Functional and Structural Segregation of Overlapping Helices in HIV-1

Maliheh Safari¹, Bhargavi Jayaraman^{1,2}, Shumin Yang^{1,3}, Cynthia Smith¹, Jason D. Fernandes^{1,4*}, Alan D. Frankel^{1*}

¹Department of Biochemistry and Biophysics, University of California, San Francisco, CA 94158, USA

²Present address: Synthekine, Menlo Park, CA 94025, USA

³School of Medicine, Tsinghua University, Beijing, 100084, China

⁴Present address: Scribe Therapeutics, Alameda, CA 94501, USA

*Correspondence: Alan D. Frankel

Department of Biochemistry and Biophysics

University of California, San Francisco

600 16th Street

San Francisco, CA 94158-2517

Telephone: 415-476-9994

Fax: 415-514-4112

e-mail: frankel@cgl.ucsf.edu

*Correspondence: Jason D. Fernandes

Scribe Therapeutics

1150 Marina Village Pkwy, Alameda, CA 94501

jason@scribetx.com

Summary

Overlapping coding regions balance selective forces between multiple genes. One possible division of nucleotide sequence is that the predominant selective force on a particular nucleotide can be attributed to just one gene. While this arrangement has been observed in regions in which one gene is structured and the other is disordered, we sought to explore how overlapping genes balance constraints when both protein products are structured over the same sequence. We use a combination of sequence analysis, functional assays and selection experiments to examine an overlapped region in HIV-1 that encodes helical regions in both Env and Rev. We find that functional segregation occurs even in this overlap, with each protein spacing its functional residues in a manner that allows a mutable non-binding face of one helix to encode important functional residues on a charged face in the other helix. Additionally, our experiments reveal novel and critical functional residues in Env and have implications for the therapeutic targeting of HIV-1.

Keywords

HIV-1 evolution, Rev NES, Env LLP-2, overlapping genes

Introduction

HIV-1 contains nine genes encoded within a relatively short genome of 10,000 nucleotides (Frankel and Young, 2002). Eight of these nine genes overlap with another on the DNA level and consequently encode protein products in different reading frames. Despite the prevalence of overlapping genes, the evolution and function of overlapping genes remains understudied. However, evidence establishing the importance of overlaps continues to grow: the SARS-CoV-2 pandemic has demonstrated the importance of identifying and studying overlapping reading frames, as many immune epitopes to overlapping ORFs have been identified in patients (Nelson *et al.*, 2020; Weingarten-Gabbay *et al.*, 2020; Finkel *et al.*, 2021).

We previously showed that, in the case of the essential regulatory genes *tat* and *rev*, HIV-1 dedicates the nucleotide sequence to one protein or the other so that critical

functional residues do not overlap with one another (Fernandes *et al.*, 2016). For *tat* and *rev*, this arrangement is possible because both proteins consist of functional motifs spaced by disordered linkers, and thus functional motifs of one protein can be encoded in regions that overlap with disordered linkers in the other. Although the vast majority of overlapped proteins are predicted to be disordered (Rancurel *et al.*, 2009) and thus have the potential for this arrangement, it is possible that some overlapped regions have more stringent structure and function requirements. Indeed, HIV-1 contains an intriguing overlapping region, between the Rev and Env genes (Figure 1A), in which both protein segments are structured and functionally important. This region might represent a viral Achilles' heel, vulnerable to therapeutic targeting, as the mutational landscape for escape mutations would be constrained by the simultaneous requirements of both proteins.

Rev is required for the nuclear export of partially spliced and un-spliced viral RNAs that encode necessary late stage viral proteins as well as the genomic RNA for encapsidation (Malim and Cullen, 1991; Pollard and Malim, 1998; Rausch and Le Grice, 2015). Notably, Rev contains a quasi-helical nuclear export sequence (NES), that binds to the NES-binding groove of the host nuclear export factor Crm1 (Fornerod *et al.*, 1997; Booth, Cheng and Frankel, 2014). Env is expressed as a polyprotein, composed of the gp120 surface and gp41 transmembrane proteins, which forms a trimeric viral envelope that mediates viral particle binding to cell entry receptors (Checkley, Luttge and Freed, 2011; Murphy *et al.*, 2017). The cytoplasmic tail of gp41 is the most C-terminal portion of Env and is highly variable in size (Tedbury and Freed, 2015). In contrast to other retroviruses, lentiviruses such as HIV-1 have an extended cytoplasmic tail containing the Kennedy epitope (KE) followed by three membrane associated lentiviral lytic peptide motifs (LLP-1, LLP-2 and LLP-3) (Steckbeck, Kuhlmann and Montelaro, 2013) (Figure 1A). HIV-1 LLP-1 and LLP-2 contain highly conserved arginine and lysine residues that are involved in cation- π interactions and have been suggested to stabilize the cytoplasmic tail α -helical structure (Murphy *et al.*, 2017). LLPs are amphipathic α helices that embed in the virus membranes (Kalia *et al.*, 2003; Steckbeck *et al.*, 2011; Murphy *et al.*, 2017). LLP-2 may be transiently exposed on the virus

surface to interact with the gp41 core and regulate virus-cell membrane fusion (Lu *et al.*, 2008; Murphy *et al.*, 2017; Piai *et al.*, 2020). A recent nuclear magnetic resonance (NMR) spectroscopy study revealed that LLP-2 forms two amphipathic helices that wrap around the transmembrane domain of Env, arranging a “baseplate” that supports the transmembrane domain and the rest of the Env trimer (Piai *et al.*, 2020). Additional evidence suggests that the C-terminal tail makes critical contacts with the matrix protein (Alfadhli *et al.*, 2019) to facilitate Env incorporation into virions.

The Rev NES and a portion of the Env LLP-2 are encoded by the same nucleotides in the viral genome, with Env encoded in the +1 reading frame relative to Rev (Figure 1A). As both these domains have important functions in viral replication and are structured, they provide an ideal system to examine how viruses balance the evolution of these two genes with critical, overlapping structured motifs. Here we perform conservation analyses, functional assays and selection experiments to examine how the virus balances selection pressures within the Rev-Env overlap. Surprisingly, we find that functional residues are also segregated at the nucleotide level, as for the Rev-Tat overlap. In Rev-Env case, the two-dimensional segregation of nucleotides in the genome is mirrored in the three-dimensional functional surfaces of the proteins such that the nucleotides encoding a functional surface in the helix of one protein encode residues on the non-binding face of the overlapping helix. Furthermore, our analyses reveal three important charged residues on the charged surface of LLP-2 that are required for Env function.

Results

Conservation of chemical properties in the Rev NES and Env LLP-2

The Rev NES (residues 75-83) binds to Crm1 using several critical hydrophobic residues and non-canonical prolines that change the helical register and position hydrophobic side chains on one face of the helix to fit within the NES recognition groove (Dong *et al.*, 2009; Güttler *et al.*, 2010) (Figure 1B). The Env LLP-2 forms an

amphipathic helix about 33 amino acids long (Eisenberg and Wesson, 1990; Kalia *et al.*, 2003; Murphy *et al.*, 2017) that spans residues 767-788 of Env (242-274 of gp41) (Figure 1A). Nine of the LLP-2 residues (767-776) overlap with the quasi-helical Rev NES (Figure 1B). Mutating charged residues within LLP-2 or hydrophobic residues in the Rev NES negatively impacts viral replication (Kalia *et al.*, 2003; Newman *et al.*, 2007; Güttler *et al.*, 2010; Fernandes *et al.*, 2016), prompting us to examine in detail the selective forces that guide the evolution of this overlapping region.

In order to measure the relative conservation of these regions within HIV patient sequences, we first downloaded curated alignments of the Los Alamos Database (hiv.lanl.org) (Foley *et al.*, 2013) and calculated Shannon entropies for each amino acid in both proteins within the NES/LLP-2 overlap (Pan and Deem, 2011; Fernandes *et al.*, 2016). Sites with low entropy (0 = perfect conservation) indicate strong conservation of the amino acid in that particular protein, but it is impossible to determine if the observed conservation is simply driven by codon constraints imposed by the alternative frame. To compare the amino acid conservation of corresponding sites in both proteins, we grouped amino acids that share two nucleotides in the genome and plotted their entropies against one another (Fernandes *et al.*, 2016) (Figure 2A). Notably, the one site that contributes to both functional surfaces (H769 in Env and P76 in Rev) is relatively well conserved in both proteins whereas the one site that contributes to neither interface (L755 in Env and T82 in Rev) is weakly conserved (Figure 2A). Unexpectedly, two known essential Crm1 binding residues (L78 and L83) displayed low conservation (high entropy) in the Rev reading frame, but the consensus sequence (Figure 2B) indicates that chemically conservative isoleucine substitutions are frequently found, thus explaining the high entropy. Moreover, entropy analysis using a reduced amino acid alphabet in which the chemically similar amino acids I/L/V are grouped, produced low entropy values for all Crm1 binding residues (Figure 2C). The use of chemically conservative side chains to form critical contacts drastically increases the number of “synonymous” codons (e.g. the six leucine codons can substitute with the three isoleucine codons with essentially equal fitness) and may reflect the kinds of interaction motifs that are most suited to evolve in constrained overlapping regions. The

three charged residues (Figure 2C; green dots) in LLP-2 were relatively strongly conserved (Figure 2B) but showed lower conservation (high entropy) than the Rev NES binding residues, suggesting that Rev may more greatly influence the selection of amino acids in the overlapping region.

Deep mutational scanning of non-overlapped viruses selects for important LLP-2 and NES residues

In order to deconvolute the functional constraints for every residue in both proteins absent the overlap, we created libraries of viruses in which Rev and Env were decoupled from one another. Although HIV-1 *env* has been exposed to deep mutational scanning previously, the overlap with *rev* has prevented exploration of the LLP-2 (Haddox, Dingens and Bloom, 2016; Haddox *et al.*, 2018). Briefly, we used the NL4-3 *rev-in-nef* virus in which *rev* is moved to the non-essential, single-frame *nef* locus (Fernandes *et al.*, 2016), and the endogenous *rev* locus is ablated, thus allowing the endogenous Env LLP-2 region and the *rev-in-nef* NES region to evolve independently from one another. We individually randomized each of the LLP-2 residues in the endogenous *env* locus, creating 10 Env libraries that each represent 64 alleles (every possible codon). Next, we infected SupT1 cells at low MOI (multiplicity of infection) with each library individually and measured the change in allele frequencies after approximately 6 generations of viral spread (Figure 3A). We then combined analyses of these libraries with previously generated deep mutational scanning data produced for the Rev NES from a *rev-in-nef* virus (Fernandes *et al.*, 2016). These fitness landscapes (Figures 3B and 3C) represent the amino acid preference of each gene when the overlap is removed.

The LLP-2 fitness landscape shows that while LLP-2 is generally highly mutable, it has a strong preference for a negatively charged side chain (D or E) at residue 773 (Figure 3B). Although the importance of positive charges in LLP-2 has been suggested previously (Kalia *et al.*, 2003; Steckbeck *et al.*, 2011), our experiment is the first to identify strong selection for a negative charge. The data also show strong selection

against the “helix-breaker” proline (Li *et al.*, 1996) at all positions, further underscoring the importance of its helical structure. Stop codons are robustly selected against as well, demonstrating the importance of the cytoplasmic tail in HIV-1 replication. Additionally, L771, which is surrounded by two positively charged arginines displays a strong preference against charged residues. The Rev NES fitness landscape is consistent with many mutational studies, showing selection for large hydrophobic side chains (ILV) at each Crm1 contacting residue. The patient isolates do not show such a striking preference (Figure 2B), perhaps reflecting constraints imposed by overlapping gene mutational paths (e.g. any transitional mutation from L to V would have to be fit in Rev and also in Env) or the existence of other *in vivo* pressures not captured in our experiments (e.g. immune or cell-type specific responses). Unlike LLP-2, we do not observe selection against proline in the semi-helical Rev NES, consistent with the presence of prolines in its reference sequence.

Functional assays confirm the importance of Rev NES hydrophobic residues and LLP-2 charged residues

To confirm that our selection data reflect Rev functional requirements, we assayed individual alanine mutants at the nine NES residues in a p24 Rev-dependent RRE reporter assay (Smith *et al.*, 1990). Mutation of any of the leucines (positions 75, 78, 81, and 83) or proline at position 76 resulted in loss of function (Figure 4A; high p24 = active Rev), consistent with previous data of Crm1 binding residues (Figure 1C) (Fornerod *et al.*, 1997; Güttler *et al.*, 2010; Booth, Cheng and Frankel, 2014; Jayaraman *et al.*, 2019). Consistent with our selection data, mutation of the NES non-binding face had limited impact on Rev function.

To confirm that our selection data reflect LLP-2 functional requirements, we measured the infectivity of virions generated from pseudotyped *env*-deficient viruses complemented with Env LLP-2 alanine mutants *in trans*. The D773A mutation produced the most drastic phenotype (15% of wild-type Env), consistent with our selection data (Figure 4B). Previous work noted the conservation of arginines within the LLPs (Kalia *et*

al., 2005; Steckbeck *et al.*, 2011) and our data confirms that the first arginine in LLP-2 (R770) is also important for Env function, in addition to D773.

Charged LLP-2 residues facilitate Env function in virus replication assays

As D773 in LLP-2 is clearly an important residue based on the selection and reporter data, we examined the effects of several substitutions at this position on viral replication. D773E, a conservative change predicted to maintain function based on our fitness data, showed ~50% of wild-type activity in the reporter assay, but substantially higher than the non-conservative D773A and D773N substitutions (Figure 5A). These results correlated well with the replication of rev-in-nef viruses engineered with the mutations (Figure 5A). It is interesting that D773A viruses still can replicate, albeit poorly, despite its very low activity; these results suggest that the virus can partially compensate for the loss of function, particularly when not competing with the wild-type sequence. The weakly active D773N allele is observed in ~2.6% of patient isolates (Figure 2B), further suggesting that the virus has alternative mechanisms to achieve the function normally performed by the negative charge at this position.

Given the importance of D773 and the neighboring positively charged R770 and R772 (Figure 4B), we sought to establish the importance of charge and spacing along the LLP-2 helix, with the goal of understanding what evolutionary constraints - absent the overlap - affect this region. We generated two double mutants with D773/R770 or D773/R772 changed to alanine (R1D→AA or R2D→AA) as well as the triple mutant (RRD→AAA) and found that mutation of R770 was particularly detrimental, in combination with D773 which is about one helical turn away (Figure 5B). Both of the double mutants showed significantly decreased virus replication (Figure 5B), again suggesting loss of function can be partially compensated in the viral context for the loss of function. The triple mutant was highly defective in both reporter and replication assays indicating the importance of these charged residues.

We constructed a set of additional charged residue mutants, including H769 as well as substitutions with lysine, to determine if their contributions could be attributed solely to charge. The results confirm the importance of R770 and R773 but not H769 (Figure 5C). It is interesting that substitution of arginine with lysine is insufficient to maintain function or replication, and indeed the (RR→KK) mutant showed the most severe replication defect. Western blots of these virions indicated decreased Env protein stability, and it is possible that lysine substitutions might serve as a target for host restriction factors (Zhang *et al.*, 2021) or introduce ubiquitination sites leading to protein turnover. Regardless of the mechanism, the data suggest that these particular residues are subject to significant evolutionary constraints. The results are also consistent with analyses of the Env cytoplasmic tail, which has shown enrichment for arginine and depletion for lysine when compared to the rest of Env and all other proteins found in the Uniprot database (Steckbeck *et al.*, 2011).

Helical requirement for LLP-2 function

LLP-2 forms two helices (Murphy *et al.*, 2017) and our deep mutational scanning data suggest strong selection against helix-breaking prolines, implying a structural constraint during evolution. To better understand the extent of this constraint, we replaced residues 767–776 with glycines and serines (10GS) to create a flexible linker. As anticipated, the 10GS variant had no activity in the reporter assay and was unable to replicate even when D773 was reintroduced (Figure 6A). We next replaced residues 767–776 with alanines (10A) which are often used to form rigid helical linkers (Chen, Zaro and Shen, 2013). This 10A variant was similarly defective (Figure 6B). However, in this case, reintroducing important charged residues (R770 or D773 or the 5-residue HRLRD segment) resulted in substantial restoration of activity and rescue of virus replication, particularly in the cases where D773 was present (Figure 6B). These data further confirm the importance of charge and helical structure in LLP-2 function (Kalia *et al.*, 2005; Boscia *et al.*, 2013; Steckbeck, Kuhlmann and Montelaro, 2014).

Functional segregation of residues in overlapping structured segments

Our experiments establish the functional restrictions in Env LLP-2 without competing selection pressure from Rev, permitting us to map the constraints of each reading frame onto the nucleotide sequence of the virus (Figure 7). Specifically, our analysis showed that in Env, D773 has a strong preference for negative charge, R770 and R772 have an intermediate preference for arginine side chains, H769 has a weak preference for a positive charge, and L771 selects against charged side chains. When combined with the preferences of the Rev NES Crm1-binding residues (Fernandes *et al.*, 2016), it is apparent that the functional surfaces of each domain are positioned so as to minimize negative consequences of the overlap. The non-binding face of the Rev NES largely corresponds to the charged surface of the Env LLP-2 and the hydrophobic surface of the LLP-2 corresponds to the binding residues of the Rev NES (Figure 7). The majority of overlaps between opposing critical residues mostly involve mutable wobble bases, with the sole exception being the Env H769/Rev P76 overlap. However, in this case Rev appears to be the major driver of evolution as mutation of P76 results in a severe loss of function whereas H769 appears to play a minor, supporting role in the function of the charged surface of LLP-2. Env L771 also overlaps Rev L78, but contributes relatively minor constraints at the nucleotide level, as its selection against polar residues, rather than for a particular side chain, permits multiple codons with high fitness in LLP-2.

LLP-2/NES overlaps in other lentiviruses

The segregation of functional residues in LLP-2 and NES prompted us to examine the Env/Rev overlap among primate lentiviruses. The Env cytoplasmic tail is moderately conserved between HIV-1 clades and the extent of their sequence variation is similar in regions that do or do not overlap with Rev (Steckbeck *et al.*, 2011). This observation is consistent with our finding of functional segregation, where the overlapping region essentially acts like two single reading frames. The four distinct HIV-1 lineage groups (M, N, O, P) share similarity with other primate lentiviruses that originated from two different simian immunodeficiency viruses (SIV) and all have overlapping LLP-2/NES regions (Figure 8).

To explore these regions in depth, we extracted Rev NES and Env LLP-2 sequences from curated amino acid alignments of HIV-1/SIVcpz, HIV-2/SIVsmm, and other SIVs in the Los Alamos Database (2018; hiv.lanl.org) and compared sequence logos for each virus (Crooks et al., 2004; Foley et al., 2013) (Figure 8). The hydrophobic Crm1-binding residues of the NES are highly conserved across different HIV-1, HIV-2 and SIVs (Figure 8) but the NESs can be broadly grouped into two categories: one contains the hydrophobic/proline arrangement observed in HIV-1 and the other contains HIV-2-like characteristics that preserve the spaced hydrophobic contacts but lack consecutive prolines and instead are glutamine-rich. LLP-2 is more varied and viruses fall into three groupings: 1) HIV-1 M, N and SIVcpz*Ptt* conserve the charge cluster dissected in our assays; 2) HIV-1 O, HIV-1 P, SIVgor, and SIVcpz*Pts* have minimal charge; and 3) SIVmac, HIV-2, SIVsmm and SIVrcm conserve R772 and occasionally D773. The high conservation of hydrophobic Crm1 contacts appears to dominate selection for the NES while maintaining sufficient sequence flexibility for LLP-2 to retain its structure. Indeed, the size of the Env cytoplasmic tail varies among different retroviruses (Tedbury and Freed, 2015), and the loss of the tail is commonly observed when SIV is propagated in human cells, suggesting that both host and viral determinants influence the function and variability of this region (Kodama *et al.*, 1989; Shimizu *et al.*, 1992; Shacklett *et al.*, 2000).

Discussion

We previously found that overlapped genes (*rev* and *tat*) can segregate functionally important residues by having a disordered segment of one protein overlapped with an ordered segment of the other (Fernandes *et al.*, 2016). The present study shows that residues can functionally segregate even when both overlapping segments are structured. We probed residues in the Env gp41 LLP-2 that overlap with the Rev NES and compared the limits to which this domain can tolerate mutation to the already well-established mutational limits of Rev (Jayaraman *et al.*, 2019). Interestingly, we observed that critical residues in both proteins are located on one face of a helix such that they

are spaced periodically in alternative reading frames (Figure 7). This arrangement is a consequence of the genetic code where the wobble bases of functionally critical codons in one reading frame can be chosen to accommodate critical residues in the alternative frame, while the first two codon bases overlap with more mutation-tolerant residues located on a non-binding surface of the overlapped protein (Figure 7). Specifically, the important hydrophobic residues of the Rev NES are oriented on one surface towards the Crm1 binding groove, with consecutive prolines altering the helical register compared to other Crm1 cargoes (Booth, Cheng and Frankel, 2014; Jayaraman *et al.*, 2019). The non-binding face of the NES helix is highly mutable and capable of sampling nearly every amino acid with roughly equal fitness. These highly mutable residues correspond to nucleotides that encode critical charged residues on the functional surface of gp41 LLP-2 in the alternative frame. Correspondingly, residues on the non-charged face of the LLP-2 helix minimally constrain codon usage of important Rev NES contact residues.

The striking segregation of function we observe for Env LLP-2/Rev NES is made possible by the atypical proline-rich NES and partially redundant function of the LLP-2 positive charges. This manner of segregation in HIV-1 is not necessarily the case for similarly overlapped regions in other lentiviruses, which appear to have different functional requirements based on their sequence conservation (Figure 8). The findings are another indication of how efficiently HIV-1 makes use of its limited coding capacity.

Our results show that Env D773 is under strong selective pressure although the precise function of this residue is not known. In addition to this negative charge, positive charges positioned on the same face of the helix are also important. Previous studies have suggested several possible functions for the LLP-2, including maintaining Env in a prefusogenic state (Lu *et al.*, 2008; Postler and Desrosiers, 2013; Murphy *et al.*, 2017), regulating viral assembly and egress (Snetkov *et al.*, 2021), or mediating matrix-Env interactions (Alfadhli *et al.*, 2019). The function of these charged residues in viral fusion or assembly remains to be determined.

HIV-1 has evolved in an elegant manner to balance the functions of overlapped proteins, however the strategy may also expose a weakness whereby immune targeting of the LLP-2 could provide a fruitful therapeutic route. The LLP-2 is transiently surface exposed on the cell membrane during infection (Kalia *et al.*, 2005; Lu *et al.*, 2008; Murphy *et al.*, 2017), and the cytoplasmic tail can modulate the antigenic properties of Env as a whole (Piai *et al.*, 2020). Targeting of the LLP-2 surface with combinations of rationally designed immunotherapies may force the virus to mutate in ways that break the segregation between Env and Rev. Although our analysis of other lentiviruses suggests that HIV-1 might evolve other ways to establish Rev function under such pressure, it is likely that escape would still require multiple mutations and may be less likely in combination with other therapies. Deep mutational scanning has proven effective at mapping probable mutational escape pathways from immune pressures such as neutralizing antibodies (Starr *et al.*, 2020; Greaney *et al.*, 2021) and the Env cytoplasmic tail has proven to be a viable target in primate animal models (Wang *et al.*, 2020); for HIV-1 the combined knowledge of both functional constraints in the proteins and coding constraints in the genome may allow targeted immune therapies that drastically limit (or prevent altogether) escape mutations. Indeed, immune targeting of overlapped reading frames in in-vivo models of has been shown to drive predictable synonymous mutations in the alternative, non-targeted frame (Hughes *et al.*, 2001). Similar approaches in this region of dual structure and function may provide a novel way of limiting HIV's otherwise unpredictable mutational escape patterns.

Acknowledgments

We thank all members of the Frankel lab for helpful comments and review of the manuscript and members of the John Gross lab for beneficial discussion. We also thank Henni Zommer for making some of the plasmids containing different Env LLP2 mutations. This work was supported by NIH grant P50AI150476, and MS was supported in part by NIH training grant T32AI0605357.

Author Contributions

Conceptualization: M.S., B.J., J.D.F., and A.D.F.

Investigation: M.S., B.J., J.D.F., S.Y., C.S.

Writing: M.S., B.J., J.D.F. and A. D. F.

Reviewing & Editing: M. S., B.J., J.D.F., and A.D.F.

Supplemental Information

Table S1: Raw data used to generate graphs for Figure 2.

Table S2: Raw data used to generate heatmap for Figure 3A.

Table S3: Raw data used to generate heatmap for Figure 3B.

Table S4: Raw data used to generate graphs for Figure 4.

Table S5: Raw data for Env Reporter assays in Figures 5 and 6.

Table S6: Raw data for Env Spreads in Figures 5 and 6.

Figure legend

Figure 1. Organization of Rev NES/Env LLP-2 overlap

A) Genetic organization of the *rev/env* overlap. (Top) The full genome organization of HIV-1 (HXB2 numbering) demonstrates that the second exon of *rev* overlaps *tat* (second exon) and *env*. A detailed annotation of this overlap depicts the overlap between the quasi-helical NES of Rev with the helical LLP-2 of Env. The ARM (arginine rich motif, nuclear localization sequence), OD (oligomerization domain), and NES (nuclear export sequence) are other functional domains of Rev that overlap with Env-gp41(Cytoplasmic tail), including the KE (Kennedy epitope) and two lentiviral lytic peptide motifs (LLP-3 in addition to LLP-2).

B) The structure of the Rev NES (red: binding residues, white: non-binding residues) bound to Crm1 (yellow) (PDB: 3nc0). A helical wheel diagram demonstrates how the arrangement of the 2D amino acid sequence creates a distinct binding surface in 3D.

C) Composite modeled structure (PDB: 6ujv) of the Env C-terminus trimer, including the first 21 residues of the LLP-2. The LLP-2 is labeled with the hydrophobic face in gray and the charged surface (positively charged residues in red and negative in blue). A corresponding helical wheel diagram derived from (Murphy *et al.*, 2017) for the Env LLP-2 is also shown with charged residues colored.

Figure 2. Rev NES and Env LLP-2 conservation

A) Shannon entropies (low entropy = high conservation) from alignments of patient sequences (see panel B). Residues in Rev and Env are grouped by positions that share two nucleotides of overlap. Red points contain Crm1 contact residues, green points are charged LLP-2 residues and yellow points contain both.

B) Sequence logos of Rev NES and Env LLP-2 showing sequence conservation at each position computed from a sequence alignment of patient data from the Los Alamos Database (<https://www.hiv.lanl.gov/content/index>; 2018). Positively charged residues (KRH) are in blue; negatively charged residues (DE) are in red.

C) Shannon entropies recomputed for a simplified amino acid alphabet (D = E, R=K, I = L & V, M, F = W & Y, N = Q, S = T). Coloring is as in 2A.

Data for Figure 2 is provided in Table S1.

Figure 3. Deep mutational scanning of Rev NES and Env LLP-2

A) Schematic of the experiment. Non-overlapped viruses with *rev* inserted in the *nef* locus were created and the endogenous LLP-2 region randomized to NNN to create 10 independent libraries. These libraries were then arrayed and allowed to spread (~6 generations of viral spread) and next-generation sequencing was used to determine frequencies of each variant in a replication competition experiment and to calculate experimental fitness.

B) Fitness (blue = unfit, white = neutral and gold = fit) for every allele in the region of the Env LLP-2 that overlaps the Rev NES. Row headers denote the reference sequence while column headers specify the allele.

C) Fitness for every allele in the region of the Rev NES that overlaps the Env LLP-2.

Data for Figure 3 is provided in Table S2 & Table S3.

Figure 4. Functional dissection of Rev NES and Env LLPs

A) Schematic of the Rev reporter assay and activities of individual NES alanine mutants. To monitor Rev export activity, a RRE-containing gag-pol reporter system that produces viral capsid (p24) in a Rev-dependent manner was used. Data are mean $\pm$ standard deviation (s.d.) of three biological replicates. Western blot shows expression levels of Strep-tagged Rev variants and GAPDH as loading controls.

B) Schematic of the Env reporter assay in the TZM-bL system and results with individual Env LLP-2 alanine mutants. Data are mean $\pm$ s.d. of three biological replicates. Western blots show Env incorporation (gp120 antibody) into pseudoviruses and p24 as loading controls. For these experiments, we ablated the *env* reading frame by mutating the start codon and introducing multiple downstream stop codons. Plasmids were co-transfected into 293T cells with each alanine mutant to generate Env pseudotyped viruses. The resulting virions were used to infect TZM-bL reporter cells which express the HIV-1 surface receptors CD4, CCR5, CXCR4, and produce luciferase under the control of the *tat*-dependent HIV-1 LTR (Sarzotti-Kelsoe *et al.*, 2014). A functional Env protein permits binding and entry and expression of Tat and thus reflects the number of infectious viral particles in a single round of infection, but the integrated virus lacks the *env* gene and cannot initiate additional rounds of replication. Incorporation of Env mutants into virions was also measured by Western blot. Data for Figure 4 is provided in Table S4 & S5.

Figure 5. Importance of LLP-2 charged residues for replication

A) Left: Env functional assays as described in Figure 4 with different D773 mutations. Right: Viral replication kinetics in spreading assays using NL4–3 rev-in-nef proviral plasmids (Fernandes *et al.*, 2016) engineered with Env D773 mutations.

B) Left: Env functional assays showing the role of three charged residues. Western blots show incorporation of Env mutants in virions (dark boxes represent a break in the lane loading presentation, with all samples loaded on a single gel).

Right: Viral replication kinetics of viruses engineered with the mutations.

C) Left: Env functional assays showing effects of positive charge mutations with corresponding Western blots of Env incorporation in virions.

Right: Viral replication kinetics of viruses engineered with the mutations.

Replication assays are represented on a log scale. All data are mean $\pm$ standard deviation (s.d.) of biological triplicates.

Data for Figure 5 is provided in Table S4 &S5.

Figure 6. Importance of LLP-2 structure and charge for Env activity

A) Left: Env functional assays as described in Figure 4 when LLP-2 helical structure was disrupted by replacing 10 residues (767–776) with alternating glycine/serine residues (10GS) to create a flexible region. Western blots show expression of Env mutants (dark boxes represent a break in the lane loading presentation, with all samples loaded on a single gel).

Right: Viral replication kinetics of WT, 10GS and 10GS+D262 viruses.

B) Left: Env functional assays with charge mutants within a 10-alanine (10A) context designed to maintain a helical structure with corresponding Western blots of Env incorporation in virions .

Right: Viral replication kinetics with helical structure mutant viruses.

Data points in both panels are mean $\pm$ s.d. of biological triplicates. Replication assays are represented on a log scale.

Data for Figure 6 is provided in Table S4 &S5.

Figure 7. Nucleotide segregation and protein structure of overlapping Rev NES and Env LLP-2.

Functionally important residues of the Rev NES (red, CRM1 binding residues) and Env LLP-2 (green, charged residues) are highlighted in their respective structures (PDB: 3nc0 and PDB: 6ujv) and shown with the corresponding nucleotide sequence.

Figure 8. LLP-2/NES overlaps in primate lentiviruses reveals variation of NES and LLP-2

Phylogenetic relationship of different primate lentiviruses, including four HIV-1 groups (M, N, O, P), SIVcpz*Pts* (from eastern chimpanzees), SIVgor (from gorillas), SIVcpz*Ptt* (from central chimpanzees), HIV-2, SIVsmm (from sooty mangabeys), SIVmac (from macaques), and SIVrcm (from red-capped mangabey).

Sequence logos showing conservation of the corresponding Rev-NES and Env-LLP2 regions of each virus computed from their sequence alignments (extracted from 2018 curated amino acid alignments of HIV-1/SIVcpz, HIV-2/SIVsmm, and other SIV from the Los Alamos Database (hiv.lanl.org)).

Positively charged residues (KRH) are indicated in blue and negatively charged residues (DE) in red.

Star Methods

Key Resources Table

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Cell Lines		
293T Cell Line	American Type Cell Culture Collection	ATCC CRL-1573
Sup-T1 Cell Line	American Type Cell Culture Collection	ATCC CRL-1942
TZM-bL Cell Line	NIH AIDS Research and Reference Reagent Program	Cat # 8129
Recombinant DNA		
Rev Mutants	(Fernandes et al., 2016)	pcDNA4 NStrep-Rev M1A, etc.
Env Mutants	This paper	pcDNA4 NStrep-Env M1A, etc.
pNL4-3 ΔEnv	This paper	pNL4-3 ΔEnv
Viruses for Rev deep mutational scanning (rev-in-nef)	(Fernandes et al., 2016)	pNL4-3 TxR M1X

Viruses for Env deep mutational scanning (rev-in-nef)	This paper	pNL4-3 TxR Env
Rev Reporter Plasmid	(Smith <i>et al.</i> , 1990)	pCMV gag-pol-RRE
Software and Algorithms		
MiSeq Reporter	Illumina, Inc	MiSeq Reporter v2.6
Geneious Prime	(Kearse <i>et al.</i> , 2012)	v2019.1.1
Other		
Patient Sequence Alignments	HIV LANL (https://www.hiv.lanl.gov/content/index)	2018 curated alignments
WebLogo 3: Create	http://weblogo.threeplusone.com/create.cgi	

Resource Availability

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contacts, Alan D Frankel (frankel@cgl.ucsf.edu) and Jason D Fernandes (jason@scribetx.com).

Materials Availability

All requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact authors.

Data and code availability

All data supporting the findings of this study have been included as supplemental tables and code is available on <https://github.com/jferna10/EnvPaper>. Requests for additional data or alternative formats should be addressed to the lead contacts. Sequencing data has been deposited in GEO: accession GSE179046. All graphs were generated in Rstudio with associated files available on github.

Cells

HEK 293 and TZM-bL cells were maintained in humidified incubators at 37°C and 5% CO₂ in DMEM, high glucose (with 10% heat-inactivated fetal bovine serum (FBS) and 1% penicillin/streptomycin). SupT1 human T-cell lines were maintained in RPMI

(supplemented with 25 mM HEPES pH 7.4, 10% heat-inactivated fetal bovine serum (FBS) and 1% penicillin/streptomycin) at 37°C and 5% CO₂.

Plasmids

Different mutants were constructed using primers (synthesized by IDT) harboring the appropriate mutations, and then used in a 2-step PCR reaction and inserted into pcDNA4TO-2xStrep vector (described in (Fernandes *et al.*, 2016)) by Gibson Assembly for testing in export assays. Viral mutants were created by the same strategy in pNL4-3 rev-in-nef construct (Fernandes *et al.*, 2016).

Method Details

Rev and Env reporter assays

Nine residues of Rev and ten residues of Env, overlapping region of Rev NES and Env LLP-2, were mutated to alanine in independent site directed mutagenesis reactions in mammalian expression vectors. For the Rev reporter assays, 293T cells in a 24-well plate were transfected with 250 ng of pCMV gagPol-RRE and 25 ng of a pcDNA4TO 2xStrep NL4-3 Rev using PolyJet (SignaGen) reagent. Cells were lysed in Tris-buffered saline containing 1% Triton X-100 and protease inhibitors, 48 hours after transfection, and intracellular p24 levels was measured by ELISA. The assay was performed in biological triplicate. For all mutants, activity was normalized to the unmutated reference activity. Rev expression levels were verified by Western blot using HRP-conjugated-anti-StrepTagII antibody (IBA biosciences) with GAPDH used as a loading control. For Env reporter assays, the relative infectivity of Env pseudoviruses of LLP-2 mutants was measured in TZM-bl cells assay. To prepare HIV-1 Env pseudoviruses of LLP-2 mutants, 293T cells were co-transfected with pNL4-3 ΔEnv and an envelope expressing plasmid pcDNA 4TO Env. Pseudovirus containing culture supernatant was harvested 48 hours after transfection and concentrated by PEG purification (the final concentrations for PEG-6000 and NaCl were 8.5% (w/v) and 0.3M, respectively (Kutner, Zhang and Reiser, 2009)). The infectivity of mutants was determined by infecting TZM-bl cells with p24-normalized pseudoviruses. 48 hours post-infection, luciferase activity of the reporter gene was quantified using Bright-Glo luciferase reagent (Promega, Madison,

WI). Experiments were done in biological triplicate and relative infectivity was normalized to the unmutated reference infectivity. The incorporation of Env LLP-2 mutants into the pseudovirus was measured by Western blot using gp120 antibody and p24 as loading control.

Viral replication assays

NL4–3 rev-in-nef proviral constructs (Fernandes *et al.*, 2016) were used to introduce mutations in the Rev NES or Env LLP-2 in non-overlapped contexts. For generation of each virus, HEK 293T cells in 6 well plates were transfected with 1000 ng of NL4–3 proviral plasmid and complemented with reference Rev or Env using PolyJet (SignaGen) reagent at a 1:3 ratio of the DNA:Polyjet. Viral supernatants were harvested 48 hours after transfection, stored at –80 °C, and then virus titer quantified by p24 capsid ELISA. One million SupT1 cells were incubated with 5 ng of p24 virus in 250 ul media containing 1 ug/ml PEI and 8 ug/ml polybrene and centrifuged 1200g for 2 hours at 37 °C. Input virus was removed and then washed cells with 250 ul of PBS. Cells were resuspended in 250 ul of media and placed in an incubator at 37 °C, 5% CO₂. Every 48–72 hours for 20 days 100 ul of virus-containing supernatants were collected for quantification by p24 ELISA and added 100ul of fresh media to cells. Viral p24 levels below 10 pg/ml are shown as 10 pg/ml in the plots.

Competitive deep mutational scanning analysis

The Rev NES fitness values were extracted from previous work (Fernandes *et al.*, 2016). Env LLP-2 proviral libraries were created as described (Fernandes *et al.*, 2016). Briefly, each of the 10 individual residues of Env LLP-2 that overlap with the Rev NES was randomized. For this purpose, this region was amplified using 10 separate pools of 33 nt primers containing NNN in the middle of the primer and using Gibson Assembly Master Mix (NEB) was cloned into pNL4-3 rev-in-nef after digesting by BamHI and XhoI. The proviral library of each residue was generated by scraping together at least 1000 colonies of each primer pool for plasmid purification. High codon diversity of each randomized proviral library (10 total) was confirmed by deep sequencing. Virus generation for each proviral plasmid library was done as described above in 293T cells.

Selection assays were performed via spinoculation of one million SupT1 cells as above. The infections were examined by IFA for cellular HIV antigen synthesis and also by quantification of progeny virion release into supernatant using p24 ELISA. Once the infection reference or defective viral pools reach their peak, 150 uL of viral supernatant was collected and centrifuged to remove cells, and then RNA was extracted from virions by ZR-96 Viral RNA extraction kits (Zymo) and eluted in 15 ul water. cDNA was amplified from 5uL of extracted RNA using the iScript Advanced (BioRad) cDNA synthesis. The selection experiment was performed in biological duplicate for each pool/residue. Samples were sequenced on an Illumina MiSeq with PE100 and demultiplexed on instrument. Raw fastqs were aligned to the reference sequence with bowtie2 and the flags --fast-local --rdg 100,3 --rfg 100,3. Codon and amino acid counts were generated via the countDMS program available on github. These counts were then log2 normalized by fractions calculated from sequencing the input proviral plasmids to generate experimental fitness scores. The resulting fitness matrices were then processed and plotted using the ComplexHeatmap package in R.

Entropy

Entropy values were calculated using hiv.lanl.gov “Entropy One” tool and the 2018 filtered web alignments (amino acid) for Rev and Env, discounting ambiguous characters. The “amino acid equivalents” setting was used to analyze chemically conservative changes with the addition that I, L, and V were all treated equivalently.

LLP-2/NES overlap in primate lentiviruses

The 2018 curated amino acid alignments of HIV-1/SIVcpz, HIV-2/SIVsmm, and other SIV were downloaded from the Los Alamos Database (hiv.lanl.org). The corresponding Rev NES and Env LLP-2 region of each virus was extracted and modified as needed in Geneious Prime v2019.1.1. The sequence conservation for each domain of each virus was generated using WebLogo3.

References

- Alfadhli, A. *et al.* (2019) 'Analysis of HIV-1 Matrix-Envelope Cytoplasmic Tail Interactions', *Journal of Virology*. Edited by F. Kirchhoff, 93(21), pp. 1–10. doi: 10.1128/jvi.01079-19.
- Booth, D. S., Cheng, Y. and Frankel, A. D. (2014) 'The export receptor Crm1 forms a dimer to promote nuclear export of HIV RNA', *eLife*, 3, pp. 1–17. doi: 10.7554/elife.04121.
- Boscia, A. L. *et al.* (2013) 'Membrane structure correlates to function of LLP2 on the cytoplasmic tail of HIV-1 gp41 protein', *Biophysical Journal*. Biophysical Society, 105(3), pp. 657–666. doi: 10.1016/j.bpj.2013.06.042.
- Checkley, M. A., Luttge, B. G. and Freed, E. O. (2011) 'HIV-1 envelope glycoprotein biosynthesis, trafficking, and incorporation.', *Journal of molecular biology*. NIH Public Access, 410(4), pp. 582–608. doi: 10.1016/j.jmb.2011.04.042.
- Chen, X., Zaro, J. L. and Shen, W. C. (2013) 'Fusion protein linkers: Property, design and functionality', *Advanced Drug Delivery Reviews*. Elsevier B.V., 65(10), pp. 1357–1369. doi: 10.1016/j.addr.2012.09.039.
- Crooks, G. E. *et al.* (2004) 'WebLogo: A sequence logo generator', *Genome Research*, 14(6), pp. 1188–1190. doi: 10.1101/gr.849004.
- Dong, X. *et al.* (2009) 'Structural basis for leucine-rich nuclear export signal recognition by CRM1', *NATURE*, 458(7242), pp. 1136–U71. doi: 10.1038/nature07975.
- Eisenberg, D. and Wesson, M. (1990) 'The most highly amphiphilic alpha-helices include two amino acid segments in human immunodeficiency virus glycoprotein 41.', *Biopolymers*, 29(1), pp. 171–7. doi: 10.1002/bip.360290122.
- Fernandes, J. D. *et al.* (2016) 'Functional Segregation of Overlapping Genes in HIV', *Cell*. Elsevier Inc., 167(7), pp. 1762–1773.e12. doi: 10.1016/j.cell.2016.11.031.
- Finkel, Y. *et al.* (2021) 'The coding capacity of SARS-CoV-2', *Nature*. Springer US, 589(7840), pp. 125–130. doi: 10.1038/s41586-020-2739-1.
- Foley, B. *et al.* (2013) *HIV Sequence Compendium 2013*. Los Alamos National Laboratory, NM: Theoretical Biology and Biophysics Group.
- Fornerod, M. *et al.* (1997) 'CRM1 is an export receptor for leucine-rich nuclear export signals.', *Cell*, 90(6), pp. 1051–60.

- Frankel, A. D. and Young, J. A. T. (2002) 'HIV-1: Fifteen Proteins and an RNA', *Annual Review of Biochemistry*, 67(1), pp. 1–25. doi: 10.1146/annurev.biochem.67.1.1.
- Greaney, A. J. *et al.* (2021) 'Complete Mapping of Mutations to the SARS-CoV-2 Spike Receptor-Binding Domain that Escape Antibody Recognition', *Cell Host and Microbe*. Elsevier Inc., 29(1), pp. 44-57.e9. doi: 10.1016/j.chom.2020.11.007.
- Güttler, T. *et al.* (2010) 'NES consensus redefined by structures of PKI-type and Rev-type nuclear export signals bound to CRM1', *Nature Structural & Molecular Biology*. Nature Publishing Group, 17(11), pp. 1367–1376. doi: 10.1038/nsmb.1931.
- Haddox, H. K. *et al.* (2018) 'Mapping mutational effects along the evolutionary landscape of HIV envelope', *eLife*, 7, pp. 1–29. doi: 10.7554/eLife.34420.
- Haddox, H. K., Dingens, A. S. and Bloom, J. D. (2016) 'Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture', *PLOS Pathogens*. Edited by R. Swanstrom, 12(12), p. e1006114. doi: 10.1371/journal.ppat.1006114.
- Hughes, A. L. *et al.* (2001) 'Simultaneous Positive and Purifying Selection on Overlapping Reading Frames of the tat and vpr Genes of Simian Immunodeficiency Virus', *Journal of Virology*. American Society for Microbiology, 75(17), pp. 7966–7972. doi: 10.1128/jvi.75.17.7966-7972.2001.
- Jayaraman, B. *et al.* (2019) 'Highly Mutable Linker Regions Regulate HIV-1 Rev Function and Stability', *Scientific Reports*, 9(1), pp. 1–17. doi: 10.1038/s41598-019-41582-7.
- Kalia, V. *et al.* (2003) 'Rational site-directed mutations of the LLP-1 and LLP-2 lentivirus lytic peptide domains in the intracytoplasmic tail of human immunodeficiency virus type 1 gp41 indicate common functions in cell-cell fusion but distinct roles in virion envelope incorporation', *Journal of virology*, 77(6), pp. 3634–46. doi: 10.1128/JVI.77.6.3634.
- Kalia, V. *et al.* (2005) 'Antibody Neutralization Escape Mediated by Point Mutations in the Intracytoplasmic Tail of Human Immunodeficiency Virus Type 1 gp41', *Journal of Virology*, 79(4), pp. 2097–2107. doi: 10.1128/jvi.79.4.2097-2107.2005.
- Kearse, M. *et al.* (2012) 'Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data', *Bioinformatics*, 28(12), pp. 1647–1649. doi: 10.1093/bioinformatics/bts199.

- Kodama, T. *et al.* (1989) 'Significance of premature stop codons in env of simian immunodeficiency virus.', *Journal of Virology*, 63(11), pp. 4709–4714. doi: 10.1128/JVI.63.11.4709-4714.1989.
- Kutner, R. H., Zhang, X.-Y. and Reiser, J. (2009) 'Production, concentration and titration of pseudotyped HIV-1-based lentiviral vectors', *Nature Protocols*, 4(4), pp. 495–505. doi: 10.1038/nprot.2009.22.
- Li, S.-C. *et al.* (1996) 'alpha-Helical, but not beta-3-sheet, propensity of proline is determined by peptide environment', *Proc Natl Acad Sci USA*, 93(June), pp. 6676–6681. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC39085/pdf/pnas01517-0472.pdf%0Apapers://26d6e71c-adea-4ad5-a5ab-deff7e0ad63e/Paper/p446%0Ahttps://www.pnas.org/content/pnas/93/13/6676.full.pdf>.
- Lu, L. *et al.* (2008) 'Surface exposure of the HIV-1 Env cytoplasmic tail LLP2 domain during the membrane fusion process: Interaction with gp41 fusion core', *Journal of Biological Chemistry*, 283(24), pp. 16723–16731. doi: 10.1074/jbc.M801083200.
- Malim, M. H. and Cullen, B. R. (1991) 'HIV-1 structural gene expression requires the binding of multiple Rev monomers to the viral RRE: Implications for HIV-1 latency', *Cell*, 65(2), pp. 241–248. doi: 10.1016/0092-8674(91)90158-U.
- Murphy, R. E. *et al.* (2017) 'Solution Structure and Membrane Interaction of the Cytoplasmic Tail of HIV-1 gp41 Protein', *Structure*. Cell Press, 25(11), pp. 1708–1718.e5. doi: 10.1016/J.STR.2017.09.010.
- Nelson, C. W. *et al.* (2020) 'Dynamically evolving novel overlapping gene as a factor in the SARS-CoV-2 pandemic', *eLife*, 9, pp. 1–45. doi: 10.7554/eLife.59633.
- Newman, J. T. *et al.* (2007) 'Differential functional phenotypes of two primary HIV-1 strains resulting from homologous point mutations in the LLP domains of the envelope gp41 intracytoplasmic domain.', *Virology*, 367(1), pp. 102–16. doi: 10.1016/j.virol.2007.05.027.
- Pan, K. and Deem, M. W. (2011) 'Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza', *Journal of the Royal Society Interface*, 8(64), pp. 1644–1653. doi: 10.1098/rsif.2011.0105.
- Piai, A. *et al.* (2020) 'Structural basis of transmembrane coupling of the HIV-1 envelope

- glycoprotein', *Nature Communications*. Springer US, 11(1), pp. 1–12. doi: 10.1038/s41467-020-16165-0.
- Pollard, V. W. and Malim, M. H. (1998) 'THE HIV-1 REV PROTEIN', *Annual Review of Microbiology*, 52(1), pp. 491–532. doi: 10.1146/annurev.micro.52.1.491.
- Postler, T. S. and Desrosiers, R. C. (2013) 'The Tale of the Long Tail: the Cytoplasmic Domain of HIV-1 gp41', *Journal of Virology*, 87(1), pp. 2–15. doi: 10.1128/jvi.02053-12.
- Rancurel, C. *et al.* (2009) 'Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation', *Journal of Virology*. American Society for Microbiology (ASM), 83(20), pp. 10719–10736.
- Rausch, J. W. and Le Grice, S. F. J. (2015) 'HIV Rev Assembly on the Rev Response Element (RRE): A Structural Perspective.', *Viruses*, 7(6), pp. 3053–75. doi: 10.3390/v7062760.
- Sarzotti-Kelsoe, M. *et al.* (2014) 'Optimization and validation of the TZM-bl assay for standardized assessments of neutralizing antibodies against HIV-1', *Journal of Immunological Methods*, 409(919), pp. 131–146. doi: 10.1016/j.jim.2013.11.022.
- Shacklett, B. L. *et al.* (2000) 'The Intracytoplasmic Domain of the Env Transmembrane Protein Is a Locus for Attenuation of Simian Immunodeficiency Virus SIVmac in Rhesus Macaques', *Journal of Virology*, 74(13), pp. 5836–5844. doi: 10.1128/jvi.74.13.5836-5844.2000.
- Shimizu, H. *et al.* (1992) 'Analysis of a human immunodeficiency virus type 1 isolate carrying a truncated transmembrane glycoprotein', *Virology*, 189(2), pp. 534–546. doi: 10.1016/0042-6822(92)90577-C.
- Smith, A. J. *et al.* (1990) 'Human immunodeficiency virus type 1 Pr55gag and Pr160gag-pol expressed from a simian virus 40 late replacement vector are efficiently processed and assembled into viruslike particles.', *Journal of Virology*, 64(6), pp. 2743–2750.
- Snetkov, X. *et al.* (2021) 'The envelope cytoplasmic tail regulates HIV-1 assembly and spread'.
- Starr, T. N. *et al.* (2020) 'Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding', *Cell*. Elsevier, 182(5), pp. 1295-1310.e20. doi: 10.1016/j.cell.2020.08.012.

- Steckbeck, J. D. *et al.* (2011) 'Highly conserved structural properties of the C-terminal tail of HIV-1 gp41 protein despite substantial sequence variation among diverse clades: Implications for functions in viral replication', *Journal of Biological Chemistry*, 286(31), pp. 27156–27166. doi: 10.1074/jbc.M111.258855.
- Steckbeck, J. D., Kuhlmann, A.-S. and Montelaro, R. C. (2013) 'C-terminal tail of human immunodeficiency virus gp41: functionally rich and structurally enigmatic.', *The Journal of general virology*, 94(Pt 1), pp. 1–19. doi: 10.1099/vir.0.046508-0.
- Steckbeck, J. D., Kuhlmann, A. S. and Montelaro, R. C. (2014) 'Structural and functional comparisons of retroviral envelope protein C-terminal domains: Still much to learn', *Viruses*, 6(1), pp. 284–300. doi: 10.3390/v6010284.
- Tedbury, P. R. and Freed, E. O. (2015) 'The Cytoplasmic Tail of Retroviral Envelope Glycoproteins', in, pp. 253–284. doi: 10.1016/bs.pmbts.2014.10.009.
- Wang, Q. *et al.* (2020) 'An amphipathic peptide targeting the gp41 cytoplasmic tail kills HIV-1 virions and infected cells', *Science Translational Medicine*, 12(546), p. eaaz2254. doi: 10.1126/scitranslmed.aaz2254.
- Weingarten-Gabbay, S. *et al.* (2020) 'SARS-CoV-2 infected cells present HLA-I peptides from canonical and out-of-frame ORFs.', *bioRxiv : the preprint server for biology*. doi: 10.1101/2020.10.02.324145.
- Zhang, Y. *et al.* (2021) 'MARCH8 targets cytoplasmic lysine residues of various viral envelope glycoproteins', *bioRxiv*, p. 6. Available at: <https://doi.org/10.1101/2021.04.20.440588>.

A

Entropy Analysis (All Amino Acids)

1. Calculate protein entropies independently
2. Group sites that share nt overlap & plot

(Y768 , L75)
(H769, P76)
etc.

Rev NES Crm1-binding Residue
Env LLP-2 Charged Residue
NES binding & LLP-2 Charge
Neither

B

C

Entropy Analysis (Reduced Alphabet)

Rev NES Crm1-binding Residue
Env LLP-2 Charged Residue
NES binding & LLP-2 Charge
Neither

Figure 2

Figure 3

Figure 4

A

B

Env Reporter Assay

Viral Spread

Env Reporter Assay

Viral Spread

Figure 7