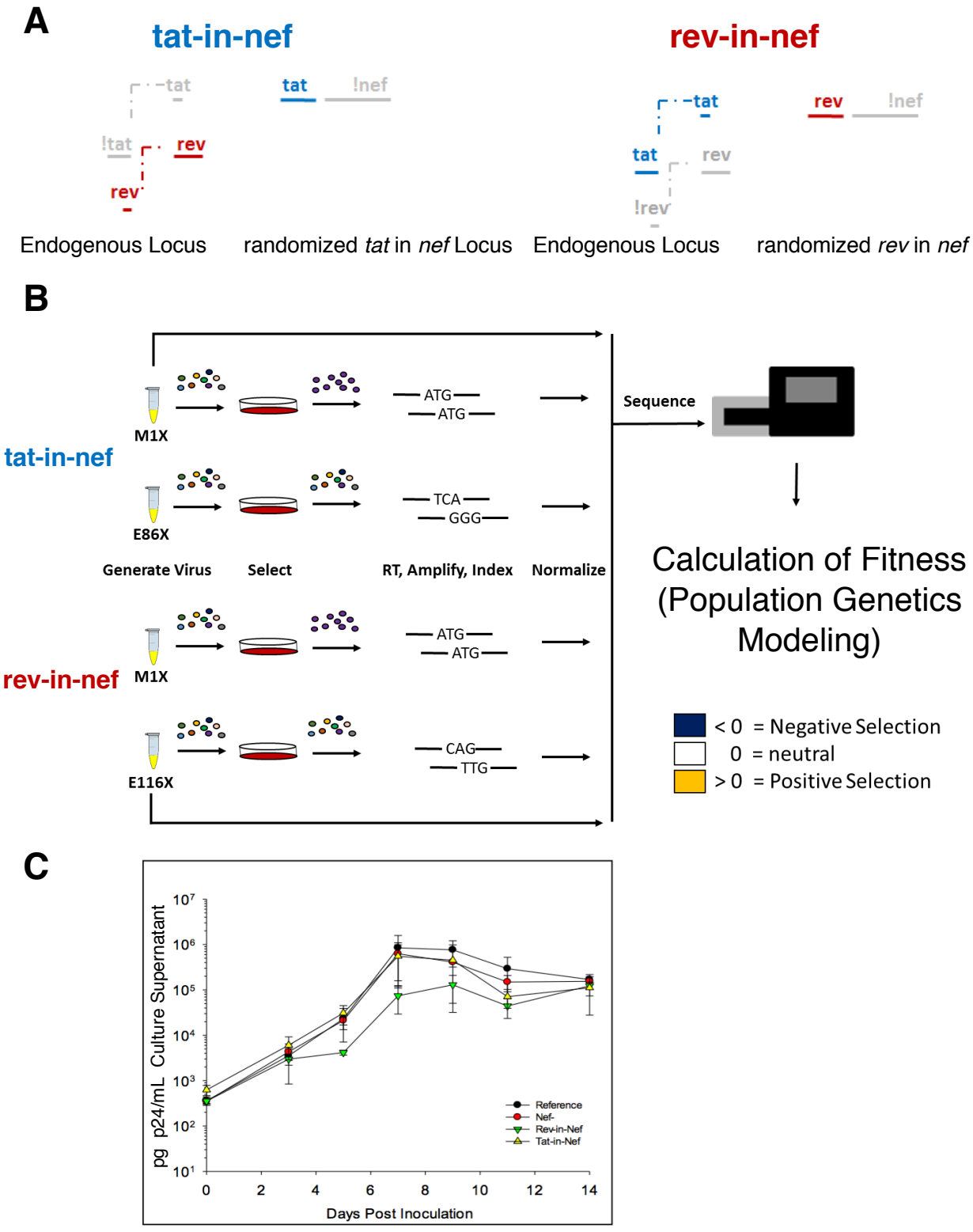


Figure S4: Generation and Selection of Uncoupled Viruses

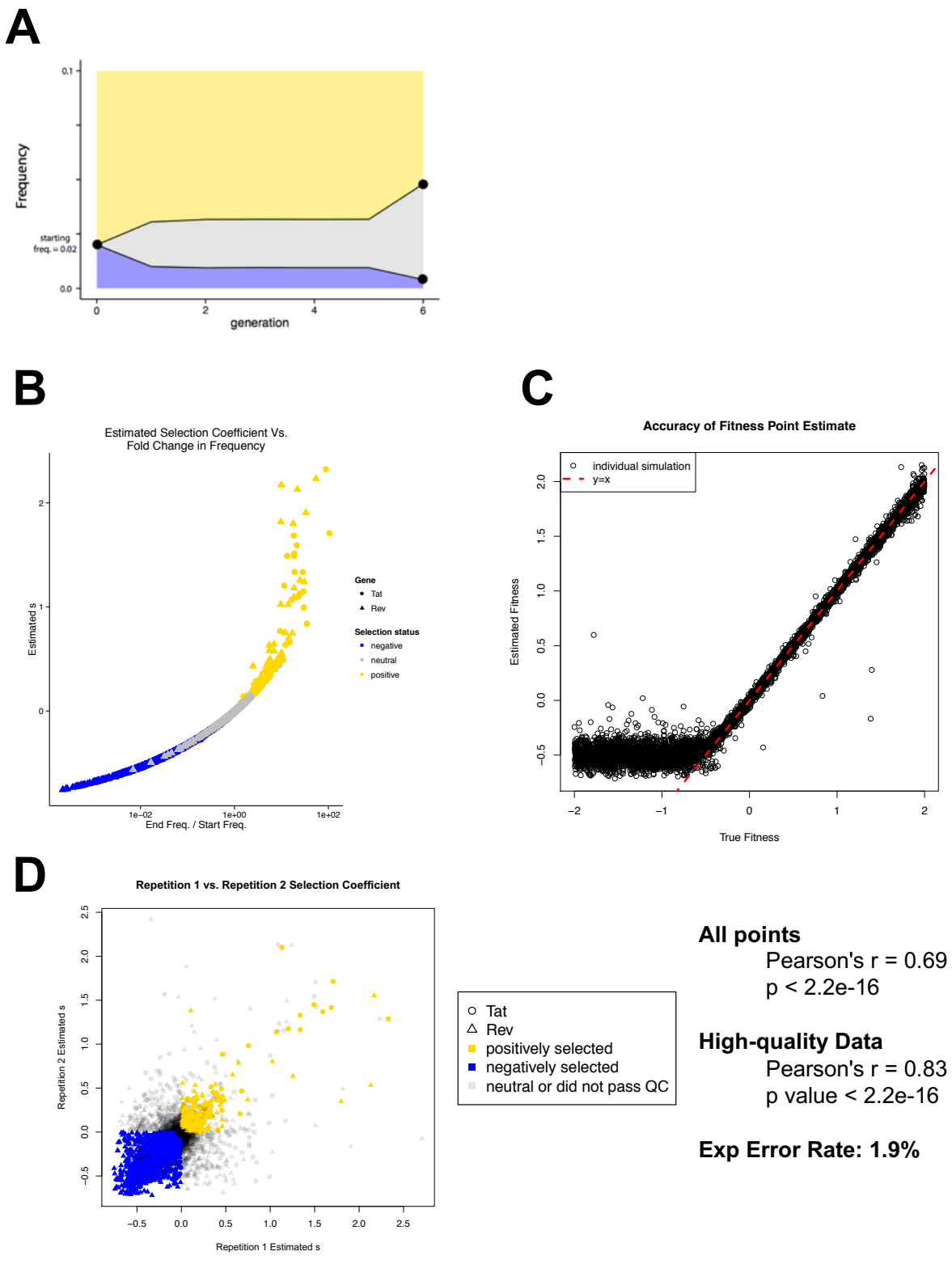


Figure S5: Population Genetics Modeling of Fitness Datasets

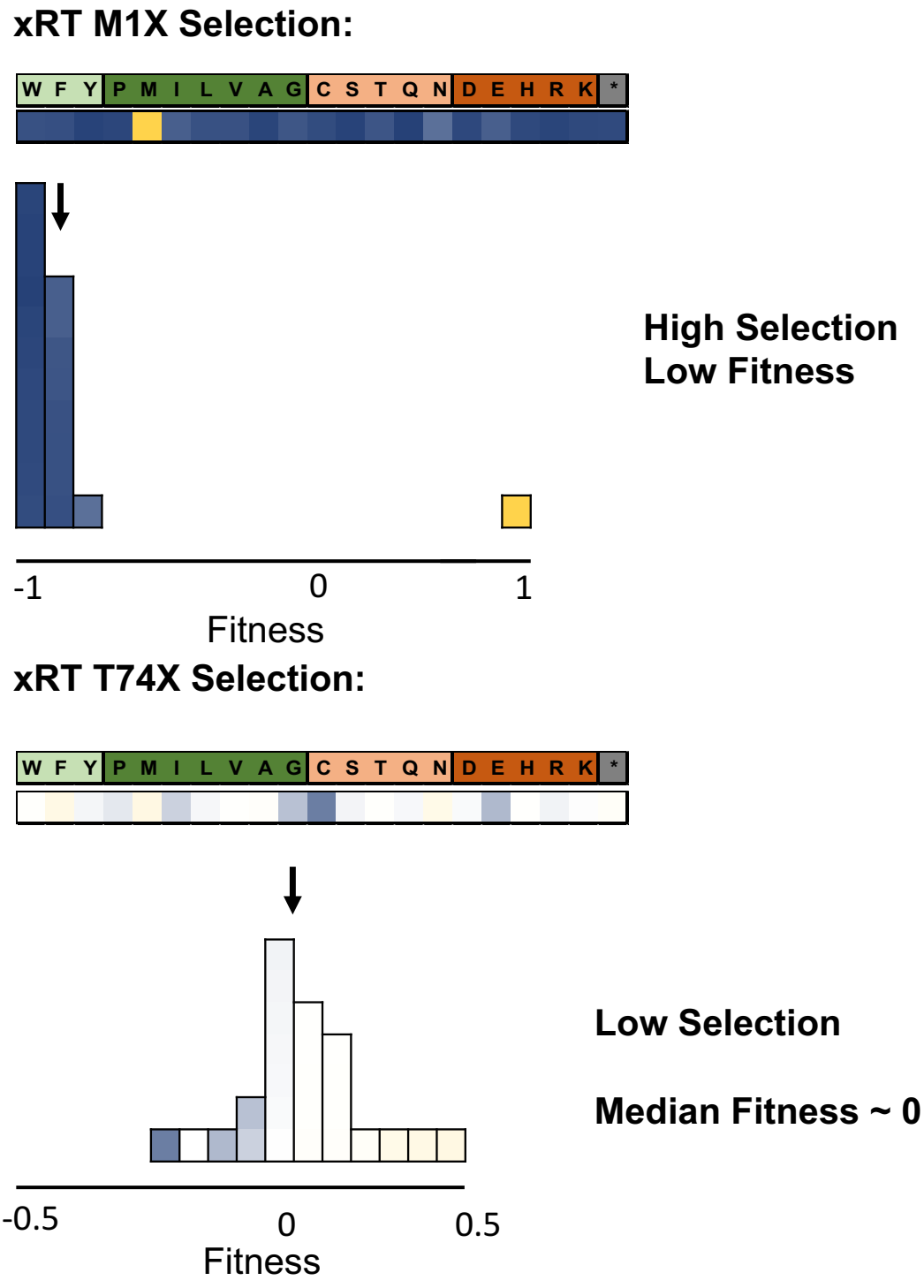
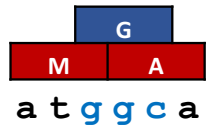
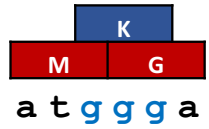


Figure S6: Median Selection as an indicator of strength of selection

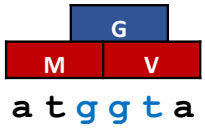
Influence of Tat G48 on Rev



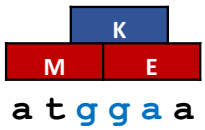
Rev M1
Rev A2



Rev M1
Rev G2



Rev M1
Rev V2



Rev M1
Rev E2

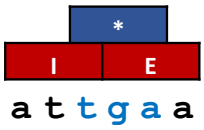
$$Sel\ Coeff_{Rev,(TatG48)} = Sel\ Coeff_{Rev\ M1} + Sel\ Coeff_{Rev\ A2}$$

$$Sel\ Coeff_{Rev,(TatG48)} = Sel\ Coeff_{Rev\ M1} + Sel\ Coeff_{Rev\ V2}$$

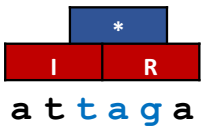
$$Sel\ Coeff_{Rev,(TatG48)} = Sel\ Coeff_{Rev\ M1} + Sel\ Coeff_{Rev\ V2}$$

$$Sel\ Coeff_{Rev,(TatG48)} = Sel\ Coeff_{Rev\ M1} + Sel\ Coeff_{Rev\ E2}$$

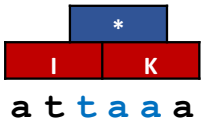
Influence of Tat *48 on Rev



Rev I1
Rev E2



Rev I1
Rev R2

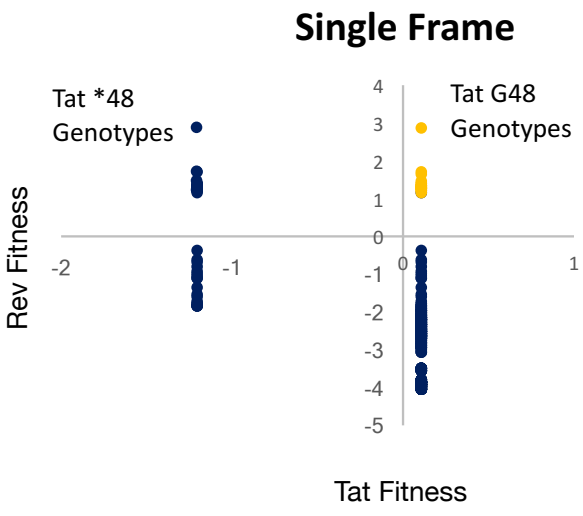


Rev I1
Rev K2

$$Sel\ Coeff_{Rev,(Tat\ *48)} = Sel\ Coeff_{Rev\ I1} + Sel\ Coeff_{Rev\ E2}$$

$$Sel\ Coeff_{Rev,(Tat\ *48)} = Sel\ Coeff_{Rev\ I1} + Sel\ Coeff_{Rev\ R2}$$

$$Sel\ Coeff_{Rev,(Tat\ *48)} = Sel\ Coeff_{Rev\ I1} + Sel\ Coeff_{Rev\ K2}$$



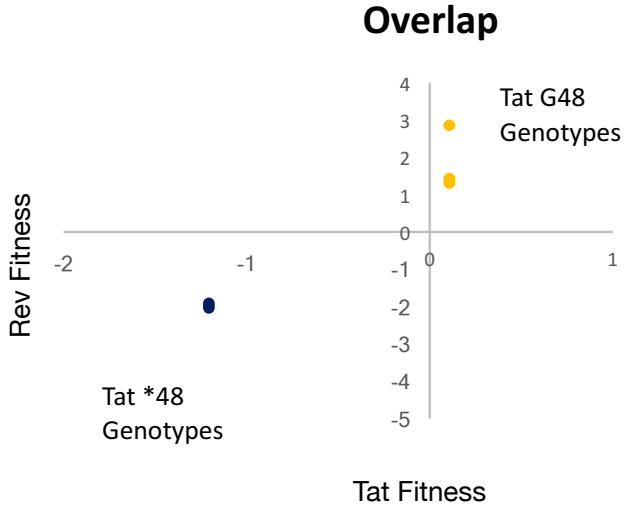


Figure S7: Inferring the Effect of Mutations in One Protein on the Other

Supplementary Figure Legends

Supplementary Figure S1: (Related to Figure 1) Models for Overlap Evolution. In the segregated model one gene drives the evolution of the other. Strict requirements of M and A in a dominant gene Y allow several equivalent changes in the sequence of accommodating gene X (D/H/N/Y) but occasionally fix particular amino acids (G in position 2 of X). In the sharing model, both genes have strong requirements of particular amino acids constraining the DNA sequence so that even synonymous codon mutations are not allowed. The sharing model is marked by stronger conservation of both proteins at the amino acid level while the segregated model allows for comparatively greater diversity.

Supplementary Figure S2: (Related to Figure 1) A) Organization of the HIV genome by reading frame. Two-frame overlaps are shaded with gray boxes and the three frame overlap with light blue. B) Partial 3D structures of Tat and Rev. Motifs are colored as labeled in Fig 1A. PDB structures 3MI9 (Tat) and 3LPH (Rev) were used. C) Genome wide entropy comparison between 1,2 and 3-frame regions.

Supplementary Figure S3: (Related to Figure 2) A) Schematic for the alanine scan for Tat and Rev (B). Tat activates transcription of firefly luciferase under the control of the HIV LTR. B) Rev allows expression of intronic p24 by exporting the message before splicing can occur. SD = Splice Donor, SA= Splice Acceptor. C) Tat activity for all mutants was reported as mutant induction of firefly luciferase relative to reference induction. D) Rev activity for all mutants was reported as mutant induction of intracellular p24 levels relative to reference induction. M* is an N-terminally Strep-

tagged M1A mutant. Error bars represent standard deviations from five (Tat) or three (Rev) biological replicates.

Supplementary Figure S4: (Related to Figures 3 & 4) A) Creation of uncoupled viruses involved mutations (ablation of start codon and introduction of downstream stops) to the endogenous frame. A synonymously substituted codon version of the knocked out gene was then introduced into the *nef* locus. The introduced gene was randomized equally amongst all codons at a single position. B) Schematic for selection experiment. Each library, representing a randomized and unconstrained version of Tat (tat-in-nef) or Rev (rev-in-nef) at a single position, was competed against itself, raised in 293 cells, and then competed against itself in SupT1 cells. Viral supernatants were harvested, genomic RNA isolated, and reverse transcribed. Amplicons targeting the randomized site were generated, indexed, normalized and then pooled and sequenced. Relative fitness was reported as a logarithmic ratio of the allele frequencies post- and pre-selection. C) Replication curves for the TRx, xRT, TxR and NL4-3 viruses demonstrate effects on viral kinetics after removing the overlap. These curves were also used to generate models for our population genetics simulations.

Supplementary Figure S5: (Related to Figures 3 & 4) A) An illustration of the neutral simulations for a hypothetical allele with a starting frequency of 0.02, an ending read depth of 500 reads, and an amino acid identity of Arginine. The grey area depicts the range of trajectories that this allele could take if it were neutral. If an ending allele frequency were observed to be above or below this neutral expectation, it is deemed positively or negatively selected, respectively. The black dots indicate the upper and lower bounds for the ending allele frequency that would still be considered neutral.

These upper and lower bounds correspond to relative fitness values of 0.380 and -0.699, respectively, which means neutrality cannot be rejected for any observed fitness value that resides between this interval. B) The observed distribution of fitness values for alleles found to be under negative (blue), neutral (grey), or positive (gold) selection. Neutral alleles were sometimes found to have relatively extreme fitness estimates (left and right tails of grey distribution). Likewise, alleles under significant positive or negative selection were sometimes found to have fitness estimates close to zero (right tail of blue distribution, and left tail of gold distribution). C) Correlation between the estimated fitness and the true fitness under our simulation framework. Each point corresponds to one simulation. D) Correlation plot of alleles between biological replicates. Alleles that have strong experimental evidence of selection show high repeatability. Correlation coefficients are shown for the whole data-set and those passing our QC criteria (note that these criteria have inherent biases, such as requiring the sign of the selection coefficient to be consistent, towards increasing the correlation coefficient). The experimental error rate (amino acids that appear to mutate outside the randomization site) is ~2%. Note that this error rate is calculated on the amino acid level and does not consider synonymous mutations, or multiple mutations within the same codon.

Supplementary Figure S6: (Related to Figures 5-6) Median Selection Coefficient as an Indicator of Strength of Selection. We take all possible alleles for a given position (e.g. tat-in-nef M1X or tat-in-nef T74X) and calculate the median for the distribution of their fitness effects (shown in the histogram). Sites like M1X have strong preference for and against particular alleles (strong selection) and a low median fitness. Sites like T74X

with near equal preference for most alleles display a more normal distribution and have a median fitness near zero.

Supplementary Figure S7: (Related to Figure 7) Inferring the Effects of Mutations in the Context of the Overlap. If one gene encodes a particular allele (i.e. Tat G48 or Tat *48) the overlapped gene is restricted to a subset of alleles based on the synonymous codons of the fixed gene. As Tat and Rev are both essential for viral replication only viral genotypes which harbor fit alleles for both genes produce fit viruses. We can explicitly calculate the fitness of every genotype in both the single frame and overlapped context for every allele in every position of each gene. Pictured in this example are explicit calculations for the effect of Tat G48 or Tat *48 on Rev.

Supplemental Table 1: Deep Mutational Scanning Data. This table contains selection coefficients (experimental fitness) for Tat and Rev alleles for each replicate as well as the mean coefficient. For several positions only one of the biological replicates gave sufficient read depth to calculate meaningful allelic frequencies – for these cases the value for a single replicate was used. Also provided are p-values for each selection coefficient (as determined by our simulation's test against neutrality), and a binary value of whether the allele's statistics passed our QC filters. Starting and ending frequencies (as well as corresponding read depth) for each allele are also provided.

Supplemental Table 2: DNA sequences of primers and genes. This table contains reference sequences for *tat* and *rev* used in our reporter assays and non-overlapped viruses. Primer sequences for amplicon sequencing are also provided.

Chapter 3: PJA2 ubiquitinates the HIV-1 Tat protein with atypical chain linkages to activate viral transcription

Abstract

Transcription complexes that assemble at the HIV-1 promoter efficiently initiate transcription but generate paused RNA polymerase II downstream from the start site. The virally encoded Tat protein hijacks positive transcription elongation factor b (P-TEFb) to phosphorylate and activate this paused polymerase. In addition, Tat undergoes a series of reversible post-translational modifications that regulate distinct steps of the transcription cycle. To identify additional functionally important Tat cofactors, we performed RNAi knockdowns of sixteen previously identified Tat interactors and found that a novel E3 ligase, PJA2, ubiquitinates Tat in a non-degradative manner and specifically regulates the step of HIV transcription elongation. Interestingly, several different lysine residues in Tat can function as ubiquitin acceptor sites, and variable combinations of these lysines support both full transcriptional activity and viral replication. Further, the polyubiquitin chain conjugated to Tat by PJA2 can itself be assembled through variable ubiquitin lysine linkages. Importantly, proper ubiquitin chain assembly by PJA2 requires that Tat first binds its P-TEFb cofactor. These results highlight that both the Tat substrate and ubiquitin modification have plastic site usage, and this plasticity is likely another way in which the virus exploits the host molecular machinery to expand its limited genetic repertoire.

Introduction

The HIV-1 Tat protein enhances transcription elongation at the viral promoter by recruiting the cellular kinase P-TEFb to paused RNA polymerase II (RNAPII) complexes(1,2). Once recruited, P-TEFb, composed of Cyclin T1 (CCNT1) and cyclin-dependent kinase 9 (CDK9) subunits, phosphorylates the C-terminal domain of RNAPII,

facilitating the transition from transcription initiation to elongation(3,4). Unlike the many DNA-binding transcription factors that directly associate with *cis* elements in promoters and enhancers(5,6), Tat activates transcription by employing an RNA-binding domain to contact a nascent structured RNA, TAR, at the 5' end of the HIV mRNA(7). A considerable fraction of P-TEFb in the cell is sequestered in an inactive state by forming an RNA-scaffolded complex with the 7SK snRNP(8–10). This inhibited complex and Tat are directly recruited to the HIV promoter, and once nascent viral RNA is synthesized, Tat and TAR competitively displace the snRNP to activate P-TEFb kinase activity(11–13). Supporting this model of activation, recent genome-wide studies have demonstrated that P-TEFb-7SK snRNP occupies the majority of promoters containing paused RNAPII(14,15). Furthermore, the splicing protein, SC35, is also enriched at cellular promoters and has been found to release the 7SK snRNP from P-TEFb in a nascent RNA-binding step(14).

Tat is regulated by several post-translational modifications that greatly expand the activity of this small viral protein. A series of acetyl transfer reactions control the RNA-binding activity of Tat, where acetylation of Lys28 by P300/CBP-associated factor (PCAF) increases the affinity of the Tat-Cyclin T1-TAR complex(16) and acetylation of Lys50 by p300 disassembles this ternary complex(17). Tat is also a substrate for methylation(18,19) and ubiquitination(20).

Ubiquitination is the process whereby the 76 amino acid protein ubiquitin becomes covalently attached via its C-terminus to a primary amine on a substrate protein. This reaction is carried out by an enzyme cascade composed of an E1 ubiquitin-activating enzyme, an E2 ubiquitin-conjugating enzyme, and an E3 ubiquitin

ligase, which generally provide substrate specificity(21). With over 600 E3 ligases targeting diverse substrates(21), ubiquitination is a major signaling pathway in the cell. In addition to substrate modification with a single ubiquitin, termed monoubiquitination, a polyubiquitin chain can be generated when ubiquitin molecules are covalently linked generally through one of the seven ubiquitin lysines. Several distinct polyubiquitin linkages are associated with unique signaling outputs, with Lys48-linked chains functioning as a general degradation signal and Lys11 linkages serving as a mitosis-specific degradation signal, while Lys63-linked chains often having positive signaling roles(22). Lys11, Lys48, and Lys63 are the most abundant and well-studies linkages in the cell(23), therefore the generalized function of the remaining ubiquitin linkages and the ligases that generate them is currently unclear.

Ubiquitin signaling has many regulatory functions in transcription(24,25). One well-studied example is the mono-ubiquitination of histone H2B, which is associated with transcriptionally active chromatin regions and regulates subsequent histone modifications(26). The tumor suppressor, p53, and proto-oncogene, Myc, are both tightly regulated by numerous ubiquitin ligases that modify multiple lysine residues in each protein. Ubiquitin modification of these transcription factors can lead to proteasomal degradation(27–30), protein stabilization(31,32), or protein re-localization(33,34). This diversity of functional outputs makes ubiquitination an attractive means to expand the function of a viral protein as it hijacks the host cell machinery, and especially to enhance the role of Tat in regulating transcription.

In the current study, we describe a candidate RNAi screen of previously identified high-confidence Tat interacting proteins(35) and report that several proteins from

ubiquitin signaling pathways are essential for Tat activity. One of these, PJA2, is a RING finger E3 ligase that polyubiquitinates Tat. Interestingly, these modifications are non-degradative and rather serve to regulate the transcriptional activity of Tat. Furthermore, Tat ubiquitination by PJA2 is an incredibly diverse signal, both at the substrate level, where multiple Tat lysines can function as ubiquitin acceptor sites, and at the level of the modification itself, where polyubiquitin chains can be generated through multiple linkages. This immense signal plasticity likely allows the virus to potently activate transcription in the face of a high mutation rate.

Results

A targeted RNAi screen establishes a significant role of ubiquitin signaling in HIV-1 transcription

Our recent proteomic interaction mapping uncovered several high-confidence host binding partners for the Tat protein(35). However, the function of these proteins in Tat-dependent transcription remains unclear. To resolve this issue, we performed a targeted RNAi screen of sixteen high-confidence interactors in a functional assay of Tat activity. For the screen, we generated a HeLa cell line containing a single, full-length HIV provirus (Supplementary Fig. 1a) deleted for the *tat* gene ($HeLa^{provirus\Delta tat}$), which allowed us precise control of transcriptional activation by transfection of a Tat-expressing plasmid (Fig. 1a). The readout of the screen was the Tat-dependent production of the HIV-1 capsid (p24) protein, allowing interrogation of Tat function in a relatively natural viral context. To validate the approach and calibrate signal strength, knockdown of CCNT1 or CDK9 resulted in a greater than 2-fold decrease in p24 production compared to a non-silencing (N.S.) siRNA (Fig. 1b, Supplementary Fig. 1c

for protein knockdown), effects comparable to other studies of P-TEFb knockdown on Tat activity(36). We therefore classified siRNAs as hits if they similarly produced a 2-fold or greater decrease in p24. Of the sixteen factors tested, nine had at least one siRNA that met this threshold (Fig. 1b, Supplementary Fig. 1b). One of these, PPM1G, has recently been shown to dephosphorylate CDK9 to locally disassemble the 7SK snRNP at the HIV promoter(12), supporting that the screen can readily identify functionally important Tat host factors**.** Examining the functions of these nine positives revealed an unexpected enrichment for ubiquitin signaling proteins, with three E3 ligases (ZFP91, PJA2, and UBE2O)(37–41), two substrate receptors for multi-subunit E3 ligases (DCAF16, FBX3)(42,43), and one de-ubiquitinating enzyme (USP11)(44) (Fig. 1c). Thus, our functional screen of Tat interacting proteins suggested an important role for ubiquitin machinery in HIV-1 transcriptional regulation.

The RING-H2 finger protein PJA2 ubiquitinates Tat

Tat is known to be ubiquitinated(20) so we surmised that one or more of these interacting, functionally important factors might be involved directly in Tat ubiquitination. To test this, we measured the ubiquitination of Tat *in vivo* upon co-expression of each of the six ubiquitin pathway proteins. We first performed these experiments in the absence of proteasome inhibitors, anticipating that non-degradative sites of ubiquitination in Tat might be used to regulate its transcriptional activity. Indeed, PJA2 co-expression dramatically increased Tat ubiquitination compared to the other candidate proteins (Fig. 2a). PJA2 is a RING-H2 E3 ligase and is a critical regulator in many kinase signaling pathways(38,39,45,46). The RING domain is a commonly used E2 interaction module in E3 ligases(21). Consistent with a role in Tat ubiquitination, expression of a PJA2 r2m

RING domain mutant (C634A, C671A) failed to stimulate Tat ubiquitination (Fig. 2b). Importantly, RNAi knockdown of endogenous PJA2 strongly reduced Tat ubiquitination (Fig. 2b), which, together with the co-expression experiments, demonstrate that PJA2 is a novel ubiquitin ligase for Tat. Interestingly, PJA2 appears to stabilize Tat protein levels, as fivefold excess of Tat plasmid was required in cells knocked down for PJA2 to yield comparable protein to non-silencing cells (Fig. 2b), consistent with Tat ubiquitination being non-degradative. Tat is a highly specific substrate as PJA2 expression did not increase the ubiquitination of two other small HIV proteins, Rev and Nef (Supplementary Fig. 2).

To demonstrate that PJA2 directly modifies Tat, we established an *in vitro* ubiquitination system. Using purified components, PJA2 readily ubiquitinates Tat only in the presence of ubiquitin, E1, and E2, generating high molecular weight ubiquitinated species (Fig. 2c). Importantly, the PJA2 r2m mutant did not modify Tat, supporting our *in vivo* results. To ensure the functional significance of PJA2-mediated Tat ubiquitination in regulating transcription elongation, we monitored the effect of PJA2 knockdown on the production of HIV-1 promoter-proximal and distal RNA transcripts using the HeLa$^{provirus\Delta tat}$ cell line. While PJA2 knockdown did not affect promoter-proximal RNA production, it significantly decreased the level of Tat-dependent distal transcripts, an effect as potent as the knockdown of CDK9 (Fig. 2d). Thus, ubiquitination of Tat by PJA2 is a novel signal that regulates its transcriptional activity. The additional ubiquitin pathway proteins that are critical for Tat-dependent activity (Fig. 1b) likely modulate the ubiquitination of other Tat cofactors. Supporting this idea, the UBE2O ligase ubiquitinates proteins of the 7SK snRNP to regulate the activity of this complex

(accompanying manuscript). However, here we further analyzed the direct PJA2-dependent modification of this essential viral protein.

Tat ubiquitination by PJA2 is non-degradative

All known substrates of PJA2 are targeted to the proteasome by ubiquitination(38,39,45,46), however our experiments suggested that PJA2 ubiquitination of Tat is not degradative. For example, PJA2-dependent ubiquitinated Tat species are readily detectable in the absence of proteasome inhibitors (Fig. 2a), and PJA2 co-expression does not decrease steady-state Tat protein levels (Fig. 2a, STREP blot), which is characteristic of degradative ubiquitination, including previously defined PJA2 substrates. Indeed, PJA2 actually stabilizes Tat, as RNAi-mediated depletion of the ligase resulted in lower Tat protein accumulation (Fig. 2b). However, when we treated cells with the proteasome inhibitor MG132, both unmodified and ubiquitinated Tat levels increased, suggesting that Tat can be degraded by the proteasome (Supplementary Fig. 3).

To determine if this apparent proteasome-mediated degradation is ubiquitin-dependent, we asked whether preventing ubiquitination would increase Tat protein levels. To block ubiquitination, we generated a Tat mutant in which all lysine side chains were mutated to arginine (Tat ΔK). A similar lysine-free mutant of Myc results in protein stabilization(30), however in the case of Tat ΔK, not only was no stabilization observed but Tat protein levels became undetectable (Fig. 3a, lane 3). One likely explanation is that mutating Lys41, which is critical for Tat folding(47), completely destabilizes Tat and provides an unfolded substrate for ubiquitin-independent 20S core proteasomal degradation. Consistent with this, MG132 treatment fully restored Tat ΔK protein levels

(Fig. 3a, lane 4). This result is also consistent with studies showing that Tat levels are decreased by molecules that activate the 20S core proteasome(48,49). While MG132 restored the level of unmodified Tat ΔK to that of wild type Tat, no ubiquitinated species were observed (Fig. 3a, lane 4), confirming that Tat ubiquitination is lysine-dependent. These experiments conclusively demonstrate that Tat protein levels can be controlled by the proteasome but not through lysine-dependent ubiquitination, and therefore PJA2-mediated ubiquitination of Tat is non-degradative.

PJA2 polyubiquitinates Tat with diverse, atypical linkages

It is generally believed that the Lys63 polyubiquitin linkage is the major signaling modality among ubiquitin systems, whereas all other linkages generally result in proteasomal degradation(50). However some recent examples point to positive signaling functions for non-Lys63 linkages(34,51). Because Tat ubiquitination is non-degradative, we wished to determine the nature of its linkage. We generated all single lysine-to-arginine mutants of ubiquitin as well as a lysine-free mutant (Ub ΔK), and co-transfected each with Tat, expecting that a point mutation would prevent addition of polyubiquitin chains through that specific residue while Ub ΔK would prevent all chain addition. As expected, Ub ΔK significantly inhibited Tat polyubiquitination (Fig. 3b), indicating that the modified species of Tat observed are indeed polyubiquitinated and not the result of multiple monoubiquitin additions. Interestingly, no single lysine point mutant diminished Tat polyubiquitination, suggesting that diverse chain linkages can be used to modify Tat in the cell. To test if this apparent plasticity of ubiquitination linkages reflects an underlying functional plasticity, we expressed each ubiquitin mutant and measured their effects in a Tat reporter assay(11). Correlating with the ubiquitination

data, expression of Ub ΔK, and not the point mutants, strongly inhibited Tat activity (Fig. 3c), supporting a functional role for diverse polyubiquitination in Tat-mediated transcription.

This diversity of chain linkages is rather surprising for ubiquitin signaling and prompted us to test if ubiquitins engineered with single lysines (Lys^{only}) could restore full Tat polyubiquitination *in vivo*. Remarkably, $Lys27^{only}$ and $Lys29^{only}$ variants allowed full ubiquitination of Tat and $Lys33^{only}$ showed partial recovery (Fig. 3d). These three lysines are contained on a single alpha helix in ubiquitin (Fig. 3f), suggesting that any one of these residues can serve as an acceptor site during chain formation on Tat. To further corroborate that this linkage diversity was being generated by PJA2, we used purified PJA2 to ubiquitinate free Tat *in vitro* and unexpectedly found no linkage preference with the single lysine ubiquitin variants, each being similar to wild type ubiquitin (Fig. 3e, top). Remarkably, however, when Tat in the context of the P-TEFb complex was used as the substrate, where Tat folds using the host proteins as a template(47), we observed a striking preference for Lys27 and Lys29 linkages (Fig. 3e, bottom) that mirrored the linkage preference *in vivo* (Fig. 3d). These results not only highlight that the Tat-P-TEFb complex is the preferred ubiquitination substrate in the cell, but also support that PJA2 is the functional E3 ligase. The experiments thus far demonstrate that Tat can be modified with diverse, rare polyubiquitin chains by PJA2, highlighting an unanticipated activity of this host ligase, and also reveal how the substrate can exert such a large effect on the specificity of polyubiquitin acceptor site usage.

Multiple lysine residues in Tat function as ubiquitin acceptor sites

Having shown that Tat ubiquitination is lysine-dependent (Fig. 3a), we wished to determine its ubiquitin acceptor sites. Degradative ubiquitination often displays low lysine specificity on the substrate, as the major signal is the ubiquitin chain itself, whereas non-degradative, signaling ubiquitination generally requires an additional layer of specificity in acceptor site position(52). Given the non-degradative nature of Tat ubiquitination, we expected to find a specific site of modification. Unexpectedly, no Tat lysine-to-alanine point mutant inhibited ubiquitination (Fig. 4a). The apparent lack of modification of K41A can be explained by low overall protein accumulation (Fig. 4a, STREP), as this mutation is known to destabilize Tat structure. This expression defect also explains why Tat ΔK was undetectable in the absence of MG132 (Fig. 3a).

One explanation for these results is that Tat can be ubiquitinated on multiple lysines, such that a single mutation does not inhibit ubiquitin conjugation. To explore this possibility, we used mass spectrometry (MS) to identify all *in vivo* ubiquitinated side chains and detected di-glycine ubiquitin remnants on three of the nine Tat lysine residues (Lys12, Lys29, and Lys71) (Supplementary Fig. 4). This result is consistent with Lys71 being previously identified as an acceptor residue(20). The Tat-P-TEFb crystal structure, which contains five lysines (Lys12 to Lys41) within the structured fragment of Tat, permits us to further interpret the results. Lys12 faces CDK9 in the complex but is surface-exposed and, therefore, a suitable acceptor site (Fig. 4b). Lys19 is at the interface with CycT1, which likely hinders its modification. Interestingly, of the next two adjacent, surface-exposed lysines (Lys28 and Lys29), only Lys29 is detectably ubiquitinated (Supplementary Fig. 4b). Lys28 is a known site of acetylation(16), which may compete with its ubiquitination. Finally, Lys41 is a buried residue and is not

detectably ubiquitinated (Fig. 4b). These data support a model in which multiple Tat lysines can function as acceptor residues. Together with the observed plasticity of lysine linkages in the polyubiquitin chain described above, the results suggest that an extraordinary degree of ubiquitin signaling diversity is possible on this one viral transcription factor.

A Tat protein with only three lysines is fully functional and ubiquitinated

While Tat can be ubiquitinated on multiple lysines, it is unclear whether one or more of these modifications are important for function. To examine the relationship between ubiquitination and function, we first measured the activities of all single Tat Lys-to-Arg mutants (Fig. 5a) and found that only mutation of Lys28 or Lys41 significantly inhibited Tat activity, in agreement with previous reports(13). However, because multiple lysines can serve as acceptor sites (Fig. 4a, b), we reasoned that there may be redundancy among ubiquitination sites such that no single mutation eliminates Tat function. To limit this possible masking effect, we utilized the lysine-free Tat ΔK background and restored individual lysines, observing that only Lys41 recovered activity (Fig. 5b). This result highlights that Lys41 is required to create a proper structural framework, but even in this case activity was still 2.5-fold lower than wild type Tat, indicating that other lysines also are important. We next restored individual lysines within the Lys41-only background and found that Lys28 recovered the most activity (Fig. 5c), supporting the Lys-to-Arg mutant data (Fig. 5a). To fully recover function, we introduced the remaining seven lysines individually to the Tat ΔK (+K41, +K28) background and found, surprisingly, that any one of four different lysines (at positions 12, 19, 51, or 85) was sufficient to restore wild type activity (Fig. 5d). Thus, Tat requires

just three lysines for full function, and the location of the presumed ubiquitination site is flexible. Interestingly, all of the three-lysine Tat variants show near wild type levels of ubiquitination (Fig. 5e) even though just four of these sites preferentially restore activity. It is worth noting that adding Lys28 to the Lys41-only background recovers some ubiquitination (Fig. 5e, compare lanes 2 and 3), suggesting that Lys28 may be acetylated or ubiquitinated, analogous to the overlapping modifications of two lysines in Smad7(53).

Finally, we used the more sensitive viral spreading assay to examine these lysine requirements in the context of HIV-1 replication. Tat lysine mutants were cloned into the Nef locus, which is dispensable for viral replication in tissue culture, of a fully replication-competent HIV-1 NL4-3 clone. As expected, the lysine-free Tat ΔK virus was completely replication incompetent (Fig. 5f) whereas both Tat ΔK (+K41) and Tat ΔK (+K41, +K28) displayed slower replication kinetics than wild type. The Tat ΔK (+K41) virus in particular, which contains no modifiable lysines, never reached the peak titer of wild type virus, even at 10 days post infection. Notably, two different three-lysine Tat viruses replicated faster than the two-lysine Tat ΔK (+K41, +K28) virus, corroborating that the third, presumably ubiquitinated, lysine results in a more active Tat and that its precise location is flexible. These viral replication experiments underscore an even more pronounced phenotype for the ubiquitinated Tat lysines.

Tat enhances the interaction between PJA2 and P-TEFb

Given that the Tat-P-TEFb complex is the preferred substrate for PJA2, we examined the interactions of these proteins in more detail. Transiently transfected wild type Tat strongly precipitated endogenous PJA2 in an RNA-independent manner (Fig.

6a). Importantly, Tat C22A, an unfolded Tat mutant unable to interact with P-TEFb(13), did not pull down PJA2, supporting that Tat folded in the P-TEFb complex is the preferred substrate for PJA2 (Fig. 3f). Additionally, Tat ΔK (+K41), which is not detectably ubiquitinated (Fig. 5e), interacted with PJA2 similarly to wild type Tat, demonstrating that Tat does not need ubiquitin acceptor sites to interact with its ligase.

We next mapped the interaction determinants in PJA2 using a series of N-terminal truncation mutants. Unexpectedly, the full-length but catalytically inactive PJA2 r2m pulled down substantially more Tat than wild type PJA2 (Fig. 6b), likely due to r2m expression also stabilizing Tat protein levels (Fig. 6b, input lane 2). This stabilization of Tat is supported by the observation that Tat protein levels decreased with siRNA knockdown of PJA2 (Fig. 2b). The Tat-interaction region of PJA2 maps between amino acids 200-300, highly conserved between human and mouse, and different from the 400-530 or 530-630 regions that interact with other protein partners(38,39,45). Like the result with r2m, binding of the truncation mutants also correlates with Tat stabilization. Given that Tat complexed with P-TEFb is the preferred ubiquitination substrate for PJA2 in the cell (Fig. 3d, e), we also examined the interaction of PJA2 with P-TEFb. Transiently transfected PJA2-FLAG pulled down P-TEFb, and, to our surprise, the co-expression of Tat dramatically increased the recovery of CCNT1 and CDK9 subunits in the wild type and r2m PJA2-FLAG precipitate (Fig. 6c). These data further support a novel signaling axis composed of a Tat-PJA2-P-TEFb complex (Fig. 6d) that controls a currently undefined step in the transcription pathway.

Discussion

Considerable effort has been dedicated to identifying the cellular protein complexes hijacked by HIV to complete its life cycle. HIV infection is currently incurable as the virus integrates into the human genome, and, depending on the cellular context, can enter a “latent” state of infection. In this latent state, the virus remains stably integrated, yet minimal or no viral RNA is transcribed, preventing clearance and resulting in a persistent infection(54). The HIV Tat protein controls viral gene expression and, therefore, provides an excellent framework to understand mechanisms that regulate and activate latency(55). Of particular importance, then, are the host proteins that Tat co-opts to induce transcription.

Here we report on the novel function of the RING finger ubiquitin ligase PJA2 to ubiquitinate Tat and control viral transcription. PJA2 has been shown to induce the degradative ubiquitination of core and regulatory components of several kinase complexes(38,39,45,46). This is the first report of non-degradative ubiquitination for this ligase. Further, we show that PJA2 specifically targets Tat bound to P-TEFb to generate diverse, atypical polyubiquitin linkages through Lys27, Lys29, and Lys33. It is intriguing that Tat ubiquitination has the signature of a degradative mark, as it is targeted by a typically degradative ligase and the site specificity of ubiquitination is low. Similar to the other known PJA2 substrates, Tat functions as a regulator of a kinase complex (P-TEFb) and even stimulates the interaction between PJA2 and P-TEFb. One exciting possibility is that Tat rewired degradative ubiquitination by PJA2 into a positive signaling function. Ubiquitination appears to stabilize Tat protein levels as PJA2 co-expression increases Tat protein accumulation while PJA2 knockdown destabilizes Tat. PJA2 may therefore act to delay the transition to latency by increasing the half-life of Tat

molecules. In addition to its own ubiquitination, Tat may even employ PJA2 to regulate the kinase activity of CDK9. Future work will uncover the amount of cross talk between the kinase and ubiquitin signaling pathways in HIV transcription.

HIV encodes a small 9kb RNA genome with limited coding capacity. However, the virus has found many solutions to maximize the information contained in its genome, such as using an adaptable Rev monomer to assemble a higher order oligomeric RNA export complex(56). We believe that the post-translational ubiquitination of Tat represents another viral strategy of genetic multi-tasking. In this case, one lysine residue encodes at least two possible protein states, modified and unmodified, each with an ability to execute a distinct function. Ubiquitination in combination with other known Tat post-translational modifications would allow a single viral protein to generate a large regulatory network to finely control transcription. The observation that re-introducing Lys28 restores substantial ubiquitination (Fig. 5e, lane 3) demonstrates that one amino acid can even encode three states (unmodified, acetylated(16), ubiquitinated). It is unclear at the moment how Tat transitions through these three states to activate transcription. Our current work suggests a model where Tat first binds P-TEFb, which creates a substrate for ubiquitination by PJA2. This ubiquitinated Tat species might facilitate the formation of another protein-protein complex (e.g. the super elongation complex(36)) required for transcription. As Lys28 acetylation is eventually required for TAR binding, any ubiquitin mark on Lys28 would need to be removed. Interestingly, recent analyses of the linkage specificity of the OTU family DUBs revealed that OTUD2 and OTUD6A preferentially cleave Lys27, Lys29, and Lys33 di-ubiquitin chains(57), demonstrating the presence of cellular machinery dedicated to recognizing

and cleaving the atypical ubiquitin chains of the types found on Tat. While a lysine deacetylase dedicated to specifically removing Tat Lys28-Ac is not known, it is likely that this modification is also dynamically regulated to allow coordinated rounds of modification of a single lysine residue. It will also be interesting to see if post-translational modifications at different sites (e.g. Lys28 and Lys50) act in an exclusive or combinatorial manner.

Compared to other transcription factors, Tat is a relatively small protein. The covalent attachment of a polyubiquitin chain drastically increases Tat's effective size, generating a large signaling scaffold with multiple protein-protein interaction surfaces. At the moment, however, it is unclear which proteins in the Tat transcriptional circuit possess ubiquitin-binding domains. Determining which proteins can read the ubiquitin mark would help uncover the exact output of the ubiquitin signal. CDK9 has been shown to be ubiquitinated by the SKP2 ligase(58), resulting in an increased affinity of the Tat-TAR-P-TEFb complex. In this case, it is uncertain whether the ubiquitin moiety itself contributes to the increased affinity or whether a protein with a ubiquitin-binding domain is recruited to CDK9-Ub to stimulate the affinity. In any case, identifying "readers" of the ubiquitin mark more generally is an exciting area of research.

The lysine mutant reporter data (Fig. 5 a-d) highlight a striking result in which a Tat protein with different combinations of only three lysines is fully functional. However, there is a clear epistasis to this plasticity, as a functional Tat strictly requires Lys41, which helps it adopt its structure by folding on the CCNT1 scaffold via a set of hydrogen bond interactions(47). A plastic ubiquitination mark, as detailed here, can then be layered over this partially folded scaffold. The functional data are supported by

conservation analysis, which shows that Lys28 and Lys41 are the two most conserved lysines in the Tat activation domain(47). The additional ubiquitinated lysines appear to tolerate mutation since multiple lysines can fully substitute as the acceptor site. These results are reminiscent of the plasticity of Tat for TAR RNA recognition, where a single arginine within a background of lysines could substitute for its arginine-rich motif when located at two different positions(59). The virus appears to have encoded Tat in a way that balances strict amino acid requirements with relaxed positional requirements for distinct functions, thereby producing a protein that is more resilient to the high mutation rate of the virus.

It has become increasingly clear that HIV-1 interacts extensively with ubiquitin conjugation pathways during the course of infection. Degradative ubiquitin signaling is hijacked by the Vif protein, which recruits the cellular cytidine deaminase APOBEC3G to a CBFβ-Elongin B/C-Cullin 5 ubiquitin ligase for cleavage(60,61), and the Vpu protein, which assembles a βTRCP-SKP1-Cullin1 complex to target the cytoplasmic domain of the CD4 receptor for ubiquitin-dependent degradation(62,63). A non-degradative, signaling function has been shown for Gag, where monoubiquitination of the p6 domain contributes to proper budding of the virus(64). Here we demonstrate an example of non-degradative ubiquitination for the Tat protein by the host E3 ligase, PJA2. The incredible diversity of modifiers (>600 E3 ligases in mammalian cells(21)), modified states (monoubiquitin, polyubiquitin with diverse linkages), and signaling outputs (degradation, re-localization, altered protein-protein interactions) makes the ubiquitin modification system an attractive target for an invading virus to hijack in order to rewire the cellular environment for sustained infection. Future research will uncover how the remaining

ubiquitin signaling proteins identified here also contribute to regulating viral transcription.

Methods

Cell culture, plasmids, reagents: HeLa cells and derived cell lines (HeLa$^{provirus\Delta tat}$, HeLa$^{LTR:FFL}$ (11)) (ATCC, CCL-2) and HEK 293T cells (ATCC, CRL-3216) were maintained in DMEM (10% FBS, 1% pen/strep) at 37°C with 5% CO_2. Jurkat E6-1 cells (ATCC, TIB-152) were maintained in RPMI (10% FBS, 1% pen/strep) at 37°C with 5% CO_2. All proteins were cloned into pcDNA4-TO (Thermo) unless otherwise noted. Ubiquitin signaling proteins were cloned with a C-terminal 3x-FLAG tag, the HIV-1 Tat protein (HXB2 isolate) was cloned with a C-terminal 2x-STREP tag, and human ubiquitin was cloned with an N-terminal HA tag. All point mutants were generated by standard quick change mutagenesis. The Tat ΔK and ubiquitin ΔK mutants were ordered as a gBlock gene fragment from IDT. PJA2 truncation mutants were generated by PCR and cloned with a C-terminal 3xFLAG tag vector. DNA was transfected with Poly Jet (Signa Gen) or PEI and siRNA was transfected with HiPerFect (Qiagen) according to the manufacturer's protocols. MG132 was purchased from Selleckchem (S2619). 2x and 4x Laemmli sample buffer were purchased from BioRad. siRNA was purchased as a FlexiPlate from Qiagen.

Antibodies: The HA antibody (MMS-101P) was purchased from Covance. The PJA2 antibody (A302-991A) was purchased from Bethyl. The CCNT1 (sc-8127), CDK9 (sc-8338), and Tat (sc-65913) antibodies were purchased from Santa Cruz Biotechnology. The STREP-HRP antibody (71591) was purchased from Millipore. The FLAG antibody

(F1804) was purchased from Sigma. The GAPDH antibody (AM4300) was purchased from Ambion.

Cloning, expression, and purification of recombinant proteins: Full length PJA2 with a C-terminal 6x His-tag, Tat (1-86) with a C-terminal 2x STREP tag, human ubiquitin and mutants (ΔK, $K6^{only}$, $K11^{only}$, $K27^{only}$, $K29^{only}$, $K33^{only}$, $K48^{only}$ and $K63^{only}$) with an N-terminal HA tag were cloned into the pGEX6p-1 vector (GE Healthcare) and expressed as a fusion protein with an N-terminal GST tag. The fusion proteins were expressed in *E. coli* BL21 (DE3). The GST-fusion proteins were digested on column with PreScission protease. The released ubiquitin proteins were collected and further purified by Superdex 200 Increase (GE Healthcare). The released PJA2 and Tat were further purified by nickel affinity chromatography or Strep-Tactin affinity purification (IBA), respectively. CDK9 (7–332) with a 6x His-tag and cyclin T1 (1–266) were cloned into pFastBac Dual vector, and HIV-1 Tat (1–86) was cloned into the pFastBac1 vector. The complex of CDK9/cyclin T1/Tat was expressed and purified as previously described (47). Human E1 and UbcH5b protein were a gift from the lab of John Gross.

Targeted RNAi screen of high confidence Tat interactors: HeLa cells were infected with a VSV-G pseudo-typed HIV virus deleted for the Env, Tat, and Nef proteins by start site mutations. GFP was cloned in the Nef locus, which is dispensable in cell culture, to facilitate clonal selection. One clone with a single integrant, $HeLa^{provirus\Delta tat}$, was used for the screen. Cells were seeded in a 96-well plate and transfected with non-silencing (N.S.) or host-specific siRNA to obtain 5nM-10nM final concentration using HiPerFect. All siRNA sequences are listed in Supplemental Table S1. After 48h, cells were transfected with a Tat-expressing plasmid and a CMV-firefly luciferase plasmid as a

transfection and toxicity control using PEI . Care was taken to transfect Tat-expressing plasmid in the linear range of activation. After an additional 48 hours, cells were lysed in cell lysis buffer (5mM Tris-HCl pH 8.0, 85mM KCl, 0.5% IGEPAL CA-630). Cell-associated HIV p24 production was detected by p24 ELISA. In parallel, 10uL of lysate was used in a standard luciferase assay. All p24 values were normalized to firefly luciferase activity for every siRNA to generate a relative Tat activity score, defined as [$p24^{\text{host RNAi}}/FFL^{\text{host RNAi}}$] / [$p24^{\text{N.S. RNAi}}/FFL^{\text{N.S. RNAi}}$]. For analysis of protein knockdown, RNAi cells were lysed with 2x Laemmli sample buffer followed by SDS-PAGE/Western blotting with specific antibodies.

***In vitro* ubiquitination assay:** In vitro ubiquitination assays were performed with purified recombinant human E1 (0.1 μM), recombinant human UbcH5b (0.2 μM), 0.1 μM E3 (full length PJA2 or mutant r2m), 5 μM HA-tagged ubiquitin (wild type or mutants) and 0.3 μM Tat or Tat/CCNT1/CDK9 complex in reaction buffer containing 50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1mM DTT, 5 mM $MgCl_2$, 4 mM ATP at 37°C. Reactions were quenched by addition of 4x Laemmli sample buffer with b-mercaptoethanol before loading onto a gel for SDS-PAGE. Ubiquitination of Tat was verified by immunoblotting with anti-STREP or anti-Tat antibody, as indicated.

Denaturing *in vivo* ubiquitination assay: HEK 293T or HeLa cells were seeded in 10 cm dishes and transfected with 7μg total DNA containing STREP-tagged bait protein and HA-Ubiquitin with Poly Jet. After 30 hours, cells were lysed in an equal volume SDS Lysis Buffer (2% SDS, 50mM Tris-HCl pH 8.0, 150mM NaCl, 0.5mM DTT) and boiled. After boiling, SDS Dilution Buffer (1% SDS, 50mM Tris-HCl pH 8.0, 150mM NaCl, 0.5mM DTT) was added to samples, which were then sonicated and centrifuged.

Cleared supernatant was incubated with Streptactin Agarose for 3 hours at room temperature. The Streptactin resin was then washed 4x with RIPA buffer (1% Triton X-100, 50mM Tris-HCl pH 8.0, 150mM NaCl, 0.5% Na-deoxycholate, 0.1% SDS) at room temperature. Bound STREP-tagged protein was eluted with 1x STREP elution buffer (IBA-GmbH). Ubiquitination of immunoprecipitated material was detected by anti-HA Western. For RNAi-ubiquitination experiments, cells were first transfected with siRNA for 48 hours prior to DNA transfection. For MG132 treatment, cells that had been transfected with DNA for 24 hours were then incubated with 2.5μM MG132 for an additional 16 hours before harvesting.

RNA elongation assay: HeLa$^{provirus\Delta tat}$ cells seeded at 40% confluence were transfected with siRNA to obtain 10nM final concentration. After 48 hours, cells were transfected -/+ a Tat-expressing plasmid for an additional 10 hours. RNA was purified on Zymo Quick-RNA Mini Prep columns, DNAse-treated with TURBO DNAse (Thermo), and used in RT-qPCR reactions with proximal (Prox F 5'-CTCTCTGGTTAGACCAGATCTGA-3', Prox R 5'-GTGGGTTCCCTAGTTAGCCAGA-3') and distal (Dist F 5'-CACATGAAGCAGCACGACTT-3', Dist R 5'-GGTCTTGTAGTTGCCGTCGT-3') HIV primer sets. All values were normalized to the abundance of RPLP0 mRNA.

Transcriptional reporter assay: We used a previously characterized HeLa cell line containing an integrated copy of the HIV LTR driving the expression of firefly luciferase (11). Reporter assays were carried out as previously described (13). Briefly, HeLa$^{LTR:FFL}$ cells were transfected with wild type or mutant Tat-expressing plasmid with PolyJet. CMV-Renilla luciferase was included as a control for transfection efficiency. Cells were

lysed in 1x Lysis Buffer (Promega) and firefly and Renilla luciferase activities were measured with the Promega DualGlo luciferase system. Different quantities of Tat plasmid were initially transfected to determine the linear range of activity. 100ng of pcDNA4-HA-Ub plasmid was used in ubiquitin mutant overexpression activity assays.

Identification of ubiquitinated Tat lysines by MS: HEK 293T cells were seeded in a 15cm cell culture dish and transfected with 2.5ug of Tat-expressing plasmid by PolyJet. After 48 hours, a denaturing STREP IP was performed on the lysate. The eluate was run on a gel for SDS-PAGE, and a band corresponding to ubiquitinated Tat (~25 kDa) was excised from the gel and digested with trypsin in gel as previously reported (65,66). Samples were analyzed on a Thermo Scientific Orbitrap Elite MS system. Peptides were then separated by an organic gradient from 5% to 30% ACN in 0.1% formic acid. Data were analyzed by the Protein Prospector analysis suite (67). Data were searched against the HIV Tat protein sequence (Uniprot accession P04608) concatenated with a HIV Tat protein sequence that was randomized. Fully tryptic peptides only were considered, with up to 2 missed cleavage sites. A static modification for carbamidomethyl cysteine was specified as well as variable modifications for oxidation of methionine, acetylation of the protein N-terminus, and Gly-Gly modification of lysines to capture tryptic ubiquitin remnant peptides. Spectra of interest were manually validated.

HIV spreading infection in Jurkat cells: The pNL4-3 proviral vector was used to generate all viruses. Tat mutants were cloned into the Nef locus of the virus with engineered SacII/XbaI sites. A codon-swapped variant of Tat was used to generate all mutant alleles to prevent recombination with the endogenous Δtat sequence. Wild type

and mutant Tat viruses were generated by transfection of pNL4-3 constructs into HEK 293T cells for 48 hours. Supernatant was collected, and viruses were purified and titered by p24 ELISA. For the spreading infection, Jurkat cells were seeded in a 96 well plate. Next day, cells were infected with wild type and mutant Tat viruses by spinoculation at an MOI of 0.5. A day 0 time point was taken after spinoculation. Supernatant was removed every 2-3 days and lysed 1:1 with PBS-2% Triton X-100. Jurkat cells were split 1:1 when reaching high cell density. The supernatant p24 was analyzed by ELISA.

Affinity purification and Western blotting: HEK 293T or HeLa cells at 60% confluence were transfected with STREP-tagged or FLAG-tagged bait plasmid using PolyJet. After 40 hours, cells were lysed in Lysis Buffer (50mM Tris-HCl pH 8.0, 150mM NaCl, 2mM $MgCl_2$, 0.5% Triton X-100, 0.2% Na-deoxycholate, Roche protease inhibitor cocktail) for 30 minutes at 4°C. Lysate was centrifuged and added to Streptactin Superflow resin or magnetic M2-FLAG resin and rotated for 3 hours at 4°C. The resin was washed 4x with WASH buffer (50mM Tris-HCl pH 8.0, 150mM NaCl, 2mM $MgCl_2$, 0.5% Triton X-100). Washed resin was eluted with 1x STREP elution buffer or FLAG elution buffer (200 µg/mL 3x FLAG peptide, diluted in WASH buffer) at room temperature for 30 minutes. The eluate was mixed 1:1 with 2x Laemmli sample buffer (BioRad), boiled for 5 minutes, and loaded on a SDS-PAGE gel. Proteins were detected with specific antibodies.

Author Contributions

Conceptualization, Tyler B Faust and Alan D Frankel; Investigation, T.B.F., Gwendolyn M. Jang, Yang Li, Shumin Yang, and Jeffrey R. Johnson; Writing-Original

Draft, T.B.F. and A.D.F.; Writing-Review & Editing, T.B.F., G.M.J., and A.D.F.; Visualization, T.B.F., G.M.J., Y.L., and A.D.F.; Supervision, A.D.F.; Funding Acquisition, A.D.F. and Nevan J. Krogan.

Acknowledgements

We thank Dr. J.J. Miranda and members of the Frankel laboratory for useful input on the manuscript. This work was supported by an N.I.H. grant (P50GM088250) to A.D.F. and N.J.K.

References

1. Wei P, Garber ME, Fang SM, Fischer WH, Jones KA. 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. 1998;92(4):451–62.
2. Feinberg MB, Baltimore D, Frankel AD. The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation. Proc Natl Acad Sci U S A [Internet]. 1991;88(9):4045–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51590&tool=pmcentrez&rendertype=abstract
3. Peterlin BM, Price DH. Controlling the Elongation Phase of Transcription with P-TEFb. Mol Cell. 2006;23(3):297–305.
4. Fuda NJ, Ardehali MB, Lis JT. Defining mechanisms that regulate RNA polymerase II transcription in vivo. Nature [Internet]. 2009;461(7261):186–92. Available from: papers2://publication/uuid/AE0EDB89-38AA-46A1-A269-84AAB14C5E24\nhttp://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19741698&retmode=ref&cmd=prlinks\npapers2://publication/doi/10.1038/nature08449
5. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet [Internet]. Nature Publishing Group; 2012;13(9):613–26. Available from: http://dx.doi.org/10.1038/nrg3207
6. Sikorski TW, Buratowski S. The basal initiation machinery: beyond the general transcription factors. Curr Opin Cell Biol. 2009;21(3):344–51.

7. Calnan BJ, Biancalana S, Hudson D, Frankel AD. Analysis of arginine-rich peptides from the HIV Tat protein reveals unusual features of RNA-protein recognition. Genes Dev. 1991;5(2):201–10.
8. Nguyen VT, Kiss T, Michels AA, Bensaude O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. Nature [Internet]. 2001;414(6861):322–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11713533\nhttp://dx.doi.org/10.1038/35104581
9. Yang Z, Zhu Q, Luo K, Zhou Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature [Internet]. 2001;414(6861):317–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11713532\nhttp://dx.doi.org/10.1038/35104575
10. Michels A a, Fraldi A, Li Q, Adamson TE, Bonnet F, Nguyen VT, et al. Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. EMBO J. 2004;23(13):2608–19.
11. D'Orso I, Frankel AD. RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. Nat Struct Mol Biol. 2010;17(7):815–21.
12. McNamara R, McCann J, Gudipaty SA, D'Orso I. Transcription factors mediate the enzymatic disassembly of promoter-bound 7sk snrnp to locally recruit p-tefb for transcription elongation. Cell Rep. 2013;5(5):1256–68.
13. D'Orso I, Jang GM, Pastuszak AW, Faust TB, Quezada E, Booth DS, et al.

Transition step during assembly of HIV Tat:P-TEFb transcription complexes and transfer to TAR RNA. Mol Cell Biol [Internet]. 2012;32(23):4780–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23007159\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3497596

14. Ji X, Zhou Y, Pandit S, Huang J, Li H, Lin CY, et al. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. Cell [Internet]. Elsevier Inc.; 2013;153(4):855–68. Available from: http://dx.doi.org/10.1016/j.cell.2013.04.028
15. McNamara RP, Reeder JE, McMillan EA, Bacon CW, McCann JL, D'Orso I. KAP1 Recruitment of the 7SK snRNP Complex to Promoters Enables Transcription Elongation by RNA Polymerase II. Mol Cell. 2016;61(1):39–53.
16. D'Orso I, Frankel AD. Tat acetylation modulates assembly of a viral-host RNA-protein transcription complex. Proc Natl Acad Sci U S A. 2009;106(9):3101–6.
17. Kiernan RE, Vanhulle C, Schiltz L, Adam E, Xiao H, Maudoux F, et al. HIV-1 tat transcriptional activity is regulated by acetylation. EMBO J [Internet]. 1999;18(21):6106–18. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1171675&tool=pmcentrez&rendertype=abstract
18. Xie B, Invernizzi CF, Richard S, Wainberg MA. 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 [Internet]. 2007;81(8):4226–34. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1866113&tool=pmcentrez&rendertype=abstract

19. Ali I, Ramage H, Boehm D, Dirk LMA, Sakane N, Hanada K, et al. The HIV-1 Tat protein is monomethylated at lysine-71 by the lysine methyltransferase KMT7. J Biol Chem [Internet]. 2016;1–19. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27235396
20. Brès V, Kiernan RE, Linares LK, Chable-Bessia C, Plechakova O, Tréand C, et al. A non-proteolytic role for ubiquitin in Tat-mediated transactivation of the HIV-1 promoter. Nat Cell Biol. 2003;5(8):754–61.
21. Metzger MB, Pruneda JN, Klevit RE, Weissman AM. RING-type E3 ligases: Master manipulators of E2 ubiquitin-conjugating enzymes and ubiquitination. Biochim Biophys Acta - Mol Cell Res [Internet]. Elsevier B.V.; 2014;1843(1):47–60. Available from: http://dx.doi.org/10.1016/j.bbamcr.2013.05.026
22. Kulathu Y, Komander D. Atypical ubiquitylation — the unexplored world of polyubiquitin beyond Lys48 and Lys63 linkages. Nat Rev Mol Cell Biol [Internet]. Nature Publishing Group; 2012;13(8):508–23. Available from: http://dx.doi.org/10.1038/nrm3394
23. Ohtake F, Saeki Y, Ishido S, Kanno J, Tanaka K. The K48-K63 Branched Ubiquitin Chain Regulates NF- k B Signaling. Mol Cell [Internet]. Elsevier Inc.; 2016;64(2):1–16. Available from: http://dx.doi.org/10.1016/j.molcel.2016.09.014
24. Salghetti SE, Caudy AA, Chenoweth JG, Tansey WP. Regulation of Transcriptional Activation Domain Function by. 2011;293(5535):1651–3.
25. Molinari E, Gilman M, Natesan S. Proteasome-mediated degradation of

transcriptional activators correlates with activation domain potency in vivo. EMBO J. 1999;18(22):6439–47.

26. Hammond-Martel I, Yu H, Affar EB. Roles of ubiquitin signaling in transcription regulation. Cell Signal [Internet]. Elsevier Inc.; 2012;24(2):410–21. Available from: http://dx.doi.org/10.1016/j.cellsig.2011.10.009

27. Haupt Y, Maya R, Kazaz A, Oren M. Mdm2 promotes the rapid degradation of p53. Nature [Internet]. 1997;387(6630):296–9. Available from: http://www.nature.com/doifinder/10.1038/246170a0\nhttp://www.ncbi.nlm.nih.gov/pubmed/9153395

28. Dornan D, Wertz I, Shimizu H, Arnott D, Frantz GD, Dowd P, et al. The ubiquitin ligase COP1 is a critical negative regulator of p53. Nature [Internet]. 2004;429(6987):86–92. Available from: http://www.nature.com/nature/journal/v429/n6987/full/nature02514.html\npapers2://publication/doi/10.1038/nature02514

29. Kim SY, Herbst A, Tworkowski KA, Salghetti SE, Tansey WP. Skp2 regulates Myc protein stability and activity. Mol Cell. 2003;11(5):1177–88.

30. Jaenicke LA, von Eyss B, Carstensen A, Wolf E, Xu W, Greifenberg AK, et al. Ubiquitin-Dependent Turnover of MYC Antagonizes MYC/PAF1C Complex Accumulation to Drive Transcriptional Elongation. Mol Cell. 2016;61(1):54–67.

31. Fu X, Yucer N, Liu S, Li M, Yi P, Mu J-J, et al. RFWD3-Mdm2 ubiquitin ligase complex positively regulates p53 stability in response to DNA damage. Proc Natl Acad Sci U S A. 2010;107(10):4579–84.

32. Popov N, Schülein C, Jaenicke L a, Eilers M. Ubiquitylation of the amino terminus

of Myc by SCF(β-TrCP) antagonizes SCF(Fbw7)-mediated turnover. Nat Cell Biol. 2010;12(10):973–81.

33. Li M. Mono- Versus Polyubiquitination: Differential Control of p53 Fate by Mdm2. Science (80-) [Internet]. 2003;302(5652):1972–5. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1091362

34. Le Cam L, Linares LK, Paul C, Julien E, Lacroix M, Hatchi E, et al. E4F1 Is an Atypical Ubiquitin Ligase that Modulates p53 Effector Functions Independently of Degradation. Cell. 2006;127(4):775–88.

35. Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global landscape of HIV-human protein complexes. Nature [Internet]. 2012;481(7381):365–70. Available from: http://dx.doi.org/10.1038/nature10719

36. Sobhian B, Laguette N, Yatim A, Nakamura M, Levy Y, Kiernan R, et al. HIV-1 Tat Assembles a Multifunctional Transcription Elongation Complex and Stably Associates with the 7SK snRNP. Mol Cell [Internet]. Elsevier Ltd; 2010;38(3):439–51. Available from: http://dx.doi.org/10.1016/j.molcel.2010.04.012

37. Jin X, Jin HR, Jung HS, Lee SJ, Lee JH, Lee JJ. An atypical E3 ligase zinc finger protein 91 stabilizes and activates NF-κB-inducing kinase via Lys63-linked ubiquitination. J Biol Chem. 2010;285(40):30539–47.

38. Lignitto L, Carlucci A, Sepe M, Stefan E, Cuomo O, Nisticò R, et al. Control of PKA stability and signalling by the RING ligase praja2. Nat Cell Biol [Internet]. Nature Publishing Group; 2011;13(4):412–22. Available from: http://dx.doi.org/10.1038/ncb2209

39. Lignitto L, Arcella A, Sepe M, Rinaldi L, Delle Donne R, Gallo A, et al. Proteolysis

of MOB1 by the ubiquitin ligase praja2 attenuates Hippo signalling and supports glioblastoma growth. Nat Commun [Internet]. 2013;4(May):1822. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3674242&tool=pmcentrez&rendertype=abstract

40. Mashtalir N, Daou S, Barbour H, Sen NN, Gagnon J, Hammond-Martel I, et al. Autodeubiquitination protects the tumor suppressor BAP1 from cytoplasmic sequestration mediated by the atypical ubiquitin ligase UBE2O. Mol Cell [Internet]. Elsevier Inc.; 2014;54(3):392–406. Available from: http://dx.doi.org/10.1016/j.molcel.2014.03.002

41. Zhang X, Zhang J, Bauer A, Zhang L, Selinger DW, Lu CX, et al. Fine-tuning BMP7 signalling in adipogenesis by UBE2O/E2-230K-mediated monoubiquitination of SMAD6. EMBO J [Internet]. Nature Publishing Group; 2013;32(7):996–1007. Available from: http://emboj.embopress.org/content/32/7/996.abstract

42. Lee J, Zhou P. DCAFs, the Missing Link of the CUL4-DDB1 Ubiquitin Ligase. Mol Cell. 2007;26(6):775–80.

43. Kainulainen M, Habjan M, Hubel P, Busch L, Lau S, Colinge J, et al. Virulence factor NSs of rift valley fever virus recruits the F-box protein FBXO3 to degrade subunit p62 of general transcription factor TFIIH. J Virol [Internet]. 2014;88(6):3464–73. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3957945&tool=pmcentrez&rendertype=abstract

44. Harper S, Gratton HE, Cornaciu I, Oberer M, Scott DJ, Emsley J, et al. Structure

and catalytic regulatory function of ubiquitin specific protease 11 N-terminal and ubiquitin-like domains. Biochemistry. 2014;53(18):2966–78.

45. Rinaldi L, Delle Donne R, Sepe M, Porpora M, Garbi C, Chiuso F, et al. praja2 regulates KSR1 stability and mitogenic signaling. Cell Death Dis [Internet]. Nature Publishing Group; 2016;7(5):e2230. Available from: http://www.nature.com/doifinder/10.1038/cddis.2016.109

46. Sakamaki J-I, Fu A, Reeks C, Baird S, Depatie C, Al Azzabi M, et al. Role of the SIK2-p35-PJA2 complex in pancreatic β-cell functional compensation. Nat Cell Biol [Internet]. 2014;16(3):234–44. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24561619

47. Tahirov TH, Babayeva ND, Varzavand K, Cooper JJ, Sedore SC, Price DH. Crystal structure of HIV-1 Tat complexed with human P-TEFb.pdf. Nature [Internet]. Nature Publishing Group; 2010;465(7299):747–51. Available from: http://dx.doi.org/10.1038/nature09131

48. Lata S, Ali A, Sood V, Raja R, Banerjea AC. HIV-1 Rev downregulates Tat expression and viral replication via modulation of NAD(P)H:quinine oxidoreductase 1 (NQO1). Nat Commun [Internet]. Nature Publishing Group; 2015;6:7244. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26059226

49. Ali A, Banerjea AC. Curcumin inhibits HIV-1 by promoting Tat protein degradation. Sci Rep [Internet]. Nature Publishing Group; 2016;6:27539. Available from: http://www.nature.com/articles/srep27539

50. Xu P, Duong DM, Seyfried NT, Cheng D, Xie Y, Robert J, et al. Quantitative Proteomics Reveals the Function of Unconventional Ubiquitin Chains in

Proteasomal Degradation. Cell [Internet]. Elsevier Ltd; 2009;137(1):133–45. Available from: http://dx.doi.org/10.1016/j.cell.2009.01.041

51. Yu Z, Chen T, Li X, Yang M, Tang S, Zhu X, et al. Lys29-linkage of ASK1 by Skp1-Cullin 1-Fbxo21 ubiquitin ligase complex is required for antiviral innate response. Elife [Internet]. 2016;5:e14087. Available from: https://elifesciences.org/content/5/e14087v1

52. Mattiroli F, Sixma TK. Lysine-targeting specificity in ubiquitin and ubiquitin-like modification pathways. Nat Struct Mol Biol [Internet]. Nature Publishing Group; 2014;21(4):308–16. Available from: http://www.nature.com/nsmb/journal/v21/n4/full/nsmb.2792.html\nhttp://www.nature.com/nsmb/journal/v21/n4/pdf/nsmb.2792.pdf

53. Grönroos E, Hellman U, Heldin CH, Ericsson J. Control of Smad7 stability by competition between acetylation and ubiquitination. Mol Cell. 2002;10(3):483–93.

54. Siliciano RF, Greene WC. HIV latency. Cold Spring Harb Perspect Med. 2011;1(1).

55. Razooky BS, Pai A, Aull K, Rouzine IM, Weinberger LS. A hardwired HIV latency program. Cell [Internet]. Elsevier Inc.; 2015;160(5):990–1001. Available from: http://dx.doi.org/10.1016/j.cell.2015.02.009

56. Fernandes JD, Booth DS, Frankel AD. A structurally plastic ribonucleoprotein complex mediates post-transcriptional gene regulation in HIV-1. Wiley Interdiscip Rev RNA [Internet]. 2016;7(August). Available from: http://doi.wiley.com/10.1002/wrna.1342

57. Mevissen TET, Hospenthal MK, Geurink PP, Elliott PR, Akutsu M, Arnaudo N, et

al. OTU deubiquitinases reveal mechanisms of linkage specificity and enable ubiquitin chain restriction analysis. Cell [Internet]. The Authors; 2013;154(1):169–84. Available from: http://dx.doi.org/10.1016/j.cell.2013.05.046

58. Barboric M, Zhang F, Besenicar M, Plemenitas A, Peterlin BM. Ubiquitylation of Cdk9 by Skp2 facilitates optimal Tat transactivation. J Virol. 2005;79(17):11135–41.

59. Calnan BJ, Tidor B, Biancalana S, Hudson D, Frankel AD, Series N, et al. Arginine-Mediated RNA Recognition : The Arginine Fork. 2007;252(5009):1167–71.

60. Yu X, Yu Y, Liu B, Luo K, Kong W, Mao P, et al. Induction of APOBEC3G ubiquitination and degradation by an HIV-1 Vif-Cul5-SCF complex. Science. 2003;302(5647):1056–60.

61. Jäger S, Kim DY, Hultquist JF, Shindo K, LaRue RS, Kwon E, et al. Vif hijacks CBF-β to degrade APOBEC3G and promote HIV-1 infection. Nature [Internet]. 2012;481(7381):371–5. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3310910&tool=pmcentrez&rendertype=abstract

62. Margottin F, Bour SP, Durand H, Selig L, Benichou S, Richard V, et al. A Novel Human WD Protein, h-βTrCP, that Interacts with HIV-1 Vpu Connects CD4 to the ER Degradation Pathway through an F-Box Motif. Mol Cell [Internet]. 1998;1(4):565–74. Available from: http://www.sciencedirect.com/science/article/pii/S1097276500800568

63. Malim MH, Emerman M. HIV-1 Accessory Proteins-Ensuring Viral Survival in a

Hostile Environment. Cell Host Microbe. 2008;3(6):388–98.

64. Gottwein E, Jäger S, Habermann A, Kräusslich H-G. Cumulative mutations of ubiquitin acceptor sites in human immunodeficiency virus type 1 gag cause a late budding defect. J Virol [Internet]. 2006;80(13):6267–75. Available from: http://dx.doi.org/10.1128/JVI.02177-05\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1488962&tool=pmcentrez&rendertype=abstract

65. Hellman U, Wernstedt C, Góñez J, Heldin CH. Improvement of an “In-Gel” digestion procedure for the micropreparation of internal protein fragments for amino acid sequencing. Vol. 224, Analytical biochemistry. 1995. p. 451–5.

66. Rosenfeld J, Capdevielle J, Guillemot JC, Ferrara P. In-gel digestion of proteins for internal sequence analysis after one- or two-dimensional gel electrophoresis. Anal Biochem. 1992;203(1):173–9.

67. Clauser KR, Baker P, Burlingame a L. Role of accurate mass measurements (+-10ppm) in protein identiifcation strategies employing MS or Ms MS and database searching. Anal Chem. 1999;71(14):2871–82.

Entrez Gene Id	NCBI gene symbol	siRNA Target Sequence	Product Id	Product Name	Gene Description
9867	PJA2	CTGGCAAGTTCAGATACTAAA	SI00108325	Hs_PJA2_1	praja ring finger 2, E3 ubiquitin protein ligase
9867	PJA2	TTGGCACATATGAGGCTTAAA	SI00108332	Hs_PJA2_2	praja ring finger 2, E3 ubiquitin protein ligase
9867	PJA2	ATCCACCATACTCAAGAGTTA	SI00108339	Hs_PJA2_3	praja ring finger 2, E3 ubiquitin protein ligase
9867	PJA2	TAAGGTTAGTAAAGCATACAA	SI00108346	Hs_PJA2_4	praja ring finger 2, E3 ubiquitin protein ligase
1025	CDK9	TAGGGACATGAAGGCTGCTAA	SI00605066	Hs_CDK9_5	cyclin-dependent kinase 9
5496	PPM1G	CAGGACCTGAGGACTCAACTA	SI02658684	Hs_PPM1G_6	protein phosphatase, Mg2+/Mn2+ dependent, 1G
5496	PPM1G	CCAGAGGATGAAGTAGAACTA	SI02658691	Hs_PPM1G_7	protein phosphatase, Mg2+/Mn2+ dependent, 1G
10768	AHCYL1	CAGGGTGGTAAAGCTAAATGA	SI00090328	Hs_AHCYL1_1	adenosylhomocysteinase-like 1
10768	AHCYL1	AAACAGTTGTATCGTATGCAA	SI00090335	Hs_AHCYL1_2	adenosylhomocysteinase-like 1
10768	AHCYL1	CTGATAGAACTCTATAATGCA	SI00090342	Hs_AHCYL1_3	adenosylhomocysteinase-like 1
10768	AHCYL1	CCCACTTGGATTTATAGTATA	SI00090349	Hs_AHCYL1_4	adenosylhomocysteinase-like 1
26273	FBXO3	CTGACGATTATCGATGTTCAT	SI00097783	Hs_FBXO3_1	F-box protein 3
26273	FBXO3	CTGCCAGAACTTAGCTCTGTA	SI00097797	Hs_FBXO3_3	F-box protein 3
26273	FBXO3	CAGGTCGGGTATATGAATACA	SI03071999	Hs_FBXO3_5	F-box protein 3
26273	FBXO3	CTGGTTTACCTCTTATGTCAA	SI03100202	Hs_FBXO3_6	F-box protein 3
63893	UBE2O	TCCCGGGACCATTCCATGGAA	SI04153863	Hs_UBE2O_2	ubiquitin-conjugating enzyme E2O
63893	UBE2O	AACCAGATCATCCTGAAGCTA	SI04228679	Hs_UBE2O_3	ubiquitin-conjugating enzyme E2O
63893	UBE2O	TGGCACGGTGATCGACGTCAA	SI04280017	Hs_UBE2O_4	ubiquitin-conjugating enzyme E2O
63893	UBE2O	TGGCCTGTCGACCACGTTATA	SI04346699	Hs_UBE2O_5	ubiquitin-conjugating enzyme E2O
54876	DCAF16	CGGGAGATCTGCCGAGGATAA	SI04162718	Hs_C4orf30_1	DDB1 and CUL4 associated factor 16
54876	DCAF16	TCTGACTAAATCAGCCTATAA	SI04163901	Hs_C4orf30_2	DDB1 and CUL4 associated factor 16
54876	DCAF16	CCAGATGTTCTTGACAGCTCA	SI04284903	Hs_C4orf30_3	DDB1 and CUL4 associated factor 16
54876	DCAF16	ATGGGCCACAAGCATACTTTA	SI04373012	Hs_C4orf30_4	DDB1 and CUL4 associated factor 16
8237	USP11	CTGCGTCGGGTACGTGATGAA	SI02780813	Hs_USP11_5	ubiquitin specific peptidase 11
8237	USP11	ACCGATTCTATTGGCCTAGTA	SI02781156	Hs_USP11_6	ubiquitin specific peptidase 11
23382	AHCYL2	CAGCGGCTTCATATACAGATA	SI00110467	Hs_KIAA0828_1	adenosylhomocysteinase-like 2
23382	AHCYL2	CAGGTTCTAAACTCTATATAT	SI00110474	Hs_KIAA0828_2	adenosylhomocysteinase-like 2
23382	AHCYL2	CAACATCTATTCCACTCTCAA	SI00110481	Hs_KIAA0828_3	adenosylhomocysteinase-like 2
23382	AHCYL2	AGGCAACACCAGCAAATTATA	SI00110488	Hs_KIAA0828_4	adenosylhomocysteinase-like 2
23411	SIRT1	CAAGCGATGTTTGATATTGAA	SI00098434	Hs_SIRT1_1	sirtuin 1
23411	SIRT1	CAGGATTATTGTATTTACGTT	SI00098441	Hs_SIRT1_2	sirtuin 1
23411	SIRT1	TTGGGTCTTCCCTCAAAGTAA	SI00098448	Hs_SIRT1_3	sirtuin 1
23411	SIRT1	AGCCATCGGAATGTTAAATTA	SI04954068	Hs_SIRT1_8	sirtuin 1
10360	NPM3	CACCCGCTCCTTCACCTTTAA	SI00661122	Hs_NPM3_2	nucleophosmin/nucleoplasmin 3
10360	NPM3	CCAGATTGTTACGATGAGCAA	SI00661129	Hs_NPM3_3	nucleophosmin/nucleoplasmin 3
10360	NPM3	CCCGGTCACTATGGACAGTTT	SI04142824	Hs_NPM3_5	nucleophosmin/nucleoplasmin 3
10360	NPM3	AAAGACGAGTGTAATGTGGTA	SI04278176	Hs_NPM3_6	nucleophosmin/nucleoplasmin 3
4676	NAP1L4	AAGAAAGTATGCAGCGCTATA	SI03127383	Hs_NAP1L4_5	nucleosome assembly protein 1-like 4
4676	NAP1L4	CTGCGGGTCACCTCATATTTA	SI03208758	Hs_NAP1L4_7	nucleosome assembly protein 1-like 4
4676	NAP1L4	GTGGACATGCTGAGTGAATTA	SI04194309	Hs_NAP1L4_8	nucleosome assembly protein 1-like 4
4676	NAP1L4	TTCGCTGTGGCTCGTCCTTAA	SI04288228	Hs_NAP1L4_9	nucleosome assembly protein 1-like 4
80829	ZFP91	CTGCGGCACACTTATCTTCAA	SI05109223	Hs_ZFP91_9	ZFP91 zinc finger protein
80829	ZFP91	CAGCTCTTAAAGTGAGGGTTA	SI05109230	Hs_ZFP91_10	ZFP91 zinc finger protein
80829	ZFP91	GTGGATTTACTTGTCGACAAA	SI05150971	Hs_ZFP91_14	ZFP91 zinc finger protein
80829	ZFP91	CCGCGACTCCTATGCATAGAA	SI05150985	Hs_ZFP91_16	ZFP91 zinc finger protein
3189	HNRNPH3	AGCGACCGGGACCATATGATA	SI03144295	Hs_HNRPH3_5	heterogeneous nuclear ribonucleoprotein H3 (2H9)
3189	HNRNPH3	ATCGCCTGACAGGCATTTAAA	SI04141858	Hs_HNRPH3_7	heterogeneous nuclear ribonucleoprotein H3 (2H9)
3189	HNRNPH3	AACATTGACGATGGACTACCA	SI04306736	Hs_HNRPH3_8	heterogeneous nuclear ribonucleoprotein H3 (2H9)
163033	ZNF579	CCGCGAAGCCTTCGCCACCAA	SI00776979	Hs_ZNF579_1	zinc finger protein 579
163033	ZNF579	CCCGGTTTGCAGAATGGAGGA	SI00776986	Hs_ZNF579_2	zinc finger protein 579
163033	ZNF579	CCCAACATGTGTCTTAAGGCA	SI00777000	Hs_ZNF579_4	zinc finger protein 579
23524	SRRM2	CGCCACCTAAACAGAAATCTA	SI00733460	Hs_SRRM2_4	serine/arginine repetitive matrix 2
23524	SRRM2	CTCGACGAAGATCCCGGTCAA	SI04216898	Hs_SRRM2_6	serine/arginine repetitive matrix 2
23524	SRRM2	CTCGATCATCTCCGGAGCTAA	SI04173995	Hs_SRRM2_5	serine/arginine repetitive matrix 2
25940	FAM98A	TTGGAGTCGTTGGAAGATCTA	SI04209373	Hs_FAM98A_1	family with sequence similarity 98, member A
25940	FAM98A	AAACGTTTGGATGTCACTGTA	SI04225263	Hs_FAM98A_2	family with sequence similarity 98, member A
25940	FAM98A	CAGCCGAAACGTTCAGTCTTA	SI04315241	Hs_FAM98A_3	family with sequence similarity 98, member A
904	CCNT1	AGGCTTTGAACTAACAATTGA	SI02625707	Hs_CCNT1_5	cyclin T1

Table 1

Table Legend

Table 1**:** Sequences of siRNAs used in targeted RNAi screen.

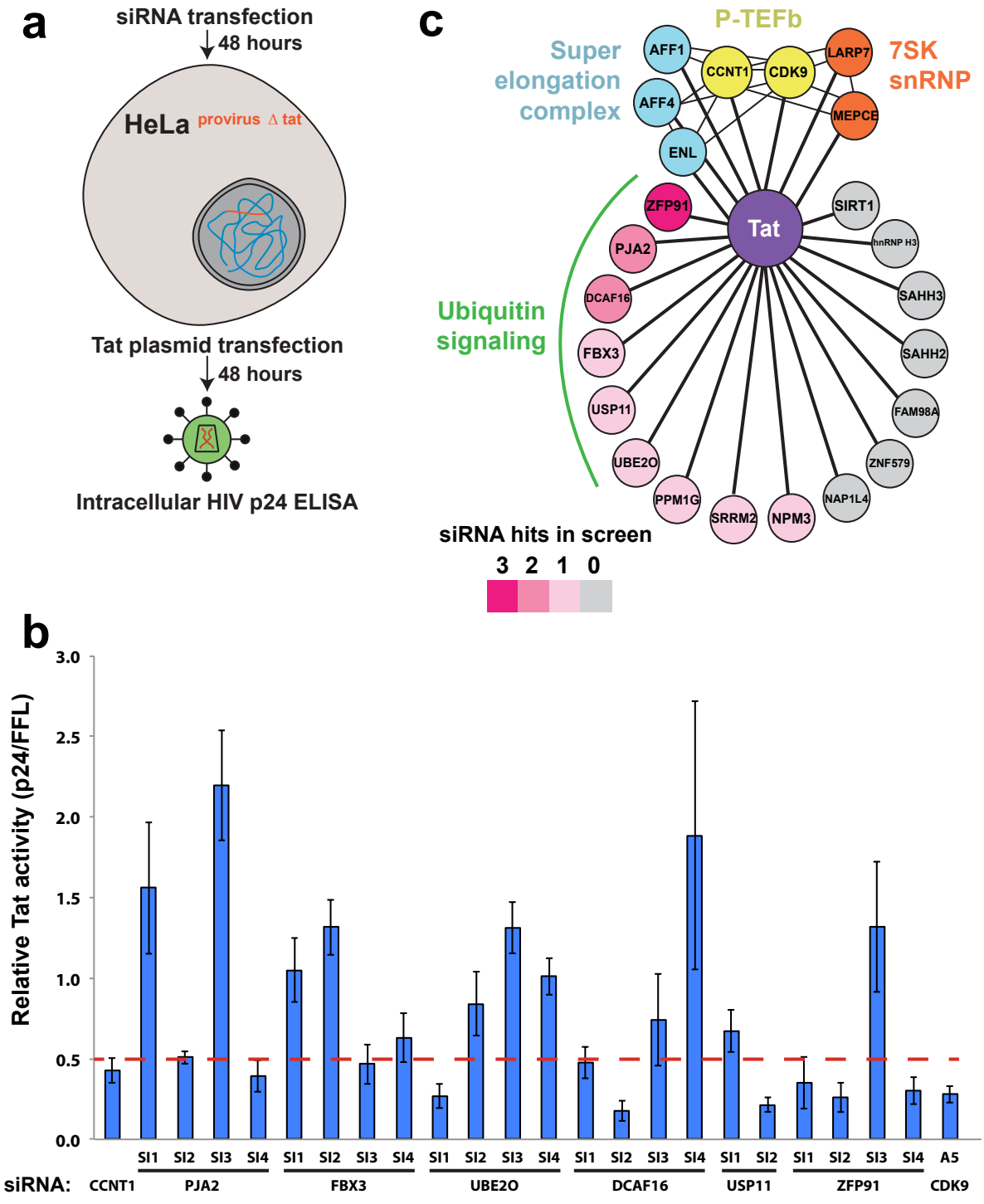

a
siRNA transfection
48 hours
HeLa provirus Δ tat
Tat plasmid transfection
48 hours
Intracellular HIV p24 ELISA
c
P-TEFb
Super elongation complex
7SK snRNP
AFF1
CCNT1
CDK9
LARP7
AFF4
MEPCE
ENL
ZFP91
Tat
SIRT1
hnRNP H3
PJA2
DCAF16
SAHH3
Ubiquitin signaling
FBX3
SAHH2
USP11
FAM98A
UBE2O
ZNF579
PPM1G
SRRM2
NPM3
NAP1L4
siRNA hits in screen
3 2 1 0
b
Relative Tat activity (p24/FFL)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
SI1 SI2 SI3 SI4 SI1 SI2 SI3 SI4 SI1 SI2 SI3 SI4 SI1 SI2 SI3 SI4 SI1 SI2 SI1 SI2 SI3 SI4 A5
siRNA: CCNT1 PJA2 FBX3 UBE2O DCAF16 USP11 ZFP91 CDK9

Fig. 1

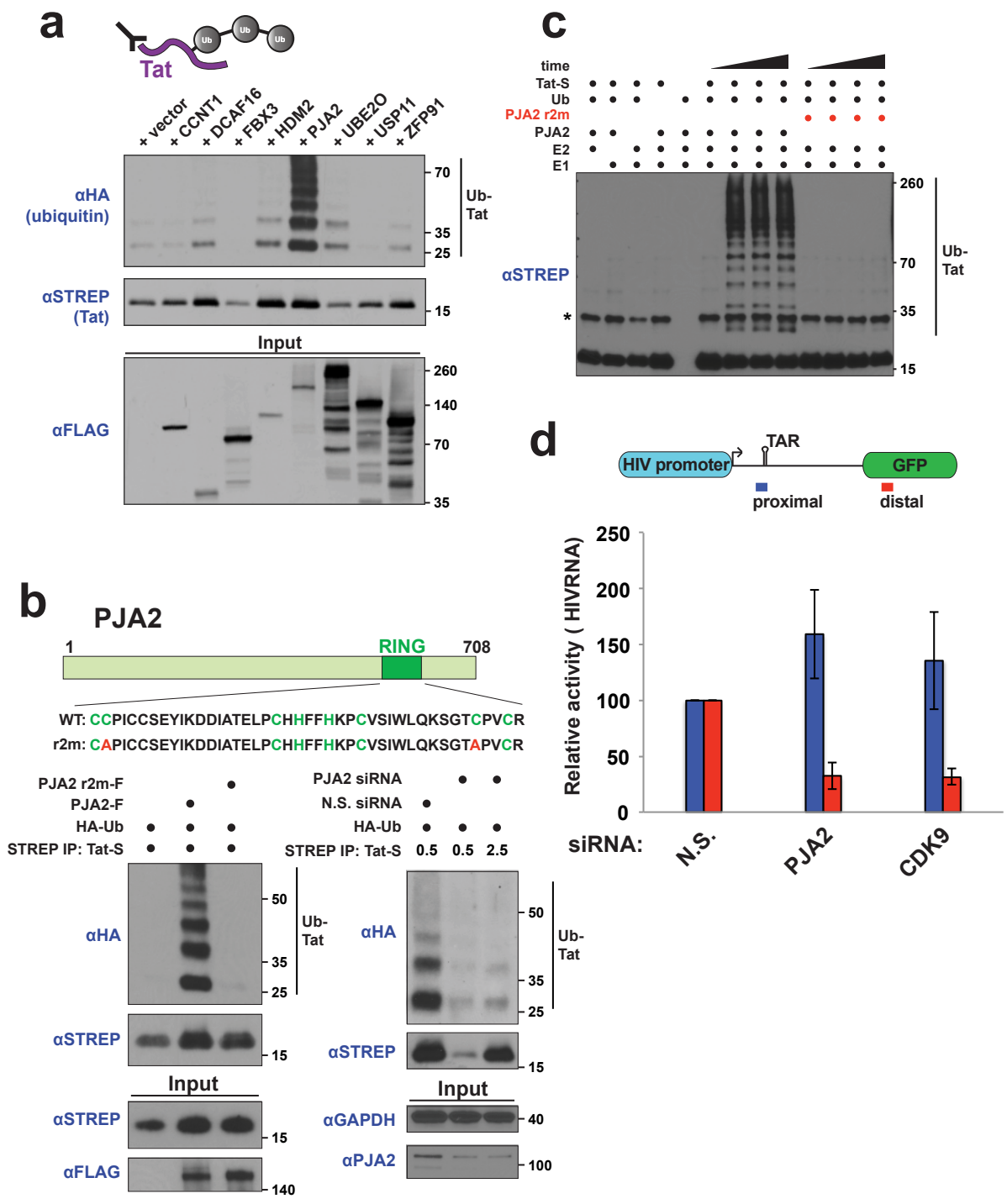

a
Tat
Ub
Ub
Ub
+ vector
+ CCNT1
+ DCAF16
+ FBX3
+ HDM2
+ PJA2
+ UBE2O
+ USP11
+ ZFP91
αHA (ubiquitin)
70
35
25
Ub-Tat
αSTREP (Tat)
15
Input
αFLAG
260
140
70
35
c
time
Tat-S
Ub
PJA2 r2m
PJA2
E2
E1
αSTREP
*
260
70
35
15
Ub-Tat
d
HIV promoter
TAR
GFP
proximal
distal
Relative activity (HIVRNA)
250
200
150
100
50
0
siRNA:
N.S.
PJA2
CDK9
b
PJA2
1
RING
708
WT: CCPICCSEYIKDDIATELPCHHFFHKPCVSIWLQKSGTCPVCR
r2m: CAPICCSEYIKDDIATELPCHHFFHKPCVSIWLQKSGTAPVCR
PJA2 r2m-F
PJA2-F
HA-Ub
STREP IP: Tat-S
αHA
50
35
25
Ub-Tat
αSTREP
15
Input
αSTREP
15
αFLAG
140
PJA2 siRNA
N.S. siRNA
HA-Ub
STREP IP: Tat-S
0.5
0.5
2.5
αHA
50
35
25
Ub-Tat
αSTREP
15
Input
αGAPDH
40
αPJA2
100

Fig. 2

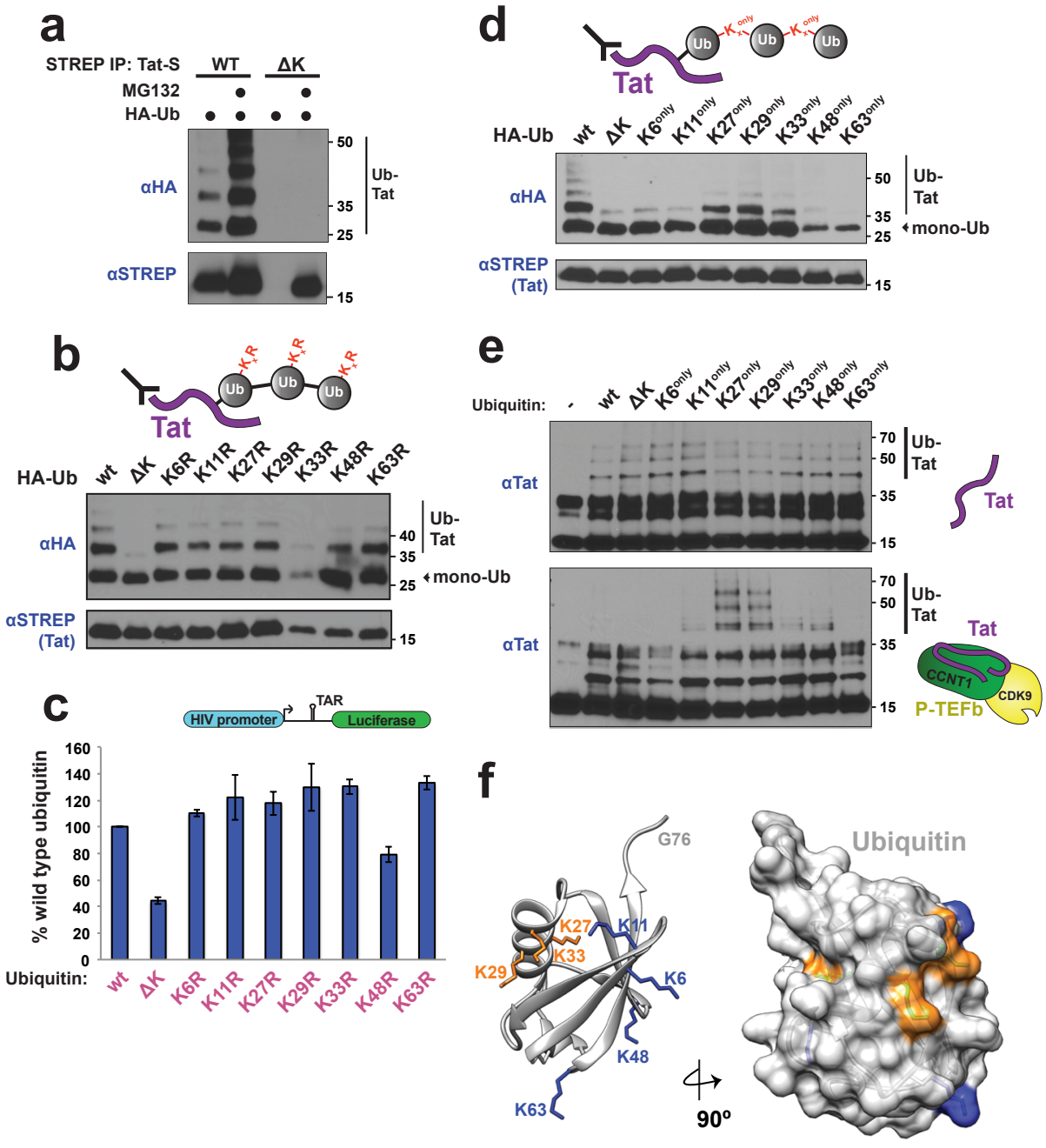

a
STREP IP: Tat-S
WT
ΔK
MG132
HA-Ub
αHA
αSTREP
Ub-Tat
50
35
25
15
b
Tat
HA-Ub
wt
ΔK
K6R
K11R
K27R
K29R
K33R
K48R
K63R
αHA
αSTREP (Tat)
Ub-Tat
mono-Ub
40
35
25
15
c
HIV promoter
TAR
Luciferase
% wild type ubiquitin
160
140
120
100
80
60
40
20
0
Ubiquitin:
wt
ΔK
K6R
K11R
K27R
K29R
K33R
K48R
K63R
d
Tat
Ub
HA-Ub
wt
ΔK
K6only
K11only
K27only
K29only
K33only
K48only
K63only
αHA
αSTREP (Tat)
Ub-Tat
mono-Ub
50
35
25
15
e
Ubiquitin:
αTat
Ub-Tat
Tat
70
50
35
15
CCNT1
CDK9
P-TEFb
f
G76
K27
K11
K29
K33
K6
K48
K63
90°
Ubiquitin

Fig. 3

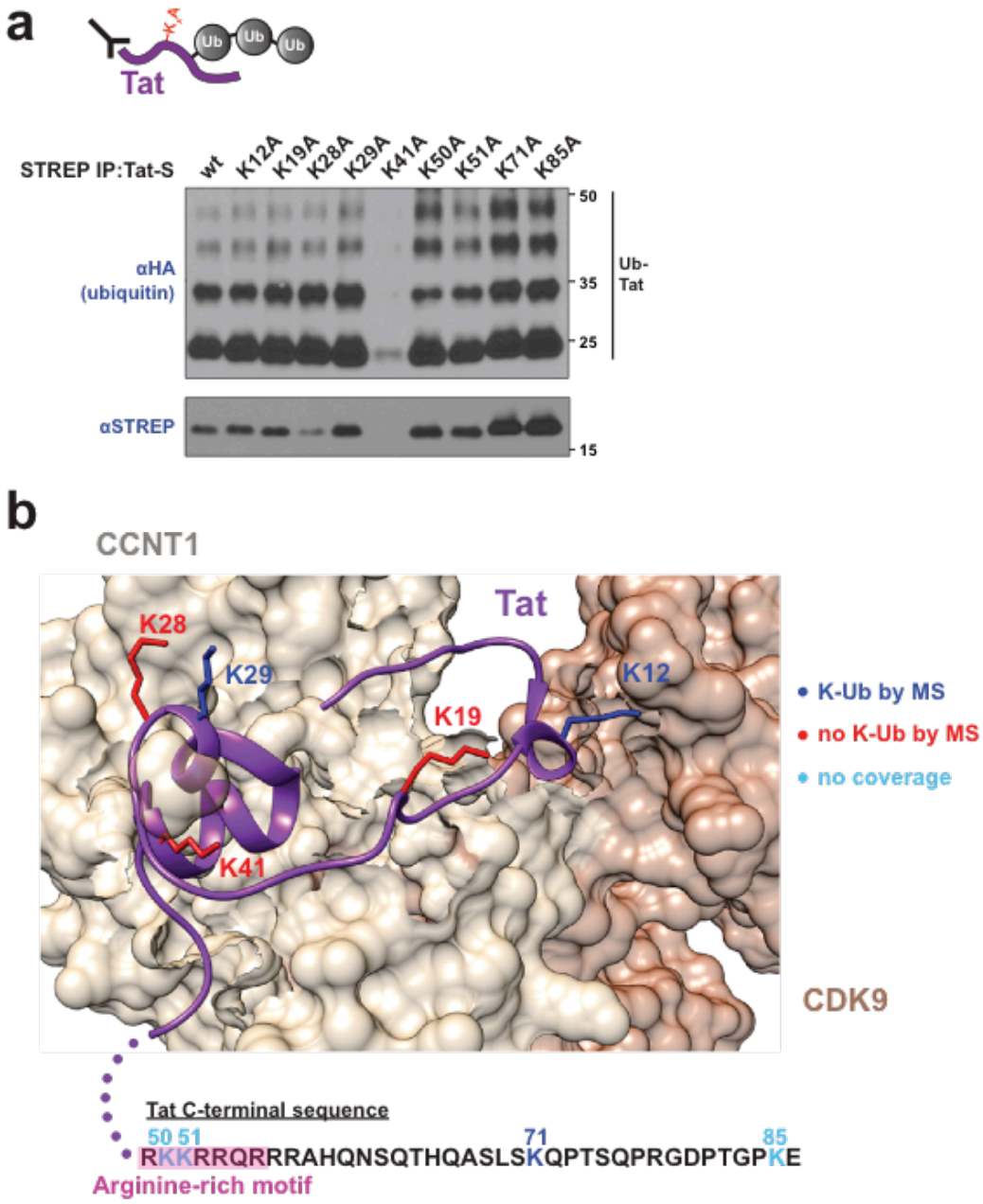

a
Ub
Ub
Ub
Tat
STREP IP:Tat-S
wt
K12A
K19A
K28A
K29A
K41A
K50A
K51A
K71A
K85A
αHA
(ubiquitin)
αSTREP
50
35
25
15
Ub-
Tat
b
CCNT1
Tat
K28
K29
K12
K19
K41
CDK9
K-Ub by MS
no K-Ub by MS
no coverage
Tat C-terminal sequence
50 51
71
85
RKKRRQRRRAHQNSQTHQASLSKQPTSQPRGDPTGPKE
Arginine-rich motif

Fig. 4

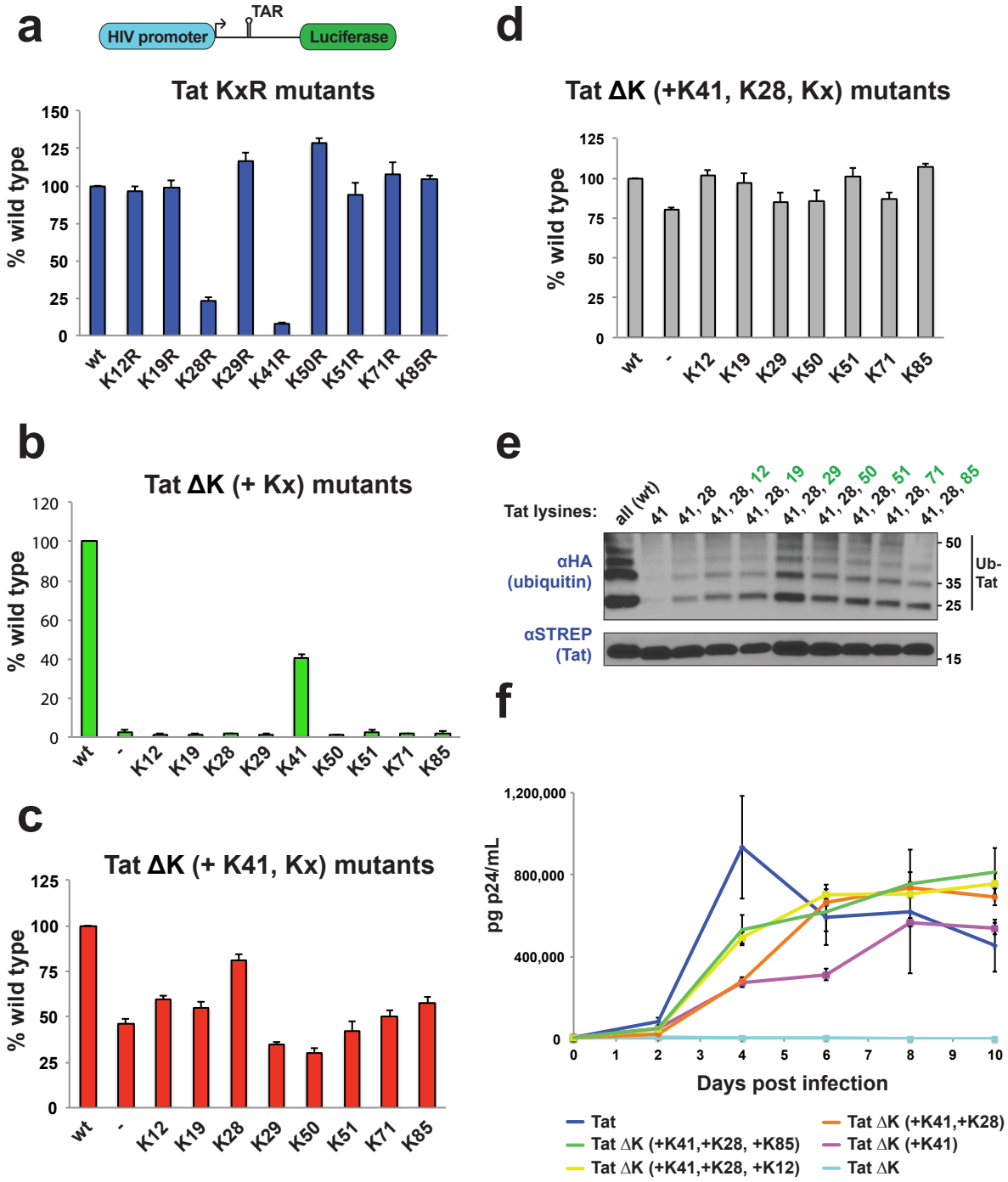

a
HIV promoter
TAR
Luciferase
Tat KxR mutants
% wild type
150
125
100
75
50
25
0
wt
K12R
K19R
K28R
K29R
K41R
K50R
K51R
K71R
K85R
b
Tat ΔK (+ Kx) mutants
% wild type
120
100
80
60
40
20
0
wt
-
K12
K19
K28
K29
K41
K50
K51
K71
K85
c
Tat ΔK (+ K41, Kx) mutants
% wild type
125
100
75
50
25
0
wt
-
K12
K19
K28
K29
K50
K51
K71
K85
d
Tat ΔK (+K41, K28, Kx) mutants
% wild type
125
100
75
50
25
0
wt
-
K12
K19
K29
K50
K51
K71
K85
e
Tat lysines:
all (wt)
41
41, 28
41, 28, 12
41, 28, 19
41, 28, 29
41, 28, 50
41, 28, 51
41, 28, 71
41, 28, 85
αHA
(ubiquitin)
αSTREP
(Tat)
50
35
25
15
Ub-
Tat
f
pg p24/mL
1,200,000
800,000
400,000
0
0
1
2
3
4
5
6
7
8
9
10
Days post infection
Tat
Tat ΔK (+K41,+K28, +K85)
Tat ΔK (+K41,+K28, +K12)
Tat ΔK (+K41,+K28)
Tat ΔK (+K41)
Tat ΔK

Fig. 5

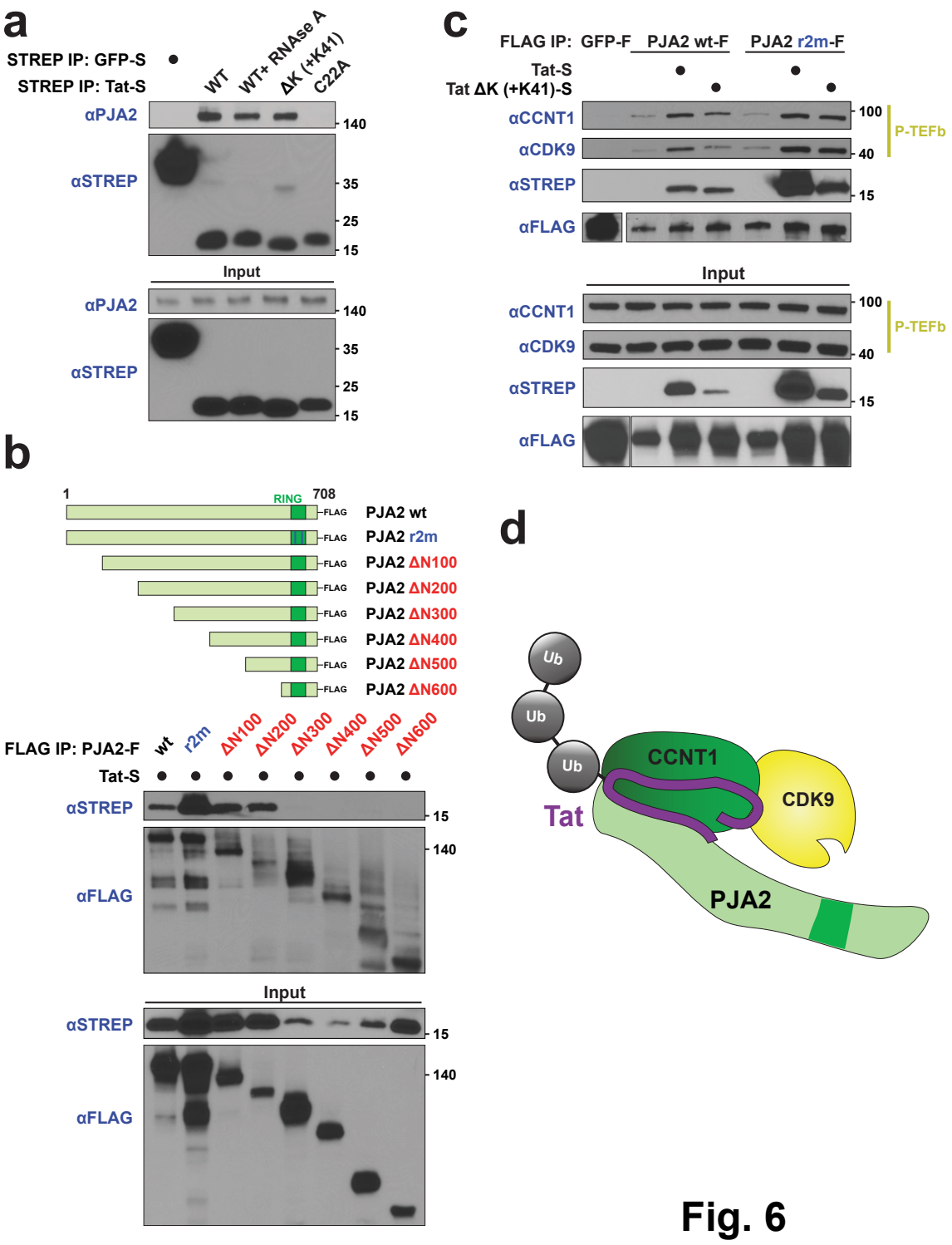

a
STREP IP: GFP-S
STREP IP: Tat-S
WT
WT+ RNAse A
ΔK (+K41)
C22A
αPJA2
αSTREP
140
35
25
15
Input
αPJA2
αSTREP
140
35
25
15
b
1
RING
708
FLAG
PJA2 wt
PJA2 r2m
PJA2 ΔN100
PJA2 ΔN200
PJA2 ΔN300
PJA2 ΔN400
PJA2 ΔN500
PJA2 ΔN600
FLAG IP: PJA2-F
wt
r2m
ΔN100
ΔN200
ΔN300
ΔN400
ΔN500
ΔN600
Tat-S
αSTREP
αFLAG
15
140
Input
αSTREP
αFLAG
15
140
c
FLAG IP: GFP-F
PJA2 wt-F
PJA2 r2m-F
Tat-S
Tat ΔK (+K41)-S
αCCNT1
αCDK9
αSTREP
αFLAG
100
40
15
P-TEFb
Input
αCCNT1
αCDK9
αSTREP
αFLAG
100
40
15
P-TEFb
d
Ub
Ub
Ub
CCNT1
CDK9
Tat
PJA2

Fig. 6

Figure Legends

Fig. 1. Targeted RNAi screen identifies multiple proteins involved in ubiquitin signaling as critical for Tat function. (a) (top) Schematic representation of the screen. A HeLa cell line containing a single, full-length integrated provirus (in red) was transfected with individual siRNAs, followed by transfection with a Tat-expressing plasmid. The readout of activity was the inhibition of Tat-dependent production of the HIV p24 protein. (b) Normalized, relative Tat activity ([$p24^{host\ RNAi}$/$FFL^{host\ RNAi}$] / [$p24^{N.S.\ RNAi}$/$FFL^{N.S.\ RNAi}$]) for CCNT1, CDK9, and the ubiquitin signaling proteins (PJA2, FBX3, UBE2O, DCAF16, USP11, and ZFP91). Data are represented as the mean -/+ S.E.M. of three biological replicates for independent siRNA transfections. The red dotted line represents the activity cutoff for the screen. (c) Network representation of sixteen novel high confidence Tat interactors from our prior large-scale proteomics study (35). Color-coding reflects the number of siRNAs for each factor that caused at least a 2-fold decrease in Tat activity. Also shown are known Tat interactors that are part of P-TEFb, 7SK RNP, and super elongation complexes.

Fig. 2. PJA2 specifically ubiquitinates Tat *in vivo* and *in vitro* (a) Tat *in vivo* denaturing ubiquitination assay. HEK 293T cells were transfected with Tat-STREP, HA-Ub, and six FLAG-tagged ubiquitin signaling proteins that showed activity in the RNAi screen, followed by a denaturing STREP IP of Tat (illustrated by schematic above). Tat ubiquitination was detected by anti-HA Western blotting and the ubiquitinated species are indicated. (b) (top) Domain organization of the PJA2 protein. The zinc-binding residues of the RING-H2 domain are highlighted in green and the alanine point mutations that constitute the r2m mutant are shown in red. (bottom left) *In vivo*

ubiquitination assay of Tat-STREP (referred to as Tat-S) co-expressed with FLAG-tagged PJA2 wild type or r2m. (bottom right) PJA2 knockdown inhibits Tat ubiquitination. The HeLa$^{provirus\Delta tat}$ cell line was transfected with N.S. or PJA2 siRNA 2 (from screen), followed by transfection with HA-Ub and 0.5ug Tat-S for N.S. RNAi cells or 0.5ug and 2.5ug Tat-S for PJA2 RNAi cells. *In vivo* ubiquitination of Tat is indicated by the HA blot. Western blots with anti-PJA2 show the extent of protein knockdown (~75%). (c) *In vitro* ubiquitination of Tat. Tat-STREP purified from *E. coli* was incubated with recombinant E1, E2, wild type PJA2 or r2m mutant, ubiquitin, and ATP. All Tat species were detected by STREP Western blotting. A time course of ubiquitination was performed for reactions containing all components of the *in vitro* ubiquitination system (t = 0, 15, 30, 60 minutes). Reactions on the left, which were missing one component of the full *in vitro* system, were incubated for 60 minutes. The band labeled with an asterisk is not a ubiquitinated species. (d) HIV RNA elongation assay with PJA2 RNAi. Purified RNA was amplified with proximal and distal primer pairs as depicted in the schematic above from HeLa$^{provirus\Delta tat}$ cells transfected with N.S. siRNA, PJA2 siRNA 2, or CDK9 siRNA. Data are presented as the mean -/+ S.E.M. of relative Tat activity (indicated siRNA/ N.S. siRNA) of at least 3 biological replicates. All qPCR values were normalized to RPLP0 mRNA.

Fig. 3. Tat is modified by PJA2 with non-degradative, atypical polyubiquitin chains. (a) *In vivo* ubiquitination assay of STREP-tagged (Tat-S) wild type Tat or Tat ΔK -/+ MG132 in HEK 293T cells. The anti-HA Western detects ubiquitinated Tat species. (b) *In vivo* ubiquitination of Tat co-transfected with wild type ubiquitin or with ΔK or individual Lys-to-Arg mutants. A schematic representation of the assay and

ubiquitin mutants is shown above. Monoubiquitinated and multi-ubiquitinated species are indicated. (c) (top) Schematic of the integrated luciferase reporter in the HeLa$^{LTR:FFL}$ cell line. (bottom) Transcriptional activity of Tat after co-transfection with the ubiquitin variants used in panel (b). Data are represented as mean + S.E.M. of three biological replicates and are normalized to co-transfected CMV-firefly luciferase. (d) *In vivo* ubiquitination of Tat co-transfected with wild type ubiquitin or the ΔK or individual Lys^{only} ubiquitin mutants. A schematic representation of the assay and ubiquitin mutants is shown above. (e) *In vitro* ubiquitination of Tat alone (top) or Tat in complex with P-TEFb (bottom) by PJA2. Wild type ubiquitin or the ΔK or individual Lys^{only} ubiquitin mutants were used and reactions were incubated for 2 hours. (f) (left) Ribbon representation of ubiquitin (PDB 1ubq). The lysine residues used to generate polyubiquitin chains on Tat are highlighted in orange (Lys27, Lys29, and Lys33) and the remaining ubiquitin lysines are shown in blue. (right) Surface representation of the ubiquitin structure rotated by 90° with identical coloring.

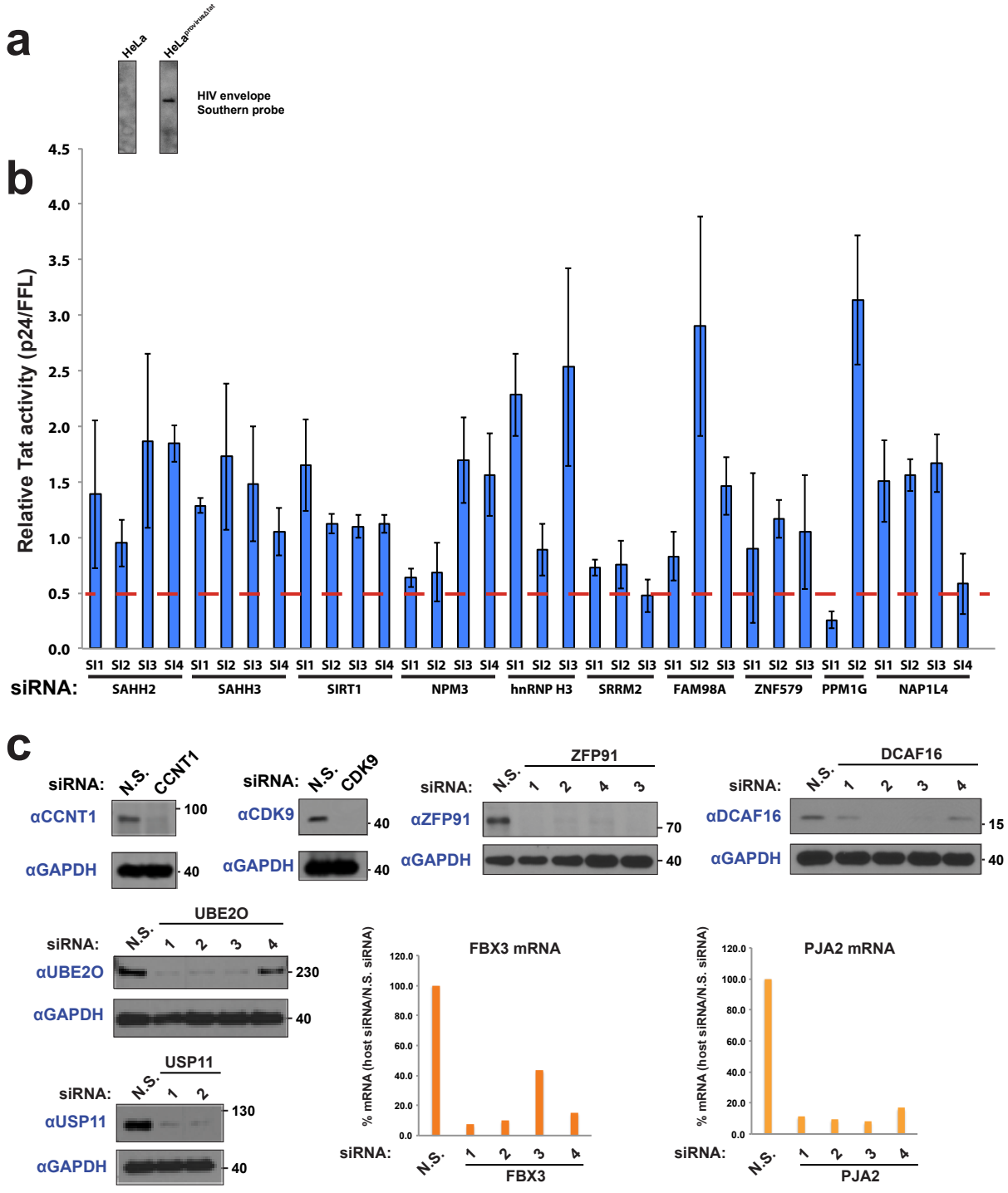
Fig. 4. Multiple lysines in Tat can function as acceptor sites for ubiquitination. (a) *In vivo* ubiquitination of STREP-tagged (Tat-S) wild type Tat or individual Lys-to-Ala mutants in HEK 293T cells. A schematic representation of the mutants is shown above. The anti-HA Western detects ubiquitinated Tat species. (b) Ubiquitinated lysines identified by MS are highlighted in blue in the Tat-P-TEFb crystal structure (PDB 3mi9) (top) or in the C-terminal Tat sequence (bottom). The color coding for modified, unmodified, or undetected lysine peptides is shown on the right.

Fig. 5. Activity and ubiquitination of minimal lysine Tat mutants. (a-d) Transcriptional reporter activities in the HeLa$^{LTR:FFL}$ reporter cell line of Tat KxR mutants

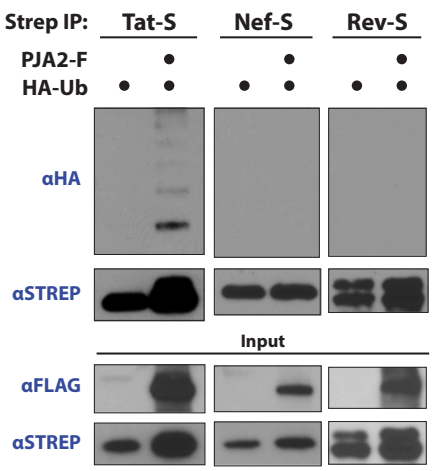
where individual lysines were mutated to arginine (a), Tat ΔK (+Kx) where individual lysines were added back to a lysine-free background (b), Tat ΔK (+K41,+Kx) where individual lysines were added back to a K41-only background (c), and Tat ΔK (+K41,+K28,+Kx) where individual lysines were added back to a K41, K28 background (d). Data are represented as mean + S.E.M. of at least three biological replicates. All values were normalized to co-transfected CMV-firefly luciferase. The schematic of the integrated LTR-FFL reporter is also represented in (a). (e) *In vivo* ubiquitination of STREP-tagged wild type Tat or various minimal lysine mutants in HeLa cells, as indicated. (f) HIV spreading replication assay in Jurkat cells with either wild type Tat or minimal lysine mutants cloned into the Nef locus of pNL4-3. Data are plotted as the average at each time point -/+ S.E.M. (n=4).

Fig. 6. Tat enhances the interaction of PJA2 with P-TEFb. (a) STREP-tagged GFP, wild type Tat, or Tat mutants were transfected into HEK 293T cells and a STREP IP was used to detect an interaction with endogenous PJA2 by Western blotting with anti-PJA2 antibody. (b) (top) Domain representation of FLAG-tagged PJA2 mutants used to map the Tat interacting region. (bottom) FLAG-tagged wild type PJA2, r2m, or truncation mutants were co-transfected with Tat-STREP (Tat-S) into HEK 293T cells. A FLAG IP of PJA2 variants was used to assess the interaction with Tat. (c) FLAG-tagged GFP, PJA2 wild type or r2m expression plasmids were co-transfected with vector, Tat-STREP or Tat ΔK (+K41)-STREP into HEK 293T cells. A FLAG IP elution was probed with anti-STREP, anti-CYCT1, and anti-CDK9 to detect PJA2 interactions with Tat and P-TEFb. FLAG blots are from the same exposure but are cropped for presentation. (d) Model of the PJA2-P-TEFb-Tat complex. PJA2 ubiquitinates Tat in the context of P-

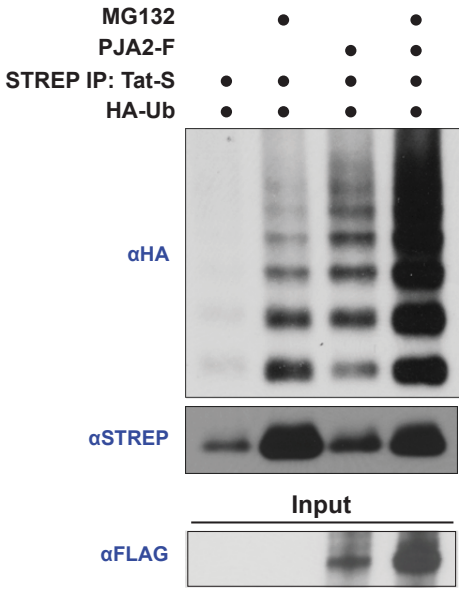
TEFb predominantly with Lys27 and Lys29 polyubiquitin chains. Tat also enhances the interaction between PJA2 and P-TEFb.


a
HeLa
HeLa pro-virus Tat
HIV envelope
Southern probe
b
Relative Tat activity (p24/FFL)
siRNA:
SAHH2
SAHH3
SIRT1
NPM3
hnRNP H3
SRRM2
FAM98A
ZNF579
PPM1G
NAP1L4
c
siRNA: N.S. CCNT1
αCCNT1
αGAPDH
siRNA: N.S. CDK9
αCDK9
αGAPDH
ZFP91
siRNA: N.S. 1 2 4 3
αZFP91
αGAPDH
DCAF16
siRNA: N.S. 1 2 3 4
αDCAF16
αGAPDH
UBE2O
siRNA: N.S. 1 2 3 4
αUBE2O
αGAPDH
USP11
siRNA: N.S. 1 2
αUSP11
αGAPDH
FBX3 mRNA
% mRNA (host siRNA/N.S. siRNA)
siRNA: N.S. 1 2 3 4
FBX3
PJA2 mRNA
% mRNA (host siRNA/N.S. siRNA)
siRNA: N.S. 1 2 3 4
PJA2

Sup. Fig. 1


Strep IP:
Tat-S
Nef-S
Rev-S
PJA2-F
HA-Ub
αHA
αSTREP
Input
αFLAG
αSTREP

Sup. Fig. 2



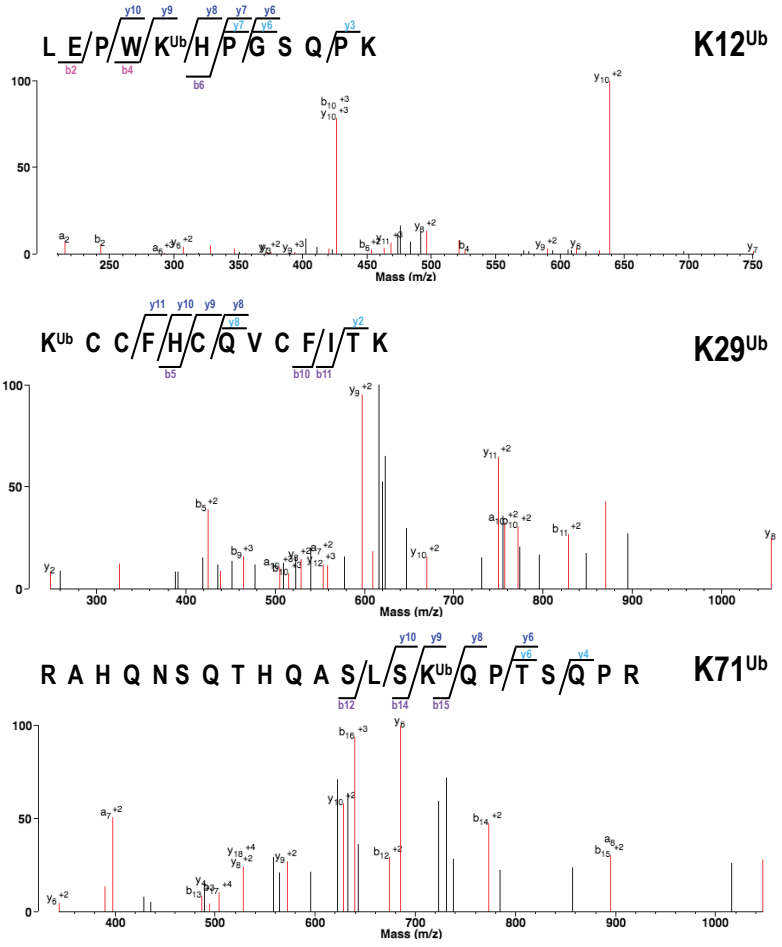
Sup. Fig. 3

a

1 MEPVDPRLEPWKHPGSQPKTACTNCYCKKCCFHCQVCFITKALGISYG
49 RKKRRQRRRAHQNSQTHQASLSKQPTSQPRGDPTGPKE

- K-Ub by MS
- no K-Ub by MS
- no coverage

b



Sup. Fig. 4

Supplemental Figure Legends

Supplementary Fig. 1. Additional relative Tat activity data from the targeted RNAi screen. (a) Southern blot of parental HeLa or HeLa$^{provirus\Delta tat}$ cell line with DIG-labeled HIV envelope. HeLa$^{provirus\Delta tat}$ contains a single integrated provirus that is responsive to Tat transfection and was used for the siRNA functional screen. (b) Normalized, relative Tat activity ([p24$^{host\ RNAi}$/FFL$^{host\ RNAi}$] / [p24$^{N.S.\ RNAi}$/FFL$^{N.S.\ RNAi}$]) for Tat-interacting host factors (SAHH2, SAHH3, SIRT1, NPM3, hnRNP H3, SRRM2, FAM98A, ZNF579, PPM1G, and NAP1L4) not shown in Fig. 1. Data are represented as the mean -/+ S.E.M. of three biological replicates for independent siRNA transfections. The red dotted line represents activity cutoff for the screen. (c) Western blots or qPCR quantitations to demonstrate protein or RNA knockdown, respectively, for CCNT1, CDK9, or ubiquitin pathway proteins.

Supplementary Fig. 2. PJA2 specifically ubiquitinates the HIV Tat protein. HeLa$^{LTR:FFL}$ cells were transfected with HA-Ub, vector or PJA2-F, and STREP-tagged Tat, Nef, or Rev. A denaturing STREP IP was used to detect viral protein ubiquitination by HA Western blotting.

Supplementary Fig. 3. Tat protein levels are stabilized by proteasome inhibition. *In vivo* ubiquitination assay of Tat-STREP (Tat-S) co-transfected with vector or PJA2-F in HEK 293T cells, -/+ MG132 treatment.

Supplementary Fig. 4. Mass spectrometry identification of ubiquitinated Tat lysines. (a) 82.6% peptide coverage of the HXB2 Tat protein sequence. Underlined and bolded is the sequence not covered by any peptide for MS. Color coding reflects whether a lysine was modified with a ubiquitin di-glycine tryptic remnant fragment,

unmodified, or undetected. (b) Mass spectra for Tat peptides containing lysine-di-glycine ubiquitin remnant tryptic fragments for Lys12, Lys29, and Lys71. Identified y and b ions are labeled on each peptide sequence.

Chapter 4: The HIV-1 Tat protein hijacks a cytoplasmic ubiquitin ligase to remodel the 7SK snRNP for transcriptional activation

Summary

HIV-1 gene expression has provided an excellent model of transcription elongation control. The HIV-1 Tat protein hijacks P-TEFb to activate paused RNAP II at the viral promoter. Tat binds additional host proteins, but it is unclear if and how they regulate RNAP II elongation. Here we identify a cytoplasmic host protein complex containing two E3 ubiquitin ligases, ZFP91 and UBE2O, which is critical for Tat transcriptional activity. Components of this complex interact with the 7SK snRNP, a major fraction of which also resides in the cytoplasm bound to P-TEFb. Further, UBE2O ubiquitinates 7SK snRNP-bound HEXIM1, leading to a remodeling of this complex. Interestingly, Tat expression or DRB treatment also remodels the cytoplasmic 7SK snRNP, releasing P-TEFb from the inhibitory complex and promoting its import into the nucleus. These results support a unique model of global elongation control where non-degradative ubiquitination in the cytoplasm drives the nuclear re-localization of activated P-TEFb.

Keywords

Host-pathogen interactions, 7SK snRNP, non-degradative ubiquitination, UBE2O, HIV, transcription elongation, nuclear import

Introduction

The advent of genome-wide methods to interrogate transcription has shown that many genes in higher eukaryotes contain stably paused RNA polymerase II (RNAP II) molecules immediately downstream of transcription start sites (1–4). The release of RNAP II into productive elongation is controlled by the kinase positive transcription elongation factor b (P-TEFb), which is composed of cyclin T1 (CCNT1) and cyclin-

dependent kinase 9 (CDK9) subunits (5). P-TEFb hyper-phosphorylates the C-terminal domain of RNAP II, resulting in a processive polymerase (6). Numerous cellular proteins have been shown to interact with P-TEFb to relieve pausing, including bromodomain-containing protein 4 (BRD4) (7), Mediator (8), nuclear factor (NF)-κB (9), and c-Myc (1). The human immunodeficiency virus-1 (HIV-1) transcription factor, Tat, also interacts with P-TEFb to activate a paused polymerase at the viral promoter (10,11).

P-TEFb activity is controlled by sequestration in an inactive complex with the 7SK snRNP (12,13). In this complex, HEXIM1 inhibits the kinase activity of CDK9 in a 7SK snRNA-dependent manner (14), while the 7SK snRNA methyl phosphate capping enzyme (MEPCE) and La-related protein 7 (LARP7) stabilize the 7SK snRNA by adding a 5' methylphosphate cap or binding to the 3' poly(U) tail of the RNA, respectively (15). Multiple signaling pathways regulate the association of P-TEFb with the 7SK snRNP, tuning the required levels of active kinase in the cell (16,17). We have previously shown that 7SK snRNP-inhibited P-TEFb is recruited to the HIV-1 promoter (18), where Tat activates transcription elongation by displacing the inhibitory subunits through binding of the nascently transcribed transactivation response element (TAR) RNA. Supporting this model of activation, recent genome-wide studies have revealed that snRNP-bound P-TEFb occupies the majority of promoters that display RNAP II pausing (19,20). The splicing protein SC35 appears to act analogously to Tat by releasing the 7SK snRNP in a nascent RNA-binding step (20).

Here we report the identification of a protein complex, containing the Tat interacting factors UBE2O, ZFP91, and NAP1L4, which is critical for regulating Tat activity. We show that UBE2O, a hybrid E2-E3 ubiquitin ligase (21,22), robustly

ubiquitinates HEXIM1 of the 7SK snRNP leading to a loss in LARP7 interaction. Interestingly, UBE2O-driven HEXIM1 ubiquitination is largely cytoplasmic, which led to the discovery of a significant cytoplasmic pool of P-TEFb-7SK snRNP. Tat expression or treatment of cells with a CDK inhibitor, DRB, releases P-TEFb from this large reservoir with the 7SK snRNP, resulting in the nuclear import of the active kinase.

Results

A targeted RNAi screen identifies ZFP91 as a critical regulator of Tat activity

Our recent proteomic interaction study identified several high-confidence host binding partners for the HIV-1 Tat protein (23). While the contributions of the P-TEFb-based super elongation complex (SEC) (24–26) and 7SK snRNP (18,27) to regulating viral transcription has become better defined, the functions of additional Tat interactors are largely unexplored. To functionally characterize the potential new Tat co-factors uncovered in our proteomic screen, we had performed a targeted RNAi screen of sixteen high-confidence physical interactors in a functional assay for Tat activity (*unpublished data*). We used a HeLa cell line harboring a single, full-length HIV provirus deleted for the *tat* gene (HeLa$^{provirus\Delta tat}$), which allowed us to precisely control transcriptional activation by transfection of a Tat plasmid and to monitor Tat activity in a relatively natural proviral context by the production of intracellular HIV capsid protein (p24). Of the sixteen proteins screened, Zinc Finger Protein 91 (ZFP91) had the strongest phenotype and three independent siRNAs inhibited Tat function (Fig. 1A). ZFP91 has been previously shown to ubiquitinate NF-κB-inducing kinase via a Lys63 linkage to activate a non-canonical NF-κB pathway (28). To confirm that ZFP91 specifically regulates transcription elongation as expected for a bona fide Tat co-factor,

we measured promoter-proximal and -distal HIV RNA production and found that ZFP91 knockdown significantly reduced the Tat-dependent production of distal, but not proximal, transcripts (Fig. 1B). These data strongly implicate ZFP91 as an essential regulator of viral transcription elongation.

ZFP91 forms a complex with two other Tat interactors, UBE2O and NAP1L4

To better understand the role of ZFP91 in Tat function, we used affinity purification-mass spectrometry (AP-MS) to identify its cellular binding partners. Surprisingly, two other Tat interacting proteins, UBE2O and NAP1L4 (23), were found in ZFP91 complexes (Fig. 1C, Table S1). UBE2O, which also scored in the RNAi screen (Fig. 1A), is a hybrid E2-E3 ubiquitin ligase that can ubiquitinate several chromatin-associated proteins within their nuclear localization signals (NLS), resulting in cytoplasmic retention (21). NAP1L4 binds core and linker histones, likely functioning as a histone chaperone (29). Reciprocal IPs confirmed the interactions between these three proteins (Fig. S1) and, importantly, the association between UBE2O and NAP1L4 was seen with the endogenous proteins (Fig. 1D). The pairwise interactions between these three functionally important Tat factors suggested that they could assemble into a single complex. To corroborate this, a double IP of transiently transfected proteins showed that ZFP91, UBE2O, and NAP1L4 could be purified as one assembly, referred to as ZUN (Fig. 1E). In parallel, a double IP of the P-TEFb subunits, CDK9 and CCNT1, pulled down LARP7 of the 7SK snRNP (Fig. 1E) but, notably, did not enrich for UBE2O compared to that observed with NAP1L4-ZFP91, establishing that the 7SK snRNP and ZUN proteins are in distinct complexes. The separate nature of these complexes agrees with the absence of P-TEFb in ZFP91 AP-MS experiments (Fig. 1C). Altogether, these

data indicate that three Tat interactors important for transcription, two of which are E3 ubiquitin ligases, form a protein complex.

Tat recruits the UBE2O ubiquitin ligase to HEXIM1 of the 7SK snRNP

We were intrigued by the interaction of Tat with a P-TEFb-independent host complex, as Tat is thought to interact with the majority of its host partners, such as the SEC, indirectly through the P-TEFb interface. While the 7SK snRNP and ZUN proteins are in distinct complexes, the AP-MS data support interactions between these complexes, as ZFP91 pulled down LARP7 and MEPCE of the snRNP while HEXIM1 pulled down UBE2O and NAP1L4 (Fig. 1C, Table S1). The HEXIM1-UBE2O interaction was also identified in a recent effort to map the human interactome by systematic AP-MS of 2,594 bait proteins (30), further corroborating the cross talk of these two complexes. Therefore, while the snRNP and ZUN proteins are not natively in a large ensemble together, they can interact in the cell, potentially in a regulated manner. Given that ZUN proteins are important for Tat-dependent transcription, one possibility is that Tat reinforces the interactions between ZUN and snRNP complexes. In support of this model, Tat dramatically increased the association between endogenous HEXIM1 and transfected UBE2O (Fig. 1F). Overall, the data support endogenous interactions between ZUN and 7SK snRNP complexes and the ability of a viral transcription factor to strengthen these interactions by hijacking a ubiquitin ligase to a major transcriptional regulatory protein.

Tat uses a novel interaction surface to bind the 7SK snRNP and ZUN proteins

Tat buries an extensive interaction surface on the CCNT1 subunit of P-TEFb (31). Given that Tat potently recruits UBE2O to HEXIM1, yet UBE2O is in a P-TEFb-

independent complex (Fig. 1E), Tat likely uses an additional binding surface to achieve this recruitment. To explore whether a novel functional interface might exist, we mapped all transcriptionally important residues of Tat based on comprehensive alanine scanning (32) onto the Tat-P-TEFb crystal structure. As expected, several known regions were clearly defined, including the cysteine-rich zinc-binding clusters (Fig. 2A) and a stretch of hydrophobic residues (Phe38, Leu43, Gly44, and Ile45) buried in CCNT1 (Fig. 2B). Remarkably, a cluster of highly conserved, severe loss-of-function mutants (Glu2, Tyr26, Lys28, and Phe32) were surface-exposed and faced away from CCNT1 and CDK9 (Fig. 2C), suggesting a distinct interaction surface.

To directly test if these residues constituted a new binding surface for the ZUN proteins, we immunoprecipitated wild type and mutant Tat proteins from transiently transfected HEK 293T cells and examined the interactions with endogenous proteins. Indeed, Tat K28A poorly bound to ZFP91, UBE2O, and NAP1L4 (Fig. 2D, Fig. S2 for inputs). Surprisingly, this Lys mutant also poorly bound to MEPCE and LARP7 of the 7SK snRNP, suggesting that it is important for interactions with both the ZUN and snRNP complexes. The interaction of both of these snRNP proteins with Tat is entirely RNA-dependent (Fig. 2D, RNAse A treatment), implicating Lys28 in 7SK snRNA binding in addition to its role in modulating the TAR RNA interaction (33). Mutation of Phe32, which is adjacent to Lys28 in the structure, significantly increased binding of UBE2O. This might indicate that Phe32 is involved in a handoff from UBE2O to another protein, such that its mutation locks Tat in the UBE2O-bound state. Importantly, the P-TEFb interaction was unaffected by Lys28 or Phe32 mutations (Fig. 2D), unlike the mutation of Phe38 and Leu43 residues that are buried in CCNT1 and disrupt P-TEFb binding,

demonstrating high specificity for the ZUN interactions and corroborating that these residues comprise a novel docking surface. The SEC scaffold protein, AFF1, had an identical binding profile to CDK9 and CCNT1, supporting recent work that AFF1 is a ubiquitous P-TEFb binding partner (34). The activity and binding data all support a model in which Tat uses a buried hydrophobic patch (Phe38, Leu43) to anchor onto P-TEFb while using two spatially adjacent residues (Lys28, Phe32) to stabilize interactions with the 7SK snRNP and ZUN proteins (Fig. 2E). Thus, Tat encodes an essential hijacking function directed at a ubiquitin ligase and the 7SK snRNP.

UBE2O monoubiquitinates HEXIM1 at multiple sites *in vivo* and *in vitro*

The interactions between the ZFP91 and UBE2O ubiquitin ligases and the 7SK snRNP led us to investigate whether they could ubiquitinate any of the snRNP proteins. To test this, we pulled down various 7SK snRNP and P-TEFb subunits and, upon blotting for HA-tagged ubiquitin, found that UBE2O co-expression dramatically increased HEXIM1 and MEPCE ubiquitination but not LARP7, CCNT1, or CDK9 (Fig. 3A). In contrast, ZFP91 expression showed no effect on any protein. Because HEXIM1 is the key regulatory kinase inhibitor of the 7SK snRNP (14) and is known to be ubiquitinated (35), we focused our efforts on this modification. As anticipated, a catalytically inactive mutant of UBE2O in the ubiquitin-conjugating domain (C1040S or CD) did not ubiquitinate HEXIM1(Fig. S3A). Because proteins can be monoubiquitinated or polyubiquitinated, leading to very different functional outputs, and recent work has shown that UBE2O monoubiquitinates numerous proteins (21,22), we next investigated the nature of the linkage. Similar to previous studies, expression of a lysine-free ubiquitin mutant (Ub ΔK), which blocks the ability to form polyubiquitin chains, did not

affect UBE2O-dependent HEXIM1 ubiquitination *in vivo* (Fig. 3B), demonstrating that HEXIM1 is monoubiquitinated at multiple sites. We next reconstituted ubiquitination of HEXIM1 by UBE2O *in vitro* and found that, like our observations *in vivo*, wild type UBE2O but not the CD mutant robustly modified HEXIM1 (Fig. 3C). As also observed *in vivo*, multiple high molecular weight HEXIM1 species were generated using the Ub ΔK variant. These results demonstrate the unbiased identification of HEXIM1 as an endogenous substrate of UBE2O that becomes directly monoubiquitinated at multiple sites.

UBE2O ubiquitinates HEXIM1 in a non-degradative manner

Targeting proteins for proteasomal destruction is a well-known function of ubiquitination, but it is increasingly appreciated that this modification also serves important signaling functions, especially using mono-ubiquitin marks. Indeed, the UBE2O-dependent ubiquitination of HEXIM1 was insensitive to treatment with the proteasome inhibitor MG132 (Fig. 3D, Fig. S3C showing MG132 control), supporting a non-degradative mechanism. Furthermore, expression of Tat, which recruits UBE2O to HEXIM1 (Fig. 1F), also stimulated non-degradative HEXIM1 ubiquitination (Fig. 3D), an effect blocked by siRNA-mediated knockdown of UBE2O (Fig. 3E). UBE2O knockdown also prevented endogenous, Tat-independent HEXIM1 ubiquitination and did not increase HEXIM1 protein levels (Fig. 3E), offering further confirmation that HEXIM1 is a substrate for non-degradative UBE2O mono-ubiquitination, providing a critical function that can be exploited by Tat.

UBE2O ubiquitinates the NLS of a 7SK snRNP-bound HEXIM1

HEXIM1 ubiquitination resembles the modification of recently described UBE2O substrates (21). Specifically, UBE2O prefers substrate acceptor lysines contained within a bipartite NLS surrounding a patch of aliphatic hydrophobic amino acids, termed the "VLI patch", which is also found in UBE2O itself and results in auto-ubiquitination. Interestingly, HEXIM1 has a similar motif (Fig. 4A), with an NLS composed of two segments of 12 and 3 basic amino acids surrounding a stretch of mostly aromatic hydrophobic (W, Y) amino acids. To identify the sites of ubiquitination in HEXIM1 and to determine if the NLS is indeed a target, we first generated a fully lysine-free HEXIM1 mutant (ΔK) and observed no modification (Fig. 4B, last lane), demonstrating that ubiquitination is lysine-dependent. When all lysines within the N- and C-termini were mutated to arginines (CTΔK+NTΔK), leaving lysines intact only within the NLS, HEXIM1 was still ubiquitinated (Fig. 4B), consistent with acceptor residues being located in the NLS. Several additional mutants suggest that lysines within the C-terminus can also be ubiquitinated (see Fig. 4B legend), but we focused on the NLS modifications because of the previously defined connections to UBE2O. To more precisely map the acceptor sites, we used mass spectrometry (MS) and identified multiple ubiquitinated lysines in the NLS and C-terminus (Fig. S4B, Table S2), supporting our initial mapping (Fig. 4B). However, the HEXIM1 NLS is lysine and arginine-rich, resulting in many undetectable residues after digestion for MS. As an alternative approach to map acceptor sites in the NLS, we restored individual lysines within the lysine-free (ΔK) background and found that four (Lys150, Lys151, Lys152, and Lys159) recovered monoubiquitination, with Lys151 and Lys159 being the most robust (Fig. 4C). These four lysines are located

within the first segment of the bipartite NLS, suggesting that they are the most accessible for UBE2O recognition.

One intriguing possibility is that HEXIM1 is modified by UBE2O in the context of the 7SK snRNP, especially given the observation that Tat also stimulates HEXIM1 modification and is known to be in complex with the snRNP (18,23,26,27). Indeed, a HEXIM1 mutant deficient for 7SK RNA binding (150-156 Ala) (14,36) displayed a complete loss of UBE2O-dependent ubiquitination, in striking contrast to another HEXIM1 mutant deficient for P-TEFb binding that was still ubiquitinated (PYNT->PDND) (14) (Fig. 4D). Because the 150-156 Ala mutant also lacks some of the lysines that become ubiquitinated (Fig. 4C), we generated a HEXIM1 150-156 Arg mutant which retains RNA binding (35) but lacks the NLS acceptor lysines and found that it was ubiquitinated equally to wild-type HEXIM1 (Fig. 4D). Thus, the 7SK snRNA-bound form of HEXIM1 is recognized by UBE2O, a model further supported by the UBE2O-dependent ubiquitination of MEPCE, another 7SK snRNP protein (Fig. 3A). RNA-binding releases HEXIM1 from an auto-inhibited state (37), which likely establishes the proper conformation of HEXIM1 for UBE2O ubiquitination.

UBE2O influences HEXIM1 subcellular localization and remodels the 7SK snRNP leading to LARP7 eviction

Given the similarities of HEXIM1 ubiquitination to other UBE2O substrates, we next asked if modification results in cytoplasmic accumulation (21). The expression of UBE2O resulted in a 1.5 fold increase of HEXIM1 in the cytoplasm with a simultaneous 2.5 fold decrease in the nuclear pellet fraction, confirming that UBE2O specifically alters the subcellular localization of endogenous HEXIM1, with negligible changes to the

distribution of CDK9.(Fig. 5A). Consistent with the cytoplasmic accumulation of HEXIM1, nearly all of the UBE2O-dependent ubiquitination of HEXIM1 was also cytoplasmic (Fig. 5B), suggesting that ubiquitination of HEXIM1 in this fraction regulates its localization and may inhibit chromatin recruitment. Interestingly, the lysine-free HEXIM1 ΔK mutant, which is not ubiquitinated by UBE2O (Fig. 4B), accumulated 3.3-fold more in the chromatin-containing nuclear pellet compared to wild type HEXIM1 when expressed in HEK 293T cells (Fig. 5C), further supporting a role of HEXIM1 lysine modification in regulating its localization. It is also intriguing that such a significant fraction of unmodified HEXIM1 natively resides in the cytoplasm, where it seems likely to be ubiquitinated by UBE2O.

As UBE2O specifically targeted snRNP-bound HEXIM1 (Fig. 4D), we next considered whether ubiquitination might also alter protein-protein interactions within the 7SK snRNP (38,39). UBE2O expression did not affect the interaction of HEXIM1 with P-TEFb but showed an unexpected 90% decrease with LARP7, one of the snRNP subunits (Fig. 5D). Binding of another snRNP protein, MEPCE, and 7SK snRNA were unaffected, suggesting that HEXIM1 ubiquitination results in partial disassembly of the snRNP complex. It is worth noting that LARP7 was the only snRNP protein not ubiquitinated by UBE2O (Fig. 3A), which is consistent with its specific loss from the snRNP. Importantly, the UBE2O CD mutant did not cause loss of LARP7 (Fig. S5), confirming that ubiquitination is required. Tat expression similarly dissociated the LARP7-HEXIM1 interaction in HEK 293T cells (Fig. S6), suggesting that both UBE2O and Tat can remodel the 7SK snRNP in a similar manner. Finally, P-TEFb is ultimately liberated from wild type HEXIM1 at increasing Tat protein levels, however HEXIM1 ΔK

was refractory to complete Tat-dependent P-TEFb release (Fig. S7), supporting that HEXIM1 lysines are critical in the full remodeling of the 7SK snRNP by Tat.

Tat releases P-TEFb from cytoplasmic 7SK snRNP complexes for nuclear import

Given the large proportion of HEXIM1 in the cytoplasm (Fig. 5A, B), we wished to determine if this reflected the presence of the full 7SK snRNP complex. Indeed, a native IP of endogenous HEXIM1 from cytoplasmic and soluble nuclear fractions revealed robust RNA-dependent interactions between HEXIM1 and proteins of the P-TEFb-7SK snRNP complex in the cytoplasm (Fig. 6A). This agrees with the previous finding that a significant fraction of the “large-form” of P-TEFb, complexed to 7SK snRNP, is cytoplasmic (40). Immunofluorescence confirmed that a significant fraction of the CCNT1 subunit of P-TEFb was also localized to the cytoplasm (Fig. 6B). Further, the interaction of HEXIM1 with 7SK snRNA was detected in the cytoplasm (Fig. 6C), thereby placing ubiquitinated, 7SK snRNP-bound HEXIM1 in this compartment.

Newly translated Tat is cytoplasmic, especially during acute infection (41), and also shuttles between the nucleus and cytoplasm (42), prompting us to examine whether Tat might utilize this fraction of the 7SK snRNP for its function. Indeed, Tat expression in HeLa cells decreased the amount of CCNT1 and CDK9 in the cytoplasm by four-fold and showed a concomitant increase in the nuclear, chromatin-containing fraction (Fig. 6D). Notably, the localization of the snRNP component HEXIM1, of which nearly 30% is in complex with P-TEFb in HeLa cells (43,44), was unaltered, consistent with previous reports that Tat expression disassembles P-TEFb from the inhibited 7SK snRNP complex (43). Of the ZUN proteins, UBE2O remained entirely cytoplasmic upon Tat expression whereas ZFP91 and NAP1L4 were reduced in the cytoplasm. Only

ZFP91 became enriched in the histone-containing fraction, like P-TEFb, potentially implicating ZFP91 in targeting the kinase to chromatin.

The transcriptional inhibitor DRB re-localizes free P-TEFb to the nucleus

Finally, we wondered whether the nuclear import of P-TEFb freed from cytoplasmic 7SK snRNP might be a more general pathway to regulate transcription elongation. To explore this idea, we used the CDK7/9 inhibitor DRB, which blocks transcription elongation by decreasing RNAP II serine 2 phosphorylation (S2-P) (45). As expected, RNAP II S2-P was dramatically reduced when HeLa cells were treated with DRB (Fig. 6E). Remarkably, DRB treatment induced extensive re-localization of P-TEFb, but not HEXIM1, from the cytoplasm to the nuclear pellet, the same fraction containing reduced RNAP II S2-P where the kinase likely restores proper polymerase phosphorylation. Supporting a role of UBE2O in this pathway, DRB treatment also increased the association between UBE2O and HEXIM1 (Fig. 6F). These results substantiate the model that nuclear re-localization of free P-TEFb is a global regulatory mechanism to sustain proper RNAP II S2-P levels.

Discussion

HIV-1 has a small 9-kilobase genome encoding only 18 proteins and thus, like other viruses, has evolved ways to economically utilize its limited genetic information and rely heavily on host protein complexes to execute all stages of its life cycle. Here we show that the viral transcription factor Tat hijacks three new host partners (ZFP91, UBE2O, and NAP1L4) to form a multi-protein complex that adds ubiquitin to the transcription machinery to enhance viral transcription. This ZUN complex binds Tat via an interaction surface distinct from the well-known surface used to bind P-TEFb (31).

We discovered that a single lysine (Lys28) and adjacent phenylalanine (Phe32) in Tat are largely responsible for the interaction with ZUN proteins, particularly with the UBE2O ligase. The requirement for just a few critical residues helps further explain how Tat can rewire large host complexes for its function using relatively little genetic information. Previous structural work has implicated Lys28 and Phe32 in directly binding AFF4 of the SEC (24) and thus may be an interaction “hot spot”. Interestingly, while AFF4 is localized exclusively to the nucleus (Fig. 6D), Tat recruits UBE2O to the 7SK snRNP in the cytoplasm and thus the “hot spot” can be used to form complexes with different host partners in distinct cellular compartments.

A classic example of transcriptional regulation being coupled to signaling events in the cytoplasm is illustrated by the NF-κB pathway, where the application of a stimulus (e.g. TNFα) leads to the cytoplasmic degradation of inhibitor of κB (IκB) and subsequent nuclear import of NF-κB to activate gene expression (46). By analogy, we find a significant pool of the inhibitory 7SK snRNP-P-TEFb complex in the cytoplasm, presumably reflecting the “large” P-TEFb complexes previously described (40), that serves as the inactive reservoir of P-TEFb. Surprisingly, Tat expression resulted in a significant redistribution of the P-TEFb subunits CCNT1 and CDK9, but not HEXIM1, from the cytoplasm to the histone-containing nuclear fraction, confirming that Tat can utilize the pool of inhibited P-TEFb. In support of this hypothesis are reports of a high FRET signal between Tat and CCNT1 in the cytoplasm (47) and an intrinsic nucleocytoplasmic shuttling activity of CDK9 (48). It is interesting that NAP1L4 also can shuttle between the nucleus and cytoplasm (49), possibly implicating this histone chaperone in the nuclear import process. Many stimuli are known to release P-TEFb

from the 7SK snRNP (13,50–52), so the nuclear import of the free kinase is likely a general mechanism to control transcription elongation and not restricted to this viral case. Indeed, treatment of cells with the CDK inhibitor DRB potently relocalized P-TEFb to the nuclear pellet fraction, phenocopying the Tat effect (Fig. 6E). Intriguingly, UBE2O has been shown to inhibit NF-κB activation by preventing the polyubiquitination of the TRAF6 signaling protein (Zhang, et al. 2013), so UBE2O is likely intercalated in a signaling pathway that regulates NF-κB and P-TEFb nuclear import, depending on the stimulus. An exciting future direction is to detail the additional cellular machinery that controls this multi-compartment signaling pathway.

The hijacking of the UBE2O E3 ligase by Tat results in ubiquitination of the HEXIM1 protein of the 7SK snRNP and adds a new dimension to viral transcriptional regulation. The monoubiquitination by UBE2O does not cause protein degradation but rather stimulates two signaling outcomes, adding to the growing list of functions for ubiquitination (39,54–56). The first outcome is the partial disassembly of the 7SK snRNP by decreased interaction between HEXIM1 and LARP7. As Tat expression also leads to HEXIM1 ubiquitination but fully releases P-TEFb from HEXIM1, we hypothesize that the UBE2O-dependent LARP7 disassembly step is an intermediate in a pathway that ultimately results in complete P-TEFb release. The second effect of HEXIM1 ubiquitination is its decreased enrichment in the chromatin fraction and accumulation in the cytoplasm, which is similar to a previously reported function of UBE2O ligase activity in regulating the localization of chromatin-associated proteins by increasing their cytoplasmic retention (21). Interestingly, UBE2O expression alone is not sufficient to also drive the nuclear import of P-TEFb. One possibility is that the ubiquitin-dependent

sequestering of HEXIM1 in the cytoplasm sustains P-TEFb activity in the nucleus by limiting the re-association of the free kinase with its inhibitor. Our findings reveal that the coordination of multiple re-localization pathways is involved in regulating transcription elongation.

Our findings lead to a revised model of Tat activation (Fig. 7). As shown previously (18), the 7SK snRNP can be recruited to the HIV-1 promoter in its basal state (without Tat), resulting in CDK9 inhibition and paused RNA polymerase II, consequently poising the promoter for activation. Upon Tat expression, Tat is also recruited to the promoter, likely through 7SK snRNP binding, as RNAse treatment reduces Tat enrichment (18). During co-transcriptional activation (Fig. 7, top), the synthesis of TAR triggers the Tat-dependent release of the 7SK snRNP, activating the P-TEFb kinase. Our hypothesis is that as Tat protein levels increase from the positive feedback loop of HIV-1 transcription (57), Tat begins to target the cytoplasmic P-TEFb-7SK snRNP complex for activation (Fig. 7, bottom). This idea is supported by the dose-response of P-TEFb nuclear import and Tat protein levels (Fig. 6D). During cytoplasmic activation, P-TEFb released from 7SK snRNP complexes is imported to the nucleus to phosphorylate the RNAP II CTD and sustain a high level of transcription elongation. We believe that the release of P-TEFb from cytoplasmic snRNP is the predominant activation mechanism as ChIP experiments show that Tat expression increases the enrichment of CCNT1 and CDK9 (free P-TEFb), but not snRNP proteins, at the HIV-1 promoter (18). Therefore, once transcription is induced, the full 7SK snRNP is not continually assembled into pre-initiation complexes. These modes of transcriptional activation are employed endogenously as both the splicing protein SC35 (20) and

helicase DDX21 (58) can release promoter-bound P-TEFb (co-transcriptional activation), while BRD4 delivers free P-TEFb to promoters (59,60), especially after signal-dependent release from the 7SK snRNP (cytoplasmic activation) (7).

Another protein of the ZUN complex, ZFP91, is a classical C2H2 zinc finger (ZnF) protein that shows some of the largest effects on Tat activation (Fig. 1). Although ZnF domains generally function as DNA-binding motifs, a previous report demonstrated that one ZFP91 ZnF domain can function as an atypical E3 ubiquitin ligase (28). We were unable to identify any substrate proteins for ZFP91 ligase activity in the Tat-P-TEFb transcriptional network, but ZFP91, like P-TEFb, became enriched in the nuclear pellet fraction upon Tat expression (Fig. 6D). One intriguing possibility is that ZFP91 is involved in targeting P-TEFb to chromatin. Many C2H2 ZnF proteins also contain a Krüppel associated box (KRAB) domain that mediates transcriptional repression, generally through association with the KAP1 (TRIM28) protein (61). KAP1 has recently been shown to activate the HIV promoter by co-occupying promoter-proximal regions with the 7SK snRNP and paused RNAP II (19) and, interestingly, we observe several KAP1 peptides in ZFP91 AP-MS experiments (Table S1). While ZFP91 does not contain a KRAB domain, it would be enlightening to investigate whether KAP1 acts in concert with ZFP91 in P-TEFb recruitment, whether other KRAB-containing ZnF proteins mediate repressive functions of KAP1, and whether ZFP91 regulates its activating function.

Experimental Procedures

Cell culture, plasmids, reagents: HeLa$^{provirus\Delta tat}$, HeLa, and HEK 293T cells were maintained in DMEM (10% FBS, 1% pen/strep) at 37C with 5% CO_2. ZFP91 and

UBE2O were cloned into pcDNA4-TO (Thermo) with a C-terminal 3x-FLAG tag. ZFP91, NAP1L4, UBE2O, CYCT1, CDK9, HEXIM1, LARP7, MEPCE, and HIV-1 HXB2 Tat were cloned into pcDNA4-TO with a C-terminal 2xSTREP tag. Human ubiquitin was cloned into pcDNA4-TO containing an N-terminal HA tag. All HEXIM1 and HIV-1 Tat point mutants were generated by standard quick-change mutagenesis. The HEXIM1 ΔK mutants were ordered as a gBlock gene fragment from IDT with restriction sites and cloned into appropriate vectors. DNA was transfected with Poly Jet (Signa Gen) and siRNA was transfected with HiPerFect (Qiagen) or RNAi MAX (Thermo). All transfections were performed according to manufacturer's protocol. MG132 was purchased from Selleckchem (S2619). DRB was purchased from Abcam (ab120939). 2x and 4x Laemmli buffer were purchased from BioRad. Human E1 protein was a kind gift from the laboratory of John Gross.

Antibodies: Normal rabbit IgG (sc-2027), RNA polymerase II (sc-899x), CCNT1 (sc-8127 and sc-10750), MEPCE (sc-82542), and SP1 (sc-14027) were purchased from Santa Cruz Biotechnology. AFF1 (A302-344A), AFF4 (A302-538A), NAP1L4 (A304-578A), UBE2O (A301-873A), and ZFP91 (A303-245A) antibodies were purchased from Bethyl. The HA antibody (MMS-101P) was purchased from Covance. LARP7 antibodies were purchased from Aviva (ARP40847_P050) and Santa Cruz (sc-515209). HEXIM1 antibodies were purchased from Abcam (ab25388) and Everest (EB06964). CDK9 antibodies were purchased from Santa Cruz (sc-8338) and ProMab (30212). The tubulin (T6199) and FLAG (F1804) antibodies were purchased from Sigma. The RNA pol II phosphor-Ser2 (ab5095) was purchased from Abcam. The STREP-HRP antibody

(71591) was purchased from Millipore. The histone H4 antibody (39269) was purchased from Active Motif.

Targeted RNAi screen of high confidence Tat interactors: HeLa$^{provirus\Delta tat}$ cells at 40% confluency were transfected with non-silencing or specific siRNAs (Qiagen) at 5-10 nM final concentrations using HiPerFect. After 48 hours cells were transfected with 0.25 ng Tat-expressing plasmid and 1 ng CMV-firefly luciferase plasmid as a transfection control using PEI. After an additional 48 hours, cells were lysed in cell lysis buffer (5 mM Tris-HCl pH 8.0, 85 mM KCl, 0.5% IGEPAL CA-630). Cell-associated HIV p24 production was detected by p24 ELISA. In parallel, 10 µL of lysate was used in a standard luciferase assay. All p24 values were normalized to firefly luciferase activity for every siRNA to generate a relative Tat activity score, defined as [p24$^{host\ RNAi}$/FFL$^{host\ RNAi}$] / [p24$^{N.S.\ RNAi}$/FFL$^{N.S.\ RNAi}$]. For analysis of protein knockdown, 2x Laemmli sample buffer was added to siRNA-treated cells and then loaded onto gels. Knockdowns were determined by Western blot with the indicated antibodies.

Denaturing *in vivo* ubiquitination assay: HEK 293T or HeLa cells at 60-70% confluency were transfected with 7 µg total DNA containing STREP-tagged bait protein and HA-Ubiquitin using Poly Jet. After 30 hours, cells were collected and flash frozen in liquid nitrogen. Cells were thawed on ice, lysed in an equal volume SDS Lysis Buffer (2% SDS, 50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5 mM DTT) and boiled for 10 minutes. Boiled lysates were sonicated for 5 cycles (17% amplitude, 20 second pulse) on a Fisher Sonic Dismembrator Model 500 and centrifuged in a microfuge at full speed. SDS Dilution Buffer (1% SDS, 50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5 mM DTT) was added to the cleared lysate to obtain final SDS concentration of 1% and incubated with

20 µL Streptactin Superflow Agarose for 3 hours at room temperature. The Streptactin resin was then washed 4x with RIPA buffer (1% Triton X-100, 50 mM Tris-HCl pH 8.0, 150 mM NaCl, 0.5% Na-deoxycholate, 0.1% SDS) at room temperature. Bound material was eluted with 1x STREP elution buffer (IBA-GmbH) shaking at room temperature. Ubiquitination of immunoprecipitated material was detected by HA Western blot. For RNAi-ubiquitination experiments, cells were first transfected with siRNA for 48 hours prior to DNA transfection. For MG132 treatment, cells that had been transfected with DNA for 24 hours were then incubated with 2.5 µM MG132 for an additional 16 hours before harvesting.

Cloning, expression, and purification of recombinant proteins: Human wild type and ΔK ubiquitin with an N-terminal HA tag were cloned into the pGEX6p-1 vector (GE Healthcare) and expressed as a fusion protein with an N-terminal GST tag. The fusion proteins were expressed in *E. coli* BL21 (DE3). The GST-fusion proteins were digested on column with PreScission protease. The released ubiquitin proteins were collected and further purified by Superdex 200 Increase (GE Healthcare).

***In vitro* ubiquitination assay**: Assays were performed in a total volume of 30 µL with 0.1 µM recombinant human E1, 0.1 µM wild type UBE2O or catalytically inactive UBE2O (both wild-type UBE2O and UBE2O CD were purified from HEK 293T cells with high salt), 5 µM HA-tagged ubiquitin (wild-type or ΔK) and 0.2 µM HEXIM1 (purified from HEK 293T cells) in reaction buffer containing 50 mM Tris-HCl pH7.4, 100 mM NaCl, 1 mM DTT, 5 mM $MgCl_2$, 4 mM ATP at 37°C with constant shaking for 2 hours. Samples were quenched with 4x Laemmli sample buffer (containing 2-Mercaptoethanol)

before analysis by SDS-PAGE. Ubiquitination of HEXIM1 was verified by immunoblotting with anti-HEXIM1 antibody.

RNA elongation assay: HeLa$^{provirus\Delta tat}$ cells at 40% confluency were transfected with siRNAs to obtain 10nM final concentration. After 48 hours, cells were transfected -/+ a Tat-expressing plasmid for an additional 10 hours. RNA was purified on Zymo Quick-RNA Mini Prep columns according to the manufacturer's protocol, DNAse-treated with TURBO DNAse (Thermo), and used in RT-PCR reactions with a primer immediately downstream of the HIV TAR sequence to detect short RNA species and oligo dT to detect Tat-activated poly-adenylated messages. Message abundance was then quantified by qPCR with proximal (Prox F 5'-CTCTCTGGTTAGACCAGATCTGA-3', Prox R 5'-GTGGGTTCCCTAGTTAGCCAGA-3') and distal (Dist F 5'-CACATGAAGCAGCACGACTT-3', Dist R 5'-GGTCTTGTAGTTGCCGTCGT-3') HIV primer sets and normalized to the abundance of RPLP0 mRNA.

Affinity purification and Western blot: HEK 293T or HeLa cells at 60-70% confluency were transfected with 7 µg total DNA containing STREP-tagged or FLAG-tagged bait protein. After 40 hours, cells were collected, lysed in Lysis Buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM $MgCl_2$, 0.5% Triton X-100, 0.2% Na-deoxycholate, Roche protease inhibitor cocktail) and rocked for 30 minutes at 4°C. Lysate was cleared and incubated with Streptactin Superflow resin slurry (IBA-GmBH) or magnetic M2-FLAG resin slurry (Sigma) for 3 hours at 4°C. The resin was washed 4x with WASH buffer (50 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM $MgCl_2$, 0.5% Triton X-100). Washed resin was eluted with 1x STREP elution buffer (IBA-GmbH) diluted in WASH buffer or FLAG elution buffer (200 µg/mL 3x FLAG peptide, diluted in WASH buffer), shaking at RT for

30 minutes. The eluate was mixed 1:1 with 2x Laemmli sample buffer, boiled, and loaded on a SDS-PAGE gel. Proteins were transferred to PVDF membranes and detected with specific antibodies. Pixel intensity quantification was performed in ImageJ.

Double IP: HEK 293T cells at 60-70% confluency in a 15cm dish (6x per condition) were transfected with 3 µg each of a STREP-tagged and a FLAG-tagged protein. After 48 hours, cells were lysed in 10 mL Lysis Buffer. Cleared lysate was added to 400 µL of Streptactin Superflow and rotated at 4°C for 3 hours. After STREP elution, supernatant was added to 100 µL M2-FLAG resin and rotated at 4°C for 3 hours. The FLAG eluate was examined by SDS-PAGE and probed by Western blot using specific antibodies.

RNA-IP: A fraction of the elution from a STREP IP was resuspended in RNA Lysis Buffer (Zymo). RNA was then processed with Zymo Quick-RNA Mini Prep kit as detailed above. Primers specific for the 7SK snRNA were used in RT-qPCR reactions (7SK F: 5'-GGATGTGAGGCGATCTGGC-3', 7SK R: 5'- AAAAGAAAGGCAGACTGCCAC-3')

High salt IP: For high salt purification of UBE2O WT and CD from HEK 293T cells, the same lysis and IP was performed as above. For washing, beads were washed 4x in High Salt Wash Buffer (25 mM Tris-HCl pH 8.0, 0.1% Triton X-100, 2mM $MgCl_2$, 1 M NaCl, 0.5 mM DTT) and 2x in IP Dilution Buffer (25 mM Tris-HCl pH 8.0, 0.1% Triton X-100, 2 mM $MgCl_2$, 150 mM NaCl, 0.5 mM DTT) and eluted as described.

Endogenous IP: HeLa or HEK 293T cells were lysed in Lysis Buffer. Cleared cell lysates were incubated with control IgG or specific antibodies (1-10 µg) for 3 hours. Protein A Dynabeads slurry (Life Technologies) was washed and added to antibody-lysate mixture for an additional 3 hours. Dynabeads were then washed 3x with 1 mL of WASH buffer. For elution, 50 µL of 2x Laemmli Buffer (BioRad) was added to beads

and incubated at 65C for 15 minutes, with occasional mixing. After elution, 1.5 µL of BME was added to eluate, which was then processed by SDS-PAGE and Western blotting.

Endogenous RNA-IP: Immunocomplexes were eluted with 50 µL of RNA elution buffer (10 mM EDTA pH 8.0, 1% SDS, 50 mM Tris-HCl pH 8.0) for 15 minutes at 65°C. RNA was purified from the elution on Quick-RNA MiniPrep columns as detailed above. 4 µL of DNAse-treated RNA was used in an RT reaction with primers specific for the 7SK snRNA and quantitated by qPCR.

Affinity purification-mass spectrometry for ZFP91 and HEXIM1: AP-MS experiments were performed as previously described (23,62). Slight modifications were as follows: HEK 293T cells were transfected with 5-20 µg plasmid complexed with PolyJet Reagent. After 40 hours, cells were harvested and washed in PBS, and then resuspended in Lysis Buffer (0.5% Nonidet P 40 Substitute, 50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, cOmplete mini EDTA-free protease and PhosSTOP phosphatase inhibitor cocktails (Roche)). Following incubation at 4°C for 30 minutes on a tube rotator, lysates were centrifuged (3,500 x g) for 20 minutes at 4°C. Cleared lysates and 2µg of 1x FLAG peptide (Sigma) were added to 20 µL anti-FLAG M2 magnetic beads and rotated for 2 hours at 4°C. FLAG beads were sequentially washed 3x with Wash Buffer (0.05% NP-40, 50 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM EDTA) and 1x with Wash Buffer without NP-40. Beads were incubated with Elution Buffer (50 mM Tris-HCl, pH 7.4, 150 mM NaCl, 1 mM EDTA, 0.05% RapiGest SF Surfactant (Waters)) on a vortex mixer at room temperature for 30 minutes. Eluates

(10 µL) were processed and analyzed by mass spectrometry as previously described (63,64).

Identification of ubiquitinated HEXIM1 lysines by MS: HEK 293T cells at 60% confluency in 11x 15cm plates were transfected with 1 µg of HEXIM1-S, 5 µg of UBE2O-F, and 2 µg of HA-Ub (to 15 µg DNA total with pBluescript) by PolyJet. After 48 hours, a denaturing STREP IP was performed on the lysate. Sample eluate was digested with trypsin or Arg-C (according to manufacturer's protocol) and analyzed by mass spectrometry as previously described (65) with slight modifications. Briefly, the raw data was matched to protein sequences by the MaxQuant algorithm (version 1.5.2.8)(66). Data were searched against a database containing SwissProt Human (downloaded 12/2014), and the APEX2 sequence, concatenated to a decoy database where each sequence was randomized in order to estimate the false discovery rate (FDR). The first search was performed with a mass accuracy of +/- 20 parts per million and the main search was performed with a mass accuracy of +/- 4.5 parts per million. A maximum of 5 modifications were allowed per peptide. A maximum of 5 missed cleavages were allowed. The maximum charge allowed was 7+. Individual peptide mass tolerances were allowed. For MS/MS matching, a mass tolerance of 0.8 Da was allowed and the top 8 peaks per 100 Da were analyzed. MS/MS matching was allowed for higher charge states, water and ammonia loss events. The data were filtered to obtain a peptide, protein, and site-level false discovery rate of 0.01. The minimum peptide length was 7 amino acids. Results were matched between runs with a time window of 2 minutes for biological replicates.

Subcellular fractionation: HEK 293T or HeLa cells were collected and the pelleted cell volume (PCV) was measured. For DRB treatment, cells were treated with DMSO control or 100 μM DRB for 1 hour prior to harvesting. Cells were washed in 3x PCV of Buffer A (10 mM Tris-HCl, pH 8.0, 1.5 mM $MgCl_2$, 10 mM KCl, 0.5 mM DTT, Roche protease inhibitor cocktail), resuspended in 5x PCV Buffer A, and allowed to swell for 10 minutes on ice. Triton X-100 was then added to the cells to achieve a final concentration of 0.5% and rocked for 1 minute. The cellular material was passed through a 25G needle 5x and spun at 1500 x g for 5 minutes. The supernatant was collected as the cytoplasmic fraction. The nuclei were washed 1x in Buffer A and resuspended in ¼ pelleted nuclear volume (PNV) of Buffer B (20 mM Tris-HCl, pH 8.0, 25% glycerol, 1.5 mM $MgCl_2$, 20 mM KCl, 0.2 mM EDTA, 0.5 mM DTT, Roche protease inhibitors). ¼ PNV of Buffer C (20 mM Tris-HCl, pH 8.0, 25% glycerol, 1.5 mM $MgCl_2$, 1.2 M KCl, 0.2 mM EDTA, 0.5 mM DTT, Roche protease inhibitors) was then added drop-wise to nuclei, which were then rocked for 30 minutes. The suspension was centrifuged at 20,000 x g for 15 minutes. The supernatant was collected as the high salt nuclear extract (nucleoplasm). The nuclear pellet was washed 1x in Buffer A and resuspended in RIPA buffer and sonicated for 5 cycles (17% amplitude, 20 second pulse) on a Fisher Sonic Dismembrator Model 500. The sonicated lysate was spun at full speed and the supernatant was collected as the nuclear pellet. A Bradford assay (BioRad) was used to determine the protein concentration in each fraction. 10 μg of protein from each fraction was analyzed by SDS-PAGE and Western blot.

Immunofluorescence: HeLa$^{provirus\Delta tat}$ cells at 90% confluency on coverslips were fixed in 4% formaldehyde for 20 minutes, quenched with 0.15 M glycine, and permeabilized

with 0.1% Triton X-100 for 15 minutes. Permeabilized cells were blocked with blocking buffer (10 mM Tris-HCl pH 7.5, 150 mM NaCl, 2% BSA, 10% goat serum) for 1 hour, and then incubated with anti-CCNT1 antibody (sc-10750) at dilution of 1:100 for 1 hour. Cells were then washed with PBS and incubated with goat anti-rabbit IgG-Alexa Fluor 555 (1:1000) and 100 ng/µl DAPI in blocking buffer. Cover slips were washed in PBS and mounted on slides with ProLong Gold antifade mounting solution (Thermo). Confocal images were obtained with a spinning disk confocal microscope (Nikon Ti-E microscope equipped with Yokagawa CSU-22) by using the 561 nm and 405 nm excitation for Alexa Fluor 555 epifluorescence and 4,6-diamidino-2-phenylindole (DAPI), respectively.

Author Contributions

Conceptualization, Tyler B. Faust and Alan D. Frankel; Investigation, T.B.F., Gwendolyn M. Jang, Yang Li, and Billy W. Newton; Writing-Original Draft, T.B.F. and A.D.F.; Writing-Review & Editing, T.B.F., G.M.J., and A.D.F.; Visualization, T.B.F., G.M.J., Y.L., and A.D.F.; Supervision, A.D.F.; Funding Acquisition, A.D.F. and Nevan J. Krogan.

Acknowledgements

We thank Dr. J.J. Miranda, Dr. John Gross, and members of the Frankel laboratory for critically reading the manuscript and Dr. Orly Laufman for technical assistance with the immunofluorescence. This work was supported by an N.I.H. grant (P50GM088250) to A.D.F. and N.J.K. All the authors declare no conflict of interest.

References

1. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, et al. C-Myc regulates transcriptional pause release. Cell [Internet]. Elsevier Ltd; 2010;141(3):432–45. Available from: http://dx.doi.org/10.1016/j.cell.2010.03.030
2. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science [Internet]. 2008;322(5909):1845–8. Available from: http://science.sciencemag.org/content/322/5909/1845.abstract
3. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science [Internet]. 2013;339(6122):950–3. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3974810&tool=pmcentrez&rendertype=abstract
4. Henriques T, Gilchrist D, Nechaev S, Bern M, Muse G, Burkholder A, et al. Stable pausing by rna polymerase II provides an opportunity to target and integrate regulatory signals. Mol Cell [Internet]. Elsevier Inc.; 2013;52(4):517–28. Available from: http://dx.doi.org/10.1016/j.molcel.2013.10.001
5. Kwak H, Lis JT. Control of transcriptional elongation. Annu Rev Genet [Internet]. 2013;47:483–508. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3974797&tool=pmcentrez&rendertype=abstract
6. Czudnochowski N, Bösken C a., Geyer M. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. Nat

Commun. 2012;3(May):842.

7. Yang Z, Yik JHN, Chen R, He N, Moon KJ, Ozato K, et al. Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. Mol Cell. 2005;19(4):535–45.

8. Takahashi H, Parmely TJ, Sato S, Tomomori-Sato C, Banks CAS, Kong SE, et al. Human mediator subunit MED26 functions as a docking site for transcription elongation factors. Cell [Internet]. Elsevier Inc.; 2011;146(1):92–104. Available from: http://dx.doi.org/10.1016/j.cell.2011.06.005

9. Barboric M, Nissen RM, Kanazawa S, Jabrane-Ferrat N, Peterlin BM. NF-kappaB binds P-TEFb to stimulate transcriptional elongation by RNA polymerase II. TL - 8. Mol Cell [Internet]. 2001;8 VN-re(2):327–37. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11545735

10. Wei P, Garber ME, Fang SM, Fischer WH, Jones KA. 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. 1998;92(4):451–62.

11. Feinberg MB, Baltimore D, Frankel AD. The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation. Proc Natl Acad Sci U S A [Internet]. 1991;88(9):4045–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51590&tool=pmcentrez&rendertype=abstract

12. Yang Z, Zhu Q, Luo K, Zhou Q. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. Nature [Internet].

2001;414(6861):317–22. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11713532\nhttp://dx.doi.org/10.1038/35104575

13. Nguyen VT, Kiss T, Michels AA, Bensaude O. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. Nature [Internet]. 2001;414(6861):322–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11713533\nhttp://dx.doi.org/10.1038/35104581
14. Michels A a, Fraldi A, Li Q, Adamson TE, Bonnet F, Nguyen VT, et al. Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. EMBO J. 2004;23(13):2608–19.
15. He N, Jahchan NS, Hong E, Li Q, Bayfield MA, Maraia RJ, et al. A La-Related Protein Modulates 7SK snRNP Integrity to Suppress P-TEFb-Dependent Transcriptional Elongation and Tumorigenesis. Mol Cell. 2008;29(5):588–99.
16. Chen R, Liu M, Li H, Xue Y, Ramey WN, He N, et al. PP2B and PP1alpha cooperatively disrupt 7SK snRNP to release P-TEFb for transcription in response to Ca2+ signaling. Genes Dev. 2008;22(10):1356–68.
17. Liu P, Xiang Y, Fujinaga K, Bartholomeeusen K, Nilson KA, Price DH, et al. Release of positive transcription elongation factor b (P-TEFb) from 7SK small nuclear ribonucleoprotein (snRNP) activates hexamethylene bisacetamide-inducible protein (HEXIM1) transcription. J Biol Chem. 2014;289(14):9918–25.
18. D'Orso I, Frankel AD. RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. Nat Struct Mol Biol. 2010;17(7):815–

21.

19. McNamara RP, Reeder JE, McMillan EA, Bacon CW, McCann JL, D'Orso I. KAP1 Recruitment of the 7SK snRNP Complex to Promoters Enables Transcription Elongation by RNA Polymerase II. Mol Cell. 2016;61(1):39–53.

20. Ji X, Zhou Y, Pandit S, Huang J, Li H, Lin CY, et al. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. Cell [Internet]. Elsevier Inc.; 2013;153(4):855–68. Available from: http://dx.doi.org/10.1016/j.cell.2013.04.028

21. Mashtalir N, Daou S, Barbour H, Sen NN, Gagnon J, Hammond-Martel I, et al. Autodeubiquitination protects the tumor suppressor BAP1 from cytoplasmic sequestration mediated by the atypical ubiquitin ligase UBE2O. Mol Cell [Internet]. Elsevier Inc.; 2014;54(3):392–406. Available from: http://dx.doi.org/10.1016/j.molcel.2014.03.002

22. Zhang X, Zhang J, Bauer A, Zhang L, Selinger DW, Lu CX, et al. Fine-tuning BMP7 signalling in adipogenesis by UBE2O/E2-230K-mediated monoubiquitination of SMAD6. EMBO J [Internet]. Nature Publishing Group; 2013;32(7):996–1007. Available from: http://emboj.embopress.org/content/32/7/996.abstract

23. Jäger S, Cimermancic P, Gulbahce N, Johnson JR, McGovern KE, Clarke SC, et al. Global landscape of HIV-human protein complexes. Nature [Internet]. 2012;481(7381):365–70. Available from: http://dx.doi.org/10.1038/nature10719

24. Schulze-Gahmen U, Lu H, Zhou Q, Alber T. AFF4 binding to TAT-P-TEFb indirectly stimulates TAR recognition of super elongation complexes at the HIV

promoter. Elife. 2014;2014(3):1–13.

25. He N, Liu M, Hsu J, Xue Y, Chou S, Burlingame A, et al. HIV-1 Tat and Host AFF4 Recruit Two Transcription Elongation Factors into a Bifunctional Complex for Coordinated Activation of HIV-1 Transcription. Mol Cell [Internet]. Elsevier Ltd; 2010;38(3):428–38. Available from: http://dx.doi.org/10.1016/j.molcel.2010.04.013
26. Sobhian B, Laguette N, Yatim A, Nakamura M, Levy Y, Kiernan R, et al. HIV-1 Tat Assembles a Multifunctional Transcription Elongation Complex and Stably Associates with the 7SK snRNP. Mol Cell [Internet]. Elsevier Ltd; 2010;38(3):439–51. Available from: http://dx.doi.org/10.1016/j.molcel.2010.04.012
27. Muniz L, Egloff S, Ughy B, Jády BE, Kiss T. Controlling cellular P-TEFb activity by the HIV-1 transcriptional transactivator tat. PLoS Pathog. 2010;6(10).
28. Jin X, Jin HR, Jung HS, Lee SJ, Lee JH, Lee JJ. An atypical E3 ligase zinc finger protein 91 stabilizes and activates NF-κB-inducing kinase via Lys63-linked ubiquitination. J Biol Chem. 2010;285(40):30539–47.
29. Rodriguez P, Pelletier J, Price GB, Zannis-Hadjopoulos M. NAP-2: histone chaperone function and phosphorylation state through the cell cycle. J Mol Biol [Internet]. 2000;298(2):225–38. Available from: http://www.ncbi.nlm.nih.gov/pubmed/10764593
30. Huttlin EL, Ting L, Bruckner RJ, Gebreab F, Gygi MP, Szpyt J, et al. The BioPlex Network: A Systematic Exploration of the Human Interactome. Cell. 2015;162(2):425–40.
31. Tahirov TH, Babayeva ND, Varzavand K, Cooper JJ, Sedore SC, Price DH. Crystal structure of HIV-1 Tat complexed with human P-TEFb.pdf. Nature

[Internet]. Nature Publishing Group; 2010;465(7299):747–51. Available from: http://dx.doi.org/10.1038/nature09131

32. D'Orso I, Jang GM, Pastuszak AW, Faust TB, Quezada E, Booth DS, et al. Transition step during assembly of HIV Tat:P-TEFb transcription complexes and transfer to TAR RNA. Mol Cell Biol [Internet]. 2012;32(23):4780–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23007159\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3497596

33. D'Orso I, Frankel AD. Tat acetylation modulates assembly of a viral-host RNA-protein transcription complex. Proc Natl Acad Sci U S A. 2009;106(9):3101–6.

34. Lu H, Li Z, Xue Y, Schulze-Gahmen U, Johnson JR, Krogan NJ, et al. AFF1 is a ubiquitous P-TEFb partner to enable Tat extraction of P-TEFb from 7SK snRNP and formation of SECs for HIV transactivation. Proc Natl Acad Sci [Internet]. 2014;111(1):E15–24. Available from: http://www.pnas.org/content/111/1/E15.abstract

35. Lau J, Qiao JL, Diribarne G, Michels AA, Dey A, Bensaude O, et al. Ubiquitination of HEXIM1 by HDM2. Cell Cycle. 2009;8(14):2247–54.

36. Yik JHN, Chen R, Pezda AC, Samford CS, Zhou Q. A human immunodeficiency virus type 1 Tat-like arginine-rich RNA-binding domain is essential for HEXIM1 to inhibit RNA polymerase II transcription through 7SK snRNA-mediated inactivation of P-TEFb. Mol Cell Biol. 2004;24(12):5094–105.

37. Barboric M, Kohoutek J, Price JP, Blazek D, Price DH, Peterlin BM. Interplay between 7SK snRNA and oppositely charged regions in HEXIM1 direct the

inhibition of P-TEFb. EMBO J [Internet]. 2005;24(24):4291–303. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1356324&tool=pmcentrez&rendertype=abstract

38. Ray S, Zhao Y, Jamaluddin M, Edeh CB, Lee C, Brasier AR. Inducible STAT3 NH2 terminal mono-ubiquitination promotes BRD4 complex formation to regulate apoptosis. Cell Signal [Internet]. Elsevier Inc.; 2014;26(7):1445–55. Available from: http://dx.doi.org/10.1016/j.cellsig.2014.03.007

39. Werner A, Iwasaki S, McGourty CA, Medina-Ruiz S, Teerikorpi N, Fedrigo I, et al. Cell-fate determination by ubiquitin-dependent regulation of translation. Nature [Internet]. 2015;525(7570):523–7. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26399832

40. Ramanathan Y, Reza SM, Young TM, Mathews MB, Pe'ery T. Human and rodent transcription elongation factor P-TEFb: interactions with human immunodeficiency virus type 1 tat and carboxy-terminal domain substrate. J Virol [Internet]. 1999;73(7):5448–58. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=112601&tool=pmcentrez&rendertype=abstract

41. A. Ranki, A. Lagerstedt, V. Ovod, E. Aavik and KJEK. Expression kinetics and subcellular localization of HIV-1 regulatory protein Nef, Tat and Rev in acutely and chronically infected lymphoid cell lines. Arch Virol. 1994;(139):365–78.

42. Stauber RH, Pavlakis GN. Intracellular Trafficking and Interactions of the HIV-1 Tat Protein. 1998;136:126–36.

43. Barboric M, Yik JHN, Czudnochowski N, Yang Z, Chen R, Contreras X, et al. Tat

competes with HEXIM1 to increase the active pool of P-TEFb for HIV-1 transcription. Nucleic Acids Res. 2007;35(6):2003–12.

44. Byers SA, Price JP, Cooper JJ, Li Q, Price DH. HEXIM2, a HEXIM1-related protein, regulates positive transcription elongation factor b through association with 7SK. J Biol Chem. 2005;280(16):16360–7.

45. Koga M, Hayashi M, Kaida D. Splicing inhibition decreases phosphorylation level of Ser2 in Pol II CTD. Nucleic Acids Res. 2015;43(17):8258–67.

46. Napetschnig J, Wu H. Molecular basis of NF-κB signaling. Annu Rev Biophys [Internet]. 2013;42:443–68. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3678348&tool=pmcentrez&rendertype=abstract

47. Marcello A, Cinelli RAG, Ferrari A, Signorelli A, Tyagi M, Pellegrini V, et al. Visualization of in Vivo Direct Interaction between HIV-1 TAT and Human Cyclin T1 in Specific Subcellular Compartments by Fluorescence Resonance Energy Transfer. J Biol Chem. 2001;276(42):39220–5.

48. Napolitano G, Licciardo P, Carbone R, Majello B, Lania L. CDK9 has the intrinsic property to shuttle between nucleus and cytoplasm, and enhanced expression of cyclinT1 promotes its nuclear localization. J Cell Physiol. 2002;192(2):209–15.

49. Rodriguez P, Munroe D, Prawitt D, Chu LL, Bric E, Kim J, et al. Functional characterization of human nucleosome assembly protein-2 (NAP1L4) suggests a role as a histone chaperone. Genomics [Internet]. 1997;44(3):253–65. Available from: http://www.sciencedirect.com/science/article/pii/S0888754397948680

50. Bartholomeeusen K, Xiang Y, Fujinaga K, Peterlin BM. Bromodomain and extra-

terminal (BET) bromodomain inhibition activate transcription via transient release of Positive Transcription Elongation Factor b (P-TEFb) from 7SK small nuclear ribonucleoprotein. J Biol Chem. 2012;287(43):36609–16.

51. Biglione S, Byers S a, Price JP, Nguyen VT, Bensaude O, Price DH, et al. Inhibition of HIV-1 replication by P-TEFb inhibitors DRB, seliciclib and flavopiridol correlates with release of free P-TEFb from the large, inactive form of the complex. Retrovirology. 2007;4:47.

52. Contreras X, Barboric M, Lenasi T, Peterlin BM. HMBA releases P-TEFb from HEXIM1 and 7SK snRNA via PI3K/Akt and activates HIV transcription. PLoS Pathog. 2007;3(10):1459–69.

53. Zhang X, Zhang J, Zhang L, van Dam H, ten Dijke P. UBE2O negatively regulates TRAF6-mediated NF-κB activation by inhibiting TRAF6 polyubiquitination. Cell Res [Internet]. Nature Publishing Group; 2013;23(3):366–77. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3587711&tool=pmcentrez&rendertype=abstract

54. Thorslund T, Ripplinger A, Hoffmann S, Wild T, Uckelmann M, Villumsen B, et al. Histone H1 couples initiation and amplification of ubiquitin signalling after DNA damage. Nature [Internet]. 2015;527(7578):389–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26503038

55. Yuan WC, Lee YR, Lin SY, Chang LY, Tan YP, Hung CC, et al. K33-Linked Polyubiquitination of Coronin 7 by Cul3-KLHL20 Ubiquitin E3 Ligase Regulates Protein Trafficking. Mol Cell [Internet]. Elsevier Inc.; 2014;54(4):586–600. Available from: http://dx.doi.org/10.1016/j.molcel.2014.03.035

56. Gatti M, Pinato S, Maiolica A, Rocchio F, Prato M, Aebersold R, et al. RNF168 promotes noncanonical K27ubiquitination to signal DNA damage. Cell Rep [Internet]. The Authors; 2015;10(2):226–38. Available from: http://dx.doi.org/10.1016/j.celrep.2014.12.021

57. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer D V. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. Cell. 2005;122(2):169–82.

58. Calo E, Flynn RA, Martin L, Spitale RC, Chang HY, Wysocka J. RNA helicase DDX21 coordinates transcription and ribosomal RNA processing. Nature [Internet]. Nature Publishing Group; 2014;518(7538):249–53. Available from: http://www.nature.com/nature/journal/v518/n7538/full/nature13923.html?WT.ec_id=NATURE-20150212

59. Hargreaves DC, Horng T, Medzhitov R. Control of Inducible Gene Expression by Signal-Dependent Transcriptional Elongation. Cell [Internet]. Elsevier Ltd; 2009;138(1):129–45. Available from: http://dx.doi.org/10.1016/j.cell.2009.05.047

60. Moon KJ, Mochizuki K, Zhou M, Jeong HS, Brady JN, Ozato K. The bromodomain protein Brd4 is a positive regulatory component of P-TEFb and stimulates RNA polymerase II-dependent transcription. Mol Cell. 2005;19(4):523–34.

61. Rowe HM, Jakobsson J, Mesnard D, Rougemont J, Reynard S, Aktas T, et al. KAP1 controls endogenous retroviruses in embryonic stem cells. Nature [Internet]. 2010;463(7278):237–40. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20075919\nhttp://www.nature.com/nature/journal/v463/n7278/pdf/nature08674.pdf

62. Jaeger S, Gulbahce N, Cimermancic P, Kane J, He N, Chou S, et al. Purification and characterization of HIV-human protein complexes. Methods. 2011;53(1):13–9.

63. Davis ZH, Verschueren E, Jang GM, Kleffman K, Johnson JR, Park J, et al. Global mapping of herpesvirus-host protein complexes reveals a transcription strategy for late genes. Mol Cell. 2015;57(2):349–60.

64. Mirrashidi KM, Elwell CA, Verschueren E, Johnson JR, Frando A, Von Dollen J, et al. Global mapping of the inc-human interactome reveals that retromer restricts chlamydia infection. Cell Host Microbe. 2015;18(1):109–21.

65. Kane JR, Stanley DJ, Hultquist JF, Johnson JR, Mietrach N, Binning JM, et al. Lineage-Specific Viral Hijacking of Non-canonical E3Ubiquitin Ligase Cofactors in the Evolution of Vif Anti-APOBEC3 Activity. Cell Rep. 2015;11(8):1236–50.

66. Cox J, Mann M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat Biotechnol [Internet]. 2008;26(12):1367–72. Available from: http://dx.doi.org/10.1038/nbt.1511\nhttp://www.ncbi.nlm.nih.gov/pubmed/19029910

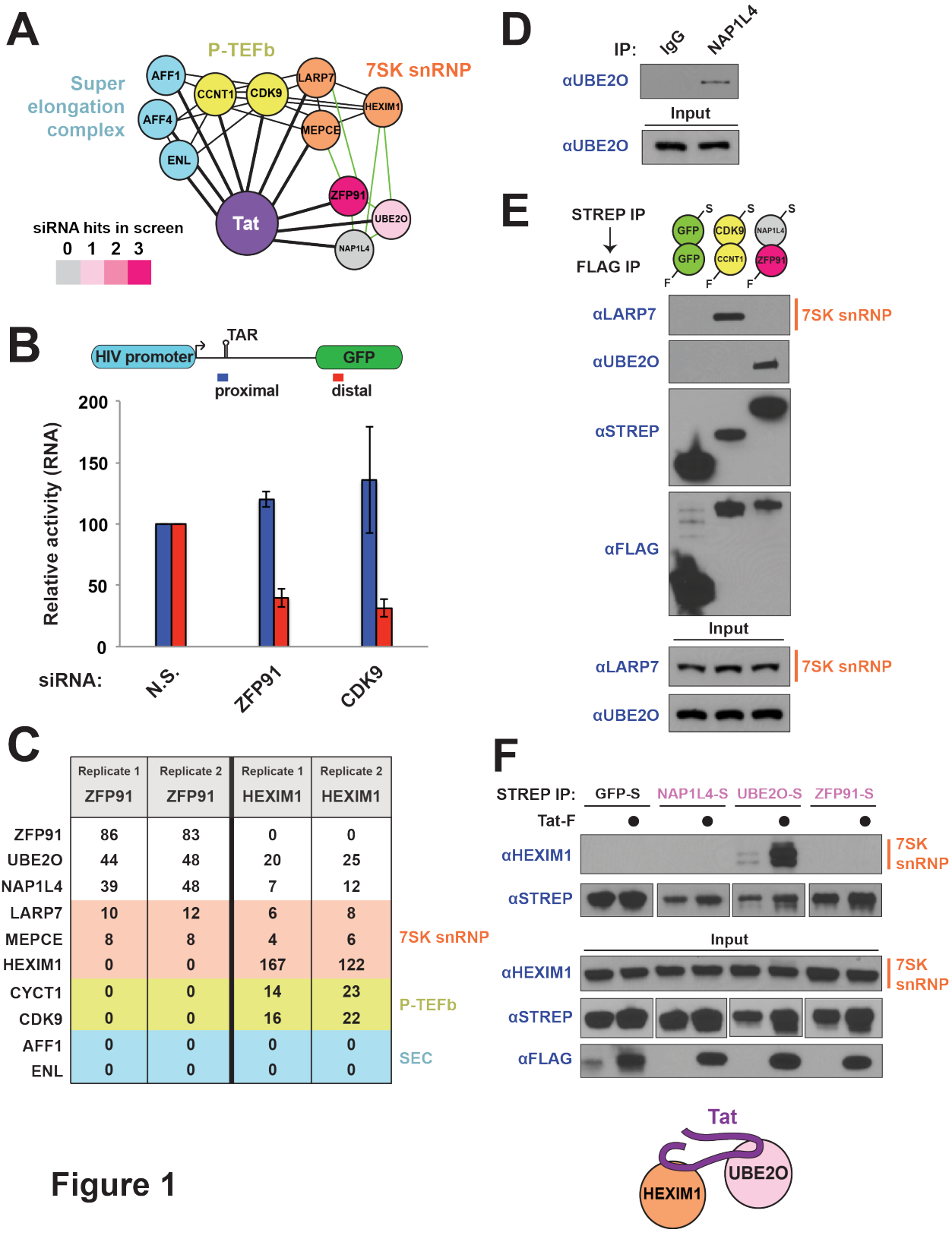

A
P-TEFb
7SK snRNP
Super elongation complex
AFF1
AFF4
ENL
CCNT1
CDK9
LARP7
HEXIM1
MEPCE
ZFP91
UBE2O
NAP1L4
Tat
siRNA hits in screen
0 1 2 3
B
TAR
HIV promoter
GFP
proximal
distal
Relative activity (RNA)
200
150
100
50
0
siRNA:
N.S.
ZFP91
CDK9
C
Replicate 1 ZFP91
Replicate 2 ZFP91
Replicate 1 HEXIM1
Replicate 2 HEXIM1
ZFP91 86 83 0 0
UBE2O 44 48 20 25
NAP1L4 39 48 7 12
LARP7 10 12 6 8
MEPCE 8 8 4 6
HEXIM1 0 0 167 122
CYCT1 0 0 14 23
CDK9 0 0 16 22
AFF1 0 0 0 0
ENL 0 0 0 0
7SK snRNP
P-TEFb
SEC
D
IP: IgG NAP1L4
αUBE2O
Input
αUBE2O
E
STREP IP
FLAG IP
GFP
CDK9
NAP1L4
GFP
CCNT1
ZFP91
S
F
αLARP7
7SK snRNP
αUBE2O
αSTREP
αFLAG
Input
αLARP7
7SK snRNP
αUBE2O
F
STREP IP: GFP-S NAP1L4-S UBE2O-S ZFP91-S
Tat-F
αHEXIM1
7SK snRNP
αSTREP
Input
αHEXIM1
7SK snRNP
αSTREP
αFLAG
Tat
HEXIM1
UBE2O

Figure 1

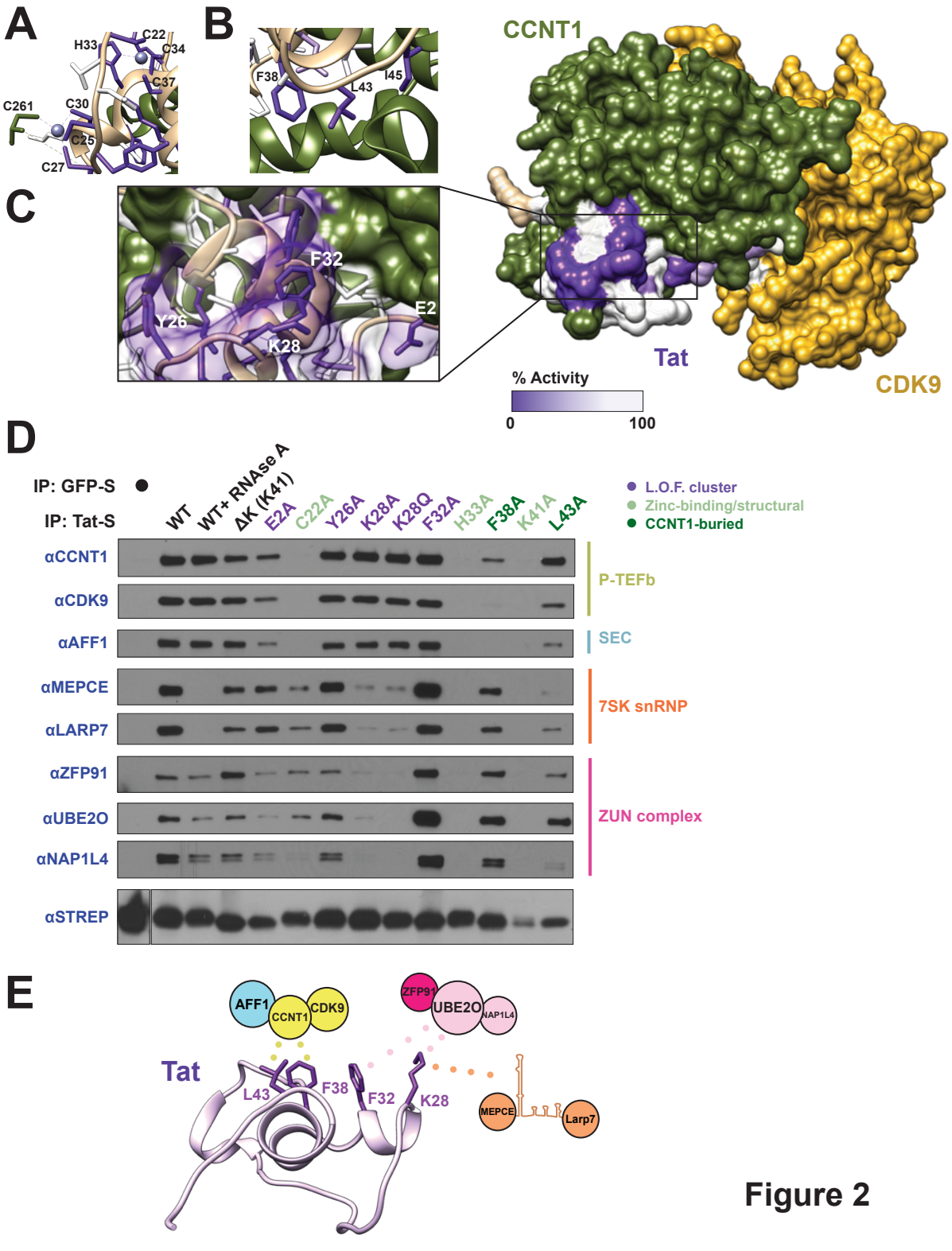

A
H33
C22
C34
C37
C261
C30
C25
C27
B
F38
L43
I45
CCNT1
C
F32
E2
Y26
K28
Tat
CDK9
% Activity
0
100
D
IP: GFP-S
IP: Tat-S
WT
WT+ RNAse A
ΔK (K41)
E2A
C22A
Y26A
K28A
K28Q
F32A
H33A
F38A
K41A
L43A
L.O.F. cluster
Zinc-binding/structural
CCNT1-buried
αCCNT1
αCDK9
αAFF1
αMEPCE
αLARP7
αZFP91
αUBE2O
αNAP1L4
αSTREP
P-TEFb
SEC
7SK snRNP
ZUN complex
E
AFF1
CCNT1
CDK9
ZFP91
UBE2O
NAP1L4
Tat
L43
F38
F32
K28
MEPCE
Larp7

Figure 2

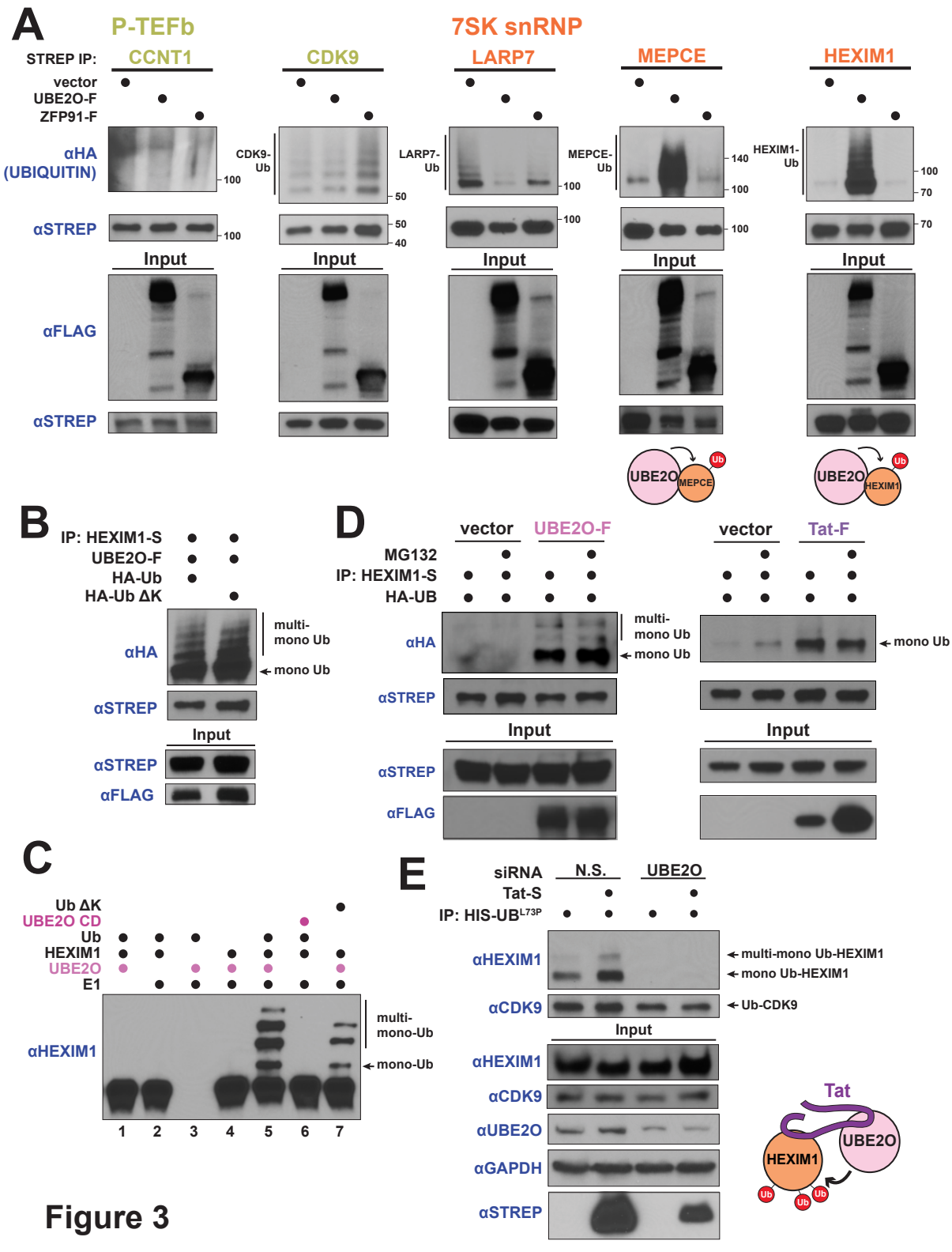

A
P-TEFb
7SK snRNP
STREP IP:
CCNT1
CDK9
LARP7
MEPCE
HEXIM1
vector
UBE2O-F
ZFP91-F
αHA (UBIQUITIN)
CDK9-Ub
LARP7-Ub
MEPCE-Ub
HEXIM1-Ub
αSTREP
Input
αFLAG
UBE2O
MEPCE
HEXIM1
B
IP: HEXIM1-S
UBE2O-F
HA-Ub
HA-Ub ΔK
αHA
multi-mono Ub
mono Ub
αSTREP
Input
αFLAG
C
Ub ΔK
UBE2O CD
Ub
HEXIM1
UBE2O
E1
αHEXIM1
multi-mono-Ub
mono-Ub
1 2 3 4 5 6 7
D
vector
UBE2O-F
Tat-F
MG132
IP: HEXIM1-S
HA-UB
αHA
αSTREP
Input
αFLAG
E
siRNA
N.S.
UBE2O
Tat-S
IP: HIS-UBL73P
αHEXIM1
multi-mono Ub-HEXIM1
mono Ub-HEXIM1
αCDK9
Ub-CDK9
αUBE2O
αGAPDH
Tat

Figure 3

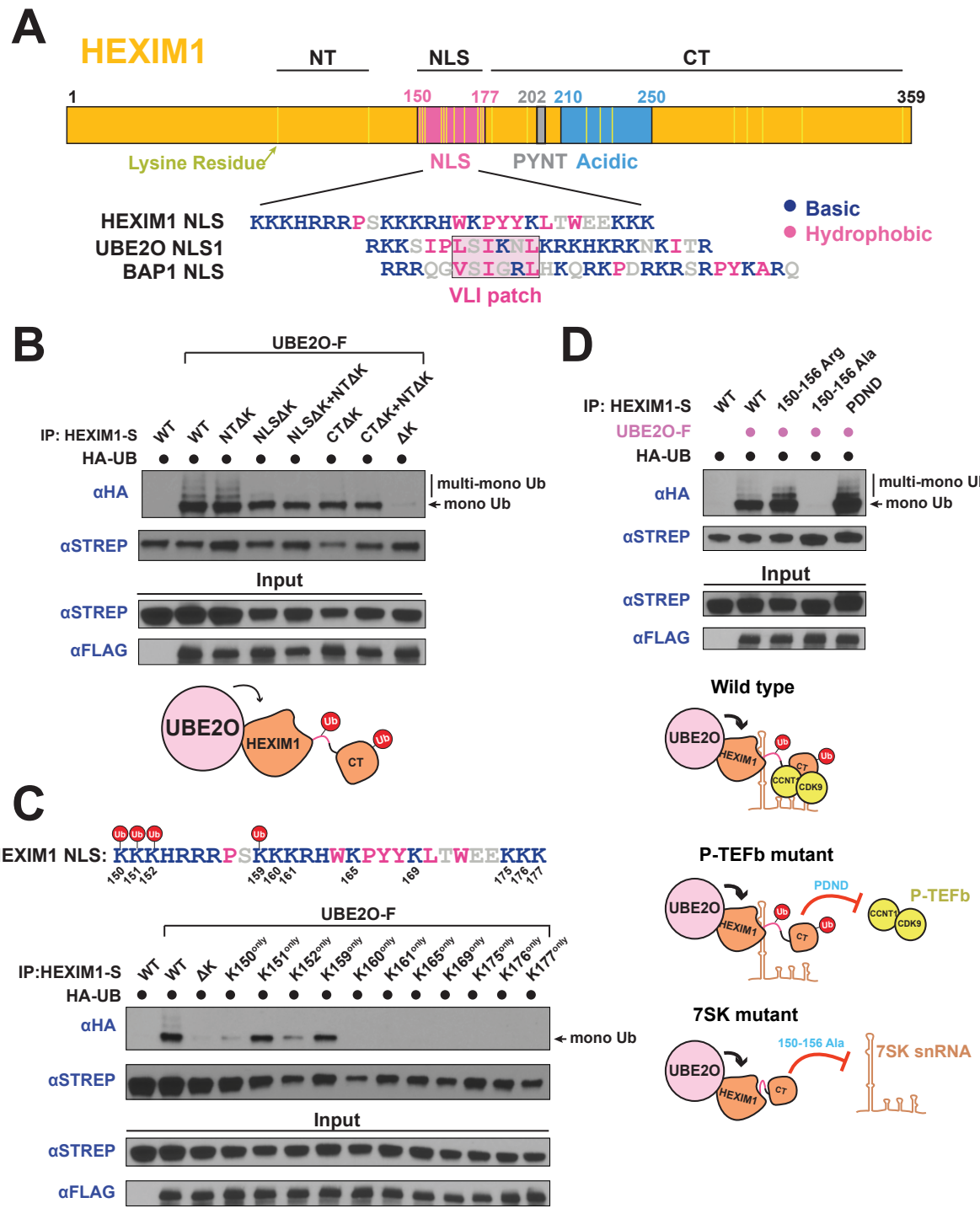

A
HEXIM1
NT
NLS
CT
1
150
177
202
210
250
359
Lysine Residue
NLS
PYNT
Acidic
HEXIM1 NLS
UBE2O NLS1
BAP1 NLS
KKKHRRRPSKKKRHWKPYYKLTWEEKKK
RKKSIPLSIKNLKRKHKRKNKITR
RRRQGVSIGRLHKQRKPDRKRSRPYKARQ
VLI patch
Basic
Hydrophobic
B
UBE2O-F
IP: HEXIM1-S
WT
WT
NTΔK
NLSΔK
NLSΔK+NTΔK
CTΔK
CTΔK+NTΔK
ΔK
HA-UB
αHA
multi-mono Ub
mono Ub
αSTREP
Input
αSTREP
αFLAG
UBE2O
HEXIM1
Ub
CT
C
HEXIM1 NLS:
KKKHRRRPSKKKRHWKPYYKLTWEEKKK
150
151
152
159
160
161
165
169
175
176
177
UBE2O-F
IP:HEXIM1-S
WT
WT
ΔK
K150only
K151only
K152only
K159only
K160only
K161only
K165only
K169only
K175only
K176only
K177only
HA-UB
αHA
mono Ub
αSTREP
Input
αSTREP
αFLAG
D
IP: HEXIM1-S
WT
WT
150-156 Arg
150-156 Ala
PDND
UBE2O-F
HA-UB
αHA
multi-mono Ub
mono Ub
αSTREP
Input
αSTREP
αFLAG
Wild type
UBE2O
HEXIM1
Ub
CT
CCNT1
CDK9
P-TEFb mutant
PDND
P-TEFb
7SK mutant
150-156 Ala
7SK snRNA

Figure 4

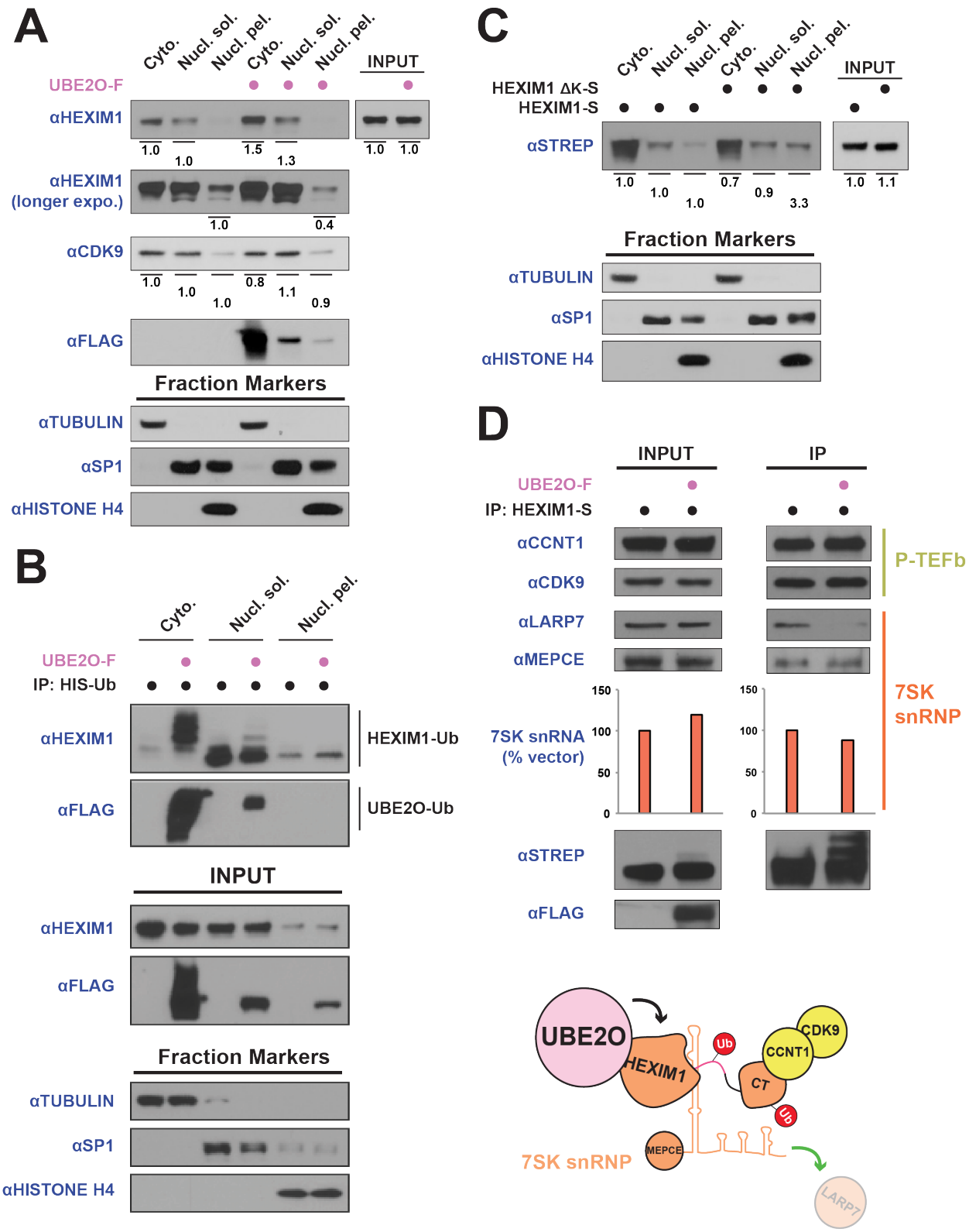

A
Cyto.
Nucl. sol.
Nucl. pel.
Cyto.
Nucl. sol.
Nucl. pel.
INPUT
UBE2O-F
αHEXIM1
αHEXIM1 (longer expo.)
αCDK9
αFLAG
Fraction Markers
αTUBULIN
αSP1
αHISTONE H4
B
Cyto.
Nucl. sol.
Nucl. pel.
UBE2O-F
IP: HIS-Ub
αHEXIM1
αFLAG
HEXIM1-Ub
UBE2O-Ub
INPUT
αHEXIM1
αFLAG
Fraction Markers
αTUBULIN
αSP1
αHISTONE H4
C
Cyto.
Nucl. sol.
Nucl. pel.
Cyto.
Nucl. sol.
Nucl. pel.
INPUT
HEXIM1 ΔK-S
HEXIM1-S
αSTREP
Fraction Markers
αTUBULIN
αSP1
αHISTONE H4
D
INPUT
IP
UBE2O-F
IP: HEXIM1-S
αCCNT1
αCDK9
αLARP7
αMEPCE
P-TEFb
7SK snRNP
7SK snRNA (% vector)
αSTREP
αFLAG
UBE2O
HEXIM1
Ub
CCNT1
CDK9
CT
MEPCE
7SK snRNP
LARP7

Figure 5

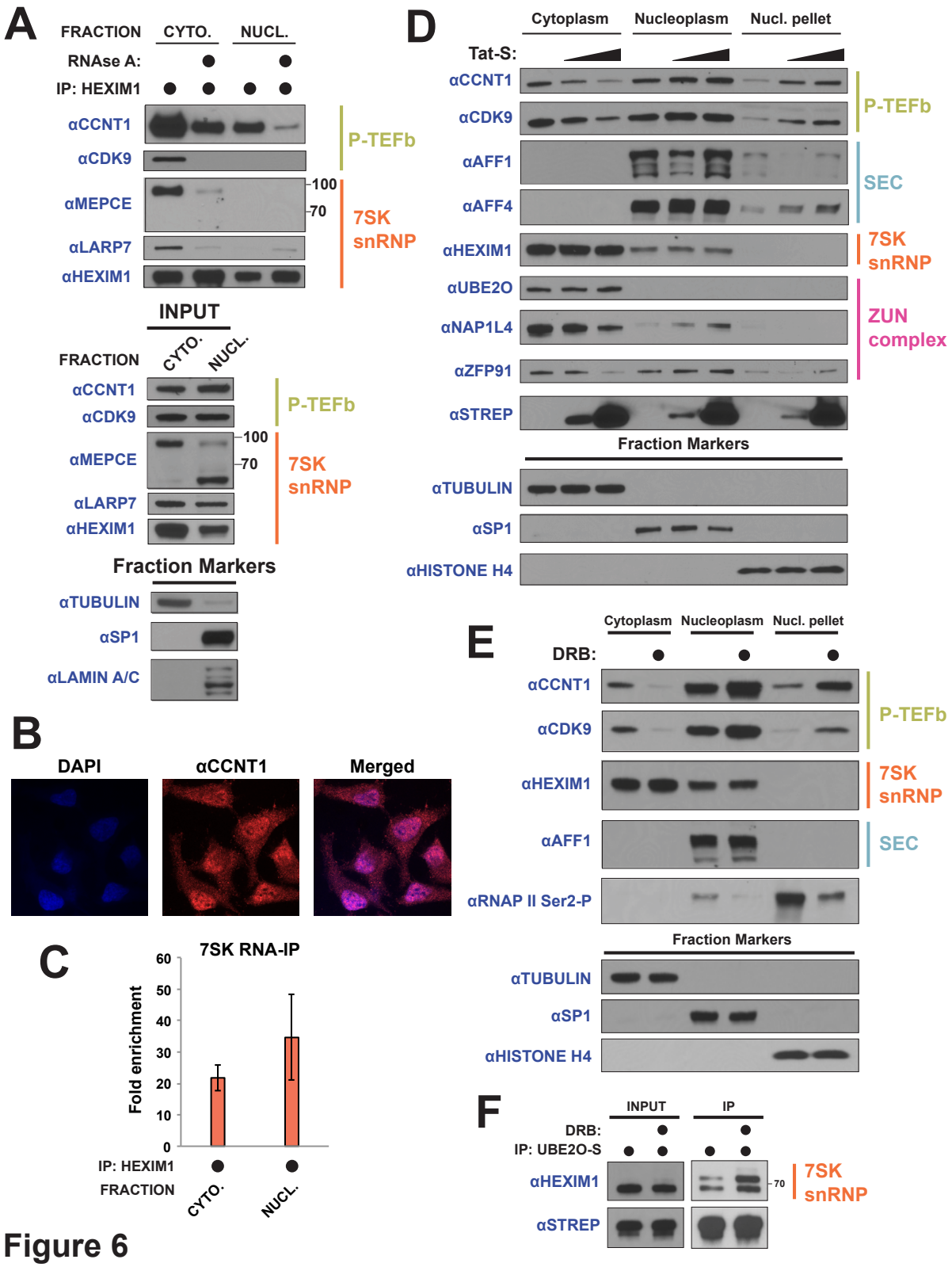

A
FRACTION
CYTO.
NUCL.
RNAse A:
IP: HEXIM1
αCCNT1
αCDK9
αMEPCE
αLARP7
αHEXIM1
P-TEFb
7SK snRNP
100
70
INPUT
FRACTION
CYTO.
NUCL.
αCCNT1
αCDK9
αMEPCE
αLARP7
αHEXIM1
P-TEFb
7SK snRNP
100
70
Fraction Markers
αTUBULIN
αSP1
αLAMIN A/C
B
DAPI
αCCNT1
Merged
C
7SK RNA-IP
Fold enrichment
60
50
40
30
20
10
0
IP: HEXIM1
FRACTION
CYTO.
NUCL.
D
Cytoplasm
Nucleoplasm
Nucl. pellet
Tat-S:
αCCNT1
αCDK9
αAFF1
αAFF4
αHEXIM1
αUBE2O
αNAP1L4
αZFP91
αSTREP
P-TEFb
SEC
7SK snRNP
ZUN complex
Fraction Markers
αTUBULIN
αSP1
αHISTONE H4
E
Cytoplasm
Nucleoplasm
Nucl. pellet
DRB:
αCCNT1
αCDK9
αHEXIM1
αAFF1
αRNAP II Ser2-P
P-TEFb
7SK snRNP
SEC
Fraction Markers
αTUBULIN
αSP1
αHISTONE H4
F
INPUT
IP
DRB:
IP: UBE2O-S
αHEXIM1
αSTREP
70
7SK snRNP

Figure 6

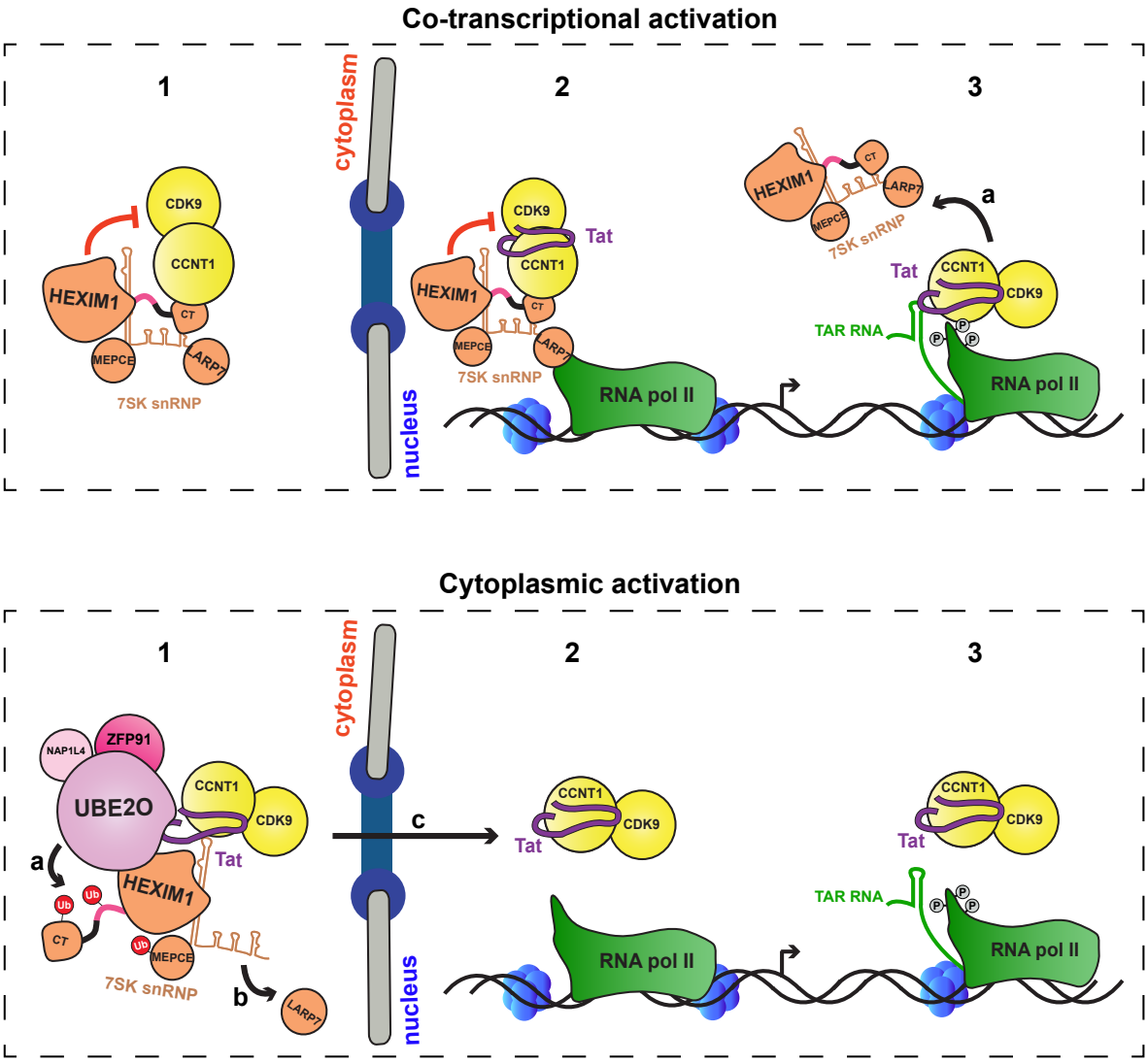

Co-transcriptional activation
1
2
3
cytoplasm
nucleus
CDK9
CCNT1
HEXIM1
CT
MEPCE
LARP7
7SK snRNP
Tat
RNA pol II
TAR RNA
a
Cytoplasmic activation
NAP1L4
ZFP91
UBE2O
Ub
b
c

Figure 7

Figure Legends

Figure 1: ZFP91 is a critical regulator of Tat-dependent transcription elongation and forms a complex with the Tat interactors UBE2O and NAP1L4. (A) Network representation of known Tat interactions (thick black edges) with P-TEFb, 7SK snRNP, the super elongation complex, and select host interactors from our recent proteomic screen (23). Thin black edges represent known host-host interactions and green edges indicate interactions that are detailed in the current manuscript. Gray and pink shading indicate the number of positive siRNA hits in a recent functional screen for the select interactors (*unpublished data)*. (B) HIV RNA elongation assay with ZFP91 (siRNA 2 from screen, unpublished data) and CDK9 knockdowns in HeLa$^{provirus\Delta tat}$ cell line, which contains GFP cloned into the Nef locus. qPCR was used to monitor the production of proximal and distal RNA transcripts, with the relative positions of the primer sets shown above. Data are presented as the mean -/+ S.E.M. of relative Tat activity (indicated siRNA/ N.S. siRNA) of at least 3 biological replicates. (C) AP-MS results using ZFP91-FLAG (two left columns) or HEXIM1-FLAG (two right columns) proteins as baits. Peptide counts are shown for selected prey proteins from duplicate biological purifications. See also Table S1 and Fig. S1. (D) Endogenous NAP1L4 IP detects an interaction with UBE2O, with IgG as a background control. (E) Double IP of transfected STREP- and FLAG-tagged proteins as indicated (-S denotes STREP while –F denotes FLAG). GFP contained a tandem STREP-FLAG tag. A STREP IP was performed first, followed by FLAG purification. Eluates from the FLAG IP were probed by Western blot using anti-LARP7 or anti-UBE2O antibodies. (F) STREP purification of transfected proteins -/+ Tat co-transfection. The endogenous HEXIM1 interaction was monitored by

Western blot using an anti-HEXIM1 antibody. Schematic illustrates that Tat enhances the interaction between UBE2O and HEXIM1. All STREP blots are from the same exposure but are cropped for presentation.

Figure 2: Tat uses a novel interaction surface to interact with the 7SK snRNP and ZUN complex. (A) Tat transcriptional activity data collected from our systematic set of alanine point mutants (32) was mapped onto the crystal structure of the Tat-P-TEFb complex (PDB i.d. 3MI9). The most severe loss-of-function (L.O.F.) mutations are shaded in deep purple while those with no effect on activity are shaded in white. The cysteine-rich zinc-binding clusters (A) and a hydrophobic stretch buried in CCNT1 (B) are critically important for Tat function. (C) A surface-exposed patch of severe loss-of-function mutants (Glu2, Tyr26, Lys28, and Phe32) does not contact CCNT1 or CDK9. (D) IP of transiently transfected wild type or mutant STREP-tagged Tat proteins. GFP-S was used as a negative control. Western blots were performed on STREP elutions using antibodies to the endogenous proteins indicated. Important residues are color coded as indicated. L.O.F. cluster is the four surface-exposed mutants from (C). See also Fig. S2 for inputs. (E) Model of Tat interaction surfaces. Tat uses a hydrophobic patch including Leu43 and Phe38 buried at the CCNT1 interface, which also results in the binding of CDK9 and AFF1. Mutation of Lys28 reduced the interactions with both the 7SK snRNP and ZUN proteins, whereas mutation of Phe32 dramatically increases the UBE2O interaction, associating the ligase with this residue.

Figure 3: UBE2O multi-monoubiquitinates HEXIM1 *in vivo* and *in vitro*. (A) *In vivo* ubiquitination of transfected, STREP-tagged P-TEFb and 7SK snRNP proteins co-transfected with HA-ubiquitin -/+ co-transfection of UBE2O-F or ZFP91-F (-F designates

FLAG tagged protein). Denaturing lysis and IPs were performed to detect ubiquitination by HA Western blot. Ubiquitinated species are indicated. Schematics illustrate that UBE2O increases ubiquitination of the 7SK snRNP proteins MEPCE and HEXIM1. (B) *In vivo* ubiquitination of HEXIM1-S in HEK 293T cells co-transfected with UBE2O-F and either wild type or lysine-free (ΔK) HA-ubiquitin. Monoubiquitinated HEXIM1 and additional ubiquitinated species are indicated. (C) *In vitro* ubiquitination of HEXIM1 by UBE2O. HEXIM1-S purified from HEK 293T cells was incubated with recombinant E1 and ubiquitin, UBE2O-F purified from HEK 293T cells, and ATP. Control reactions lacked individual components as indicated. Additionally, one reaction was incubated with UBE2O CD (catalytically dead) purified from HEK 293T cells in place of wild-type UBE2O (lane 6) or recombinant ubiquitin ΔK in place of wild-type ubiquitin (lane 7). See also Fig. S3A, B. (D) *In vivo* ubiquitination of HEXIM1-S co-transfected with UBE2O-F or Tat-F and subsequently treated with DMSO control or MG132 to block proteasomal degradation. See also Fig. S3 C. (E) HEK 293T cells were transfected with non-silencing or UBE2O-directed siRNA. Knockdown cells were then transfected with HIS-Ubiquitin L73P followed by a denaturing nickel purification. Ubiquitinated species were detected by Western blot using specific antibodies (anti-CDK9, anti-HEXIM1) on the nickel-purified elution and are indicated. Schematic illustrates that Tat enhances HEXIM1 ubiquitination through UBE2O.

Figure 4: UBE2O ubiquitinates the NLS of 7SK snRNP-bound HEXIM1. (A) Domain organization of HEXIM1. The positions of the 23 lysines in HEXIM1 are indicated by yellow bars in the C-terminal (CT) and N-terminal (NT) regions and the NLS. (Below) Alignment of the NLS sequences from HEXIM1, UBE2O, and a known UBE2O

substrate, BAP1 (21). UBE2O and several of its target proteins contain a consensus bipartite NLS architecture that shares similarities with the HEXIM1 NLS. Basic residues are labeled in blue while hydrophobic residues are labeled in pink. (B) *In vivo* ubiquitination of HEXIM1 lysine mutants. Combinations of STREP-tagged lysine-to-arginine mutants were generated in the HEXIM1 N-terminal region (NTΔK), NLS (NLSΔK), and C-terminal region (CTΔK). A complete lysine-free HEXIM1 is designated as ΔK. UBE2O-F was co-transfected where indicated, and ubiquitination was detected by anti-HA Western after denaturing lysis and IP. Monoubiquitin and multi-monoubiquitin species are indicated. The NTΔK mutant had no effect on ubiquitination, suggesting that there are no acceptor sites in the N-terminal region. However, both the NLSΔK and CTΔK mutants displayed decreased ubiquitination, demonstrating the presence of acceptor lysines in these regions. Schematic illustrates that lysines in the NLS (pink connection between domains) and in the C-terminal region function as acceptor sites. See also Fig. S4 and Table S2. (C) *In vivo* ubiquitination of individual lysine residues in the HEXIM1 NLS re-introduced in the context of the lysine-free (ΔK) HEXIM1 background. Ubiquitination of single-lysine HEXIM1 proteins was detected by HA Western blot. (top) All lysines are numbered in the HEXIM1 NLS and those that restore ubiquitination are indicated. (D) *In vivo* ubiquitination of wild type STREP-tagged HEXIM1 or mutants deficient in binding P-TEFb (PDND) or the 7SK snRNA (150-156 Ala). HEXIM1-S was co-transfected with UBE2O-F as indicated. Schematic below illustrates the effects of these mutations on complex formation and ubiquitination.

Figure 5: UBE2O influences HEXIM1 subcellular localization and remodels the 7SK snRNP by ubiquitin-dependent loss of the HEXIM1-LARP7 interaction. (A)

Subcellular localization of endogenous HEXIM1. HEK 293T were transfected -/+ 3 μg UBE20-F plasmid and fractionated into cytoplasmic (Cyto.), soluble nuclear (Nucl. sol.), and nuclear pellet (Nucl. pel.) fractions. Numbers represent quantified pixel intensity for HEXIM1 and CDK9 bands relative to un-transfected cells for the same fraction. Tubulin, SP1, and histone H4 were used as markers for each fraction and show the quality of the separations. All HEXIM1 and CDK9 bands were first normalized to the pixel intensity of appropriate fraction controls (i.e. Cyto. normalized to tubulin, Nucl. sol. normalized to SP1, Nucl. pel. normalized to histone H4) prior to comparison between treatments. Two exposures of HEXIM1 from the same blot are shown. (B) Subcellular fractionation of HEK 293T cells transfected with HIS-Ubiquitin. UBE2O-F was co-transfected as indicated and denaturing nickel pull-downs were used to detect HEXIM1 ubiquitination by anti-HEXIM1 Western in the various subcellular fractions. Ubiquitinated species are indicated. (C) Subcellular localization of 0.25 μg of transfected HEXIM1-S or HEXIM1 ΔK-S plasmid in HEK 293T cells. Numbers represent quantified pixel intensity for HEXIM1 (STREP) bands relative to wild type expression for the same fraction. Bands were normalized as in (A). (D) IPs of HEXIM1-S -/+ UBE2O-F co-transfection in HEK 293T cells. Western blots were performed with the specific endogenous antibodies indicated or RT-qPCR was performed with 7SK snRNA primers to detect HEXIM1 protein-protein or protein-RNA interactions, respectively. Schematic illustrates that UBE2O ubiquitination of HEXIM1 decreases the LARP7 interaction, effectively ejecting LARP7 from the complex. See also Fig. S5, S6, and S7.

Figure 6: Tat and DRB release P-TEFb from a cytoplasmic pool of 7SK snRNP. (A) Endogenous HEXIM1 IP using cytoplasmic and soluble nuclear fractions. One sample

from each fraction was treated with RNAse A prior to the IP to evaluate RNA-dependence of interactions. (B) Immunofluorescence analysis of endogenous CCNT1 in HeLa$^{provirus\Delta tat}$ cells, demonstrating nuclear and cytoplasmic staining. Nuclei were stained with DAPI. (C) RNA-IP assay of the 7SK snRNA. An endogenous HEXIM1 IP was performed using cytoplasmic and soluble nuclear fractions. RNA was then purified from the IPs and 7SK snRNA was detected by RT-qPCR. Fold enrichments were calculated as relative amounts of 7SK snRNA [HEXIM1 IP/ rabbit IgG IP], with the average fold enrichments shown -/+ SEM (n=3). (D) HeLa cells were transfected with increasing amounts of Tat-S plasmid (0, 0.5 μg, or 5 μg) followed by subcellular fractionation, and Western blots of endogenous proteins were performed using the indicated antibodies. (E) HeLa cells were treated with 100 μM DRB or an equal volume of DMSO for 1 hour prior to subcellular fractionation and Western blotting. (F) HEK 293T cells were transfected with UBE2O-S. Cells were treated with 100 μM DRB or DMSO 1 hour prior to lysis and STREP IP. The endogenous HEXIM1 interaction was detected by anti-HEXIM1 Western.

Figure 7: Model of two-stage HIV-1 Tat transcriptional activation. (Top) Co-transcriptional activation. The inhibited P-TEFb-7SK snRNP complex resides in the cytoplasm (1) and bound to RNAP II at the viral promoter (2). During co-transcriptional activation, Tat assembles with the inhibited complex at the HIV-1 promoter (2). The synthesis of TAR RNA results in the displacement of the inhibitory snRNP (3a) to activate the kinase activity of CDK9 (3) (18). (Bottom) During cytoplasmic activation (1), Tat hijacks UBE2O to ubiquitinate HEXIM1 (1a), resulting in the eviction of LARP7 from the snRNP (1b) and ultimately the nuclear import of free P-TEFb (1c). The released,

active P-TEFb is then either recruited to transcription initiation complexes at the viral promoter (2) or the nascent TAR RNA (3) for RNAP II CTD phosphorylation. The nuclear import of free P-TEFb is also triggered by DRB treatment.

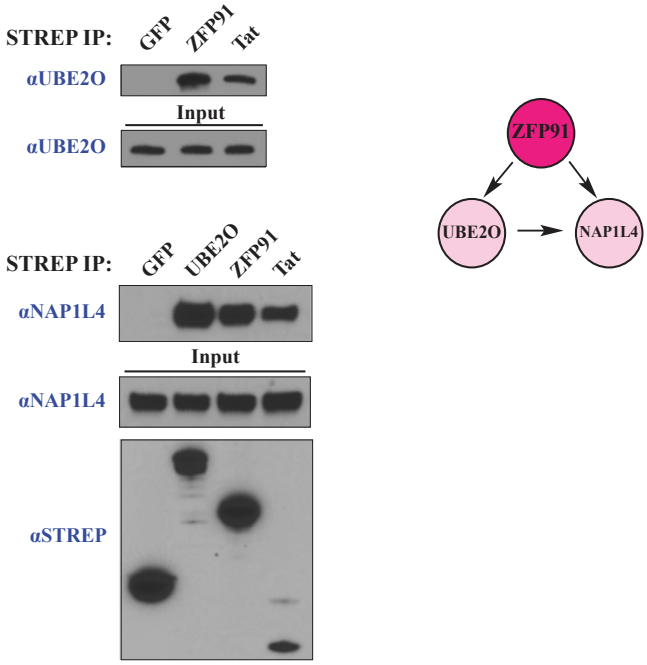


Figure S1: Reciprocal IPs confirm interactions between ZFP91, UBE2O, and NAP1L4.

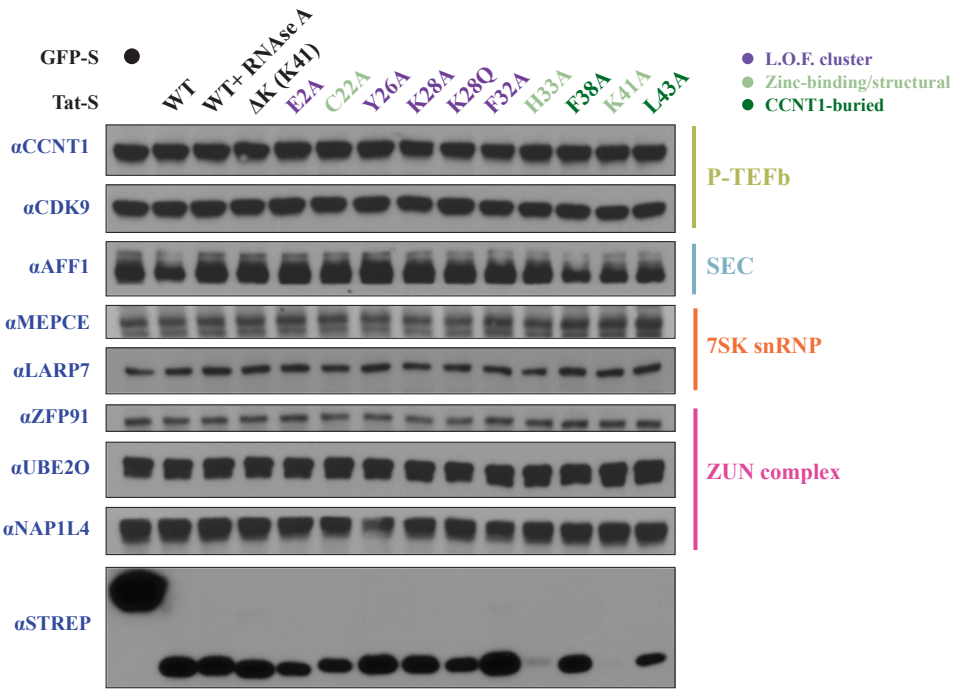


Figure S2: Protein input blots for Tat mutant IPs.

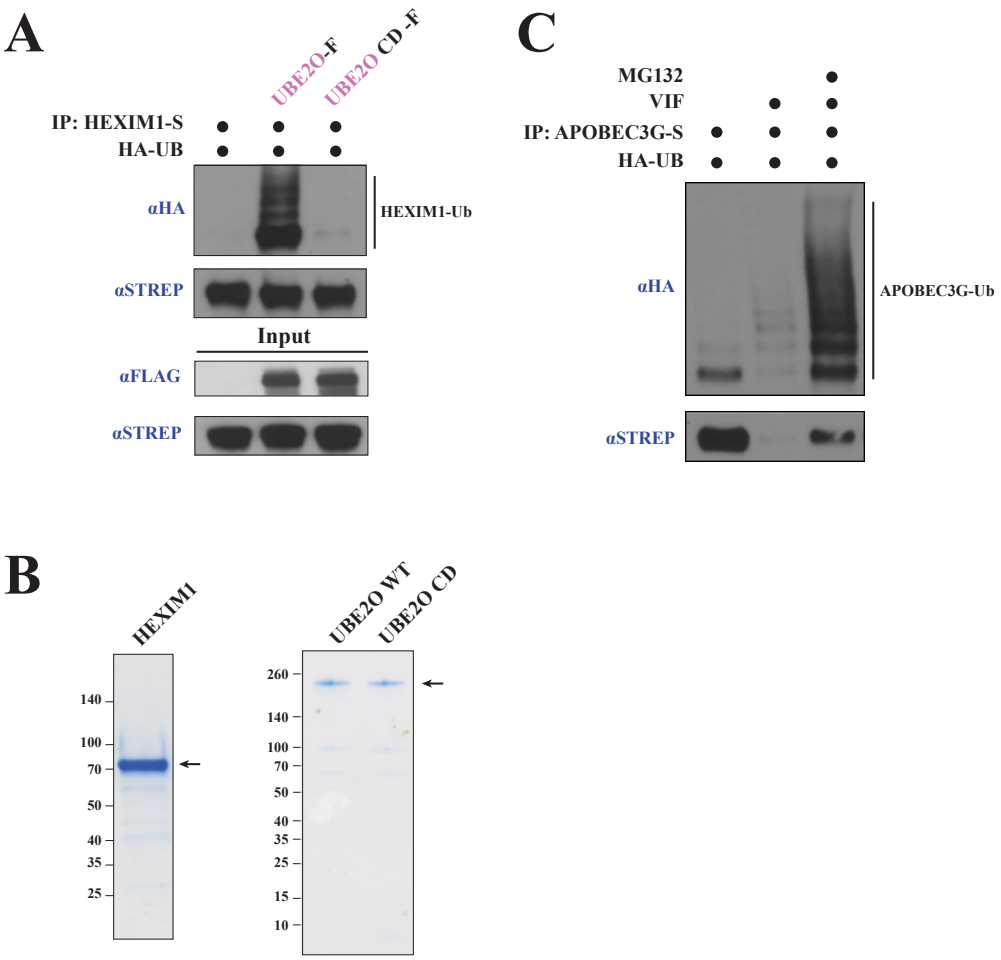


Figure S3: A catalytically inactive UBE2O does not induce HEXIM1 ubiquitination.

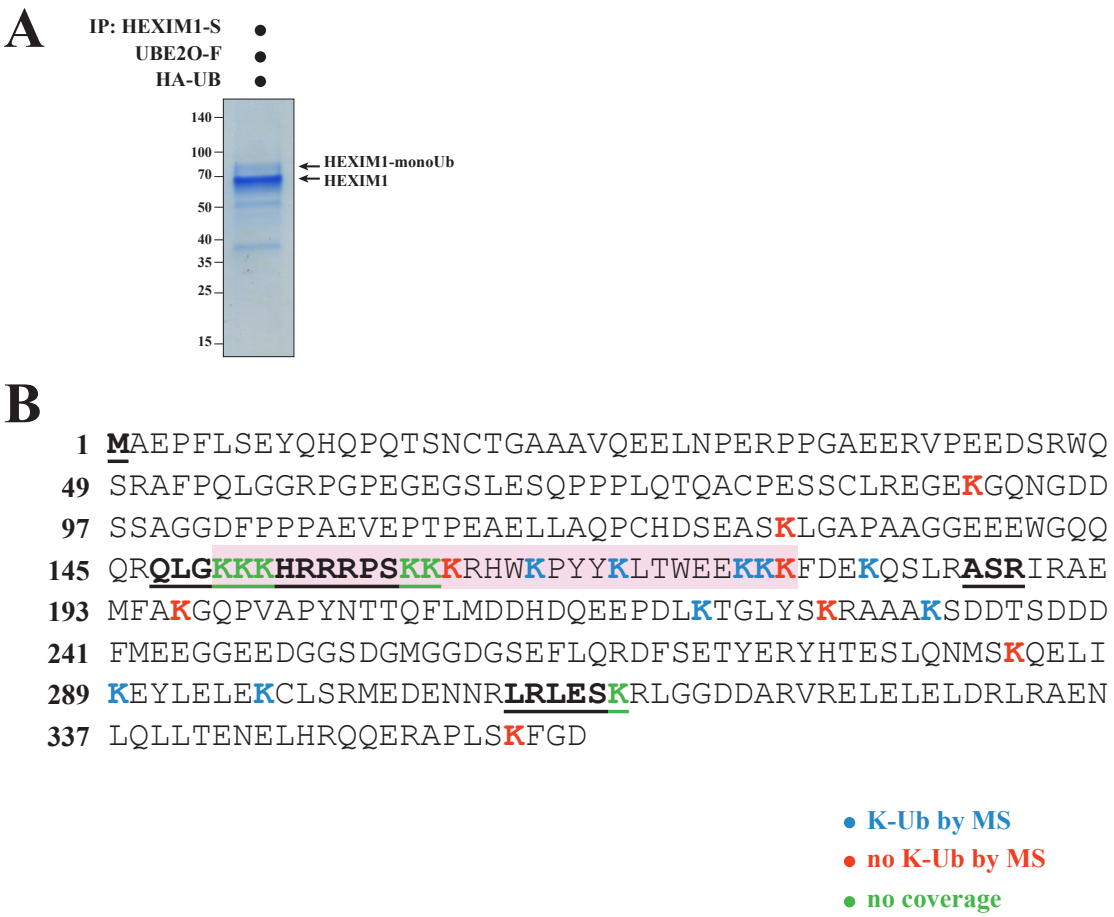


Figure S4: Mass spectrometry identification of ubiquitinated HEXIM1 lysines with UBE2O co-expression.

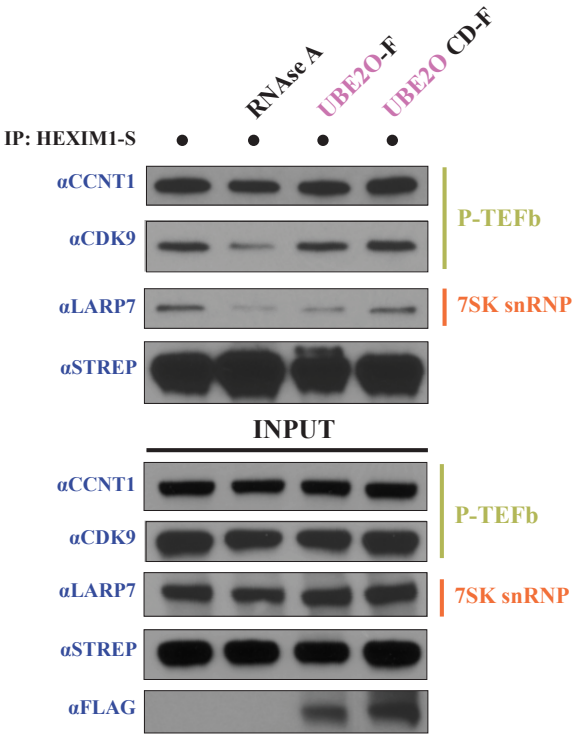


Figure S5: The catalytic activity of UBE2O is required to decrease the HEXIM1-LARP7 interaction.

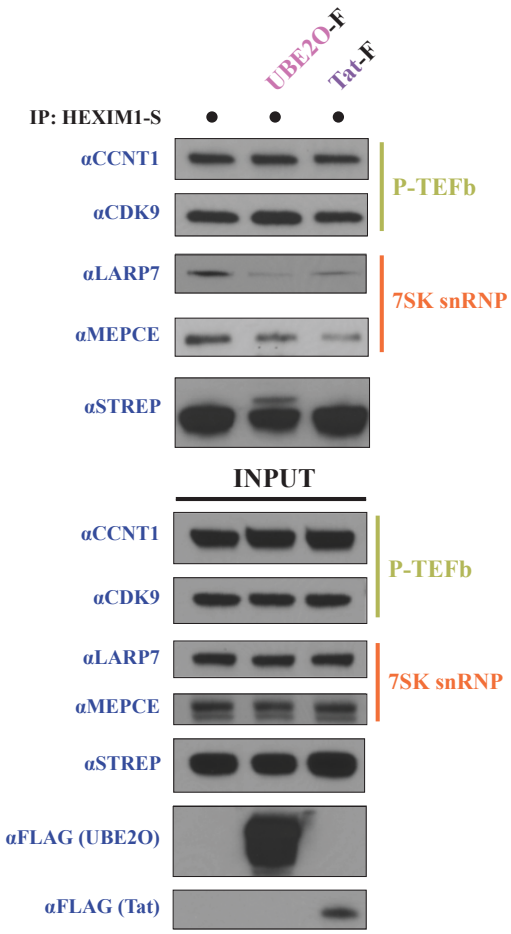


Figure S6: Tat expression decreases the HEXIM1-LARP7 interaction.

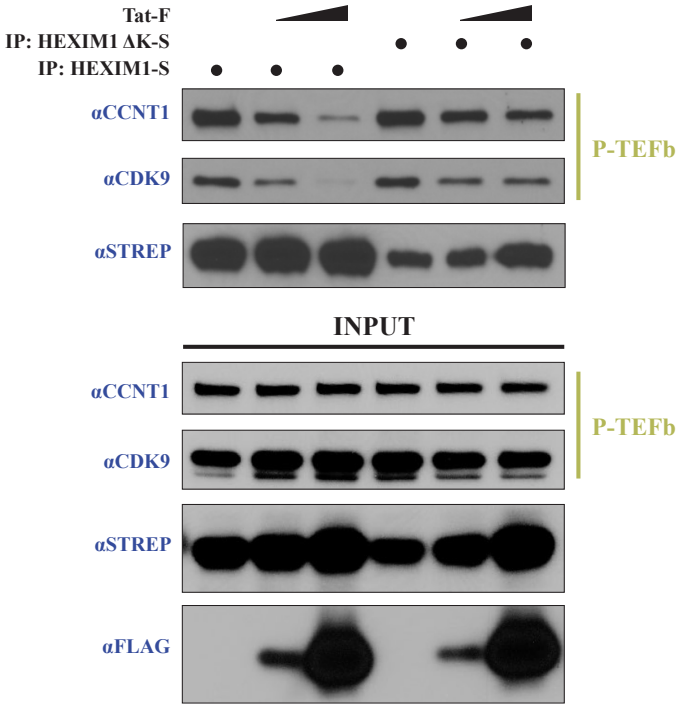


Figure S7:The HEXIM1 ΔK mutant is refractory to P-TEFb release by increasing Tat expression

Supplemental Figure Legends

Figure S1. Reciprocal IPs confirm interactions between ZFP91, UBE2O, and NAP1L4: STREP IP of transfected GFP-S, ZFP91-S, UBE2O-S, or Tat-S from HEK 293T cells (-S designates STREP tagged protein) to confirm interactions identified from mass spectrometry, using anti-UBE2O and anti-NAP1L4 antibodies. Schematic represents detected host protein interactions as arrows.

Figure S2. Protein input blots for Tat mutant IPs: Input material for STREP IPs analyzed with endogenous antibodies (Fig. 2D).

Figure S3. A catalytically inactive UBE2O does not induce HEXIM1 ubiquitination: (A) *In vivo* ubiquitination of transfected HEXIM1 with the co-transfection of wild-type or catalytically dead (CD) UBE2O from HEK 293T cells. HA Western blot indicates the ubiquitinated species. (B) Protein gels demonstrating purity of HEXIM1, UBE2O WT, and UBE2O CD purified from HEK 293T cells and used in the *in vitro* ubiquitination assay (Fig. 3C). Arrows indicate HEXIM1 or UBE2O species. (C) *In vivo* ubiquitination of APOBEC3G-S -/+ HIV-1 Vif co-transfection and -/+ MG132 treatment. This control indicates that MG132 treatment restores APOBEC3G protein levels from Vif-mediated degradation (Yu et al. 2003).

Figure S4. Mass spectrometry identification of ubiquitinated HEXIM1 lysines with UBE2O co-expression: (A) Protein gel showing HEXIM1-S purified under denaturing conditions from HEK 293T cells with the co-expression of UBE2O-F. Arrows indicate unmodified and monoubiquitinated HEXIM1 species. (B) Coverage of the HEXIM1 sequence by MS. Sequences not identified by MS are bolded and underlined. Shaded

in red is the HEXIM1 bipartite NLS. Color coding reflects whether a lysine was identified with a ubiquitin di-glycine tryptic remnant fragment.

Figure S5. The catalytic activity of UBE2O is required to decrease the HEXIM1-LARP7 interaction: STREP IP of transiently transfected HEXIM1-S. Wild-type or catalytically dead (CD) UBE2O-F was co-transfected as indicated. One HEXIM1-S sample was treated with RNAse A to determine if the protein interactions were RNA-dependent. Western blotting of the eluate was used to detect interactions with endogenous proteins.

Figure S6. Tat expression decreases the HEXIM1-LARP7 interaction: HEK 293T cells were transiently transfected with HEXIM1-S alone or co-transfected with UBE2O-F or Tat-F and Western blots were performed with endogenous antibodies as indicated to assess protein-protein interactions.

Figure S7. The HEXIM1 ΔK mutant is refractory to P-TEFb release by increasing Tat expression: HeLa cells were transfected with 0.5 μg HEXIM1-S or HEXIM1 ΔK-S and either 0, 0.5, or 3 μg of Tat-F. A STREP IP of transfected HEXIM1 proteins was used to assess the interaction with endogenous P-TEFb by Western.

Chapter 5: Discussion

Genome-wide methods to interrogate transcription, including ChIP-seq and global run-on (GRO)-seq, have shown that the majority genes in higher eukaryotes contain stably paused RNAP II molecules immediately downstream of transcription start sites (1–4). HIV-1 gene expression is similarly controlled at the level of transcription elongation (5), and has therefore served as an excellent model to study this regulatory phenomenon. In a landmark paper, it was demonstrated CCNT1 directly binds Tat and is required for proper binding of the TAR RNA (6). We now know that CCNT1 is one protein of the heterodimeric P-TEFb complex, which regulates the majority of RNAP II pause release genome-wide (1). Similarly for the 7SK snRNP and super elongation complex (7–9), studying Tat-dependent HIV-1 transcription has revealed that these transcription complexes are employed more broadly in the cell for global regulation. This thesis expands these insights by revealing that multiple proteins in ubiquitin conjugation pathways are important for full Tat activity. One ubiquitin ligase in particular, UBE2O, appears to function more broadly as it ubiquitinates the HEXIM1 protein of the 7SK snRNP independently of Tat. This modification functions to partially disassemble the snRNP complex and further sequesters HEXIM1 in the cytoplasm.

Given the importance of this HEXIM1 modification to more global regulatory mechanisms, it will be critical to understand the molecular mechanism of this modification by reconstituting UBE2O-dependent ubiquitination of HEXIM1 *in vitro.* Experiments *in vivo* suggest that UBE2O preferentially modifies snRNP-bound HEXIM1, and this could be further evaluated *in vitro* by comparing the ubiquitination of free HEXIM1 versus HEXIM1 bound to the 7SK snRNA. One exciting possibility is that RNA-binding alters the conformation of HEXIM1 such that it becomes accessible for UBE2O,

a mechanism that ensures that the ligase specifically controls the regulatory snRNP complex. It will also be worth exploring the disassembly of the snRNP *in vitro.* Results from the thesis show that UBE2O ubiquitination inhibits the interaction between HEXIM1 and LARP7. The 7SK snRNP is a dynamic complex that responds to many stimuli to release P-TEFb for transcriptional activation (10). It is unclear if HEXIM1 ubiquitination directly evicts LARP7 from the complex or whether a snRNP reassembling with ubiquitinated HEXIM1 prevents the incorporation of LARP7. These models could be reconciled *in vitro* by assembling a minimal 7SK snRNA-HEXIM1-LARP7 snRNP, incubating the complex with wild type or catalytically dead UBE2O and monitoring the release of LARP7. Release of LARP7 with wild type UBE2O incubation would support that ubiquitination directly evicts LARP7 from the complex.

A common theme in ubiquitination and post-translational modification more generally is substrate recognition. In the crowded cellular environment, it is essential that modifiers properly identify target proteins to precisely control signaling pathways. For certain ubiquitin ligases, there are sequence motifs, or degrons, that target the ligase to the proper substrate for modification. For instance, the anaphase-promoting complex (APC/C) recognizes several short linear motif degrons in its substrates to properly coordinate the cell cycle (11). This thesis details two interesting cases of context-dependent substrate recognition, the first being that PJA2 properly modifies Tat only in complex with P-TEFb and the second that UBE2O modifies HEXIM1 that is incorporated into the 7SK snRNP. This requirement for proper complex formation ensures the correct staging of signal transduction. Given these context-dependent requirements, it will be important to explore the structural features of these complexes.

For instance, it is unclear if PJA2 only recognizes Tat or a shared surface between Tat and either CCNT1 or CDK9. Results from Chapter 3 demonstrate that PJA2 can bind P-TEFb, but this interaction is strengthened by Tat, suggesting that there is likely a shared recognition surface. For UBE2O, HEXIM1 is known to exist in an auto-inhibited state in which a basic region containing its NLS and an adjacent acidic region are thought to bury the CCNT1-binding PYNT motif (12). 7SK snRNA binding relieves HEXIM1 from this inhibited state allowing P-TEFb inhibition and appears to also present the proper orientation of HEXIM1 for UBE2O modification. As UBE2O is a hybrid E2-E3 ligase, it carries out the function of both ubiquitin transfer (E2) and substrate binding (E3). It would therefore be important to first determine if snRNP-bound HEXIM1 is a better substrate purely because of binding, which could be analyzed by pull down assays between UBE2O and either free or 7SK-bound HEXIM1. If binding is similar, it might then be the case that RNA-binding exposes otherwise buried acceptor lysines that are then properly oriented for modification. Obtaining an atomic structure of UBE2O-HEXIM1-7SK would further clarify how UBE2O both recognizes HEXIM1 and targets lysines for ubiquitination. Given that the 7SK snRNP is a major inhibitory reservoir for P-TEFb in the cell, it would also be useful to explore structures of free HEXIM1 and a minimal 7SK snRNA-HEXIM1 complex to examine how 7SK snRNA-dependent changes in HEXIM1 conformation allow P-TEFb binding.

In addition to context-dependent substrate recognition, another intriguing feature of Tat ubiquitination by PJA2 is that the ligase generates atypical polyubiquitin linkages. A ubiquitin chain that is generated through the recurrent use of a single ubiquitin lysine is referred to as a homotypic chain. In the cell, Lys11, Lys48, and Lys63 polyubiquitin

chains account for nearly 80% of homotypic conjugates (13), and these linkages have generally well-defined functions. However, through the use of single lysine (Lys^{only}) ubiquitin mutants, we were able to show that PJA2 can generate polyubiquitin chains on Tat-P-TEFb through Lys27 and Lys29, and to a lesser extent Lys33. Point mutations of these lysines in a wild type ubiquitin background did not diminish ubiquitination, suggesting a certain degree of redundancy in ubiquitin acceptor lysine choice. At the moment, the exact nature of these PJA2-dependent polyubiquitin chains is unclear, and this knowledge would eventually be important for answering future biochemical questions concerning the function of this modification. For example, PJA2 may generate homotypic Lys27, Lys29, or Lys33 chains, however the data are also consistent with the generation of heterotypic chains in which all three linkages are used in one polyubiquitin conjugate. One strategy to analyze ubiquitin chain linkage is to employ linkage-specific deubiquitinases (DUB) on modified substrates (14). At the moment, however, no DUB is reported to have specific, exclusive activity on Lys27, Lys29, or Lys33 homotypic chains, although several can cleave all three chain types (OTUD2 and OTUD6A) or both Lys29 and Lys33 chains (TRABID). Therefore, while employing these DUBs on PJA2-modified Tat could confirm the use of atypical linkages, it would not distinguish the architecture of the chain. Another way to explore polyubiquitin linkage is to examine the lysine acceptor usage by mass spec in which modified (diGlycine tryptic remnant) ubiquitin peptides could be monitored. This analysis would reveal if a certain chain type predominates. It will also be interesting to explore whether PJA2 generates Lys27, Lys29, and/or Lys33 chains on its other reported substrates (15–17).

While this thesis thoroughly explores the molecular determinants of Tat ubiquitination by PJA2, it is currently unclear the exact function of this modification. Given its non-degradative nature, one possibility is that the mark alters a protein-protein interaction e.g. recruitment of the super elongation complex. This model can be addressed by analyzing changes in Tat's protein-protein interactions upon PJA2 knockdown. Additionally, it is unclear if Tat is the only target of PJA2 that is important for regulating transcription or even if Tat redirects PJA2 to new substrates. It would therefore be valuable to identify all proteins in the Tat-PJA2 complex, which could be accomplished by Tat-PJA2 double IP coupled to mass spec. Any identified protein could then be evaluated as a PJA2 substrate either *in vivo* or *in vitro.* Of particular interest would be identifying the E2 used by PJA2 to modify Tat. The E2 UbcH5b used in our *in vitro* ubiquitination experiments is a generic E2 that is compatible with many E3 ligases. While UbcH5b-PJA2 can recapitulate most of the signaling activity observed *in vivo,* it would be ideal to uncover the native E2. Structural studies of Tat-P-TEFb-E2-PJA2 would then address how the Tat-P-TEFb substrate restricts PJA2 activity to generate the atypical polyubiquitin chains. As Lys27, Lys29, and Lys33 are proximal to each other in sequence and are also all contained on a single helix in the ubiquitin structure, one possibility is that CCNT1 or CDK9 blocks access to additional lysines to facilitate Lys27/29/33 usage. This structure could also provide insight into the degenerate nature of Tat acceptor lysine selection by PJA2.

Despite its potent activity, Tat is a particularly small protein able to commandeer the activity of several large cellular protein complexes. This thesis offers two solutions employed by Tat to maximize its limited information content. The first is the use of post-

translation modifications, in particular ubiquitination. The covalent attachment of a polyubiquitin chain drastically increases Tat's effective size, generating a large signaling scaffold with multiple possible protein-protein interaction surfaces. As Tat is also acetylated and methylated (18,19), a large combination of Tat protein states is available, each with a possible unique functional output. The second strategy employed by Tat to maximize its limited genetic content is the requirement of only a few amino acids to engage large protein complexes. In particular, the mutation of Lys28 decreases association with both the 7SK snRNP and ZUN complex, while mutation of the adjacent Phe32 dramatically increases association with UBE2O. In the Tat-P-TEFb crystal structure, Lys28 is positioned in an unstructured segment while Phe32 is on the boundary of a 3_{10}-helix (20), which is reminiscent of the short linear motifs used for protein-protein interactions found in many cellular and viral proteins (21). On the same surface that contains Lys28 and Phe32 are two additional residues (Glu2 and Tyr26) that have severe loss-of-function transcriptional phenotypes but no obvious interaction deficiencies. Tyr26 has been implicated in TAR-binding as mutation of this residue has no effect in a virus engineered to be TAR-independent (22). Activity measurements of all Tat alanine mutants in native and hybrid reporter assays uncovered a tight phenotypic clustering between Tyr26 and Lys28 (23). As Lys28 acetylation is important for tight TAR binding (18) and appears to also function in 7SK RNA binding as detailed in this thesis, Tyr26 may similarly regulate host and viral RNA interactions. Glu2 may be required for an interaction with a novel host protein not examined in this thesis (e.g. FBX3, USP11, DCAF16) and this could be explored with proteomics of wild type and E2A Tat proteins.

Tat Lys28 is highly conserved and highly selected for lysine during replication, Tyr26 and Phe32 are both conserved and selected for aromatics during replication, yet Glu2 is highly conserved with predominantly neutral selection in the virus. Several other Tat positions that are non-overlapping with other viral sequence, notably Glu9, Pro10, Trp11, Gln17, Thr20, and Ala21, are highly conserved in patients yet display similar neutral selection during viral replication with randomized Tat libraries. One possibility is that a strong selection pressure during organismal replication is absent during viral spread in tissue culture, such that these neutrally selected positions are not essential for Tat activity. Further, besides Glu2, none of the above-mentioned positions display any loss-of-function phenotypes in the Tat reporter assay when mutated to alanine. One possibility is that Tat utilizes an additional transcriptional pathway in certain infected cell types (i.e. macrophages), that is not highly expressed in the reporter system in HeLa cells or during replication in $CD4^+$ T cells. It could also be the case that Tat encodes an excess of P-TEFb targeting activity, such that the fitness of these neutrally selected residues is only apparent when Tat is operating under a strong transcriptional selection pressure, such as P-TEFb inhibition (24). To explore these ideas, T cells could be treated with drugs at increasing dosage that inhibit the canonical P-TEFb pathway (e.g. DRB, flavopiridol) and then individual residue fitness with randomized libraries during replication could be examined. It would be exciting if Glu2, Glu9, Pro10, Trp11, Gln17, Thr20, or Ala21 display strong positive selection only during a transcription selection pressure, supporting that there is a “core” P-TEFb-targeting activity composed of highly conserved and structural residues and a “secondary” P-TEFb activity that displays fitness defects only under strong selection pressure. Another possibility for neutral

selection of these highly conserved residues is that Tat executes a non-transcriptional function during organismal replication that is not recapitulated during viral spread in cell culture.

Finally, Tat ubiquitination is a mutationally robust transcriptional activity as multiple lysines can fully function as the acceptor site. As such, these modified lysines (Lys12, Lys19, Lys29, Lys50, Lys51, Lys71, and Lys85) display neutral selection with randomized Tat libraries. Further, Lys12, Lys19, and Lys29 are some of the least conserved positions in the non-overlapped Tat sequence, supporting the plasticity of these ubiquitinated positions. Replication with defined Tat mutants in Chapter 3 corroborates the function of the third, plastic lysine as a Tat ΔK (+K41, +K28) virus was delayed from wild type and the introduction of two different lysines (Lys12 or Lys85) to the Tat ΔK (+K41, +K28) background improved replication kinetics. As a better test of a third, plastic lysine requirement, the selection experiments with randomized Tat libraries at positions 12 and 85 could be repeated but in the Tat ΔK (+K41, +K28) background. The requirement of this flexible ubiquitin acceptor site would manifest itself as strong positive selection for a lysine in this sensitized background.

References

1. Rahl PB, Lin CY, Seila AC, Flynn RA, McCuine S, Burge CB, et al. C-Myc regulates transcriptional pause release. Cell [Internet]. Elsevier Ltd; 2010;141(3):432–45. Available from: http://dx.doi.org/10.1016/j.cell.2010.03.030
2. Core LJ, Waterfall JJ, Lis JT. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science [Internet]. 2008;322(5909):1845–8. Available from: http://science.sciencemag.org/content/322/5909/1845.abstract
3. Kwak H, Fuda NJ, Core LJ, Lis JT. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. Science [Internet]. 2013;339(6122):950–3. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3974810&tool=pmcentrez&rendertype=abstract
4. Henriques T, Gilchrist D, Nechaev S, Bern M, Muse G, Burkholder A, et al. Stable pausing by rna polymerase II provides an opportunity to target and integrate regulatory signals. Mol Cell [Internet]. Elsevier Inc.; 2013;52(4):517–28. Available from: http://dx.doi.org/10.1016/j.molcel.2013.10.001
5. Feinberg MB, Baltimore D, Frankel AD. The role of Tat in the human immunodeficiency virus life cycle indicates a primary effect on transcriptional elongation. Proc Natl Acad Sci U S A [Internet]. 1991;88(9):4045–9. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=51590&tool=pmcentrez&rendertype=abstract

6. Wei P, Garber ME, Fang SM, Fischer WH, Jones KA. 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. 1998;92(4):451–62.

7. Sobhian B, Laguette N, Yatim A, Nakamura M, Levy Y, Kiernan R, et al. HIV-1 Tat Assembles a Multifunctional Transcription Elongation Complex and Stably Associates with the 7SK snRNP. Mol Cell [Internet]. Elsevier Ltd; 2010;38(3):439–51. Available from: http://dx.doi.org/10.1016/j.molcel.2010.04.012

8. He N, Liu M, Hsu J, Xue Y, Chou S, Burlingame A, et al. HIV-1 Tat and Host AFF4 Recruit Two Transcription Elongation Factors into a Bifunctional Complex for Coordinated Activation of HIV-1 Transcription. Mol Cell [Internet]. Elsevier Ltd; 2010;38(3):428–38. Available from: http://dx.doi.org/10.1016/j.molcel.2010.04.013

9. D'Orso I, Frankel AD. RNA-mediated displacement of an inhibitory snRNP complex activates transcription elongation. Nat Struct Mol Biol. 2010;17(7):815–21.

10. Van Herreweghe E, Egloff S, Goiffon I, Jády BE, Froment C, Monsarrat B, et al. Dynamic remodelling of human 7SK snRNP controls the nuclear level of active P-TEFb. EMBO J [Internet]. 2007;26(15):3570–80. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1949012&tool=pmcentrez&rendertype=abstract

11. Davey NE, Morgan DO. Building a Regulatory Network with Short Linear Sequence Motifs: Lessons from the Degrons of the Anaphase-Promoting Complex. Mol Cell [Internet]. Elsevier Inc.; 2016;64(1):12–23. Available from: http://linkinghub.elsevier.com/retrieve/pii/S1097276516305238

12. Barboric M, Kohoutek J, Price JP, Blazek D, Price DH, Peterlin BM. Interplay between 7SK snRNA and oppositely charged regions in HEXIM1 direct the inhibition of P-TEFb. EMBO J [Internet]. 2005;24(24):4291–303. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1356324&tool=pmcentrez&rendertype=abstract

13. Ohtake F, Saeki Y, Ishido S, Kanno J, Tanaka K. The K48-K63 Branched Ubiquitin Chain Regulates NF- k B Signaling. Mol Cell [Internet]. Elsevier Inc.; 2016;64(2):1–16. Available from: http://dx.doi.org/10.1016/j.molcel.2016.09.014

14. Mevissen TET, Hospenthal MK, Geurink PP, Elliott PR, Akutsu M, Arnaudo N, et al. OTU deubiquitinases reveal mechanisms of linkage specificity and enable ubiquitin chain restriction analysis. Cell [Internet]. The Authors; 2013;154(1):169–84. Available from: http://dx.doi.org/10.1016/j.cell.2013.05.046

15. Rinaldi L, Delle Donne R, Sepe M, Porpora M, Garbi C, Chiuso F, et al. praja2 regulates KSR1 stability and mitogenic signaling. Cell Death Dis [Internet]. Nature Publishing Group; 2016;7(5):e2230. Available from: http://www.nature.com/doifinder/10.1038/cddis.2016.109

16. Lignitto L, Carlucci A, Sepe M, Stefan E, Cuomo O, Nisticò R, et al. Control of PKA stability and signalling by the RING ligase praja2. Nat Cell Biol [Internet]. Nature Publishing Group; 2011;13(4):412–22. Available from: http://dx.doi.org/10.1038/ncb2209

17. Lignitto L, Arcella A, Sepe M, Rinaldi L, Delle Donne R, Gallo A, et al. Proteolysis of MOB1 by the ubiquitin ligase praja2 attenuates Hippo signalling and supports glioblastoma growth. Nat Commun [Internet]. 2013;4(May):1822. Available from:

http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3674242&tool=pmcentrez&rendertype=abstract

18. D'Orso I, Frankel AD. Tat acetylation modulates assembly of a viral-host RNA-protein transcription complex. Proc Natl Acad Sci U S A. 2009;106(9):3101–6.
19. Ali I, Ramage H, Boehm D, Dirk LMA, Sakane N, Hanada K, et al. The HIV-1 Tat protein is monomethylated at lysine-71 by the lysine methyltransferase KMT7. J Biol Chem [Internet]. 2016;1–19. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27235396
20. Tahirov TH, Babayeva ND, Varzavand K, Cooper JJ, Sedore SC, Price DH. Crystal structure of HIV-1 Tat complexed with human P-TEFb.pdf. Nature [Internet]. Nature Publishing Group; 2010;465(7299):747–51. Available from: http://dx.doi.org/10.1038/nature09131
21. Hagai T, Azia A, Babu MM, Andino R. Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions. Cell Rep [Internet]. The Authors; 2014;7(5):1729–39. Available from: http://dx.doi.org/10.1016/j.celrep.2014.04.052
22. Das AT, Harwig A, Berkhout B. The HIV-1 Tat protein has a versatile role in activating viral transcription. J Virol [Internet]. 2011;85(18):9506–16. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3165771&tool=pmcentrez&rendertype=abstract
23. D'Orso I, Jang GM, Pastuszak AW, Faust TB, Quezada E, Booth DS, et al. Transition step during assembly of HIV Tat:P-TEFb transcription complexes and

transfer to TAR RNA. Mol Cell Biol [Internet]. 2012;32(23):4780–93. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23007159\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3497596

24. Stiffler MA, Hekstra DR, Ranganathan R. Evolvability as a Function of Purifying Selection in TEM-1 ??-Lactamase. Cell [Internet]. Elsevier Inc.; 2015;160(5):882–92. Available from: http://dx.doi.org/10.1016/j.cell.2015.01.035

Publishing Agreement

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

Please sign the following statement:

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

Tyler Faust

Author Signature

12/7/16

Date