FLORIDA INTERNATIONAL UNIVERSITY

Miami, Florida

IDENTIFICATION OF FITNESS-CRITICAL REGIONS IN THE HIV PROTEOME BASED ON EVOLUTIONARY AND STRUCTURAL ANALYSIS AS TARGETS FOR BROADLY NEUTRALIZING HIV THERAPIES

An Undergraduate Honors Thesis submitted in partial fulfillment of the

requirements for the degree of

UNDERGRADUATE HONORS

in

BIOLOGICAL SCIENCES

by

Richard Suarez

2019

To: Dr. Steven Oberbauer, Chairperson

Department of Biological Sciences

This Undergraduate Honors Thesis in Biological Sciences, written by Richard Suarez entitled "Identification of Fitness-Critical Regions in the HIV Proteome based on Evolutionary and Structural Analysis as Targets for Broadly Neutralizing HIV Therapies", is submitted to you in partial fulfillment of the requirements for the Undergraduate Honors in Biological Sciences. The Biological Sciences Undergraduate Honors Committee and the candidate's research supervisor have read this thesis. We recommend that it be approved.

Dr. Jessica Siltberg-Liberles
Honors Research Supervisor

Dr. Walter M. Goldberg, Chairperson
Undergraduate Honors Committee

Dr. Steven Oberbauer, Chairperson
Department of Biological Sciences

Date of Honors Research Presentation: April 15, 2019
This thesis by Richard Suarez is approved.

Department of Biological Sciences
Florida International University
2019

DEDICATION

I dedicate this thesis to my family, especially my mother for always inspiring me to continue learning and growing as an individual. My family is my motivation.

ACKNOWLEDGEMENTS

I would like to thank Dr. Jessica Siltberg-Liberles for her willingness to bring me into her lab and guide me throughout the process. Without knowing much about me, she offered me the opportunity to conduct the research necessary for this thesis with kindness and a smile. In my three years here at FIU, I have never met a better professor mentor than her, and for that I am forever grateful.

I would like to thank the Siltberg-Liberles lab members, specifically Janelle Nunez-Castilla, who helped me understand certain concepts I struggled with, and explained how the different techniques worked and how to use them.

I would like to thank the FIU Honors College for providing the funds to attend different conferences to present my research, including: the 2019 Miami Winter Symposium, Florida Undergraduate Research Conference, and National Conference for Undergraduate Research.

I would like to thank the Honors in Biology Program within the Biological Sciences Department for allowing me to construct this thesis, as well as present and defend it. I would also like to thank Dr. Goldberg for his guidance and patience.

I would like to thank my friends at the PLTL program for informing me of the Honors in Biology program and supporting me throughout the process.

Lastly, I would like to thank my family for all their love, support and understanding while I was working on this project.

ABSTRACT

While we have known that HIV causes AIDS for almost 40 years and despite numerous efforts to develop a vaccine, a vaccine to prevent HIV infection is not yet available. Currently, combination treatments with multiple antiretroviral drugs are used to keep HIV infection at bay, but due to natural evolutionary processes in HIV populations, ongoing efforts to catch up with resistance-causing mutations are necessary and there is a potential for diverse drug interactions. To better understand the evolution of protein structure and function in HIV, I studied HIV from an evolutionary perspective using molecular evolution methodology and structural bioinformatics tools. I identified potential antiretroviral target regions of five or more consecutive residues that must remain conserved in sequence and structure to avoid losing viral fitness. By targeting these fitness-critical sites, the time to onset of resistance-causing mutations for antiretrovirals should be prolonged and the need for combination therapeutics reduced. I propose that by targeting these sites the function of the proteins should be greatly affected by preventing proper assembly or function of the HIV virus. Thus, drugs targeting these sites are potential treatment options that can avoid the rapid mutation of the virus and will consequently limit the current need for multiple therapeutics. Based on my results, I propose that the target site regions in the capsid are of utmost interest as a future inhibitor target. Several of these regions in the capsid are located in close vicinity of each other and appear vital in the packing of the capsid hexamer.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

INTRODUCTION

Since the 1980s, human immunodeficiency virus (HIV) has been a prevalent pandemic, resulting in millions of deaths. A product of multiple successful cross-species transmission events of simian immunodeficiency virus (SIV), animal-to-human, and further human-to-human, the virus continues to pose a public health threat (Gao et al. 1999; Wertheim and Worobey 2009). Currently, individuals infected with HIV are treated with antiretroviral therapy (ART) and combination ART (cART), which are regimens of antiretroviral drugs. Along with those therapeutics, recent advancements in the field, such as the recommendation of rapid initiation of treatment (Rosen et al. 2016), Pre-exposure Prophylaxis (PrEP) (Heneine and Kashuba 2012), and the use of monoclonal antibodies for viral suppression and decreased virus reservoirs (Lynch et al. 2015), have drastically reduced the mortality rate (Wang et al. 2015). Moreover, strategies such as "treatment as prevention" involving PrEP (Montaner et al. 2010) and high coverage of ART (Tanser et al. 2013) have been shown to decrease rates of transmission by having undetectable viral loads.

Today, the antiretroviral drugs on the market target different aspects of the HIV life cycle using various mechanisms of action. Each drug is classified according to the step it attacks, ranging from fusion and maturation inhibitors to reverse transcriptase and integrase blockers (Palmisano and Vella 2011). Commonly, antiretroviral chemotherapies are combined with one another because of drug resistance due to the accumulation of mutations associated with the evolution of the virus. These in turn affect ligand binding in the host (Fernandez et al 2005; Guo et al 2015). Examples, such as the conformational flexibility of the trimeric envelope glycoprotein 120, gp120 (Do Kwon et al. 2015) and intrinsic disorder of proteins including nucleocapsid and matrix (Xue et al. 2014),

illustrate the variability of HIV structural and functional components. However, receiving these drug combinations has been associated with toxicity and drug interactions (Di Giambenedetto et al. 2017). Although recent studies have shown that fewer simultaneous antiretrovirals in combination therapy can still be effective, this treatment is still needed (Di Giambenedetto et al. 2017; Perez-Molina et al. 2017; Trottier et al. 2017). Furthermore, researchers have attempted to improve the efficacy of the current therapeutics by using monoclonal antibodies to enhance viral suppression (Lynch et al. 2015; Byrareddy et al. 2016).

While these approaches have changed this once fatal ailment to a chronic but manageable infection, there are problems with the present means of treatment. It is also important to note that this transition is not felt worldwide as accessibility and availability for medications vary country to country. The rapid evolutionary rate of HIV has consequences for the use of antiretrovirals (Little et al. 2002; Simon et al. 2002). Combination ART is primarily used to prevent relapse in response to HIV becoming drug-resistant as a result of its accumulation of fitness-neutral non-synonymous mutations. Thus, ligand binding is modified as the residues are altered. Moreover, the conformational flexibility of the viral structures increases the difficulty of proper targeting by therapeutics (Tirado-Rives and Jorgensen 2006). As a result, developing antiretrovirals and potentially a vaccine for HIV constitute a problem. For example, with the initial aggregation of mutations, the HIV protease decreases in fitness and function; more compensatory mutations result in the protein returning to its native conformation along with the acquisition of drug resistance (Fernandez et al. 2005). Likewise, the conformational flexibility of gp120 creates an obstacle because the form it holds at any moment determines the binding affinity of an antibody (Do Kwon et al. 2015).

The strategy presented here entails a comparative and evolutionary approach that utilizes publicly available HIV protein sequence data and structural prediction methods to rein in regions in the HIV proteome that are critical to fitness of the virus. Due to the global concern of HIV, its sequence space is well represented. The generation of broadly neutralizing vaccines commonly incorporate the genomic sequences of multiple variants of the virus, as is used to produce influenza immunization (Giles and Ross 2011, 2012). In order to develop these vaccines, the consensus or ancestral sequences of living viruses are used to produce greater coverage (Kesturu et al. 2006). However, these can result in errors when dealing with sequences that quickly diverge such as those of HIV (McCloskey et al. 2014). The methods above assume that conserved sequence sites are important for viral fitness, but not all conserved sequence sites make optimal drug targets. A prime choice for medication must be accessible and its success also depends on the physicochemical properties and orientation of the amino acids in the binding pocket.

When designing a vaccine or an antiviral, the ability to occupy dynamic conformational ensembles under physiological conditions associated with the viral proteins (i.e. structural disorder) complicates the process because of the structural variability and internal disorder (Yu et al 2016). In order to resolve the complications posed by conformational flexibility, potential antiretroviral, or vaccine, sites that consider the structural plasticity of the viral proteins should be identified. In addition, to protect against potential mutations, the evolutionary context of sequence and structure must be evaluated to further ensure efficacy (Rahaman and Siltberg-Liberles 2016).

HIV is among a family of retroviruses known as *Lentivirus* (Figure 1). These viruses are linked by their similarities in their biological interactions with the organisms they infect, mechanism of replication, genetic composition and presence in slow disease

syndromes (Narayan and Clements 1989). Understanding the relationship between these groups will provide the evolutionary context necessary for this study.

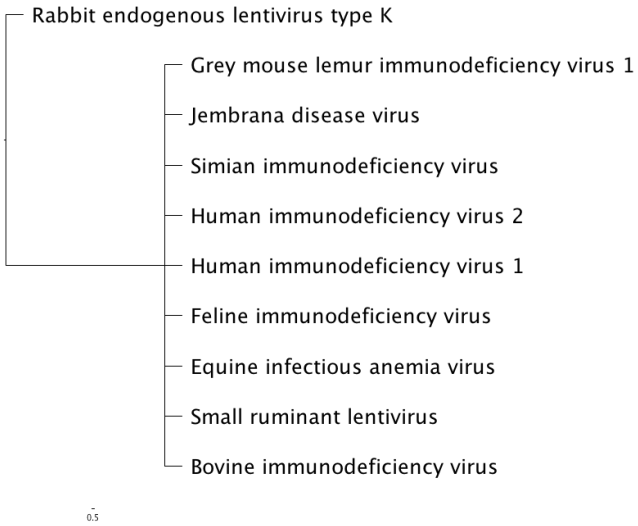


Figure 1. *Lentivirus* Family Tree
The relations between the groups of the *Lentivirus* family are mostly unresolved.

Here, I study the evolution of the HIV proteome across its genus *Lentivirus* in combination with predictions of secondary structure and structural disorder. Sites that are conserved in order, sequence, and secondary structure across various virus protein homologs are likely to be constrained from 1) changing sequence on evolutionary time scales and 2) undergoing real-time structural transitions (Rahaman and Siltberg-Liberles 2016). Thus, these regions can be considered vulnerable and druggable in the HIV and, potentially, the SIV proteome. I will refer to these areas as target sites. Such sites have potential to display regions with conservation in sequence and in structure that can be explored for binding broadly neutralizing antivirals or vaccines.

METHODS

Protein Family Reconstruction

Sequences were identified from the NCBI HIV-1 RefSeq database, resulting in 28 proteins for the initial analysis (Table 1). BLAST searches of each protein using the accession numbers were performed against the refseq_protein database. Sequences with >30% sequence identity and >60% query coverage relative to those of HIV were identified as homologous.

Table 1. Extracted Proteins and Accession Numbers
28 HIV-1 proteins from NCBI HIV-1 RefSeq database, along with the corresponding accession numbers.

Protein	Accession Number	Protein	Accession Number	Protein	Accession Number
Pr55 (Gag)	NP_057850	Env gp160	NP_057856	Matrix	NP_579876
Gag-Pol	NP_057849	Asp	YP_009028572	Pol	NP_789740
VPR	NP_057852	Integrase	YP_001856243	p1	NP_787042
Env gp41	NP_579895	RT	YP_001856242	Integrase	NP_705928
RTp51	NP_789739	Retropepsin	YP_001856241	RT	NP_705927
Gag-Pol Transframe Peptide	NP_787043	Env gp120	NP_579894	Retropepsin	NP_705926
Nef	NP_057857	Unnamed Protein Product	NP_579893	p6	NP_579883
p2	NP_579882	Nucleocapsid	NP_579881	Capsid	NP_579880
Vpu	NP_057855	Vif	NP_057851	Rev	NP_057854
Tat	NP_057853				

For each query, a multiple sequence alignment was built for the identified set of sequences using Muscle (Edgar 2004) as implemented in Jalview (Waterhouse et al. 2009). The initial exploration included all 28 proteins extracted from NCBI. Some of these proteins were polyproteins, such as Gag, Pol and Env, that contain 4, 3, and 2 final protein products, respectively (Figure 2). For these polyproteins, I found that there was

redundancy in the sense that different combinations of these proteins were included within the 28 sequences in the initial dataset. To avoid this redundancy, only the final protein products were included. Further, the proteins unique to HIV were not included due to their narrow evolutionary context. The remaining seven protein sequence datasets were used for phylogenetic reconstruction.

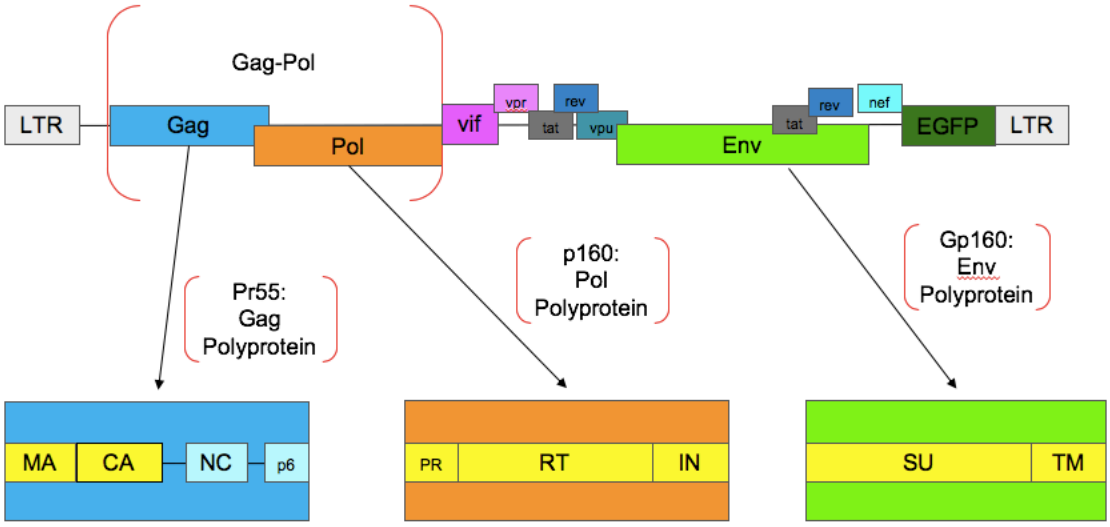


Figure 2. Simplified Schematic of the HIV Proteome
A viral map outlining the major polyproteins of HIV. The proteins colored in yellow were investigated further (MA-Matrix, CA-Capsid, PR-Protease, RT-Reverse Transcriptase, IN-Integrase, SU-Envelope Surface Protein, TM-Envelope Transframe Protein).

MrBayes 3.2.6 with a four-category gamma distribution and mixed model of amino acid substitution was used to construct phylogenetic trees (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Each analysis extended to one million generations, with a sample frequency of 100, or until a preset stop value (defined as the average standard deviation of split frequencies below 0.005) was reached. The final tree was produced by discarding the first 25% of the samples as the default burnin, thus structuring it from the last 75% of samples, and using the half-compatible parameter, to

avoid weakly supported nodes (i.e., with a posterior probability <0.5). The generated trees were midpoint rooted.

Protein Family Reconstruction in Lentivirus

To narrow down the evolutionary context*, Lentivirus* subsets to be used for the conservation analysis were generated. For each individual component of the major polyproteins, the respective protein sequences for the four lentiviruses HIV-1, HIV-2, SIV, and SIV-m were used to build a set of additional multiple sequence alignments using accuracy-oriented MAFFT (Katoh and Standley 2013) as implemented in Jalview (Waterhouse et al. 2009). Phylogenetic trees were assembled using the same procedures described above.

Predictions of Intrinsic Disorder and Secondary Structure Elements

Intrinsic disorder propensity for each site in each protein was predicted using IUPred with the 'long' option as the default setting (Dosztányi et al. 2005a, 2005b). This method employs a statistical interaction potential, which aims to assess the capacity of inter-residue interactions as disordered regions that lack the ability to form sufficient contacts (Dosztányi et al. 2005a). The site-specific disorder propensities of each protein were plotted according to the multiple sequence alignment and phylogenetic tree as raw disorder propensities and as binary states, order or disorder, using the cutoff of 0.5. If the propensities fell below the cutoff, it was assigned 'ordered' and if was at the cutoff or above, it was assigned disorder.

PSIPRED was used to predict secondary structure elements (McGuffin et al. 2000). The predictor performs an analysis of the output received from PSI-BLAST (Position Specific Iterated BLAST) (Altschul et al. 1997; McGuffin et al. 2000). PSIPRED used the uniref90 database, meaning the sequences were clustered at the 90%

sequence identity level. Sites were classified into three states: alpha helices, beta strands, or loops. These states were mapped onto their corresponding sites in the multiple sequence alignment and visualized according to the phylogenetic tree.

Conservation Analysis

Potential target sites were identified as conserved in sequence, order, and structure if the following three criteria were met for all sequences at a site: no change in amino acid, disorder propensity below 0.5, and no change in secondary structure elements (alpha helix or beta strand). Five or more consecutive target sites are termed a target region. The output files of IUPred, PSIPRED and amino acid information were juxtaposed to evaluate if the areas of conservation were in the same positions across the three criteria. If so, they were labeled as a potential target site; if they were conserved across the groups, but not in the same positions for sequence and structure, they were not labeled as such.

Strain Data Analysis

To further analyze the potential of the recognized target sites, sequence information from different strains were used. A local BLAST database was built containing sequences extracted from NCBI. In total, the local database contained 1014395, 6555, and 15770 sequences for HIV-1, HIV-2, and SIV, respectively. The HIV-1 sequence for the protein families that harbored target sites was used to complete BLAST searches against the local database. A multiple sequence alignment was built for the yield of each search using Clustal Omega (Sievers et al 2011). These alignments were used to further evaluate the conservation at target regions.

Protein Visualization

Each protein family was displayed through the incorporation of the phylogenetic tree with the multiple sequence alignment or predictor-based matrix and portrayed on a

heatmap via the in-house pipeline (Rahaman and Siltberg-Liberles 2016). In order to visualize the location of potential target sites in 3D, PyMOL was used (Schrodinger 2015). The protein structure representatives for capsid, reverse transcriptase, protease and integrase were identified by NCBI-BLAST against the PDB database.

RESULTS

Phylogenetic Trees

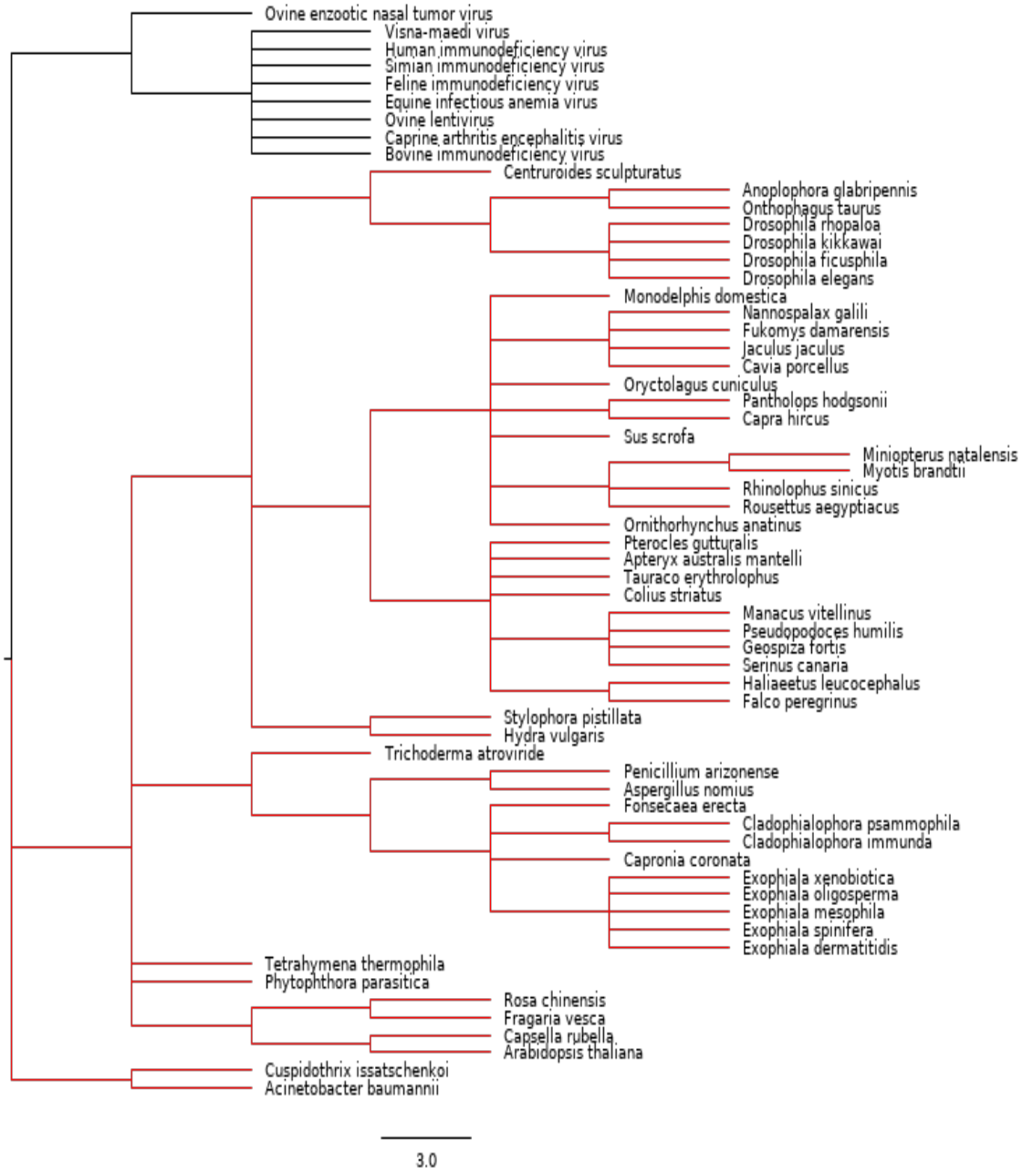
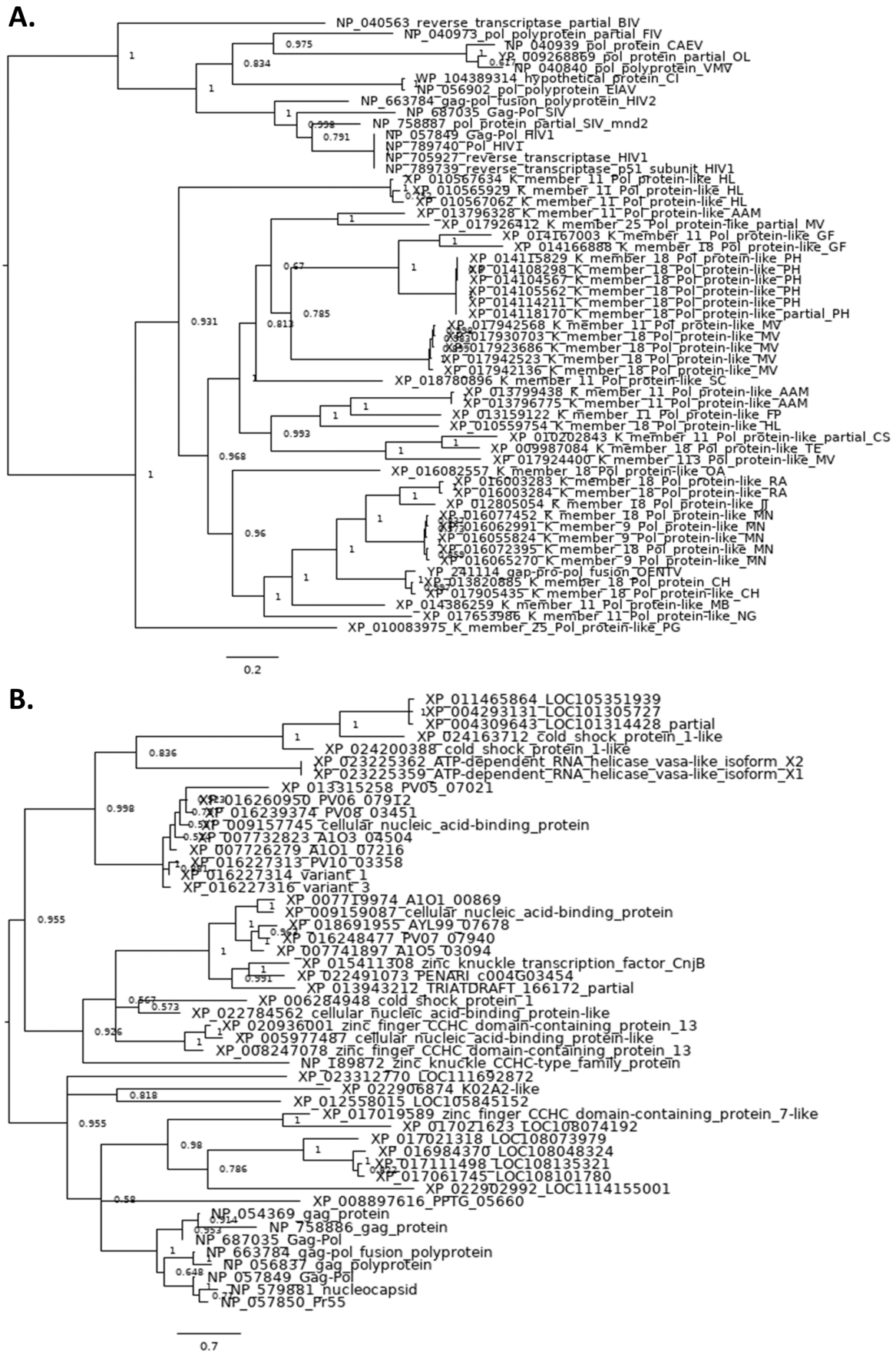


Figure 3. Species Tree
Portrays the species and viruses in which the extracted sequences belong, using Common Tree on NCBI (Sayers et al. 2009; Benson et al. 2009). The red branches indicate non-viral entities. The black branches represent the *Lentiviruses*.

The BLAST searches identified homologs of HIV proteins in a diverse group of lentiviruses and other taxa (Figure 3). For each of the seven final queries and its identified homologs, the protein sequences were extracted, aligned and examined using phylogenetic trees constructed with MrBayes 3.2.6 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) resulting in seven protein families (Figure 4). The trees display the evolutionary relationships among the different sequences in each protein family. Certain proteins, such as reverse transcriptase (Figure 4A) and retropepsin (a.k.a. protease) (Figure 4H), have homologs present in other organisms aside from lentiviruses; this offers greater insight into the evolution of the virus and provides potential models for further study of these proteins. Moreover, many of them have unresolved or flat branches indicating that the evolutionary history for those groups cannot be determined based on the given sequence set, its alignment, and the tree building method. The Bayesian inference method attempts to estimate relationships by constructing the most probable tree, but when there appears to be no apparent agreement between the generated trees, the branches remain unresolved. Further, the lentivirus group was reduced to the following four viruses: HIV-1, HIV-2, SIV and SIVm (SIV of mandrills) for an additional set of smaller trees (Figure 5).


A.
NP_040563_reverse_transcriptase_partial_BIV
NP_040973_pol_polyprotein_partial_FIV
NP_040939_pol_protein_CAEV
YP_009268869_pol_protein_partial_OL
NP_040840_pol_polyprotein_VMV
WP_104389314_hypothetical_protein_CI
NP_056902_pol_polyprotein_EIAV
NP_663784_gag-pol_fusion_polyprotein_HIV2
NP_687035_Gag-Pol_SIV
NP_758887_pol_protein_partial_SIV_mnd2
NP_057849_Gag-Pol_HIV1
NP_789740_Pol_HIV1
NP_705927_reverse_transcriptase_HIV1
NP_789739_reverse_transcriptase_p51_subunit_HIV1
XP_010567634_K_member_11_Pol_protein-like_HL
XP_010565929_K_member_11_Pol_protein-like_HL
XP_010567062_K_member_11_Pol_protein-like_HL
XP_013796328_K_member_11_Pol_protein-like_AAM
XP_017926412_K_member_25_Pol_protein-like_partial_MV
XP_014167003_K_member_11_Pol_protein-like_GF
XP_014166888_K_member_18_Pol_protein-like_GF
XP_014115829_K_member_18_Pol_protein-like_PH
XP_014108298_K_member_18_Pol_protein-like_PH
XP_014104567_K_member_18_Pol_protein-like_PH
XP_014105562_K_member_18_Pol_protein-like_PH
XP_014114211_K_member_18_Pol_protein-like_PH
XP_014118170_K_member_18_Pol_protein-like_partial_PH
XP_017942568_K_member_11_Pol_protein-like_MV
XP_017930703_K_member_18_Pol_protein-like_MV
XP_017923686_K_member_18_Pol_protein-like_MV
XP_017942523_K_member_18_Pol_protein-like_MV
XP_017942136_K_member_18_Pol_protein-like_MV
XP_018780896_K_member_11_Pol_protein-like_SC
XP_013799438_K_member_11_Pol_protein-like_AAM
XP_013796775_K_member_11_Pol_protein-like_AAM
XP_013159122_K_member_11_Pol_protein-like_FP
XP_010559754_K_member_18_Pol_protein-like_HL
XP_010202843_K_member_11_Pol_protein-like_partial_CS
XP_009987084_K_member_18_Pol_protein-like_TE
XP_017924400_K_member_113_Pol_protein-like_MV
XP_016082557_K_member_18_Pol_protein-like_OA
XP_016003283_K_member_18_Pol_protein-like_RA
XP_016003284_K_member_18_Pol_protein-like_RA
XP_012805054_K_member_18_Pol_protein-like_JI
XP_016077452_K_member_18_Pol_protein-like_MN
XP_016062991_K_member_9_Pol_protein-like_MN
XP_016055824_K_member_9_Pol_protein-like_MN
XP_016072395_K_member_18_Pol_protein-like_MN
XP_016065270_K_member_9_Pol_protein-like_MN
YP_241114_gag-pro-pol_fusion_QENTV
XP_013820885_K_member_18_Pol_protein_CH
XP_017905435_K_member_18_Pol_protein-like_CH
XP_014386259_K_member_11_Pol_protein-like_MB
XP_017653986_K_member_11_Pol_protein-like_NG
XP_010083975_K_member_25_Pol_protein-like_PG
0.2
B.
XP_011465864_LOC105351939
XP_004293131_LOC101305727
XP_004309643_LOC101314428_partial
XP_024163712_cold_shock_protein_1-like
XP_024200388_cold_shock_protein_1-like
XP_023225362_ATP-dependent_RNA_helicase_vasa-like_isoform_X2
XP_023225359_ATP-dependent_RNA_helicase_vasa-like_isoform_X1
XP_013315258_PV05_07021
XP_016260950_PV06_07912
XP_016239374_PV08_03451
XP_009157745_cellular_nucleic_acid-binding_protein
XP_007732823_A1O3_04504
XP_007726279_A1O1_07216
XP_016227313_PV10_03358
XP_016227314_variant_1
XP_016227316_variant_3
XP_007719974_A1O1_00869
XP_009159087_cellular_nucleic_acid-binding_protein
XP_018691955_AYL99_07678
XP_016248477_PV07_07940
XP_007741897_A1O5_03094
XP_015411308_zinc_knuckle_transcription_factor_CnjB
XP_022491073_PENARI_c004G03454
XP_013943212_TRIATDRAFT_166172_partial
XP_006284948_cold_shock_protein_1
XP_022784562_cellular_nucleic_acid-binding_protein-like
XP_020936001_zinc_finger_CCHC_domain-containing_protein_13
XP_005977487_cellular_nucleic_acid-binding_protein-like
XP_008247078_zinc_finger_CCHC_domain-containing_protein_13
NP_189872_zinc_knuckle_CCHC-type_family_protein
XP_023312770_LOC111692872
XP_022906874_K02A2-like
XP_012558015_LOC105845152
XP_017019589_zinc_finger_CCHC_domain-containing_protein_7-like
XP_017021623_LOC108074192
XP_017021318_LOC108073979
XP_016984370_LOC108048324
XP_017111498_LOC108135321
XP_017061745_LOC108101780
XP_022902992_LOC111415500
XP_008897616_PPTG_05660
NP_054369_gag_protein
NP_758886_gag_protein
NP_687035_Gag-Pol
NP_663784_gag-pol_fusion_polyprotein
NP_056837_gag_polyprotein
NP_057849_Gag-Pol
NP_579881_nucleocapsid
NP_057850_Pr55
0.7

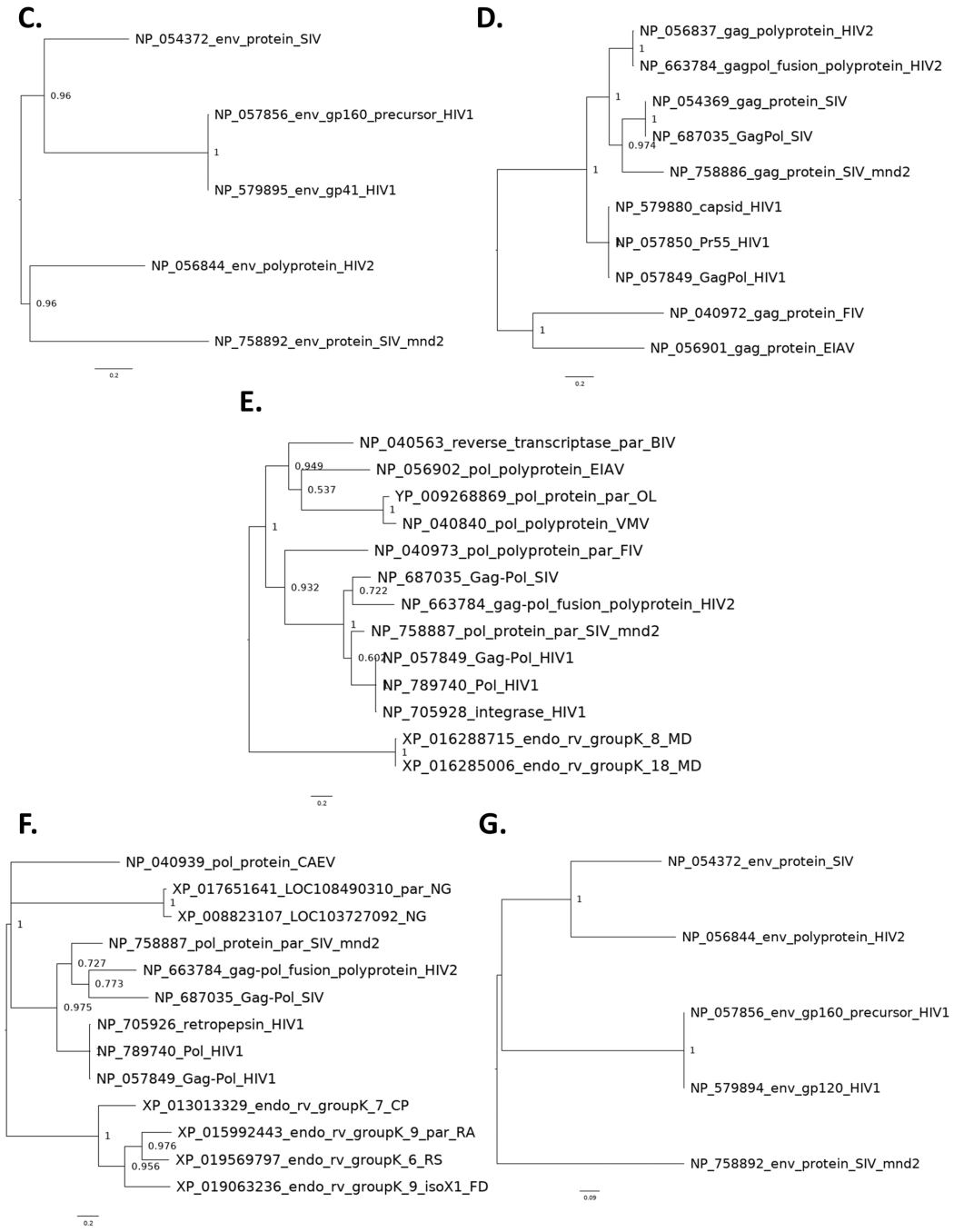


Figure 4. MrBayes-3.2.6 Phylogenetic Trees
Prediction of the evolutionary relationships of the different proteins. In each tree, each sequence is named with its refseq_protein accession number, followed by its functional annotation, and last, its taxa. The majority of the support values at each node are >0.9. **A**: nucleocapsid; **B**: reverse transcriptase; **C**: gp41; **D**: capsid; **E**: integrase; **F**: retropepsin; **G**: gp120 - TM.

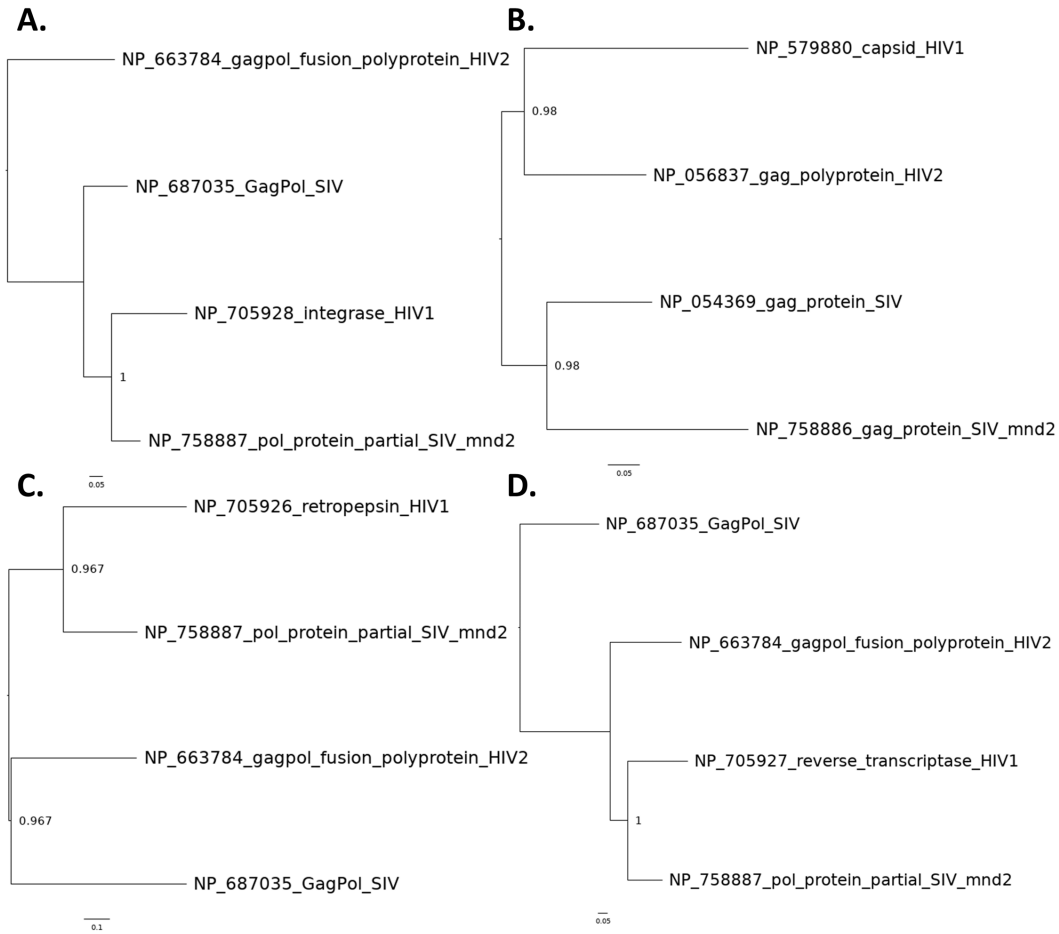


Figure 5. MrBayes-3.2.6 Reduced Phylogenetic Trees The phylogenetic trees of the sequences from HIV1, HIV2, SIV, and SIVmnd2 from Figure 4. The support values at each node are >0.9. **A**: integrase; **B**: capsid; **C**: retropepsin; **D**: reverse transcriptase.

Predictions of Intrinsic Disorder and Secondary Structure Elements

The prediction data for the disorder propensity and secondary structures for each protein family were illustrated on a heatmap via an in-house pipeline (Rahaman and Siltberg-Liberles 2016). The figures are displayed through the incorporation of the phylogenetic tree with the multiple sequence alignment or predictor-based matrix (Figure 6). Based on the predictions of intrinsic disorder, conservation was inferred for sites where all sequences were predicted to have disorder propensity below 0.5. Similarly, conserved

structure was inferred for sites where all sequences were predicted to have the same secondary structure element.

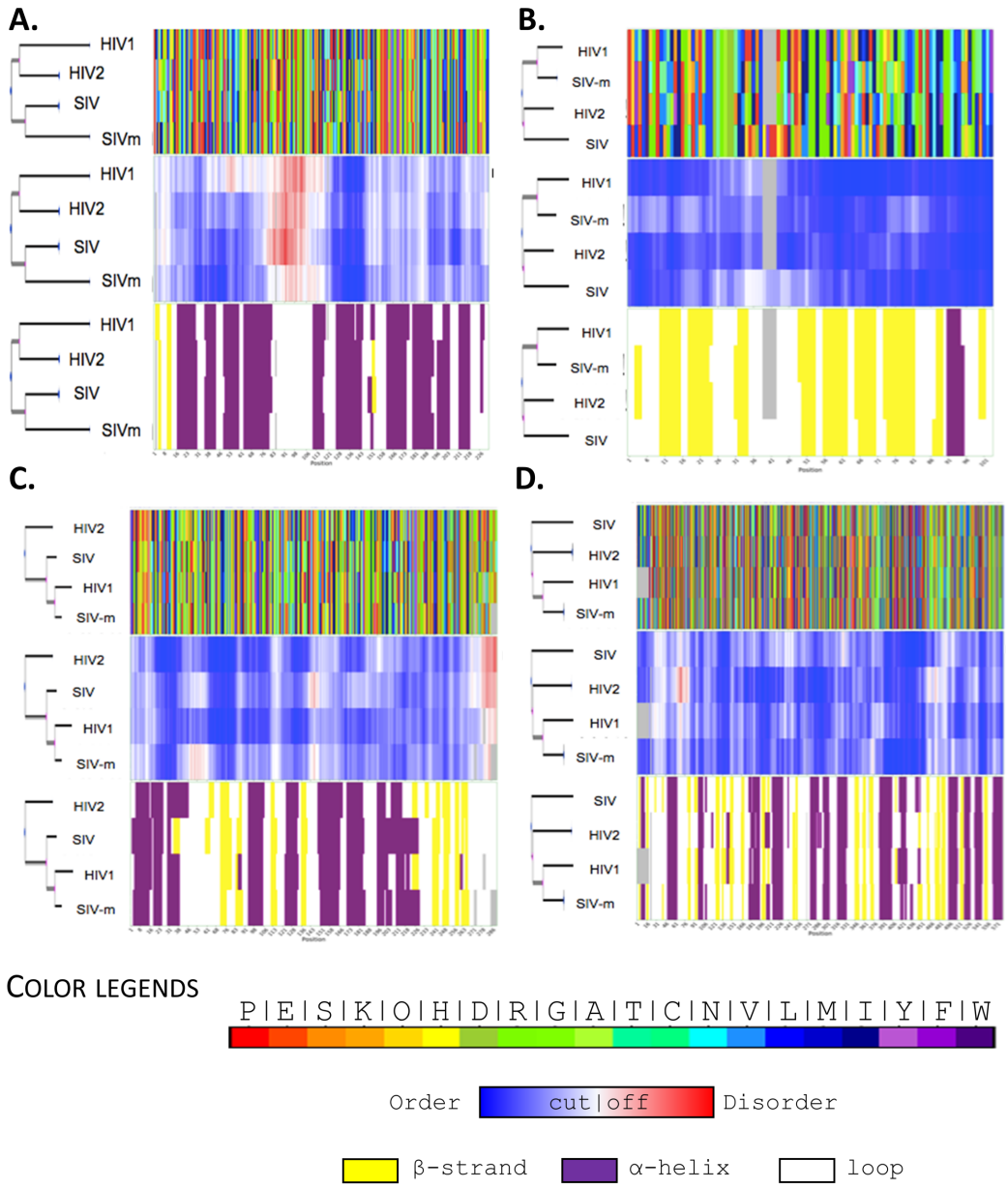


Figure 6: Prediction Heatmaps
Heatmaps were constructed by aligning the multiple sequence alignment or predictor-based matrix according to the phylogenetic tree. Capsid (**A**); Retropepsin (**B**); Integrase (**C**); Reverse Transcriptase (**D**). The top portion provides the amino acid information, colored as described in the legend. The middle portion presents the disorder propensity (cutoff 0.5) according to legend. The bottom portion shows secondary structure predictions with the color scheme outlined in the legend.

Conservation Analysis

Identification of target sites was based on conservation of amino acid identity, low intrinsic disorder propensity (i.e. order), and secondary structure information for each site. A target region is one that is five or more consecutive residues with conservation in the three criteria described above. No target regions were found in the first set of trees that included homologs from all possible species based on the database used.

Table 2. Number of Epitopes

Proteins	**Region**	**Target Sites**	**Number of Epitopes**
Capsid	1	SPRTLNAWVK	2
	2	INEEA	2
	3	PKEPF	3
	4	YVDRFYK	8
	5	ACQGVGGP	1
Retropepsin	1	DTGADD	0
	2	GGIGG	4
Reverse Transcriptase	1	KQWPL	2
	2	QLGIPHP	0
	3	PQGWKGS	1
	4	WVPAHKGIG	4
Integrase	1	FLLKL	4

The conservation analysis completed on the reduced phylogenies that only include HIV-1, HIV-2, SIV, and SIVm (Figure 5) yielded target regions for retropepsin, reverse transcriptase, capsid and integrase. These areas are likely to possess greater structural stability and to be critical for function resulting in decreased rate of mutation. The potential for these sites to be attacked by protein/peptide products is strengthened by their presence in known epitopes (Table 2) because this shows at least some portion of the site has already been bound by antibodies and is accessible. Although most of the sites have been found in known epitopes, no determinant contains the entire target area. Only two

proteins, retropepsin and reverse transcriptase, have one site that has not been found in any known epitope.

Target Region Visualization

To complement the linear visualization of the preditions shown above (Figure 6), the target regions were mapped onto a three-dimensional structure. Although these regions have potential as viable target sites, their location within the three-dimensional structure plays an important role in this designation. Therefore, with the exception of integrase, the proteins were labeled with different colors to highlight the distinct target regions (Figure 7). The blue structures (Figures 7A, 8A, 9A, 10A), and individualized monomer colors (Figure 7C), were used to distinguish these regions from the rest of the protein structure; the grey structures (Figures 7B, 7D, 7E, 8B, 9B, 10B) have each region in its own color to identify each separately.

In the capsid, region 1 is a large area within the helix (Figure 7B); and regions 3-5 (Figure 7B) cover an overall larger space due to their close proximity, thus this broad region may be a viable target as it influences more than one component of the capsid. As

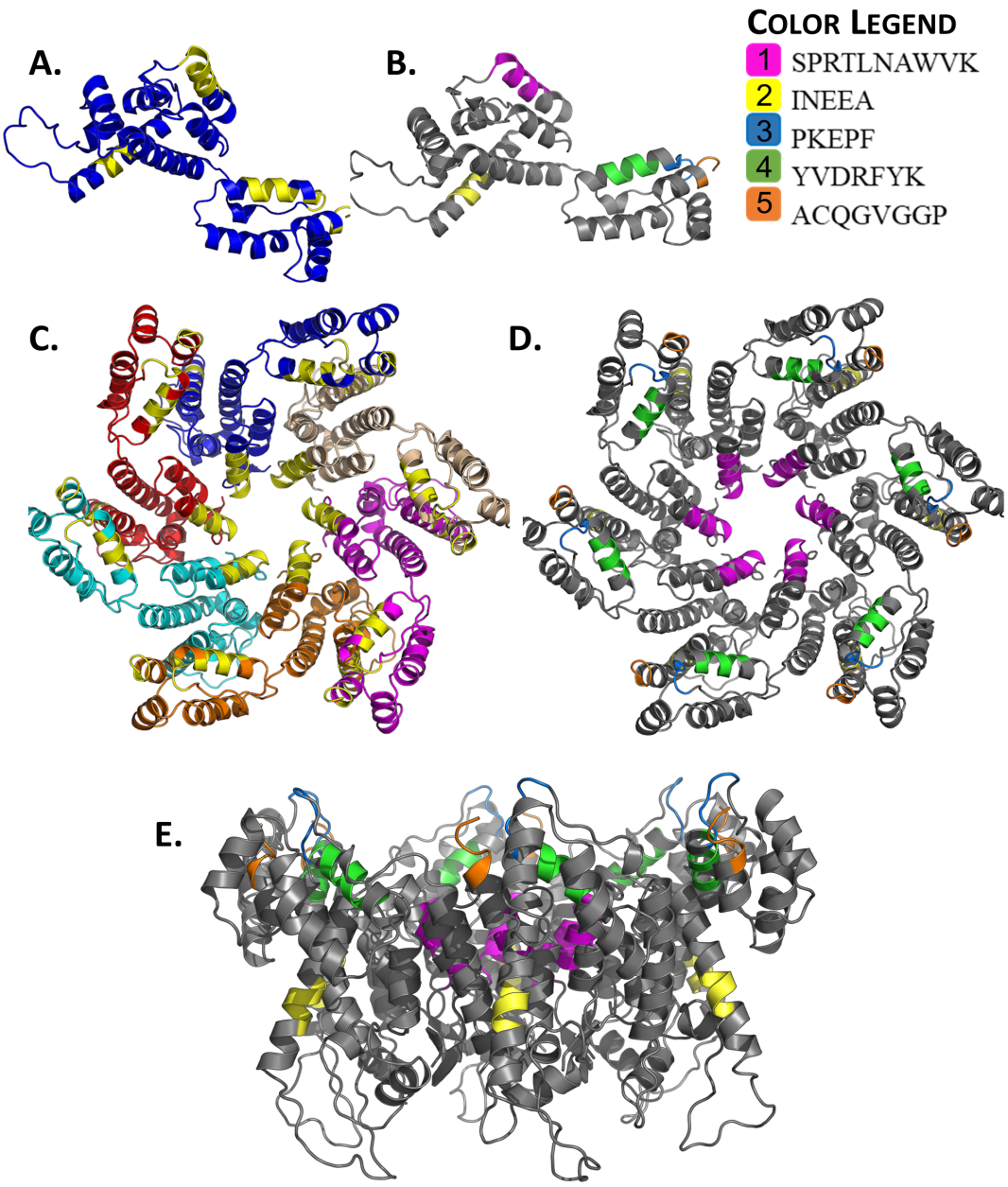


Figure 7. Capsid 3D Visualization 3D-capsid monomer structures labeled with the target sites in yellow (**A**) and with each site in a distinguishing color (**B**) as outlined by the legend. 3D-capsid hexamer structures showing six capsid monomers in different colors with the target regions in yellow (**C**) and with the monomers in grey and the regions in distinguishing colors (**D and E (sideview)**) according to the legend. Region 1 = pink, region 2 = yellow, region 3 = blue, region 4 = green and region 5 = orange. PDB id: 4XFX (Gres et al. 2015). 3D-visualizations were done with PyMOL (Schrodinger 2015).

seen in the capsid hexamer (Figure 7C, 7D, 7E), the sites are situated at not only the subunit interfaces between the monomers (regions 1 and 3-5), but also between the hexamer-hexamer interactions (region 2), indicating their importance in proper assembly of the mature capsid.

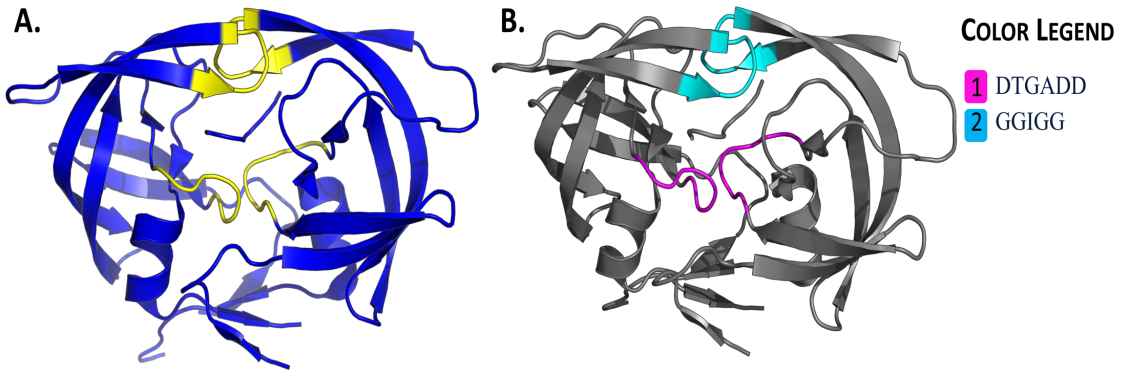


Figure 8. Retropepsin 3D Visualization
3D-retropepsin structures labeled with the target regions in yellow (**A**) and with each region in a distinguishing color (**B**) as outlined by the legend. Region 1 = pink, and region 2 = blue. PDB id: 5YRS (Das et al. 2010). 3D-visualizations were done with PyMOL (Schrodinger 2015).

Regions 1 and 2 (Figure 8) are on the outer surface of retropepsin, illustrating no obstructions to protein/peptide products that may be developed to target them. Region 1 appears in a loop. Region 2 is located in the middle of a beta hairpin, including the connecting loop. Loops are typically more conformationally flexible, but these regions were predicted to be ordered. The two regions are rather close to each other and accessible from the same side. This may allow the production of a drug that can bind both at the same time. Also, the protein represents a dimer; regions 1 and 2 are situated at the interface between the monomers indicating its importance in the function of the protein.

The target regions of reverse transcriptase (Figure 9) are in the area of interaction

between DNA and the protein. These residues may affect the way in which the proteins function on the DNA, implicating the role they may play in the generation of viral DNA. Moreover, the regions for integrase (Figure 10) is on the exterior of the protein. Similar to retropepsin, protein/peptide products may be created to target the accessible site

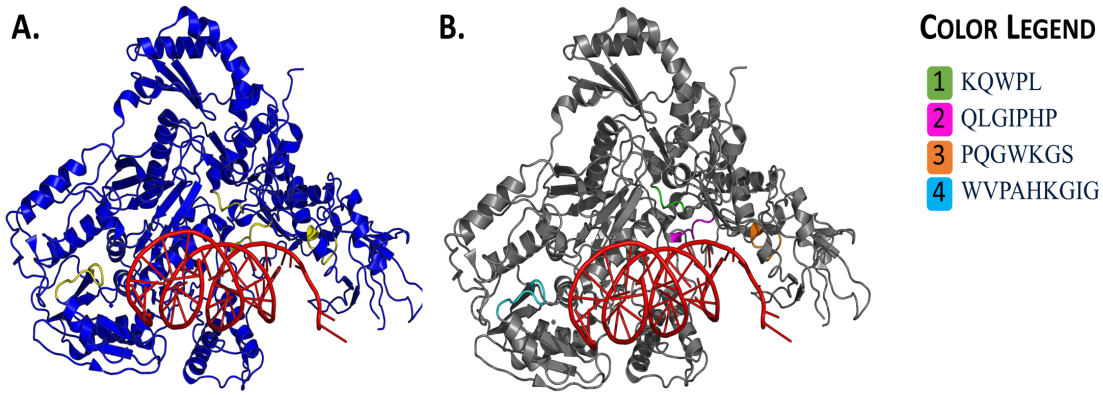


Figure 9. Reverse Transcriptase 3D Visualization
3D-reverse transcriptase with DNA structures labeled with the target regions in yellow (**A**) and with each region in a distinguishing color (**B**) as outlined by the legend. Region 1 = green, region 2 = pink, region 3 = orange, and region 4 = blue. PDB id: 3KJV (Lansdon et al. 2010). 3D-visualizations were done with PyMOL (Schrodinger 2015).

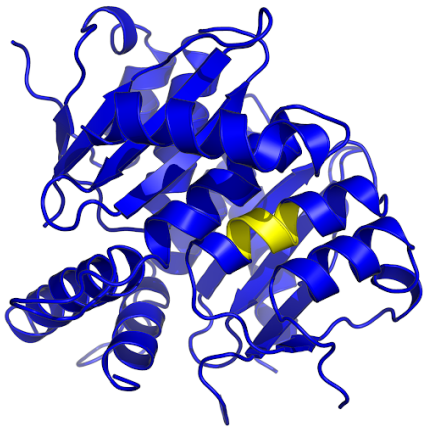


Figure 10. Integrase 3D Visualization
3D-integrase structure labeled with the target site in yellow. PDB id: 1BIZ (Goldgur et al. 1998). 3D-visualization was done with PyMOL (Schrodinger 2015).

Strain Conservation Count

To further analyze the potential of the recognized target regions, sequence information from different strains were used. The HIV-1 sequences for the protein families that harbored target regions were used to perform BLAST searches against the local database constructed of HIV-1, HIV-2 and SIV strains from GenBank. The resulting alignments were used to evaluate conservation at the target regions, and the evolutionary distribution of how often the target regions change judging from the frequency of region appearance in the sequenced strains (Table 3). Capsid, retropepsin and regions 1 and 2 of reverse transcriptase present a large percentage of the sequenced strains containing these sequence motifs, which further supports some form of resistance to frequent mutation at these sites. The region of integrase presents low conservation demonstrating this region does not withstand the effects of mutations as well as others. Regions 3 and 4 of reverse transcriptase show no constraint and appear to be accumulating mutations.

Table 3. Strain Conservation Count Percentage

Proteins	**Region**	**Target Sites[1]**	**SCC[2] (%)**
Capsid	1	SPRTLNAWVK	95/98
	2	INEEA	97
	3	PKEPF	98
	4	YVDRFYK	60/99
	5	ACQGVGGP	81
Retropepsin	1	DTGADD	99
	2	GGIGG	99
Reverse Transcriptase	1	KQWPL	95
	2	QLGIPHP	99
	3	PQGWKGS	0
	4	WVPAHKGIG	0/0.12
Integrase	1	FLLKL	44

[1] Underlined portions indicate a change in percentage when reduced to the minimum number required for a target site (five consecutive residues).
[2] Percentage of sequenced strains containing the motifs based on the BLAST searches against the local database according to each site for the respective proteins (SCC%).

DISCUSSION

RNA viruses are notorious for their rapid sequence evolution that contributes to drug resistance (Holland et al. 1982; Drake and Holland 1999). These viruses also suffer the adverse effects of antibody-dependent enhancement where antibodies for closely related strains bind with altered affinity, promoting, instead of demoting, viral infection (Robinson et al. 1989; Sol, Tirado and Yoon 2004). Antibodies bind to antigenic regions due to their amino acid sequence, often in a specific 3D conformation. If the amino acids of an antigenic region are altered, due to a modification or mutation, the affinity of the antibody is likely affected. Similarly, if the 3D conformation of the antigenic region is changed due to nearby amino acid substitutions, it can also alter antibody affinity (Rudikoff et al. 1982). Some closely related viral proteins present intrinsic structural disorder in different areas, indicating rapid evolution of disorder (Ortiz et al. 2013; Gitlin et al. 2014; Rahaman and Siltberg-Liberles 2016). The presence of intrinsic disorder regions in a protein means that region is conformationally flexible. Disorder in a region harboring an antigenic site means the antigenic site may not always be in a conformation to which the antibody can bind. Further, due to high evolutionary dynamics of disordered regions in viral proteins, an antibody that is effective against one virus may be ineffective against a similar strain even if the antigenic sequence itself has not yet changed (Childs, Baskerville and Cobey 2015). In the case of antiviral drugs, the drug-binding pocket also depends on its conformation and the interactions between the drug and the amino acid residues lining the pocket. If the conformation or the amino acid identities of the binding pocket change, it can alter the binding of the drug and consequently, the efficacy of the drug (Ortiz et al. 2013; Gitlin et al. 2014).

It is therefore important to not only consider sequence conservation when attempting to develop broadly neutralizing antibodies and antivirals, but to also consider structural conservation. To accomplish this, my study relied on predictions for secondary structure and intrinsic disorder to consider structural conservation: sites with changes in secondary structure or intrinsic disorder across all sequences in the alignment were considered to not be structurally conserved (Rahaman and Siltberg-Liberles 2016). as a (Rahaman and Siltberg-Liberles 2016). For a site to be considered a target, it must be conserved in structure and sequence and found it a linear sequence of at least 5 sites. Here, I refer to these as target regions, a general term that includes both antigenic regions and drug binding pockets. Four proteins (the capsid, reverse transcriptase, retropepsin, and integrase) were found to have target regions, conserved across HIV-1, HIV-2, SIV and SIVm (SIV of mandrills).

Observing high conservation of the regions found in sequenced strains supports the evolutionary importance of these sites. Here, the analysis demonstrates that the regions for capsid are found in 60-98% of the different strains, but if 5 residue long subsequences are considered, such as YVDRF (YVDRFYK) and PRTLN (SPRTLNAWVK), the percentage changes to almost 100%. Retropepsin possesses two sites, both with 99% strain conservation. Based on the high conservation and the location of region 2, one can speculate that this beta hairpin provides an interesting drug target because it is also located at the dimer interface. The areas within reverse transcriptase (KQWPL and QLGIPHP) are highly conserved across sequenced strains. Again, this suggests that these regions possess some mechanism of resistance to mutation, while also maintaining structural stability, indicating their high potential for being targeting by broadly neutralizing drugs. The latter two regions of reverse transcriptase and of

integrase have low sequence conservation within the sequenced strains, suggesting that these regions should be avoided as broadly neutralizing targets.

Mapping the target regions onto the three-dimensional protein structures reveals which of them are potentially accessible to drugs or antibodies. In particular, all the regions except one (INEEA) in the capsid, and all in retropepsin and integrase are found on the exterior of the proteins or in a space with no obstructions as seen with the region SPRTLNAWVK in the capsid. Further regions, including all regions in the capsid and one in retropepsin, have functional roles in protein-protein interactions such as the packing of the capsid hexamer and of the retropepsin dimer, while the regions in reverse transcriptase seem important for interacting with DNA. This demonstrates the potential of these sites as critical for proper HIV pathogenicity because these regions are involved in the assembly and interactions of the proteins and DNA; thus, if their functions are disrupted, the virus could lose its ability to successfully replicate.

Several of the target regions in capsid are in close vicinity of each other and occupy a rather large area in the subunit interfaces of the monomers, and an area involved in the space promoting hexamer-hexamer interactions (Yufenyuy and Aiken 2013; Gres et al. 2015). The immature and mature capsids in retroviruses assemble into a protein lattice surrounding the budding virion (Qu et al. 2018). The HIV capsid hexagonal lattice structure exhibits structural variability and variation at the inter-hexamer interfaces, which may disrupt capsid assembly or uncoating (Gres et al. 2015). Here, the visualization of the region involved in hexamer-hexamer interactions demonstrates the importance of binding mechanics in promoting viability as a site to prevent proper HIV replication.

Most of these target sites have been seen, to varying degrees, in known epitopes. This provides support for their possible use as areas of interest because they appear to be accessible, as described above. Particular sites, such as the target region SPRTLNAWVK located at N-terminal helix that forms the center of the capsid hexamer (Figure 7), have previously been described as "promiscuous" with potential to bind to several biomolecules and thus, making them viable targets for a broadly-neutralizing therapy (Fernandez et al. 2005; Frahm et al. 2007). This analysis indicates that surface-accessible sites, which are conserved in different strains and are relevant to protein-protein interactions, have the attributes to function as a viable target (Figure 11). This is further confirmed when some portion of the targets is found within known epitopes, such as the sites within the capsid, which has been previously described as a target of interest (Carnes et al. 2018). The other protein regions are vital areas for further analysis and do present importance for interrupting the viral life cycle as shown here.

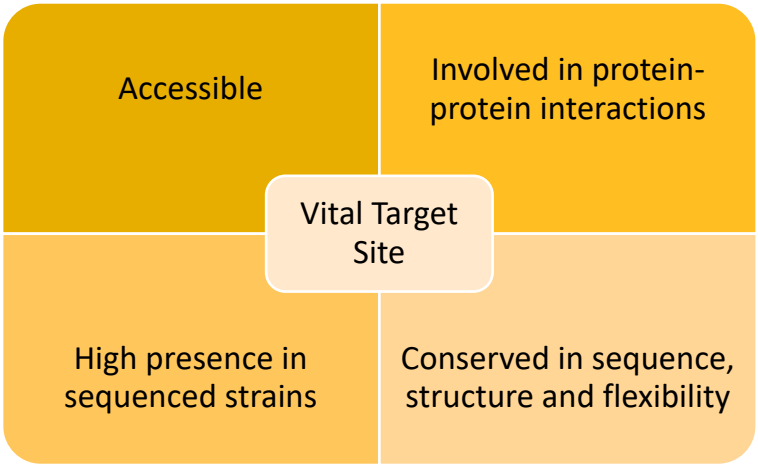


Figure 11. Vital Target Site Criteria
Four key aspects that indicate the viability of a site to be targeted by a drug.

In order to strengthen the conclusions drawn from this study, further investigation needs to be done in terms of developing epitopes to target the entire range of the sites

presented here. The physicochemical properties of the residues in these sites and their response to environment-specific factors require experimentation to ensure their efficacy. Further exploration into the capsid could produce fruitful results as they are potentially vital inhibitor targets (Carnes et al. 2018). Trials done to verify the importance of these sites in the function, assembly, and in general, the replication of the virus will offer a greater understanding of the virus.

Overall, drugs targeting these sites are potential treatment options that avoid issues with the rapid mutation of the virus and the consequent need for multiple therapeutics that risk potential drug interactions. Altogether, these areas can be considered vulnerable and druggable in the HIV and, potentially, the SIV proteome. Such sites have potential as targets for broadly neutralizing antivirals or vaccines because sequence conservation makes them broadly specific and avoids targeting conformationally flexible regions, offering hope for the future of HIV treatment.

LITERATURE CITED

Altschul, SF, Madden, TL, Schaffer, AA, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25:3389-3402.

Bahir, I, Fromer, M, Prat, Y, Linail, M. 2009. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Molecular Systems Biology 5:1-14.

Benson, DA, Karsch-Mizrachi, I, Lipman, DJ, et al. 2009. GenBank. Nucleic Acids Research 37:D26-31. DOI: 10.1093/nar/gkn723.

Byrareddy, SN, Arthos, J, Cicala, C, et al. 2016. Sustained virologic control in SIV+ macaques after antiretroviral and alpha(4)beta(7) antibody therapy. Science: 354:197-202.

Carnes, SK, Sheehan, JH, Aiken, C. 2018. Inhibitors of the HIV-1 capsid, a target of opportunity. Current Opinion in HIV and AIDS 13:359-365.

Childs, LM, Baskerville, EB, Cobey, S. 2015. Trade-offs in antibody repertoires to complex antigens. Philosophical Transactions of the Royal Society B doi: https://doi.org/10.1098/rstb.2014.0245.

Danielson, ML, Lill, MA. 2012. Predicting flexible loop regions that interact with ligands: the challenge of accurate scoring. Proteins 80:246-260.

Das, A, Mahale, S, Prashar, V, et al. 2010. X-ray Snapshot of HIV-1 Protease in Action: Observation of Tetrahedral Intermediate and Its SIHB with Catalytic Aspartate. Journal of American Chemical Society 132:6366-6673.

Di Giambenedetto, S, Fabbiani, M, Roldan, EQ, et al. 2017. Treatment simplification to atazanavir/ritonavir plus lamivudine versus maintenance of atazanavir/ritonavir plus two NRTIs in virologically suppressed HIV-1-infected patients: 48 week results from a randomized trial (ATLAS-M). Journal of Antimicrobial Chemotherapy 72:1163-1171.

Do Kwon, Y, Pancera, M, Acharya, P, et al. 2015. Crystal structure, conformational fixation, and entry-related interactions of mature ligand-free HIV-1 Env. Nature Structural & Molecular Biology 22:522-531.

Dosztányi, Z, Csizmok, V, Tompa, P, Simon, I. 2005a. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21:3433–3434.

Dosztányi, Z, Csizmók, V, Tompa, P, Simon, I. 2005b. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. Journal of Molecular Biology 347:827–839.

Drake, JW, Holland, JJ. Mutation rates among RNA viruses. Proceedings of the National Academy of Sciences 96:13910-13913. doi: https://doi.org/10.1073/pnas.96.24.13910.

Edgar, RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797. doi: 10.1093/nar/gkh340.

ElHefnawi, M, AlAidi, O, Mohamed, N, et al. 2011. Identification of novel conserved functional motifs across most Influenza A viral strains. Virology Journal 8:1-10.

Fernandez, A, Tawfik, DS, Berkhout, B, et al. 2005. Protein promiscuity: Drug resistance and native functions - HIV-1 case. Journal of Biomolecular Structure & Dynamics 22:615-624.

Fiser, A, Kinh Gian Do, R, Sali, A. 2000. Modeling of loops in protein structures. Protein Science 9:1753-1773.

Frahm, N, Yusim, K, Suscovich, TJ, Adams, S, Sidney, J, et al. 2007. Extensive HLA class I allele promiscuity among viral CTL epitopes. European Journal of Immunology 37:2419-2433.

Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, et al. 1999. Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. Nature 397:436–441.

Giles, BM, Ross, TM. 2011. A computationally optimized broadly reactive antigen (COBRA) based H5N1 VLP vaccine elicits broadly reactive antibodies in mice and ferrets. Vaccine 29:3043–3054.

Giles BM, Ross TM. 2012. Computationally optimized antigens to overcome influenza viral diversity. Expert Review of Vaccines 11:267–269.

Gill, SR, Fouts, DE, Archer, GL, et al. 2005. Insights on Evolution of Virulence and Resistance from the Complete Genome Analysis of an Early Methicillin-Resistant *Staphylococcus aureus* Strain and a Biofilm-Producing Methicillin-Resistant *Staphylococcus epidermidis* Strain. Journal of Bacteriology 187:2426-2438.

Gitlin, L, Hagai, T, LaBarbera, A, Solovey, M, Andino, R. 2014. Rapid evolution of virus sequences in intrinsically disordered protein regions. PLOS Pathology 10:e1004529. doi: 10.1371/journal.ppat.1004529.

Goldgur, Y, Dyda, F, Hickman, AB, et al. 1998. HIV-1 integrase core domain. Proceedings of the National Academy of Science, USA 95:9150-9154.

Gres, AT, Kirby, KA, KewalRamani, VN, et al. 2015. STRUCTURAL VIROLOGY. X-ray crystal structures of native HIV-1 capsid protein reveal conformational variability. Science 349:99-103.

Guo, W, Han, JW, Zhuang, DM, et al. 2015. Characterization of two HIV-1 infectors during initial antiretroviral treatment, and the emergence of phenotypic resistance in reverse transcriptase-associated mutation patterns. Virology Journal 12:187-195.

Heneine, W, Kashuba, A. 2012. HIV prevention by oral preexposure prophylaxis. Cold Spring Harbor Perspectives in Medicine 2:4140-4150.

Holland, J, Spindler, K, Horodyski, F, et al. 1982. Rapid evolution of RNA genomes. Science 215:1577-1585.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754–755.

Katoh, K, Standley, DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution 30: 772-780. doi: 10.1093/molbev/mst010.

Kesturu GS, et al. 2006. Minimization of genetic distances by the consensus, ancestral, and center-of-tree (COT) sequences for HIV-1 variants within an infected individual and the design of reagents to test immune reactivity. Virology 348:437–448.

Lansdon, EB, Samuel, D, Lagpacan, L, et al. 2010. HIV-1 reverse transcriptase in complex with DNA. Journal of Molecular Biology 397:967-978.

Little, SJ, Holte, S, Routy, JP, et al. 2002. Antiretroviral-drug resistance among patients recently infected with HIV. New England Journal of Medicine 347:385-394.

Lynch, RM, Boritz, E, Coates, EE, et al. 2015. Virologic effects of broadly neutralizing antibody VRC01 administration during chronic HIV-1 infection. Science Translational Medicine 7:319ra206.

McCloskey RM, Liang RH, Harrigan PR, et al. 2014. An evaluation of phylogenetic methods for reconstructing transmitted HIV variants using longitudinal clonal HIV sequence data. Journal of Virology 88:6181–6194.

McGuffin, LJ, Bryson, K, Jones, DT. 2000. The PSIPRED protein structure prediction server. Bioinformatics 16:404–405.

Montaner, JS, Wood E, Kerr T, et al. 2010. Expanded highly active antiretroviral therapy coverage among HIV-positive drug users to improve individual and public health outcomes. Journal of Acquired Immune Deficiency Syndrome 55:S5-9.

Narayan, O, Clements, JE. 1989. Biology and Pathogenesis of Lentiviruses. Journal of General Virology 70:1617-1639.

Ortiz, JF, MacDonald, ML, Masterson, P, Uversky, VN, Siltberg-Liberles, J. 2013. Rapid evolutionary dynamics of structural disorder as a potential driving force for biological divergence in flaviviruses. Genome Biology Evolution 5:504-513.

Palmisano, L, Vella, S. 2011. A brief history of antiretroviral therapy of HIV infection: success and challenges. Annali dell'Istituto Superiore di Sanità 47: 44-48.

Palmenberg, AC, Spiro, D, Kuzmickas, R, et al. 2009. Sequencing and Analyses of All Known Human Rhinovirus Genomes Reveal Structure and Evolution. Science 324:55-59.

Perez-Molina, JA, Rubio, R, Rivero, A, et al. 2017. Simplification to dual therapy (atazanavir/ritonavir plus lamivudine) versus standard triple therapy [atazanavir/ritonavir plus two nucleos(t)ides] in virologically stable patients on antiretroviral therapy: 96 week results from an open-label, non-inferiority, randomized clinical trial (SALT study). Journal of Antimicrobial Chemotherapy 72:246-253.

Qu, K, Glass, B, Dolezal, M, et al. 2018. Structure and architecture of immature and mature murine leukemia virus capsids. Proceedings of the National Academy of the Sciences, USA 115:E11751-E11760. doi: 0.1073/pnas.1811580115.

Rahaman, J, Siltberg-Liberles, J. 2016. Avoiding regions symptomatic of conformational and functional flexibility to identify antiviral targets in current and future coronaviruses. Genome Biology and Evolution 8:3471-3484.

Ronquist F, Huelsenbeck JP. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572–1574.

Rosen, S, Maskew, M, Fox, MP, et al. 2016. Initiating antiretroviral therapy for HIV at a patient's first clinic visit: the RapIT randomized controlled trial. PLOS Medicine 13. DOI: 10.1371/journal.pmed.1002015.

Rudikoff, S, Giusti, AM, Cook, WD, Schraff, MD. 1982. Single amino acid substitution ltering antigen-binding specificity. Proceedings of the National Academy of Sciences 79:1979-1983. doi: https://doi.org/10.1073/pnas.79.6.1979.

Sayers, EW, Barrett, T, Benson, DA, et al. 2009. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 37:D5-15. doi: 10.1093/nar/gkn741.

Schrodinger, LLC. 2015. The {PyMOL} Molecular Graphics System, Version~1.8.

Sievers, F, Wilm A, Dineen D, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology **7**:539. doi: 10.1038/msb.2011.75.

Siltberg-Liberles, J, Grahnen, JA, Liberles, DA. 2011. The evolution of protein structures and structural ensembles under functional constraint. Genes (Basel) 2:748-762.

Simon, V, Vanderhoeven, J, Hurley, A, et al. 2002. Evolving patterns of HIV-1 resistance to antiretroviral agents in newly infected individuals. Aids 16:1511-1519.

Sol, M, Tirado, C, Yoon, K. 2003. Antibody-Dependent Enhancement of Virus Infection and Disease. Viral Immunology 16. doi: https://doi.org/10.1089/088282403763635465.

Tanser, F, Barnighausen, T, Grapsa, E, et al. 2013. High coverage of ART associated with decline in risk of HIV acquisition in rural KwaZulu-Natal, South Africa. Science 339:966-971.

Tirado-Rives, J, Jorgensen, WL. 2006. Contribution of conformer focusing to the uncertainty in predicting free energies for protein-ligand binding. Journal of Medicinal Chemistry 49:5880-5884.

Trottier, B, Lake, JE, Logue, K, et al. 2017. Dolutegravir/abacavir/lamivudine versus current ART in virally suppressed patients (STRIIVING): a 48-week, randomized, non-inferiority, open-label, Phase IIIb study. Antiviral Therapy 22:295-305.

Wang, H, Wolock, TM, Carter, A, et al. 2015. Estimates of global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2015: the global burden of disease study 2015. The Lancet HIV 3:e363. doi: 10.1016/S2352-3018(16)30087-X.

Waterhouse, AM, Procter, JB, Martin, DMA, et al. 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191. doi:10.1093/bioinformatics/btp033.

Werthiem, JO, Worobey, M. 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. PLOS Computational Biology 5:e1-9. doi: 10.1371/journal.pcbi.1000377.

Xue, B, Blocquel, D, Habchi, J, et al. 2014. Structural disorder in viral proteins. Chemical Reviews 114:6880-6911.

Yu, C, Niu, X, Jin, F, et al. 2016. Structure-based inhibitor design for the intrinsically disordered protein c-Myc. Scientific Reports 6:22298. doi:10.1038/srep22298.

Yufenyuy, EL, Aiken, C. 2013. The NTD-CTD intersubunit interface plays a critical role in assembly and stabilization of the HIV-1 capsid. Retrovirology 10:1-14.