Mechanisms and evolutionary implications of integration of different functions within HIV-1 Tat

Margarita A. Kurnaeva[†,1], Arthur O. Zalevsky[†,1,2], Eugene A. Arifulin[†,3], Olga M. Lisitsyna[3], Anna V. Tvorogova[3], Maria Y. Shubina[3], Gleb P. Bourenkov[4], Yana R. Musinova[3,5], Andrey V. Golovin[1,2], Yegor S. Vassetzky[5,6], Eugene V. Sheval*[,1,3,7]

[1]Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia

[2]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow, Russia

[3]Belozersky Institute of Physico-Chemical Biology, Lomonosov Moscow State University, Moscow, Russia

[4]European Molecular Biology Laboratory, Hamburg, Germany

[5]Koltzov Institute of Developmental Biology, Russian Academy of Sciences, Moscow, Russia

[6]CNRS, UMR 9018, Université Paris-Saclay, Institut Gustave Roussy, Villejuif, France

[7]Department of Cell Biology and Histology, Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia

[†]These authors contributed equally to this work.

***Corresponding author**: E-mail: sheval_e@belozersky.msu.ru

Summary

During evolution, viruses had to adapt to an increasingly complex environment of eukaryotic cells. Viral proteins that need to enter the cell nucleus or associate with nucleoli possess nuclear localization signals (NLSs) and nucleolar localization signals (NoLSs) for nuclear and nucleolar accumulation, respectively. As viral proteins are relatively small, acquisition of novel sequences seems to be a more complicated task for viruses than for eukaryotes. Here, we carried out a comprehensive analysis of the basic domain (BD) of HIV-1 Tat to show how viral proteins might evolve with NLSs and NoLSs without an increase in protein size. The HIV-1 Tat BD is involved in several functions, the most important being the transactivation of viral transcription. The BD also functions as an NLS, although it is substantially longer than a typical NLS. It seems that different regions in the BD could function as NLSs due to its enrichment with positively charged amino acids. Additionally, the high positive net charge inevitably causes the BD to function as an NoLS through a charge-specific mechanism. The integration of NLSs and NoLSs into functional domains enriched with positively charged amino acids might be a mechanism that allows the condensation of different functional sequences in small protein regions and, as a result, to reduce protein size, influencing the origin and evolution of NLSs and NoLSs in viruses.

Key words: HIV-1 Tat, nuclear localization signal (NLS), nucleolar localization signal (NoLS), viruses, evolution

Introduction

The origin of the cell nucleus, which led to the emergence of eukaryotic cells, was one of the most fateful events during the evolution of life on Earth. Acquisition of a nucleus enabled the spatial segregation of transcription and translation and led to the evolution of sophisticated mechanisms of regulation of gene expression (Martin and Koonin 2006).

The origin of the nuclear envelope led to the evolution of a complex system for nuclear import of proteins that are produced in the cytoplasm. To enter the nucleus, proteins use short signal sequences, known as nuclear localization signals (NLSs) (Lange et al. 2007; Paci et al. 2021). Sequences similar or identical to eukaryotic NLSs are present in proteins in different prokaryotes (Rossi et al. 1993; Nederlof et al. 1995; Perić et al. 2008; Lee et al. 2012; Moon et al. 2012; Kim et al. 2016; Kwon et al. 2016; Lisitsyna et al. 2020; Melnikov et al. 2020), indicating that the origin of NLSs precedes the separation of the nucleus and the cytoplasm. This integration of NLS(s) into different (mostly DNA- and/or RNA-binding) domains was also observed in modern Eukaryota (LaCasse and Lefebvre 1995; Cokol et al. 2000; Melnikov et al. 2015; Kharitonov et al. 2020; Lisitsyna et al. 2020).

Separation of the genome from the cytoplasm was followed by the evolution of numerous biomolecular condensates with concentrated proteins and/or RNAs and compartmentalized intranuclear processes (Shin and Brangwynne 2017; Sabari et al. 2020; Feric and Misteli 2021; Schmit et al. 2021). The largest intranuclear condensate is the nucleolus, which integrates numerous functions for genome organization and function (Iarovaia et al. 2019). Some proteins possess special signal sequences, which are referred to as nucleolar localization signals (NoLSs), to enable their accumulation inside nucleoli (Emmott and Hiscox 2009). The origin of the NoLS is unclear.

Viruses have adapted amazingly to infecting and replicating in eukaryotic cells by hijacking nuclear structures and processes (J de Castro and Lusic 2019). Viruses evolved long before the emergence of Eukaryota, when the unique ancestor of living organisms commonly referred to as the last universal cellular ancestor (LUCA)) had acquired a highly complex virome (Krupovic et al. 2020). In the course of evolution, viruses adapted to an increasingly complex environment of ancestral eukaryotic cells. The size of viral particles did not substantially increase during evolution (e.g., 50-90% of virus particles detected in marine virome have an average diameter of 50 nm (Brum et al. 2013)). Smaller viruses have multiple fitness advantages since a greater number of viral offspring can be created from limited host resources; smaller particles also diffuse to encounter new hosts more rapidly. Virion size and genome size are tightly correlated (Cui et al. 2014); therefore, proteins encoded by viral genomes are generally small and often versatile: one viral protein may be involved in many different processes, including regulation of the viral life cycle and modulation of host cell functions. The integration of the NLS, and probably the NoLS, into

different functional domains might confer special benefits to viruses by allowing them to express small proteins.

Here, we used the trans-activator of transcription (Tat) protein of human immunodeficiency virus 1 (HIV-1) to investigate how viral proteins integrate the NLS and NoLS domains. HIV-1 Tat is a regulatory protein essential for productive and processive transcription from the HIV-1 long terminal repeat (LTR) promoter. Tat binds to viral RNA and recruits a complex of cyclin T1 and cyclin-dependent kinase 9 (CDK9) that activates RNA polymerase II, thus increasing transcription from the viral promoter (Ne et al. 2018). Tat is a small protein encoded by two exons (Kuppuswamy et al. 1989) containing several domains: (i) the N-terminal proline-rich domain (1-21 aa), (ii) cysteine-rich domain (22-37 aa), (iii) hydrophobic core domain (38-47 aa), (iv) basic domain (BD) (48-59 aa), (v) glutamine-rich domain (60-72 aa), and (vi) C-terminal domain. While Tat can tolerate sequence mutations as high as 38% without its activity significantly changing (Opi et al. 2002), the BD is highly conserved in Tat variants. It is enriched with positively charged arginine and lysine residues comprising the ^{49}RKKRRQRRR^{57} motif. The BD confers many properties to Tat (for a review, see (Kurnaeva et al. 2019)). To activate viral transcription, Tat binds *via* its BD to a short nascent stem-bulge loop leader RNA transactivation responsive region (TAR) at the 5' extremity of viral transcripts (Aboul-ela et al. 1995; Ne et al. 2018; Pham et al. 2018). The BD also functions as an NLS (Dang and Lee 1989; Kuppuswamy et al. 1989; Ruben et al. 1989; Subramanian et al. 1990; Efthymiadis et al. 1998); however, the mechanism of Tat nuclear import is still debated. While some data indicate that Tat is imported into the nucleus *via* its interaction with importin-α (Gautier et al. 2009; Yang et al. 2010; Smith et al. 2017), *in vitro* assays suggest that Tat nuclear import is mediated by the direct binding of its BD to importin-β (Truant and Cullen 1999). Additionally, the Tat BD interacts with nuclear components, probably with RNAs, which can also lead to nuclear accumulation (Efthymiadis et al. 1998). The BD also functions as an NoLS (Dang and Lee 1989; Kuppuswamy et al. 1989; Siomi et al. 1990; Musinova et al. 2015); Tat accumulation might occur *via* an electrostatic interaction of its BD with nucleolar components (Musinova et al. 2015). The accumulation of Tat inside nucleoli modulates nucleolar processes (Ponti et al. 2008; Jarboui et al. 2012).

Thus, the structure of the Tat BD allows the realization of several unrelated functions. This multifunctionality can be beneficial for the virus, as it acquired several functions without an increase in the protein size. Here, we investigated this multifunctionality of the Tat BD and demonstrated that NLS and NoLS were inevitably integrated into the HIV-1 Tat BD. Integration of an NLS and/or NoLS into functional protein domains might influence viral evolution, since this integration may lead to a reduction in protein size.

Results

Tat is actively imported into the nucleus *via* its BD

To investigate the mechanisms of nuclear and nucleolar accumulation of the HIV-1 Tat protein in living cells, we used Tat protein fused with EGFP (EGFP-Tat). To exclude the possibility that EGFP inhibited Tat transactivation activity, we developed an *in vivo* assay based on a fast-maturing TurboRFP fluorescent protein controlled by a fragment of the HIV-1 3' LTR that included the TAR (LTR-TurboRFP plasmid). We cotransfected U2OS cells with plasmids coding EGFP-Tat and LTR-TurboRFP and observed TurboRFP fluorescence in cells coexpressing EGFP-Tat but not EGFP or EGFP-BD (supplementary fig. S1). The results showed that the transactivation activity of the Tat protein was not perturbed by its fusion with EGFP.

We next analyzed the subcellular localization of EGFP-Tat transiently expressed in U2OS cells. EGFP-Tat was detected predominantly inside nuclei (fig. 1A), where it accumulated inside nucleoli (supplementary fig. S2). Tat with the deleted BD (EGFP-TatΔBD) was diffusely distributed in cells, while BD-fused EGFP (EGFP-BD) accumulated inside nuclei and nucleoli (fig. 1A). To estimate the efficiency of nuclear accumulation, the ratio of nucleoplasmic to cytoplasmic fluorescence (F_{nuc}/F_{cyt}) was measured after background correction, and the measurements confirmed that the BD was essential for nuclear localization of the Tat protein (fig. 1B).

There are two major ways that a protein uses to enter the nucleus: passive diffusion, which is possible only for small proteins (<40-60 kDa), and active transport *via* importin-dependent pathways. The size of the EGFP-Tat fusion protein in this study was within the limit for passive diffusion, and Tat protein was shown to freely diffuse through nuclear pore complexes(Stauber and Pavlakis 1998). However, diffusion *per se* cannot lead to nuclear accumulation: biologically inert EGFP freely enters the cell nucleus by diffusion, but its concentration is roughly the same in the nucleus and cytoplasm (fig. 1A and B). Nuclear accumulation might result either from nuclear retention *via* interaction with nuclear components or from active nuclear import *via* importin-dependent pathways. Nuclear import is an energy-dependent process, and ATP depletion can be used to discriminate between nuclear retention and active nuclear import(Schwoebel et al. 2002). For ATP depletion, U2OS cells expressing EGFP-Tat were incubated in a buffer containing sodium azide and 2-deoxyglucose for 50 min. Cells expressing either EGFP or SV40 T-antigen NLS fused with EGFP (NLS^{SV40}-EGFP) were used as the negative and positive controls, respectively. Live-cell imaging demonstrated that NLS^{SV40}-EGFP was relocalized from the nucleoplasm into the cytoplasm upon ATP depletion, while EGFP localization was not changed (fig. 1C). We also observed a substantial decrease in the nucleoplasmic concentrations of EGFP-Tat and EGFP-BD in the absence of ATP (fig. 1C and D). Nuclear accumulation of EGFP-Tat decreased after ATP depletion, but the level remained higher than that of EGFP and NLS^{SV40}-EGFP under similar conditions, suggesting

that nuclear accumulation of EGFP-Tat was partially due to active transport and, to a lesser extent, to nuclear retention. Importantly, after ATP depletion, the nuclear accumulation of EGFP-BD decreased to the level of NLS^{SV40}-EGFP and EGFP, indicating that while active nuclear import was provided by the BD, other domains of the Tat protein were essential for its nuclear retention.

Localization of the NLS in the Tat BD

The most common type of NLS, classical NLS (cNLS), is imported into the nucleus *via* interaction with importin-*α*. The most characterized monopartite cNLS has a consensus sequence of K(K/R)X(K/R), ensuring the optimal interaction with importin-*α (Lange et al. 2007)*. The Tat BD contains nine amino acids (^{49}RKKRRQRRR^{57}) and thus is substantially longer than necessary for binding to importin-*α*. We used several software programs to predict the potential NLS(s) in the BD. NLStradamus (Nguyen Ba et al. 2009), seqNLS (Lin and Hu 2013), NucPred (Brameier et al. 2007) and cNLS Mapper (Kosugi et al. 2009) predicted an NLS that roughly corresponded to the BD (fig. 2A), while PSORT II (Horton et al. 2007) predicted three NLSs (^{49}RKKR^{52}, ^{50}KKRR^{53} and ^{55}RRRP^{58}) overlapping the Tat BD. To map the NLS(s) in the BD, we systematically substituted each amino acid in the BD with alanine, followed by expression of the mutated Tat proteins and the analysis of their localization *in vivo* (fig. 2B). The substitution of any arginine resulted in an ~2-fold decrease in Tat nuclear accumulation. The substitution of any lysine or glutamine with alanine decreased Tat accumulation, albeit to a lesser extent, and this difference was not statistically significant. Thus, arginine residues are more important than lysine residues for the NLS function of the Tat BD. These data seem to contradict the recently obtained crystal structure of the importin-*α*/TAT-NLS (^{47}SGRKKRRQRRRAPQN^{61}) complex, in which lysine residues constituted the main interaction core (Smith et al. 2017). To resolve this discrepancy, we employed molecular modeling.

A broad range of computational approaches is routinely applied to study and design interactions between macromolecules with atomic precision (Leman et al. 2020); they proved useful in resolving the contradictory structural data in previous studies (Kuzmanic et al. 2014) and in situations with multiple potential binding modes (Mobley and Dill 2009). To predict the potential interaction sites of different parts of the BD and importin-*α*, we performed two series of docking procedures. We started with rapid coarse grain docking of the ^{47}SGRKKRRQRRR^{57} peptide into the full importin-*α* structure. The docking experiment revealed two main binding sites overlapping the major and minor NLS-binding sites (fig. 3A). Of the ten final models, two models showed binding outside the NLS-binding sites, six showed binding to the major NLS-binding sites, and two showed less favorable configurations with binding at the minor site, an interaction that was not apparent in the crystal structure.

To obtain a detailed model of the interactions between the peptides and importin-α NLS-binding sites, we performed full-atom docking of tetrapeptides derived from the Tat NLS. In agreement with the coarse-grain docking, the major NLS-binding site attracted more peptides than other protein areas (fig. 3B). Moreover, peptides with higher binding energy estimates were enriched at the major NLS-binding site, not the minor site (fig. 3B). We next analyzed tetrapeptides in the major binding site and found that the arginine-rich tetrapeptides had a higher binding energy than the lysine-rich tetrapeptides (fig. 3C). Importantly, we found that the arginine-rich peptides in different parts of the BD docked to major NLS sites in the same manner (with the full-atom backbone RMSD less than 1 Å) (fig. 3E). Despite the similarities between models with lysine and arginine, the difference in binding energies may be the result of an unusual amino acid composition at the binding sites. Major importin-α NLS binding sites are enriched in aromatic amino acids, especially tryptophan residues (Trp184, Trp 231 and Trp273), which possess electron-rich conjugated pi-systems. In this context, recognition of positively charged organic cations is realized through electrostatic interactions. Surprisingly, arginine residues were shown to demonstrate stronger binding than lysine residues in pi-cation interactions (Kumar et al. 2018). These previous observations are in agreement with the results of our mutagenesis study (fig. 2B).

As we noted above, the analysis of docking results demonstrated that regardless of the peptide sequences, the backbone position was almost constant in most of the docking poses (fig. 3E). This diversity of sidechain positions with the almost immutable backbone position in the crystal structure could be misinterpreted as a result of thermal fluctuations of atoms around their respective positions, thus causing an increase in the so-called thermal or B-factor. For instance, B-factors greater than 100 $Å^2$ or residues totally missing from the structure because of large thermal motions in the otherwise stable and well-resolved areas may indicate this type of problem region or be a sign of structure overfitting (Wlodawer et al. 2008; Kuzmanic et al. 2014; Carugo 2018). The polder map around the NLS shows excessive unattributed electron density around N-terminal residues (supplementary fig. S3). Moreover, many of the NLS residues have high B-factors, with that of the terminal arginine exceeding 100 $Å^2$. Thus, the reliability of the positions of these residues is relatively low.

The docking results indicated that different parts of the BD could interact with the major site of importin-α. To test whether the reported electron density map (Smith et al. 2017) can accommodate different peptides, we rescored the docking results in accordance with how well their heavy atoms (with the exception of hydrogen) fit in areas with higher levels of electron density (supplementary fig. 4). Two peptides, SGRK and GRKK, had the highest enrichment among the top-scoring peptides. With 58 and 22 % of occurrence in the top-3 list from every run, they greatly outperformed the other peptides. Nevertheless, RKKR, KKRR, RQRR, and KRRQ were also able to adequately fit the electron density, although KKRR had the reverse orientation (supplementary fig. S4). In the selected structures, arginine

residues occupied sites initially assigned to N-terminal glycine or both lysine residues in the crystal reported structure. We observed multiple GRKK configurations, with one being identical to the crystal structure, while in the other, glycine and arginine interchanged their positions. Considering the unexplained residual electron density peaks at the -6 to +7 sigma level over G48 and R49 and a Ramachandran plot outlier at R49, we performed a correction of the initial 5SVZ model by rotating the G48 and R49 side chains about the R49 CA-C bond by 180 degrees following refinement with Refmac (Murshudov et al. 2011), resulting in R=0.16, Rfree=0.19 at 2.0 Å with no significant residual density over the remodeled segment (supplementary fig. 5). R49 was in the preferred Ramachandran region, while its guanidine group established hydrogen bonds with side chains 234, 235, 270, and 277.

Hence, the peptide reported by (Smith et al. 2017) and several other peptides composing the Tat BD can be accommodated within the published crystal structure (fig. 3E, supplementary fig. S4).

The docking results indicated that the same positions could be occupied by either lysine or arginine. To ascertain whether this interchangeability between lysine residues and arginine residues is a unique feature of HIV-1 Tat, we analyzed all published PBD structures obtained for the complexes of importin-*α* with 38 eukaryotic and 14 viral proteins (supplementary table S1). Monopartite cNLSs have a consensus sequence of K(K/R)X(K/R)(Lange et al. 2007). Consecutive residues from the N-terminal lysine of the monopartite NLS are referred to as P1, P2, etc. Previous structural (Conti and Kuriyan 2000; Fontes et al. 2000) and thermodynamic (Hodel et al. 2001) studies demonstrated that a monopartite cNLS requires a lysine in the P1 position, followed by basic residues in positions P2 and P4. However, in three importin-*α*/NLS PDB structures obtained for viral proteins (beak and feather disease virus capsid protein (PDB ID 4HTV), HIV-1 viral protein R (Vpr) (PDB ID 5B56), and influenza A virus nucleoprotein (NP) (PDB ID 5V5O)), an arginine residue was substituted for a lysine in the P1 position (fig. 3A; supplementary fig. 6). We obtained a consensus logo structure for the region that directly interacted with importin-*α*. The P1 position was always occupied by lysine in eukaryotic proteins (fig 4B), while the presence of arginine residues in the P1 position of some viral NLSs produced a slightly different consensus sequence of the viral NLS ((K/R)(K/R)X(K/R)) (fig. 4C).

Hence, NLSs from eukaryotic proteins always (at least in all described cases) contain lysine in the P1 position in the major NLS-binding site, indicating the functional importance of this lysine; thus, the NLS sequence might be suboptimal for nuclear import in some viral proteins.

Long NLSs in viral proteins

To ascertain whether similar long regions with NLS-like activity are present in other viral proteins, we analyzed published data on experimentally reported viral NLSs in which the NLSs were precisely mapped using either site-directed mutagenesis or importin binding. We found 10 viral proteins with either long

(≥9 a.a.) or adjacent NLSs (supplementary table S2). Some of these "long" NLSs, as in the case of NLSs in the feline immunodeficiency virus Rev protein, contained a long region enriched with positively charged amino acids similar to the Tat BD (Marchand et al. 2020); others, e.g., 2b protein in cucumber mosaic virus, contained two or more adjacent NLSs (^{22}KKQRRR^{27} and ^{33}RRER^{36}) (Wang et al. 2004). Thus, viral proteins may contain extended regions that function as NLSs.

The short classical NLSs and long NLSs described above represent two extreme variants. To analyze all variants and describe the heterogeneity of NLS structure, we assessed 106 NLSs in 88 viral proteins and 269 NLSs in 228 human proteins (supplementary tables S3 and S4). All these NLSs, both in viral and human proteins, were enriched with positively charged amino acids (arginine residues and lysine residues), while NLSs in the viral proteins were also proline-rich (fig. 5A). Importantly, viral NLSs contained more arginine residues than lysine residues, while human NLSs were lysine-rich. Surprisingly, the prevalence of arginine residues over lysine residues is an overall characteristic of NLS-containing viral proteins; in contrast, the prevalence of lysine residues over arginine residues is seen in NLS-containing human proteins.

The NLSs were analyzed for the size and structure of clusters of positively charged amino acids (fig. 5B). We identified all clusters where arginine residues/lysine residues/histidine residues were interspersed with no more than one non-positively charged amino acid. The clusters containing three and four positively charged amino acids corresponded to cNLSs (K(R/K)X(R/K)). The number of these clusters in viral NLSs was slightly higher than that in human NLSs (~39% of viral clusters and ~32% of human clusters), while the number of clusters with 7-10 amino acids was twofold greater in human NLSs (~15%) than in viral NLSs (~6%), probably due to the presence of additional types of noncanonical NLSs that seem to be rare in viral proteins. Hence, long NLSs are present in both viral and human proteins.

Tat accumulation inside nucleoli depends on the charge of its BD

Previous studies showed that Tat protein accumulated inside nucleoli and that this accumulation depended on the BD (Dang and Lee 1989; Kuppuswamy et al. 1989; Siomi et al. 1990; Musinova et al. 2015). We confirmed these observations by expressing Tat-EGFP and its mutant forms in U2OS cells (fig. 6A). Deletion of the BD led to a reduction in Tat-EGFP nucleolar accumulation, and the BD alone led to the accumulation EGFP in nucleoli, indicating that the BD indeed functioned as an NoLS. While ATP depletion led to a decrease in nuclear accumulation (fig. 1C and D), the nucleolar accumulation of Tat simultaneously increased (fig. 6B). We quantified the nucleolar accumulation by measuring the ratio of EGFP fluorescence in the nucleolus to that in the nucleoplasm (F_n/F_{nuc}) and found that nucleolar accumulation was ~2-fold higher after ATP depletion. NoLS(s) facilitate protein accumulation inside nucleoli by interacting with nucleolar components (nucleolar retention) (Martin et al. 2015; Musinova et

al. 2015). To ascertain whether Tat nucleolar retention changed after ATP depletion, we analyzed the EGFP-Tat exchange between the nucleolus and the nucleoplasm using FRAP (fig. 6C). In untreated U2OS cells, full recovery of EGFP-Tat was observed within $t_{1/2}$ ~6 s. Upon depletion of cellular ATP, the recovery rate was decreased to ~65%, indicating that a substantial fraction of EGFP-Tat was immobilized inside nucleoli after ATP depletion (fig. 6D). The $t_{1/2}$ for the mobile fraction was ~9 s, and thus, the rate of Tat exchange was decreased. Thus, the retention of the Tat protein inside nucleoli led to nucleolar accumulation.

We next used site-directed mutagenesis to analyze the role of each amino acid in BD nucleolar accumulation. When any amino acid of the BD was substituted with alanine, nucleolar accumulation of EGFP-Tat decreased regardless of the residue substituted, with the exception of glutamine (fig 7A). Tat nucleolar accumulation decreased further upon simultaneous substitution of two (fig. 7B) or three (fig. 7C) amino acids with alanine residues. To compare the effect of the mutation of one, two or three amino acids on nucleolar accumulation, we combined the data on all mutants (with the exception of the Q54A mutant, which did not affect nucleolar accumulation) with the same number of substitutions (fig. 7D). Nucleolar accumulation was indeed dependent on the proportion of positively charged amino acids, indicating that HIV-1 Tat accumulated inside nucleoli *via* a charge-dependent, not a sequence-dependent, mechanism.

HIV-1 Tat interacts with nucleolin and nuclear RNA

The accumulation of proteins inside nucleoli was previously shown to result from high-affinity binding interactions with core nucleolar components, including ribosomal DNA, RNAs and proteins (Carmo-Fonseca et al. 2000). The Tat nuclear interactome includes several nucleolar proteins, including NPM1 and NCL (Gautier et al. 2009), indirectly indicating that Tat can accumulate inside nucleoli due to its interactions with nucleolar proteins. The interaction of Tat with NPM1 was also previously demonstrated *in vitro (Li 1997)*. However, NoLS(s) were shown to interact with nucleolar RNAs (Martin et al. 2015; Musinova et al. 2015). To test the potential interactions of the Tat protein with major nucleolar proteins *in vivo*, we used a fluorescent two-hybrid (F2H) assay to make use of cells with a GFP-anchoring platform in the nucleus. EGFP-fused proteins were tethered to the intranuclear GFP-anchoring platform and then assayed for colocalization with TagRFP-fused proteins that could potentially interact with GFP-fused proteins. After expression in F2H cells, EGFP-Tat was located preferentially at the GFP-anchoring platform in the majority of cells (nucleolar accumulation was clearly visible only in cells with a high expression of EGFP-Tat, indicating a robust interaction with the platform) (fig. 7E). We next analyzed the interaction of EGFP-Tat with two major nucleolar proteins, NPM1 and NCL. After coexpression of EGFP-Tat and NPM1-TagRFP, we did not observe any accumulation of

NPM1-TagRFP at the platform. In contrast, clear colocalization at the platform was seen after coexpression of EGFP-Tat and NCL-TagRFP, indicating a direct interaction between these proteins. Interestingly, some accumulation of NCL-TagRFP on the platform was detected in ~50% of the cells after EGFP-BD and NCL-TagRFP were coexpressed, indicating that the interaction between Tat and NCL can occur through the BD (fig. 7E).

We next analyzed whether Tat interacted with cellular RNAs using an RNA pull-down assay. EGFP-Tat or EGFP was expressed in U2OS cells and then immunoprecipitated with magnetic beads (fig. 7F). The beads were then incubated with total cellular RNA from U2OS cells and stained with SYBR-Gold dye. Fluorescence microscopy revealed that immunoprecipitated EGFP-Tat immobilized RNA on the surface of the agarose beads, while RNA was not detected on the surface of either the control beads or beads with EGFP (fig. 7G); therefore, Tat protein interacts with both NCL and RNA, which may account for its retention in the nucleus and nucleolus.

Overlapping of NLSs and NoLSs

Several reports previously described overlapping NLS(s) and NoLS(s) (for discussion, see (Scott et al. 2010)); however, a comprehensive mapping of both NLSs and NoLSs has never been performed. We used the datasets of experimentally established NLSs (supplementary tables S3 and S4) and predicted NoLSs (Scott et al. 2011). In addition, we used a dataset of experimentally established NoLSs for which we predicted NLSs using the NLStradamus program (supplementary table S5). Both prediction strategies demonstrated that NLSs and NoLSs more frequently overlapped in viral proteins than in human proteins (fig. 8A).

Estimates of the minimal content required for NoLS function vary. According to one report, a protein region can function as an NoLS if it contains more than either three arginine residues or five lysine residues (Musinova et al. 2015), according to the other report, six arginine residues are required (Martin et al. 2015). NLSs are enriched with positively charged amino acids, and therefore, one can assume that long NLSs may function as NoLSs. Indeed, the majority of the long NLSs (with >6-7 positively charged amino acids) were predicted to be potential NoLSs (fig. 8B).

Discussion

Many functions of HIV-1 Tat are attributed to its BD (Kurnaeva et al. 2019), the major function being the transactivation of viral transcription (Mancebo et al. 1997; Wei et al. 1998; Bieniasz et al. 1999; Rustanti et al. 2017). Additionally, the Tat BD functions as an NLS and NoLS. Here, we investigated the mechanisms of NLS and NoLS integration into the BD of HIV-1 Tat and analyzed the implications of this integration for virus evolution.

Integration of NLS into the Tat BD

The Tat protein is small, and therefore, it was previously shown to diffuse freely through nuclear pore complexes (Stauber and Pavlakis 1998). However, the presence of NLS led to a robust nuclear accumulation of HIV-1 Tat compared to biologically inert EGFP. A striking feature of the Tat BD is its length (9 amino acids), which is substantially longer than the minimum required for association with importin-*α* (4 amino acids). We carried out systematic site-directed mutagenesis of all amino acids of the Tat BD and found that the whole BD sequence functions as an NLS. *In silico* molecular docking experiments indicated that different regions of the BD might potentially interact with importin-*α*. Following these results, we proposed a model of multiple concurrent binding modes of the Tat NLS, in which different parts of the BD can interact with importin-*α* (supplementary fig. 6). It appears that a single NLS can bind with importin-*α* in a variety of ways because of the excess of charged amino acids,, preserving the overall geometry without any significant decrease in binding energy. As a result, the nuclear accumulation can be a consequence of cumulative binding between different fragments of BD and importin-*α*.

The second unusual feature of the Tat NLS of HIV-1 is the essential role of arginine residues in NLS function, which was corroborated both by the results of site-directed mutagenesis and molecular docking. These observations were in obvious contradiction with the known consensus sequence of the classical monopartite NLSs, in which the P1 position is always occupied by lysine (K(R/K)X(R/K)). However, when we analyzed PDB structures obtained for complexes of importin-*α* with 38 eukaryotic and 14 viral NLSs, we found that in three NLSs from viral proteins, the most conserved lysine in the P1 position was replaced with arginine, and in one NLS, it was replaced with valine. Thus, at least in some viral proteins, NLSs have a suboptimal organization. Notably, the Tat BD contains a sequence that corresponds to the classical eukaryotic NLS consensus sequence (^{50}KKRR^{53}), but paradoxically, this region of the Tat BD did not lead to more efficient nuclear accumulation than other sequences in the BD. Thus, in some viral proteins, the NLS sequence is not optimal for nuclear import due to enrichment with arginine residues. This might be a consequence of the integration of NLSs into functional domains, as in the case of HIV-1 Tat. In the BD, arginine residues participate in binding to the TAR, which leads to transactivation of HIV-1 transcription (Calnan et al. 1991; Chavali et al. 2020). This may be a general feature of these integrated NLSs, since 72-87% of known protein-RNA interfaces contain arginine residues (Barik et al. 2015; Krüger et al. 2018) whose substitution with lysine residues often diminishes their RNA-binding activity (Calnan et al. 1991; Su et al. 1997; Belashov et al. 2018). Thus, when an RNA-binding domain acquires the function of an NLS, its RNA-binding and importin-binding efficiencies must be equilibrated.

Moreover, we found that NLS-containing viral proteins are overall arginine-rich, and the composition of the viral NLS may simply reflect this enrichment.

Thus, integration of NLSs into functional domains may prevent the evolution of structurally optimal NLS sequences when evolution would impair the major function(s) of the domain, as in the case of HIV-1 Tat, but this might be compensated, for example, by the acquisition of multiple concurrent binding-mode mechanisms.

It appears that HIV-1 Tat is not a single viral protein with an integrated NLSs. The domains involved in RNA binding and the NLS of Herpes simplex virus type 1 nucleocytoplasmic shuttling protein UL47 overlap, and it is not possible to separate their activities (Donnelly et al. 2007). The organization of the NLS we found in Tat BD might be widespread not only among viruses but also among eukaryotes. Indeed, in some eukaryotic proteins, NLSs might also be integrated into other domains that occupy a significant portion of the protein. This situation was previously described for histones (Mosammaparast et al. 2001; Mosammaparast et al. 2002), ribosomal proteins (Moreland et al. 1985; Schaap et al. 1991; Jäkel and Görlich 1998) and the nucleolar methyltransferase fibrillarin (FBL) (Shubina et al. 2020). Histones and ribosomal proteins are small and highly conserved proteins, and the presence of extended NLSs may also be a consequence of their integration into a functional domain, which could not be uncoupled during evolution.

Integration of NoLS into the Tat BD

NoLSs facilitate the accumulation of proteins inside nucleoli *via* a charge-dependent mechanism, not a sequence-dependent mechanism (Musinova et al. 2011; Savada and Bonham-Smith 2013; Martin et al. 2015; Musinova et al. 2015); i.e., any positively charged region can potentially function as an NoLS. For example, core histones can accumulate inside nucleoli before their incorporation into chromatin or after their release from chromatin (Musinova et al. 2011; Safina et al. 2017; Apta-Smith et al. 2018; Sen Gupta et al. 2018), and it was demonstrated that histone H2B contains an NoLS in its N-terminal tail (Musinova et al. 2011).

Here, we used site-directed mutagenesis to demonstrate that each positively charged amino acid of the BD is involved in the nucleolar accumulation of Tat; thus, Tat accumulated inside nucleoli *via* a charge-dependent mechanism, similar to other NoLSs. It seems logical that electrostatic interactions with nucleolar components that lead to nucleolar accumulation would have a low specificity. Indeed, we found that the Tat protein could interact with both the nucleolar protein NCL and RNAs. RNA is charged negatively, and NCL contains several extended acidic domains and even has a negative net charge at physiological pH (pI 4.1). Interestingly, we were unable to detect any interaction between Tat and NPM1,

although it was previously demonstrated that Tat interacted *in vitro* with this nucleolar protein (Li 1997; Gautier et al. 2009).

Evolutionary implications of NLS and NoLS integration

The Tat BD is a conserved region enriched with positively charged amino acids that serves as both an NLS and NoLS. The BD evolved primarily as a domain involved in binding with TAR and the transactivation of HIV-1 transcription, but its enrichment with positively charged amino acids inevitably led to the acquisition of its functions as an NLS and NoLS (fig. 9). These two additional functions are embedded in the BD amino acid sequence and cannot be uncoupled from the major function of the BD.

Evolution of the transactivation function would have been perturbed if nuclear and/or nucleolar accumulation had disrupted the performance of the main function of this BD. However, accumulation in the nucleus most likely promotes transactivation. Although the role of the nucleolar accumulation of Tat in HIV-1 infection was not directly demonstrated, it seems that due to the rapid exchange of Tat between the nucleoli and the nucleoplasm, its nuclear accumulation cannot negatively affect the transactivation function.

The primacy of the transactivation function during evolution led to the formation of NLSswhich are substantially longer than classical NLSs. These NLSs might function through multiple concurrent binding with importin-α. Moreover, our bioinformatic analysis demonstrated that, in some viruses, the structure of the NLS may also be nonoptimal. The multiple concurrent binding mode mechanism allows viral proteins to accumulate in the nucleus due to the presence of several closely located or overlapping NLSs, even if a single NLS was not optimized for NLS function.

Examples of protein domain evolution when functions did not evolve sequentially and independently of each other but evolved directly in the form of an integrated functional complex may be widespread, and traces of this integration might be observed in modern living beings. Indeed, NLS integration into annotated domains is a widespread phenomenon in Eukaryota (Lisitsyna et al. 2020). An additional indication of the inevitable character of NLS/NoLS function acquisition is the presence of these signals integrated into cytoplasmic proteins (Kharitonov et al. 2020).

Integration of the NLS and NoLS into functional protein domains might influence viral evolution since this integration could prevent an increase in protein size. Separation of functions in different protein domains, which is fundamentally possible for eukaryotic proteins, leads to an increase in protein size that might be disadvantageous for viruses. Thus, the integration of the NLS and NoLS into functional domains could be a key phenomenon that influenced the origin and evolution of the NLS both in viruses and in eukaryotes.

Materials and methods

Cell culture

U2OS and HeLa cells (both from Russian Cell Culture Collection, Institute of Cytology of the Russian Academy of Sciences, Saint Petersburg, Russia) were grown in Dulbecco's modified Eagle's medium supplemented with alanyl glutamine (Paneco, Moscow, Russia), 10% fetal calf serum (HyClone, Logan, UT, USA) and an antibiotic and antimycotic solution (Gibco, Rockville, MD, USA).

In vivo analysis of transactivation activity of Tat protein

To detect the transactivation activity of the Tat protein, we developed an *in vivo* system based on the fast-maturing fluorescent protein TurboRFP. The fragment of the HIV-1 3' LTR, which contains all of the U3 region and a fragment of the R region, including the TAR, was amplified from the HIV-1 LTR lacZ reporter vector, which was obtained through the NIH AIDS Reagent Program from Dr. Joseph Maio (Maio and Brown 1988) using the following primers:
5'-AGTCAAGCTTTGGAAGGGCTAATTCACTCCCAAAG-3' and
5'-AGTCGGTACCAGCTTTATTGAGGCTTAAGCAGTGGG-3'. The PCR product was digested with FastDigest HindIII and KpnI restriction enzymes (Thermo Scientific, Waltham, MA, USA) and cloned into a promoter-less vector encoding the red fluorescent protein TurboRFP (pTurboRFP-PRL; Evrogen, Moscow, Russia). U2OS cells were cotransfected with the EGFP-Tat, EGFP or EGFP-BD and LTR-TurboRFP plasmids, and the fluorescence was analyzed using an AxioVert 200M microscope (Carl Zeiss, Oberkochen, Germany) equipped with an ORCAII-ERG2 cooled CCD camera (Hamamatsu Photonics K.K., Hamamatsu City, Shizuoka, Japan).

Plasmid construction

To construct EGFP-Tat, the pGST-Tat 1 86R plasmid was obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, from Dr. Andrew Rice (Herrmann and Rice 1993; Rhim et al. 1994; Herrmann and Rice 1995). Full-length Tat was PCR amplified with Phusion high-fidelity DNA polymerase (Thermo Scientific, Waltham, MA, USA). The amplified PCR products were digested with BamHI and HindIII and inserted into a EGFP-C1 vector (Clontech, Mountain View, CA, USA).

For site-directed mutagenesis, the Tat gene was cloned into a pJET1.2/blunt cloning vector (Thermo Scientific, Waltham, MA, USA) using a CloneJET PCR cloning kit (Thermo Scientific, Waltham, MA, USA). Point mutations were made using a Change-IT multiple-mutation site-directed mutagenesis kit (Affymetrix, Santa Clara, CA, USA) according to the manufacturer's instructions. Each subsequent round of mutagenesis was carried out after confirming the presence of the preceding necessary mutation by sequencing. The mutated genes were amplified by PCR; the resulting PCR products were digested by HindIII and BamHI, gel purified and cloned into a pEGFP-C1 vector (Clontech, Mountain View, CA, USA).

To construct the EGFP-BD plasmid, we designed adapters encoding the BD of HIV-1 Tat with HindIII and BamHI restriction sites. The oligonucleotides were incubated in annealing buffer (10 mM Tris, pH 8.0; 50 mM NaCl; and 10 mM $MgCl_2$) for 5 min at 94°C and then for 5 min at 70°C. The resulting product was cloned into a pEGFP-C1 vector (Clontech, Mountain View, CA, USA).

To obtain the NCL-TagRFP plasmid, full-length NCL was PCR amplified with Phusion high-fidelity DNA polymerase (Thermo Scientific, Waltham, MA, USA) using the following primers: 5'-AGTCGAATTCATGGTGAAGCTCGCGAAGGCAG-3' and 5'-AGTCGGATCCAATTCAAACTTCGTCTTCTTTCCTTGTGG-3'. The pEGFP-NCL expression plasmid (nucleolin), which was used as a template for PCR, was a kind gift from Dr J. Borowiec(Kim et al. 2005). The amplified PCR product was digested with EcoRI and BamHI and then inserted into the pTagRFP-N1 vector (Evrogen, Moscow, Russia). After cloning into an expression vector, the correctness of all constructs was confirmed by sequencing.

The construction of NLS^{SV40}-EGFP(Shubina et al. 2020) and NPM1-TagRFP (Musinova et al. 2015) plasmids was previously described.

Measurement of nuclear and nucleolar accumulation

To evaluate the nuclear and nucleolar accumulation of the Tat protein, we used a method described elsewhere with some modifications (Musinova et al. 2015). Images of at least 20 living U2OS cells expressing EGFP-fusion proteins were acquired through two different experiments using a Nikon C2 confocal laser scanning microscope with a 63x1.4NA oil immersion objective under identical conditions. A region of interest was determined within the nucleolus, within the nucleoplasm and within the cytoplasm of the same cell. The mean gray value was determined for each, background levels were subtracted, and the nucleolar-nucleoplasmic (F_n/F_{nuc}) and nucleoplasmic-cytoplasmic (F_{nuc}/F_{cyt}) ratios were determined for every value pair. The statistical analysis and graph generation were performed using Prism 6 software (GraphPad, San Diego, CA, USA). The comparisons were performed using nonparametric statistical tests: either Mann-Whitney U test or Kruskal-Wallis test followed by Dunn's multiple comparisons test.

ATP depletion assay

ATP depletion was carried out as described by (Bhattacharya et al. 2006). Briefly, HeLa cells were grown in 35-mm dishes on a coverslip (MatTek, Ashland, MA, USA). The cells were treated with 20% Hank's solution for 15 min and then transferred into Medium 1 (150 mM NaCl, 5 mM KCl, 1 mM $CaCl_2$, 1 mM $MgCl^2$ and 20 mM HEPES, pH 7.4) containing 10 mM sodium azide and 6 mM 2-deoxy-D-glucose (Sigma-Aldrich, St. Louis, MO, USA) and incubated for 50 min. Live cell imaging was performed with a Nikon C2 confocal microscope with a 60× Plan-Apo objective (NA 1.4) at 37°C; focus stabilization was performed using a PFS system (Nikon, Minato City, Tokyo, Japan).

Molecular docking

Docking was performed through two different approaches. Full-length coarse-grained docking of the SGRKKRQRRR peptides was performed with standalone CABS docking (Kurcinski et al. 2015). The importin-*α* structure with PDB ID 5SVZ was cleared of the NLS peptide and water molecules. The default parameters for sampling efficacy (simulation cycles 50) were used, and no additional restraints were applied.

Full-atom docking of tetrapeptides was performed using the importin-*α* structure with PDB ID 5SVZ (a 30 Å x 45 Å x 45 Å docking box with the center at 80.559, 21.365, 100.184 was used to cover both NLS sites). Docking was performed with the Qvina-W package (Hassan et al. 2017), which implements the AutoDock Vina (Trott and Olson 2010) scoring function optimized for blind docking. Due to the relative complexity of peptide docking (Rentzsch and Renard 2015), the exhaustiveness parameter, which affects sampling efficacy, was set to 512. Twenty independent runs for every peptide were performed with twenty poses per peptide produced.

Rescoring of the poses was performed in accordance with a 2FoFc electron density map using a modified PeptoGrid (Zalevsky et al. 2019) procedure. The density map was converted into a numerical grid with an elementary step of approximately 0.1 Å. Per-atom scores were calculated as negative log probabilities of corresponding grid cells with the corresponding sigma sign. Values less than +1σ were penalized as -3σ. Only heavy atoms were considered, with ACE and NME caps excluded from scoring. Total scores were sorted and normalized to the highest score.

Fluorescence recovery after photobleaching (FRAP)

Cells were grown in 35-mm dishes on coverslips. The medium was overlaid with mineral oil before the experiment, and the dishes were mounted onto a Nikon A1 confocal microscope. Four single scans were acquired for the FRAP experiments, followed by a single pulse for photobleaching. The recovery curves were generated from background-subtracted images. The relative fluorescence intensity (RFI) was calculated as

$$RFI = T_0 I_t / T_t I_0,$$

where T_0 is the total cellular intensity during prebleaching, T_t is the total cellular intensity at time point t, I_0 is the average intensity in the region of interest during prebleaching, and I_t is the average intensity in the region of interest at time point t. The results for at least 20 cells were averaged to obtain the final curve of fluorescence recovery.

F2H assay

F2H cells (genetically modified baby hamster kidney (BHK) fibroblasts) were cultured according to the manufacturer's instructions (ChromoTek GmbH, Planegg-Martinsried, Germany). For detection of protein-protein interactions in live cells, we transfected F2H cells with a plasmid coding the protein fused

with EGFP and a second plasmid coding the protein fused with TagRFP. Cells were fixed with 3.7% paraformaldehyde, and images were acquired using a Nikon C2 confocal laser scanning microscope with a 63x1.4NA oil immersion objective.

RNA pull-down assay

Total RNA was isolated from U2OS cells using an RNeasy mini kit (Qiagen Inc., Valencia, CA, USA) according to the manufacturer's instructions. EGFP-Tat or EGFP genes were cloned into a pLCMV lentiviral vector. Lentiviral constructs were introduced into 293T cells along with packaging plasmids. Viral particles were used for the transduction of the U2OS cells, and cells expressing target proteins were selected using a FACSAria III cell sorter (BD). EGFP-Tat and EGFP were immunoprecipitated from U2OS cells using 25 µl of GFP-Trap magnetic beads (ChromoTek GmbH, Planegg-Martinsried, Germany) according to the manufacturer's instructions.

Control of immunoprecipitation was carried out using immunoblotting. Beads were resuspended in the Laemmli sample buffer, boiled for 3 min, resolved on a 12.5% SDS-polyacrylamide gel and transferred to a nitrocellulose membrane. The membranes were blocked in 1% bovine serum albumin and incubated with either a monoclonal antibody against GFP (1:3000; a generous gift from A.G. Evstafieva) or a monoclonal antibody against B23 (1:10,000; Sigma-Aldrich, St. Louis, MO, USA) and monoclonal antibody against β-tubulin (1:10,000; Sigma-Aldrich, St. Louis, MO, USA). The membranes were washed three times with PBS (5 min each time) and were then incubated with secondary peroxidase-conjugated antibody (1:15,000; Sigma-Aldrich, St. Louis, MO, USA). The antibody-bound proteins were detected using Pierce ECL western blotting substrate (Thermo Scientific, Waltham, MA, USA), and images were acquired using a Gel DocXR system (Bio-Rad, Hercules, CA, USA).

For the pulldown assay, immobilized EGFP or EGFP-Tat was incubated with total RNA from U2OS cells for 2 h at 4°C on a rotary shaker. After three washing steps, the RNA was labeled with SYBR-Gold dye (Invitrogen). The beads were imaged using an AxioVert 200M microscope (Carl Zeiss, Oberkochen, Germany) equipped with an ORCAII-ERG2 cooled CCD camera (Hamamatsu Photonics K.K., Hamamatsu City, Shizuoka, Japan).

NLS/NoLS bioinformatic analysis

We used 106 experimentally annotated NLSs of 88 viral proteins and 269 experimentally annotated NLSs of 228 human proteins (supplementary tables S4 and S5). The dataset of the NLSs of human proteins was published elsewhere (Lisitsyna et al. 2020), and NLSs described in recent papers were included (supplementary table S6). All NLSs of viral proteins were found via a search of UniprotKB database or published papers (supplementary table S2). The NoLS dataset consisted of 24 experimentally determined NoLSs in 23 viral proteins and 71 NoLSs in 65 animal proteins (supplementary table S5).

The presence of potential NoLSs in the human and viral protein datasets with experimentally determined NLSs was predicted by the NOD server (Scott et al. 2010), and the putative NLSs of the proteins in the NoLS dataset were predicted by NLStradamus (Nguyen Ba et al. 2009).

The diversity and number of clusters of positive amino acids in an NLS were evaluated by custom R-script. The cluster was defined as a region in the NLS in which arginine residues and/or lysine residues/histidine residues were interspersed with another amino acid (there may be several such gaps, but each gap was not to be greater than one amino acid).

Artwork

The contrast and brightness of the final images were adjusted using Adobe Photoshop (Adobe, San Jose, CA, USA). The gamma value was adjusted only for several panels, and these cases are noted in the figure legends. Figure 9 was created with BioRended.com.

Code availability

R notebooks and custom scripts are available at https://github.com/lisitsynaom/NLS_Tat.

Acknowledgements

We would like to thank S. Dokudovskaya, P.V. Lidsky, A.A. Zharikova and A.A. Mironov for stimulating discussions and valuable suggestions, and D.M. Potashnikova and M.A. Tikhomirova for technical support. A.O.Z. is grateful to Dr. G. Armeev for the comments on the quality assessment of crystal structures. We are grateful to Dr J. Borowiec for EGFP-nucleolin plasmid. The following reagents were obtained through the NIH AIDS Reagent Program, Division of AIDS, NIAID, NIH: HIV-1 HXB2 GST-Tat Expression Vector (GST-Tat 1 (86R)) from Dr. Andrew Rice (cat# 2367) and HIV-1 LTR lacZ Reporter Vector (pHIVlacZ) from Dr. Joseph Maio (cat #151). Docking experiments were carried out using the equipment of the shared research facilities of HPC computing resources at Lomonosov Moscow State University. The microscopy facility became available in the framework of the Moscow State University Development Program PNR 5.13. This work was supported by the Russian Science Foundation (grant 21-74-20134 to E.V.S.) and the Russian Foundation for Basic Research (grant 18-29-08012 to A.O.Z.).

Author contributions

E.V.S. conceived and designed the study. M.A.K., E.A.A., A.V.T., M.Y.S., Y.R.M. and E.V.S. performed cloning and microscopy studies. A.O.Z. and A.V.G. performed molecular docking. G.B. performed refinement of the crystal structure. O.M.L. and E.V.S. performed bioinformatic analysis. M.A.K., A.O.Z., O.M.L., Y.R.M., Y.S.V. and E.V.S. analysed the data and wrote the manuscript.

References

Aboul-ela F, Karn J, Varani G. 1995. The structure of the human immunodeficiency virus type-1 TAR RNA reveals principles of RNA recognition by Tat protein. *J. Mol. Biol.* 253:313–332.

Apta-Smith MJ, Hernandez-Fernaud JR, Bowman AJ. 2018. Evidence for the nuclear import of histones H3.1 and H4 as monomers. *EMBO J.* [Internet] 37. Available from: http://dx.doi.org/10.15252/embj.201798714

Barik A, C N, Pilla SP, Bahadur RP. 2015. Molecular architecture of protein-RNA recognition sites. *J. Biomol. Struct. Dyn.* 33:2738–2751.

Belashov IA, Crawford DW, Cavender CE, Dai P, Beardslee PC, Mathews DH, Pentelute BL, McNaughton BR, Wedekind JE. 2018. Structure of HIV TAR in complex with a Lab-Evolved RRM provides insight into duplex RNA recognition and synthesis of a constrained peptide that impairs transcription. *Nucleic Acids Res.* 46:6401–6415.

Bhattacharya D, Mazumder A, Miriam SA, Shivashankar GV. 2006. EGFP-tagged core and linker histones diffuse via distinct mechanisms within living cells. *Biophys. J.* 91:2326–2336.

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

Brameier M, Krings A, MacCallum RM. 2007. NucPred—Predicting nuclear localization of proteins. *Bioinformatics* 23:1159–1160.

Brum JR, Schenck RO, Sullivan MB. 2013. Global morphological analysis of marine viruses shows minimal regional variation and dominance of non-tailed viruses. *ISME J.* 7:1738–1751.

Calnan BJ, Tidor B, Biancalana S, Hudson D, Frankel AD. 1991. Arginine-mediated RNA recognition: the arginine fork. *Science* 252:1167–1171.

Carmo-Fonseca M, Mendes-Soares L, Campos I. 2000. To be or not to be in the nucleolus. *Nat. Cell Biol.* 2:E107–E112.

Carugo O. 2018. How large B-factors can be in protein crystal structures. *BMC Bioinformatics* 19:61.

Chavali SS, Mali SM, Jenkins JL, Fasan R, Wedekind JE. 2020. Co-crystal structures of HIV TAR RNA bound to lab-evolved proteins show key roles for arginine relevant to the design of cyclic peptide TAR inhibitors. *J. Biol. Chem.* [Internet]. Available from: http://dx.doi.org/10.1074/jbc.RA120.015444

Cokol M, Nair R, Rost B. 2000. Finding nuclear localization signals. *EMBO Rep.* 1:411–415.

Conti E, Kuriyan J. 2000. Crystallographic analysis of the specific yet versatile recognition of distinct nuclear localization signals by karyopherin alpha. *Structure* 8:329–338.

Cui J, Schlub TE, Holmes EC. 2014. An allometric relationship between the genome length and virion

volume of viruses. *J. Virol.* 88:6403–6410.

Dang CV, Lee WM. 1989. Nuclear and nucleolar targeting sequences of c-erb-A, c-myb, N-myc, p53, HSP70, and HIV tat proteins. *J. Biol. Chem.* 264:18019–18023.

Donnelly M, Verhagen J, Elliott G. 2007. RNA binding by the herpes simplex virus type 1 nucleocytoplasmic shuttling protein UL47 is mediated by an N-terminal arginine-rich domain that also functions as its nuclear localization signal. *J. Virol.* 81:2283–2296.

Efthymiadis A, Briggs LJ, Jans DA. 1998. The HIV-1 Tat nuclear localization sequence confers novel nuclear import properties. *J. Biol. Chem.* 273:1623–1628.

Emmott E, Hiscox JA. 2009. Nucleolar targeting: the hub of the matter. *EMBO Rep.* 10:231–238.

Feric M, Misteli T. 2021. Phase Separation in Genome Organization across Evolution. *Trends Cell Biol.* [Internet]. Available from: http://dx.doi.org/10.1016/j.tcb.2021.03.001

Fontes MR, Teh T, Kobe B. 2000. Structural basis of recognition of monopartite and bipartite nuclear localization sequences by mammalian importin-alpha. *J. Mol. Biol.* 297:1183–1194.

Gautier VW, Gu L, O'Donoghue N, Pennington S, Sheehy N, Hall WW. 2009. In vitro nuclear interactome of the HIV-1 Tat protein. *Retrovirology* 6:47.

Hassan NM, Alhossary AA, Mu Y, Kwoh C-K. 2017. Protein-Ligand Blind Docking Using QuickVina-W With Inter-Process Spatio-Temporal Integration. *Sci. Rep.* 7:15451.

Herrmann CH, Rice AP. 1993. Specific interaction of the human immunodeficiency virus Tat proteins with a cellular protein kinase. *Virology* 197:601–608.

Herrmann CH, Rice AP. 1995. Lentivirus Tat proteins specifically associate with a cellular protein kinase, TAK, that hyperphosphorylates the carboxyl-terminal domain of the large subunit of RNA polymerase II: candidate for a Tat cofactor. *J. Virol.* 69:1612–1620.

Hodel MR, Corbett AH, Hodel AE. 2001. Dissection of a nuclear localization signal. *J. Biol. Chem.* 276:1317–1325.

Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35:W585–W587.

Iarovaia OV, Minina EP, Sheval EV, Onichtchouk D, Dokudovskaya S, Razin SV, Vassetzky YS. 2019. Nucleolus: A Central Hub for Nuclear Functions. *Trends Cell Biol.* [Internet]. Available from: http://dx.doi.org/10.1016/j.tcb.2019.04.003

Jäkel S, Görlich D. 1998. Importin beta, transportin, RanBP5 and RanBP7 mediate nuclear import of ribosomal proteins in mammalian cells. *EMBO J.* 17:4491–4502.

Jarboui MA, Bidoia C, Woods E, Roe B, Wynne K, Elia G, Hall WW, Gautier VW. 2012. Nucleolar protein trafficking in response to HIV-1 Tat: rewiring the nucleolus. *PLoS One* 7:e48702.

J de Castro I, Lusic M. 2019. Navigating through the nucleus with a virus. *Curr. Opin. Genet. Dev.*

55:100–105.

Kharitonov AV, Shubina MY, Nosov GA, Mamontova AV, Arifulin EA, Lisitsyna OM, Nalobin DS, Musinova YR, Sheval EV. 2020. Switching of cardiac troponin I between nuclear and cytoplasmic localization during muscle differentiation. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1867:118601.

Kim J-M, Choe M-H, Asaithambi K, Song J-Y, Lee YS, Lee JC, Seo J-H, Kang H-L, Lee KH, Lee W-K, et al. 2016. Helicobacter pylori HP0425 Targets the Nucleus with DNase I-Like Activity. *Helicobacter* 21:218–225.

Kim K, Dimitrova DD, Carta KM, Saxena A, Daras M, Borowiec JA. 2005. Novel checkpoint response to genotoxic stress mediated by nucleolin-replication protein a complex formation. *Mol. Cell. Biol.* 25:2463–2474.

Kosugi S, Hasebe M, Tomita M, Yanagawa H. 2009. Systematic identification of cell cycle-dependent yeast nucleocytoplasmic shuttling proteins by prediction of composite motifs. *Proc. Natl. Acad. Sci. U. S. A.* 106:10171–10176.

Krüger DM, Neubacher S, Grossmann TN. 2018. Protein-RNA interactions: structural characteristics and hotspot amino acids. *RNA* 24:1457–1465.

Krupovic M, Dolja VV, Koonin EV. 2020. The LUCA and its complex virome. *Nat. Rev. Microbiol.* [Internet]. Available from: http://dx.doi.org/10.1038/s41579-020-0408-x

Kumar K, Woo SM, Siu T, Cortopassi WA, Duarte F, Paton RS. 2018. Cation-π interactions in protein-ligand binding: theory and data-mining reveal different roles for lysine and arginine. *Chem. Sci.* 9:2655–2665.

Kuppuswamy M, Subramanian T, Srinivasan A, Chinnadurai G. 1989. Multiple functional domains of Tat, the trans-activator of HIV-1, defined by mutational analysis. *Nucleic Acids Res.* 17:3551–3561.

Kurcinski M, Jamroz M, Blaszczyk M, Kolinski A, Kmiecik S. 2015. CABS-dock web server for the flexible docking of peptides to proteins without prior knowledge of the binding site. *Nucleic Acids Res.* 43:W419–W424.

Kurnaeva MA, Sheval EV, Musinova YR, Vassetzky YS. 2019. Tat basic domain: A “Swiss army knife” of HIV-1 Tat? *Rev. Med. Virol.*:e2031.

Kuzmanic A, Pannu NS, Zagrovic B. 2014. X-ray refinement significantly underestimates the level of microscopic heterogeneity in biomolecular crystals. *Nat. Commun.* 5:3220.

Kwon YC, Kim S, Lee YS, Lee JC, Cho M-J, Lee W-K, Kang H-L, Song J-Y, Baik SC, Ro HS. 2016. Novel nuclear targeting coiled-coil protein of Helicobacter pylori showing Ca(2+)-independent, Mg(2+)-dependent DNase I activity. *J. Microbiol.* 54:387–395.

LaCasse EC, Lefebvre YA. 1995. Nuclear localization signals overlap DNA-or RNA-binding domains in

nucleic acid-binding proteins. *Nucleic Acids Res.* 23:1647.

Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH. 2007. Classical nuclear localization signals: definition, function, and interaction with importin alpha. *J. Biol. Chem.* 282:5101–5105.

Lee JH, Jun SH, Baik SC, Kim DR, Park J-Y, Lee YS, Choi CH, Lee JC. 2012. Prediction and screening of nuclear targeting proteins with nuclear localization signals in Helicobacter pylori. *J. Microbiol. Methods* 91:490–496.

Leman JK, Weitzner BD, Lewis SM, Adolf-Bryfogle J, Alam N, Alford RF, Aprahamian M, Baker D, Barlow KA, Barth P, et al. 2020. Macromolecular modeling and design in Rosetta: recent methods and frameworks. *Nat. Methods* 17:665–680.

Lin J-R, Hu J. 2013. SeqNLS: nuclear localization signal prediction based on frequent pattern mining and linear motif scoring. *PLoS One* 8:e76864.

Lisitsyna OM, Kurnaeva MA, Arifulin EA, Shubina MY, Musinova YR, Mironov AA, Sheval EV. 2020. Origin of the nuclear proteome on the basis of pre-existing nuclear localization signals in prokaryotic proteins. *Biol. Direct* 15:9.

Li YP. 1997. Protein B23 is an important human factor for the nucleolar localization of the human immunodeficiency virus protein Tat. *J. Virol.* 71:4098–4102.

Maio JJ, Brown FL. 1988. Regulation of expression driven by human immunodeficiency virus type 1 and human T-cell leukemia virus type I long terminal repeats in pluripotential human embryonic cells. *J. Virol.* 62:1398–1407.

Mancebo HS, Lee G, Flygare J, Tomassini J, Luu P, Zhu Y, Peng J, Blau C, Hazuda D, Price D, et al. 1997. P-TEFb kinase is required for HIV Tat transcriptional activation in vivo and in vitro. *Genes Dev.* 11:2633–2644.

Marchand C, Lemay G, Archambault D. 2020. Identification of the nuclear and nucleolar localization signals of the Feline immunodeficiency virus Rev protein. *Virus Res.* 290:198153.

Martin RM, Ter-Avetisyan G, Herce HD, Ludwig AK, Lättig-Tünnemann G, Cardoso MC. 2015. Principles of protein targeting to the nucleolus. *Nucleus* 6:314–325.

Martin W, Koonin EV. 2006. Introns and the origin of nucleus-cytosol compartmentalization. *Nature* 440:41–45.

Melnikov S, Ben-Shem A, Yusupova G, Yusupov M. 2015. Insights into the origin of the nuclear localization signals in conserved ribosomal proteins. *Nat. Commun.* 6:7382.

Melnikov S, Kwok H-S, Manakongtreecheep K, van den Elzen A, Thoreen CC, Söll D. 2020. Archaeal Ribosomal Proteins Possess Nuclear Localization Signal-Type Motifs: Implications for the Origin of the Cell Nucleus. *Mol. Biol. Evol.* 37:124–133.

Mobley DL, Dill KA. 2009. Binding of small-molecule ligands to proteins: “what you see” is not always

"what you get." *Structure* 17:489–498.

Moon DC, Gurung M, Lee JH, Lee YS, Choi CW, Kim SI, Lee JC. 2012. Screening of nuclear targeting proteins in Acinetobacter baumannii based on nuclear localization signals. *Res. Microbiol.* 163:279–285.

Moreland RB, Nam HG, Hereford LM, Fried HM. 1985. Identification of a nuclear localization signal of a yeast ribosomal protein. *Proc. Natl. Acad. Sci. U. S. A.* 82:6561–6565.

Mosammaparast N, Guo Y, Shabanowitz J, Hunt DF, Pemberton LF. 2002. Pathways mediating the nuclear import of histones H3 and H4 in yeast. *J. Biol. Chem.* 277:862–868.

Mosammaparast N, Jackson KR, Guo Y, Brame CJ, Shabanowitz J, Hunt DF, Pemberton LF. 2001. Nuclear import of histone H2A and H2B is mediated by a network of karyopherins. *J. Cell Biol.* 153:251–262.

Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr. D Biol. Crystallogr.* 67:355–367.

Musinova YR, Kananykhina EY, Potashnikova DM, Lisitsyna OM, Sheval EV. 2015. A charge-dependent mechanism is responsible for the dynamic accumulation of proteins inside nucleoli. *Biochim. Biophys. Acta* 1853:101–110.

Musinova YR, Lisitsyna OM, Golyshev SA, Tuzhikov AI, Polyakov VY, Sheval EV. 2011. Nucleolar localization/retention signal is responsible for transient accumulation of histone H2B in the nucleolus through electrostatic interactions. *Biochim. Biophys. Acta* 1813:27–38.

Nederlof PM, Wang HR, Baumeister W. 1995. Nuclear localization signals of human and Thermoplasma proteasomal alpha subunits are functional in vitro. *Proc. Natl. Acad. Sci. U. S. A.* 92:12060–12064.

Ne E, Palstra R-J, Mahmoudi T. 2018. Transcription: Insights From the HIV-1 Promoter. *Int. Rev. Cell Mol. Biol.* 335:191–243.

Nguyen Ba AN, Pogoutse A, Provart N, Moses AM. 2009. NLStradamus: a simple Hidden Markov Model for nuclear localization signal prediction. *BMC Bioinformatics* 10:202.

Opi S, Péloponèse J-M Jr, Esquieu D, Campbell G, de Mareuil J, Walburger A, Solomiac M, Grégoire C, Bouveret E, Yirrell DL, et al. 2002. Tat HIV-1 primary and tertiary structures critical to immune response against non-homologous variants. *J. Biol. Chem.* 277:35915–35919.

Paci G, Caria J, Lemke EA. 2021. Cargo transport through the nuclear pore complex at a glance. *J. Cell Sci.* [Internet] 134. Available from: http://dx.doi.org/10.1242/jcs.247874

Perić M, Schedewig P, Bauche A, Kruppa A, Kruppa J. 2008. Ribosomal proteins of Thermus thermophilus fused to beta-galactosidase are imported into the nucleus of eukaryotic cells. *Eur. J. Cell Biol.* 87:47–55.

Pham VV, Salguero C, Khan SN, Meagher JL, Brown WC, Humbert N, de Rocquigny H, Smith JL, D'Souza VM. 2018. HIV-1 Tat interactions with cellular 7SK and viral TAR RNAs identifies dual structural mimicry. *Nat. Commun.* 9:4266.

Ponti D, Troiano M, Bellenchi GC, Battaglia PA, Gigliani F. 2008. The HIV Tat protein affects processing of ribosomal RNA precursor. *BMC Cell Biol.* 9:32.

Rentzsch R, Renard BY. 2015. Docking small peptides remains a great challenge: an assessment using AutoDock Vina. *Brief. Bioinform.* 16:1045–1056.

Rhim H, Echetebu CO, Herrmann CH, Rice AP. 1994. Wild-type and mutant HIV-1 and HIV-2 Tat proteins expressed in Escherichia coli as fusions with glutathione S-transferase. *J. Acquir. Immune Defic. Syndr.* 7:1116–1121.

Rossi L, Hohn B, Tinland B. 1993. The VirD2 protein of Agrobacterium tumefaciens carries nuclear localization signals important for transfer of T-DNA to plant. *Mol. Gen. Genet.* 239:345–353.

Ruben S, Perkins A, Purcell R, Joung K, Sia R, Burghoff R, Haseltine WA, Rosen CA. 1989. Structural and functional characterization of human immunodeficiency virus tat protein. *J. Virol.* 63:1–8.

Rustanti L, Jin H, Lor M, Lin MH, Rawle DJ, Harrich D. 2017. A mutant Tat protein inhibits infection of human cells by strains from diverse HIV-1 subtypes. *Virol. J.* 14:52.

Sabari BR, Dall'Agnese A, Young RA. 2020. Biomolecular Condensates in the Nucleus. *Trends Biochem. Sci.* [Internet]. Available from: http://dx.doi.org/10.1016/j.tibs.2020.06.007

Safina A, Cheney P, Pal M, Brodsky L, Ivanov A, Kirsanov K, Lesovaya E, Naberezhnov D, Nesher E, Koman I, et al. 2017. FACT is a sensor of DNA torsional stress in eukaryotic cells. *Nucleic Acids Res.* 45:1925–1945.

Savada RP, Bonham-Smith PC. 2013. Charge versus sequence for nuclear/nucleolar localization of plant ribosomal proteins. *Plant Mol. Biol.* 81:477–493.

Schaap PJ, van't Riet J, Woldringh CL, Raué HA. 1991. Identification and functional analysis of the nuclear localization signals of ribosomal protein L25 from Saccharomyces cerevisiae. *J. Mol. Biol.* 221:225–237.

Schmit JD, Feric M, Dundr M. 2021. How Hierarchical Interactions Make Membraneless Organelles Tick Like Clockwork. *Trends Biochem. Sci.* [Internet]. Available from: http://dx.doi.org/10.1016/j.tibs.2020.12.011

Schwoebel ED, Ho TH, Moore MS. 2002. The mechanism of inhibition of Ran-dependent nuclear transport by cellular ATP depletion. *J. Cell Biol.* 157:963–974.

Scott MS, Boisvert F-M, McDowall MD, Lamond AI, Barton GJ. 2010. Characterization and prediction of protein nucleolar localization sequences. *Nucleic Acids Res.* 38:7388–7399.

Scott MS, Troshin PV, Barton GJ. 2011. NoD: a Nucleolar localization sequence detector for eukaryotic

and viral proteins. *BMC Bioinformatics* 12:317.

Sen Gupta A, Joshi G, Pawar S, Sengupta K. 2018. Nucleolin modulates compartmentalization and dynamics of histone 2B-ECFP in the nucleolus. *Nucleus* 9:350–367.

Shin Y, Brangwynne CP. 2017. Liquid phase condensation in cell physiology and disease. *Science* [Internet] 357. Available from: http://dx.doi.org/10.1126/science.aaf4382

Shubina MY, Arifulin EA, Sorokin DV, Sosina MA, Tikhomirova MA, Serebryakova MV, Smirnova T, Sokolov SS, Musinova YR, Sheval EV. 2020. The GAR domain integrates functions that are necessary for the proper localization of fibrillarin (FBL) inside eukaryotic cells. *PeerJ* 8:e9029.

Siomi H, Shida H, Maki M, Hatanaka M. 1990. Effects of a highly basic region of human immunodeficiency virus Tat protein on nucleolar localization. *J. Virol.* 64:1803–1807.

Smith KM, Himiari Z, Tsimbalyuk S, Forwood JK. 2017. Structural Basis for Importin-α Binding of the Human Immunodeficiency Virus Tat. *Sci. Rep.* 7:1650.

Stauber RH, Pavlakis GN. 1998. Intracellular trafficking and interactions of the HIV-1 Tat protein. *Virology* 252:126–136.

Subramanian T, Kuppuswamy M, Venkatesh L, Srinivasan A, Chinnadurai G. 1990. Functional substitution of the basic domain of the HIV-1 trans-activator, Tat, with the basic domain of the functionally heterologous Rev. *Virology* 176:178–183.

Su L, Radek JT, Hallenga K, Hermanto P, Chan G, Labeots LA, Weiss MA. 1997. RNA recognition by a bent alpha-helix regulates transcriptional antitermination in phage lambda. *Biochemistry* 36:12722–12732.

Trott O, Olson AJ. 2010. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* 31:455–461.

Truant R, Cullen BR. 1999. The arginine-rich domains present in human immunodeficiency virus type 1 Tat and Rev function as direct importin beta-dependent nuclear localization signals. *Mol. Cell. Biol.* 19:1210–1217.

Wang Y, Tzfira T, Gaba V, Citovsky V, Palukaitis P, Gal-On A. 2004. Functional analysis of the Cucumber mosaic virus 2b protein: pathogenicity and nuclear localization. *J. Gen. Virol.* 85:3135–3147.

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

Wlodawer A, Minor W, Dauter Z, Jaskolski M. 2008. Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* 275:1–21.

Yang SNY, Takeda AAS, Fontes MRM, Harris JM. 2010. Probing the specificity of binding to the major

nuclear localization sequence-binding site of importin-α using oriented peptide library screening. *Journal of Biological* [Internet]. Available from: http://www.jbc.org/content/285/26/19935.short

Zalevsky AO, Zlobin AS, Gedzun VR, Reshetnikov RV, Lovat ML, Malyshev AV, Doronin II, Babkin GA, Golovin AV. 2019. PeptoGrid-Rescoring Function for AutoDock Vina to Identify New Bioactive Molecules from Short Peptide Libraries. *Molecules* [Internet] 24. Available from: http://dx.doi.org/10.3390/molecules24020277

Figures

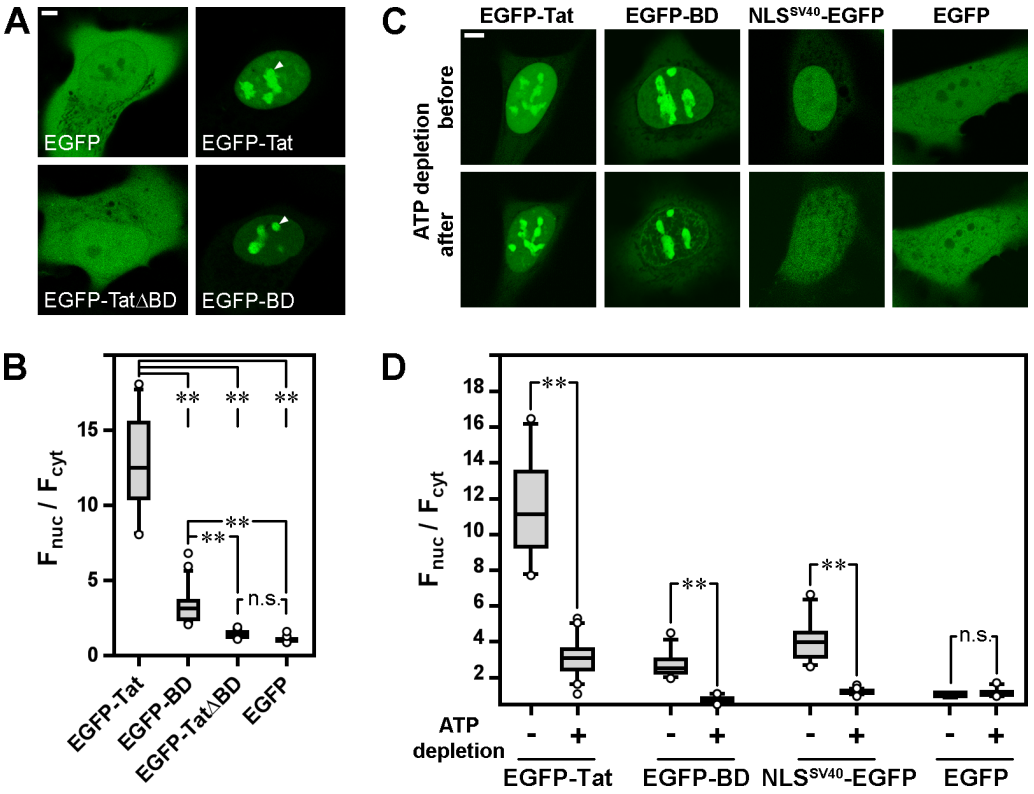

Fig. 1. Basic domain of HIV-1 Tat functions as an NLS. *(A)* The localization of EGFP, EGFP-Tat, EGFP-TatΔBD and EGFP-BD in U2OS cells (live cell imaging). The intranuclear accumulation of EGFP-Tat and EGFP-BD (arrowheads) corresponded to nucleoli (see also supplementary fig. S2). Bar – 5µm. *(B)* Estimation of the nuclear accumulation (F_{nuc}/F_{cyt}) of EGFP, EGFP-Tat, EGFP-TatΔBD and EGFP-BD in living U2OS cells. The comparisons were performed using the Kruskal-Wallis test (n.s. – not significant; ** – $p < .005$; $n > 35$). *(C)* The localization of EGFP-Tat and EGFP-TatΔBD in living U2OS cells before and after ATP depletion. EGFP-NLS^{SV40} was used as a positive control, and EGFP was used as a negative control. The cells were observed under identical conditions, and the images of untreated and treated cells were processed similarly. For image processing of cells expressing EGFP-Tat, the gamma value was set to 2.5. Bar – 5µm. *(D)* Estimation of the nuclear accumulation (F_{nuc}/F_{cyt}) of EGFP-Tat and EGFP-TatΔBD in living U2OS cells before and after ATP depletion. The comparisons were performed with Mann-Whitney U tests (n.s. – not significant; ** – $p < .005$, $n > 35$).

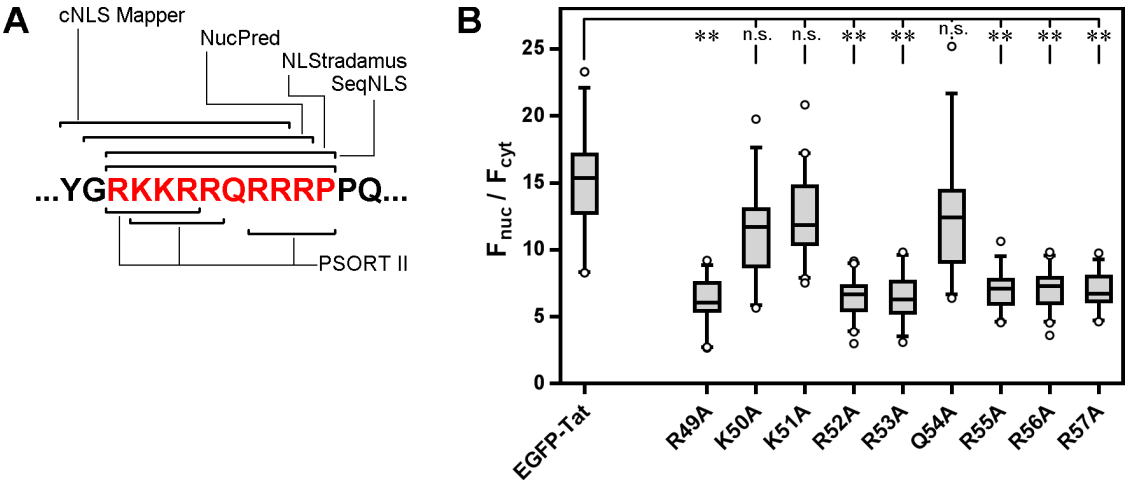

Fig. 2. Prediction and mapping amino acids essential for nuclear accumulation of HIV-1 Tat. *(A)* Prediction of NLS in the BD of HIV-1 Tat by different programs: NLStradamus (2-state HMM static, prediction cutoff 0.6), seqNLS (final-score cutoff 0.86), NucPred (colored from yellow to red), cNLS Mapper (cutoff score 7.0), PSORT II. Amino acids of the BD are highlighted in red. *(B)* Nuclear accumulation (F_{nuc}/F_{cyt}) of EGFP-Tat and its mutants in living U2OS cells. The comparisons were performed with Kruskal-Wallis tests (n.s. – not significant; ** – $p < .005$; $n > 30$).

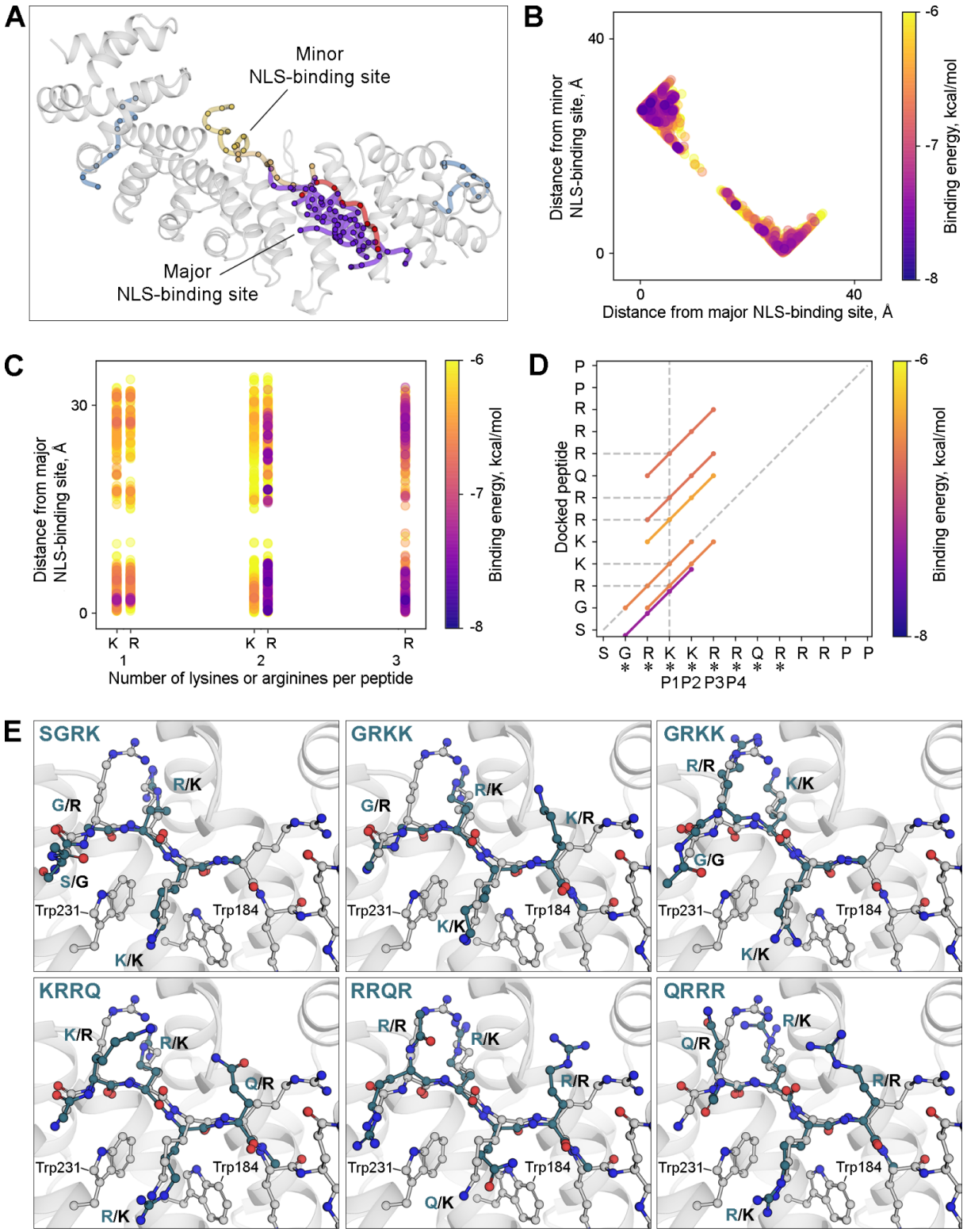

Fig. 3. Molecular docking analysis of interaction of the Tat BD with importin-*α*. *(A)* Poses reported by CABS-dock. This approach utilizes a simplified representation of amino acids (each amino acid is represented by one coarse-grained particle). Each pose corresponds to a cluster of structures obtained during Monte Carlo modeling. Importin-*α* is depicted in gray, and the Tat NLS from the crystal structure is depicted in red. Major NLS poses are colored violet, minor NLSs are colored orange, and off-site structures are colored blue. *(B)* Distribution of docking energies between NLS binding sites obtained in the full-atom docking of the tetrapeptides. *(C)* A number of lysine residues or arginine residues per

tetrapeptide. Estimated docking energies are lower (better) for peptides with more arginine residues. *(D)* Distribution of the best binding energies for tetrapeptides that comprise the NLS sequence. *(E)* The best binding poses of tetrapeptides of the NLS are in agreement with the published crystal structure. The published crystal structure is depicted in gray, and the tetrapeptides are depicted in green.

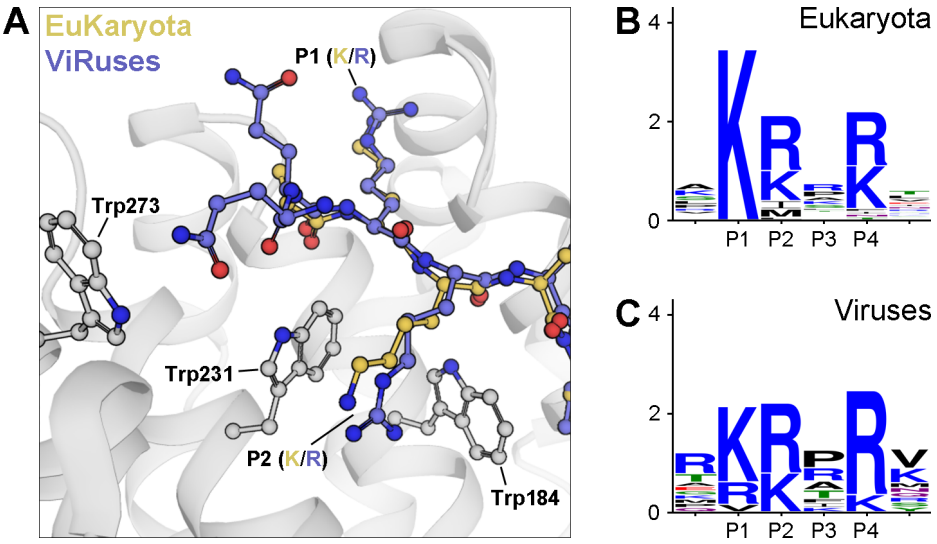

Fig. 4. Differences between the eukaryotic and viral NLS consensus sequences. *(A)* Interaction of a typical NLS in a eukaryotic protein (Ku80, PDB ID 3RZ9) and a viral protein (HIV-1 VPR, PDB ID 5B56) with importin-*α*. *(B)* LOGO sequence of the region directly interacting with importin-*α* obtained from 38 PBD structures formed by importin-*α* and eukaryotic NLSs (supplementary table S1). Consecutive residues from the N-terminal highly conserved lysine residue are referred to as P1, P2, etc. *(C)* The LOGO sequence of the region directly interacting with importin-*α* was obtained from 14 PDB structures formed by importin-*α* and viral NLSs (supplementary table S1). In viral NLSs, the P1 position, which is occupied only by lysine in eukaryotic NLSs, can contain an arginine or a valine residue.

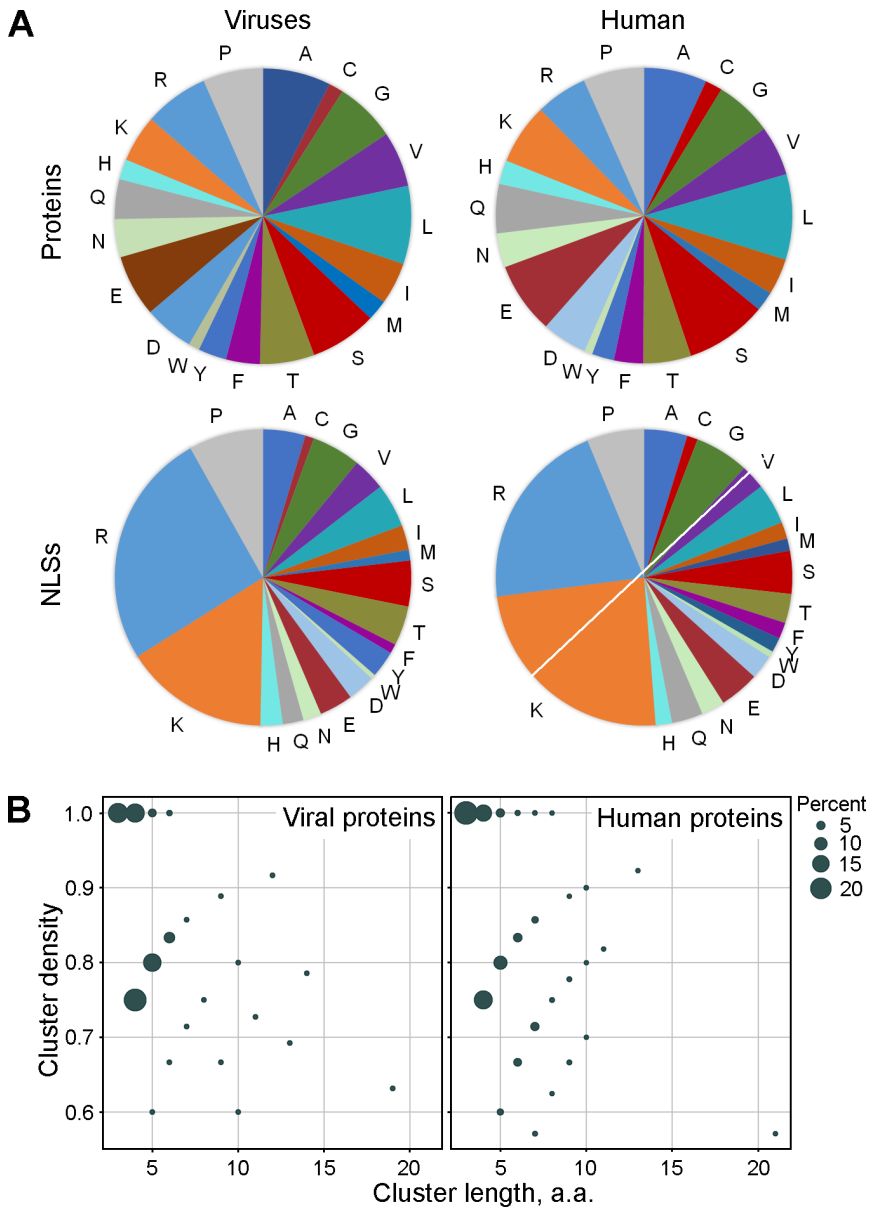

Fig. 5. Heterogeneity of NLS organization in viral and human proteins. *(A)* The average amino acid content of known NLSs in viral and human proteins (top panels) in comparison with the overall amino acid content of these proteins (bottom panels). *(B)* Size of clusters of positively charged amino acids within human and viral experimentally confirmed NLSs. The clusters are defined as sequential arginine residues/lysine residues/histidine residues that are interspersed with no more than one non-positively charged amino acid.

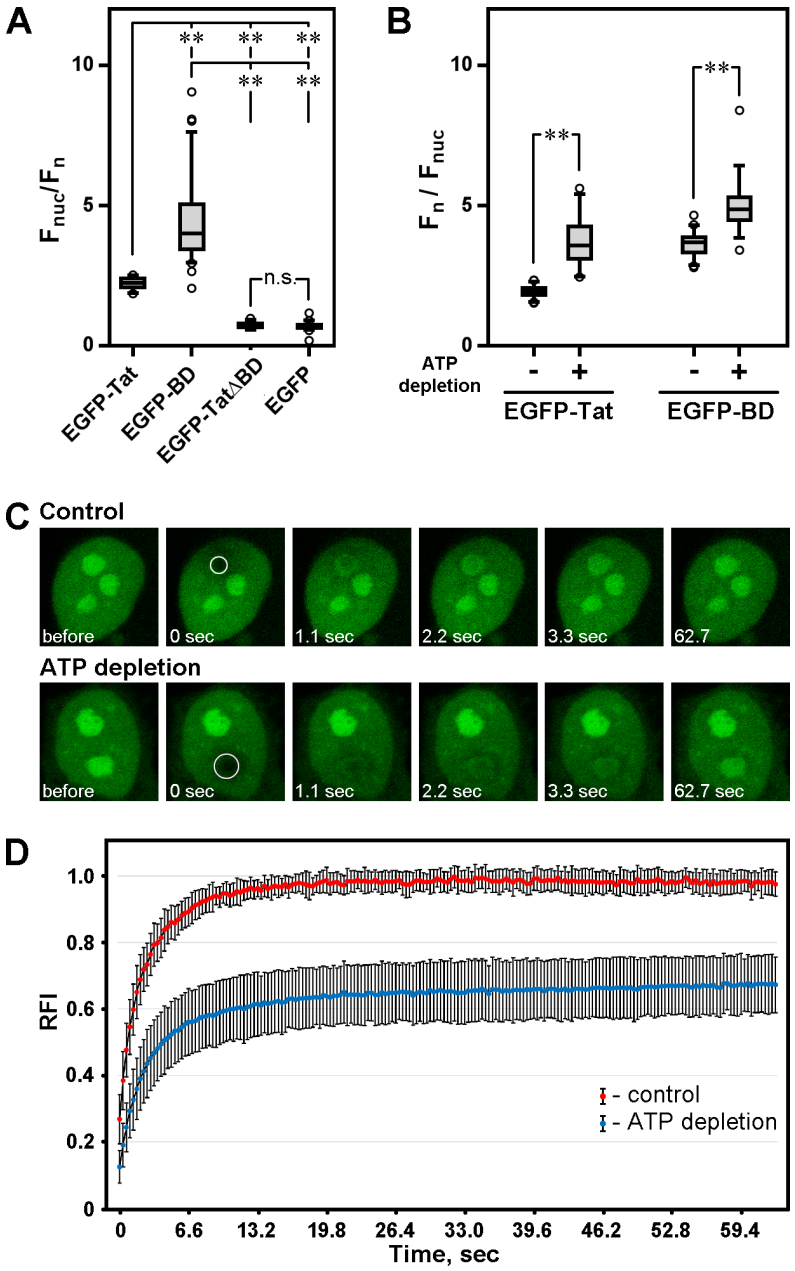

Fig. 6. Nucleolar accumulation of HIV-1 Tat depends on its BD. *(A)* Nucleolar accumulation (F_n/F_{nuc}) of EGFP, EGFP-Tat, EGFP-TatΔBD and EGFP-BD in living U2OS cells. The comparisons were performed with Kruskal-Wallis tests (n.s. – not significant; ** – $p < .005$; $n > 20$). *(B)* Nucleolar accumulation (F_n/F_{nuc}) of EGFP-Tat and EGFP-TatΔBD in U2OS cells before and after ATP depletion. EGFP-NLS^{SV40} was used as a positive control, and EGFP was used as a negative control. The comparisons were performed with Mann–Whitney U tests (n.s. – not significant; ** – $p < .005$; $n > 35$). *(C)* FRAP analysis of the EGFP-Tat interaction with the nucleoli in living U2OS cells. Cells expressing EGFP-Tat were imaged before and after photobleaching (the bleached regions are outlined). The contrast was adjusted to normalized to adjust for the loss of fluorescence during imaging. *(D)* FRAP analysis of EGFP-Tat mobility in nucleoli of the control U2OS cells and after ATP depletion. The results are presented as the means±s.d. (n=28 and 38).

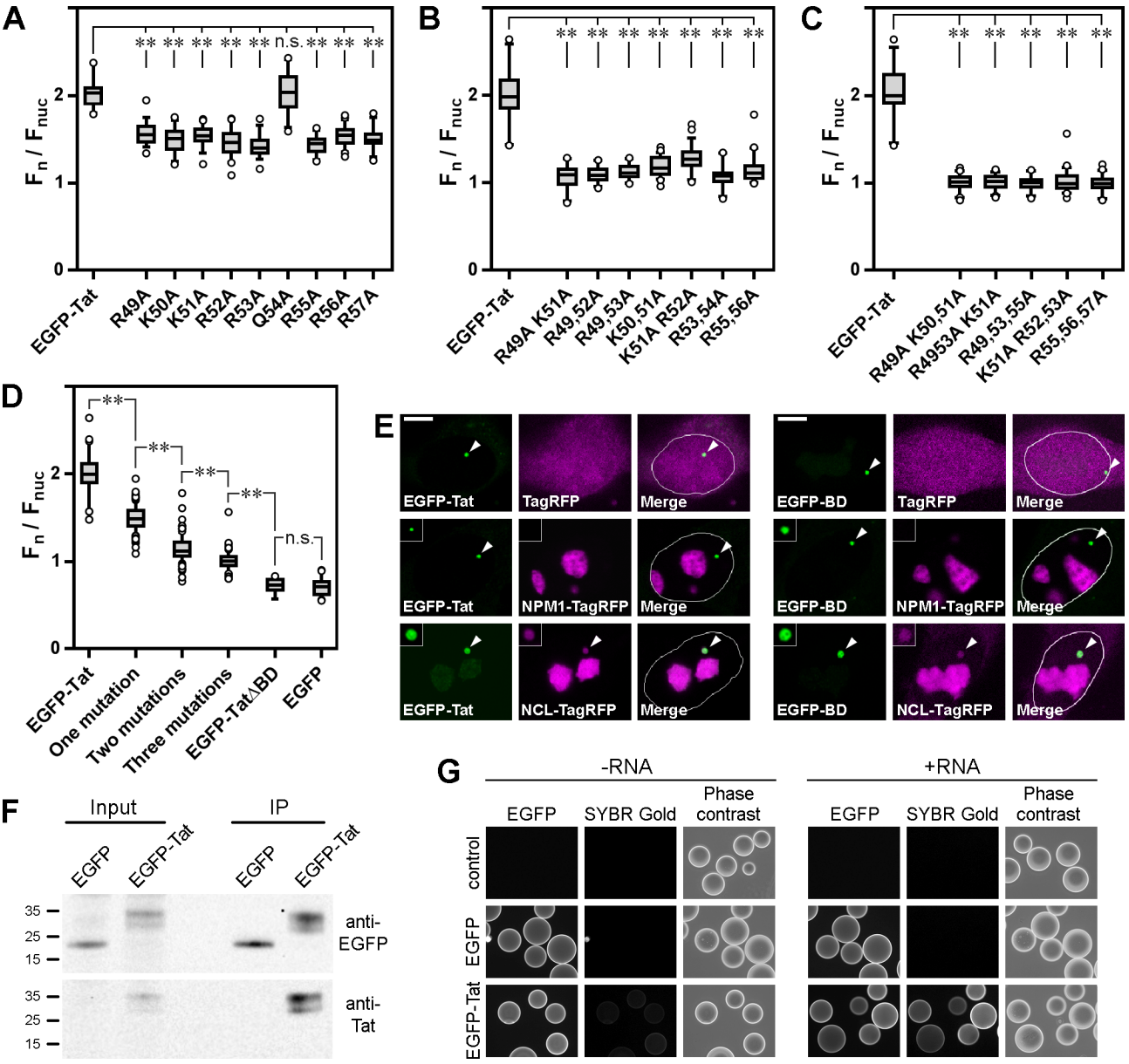

Fig. 7. HIV-1 Tat BD interacts with nucleolar components. *(A)* Nucleolar accumulation (F_n/F_{nuc}) of the EGFP-Tat protein and its mutants (one a.a. substitution in the BD) in living U2OS cells. The comparisons were all performed with Kruskal-Wallis tests (n.s. – not significant; ** – $p < .005$; $n > 20$). *(B)* Nucleolar accumulation (F_n/F_{nuc}) of the EGFP-Tat protein and its mutants (two a.a. substitutions in the BD). *(C)* Nucleolar accumulation (F_n/F_{nuc}) of the EGFP-Tat protein and its mutants (three a.a. substitutions in the BD). *(D)* Nucleolar accumulation (F_n/F_{nuc}) of EGFP-Tat and combined data for all mutants (with the exception of Q54A) with an equal number of substitutions (i.e., one, two or three substitutions irrespective of the precise position of substitution). EGFP-TatΔBD and EGFP were used as negative controls. *(E) In vivo* interaction of HIV-1 Tat and its BD with NPM1 and NCL. The expressed EGFP-fused proteins were immobilized on a platform inside F2H cells (inserts), and the colocalization of the coexpressed proteins fused with TagRFP indicates the interaction between the two proteins. The contrast and brightness were adjusted for display purposes. For image processing of the cells expressing

NCL-TagRFP, the gamma value was set to 2.0. *(F)* Western blotting of the cell lysates (input) and proteins immunoprecipitated on agar beads (IP) using antibodies against EGFP and the Tat protein. EGFP and EGFP-Tat were immobilized for a subsequent study on their interactions with cellular RNA. *(G)* Interaction of RNA with the Tat protein immobilized on agar beads. EGFP or EGFP-Tat proteins expressed in U2OS cells were immunoprecipitated on agar beads and then incubated with or without total cellular RNA from U2OS cells (+RNA or -RNA samples, respectively). The cells were observed and the images processed under identical conditions.

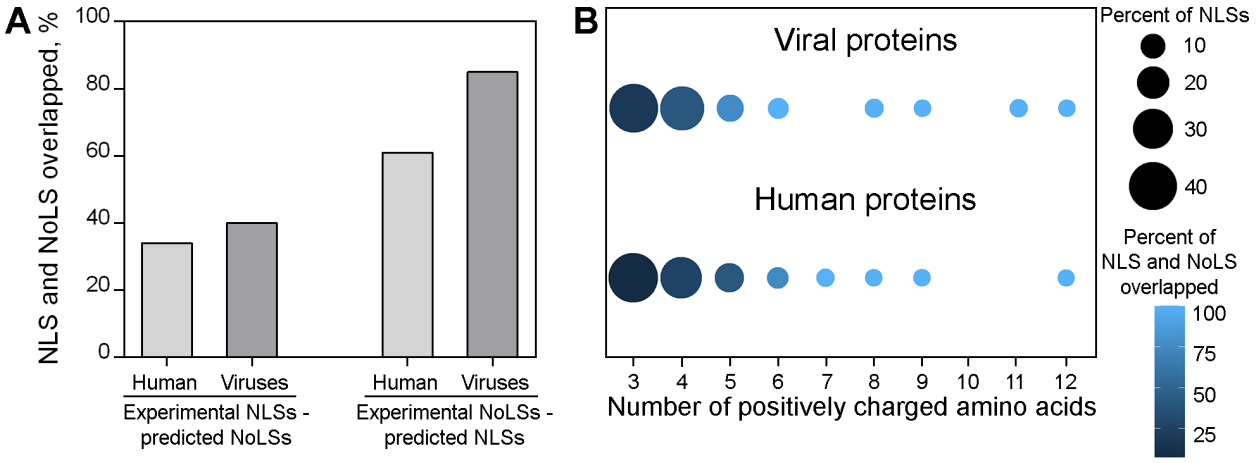

Fig. 8. NLSs overlap with NoLSs in viral and human proteins. *(A)* Overlapping of experimentally annotated NLSs with predicted NoLSs and predicted NLSs with experimental NoLSs. *(B)* Long NLSs more frequently overlapped with the predicted NoLSs.

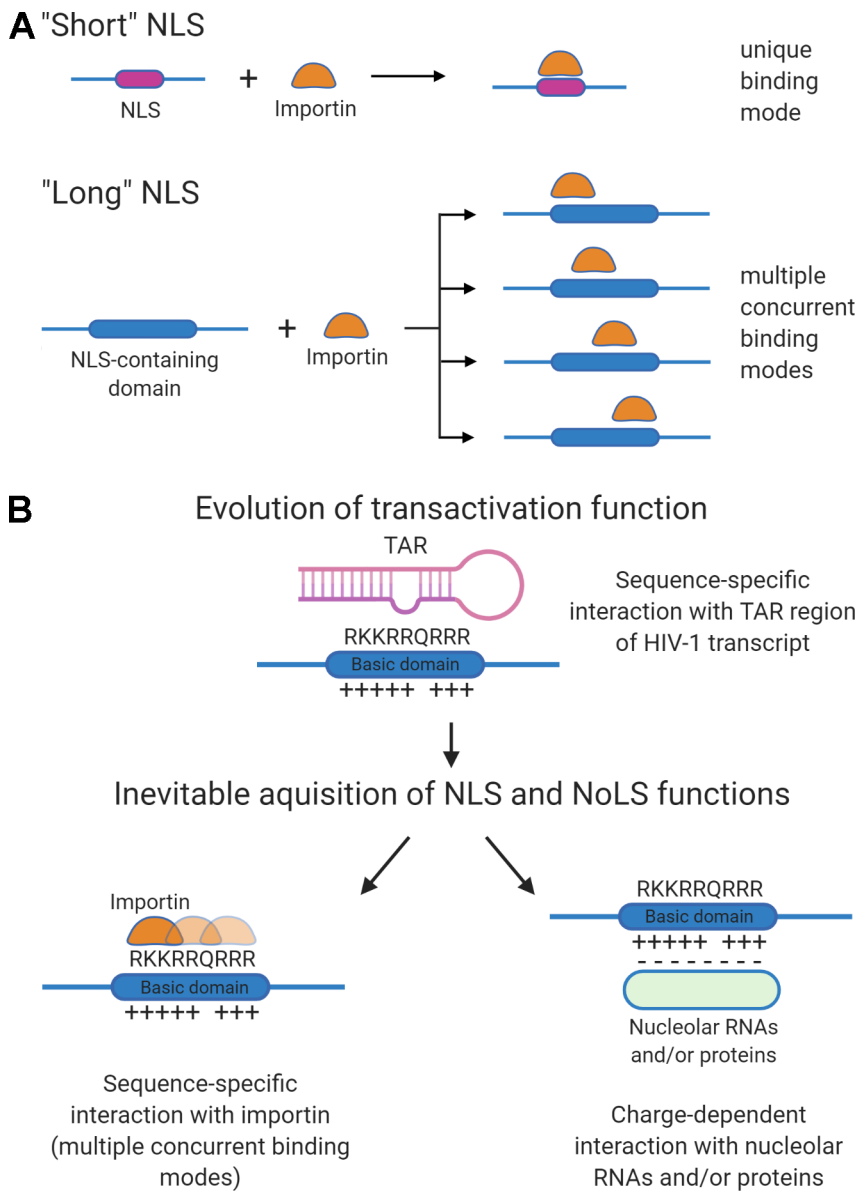

Fig. 9. *(A)* Different regions of NLS-containing domains enriched with positively charged amino acids ("long" NLSs) might interact with importin-*α* (multiple concurrent binding modes), and nuclear accumulation may be a consequence of cumulative binding between different fragments of these regions and importin-*α*. *(B)* Evolution of domains enriched with positively charged amino acids may lead to the inevitable acquisition of additional functions (as an NLS and NoLS, as in the case of HIV-1 Tat), i.e., different functions evolve simultaneously as a complex of integrated activities.