EXPLORING HIV RNA STRUCTURE DIVERSITY

By

Timothy Isham Shaw

(Under the Direction of Ming Zhang)

ABSTRACT

HIV-1 has several mechanisms that can significantly impact its genetic diversity. A high rate of mutation coupled with a strong tolerance for sequence change has allowed HIV-1 to evolve a number of machineries to evade immune control. RNA secondary structure was recently found critical in directing recombination and replication mechanisms. In our work, we assessed existing RNA structure modeling technologies for HIV application and developed a novel RNA structure prediction pipeline. In our pipeline, to compliment different prediction strategies, we combined experimental and computational methods to optimize HIV RNA structure prediction accuracy. The pipeline was applied to examine recombination and replication mechanisms. Based on the comparison of B subtype derived from complete genomes and from the recombinants CRF07-08, we found RNA structure variations at VPR-ENV splice donor/acceptor sites and at the NEF/LTR region. To quantify the RNA structure space, we further developed a measurement that uses Shannon Entropy to capture the distribution of Boltzmann un-pairing probabilities. Through our quantification, we were able to estimate the ribosomal frameshift efficiency across various HIV subtypes. Our work revealed that the frameshift element can be clustered according to different subtypes, and recombinants of the two subtypes tend to have identical frameshift elements in both HIV populations. Potential association between frameshift efficiency and disease progression was observed. This work can help us address existing

knowledge gaps on replication and recombination mechanisms that could lead to novel antiviral therapies and HIV/AIDS vaccines development.

INDEX WORDS: RNA secondary structure, HIV, evolutionary diversity

EXPLORING HIV RNA STRUCTURE DIVERSITY

By

Timothy Isham Shaw

B.S., Georgia Institute of Technology, 2008

A Dissertation Submitted to the Graduate Faculty

of the University of Georgia in Partial Fulfillment

of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2014

EXPLORING HIV RNA STRUCTURE DIVERSITY

By

Timothy Isham Shaw

Approved:

Major Professor: Ming Zhang

Committee: Russell Malmberg

Stephen Rathbun

Shaying Zhao

Electronic Version Approved:

Maureen Grasso

Dean of the Graduate School

The University of Georgia

May 2014

DEDICATION

To my parents.

ACKNOWLEDGEMENTS

Looking back, I felt truly blessed to have received my PhD training under the guidance of Dr. Ming Zhang. PhD is probably the hardest challenge I have faced in my life. I would have probably given up half way through the PhD, if it wasn't for the constant encouragement from Dr. Zhang. Working at Dr. Ming Zhang's lab has definitely matured me as a scientist and individual. Dr. Zhang, thank you for pushing me to take on challenges, and thank you for being patient with me.

I also want to use this opportunity to thank all the committee members Dr. Russell Malmberg, Dr. Shaying Zhao, and Dr. Stephen Rathbun. Thank you for your time serving on my dissertation committee and providing guidance on my dissertation projects. I enjoyed each of our individual discussions, and I would like to sincerely thank you for your advice and support.

I would also like to thank Dr. Liming Cai and Dr. Russell Malmberg for introducing me to bioinformatics. Prior to entering the UGA IOB program, I was fortunate enough to enroll in Dr. Liming Cai's course in Algorithms for Computational Biology and Dr. Russell Malmberg's course in Essential Biology for Quantitative Scientist. These two courses were probably the reason why I chose to become a bioinformatician. Thank you both for the valuable training I received while I was a member at the RNA Informatics lab.

I would also like to thank all the faculties at the Institute of Bioinformatics: Dr. Jessica Kissinger, Dr. Jeffrey Dean, Dr. Ying Xu, Dr. Liming Cai, Dr. Jonathan Arnold, and many others. I was able to receive a well-rounded Bioinformatics PhD education, and I look forward to the continued growth of our beloved IOB program. To all the students at the Institute of Bioinformatics College of Public Health, and Department of Computer Science: Dr. Anuj Srivastava, Dr. Wen-chi Chou, Amir Manzour, Dr. Yingfeng Wang, Ruan Zheng, Yulun Chiu,

Dr. Eric Talevich, Joshua Bridgers, Arunima Singh, Joydeep Mitra, Tess Griffin, Gretchen Parrott and many others. Thank you all and I will miss you all.

TABLE OF CONTENTS

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

Human immunodeficiency virus (HIV) is the etiological agent for acquired immunodeficiency syndrome (AIDS) (Barre-Sinoussi et al. 1983, Levy et al. 1984, Popovic et al. 1984). It has been estimated that 34 million people were infected with HIV by 2011 (UNAIDS 2012). Thirty years since the HIV discovery, the virus has continued to spread across the globe, causing public health challenges and economic burden worldwide. Currently there are no effective HIV/AIDS vaccines available, although a handful of antiretroviral drugs have been developed (Thompson et al. 2012). The major obstacle that hinders development of vaccines and improved antiviral therapies is HIV's high replication errors (Mansky and Temin 1995), high turnover rate (Ho 1997), and frequent recombination events (Zhuang et al. 2002). Therefore, the tracking of the genetic diversity of HIV and its relationship with immune response is of profound importance in global public health.

HIV RNA Structure Diversity

Across all three domains of life, RNA structure plays a key role in multiple cellular processes. Given HIV's short RNA genome, RNA structure also plays an essential role regulating HIV's life cycle. This dissertation aims to examine RNA secondary structure's contribution to HIV's genetic diversity. In particular, we are interested in using RNA secondary structure to elucidate important HIV replication mechanisms involving recombination and frameshifting. The RNA structure in HIV has been extensively studied; however, they mostly focus on the long terminal repeats (LTR) (Karn 2000, Wilkinson et al. 2008, Lever et al. 1989, Clever, Sassetti, and Parslow 1995) that occupy less than 10% of the viral genome (Los-Alamos-HIV-Sequence-Database). It is only recently that RNA structure solution for a single B subtype HIV NLM4-3 strand was

made available, revealing a complex RNA structural landscape embedded in the protein coding region (Watts et al. 2009). By taking advantage of this recently published experimental data (Watts et al. 2009), we can perform a systematic evaluation of computational RNA prediction methods in HIV modeling, which remains a knowledge gap in current HIV research. Furthermore, we will establish an RNA prediction pipeline to investigate HIV diversity through RNA secondary structure.

RNA Background

Alexander Rich made the landmark discovery in year 1956 that single-stranded RNAs are capable of hybridizing (Rich and Davies 1956), resulting in stimulated discussions of RNA structure-function implications (Judson 1979). He further predicted that segments of double stranded RNA (hairpins) could potentially control and regulate protein synthesis (Rich 1961), which was validated through the discovery of microRNA in *Caenorhabditis elegans* (Lee, Feinbaum, and Ambros 1993). RNA is now widely recognized for being a versatile molecule that possesses similar catalytic properties as proteins (Kruger et al. 1982). Non-coding RNA (ncRNA) is a common term for functional RNA molecules that do not code for protein (Mattick and Makunin 2006). Classification of ncRNA could be separated into infrastructural ncRNAs and regulatory ncRNAs (Mattick and Makunin 2006). Infrastructural ncRNA include transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), spliceosomal uRNAs (snRNAs), and small nucleolar RNAs (snoRNAs). While tRNA and rRNA are involved in central dogma role of translating protein-coding genes (Crick 1970), other infrastructural ncRNAs for instance snRNA and snoRNA are capable of directing splicing and nucleotide modifications (Matera, Terns, and Terns 2007). Regulatory ncRNAs in humans includes: microRNAs (Hobert 2008), small interfering RNA (siRNA) (Wianny and Zernicka-Goetz 2000), piwi-interacting RNA (piRNA)

(Girard et al. 2006) and riboswitches (Breaker 2008), and these ncRNAs are capable of up-and-down regulating gene expressions in both cis and trans (Hartzog and Martens 2009, Guil and Esteller 2012). For ncRNA greater than 200 nucleotides, they are categorized into a special ncRNA class called long non-coding RNAs (lncRNA) (Mattick and Makunin 2006, Ponting, Oliver, and Reik 2009). LncRNA have a complex role in host epigenetic regulation (i.e. RNA directed gene silencing, chromatin modification, structural changes to chromatin, and gene promoter regulation) (Kaikkonen, Lam, and Glass 2011).

A widely accepted view is that conserved RNA structure can imply functional ncRNA (Gruber et al. 2010), for example, lncRNAs contains low primary level sequence conservation but functional conservation is present in its RNA secondary structures (Qu and Adelson 2012). In short ncRNA structure can also determine functional interaction for ribozymes and pre-processing of precursor microRNAs (Svoboda and Di Cara 2006, Han et al. 2006). Given the conserved importance of RNA secondary structure, RNA secondary structure solutions for ncRNAs could provide clues to their function (Washietl et al. 2005) (i.e. shaping gene expression (Ameres and Zamore 2013), genome integrity (Francia et al. 2012) and cellular regulatory network (Qu and Adelson 2012)).

HIV RNA Structures

In addition to RNA structure in humans, RNA has also been discovered to play critical roles in the viral life cycle (Damgaard et al. 2004, Rodriguez-Alvarado and Roossinck 1997, Pelletier and Sonenberg 1988). In the hepatitis delta virus (HDV), long terminal repeat contains a ribozyme critical to processing replication products (Been and Wickham 1997). In HIV, the genome is punctuated by multiple RNA structure elements both in the long terminal repeat (LTR) and protein junctions (Watts et al. 2009), and HIV reverse transcription is initiated through tRNA

binding to the LTR primer binding site (Frankel and Young 1998). HIV RNA structure study has largely focused on LTR: Seven out of ten RNA element database entries are within the LTR region (See Table 1) (Gardner et al. 2009). Although the LTR RNA structures have tremendous impact on HIV life cycle, the structures and function of RNA structures in the rest of the HIV genome remains elusive. In the following section, we will review key RNA secondary structures in HIV (See figure 1.1 and for the LTR RNA secondary structures and table 1.1 for list of Rfam HIV RNA structures).

The trans-activation response element (TAR) is the first identified RNA element that resides in both 5' and 3' LTR repeats (R) and located after the +1 transcription initiation site (Karn 2000). TAR is capable of interacting with Trans-Activator of Transcription (TAT), a transactivator of HIV gene expression (Keen, Churcher, and Karn 1997). TAR RNA structure primarily consist of a hairpin loop with a bulge, which is found to be necessary for TAR-TAT binding (Roy et al. 1990). Due to structural similarity between TAR and pre-microRNAs, the possibility for TAR to have microRNA function is heatedly debated (Klase et al. 2009, Whisnant et al. 2013). In addition, TAR has been indicated to have host protein binding ability, presenting additional TAR induced regulatory functions within the host (Bannwarth and Gatignol 2005).

The polyA signal is located on both 5' and 3' LTR R regions, sitting beside the TAR RNA structure (Wilkinson et al. 2008). The polyA signal is embedded in the loop region of an RNA hairpin (Wilkinson et al. 2008). Interestingly, the polyA signal is capable of functioning without the RNA hairpin structure, and the stabilization of the hairpin structure actually have a negative effect on polyadenylation efficiency (Klasens, Das, and Berkhout 1998, Klasens et al. 1999). The result indicates that the RNA hairpin structure can regulate the polyadenylation signal (Klasens, Das, and Berkhout 1998, Klasens et al. 1999). Additional studies on the polyA

hairpin indicate potential metastable conformations of the RNA structures (Gee, Kasprzak, and Shapiro 2006); however, functional implications for these suboptimal structures have not been identified.

Lever et al first identified regions in LTR required for efficient HIV genome packaging during replication (Lever et al. 1989). The psi RNA structure contains a four hairpin stem loop. The RNA structure allows the psi element to bind to the nucleocapsid (NC) protein (Clever, Sassetti, and Parslow 1995), a protein that mediates efficient reverse transcription (Levin et al. 2010). The four hairpin loop in psi packaging is labeled as: SL1, SL2, SL3, and SL4 (Clever, Sassetti, and Parslow 1995). More recently, a three-dimensional resolution of the packaging signal was computationally predicted indicating a putative GAG polyprotein binding pocket on SL2 and SL3 with SL1 oriented away (Stephenson et al. 2013). In addition to psi packaging, the hairpin loops are multifunctional for dimerization (Clever, Sassetti, and Parslow 1995, Wilkinson et al. 2008) and splicing (Clever, Sassetti, and Parslow 1995).

The dimerization initiation site (DIS) is also known as SL1 of the psi packaging (Clever, Sassetti, and Parslow 1995, Wilkinson et al. 2008). Prior to HIV packaging, two HIV strands form a dimer at the DIS region during HIV capsid packaging (Moore and Hu 2009). The dimerization element links two HIV genomes through homodimerization based on the RNA hairpin structure (Mujeeb et al. 1999). The dimerization step includes the initial formation of a kissing loop complex and further refolding into a palindrome base pairing duplex (Mujeeb et al. 1999). Recombination events are strongly associated with the dimerization process (Sakuragi et al. 2010).

The splice donor is on the SL2 region of the psi packaging signal (Clever, Sassetti, and Parslow 1995). Abbink et al indicate that strengthened thermodynamic base pairing results in a

severely lower HIV replication efficiency, suggesting the importance of RNA structure in modulating splice efficiency(Abbink and Berkhout 2008). The splice donor signal can also be found in SIV and HIV-2 (Strappe et al. 2003). Given the complex regulation of HIV alternative splicing (Ocwieja et al. 2012), the thermodynamic stability of the RNA structure is hypothesized to be one of the major factors regulating HIV-1 splicing (Mueller, Berkhout, and Das 2013).

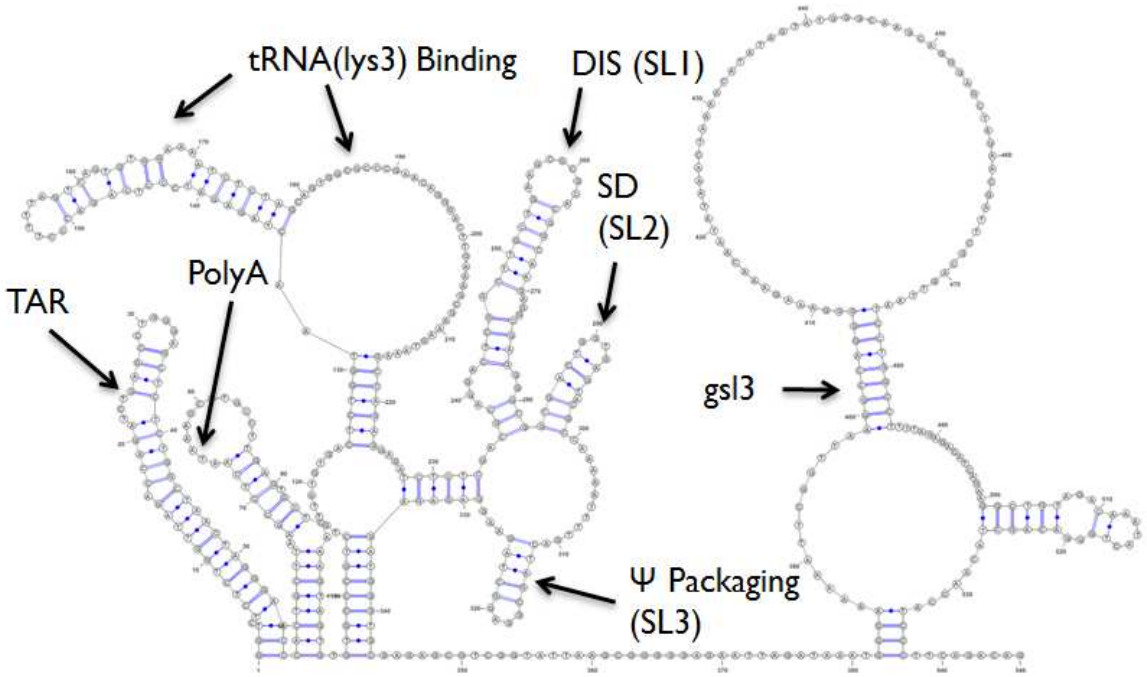


Figure 1.1. HIV Leader (beginning of the HIV genome) RNA structure.

The RNA structure is based on the NLM4-3 strain. Nearly every single hairpin loop that is present in the HIV Leader was given a name, and their specific functions are extensively studied. GSL3 RNA structure is located after the GAG start codon (Damgaard et al. 2004).

Table 1.1 RNA structure elements based on the Rfam database (Gardner et al. 2009)

Name	**Rfam Entry**	**Function (Experimentally Verified)**	**Genomic Region**

PBS	Yes	Primer binding site	LTR
GSL3	Yes	Regulation of splicing and psi packaging	LTR
DIS	Yes	RNA dimerization	LTR
SD	Yes	Splice Donor	LTR
SL3	Yes	Part of Psi Packaging	LTR
SL4	Yes	Part of Psi Packaging	LTR
FE (IPL4)	Yes	Ribosomal Frameshift Event	GAG-POL
POL (IPL6-PDJ5)	Yes	Modulate Ribosome Elongation	POL
RRE	Yes	Interacts with REV to promote nuclear export.	ENV/TAT
TAR	Yes	Interacts with TAT for transcriptional activation. Listed as putative microRNA.	LTR

Snoeck et al have indicated that HIV-1 RNA secondary structure is one of the driving forces that guide and restrict HIV's mutational patterns (Snoeck et al. 2011). Although RNA structural elements are susceptible to mutations (indicated by significant RNA structure difference between lentivirus lineages (Pollom et al. 2013)), at the same time, selective forces are acting on RNA structure to preserve key binding sites and functional domains (Pollom et al. 2013, Wang et al. 2008, Peleg, Trifonov, and Bolshoy 2003). G-to-A nucleotide substitution is a frequently observed mutation pattern in HIV (Mansky and Temin 1995). Although G-to-A nucleotide substitution can affect RNA secondary structure formation (van der Kuyl and

Berkhout 2012), van Hermert et al indicate that adenine nucleotide substitutions appear more frequent in single stranded RNA structure than double stranded RNA structure indicating evolutionary selection on RNA structures (van Hemert, van der Kuyl, and Berkhout 2013). For RNA structures that overlap with protein coding regions, RNA structure are shown to not evolve independently from that of the protein coding regions (Sanjuan and Borderia 2011) and further studies are necessary to fully comprehend the relationship between RNA structure and protein coding regions. With potential fitness cost associated with loss of RNA structure (Parthasarathi et al. 1995, Sanjuan and Borderia 2011), it is conceivable that evolutionary forces are selected on double stranded RNAs to protect HIV genome degradation while balancing fitness cost incurred by amino acid usage preferences.

Overview of HIV Diversity

HIV is categorized into HIV type 1 (HIV-1) and HIV type 2 (HIV-2) (Sharp and Hahn 2011). HIV-1 is hypothesized to result from multiple cross-species (Gao et al. 1999, Plantier et al. 2009, Damond et al. 2004, Hirsch et al. 1989) zoonotic transmission of simian immunodeficiency virus (SIV) (Plantier et al. 2009, Gao et al. 1999) into the human population. Phylogenetic analysis suggested that HIV-1 can be divided into four different groups: major (M), outlier (O), non-M non-O (N) and pending group classification (P). HIV-1 M group is responsible for over 90% of the global HIV infection (Robertson et al. 2000). It is estimated that the M group appeared after 1900s (Korber et al reported as 1931 with 95% confidence interval of 1915 to 1941; and Worobey et al reported as 1908 with 95% confidence interval of 1884-1924) (Korber et al. 2000, Worobey et al. 2008). The M group can be further divided into nine different subtypes: A, B, C, D, F, G, H, J and K with at least 55 circulating recombinant families (Osmanov et al. 2002, Peeters and Sharp 2000, Peeters, Toure-Kane, and Nkengasong 2003, Hemelaar et al. 2006).

Subtypes A and F are further divided into sub-subtypes, A1 and A2 within A clade, and F1 and F2 within F clade (Gao et al. 2001, McCutchan 2006). Figure 1.2 presents the phylogenetic relationship of M group subtypes based on the 2010 HIV reference sequences (Los-Alamos-HIV-Sequence-Database).

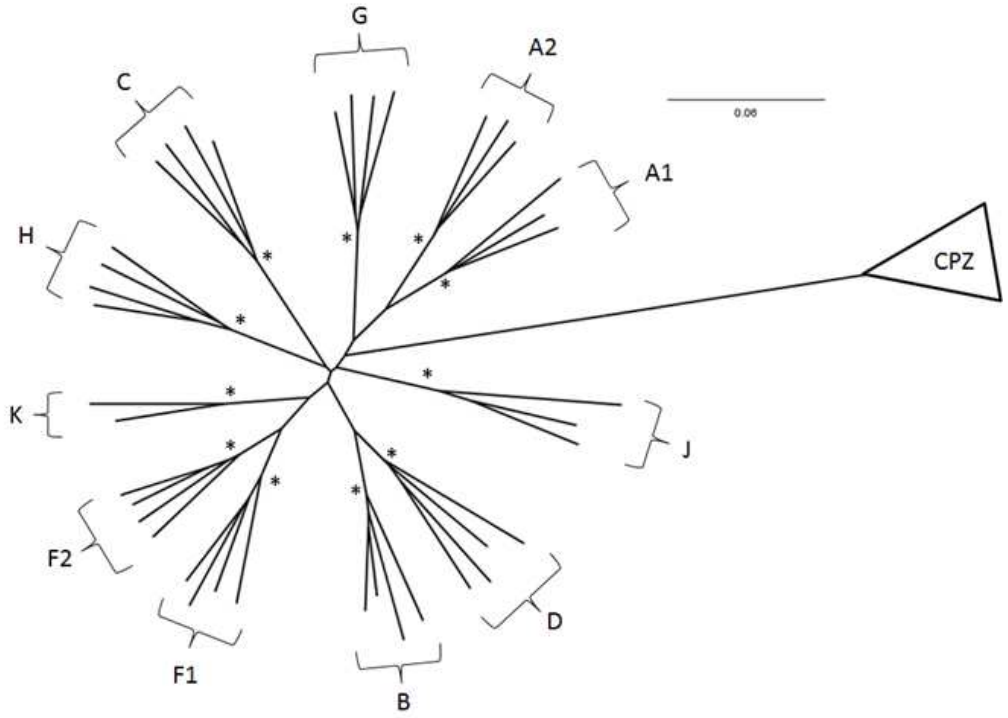


Figure 1.2. Phylogenetic relationship between HIV-1 M group subtypes.

The phylogenetic tree was constructed using neighbor-joining with 500 bootstrap resampling using the 2010 HIV full genome sequences (Los-Alamos-HIV-Sequence-Database). The "" denotes > 80% bootstrap confidence. The SIVcpz (CPZ) clade obtained from the 2010 HIV reference was used as out-group. Time scale: 0.08 nucleotide substitution per site.*

Most HIV-1 M group subtypes form distinct geographical clusters globally due to the founders effect (Rambaut et al. 2004). Subtypes A, B, C, D and F are frequently sampled

globally (Los-Alamos-HIV-Sequence-Database). Subtype B is predominant in Western countries (Los-Alamos-HIV-Sequence-Database) (Figure 1.3), particularly in North America, South America, Western Europe, and Oceana (Buonaguro, Tornesello, and Buonaguro 2007). Asia epidemic is primarily consisted of B, C, and CRF01 (a recombinant between A and E subtype), and East Europe consisted mostly of A1 and A2 sub-subtypes (Buonaguro, Tornesello, and Buonaguro 2007). Sub-Saharan Africa contains the most diverse distribution of non-B subtypes (Hemelaar et al. 2006) (Figure 1.3) and it contains majority of the HIV infected adults (Figure 1.4). Abecasis et al indicated that the M group diversified after the middle of the 20th century, placing time of the most recent common ancestor for CRF01, A1, B, C, D, and G to be 1975, 1954, 1952, 1946, and 1969 respectively (Abecasis, Vandamme, and Lemey 2009).

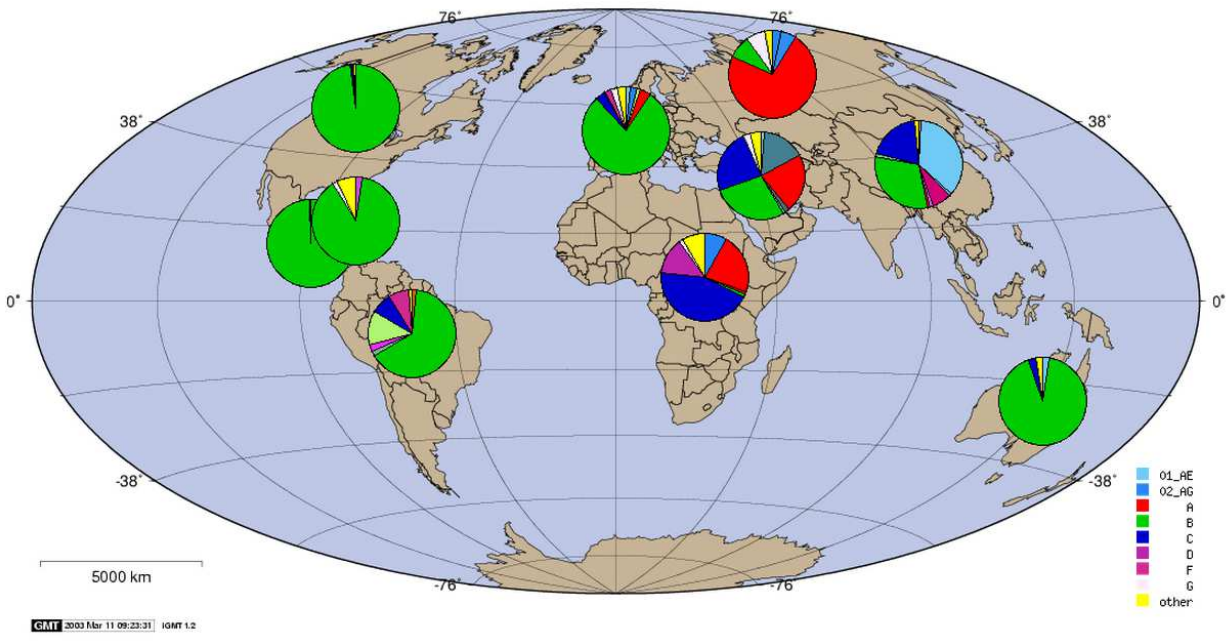


Figure 1.3. Global HIV-1 subtype distribution.

Subtype distribution based on sequences from HIV LANL Database (Kuiken et al. 2012). Image was generated from the LANL HIV geographic search interface using HIV-1 across the world.

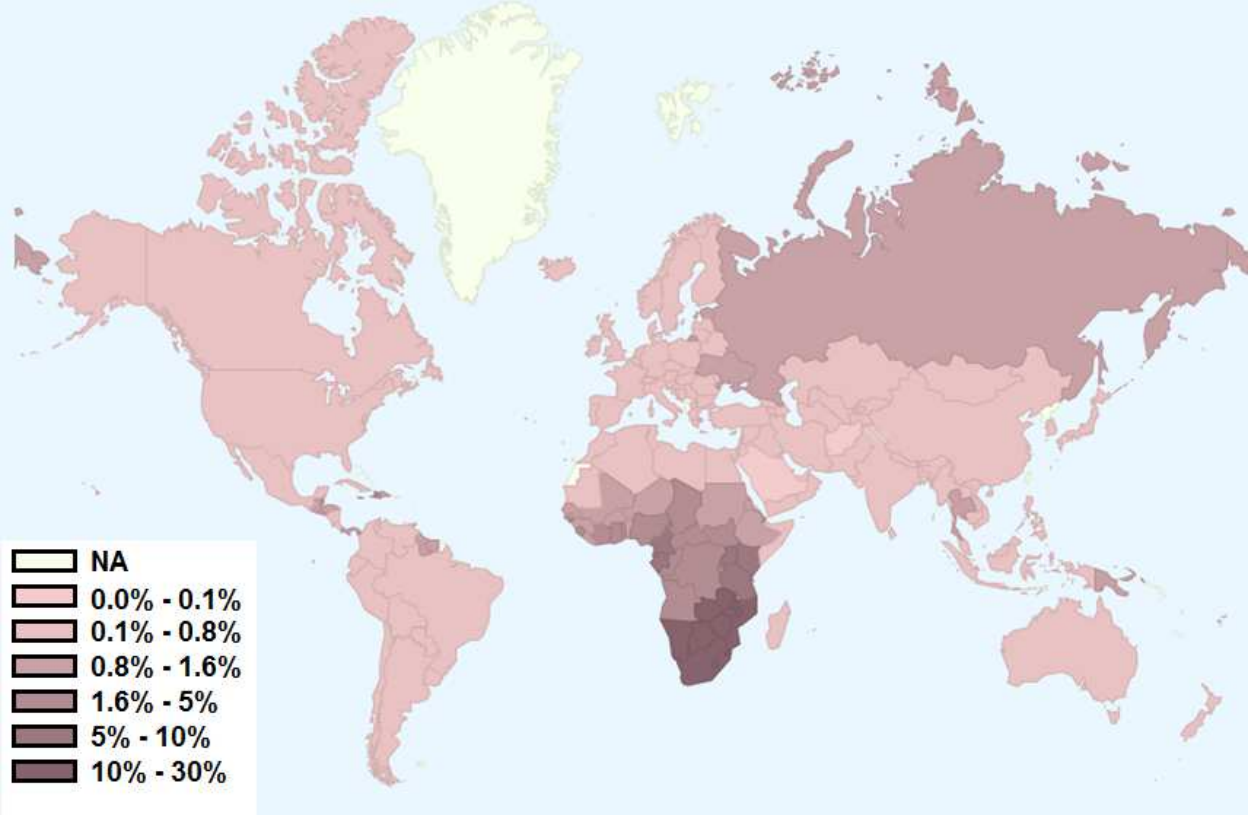


Figure 1.4. HIV-1 global prevalence.

The epidemiology data is based on percentage of adult aged 15-49 living with HIV in each country provided by UNAIDS, 2009 data (UNAIDS 2010).

HIV genome structure and function

A typical HIV genome is a 9500nt long single stranded RNA genome (Los-Alamos-HIV-Sequence-Database) with long terminal repeats (LTR) at the 5' and 3' end (Figure 1.5a). Embedded between the LTR are nine protein coding genes located on three different reading frames. The nine protein coding genes can be further processed into 19 different protein products (Los-Alamos-HIV-Sequence-Database). The protein coding genes are categorized into four major groups: structural proteins, enzymatic proteins, regulatory proteins, and accessory proteins (Los-Alamos-HIV-Sequence-Database). Particularly, structural proteins (GAG, ENV),

enzymatic protein (POL), and certain regulatory protein (TAT) are noted to be essential for HIV replication mechanisms and a transactivator for HIV gene expression (Ensoli et al. 1993). (Table 1.2 and Figure 1.5b)

HIV is coated by the envelope (ENV) protein, and selective pressure from the host immune system contributes significantly to ENV's genetic diversification (Yusim et al. 2002). The envelope polyprotein can code for two structural proteins: gp120 external glycoprotein and the gp41 trans-membrane protein. The gp120 glycoprotein contains five variable loop (V1-V5) localized between conserved regions (Wyatt et al. 1998). These variable regions experience extreme insertion and deletion, and they are typically excluded from phylogenetic analysis (Korber et al. 2001). Many of the nucleotide insertion and deletion involves N-linked glycosylation sites, the amino acid pattern of NX[S or T], with X being any amino acid except Proline (Stanley, Schachter, and Taniguchi 2009). N-linked glycosylation plays an essential role for HIV virus to induce conformational changes that diminishes binding of gp120-specific antibodies (Si, Cayabyab, and Sodroski 2001).

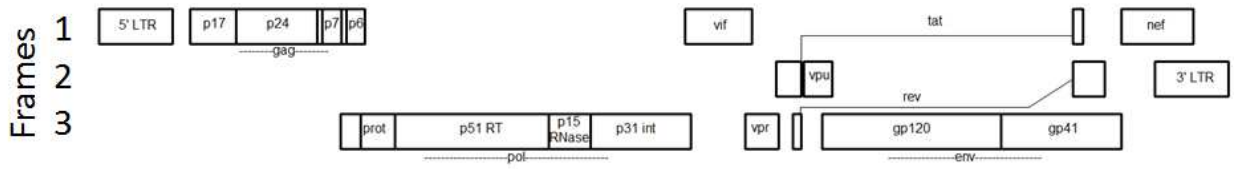


Figure 1.5A. HIV Genome Architecture.

The rectangles other than 5' LTR and 3' LTR represent different protein coding genes. There are three different reading frames that can be accessed through alternative splicing and frameshift events. The gag gene encodes: p17, p24, p7, and p6 proteins that play critical roles in

virus assembly and a driving force for virus assembly (Gheysen et al. 1989, Göttlinger 2001). Pol contains four polyproteins: protease (prot), reverse transcriptase (p51), RNase (p15), and integrase (p31) that are important for various process in the HIV life cycle: processing polyprotein, production of DNA genome, trimming of RNA genome, and integration to host (Greene 2005). Envelope protein contains two polyprotein of GP120 and GP41 that represent glycoprotein that plays a critical role of interacting with the host immune system (Greene 2005, Arrildt, Joseph, and Swanstrom 2012). The genome structure is drawn on the HXB2 (accession K03455) scale.

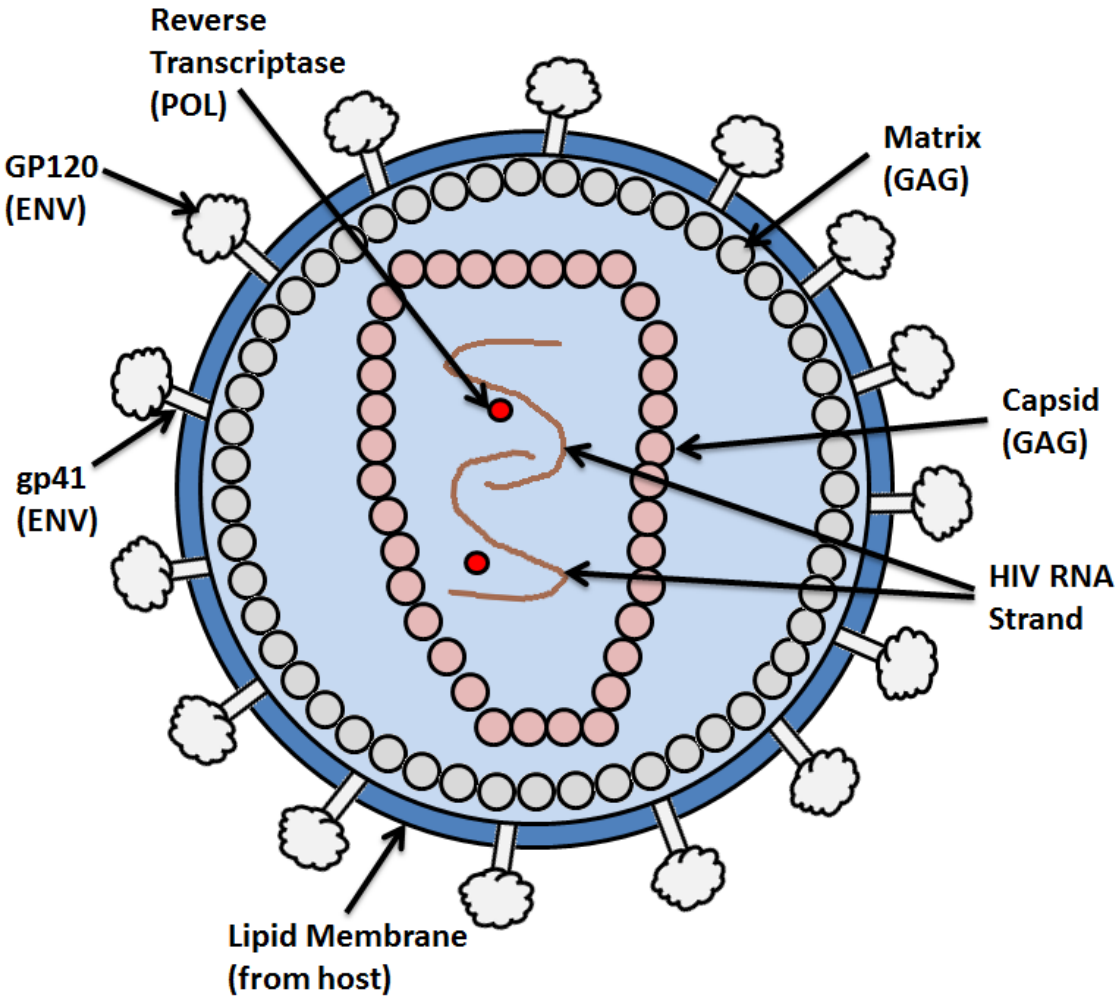


Figure 1.5B. Schematic presentation of an HIV virion.

ENV, POL, and GAG are shown. The lipid membrane was derived from the host through the budding process. Each virion packages two HIV RNA strands that would create opportunity for recombination in case the two strands are of different subtypes.

Table 1.2. Protein coding genes and their function.

Category	Protein Name	Function	Reference
Structural Proteins	Gag	Encode the capsid protein. The p55 precursor can be processed by the viral protease into p17(Matrix), p24(Capsid), p7(Nucleocapsid), and p6 protein.	(Greene 2005, Freed 1998)
	Env	Contains gp160 which can be processed into an external glycoprotein gp120 and a transmembrane glycoprotein gp41. gp120 and gp41 are bonded by a weak non-covalent interaction	(Greene 2005, Arrildt, Joseph, and Swanstrom 2012)
Enzymatic Proteins	Pol	Encodes the protease (pro), reverse transcriptase (RT), ribonuclease H(RNase H) and integrase(IN). The Pol is part of a Gag-Pol complex and only expressed through a frameshift event.	(Greene 2005, Hung et al. 1998)

Regulatory Proteins	Tat	A transactivator of HIV gene expression. Interacts with TAR a RNA element for transcription initiation and elongation.	(Keen, Churcher, and Karn 1997, Greene 2005, Roy et al. 1990)
	Rev	Interacts with RRE to promote nuclear export.	(Greene 2005, Pollard and Malim 1998)
Accessory Proteins	Vif	Highly conserved protein that promotes viral infectivity. Vif protein could also bind to HIV RNA within the cell cytoplasm.	(Greene 2005, Zhang et al. 2000)
	Vpr	Facilitate nuclear import of HIV-1 pre integration complex. It also induces cell arrest.	(Greene 2005, Bukrinsky and Adzhubei 1999)
	Vpu	Found only in HIV. Can down regulate CD4 and promote virion release	(Greene 2005,

			Klimkait et al. 1990, Willey et al. 1992)
	Nef	Essential for disease progression and viral spread. Capable of elevating viral infectivity	(Greene 2005, Das and Jameel 2005)

RNA Structure Directed Mechanisms

HIV genetic diversity plays a critical role for virus-to-human adaptation (Kawashima et al. 2009, Wain et al. 2007, Kearney et al. 2009). Factors influencing genetic diversity includes: a high replication error (3.4e-5 mutations per bp per cycle) (Mansky and Temin 1995), frequent recombination events (an estimated 2.8 recombination per genome life cycle) (Zhuang et al. 2002) and a high turnover rate (an estimated 1.03e10 virions released each per day) (Ho 1997). While external factors such as antiretroviral drugs (Beerenwinkel et al. 2002), host immunity (Deeks and Walker 2007, Chakrabarti and Simon 2010) and transmission route (Tournoud et al. 2008, Yu et al. 2009, Soto-Ramirez et al. 1996) could also impact viral diversity, in our dissertation, our focus will primarily be on understanding recombination and replication mechanisms directed by RNA secondary structures.

Recombination Mechanism

HIV recombination occurs when two genetically different RNA strands are simultaneously infecting a cell (Galetto and Negroni 2005). During reverse transcription, the reverse transcriptase (RT) shuffles between the two genetically distinct RNA templates producing a chimeric DNA strand (Negroni and Buc 2001). The emergence of a similar recombinant form can occur, and common recombinants with more than three epidemiological unlinked individuals is called a circulating recombinant form (CRF) (Robertson et al. 2000). In general, HIV recombination is conditioned on three events: (1) the two parental viruses must be able to infect the same patient and cell type (Galetto and Negroni 2005). (2) The two RNA strand must be capable of forming dimers (Galetto and Negroni 2005). (3) The chimeric structural proteins derived from the two HIV strands must not interfere with one another (Galetto and Negroni 2005).

Compared to HIV genetic diversity accumulated from replication error, recombination events allow HIV to explore a broader level of sequence space (Burke 1997). Understanding recombination mechanisms will allow us to improve our HIV genotyping model. Currently, there are three proposed mechanisms that enable recombination to occur. (1) A forced copy choice of the reverse transcript. The degradation of the RNA strand could lead to breaks on the RNA template stalling RT. The stalled RT is then rescued by ligating a RNA donor to the nascent RNA strand (Coffin 1979) leading to the transfer of RT (Peliska and Benkovic 1992, Galetto and Negroni 2005). (2) Recombination induced by RT pausing during reverse transcription (DeStefano et al. 1992). The mechanism described here is similar to a forced copy choice model in which pausing (stalling) of RT induces template switching (DeStefano et al. 1992, Galetto and Negroni 2005). Under the RT pausing model, strand break is one factor

contributing to RT pausing. Other factors such as sequence motif (DeStefano et al. 1992, DeStefano, Bambara, and Fay 1994) or decreasing nucleotide pools (Wu et al. 1995) could also induce RT pausing (Galetto and Negroni 2005) (3) RNA secondary structure directed recombination. RNA secondary structure could also potentially induce RT pausing (Harrison et al. 1998). However, recombination rarely exists at the base of the RNA hairpin, but instead they tend to occur at the loop region of the RNA hairpin (Galetto and Negroni 2005). The highly concentrated recombination in the loop region has prompted a donor/acceptor interaction model hypothesis (Moumen et al. 2003). The model propose that the donor strand's RNA structure is destabilized and forms complimentary base pairing with the structural form of the acceptor strand with a template switch occurring similar to branch migration (Moumen et al. 2003, Galetto and Negroni 2005). While key mechanisms responsible for driving the recombination event are still obscure, we believe a comparative approach of RNA structure across subtypes could help us shed light on the exact recombination mechanism.

Replication Mechanisms

A high turnover rate contributes to the increase in virus genetic diversity (Ho 1997). HIV variants with broad cell type replication efficiency are associated with AIDS development (Cheng-Mayer et al. 1988, Tersmette et al. 1989). HIV gene expression profile is one of the determining features influencing replication fitness (Malim and Emerman 2008, Stoltzfus 2009). The gene expression profile is regulated by both frameshift events and alternative splicing. Frameshift elements between Gag-Pol can determine level of expression for the Pol gene product that contains a number of replication important enzymes (Shehu-Xhilaga, Crowe, and Mak 2001, Dinman 2012). Changes in alternative splice patterns could also influence gene transcript

expression patterns, modifying the cellular pathways into a replication friendly environment (Malim and Emerman 2008, Stoltzfus 2009).

Alternative splicing allows HIV to generate more than 47 different mRNA transcripts (Purcell and Martin 1993). HIV mRNA transcript can be separated into three different size classes: (1) un-spliced 9-kb transcripts which generate Gag and Pol proteins (Tazi et al. 2010, Stoltzfus 2009). (2) A single spliced 4-kb transcript encoding Env, Vif, Vpr, and Vpu (Tazi et al. 2010, Stoltzfus 2009). (3) Multiple spliced 2-kb transcripts encode regulatory proteins like Tat, Rev, and Nef (Tazi et al. 2010, Stoltzfus 2009). The splice signal involves a 5' splice site, 3' splice site, and a branch point sequence (Stoltzfus 2009). The alternative splice event is modulated by core splicing signal's exon definition (Hoffman and Grabowski 1992, Robberson, Cote, and Berget 1990), splice enhancers/silencers (Matlin, Clark, and Smith 2005, Wang and Burge 2008), and RNA secondary structures (Buratti and Baralle 2004, Abbink and Berkhout 2008).

The RNA structure frameshift element sits between HIV Gag and Pol polyproteins, enabling the production of two polyproteins (Parkin, Chamorro, and Varmus 1992). The frameshift element consists of a heptanucleotide that is composed of a UUUUUUA slippery sequence, and an 8-nt spacer located between the slippery sequence and RNA stem loop (Mouzakis et al. 2013, Kollmus et al. 1994). During translation, the polymerase would pause and slip on the slippery sequence causing a programmed ribosomal frameshift (PRF) event producing an -1 frame protein (Jacks, Power, et al. 1988). The frameshift mechanism is found to regulate the ratio of Gag to Gag-Pol ensuring effective assembly of the infectious virus particle (Park and Morrow 1991). Jack et al first discovered the -1 frameshift mechanism in Rous sarcoma virus (Jacks, Madhani, et al. 1988) and later in HIV-1 (Jacks, Power, et al. 1988). In

addition to retroviruses, the programmed ribosomal frameshifting mechanism and structural characteristics are widely conserved across eukaryotes and prokaryotes (Farabaugh 1996, Chamorro, Parkin, and Varmus 1992, Jacks, Madhani, et al. 1988, Jacks, Power, et al. 1988, Jacks et al. 1987, Jacks and Varmus 1985).

In HIV B subtype, the -1 PRF induced frameshifting roughly 5% of the time (Jacks, Power, et al. 1988). Mutagenesis experiments indicate significant changes to frameshift efficiency that prohibits viral replication (Dulude et al. 2006, Hung et al. 1998, Shehu-Xhilaga, Crowe, and Mak 2001). These studies have resulted in an increased interest in developing frameshift drug targets (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009, Hung et al. 1998, Dinman, Ruiz-Echevarria, and Peltz 1998). For a majority of the retroviruses, the thermodynamic stability of a pseudoknot structure after the slippery heptamer is a major determinant of frameshift efficiency (Nixon and Giedroc 2000, Giedroc, Theimer, and Nixon 2000, Nixon et al. 2002). In HIV, the RNA pseudoknot is absent within the stem loop, instead, a stable stem-loop structure is found to sufficiently promote the -1 PRF event (Yu et al. 2011). Existing hypothesis states that the stability of the RNA structure can hinder the unwinding of the elongating ribosome; prompting the ribosome to pause and shift in the -1 nucleotide position (Green et al. 2008, Plant and Dinman 2005). One area that is not fully understood is the contribution of subtype variation's impact on RNA frameshift events. As frameshift element is a critical component to replication, a comparative approach of studying this genomic region could help us compare replication mechanisms among different subtypes.

RNA structure modeling

Despite tremendous advances in RNA crystallography, NMR and chemical modification, determining the RNA structure remains experimentally difficult (Schroeder 2009).

Computational methods are important tool for inferring RNA structure and function. Computational modeling of RNA structure has primarily relied on recursive algorithms to generate all base-pairing combinations given an RNA sequence (Nussinov and Jacobson 1980, Eddy 2004). Structure prediction algorithms can be separated into two categories: thermodynamic and non-thermodynamic probabilistic, based on stochastic context-free grammar (SCFG) (Schroeder 2009). SCFG establishes the rules used for parsing the RNA sequence. Backtracking of the path through maximizing the RNA grammar probabilities can be used to predict an RNA structure (Durbin 1998a, Dowell and Eddy 2004). The thermodynamic approach is similar to the SCFG method but provides ranks based on free-energy estimation through RNA neighboring thermodynamic parameters (Znosko et al. 2002, Xia et al. 1998). RNA structure prediction can be further divided into de novo method (Hofacker and Stadler 2006, Dirks and Pierce 2003, 2004, Do, Woods, and Batzoglou 2006) and evolutionary based method (Bernhart et al. 2008, Sukosd et al. 2011). The de novo based approach simply predicts the RNA structure based on a single sequence (Hofacker and Stadler 2006, Dirks and Pierce 2003, 2004, Do, Woods, and Batzoglou 2006). The evolutionary based approach incorporates compensatory mutation or evolutionary history of an RNA alignment to guide the prediction of the RNA structure (Bernhart et al. 2008, Sukosd et al. 2011). More recently, the development of selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) has resulted in the use of auxiliary chemical probing information to guide RNA folds (Sukosd et al. 2012, Washietl, Hofacker, et al. 2012, Reuter and Mathews 2010).

SHAPE Technologies and HIV NLM4-3 Strand

HIV RNA structure research has largely focused on the LTR, while the rest of the viral genome that occupy over 90% of the genome has been widely neglected. The reason is the minimal RNA

structure conservation in the protein coding region (Knoepfel and Berkhout 2013), and the algorithmic and experimental challenge of solving RNA structure that could exist in suboptimal states (Weeks 2010, Washietl, Will, et al. 2012). Watts et al. recently published a full subtype B HIV NLM4-3 strain RNA secondary structure (Watts et al. 2009), identifying complex architectural patterns in the HIV RNA structure landscape. Many HIV RNA researches have utilized this NLM4-3 RNA structure as the reference and found association with epitopes (Snoeck et al. 2011) and drug resistance mutations (Sanjuan and Borderia 2011). However, these studies were performed under the assumption that NLM4-3 RNA structure is conserved across the B subtype or the HIV M group, and we believe this assumption deserves further investigation.

Based on the "thermodynamic hypothesis", folding of macromolecules is primarily governed by the minimization of its free energy (Anfinsen 1973). However, RNA structure prediction often possess a complicated folding landscape (Manzourolajdad et al. 2013, Gonzalez 2008, Solomatin et al. 2010, Chen and Dill 2000), and suboptimal conformation can be favored over the optimal conformation(Zuker 1989) (i. e. clover leaf tRNA structure (Bernhart et al. 2006)). Existing prediction methods widely assume the minimal free energy RNA structure conformation (Schroeder 2009), but these RNA structure predictions could fall victim to (as Michael Zuker puts it) the "ill-conditioning of the folding program", the disconnect between computational assumptions and biological reality (Zuker 1986). One method to overcome the ill-conditioning is through the correction of RNA structure prediction using chemical modification information (Sukosd et al. 2013, Mathews et al. 2004) such as selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) (Deigan et al. 2009).

The SHAPE technology is a high throughput interrogation method for evaluating local RNA nucleotide flexibility (Low and Weeks 2010, Lucks et al. 2011). The technology chemically probes each nucleotide to obtain a reactivity index, the propensity for each nucleotide to base pair (Weeks 2010). For the last 25-30 years, different chemical probing techniques based on small organic molecules, metal ions, or RNase enzymes has been developed (Ehresmann et al. 1987, Stern, Moazed, and Noller 1988, Tullius and Greenbaum 2005); these chemicals probe the RNA structure free region which is detected by primer extension (Weeks 2010, Wilkinson, Merino, and Weeks 2006). The length of the resulting cDNA fragment could be tracked back to the site of the modification (Weeks 2010). In contrast, SHAPE technology is revolutionary for its ability to interrogate every single RNA nucleotide and its ability to circumvent two major problems of the base-selective chemical agent: (1) being sensitive to both solvent accessibility and electrostatic features. (2) Its difficulty to place reactivity measurements from multiple reagents (Mortimer, Johnson, and Weeks 2009, Lavery and Pullman 1984, Weeks 2010). Recent results comparing RNA prediction with and without the extended use of SHAPE information have led to the favored extended incorporation of SHAPE reactivity into the RNA structure prediction (Deigan et al. 2009, Hajdin et al. 2013). Hajdin et al show that the incorporation of SHAPE reactivity is capable of achieving close to 93% accurate prediction of base pairs as compared to 73% accurately predicted base pairs for pseudoknot structure dataset (Hajdin et al. 2013). Sukosd et al also indicated similar results on 16S ribosomal sequences, SHAPE corrected predictions were able to achieve 75% accuracy compared to 41.1% accuracy achieved by the uncorrected program (Sukosd et al. 2013).

Thesis Overview

Importance of RNA

RNA possesses diverse functions across cellular processes. RNA's three main functional roles include: translating protein-coding genes (Crick 1970), catalyzing biological reactions (Kruger et al. 1982) and regulating gene expression (Hartzog and Martens 2009, Guil and Esteller 2012). In certain cases, dysfunctional ncRNA in humans can cause disease (Taft et al. 2010) reflecting ncRNA's functional and biological functional significance. The ncRNA's RNA structure is often closely associated with its function (Gruber et al. 2010), and resolving RNA structure's solution is a critical first step to elucidating ncRNA's biological functions and mechanisms (Gruber et al. 2010, Qu and Adelson 2012, Svoboda and Di Cara 2006, Han et al. 2006).

Importance of HIV

With more than 34 million infected individuals, HIV poses a major epidemiological and economic burden worldwide (UNAIDS 2012). Although a handful of antiretroviral drugs have been developed, there are currently no effective HIV vaccines (Thompson et al. 2012). HIV vaccine development is hindered by HIV's extraordinary diversity and particularly by the rapid alteration of its genetic composition to evade immune pressure (Kawashima et al. 2009, Wain et al. 2007, Kearney et al. 2009). HIV's diversity is contributed to by its high replication error (Mansky and Temin 1995), frequent recombination events (Zhuang et al. 2002) and a high turnover rate (Ho 1997). Provided that immune clearance of HIV is nearly impossible, HIV genetic diversity has persistently expanded since the early 1900s (estimated time of origin for the HIV-1 M group) (Korber et al. 2000, Worobey et al. 2008), making the emergence of new HIV strands and drug resistance mutations an ever-present risk (Ndung'u and Weiss 2012). Therefore,

the tracking of existing HIV genotypes is critical to the global and regional surveillance of the HIV epidemic (Bennett 2005, Hu et al. 1996).

Importance of HIV RNA Structure

HIV contains a condensed single stranded RNA genome packed with multiple protein coding genes. Given HIV's short RNA genome, RNA structures embedded inside the virus's genome are essential to the HIV's life cycle (Greene 2005). RNA structure in the LTR region have been extensively studied (Gardner et al. 2009); however, RNA structure within the protein coding region with weak evolution conservation are rarely studied (Knoepfel and Berkhout 2013). Recent work by Watt et al. has resolved the RNA structure for NLM4-3, an old B subtype laboratory strand (Watts et al. 2009), revealing a complex RNA structural organization embedded between protein coding genes (Watts et al. 2009). These RNA structure's functions include but are not limited to directing recombination (Moumen et al. 2003, Galetto and Negroni 2005) and regulating gene expression (Buratti and Baralle 2004, Abbink and Berkhout 2008, Jacks, Power, et al. 1988).

Multiple recombination mechanisms have been previously proposed. These comprise of a forced copy choice of reverse transcription (Galetto and Negroni 2005), reverse transcriptase (RT) pausing directed recombination (Galetto and Negroni 2005), and an RNA structure directed recombination (Galetto and Negroni 2005). RNA structure directed recombination is hypothesized to induce RT pausing during reverse transcription (Harrison et al. 1998) allowing RT to jump between HIV strands. Another model for RNA structure recombination is a donor/acceptor interaction model (Moumen et al. 2003). Recombination could enable HIV to explore replication advantageous mutations (Burke 1997), so RNA structures throughout the

genome could potentially confer a selective advantage to enable the diversification of its genetic composition (Burke 1997, Sanjuan and Borderia 2011).

RNA structure can also perform functional regulation of HIV gene expression (Buratti and Baralle 2004, Abbink and Berkhout 2008, Jacks, Power, et al. 1988). Ribosomal frameshift element (FE) represents one of the main mechanisms for RNA structure to regulate gene expression (Jacks, Power, et al. 1988). Through the ribosomal frameshift element, the RNA structure regulates the ratio of Gag to Gag-Pol protein expression; these proteins are important for infectious virus particle assembly (Park and Morrow 1991). Proposed mechanisms for frameshifting consist of ribosomal pausing due to the FE RNA structure with -1 nucleotide slipping on the slippery sequence (Jacks, Power, et al. 1988). The FE RNA structure's base stem stability has been suspected to be a determining factor for frameshift efficiency (Mouzakis et al. 2013), but this remains to be further validated.

Current Problems and Difficulties in HIV RNA Structure Study

Multiple problems relating to HIV diversity and RNA structure modeling hamper the success of our HIV RNA structure study. These problems can be divided into two categories, the first is the underestimation of HIV diversity, and the second is the ill-conditioning of HIV RNA structure modeling. To accurately examine the relationship between RNA structure and HIV diversity, these problems must first be resolved. The underestimation of HIV diversity and the ill-conditioning of HIV RNA structure prediction can be further separated into five sub-problems: (1) An underestimation of gp120 HIV sequence diversity. Due to the hypervariable nature of the gp120 v-loop region, these regions are often removed during phylogenetic analyses resulting in the underestimation of HIV diversity. To adequately estimate the HIV diversity will require a method to preserve the variable loop architecture during sequence alignment. (2) An

inconsistent genotyping quality causes the underestimation of HIV sequence diversity. To have an accurate HIV subtype annotation is integral for proper diversity estimation. HIV LANL sequence database subtype annotation is encompassed with erroneous annotated genotype (Zhang et al. 2010); therefore, efforts are necessary to correct these erroneous subtype annotations. (3) RNA's ability to form multiple alternative structures. RNA structure prediction is generally based on the lowest free energy state assumption, but in reality, RNA structure can reside in multiple suboptimal states (Weeks 2010, Washietl, Will, et al. 2012). To accurately model RNA structure-function, a measure that captures the RNA structure folding space will be necessary. Understanding the RNA structure space can potentially reveal additional functional mechanisms relating to HIV RNA structure. (4) Structure prediction programs have been poorly tailored for HIV RNA structure. Existing RNA structure programs were generally trained on noncoding RNA from prokaryotes and eukaryotes. Therefore, reassessment of commonly used RNA structure prediction programs on HIV is necessary. (5) NLM4-3 RNA structure's overextended reference to other HIV strains. RNA structure for the B subtype NLM4-3 has been recently solved by the SHAPE technology (Watts et al. 2009). This reference has been applied to other HIV studies (Snoeck et al. 2011, Sanjuan and Borderia 2011). Since NLM4-3 is a chimeric HIV laboratory strain generated during the 1980s (Barre-Sinoussi et al. 2004, Benn et al. 1985, Adachi et al. 1986), we are uncertain of whether the NLM4-3 RNA structure can be appropriately extrapolated to other contemporary and non-B subtype sequences.

In this dissertation, we aim to examine RNA structure's relationship with replication and recombination and their contribution to HIV diversity and the HIV epidemic. In addition to the five sub-problems described previously, this dissertation will further examine the following questions: (6) What is the mechanistic impact of RNA structure on recombination?

Recombination represents one major mechanism that allows HIV to gain its tremendous diversity. HIV genetic diversity accumulated from recombination events enable the virus to explore a broader sequence space (Burke 1997). As more circulating recombinant forms are being identified (Robertson et al. 2000), understanding recombination mechanisms can eventually lead to a more accurately characterized HIV diversity. Although multiple mechanisms have been described for recombination, question remains regarding which of the mechanisms is driving the recombination event. As indicated previously, RNA structure directed recombination could either induce RT pausing or result in donor/acceptor interaction; however, which of the two mechanisms is driving the recombination event is still relatively unclear. Finally, this dissertation will also attempt to answer (7) What is RNA structure's impact on frameshift efficiency and its further impact on the HIV epidemic? The frameshift element represents an important replication mechanism enabling HIV to gain its diversity. Since frameshift RNA structure is capable of altering gene expression between Gag and Pol (Parkin, Chamorro, and Varmus 1992), growing interest has emerged to use frameshift element as a drug target (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009, Hung et al. 1998, Dinman, Ruiz-Echevarria, and Peltz 1998). Recent studies have shown RNA structure base stem stability is a determining factor for frameshift efficiency (Mouzakis et al. 2013). Given HIV's high genetic diversity, each subtype will potentially possess varied RNA stem stability resulting in varied levels of frameshift efficiency (Chang et al. 1999). By analyzing the variation of frameshift element RNA structure across HIV subtype can potentially allow us to infer each subtype's frameshift efficiency and their impact on virus pathogenicity and replication fitness.

[Chapter 2] Solution to Problem 1: An underestimation of gp120 HIV sequence diversity

Due to the hypervariable nature of the variable loop region, HIV variable loop region is often removed during phylogenetic analyses resulting in the underestimation of HIV diversity. N-linked glycosylation in the variable region plays an essential role for the HIV virus to induce conformational changes that diminish binding of gp120-specific antibodies (Si, Cayabyab, and Sodroski 2001). The placement of N-linked glycosylation sites represents an important evolutionary mechanism adopted by HIV-1 to generate its extraordinary sequence diversity (Zhang et al. 2004). We introduce a greedy algorithm capable of refining HIV's variable loop region alignment through N-linked glycosylation, preventing the removal of the variable loop region. Given an unaligned HIV sequence, the developed pipeline is also capable of performing multiple sequence alignment. Tracking the N-linked glycosylation patterns allow us to understand the changing context of antigenic structures and transmission mechanisms (Shaw and Zhang 2013). The tool can further be used to assess the distribution of N-linked glycosylation sites. As an example, figure 1.6 shows the improved alignment with the green N-linked glycosylation sites aligned together. The tool is available for users as a webserver at http://hivtools.publichealth.uga.edu/N-Glyco/.

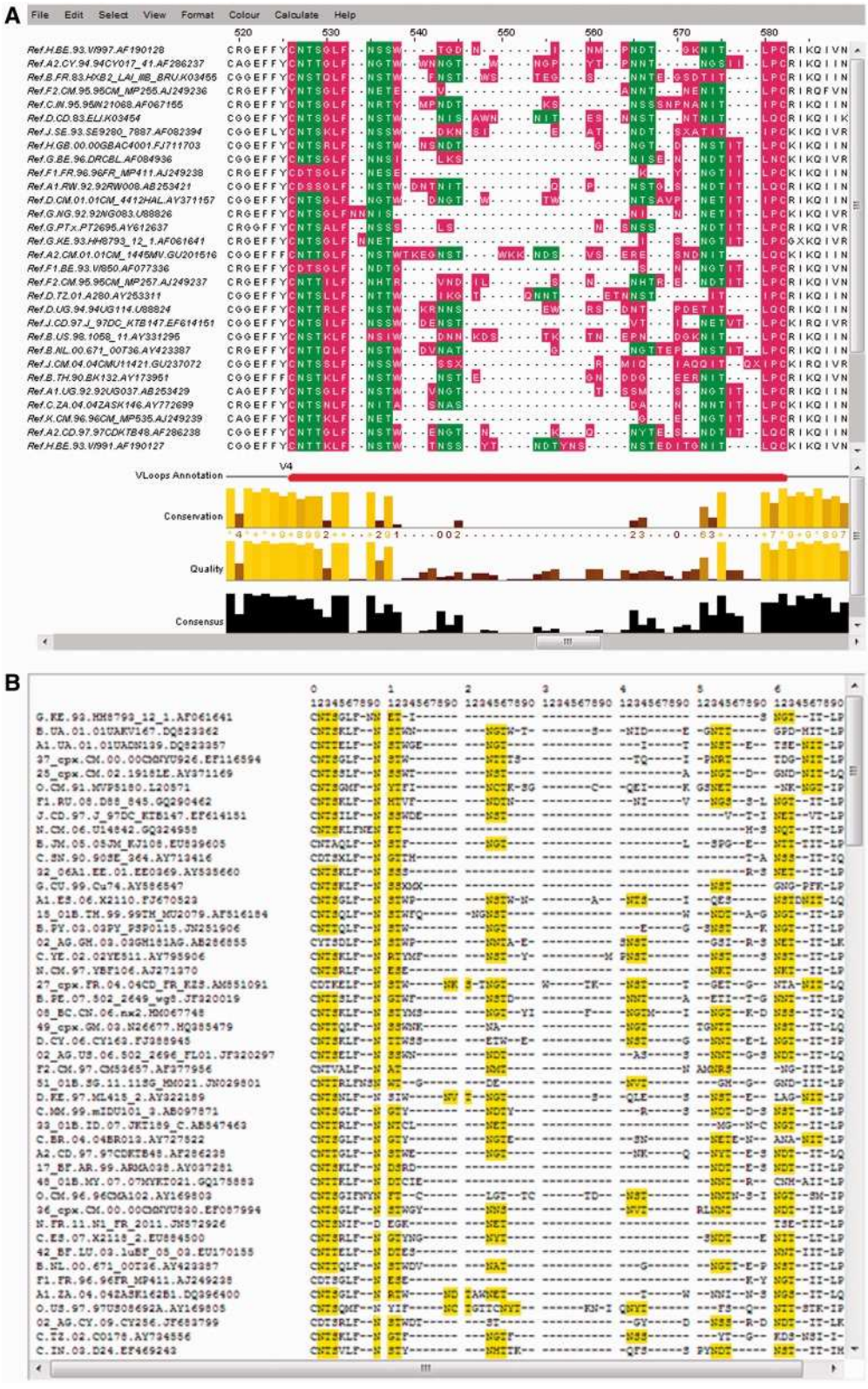


Figure 1.6. An example output of N-linked Glycosylation Site Alignment program.

(A) A Jalview-based alignment editor depicts an optimized alignment using N-linked glycosylation sites as alignment anchors. From the V loop Alignment program, V loop regions are highlighted as pink and each N-linked glycosylation sites are highlighted in green. An additional V loop annotation track is added underneath the alignment. (B) An HTML view of the optimized alignment with the N-linked glycosylation sites are highlighted in yellow.

[Chapter 3] Solution to Problem 2: Inconsistent genotyping quality causes the underestimation of HIV sequence diversity

Accurate subtype annotation of HIV sequences is integral for proper HIV diversity estimation. Previous analysis on the HIV LANL sequence database indicated the presence of erroneous annotated genotype (Zhang et al. 2010). Improper subtyping could potentially lead to an incorrect estimation of HIV diversity. To properly capture the RNA structure's impact on HIV subtypes, a comprehensive re-subtyping effort is necessary. To ensure our dataset is properly genotyped, we constructed a pipeline to re-annotate the subtype and its recombinant breakpoint for the entire HIV database of 460,000 sequences. Examples of genotyping errors for AF504640, AY586546 and FJ388956 are shown in figure 1.7. In the process, we were able to identify the existence of 5% genotyping errors across geographic regions and risk factors. Results from this genotyping assessment will benefit future subtype specific investigation of RNA structures. In addition, the result is also of epidemiological importance for retrospective genotyping research within the HIV community.

	Original Assignment From GenBank/Literature	Validated Assignment From This Study
AF504640	G	B
AY586546	BG	BDG
FJ388956	B	A1B

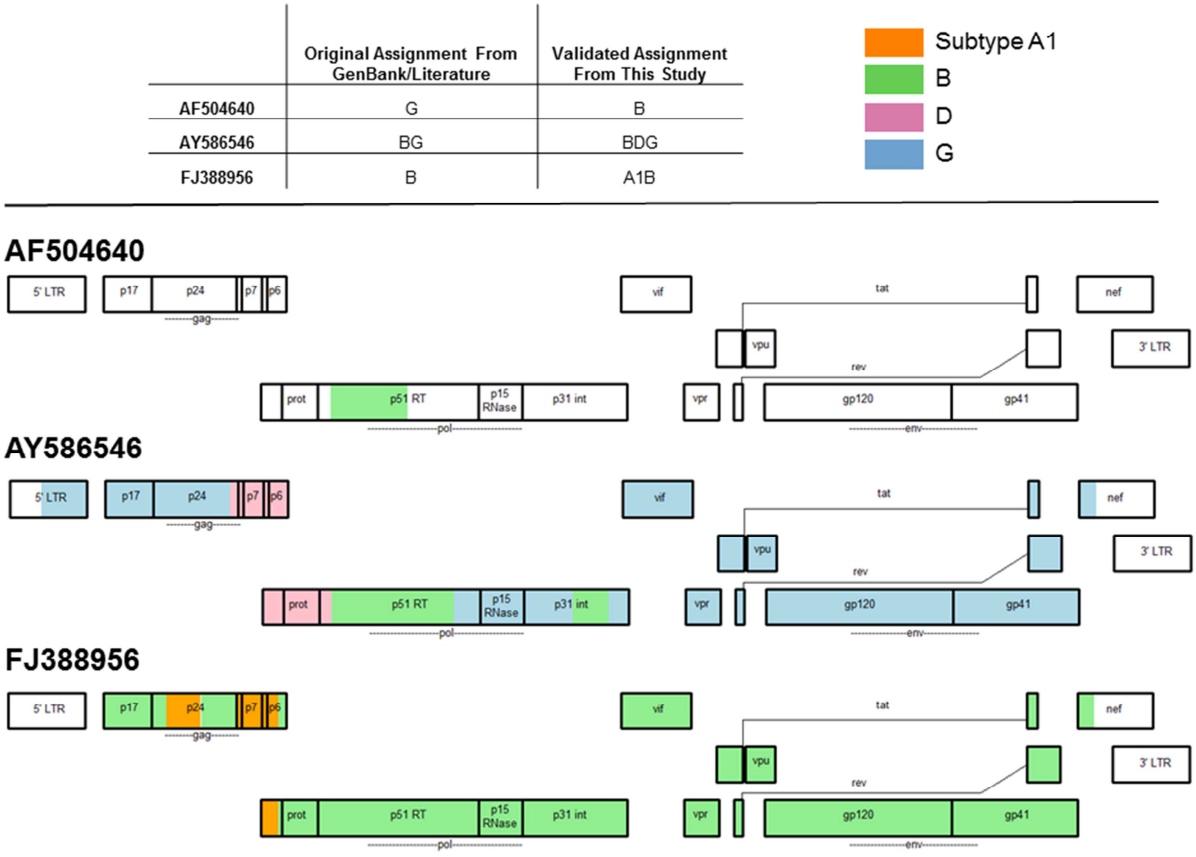


Figure 1.7. HIV-1 mis-genotyped cases in full-length genomes and fragment sequences.

Mis-genotyping is defined as a scenario in which a sequence's original genotyping assignment by GenBank/publication is inconsistent with results of jpHMM and the phylogenetic analyses, as described in the Method section. AF504640 is an incomplete genome example in which the original G assignment was found to be of subtype B. In full-length genomes AY586546 and FJ388956, our results identified more complex genomic composition than their original subtype assignment; extra parental subtype was found in each sequence.

[Chapter 4] Solution to Problem 3: RNA structure's ability to form multiple alternative structures.

An often neglected fact regarding RNA structural folding is the presence of suboptimal structures. Therefore, RNA structure measures that are capable of evaluating the diversity of the RNA structural folding can potentially reveal new insight to RNA structural research especially for HIV RNA structural studies. In this chapter, a novel Shannon information entropy approach is developed called "Unpaired Structural Entropy" (USE). The method utilizes the unpaired probability for each nucleotide and calculates its distribution based on Shannon Entropy. The USE measure is found to be a better indicator for the certainty of the RNA structure fold than positional entropy described by Hyunen et al (Huynen, Gutell, and Konings 1997). This method was initially developed to be provided for ab initio noncoding RNA prediction, and the USE measure was used to assess multiple noncoding RNAs finding. The USE measure was found to be capable of distinguishing precursor-microRNAs from the genomic background (Figure 1.8). In later chapters, the USE measure is further applied to assess HIV frameshift element's RNA structural fold stability in chapter 6.

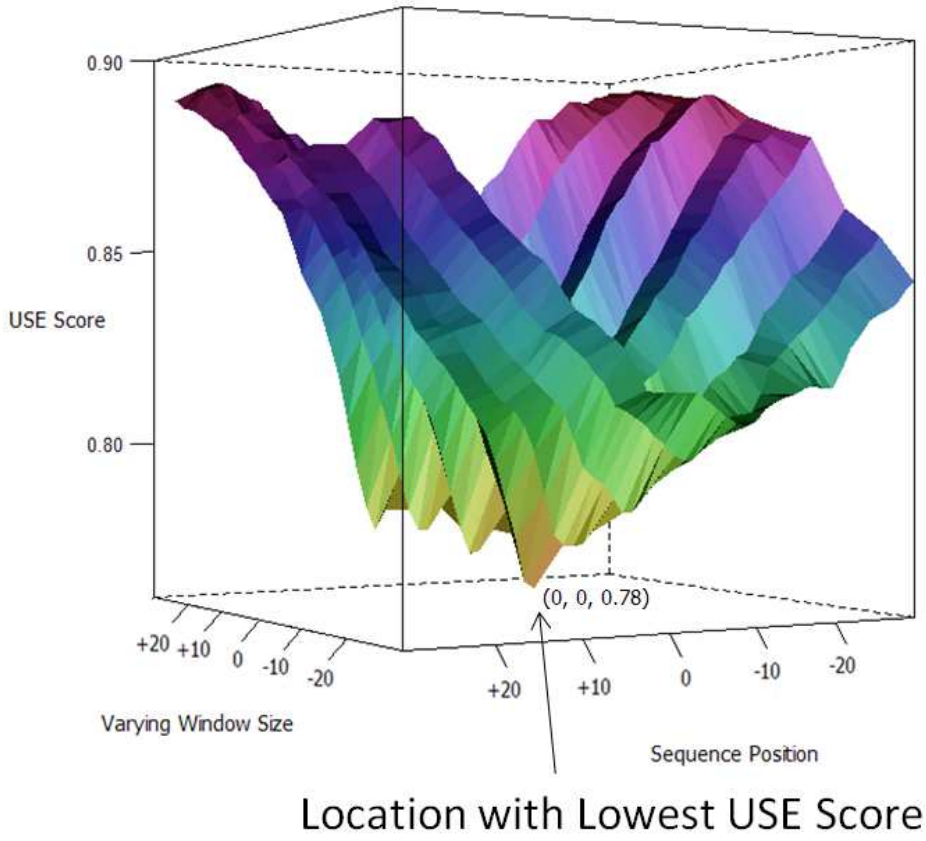


Figure 1.8. Average USE value of all miRNAs.

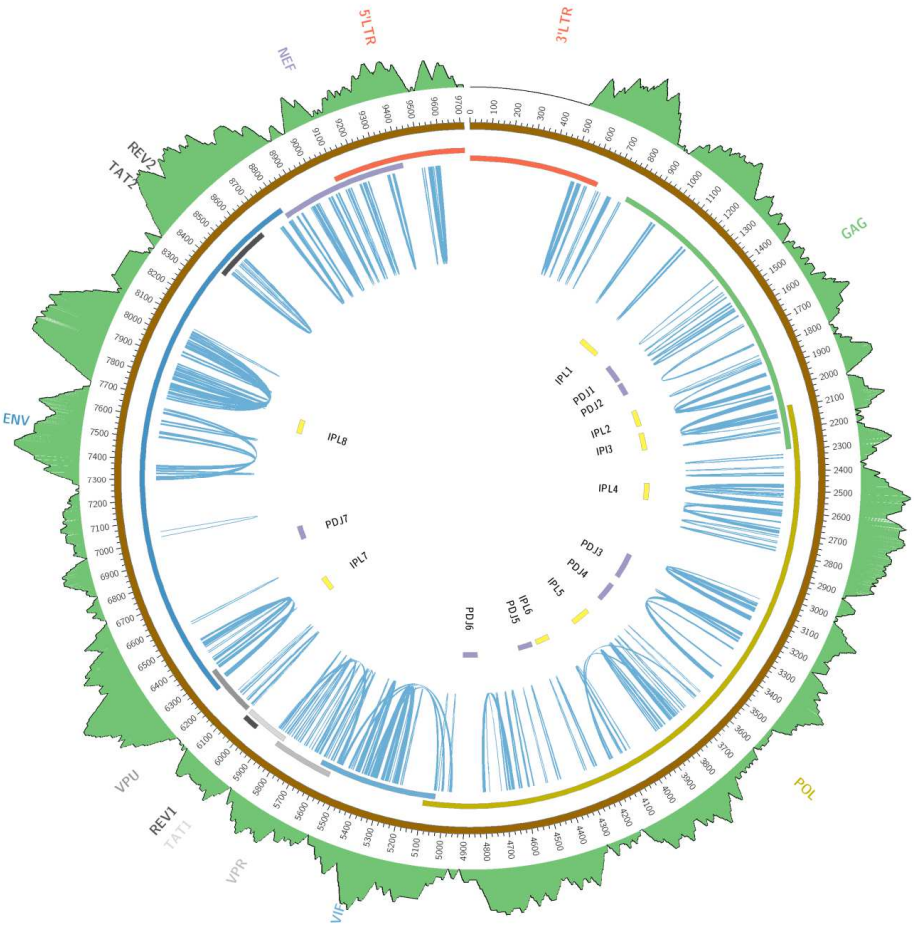
Any point on the graph corresponded to a specific window length and position and represents the USE score averaged across all 721 Human microRNA sequences. The labels on the sequence-position axis represented the relative upstream/downstream position from the location of the actual microRNA. The window-size axis represented the amount of increments/decrements of the window-length relative to the length of the actual microRNA. The point indicated by the low USE value indicates the microRNA size and location within the genome.

[Chapter 5] Solution to Problem 4 and 5: Structure prediction programs have been poorly tailored for HIV RNA structure, and NLM4-3 RNA structure's overextended reference to other HIV strains.

Chapter 5 aims to resolve the two existing problems facing HIV RNA structure modeling. As RNA structure modeling programs are generally trained using noncoding RNA structure from prokaryotes or eukaryotes (non-virus), the accuracy of existing RNA structure models' application in HIV is questionable. Therefore, an HIV specific examination of commonly used RNA structure predictions program is necessary. RNA fold methods are categorized into thermodynamic based methods and non-thermodynamic based methods. We compiled all existing HIV RNA structures from the Rfam RNA structure database and used it to assess the accuracy of the individual RNA structural fold methods. We found that CONTRAFold, a non-thermodynamic based method, was capable of achieving the best performance even without a SHAPE reference. By incorporating NLM4-3 SHAPE as auxiliary information, the CYK, a non-thermodynamic based method, was found to perform better than RNAstructure, a thermodynamic based method**.** Collectively, the HIV RNA structure prediction results indicate a slight advantage for non-thermodynamic method over thermodynamic method.

The other problem that challenges existing HIV RNA structure study is the overextended application of NLM4-3 strain as a RNA structural reference. Multiple studies are performed under the assumption that NLM4-3 RNA structure is conserved across HIV sequences (Snoeck et al. 2011, Sanjuan and Borderia 2011); however NLM4-3 is an old chimeric strand artificially generated during the 1980s (Barre-Sinoussi et al. 2004, Benn et al. 1985, Adachi et al. 1986) raising doubt whether the NLM4-3 RNA structure can be extended to other HIV sequences. We compared RNA secondary structure base pairing conservation between NLM4-3 and other HIV subtypes and found significant RNA structure variation when compared to non-B subtype sequences. Figure 1.9 present an example RNA structure comparison between B subtype and C subtype indicating base pairing variation for in multiple genomic regions. We also further

compared the RNA structure within the B subtype lineage, particularly between pure B and B' subtype from China; the B' subtype is genetically divergent from other B subtype (Kalish et al. 1995, Li, Uenishi, et al. 2010). Within our B subtype comparison, we found a region with high base pairing conservation in gag and pol, but significant base pairing difference was found for genomic region in VPR-VPU and NEF/LTR. In this chapter, we also constructed a RNA structure prediction pipeline by predicting RNA secondary structure fold through an interpolated SHAPE reactivity with further RNA structure refinement via CONTRAFold (Do, Woods, and Batzoglou 2006).



1

Figure 1.9A.

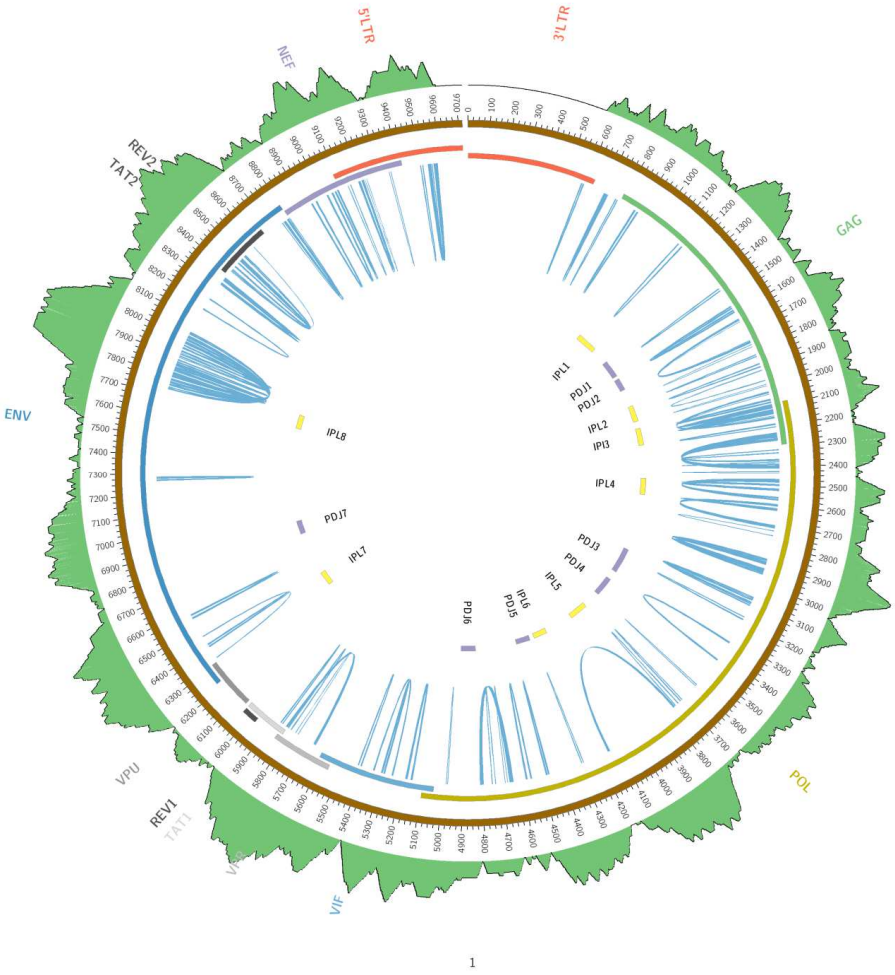


Figure 1.9B.

Figure 1.9. Circos plot comparing NLM4-3 to HIV B and C subtype sequence.

(A) RNA secondary structure for a B subtype HIV sequence AY423387 from 2000 Netherland. (B) RNA secondary structure for C subtype HIV sequence AY772699 from 2004. The blue curves indicate base pairs that are conserved between NLM4-3 and the query sequence. The mountain plot on the outer ring represents the base pairing depth for secondary structures in either B or C

subtype sequences. Compared to NLM4-3, the B subtype sequence has more conserved stems than C subtype sequences.

[Chapter 6] Solution to Problem 6 and 7: What is the mechanistic impact of RNA structure on recombination? What are RNA structure's impact on frameshift efficiency and its further impact on the HIV epidemic?

Considering ribosomal frameshift event's importance to HIV replication (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009), there is vested interest to using ribosomal frameshifting as a drug target (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009, Hung et al. 1998, Dinman, Ruiz-Echevarria, and Peltz 1998). Previous mutagenesis studies have indicated RNA secondary structure stability can contribute to RNA frameshift efficiency, impacting virus replication (Dulude et al. 2006, Hung et al. 1998, Shehu-Xhilaga, Crowe, and Mak 2001). Although RNA secondary structure plays a key role in frameshift efficiency, there has yet been an in depth evaluation of frameshift element's RNA structure across HIV-1 subtypes and their recombinants. Since the frameshift element RNA structure is highly conserved across all HIV subtypes, we examined the recombinant patterns surrounding the frameshift element. We found that two subtypes tend to recombine at the frameshift element region only if the two subtype's frameshift elements share the same sequence space. In addition, we also noticed that recombinant breakpoints tend to occur upstream and downstream of the frameshift element region indicating less contribution by the donor/acceptor model and a more suspected contribution by the RT pausing model. However, as more of the recombinations that are observed indicate high sequence similarity between the two recombinants. We believe the main mechanism driving recombination is perhaps the functional viability of the resulting HIV sequence.

In addition to examining recombination, we performed a cross-clade examination of the frameshift element's RNA structure and constructed a regression model capable of predicting frameshift efficiency for all the sequences in the HIV database. We stratified our analysis of frameshift efficiency based on HIV subtypes, risk factors, geographical regions, and sampling year. Our analysis indicates that the B subtype possesses higher frameshift efficiency (Figure 1.10). B subtype is known for its high replication fitness, allowing us to speculate regarding frameshift efficiency's potential impact on replication fitness. We found potential frameshift efficiency altercation within North American B subtype but not in Sub-Saharan Africa C subtypes. Differences in transmission route associated with multiplicity of infection seem to impact variations observed in frameshift efficiency. Our study indicates that RNA structure can be evaluated from an epidemiological perspective, and that the improved understanding of HIV RNA structure can eventually help us better understand and improve our model for HIV sequence diversity.

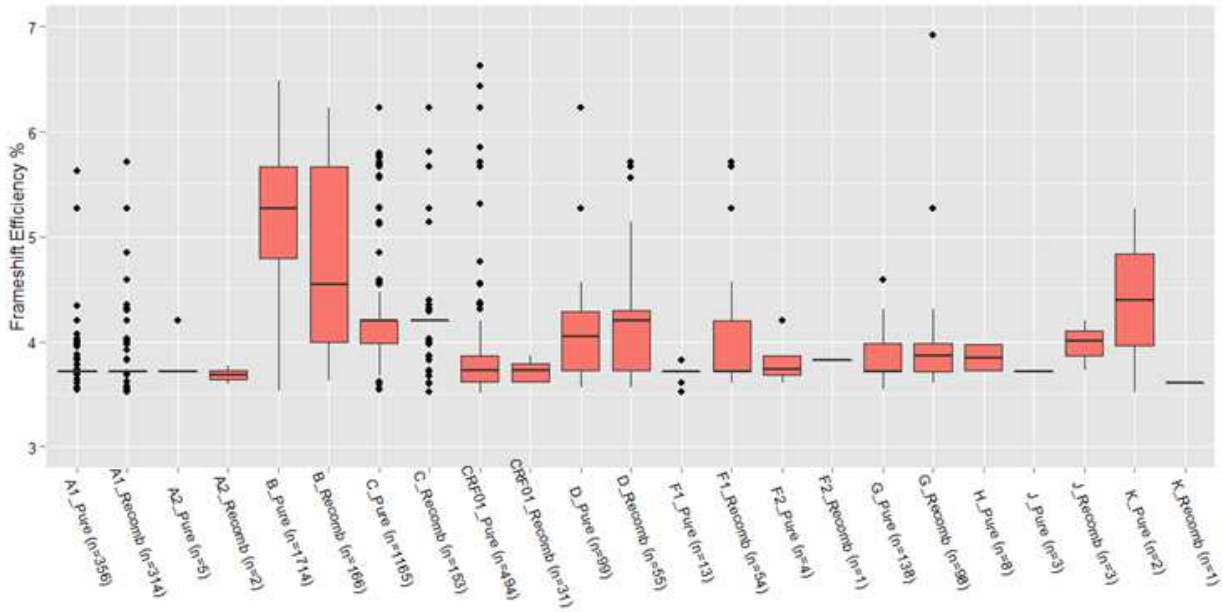


Figure 1.10. Frameshift Efficiency Comparison Across HIV Subtypes.

Boxplot compares the estimated relative frameshift efficiency across different subtypes. Noticeably B subtype and B subtype recombinant possess the highest frameshift efficiency than other subtypes. After B subtype, C subtype was indicated to be the second most efficient group of subtype with D subtype as the third most efficient. Most of the other subtypes range within 3.5-4.0% of relative frameshift efficiency. For the vast majority of cases, the frameshift efficiency for the recombinants was very similar to that of the pure subtypes.

References

Abbink, T. E., and B. Berkhout. 2008. "RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor." *J Virol* 82 (6):3090-8. doi: 10.1128/JVI.01479-07.

Abecasis, A. B., A. M. Vandamme, and P. Lemey. 2009. "Quantifying differences in the tempo of human immunodeficiency virus type 1 subtype evolution." *J Virol* 83 (24):12917-24. doi: 10.1128/JVI.01022-09.

Adachi, A., H. E. Gendelman, S. Koenig, T. Folks, R. Willey, A. Rabson, and M. A. Martin. 1986. "Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone." *J Virol* 59 (2):284-91.

Ameres, S. L., and P. D. Zamore. 2013. "Diversifying microRNA sequence and function." *Nat Rev Mol Cell Biol* 14 (8):475-88. doi: 10.1038/nrm3611.

Anfinsen, C. B. 1973. "Principles that govern the folding of protein chains." *Science* 181 (4096):223-30.

Arrildt, K. T., S. B. Joseph, and R. Swanstrom. 2012. "The HIV-1 env protein: a coat of many colors." *Curr HIV/AIDS Rep* 9 (1):52-63. doi: 10.1007/s11904-011-0107-3.

Bannwarth, S., and A. Gatignol. 2005. "HIV-1 TAR RNA: the target of molecular interactions between the virus and its host." *Curr HIV Res* 3 (1):61-71.

Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. 1983. "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." *Science* 220 (4599):868-71.

Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. 2004. "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." *Revista De Investigacion Clinica* 56 (2):126-129.

Been, M. D., and G. S. Wickham. 1997. "Self-cleaving ribozymes of hepatitis delta virus RNA." *Eur J Biochem* 247 (3):741-53.

Beerenwinkel, N., B. Schmidt, H. Walter, R. Kaiser, T. Lengauer, D. Hoffmann, K. Korn, and J. Selbig. 2002. "Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype." *Proc Natl Acad Sci U S A* 99 (12):8271-6. doi: 10.1073/pnas.112177799.

Benn, S., R. Rutledge, T. Folks, J. Gold, L. Baker, J. McCormick, P. Feorino, P. Piot, T. Quinn, and M. Martin. 1985. "Genomic heterogeneity of AIDS retroviral isolates from North America and Zaire." *Science* 230 (4728):949-51.

Bennett, D. 2005. "HIV [corrected] genetic diversity surveillance in the United States." *J Infect Dis* 192 (1):4-9. doi: 10.1086/430329.

Bernhart, S. H., I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. 2008. "RNAalifold: improved consensus structure prediction for RNA alignments." *BMC Bioinformatics* 9:474. doi: 1471-2105-9-474 [pii] 10.1186/1471-2105-9-474.

Bernhart, S. H., H. Tafer, U. Muckstein, C. Flamm, P. F. Stadler, and I. L. Hofacker. 2006. "Partition function and base pairing probabilities of RNA heterodimers." *Algorithms Mol Biol* 1 (1):3. doi: 1748-7188-1-3 [pii] 10.1186/1748-7188-1-3.

Brakier-Gingras, L., J. Charbonneau, and S. E. Butcher. 2012. "Targeting frameshifting in the human immunodeficiency virus." *Expert Opin Ther Targets* 16 (3):249-58. doi: 10.1517/14728222.2012.665879.

Breaker, R. R. 2008. "Complex riboswitches." *Science* 319 (5871):1795-7. doi: 10.1126/science.1152621.

Bukrinsky, M., and A. Adzhubei. 1999. "Viral protein R of HIV-1." *Rev Med Virol* 9 (1):39-49.

Buonaguro, L., M. L. Tornesello, and F. M. Buonaguro. 2007. "Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications." *J Virol* 81 (19):10209-19. doi: 10.1128/JVI.00872-07.

Buratti, E., and F. E. Baralle. 2004. "Influence of RNA secondary structure on the pre-mRNA splicing process." *Mol Cell Biol* 24 (24):10505-14. doi: 10.1128/MCB.24.24.10505-10514.2004.

Burke, D. S. 1997. "Recombination in HIV: an important viral evolutionary strategy." *Emerg Infect Dis* 3 (3):253-9. doi: 10.3201/eid0303.970301.

Chakrabarti, L. A., and V. Simon. 2010. "Immune mechanisms of HIV control." *Curr Opin Immunol* 22 (4):488-96. doi: 10.1016/j.coi.2010.06.006.

Chamorro, M., N. Parkin, and H. E. Varmus. 1992. "An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA." *Proc Natl Acad Sci U S A* 89 (2):713-7.

Chang, S. Y., R. Sutthent, P. Auewarakul, C. Apichartpiyakul, M. Essex, and T. H. Lee. 1999. "Differential stability of the mRNA secondary structures in the frameshift site of various HIV type 1 viruses." *AIDS Res Hum Retroviruses* 15 (17):1591-6. doi: 10.1089/088922299309892.

Chen, S. J., and K. A. Dill. 2000. "RNA folding energy landscapes." *Proc Natl Acad Sci U S A* 97 (2):646-51.

Cheng-Mayer, C., D. Seto, M. Tateno, and J. A. Levy. 1988. "Biologic features of HIV-1 that correlate with virulence in the host." *Science* 240 (4848):80-2.

Clever, J., C. Sassetti, and T. G. Parslow. 1995. "RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type 1." *J Virol* 69 (4):2101-9.

Coffin, J. M. 1979. "Structure, replication, and recombination of retrovirus genomes: some unifying hypotheses." *J Gen Virol* 42 (1):1-26.

Crick, F. 1970. "Central dogma of molecular biology." *Nature* 227 (5258):561-3.

Damgaard, C. K., E. S. Andersen, B. Knudsen, J. Gorodkin, and J. Kjems. 2004. "RNA interactions in the 5' region of the HIV-1 genome." *J Mol Biol* 336 (2):369-79.

Damond, F., M. Worobey, P. Campa, I. Farfara, G. Colin, S. Matheron, F. Brun-Vezinet, D. L. Robertson, and F. Simon. 2004. "Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification." *AIDS Res Hum Retroviruses* 20 (6):666-72. doi: 10.1089/0889222041217392.

Das, S. R., and S. Jameel. 2005. "Biology of the HIV Nef protein." *Indian J Med Res* 121 (4):315-32.

Deeks, S. G., and B. D. Walker. 2007. "Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy." *Immunity* 27 (3):406-16. doi: 10.1016/j.immuni.2007.08.010.

Deigan, K. E., T. W. Li, D. H. Mathews, and K. M. Weeks. 2009. "Accurate SHAPE-directed RNA structure determination." *Proc Natl Acad Sci U S A* 106 (1):97-102. doi: 0806929106 [pii] 10.1073/pnas.0806929106.

DeStefano, J. J., R. A. Bambara, and P. J. Fay. 1994. "The mechanism of human immunodeficiency virus reverse transcriptase-catalyzed strand transfer from internal regions of heteropolymeric RNA templates." *J Biol Chem* 269 (1):161-8.

DeStefano, J. J., L. M. Mallaber, L. Rodriguez-Rodriguez, P. J. Fay, and R. A. Bambara. 1992. "Requirements for strand transfer between internal regions of heteropolymer templates by human immunodeficiency virus reverse transcriptase." *J Virol* 66 (11):6370-8.

Dinman, J. D. 2012. "Mechanisms and implications of programmed translational frameshifting." *Wiley Interdiscip Rev RNA* 3 (5):661-73. doi: 10.1002/wrna.1126.

Dinman, J. D., M. J. Ruiz-Echevarria, and S. W. Peltz. 1998. "Translating old drugs into new treatments: ribosomal frameshifting as a target for antiviral agents." *Trends Biotechnol* 16 (4):190-6.

Dirks, R. M., and N. A. Pierce. 2003. "A partition function algorithm for nucleic acid secondary structure including pseudoknots." *J Comput Chem* 24 (13):1664-77. doi: 10.1002/jcc.10296.

Dirks, R. M., and N. A. Pierce. 2004. "An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots." *J Comput Chem* 25 (10):1295-304. doi: 10.1002/jcc.20057.

Do, C. B., D. A. Woods, and S. Batzoglou. 2006. "CONTRAfold: RNA secondary structure prediction without physics-based models." *Bioinformatics* 22 (14):e90-8. doi: 22/14/e90 [pii] 10.1093/bioinformatics/btl246.

Dowell, R. D., and S. R. Eddy. 2004. "Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction." *BMC Bioinformatics* 5:71. doi: 10.1186/1471-2105-5-71.

Dulude, D., Y. A. Berchiche, K. Gendron, L. Brakier-Gingras, and N. Heveker. 2006. "Decreasing the frameshift efficiency translates into an equivalent reduction of the replication of the human immunodeficiency virus type 1." *Virology* 345 (1):127-36. doi: 10.1016/j.virol.2005.08.048.

Durbin, R., Eddy,S., Krogh,A. and Mitchison,G. 1998a. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge, UK: Cambridge University Press.

Eddy, S. R. 2004. "How do RNA folding algorithms work?" *Nat Biotechnol* 22 (11):1457-8. doi: 10.1038/nbt1104-1457.

Ehresmann, C., F. Baudin, M. Mougel, P. Romby, J. P. Ebel, and B. Ehresmann. 1987. "Probing the structure of RNAs in solution." *Nucleic Acids Res* 15 (22):9109-28.

Ensoli, B., L. Buonaguro, G. Barillari, V. Fiorelli, R. Gendelman, R. A. Morgan, P. Wingfield, and R. C. Gallo. 1993. "Release, uptake, and effects of extracellular human

immunodeficiency virus type 1 Tat protein on cell growth and viral transactivation." *J Virol* 67 (1):277-87.

Farabaugh, P. J. 1996. "Programmed translational frameshifting." *Microbiol Rev* 60 (1):103-34.

Francia, S., F. Michelini, A. Saxena, D. Tang, M. de Hoon, V. Anelli, M. Mione, P. Carninci, and F. d'Adda di Fagagna. 2012. "Site-specific DICER and DROSHA RNA products control the DNA-damage response." *Nature* 488 (7410):231-5. doi: 10.1038/nature11179.

Frankel, A. D., and J. A. Young. 1998. "HIV-1: fifteen proteins and an RNA." *Annu Rev Biochem* 67:1-25. doi: 10.1146/annurev.biochem.67.1.1.

Freed, E. O. 1998. "HIV-1 gag proteins: diverse functions in the virus life cycle." *Virology* 251 (1):1-15. doi: 10.1006/viro.1998.9398.

Galetto, R., and M. Negroni. 2005. "Mechanistic features of recombination in HIV." *AIDS Rev* 7 (2):92-102.

Gao, F., E. Bailes, D. L. Robertson, Y. Chen, C. M. Rodenburg, S. F. Michael, L. B. Cummins, L. O. Arthur, M. Peeters, G. M. Shaw, P. M. Sharp, and B. H. Hahn. 1999. "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes." *Nature* 397 (6718):436-41. doi: 10.1038/17130.

Gao, F., N. Vidal, Y. Li, S. A. Trask, Y. Chen, L. G. Kostrikis, D. D. Ho, J. Kim, M. D. Oh, K. Choe, M. Salminen, D. L. Robertson, G. M. Shaw, B. H. Hahn, and M. Peeters. 2001. "Evidence of two distinct subsubtypes within the HIV-1 subtype A radiation." *AIDS Res Hum Retroviruses* 17 (8):675-88. doi: 10.1089/088922201750236951.

Gardner, P. P., J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. 2009. "Rfam: updates to the

RNA families database." *Nucleic Acids Res* 37 (Database issue):D136-40. doi: gkn766 [pii] 10.1093/nar/gkn766.

Gareiss, P. C., and B. L. Miller. 2009. "Ribosomal frameshifting: an emerging drug target for HIV." *Curr Opin Investig Drugs* 10 (2):121-8.

Gee, A. H., W. Kasprzak, and B. A. Shapiro. 2006. "Structural differentiation of the HIV-1 polyA signals." *J Biomol Struct Dyn* 23 (4):417-28. doi: 10.1080/07391102.2006.10531236.

Gheysen, D., E. Jacobs, F. de Foresta, C. Thiriart, M. Francotte, D. Thines, and M. De Wilde. 1989. "Assembly and release of HIV-1 precursor Pr55gag virus-like particles from recombinant baculovirus-infected insect cells." *Cell* 59 (1):103-12.

Giedroc, D. P., C. A. Theimer, and P. L. Nixon. 2000. "Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting." *J Mol Biol* 298 (2):167-85. doi: 10.1006/jmbi.2000.3668.

Girard, A., R. Sachidanandam, G. J. Hannon, and M. A. Carmell. 2006. "A germline-specific class of small RNAs binds mammalian Piwi proteins." *Nature* 442 (7099):199-202. doi: 10.1038/nature04917.

Gonzalez, R. L., Jr. 2008. "Navigating the RNA folding landscape." *Nat Chem Biol* 4 (8):451-2. doi: 10.1038/nchembio0808-451.

Göttlinger, H. 2001. HIV-1 Gag: a molecular machine driving viral particle assembly and release.

Green, L., C. H. Kim, C. Bustamante, and I. Tinoco, Jr. 2008. "Characterization of the mechanical unfolding of RNA pseudoknots." *J Mol Biol* 375 (2):511-28. doi: 10.1016/j.jmb.2007.05.058.

Greene, W. C., Peterlin. 2005. *Molecular insights into HIV biology*. Edited by S. Coffey In L. Peiperl, O. Bacon, and P. Volberding (ed.), *HIV Insite Knowledge Base.* Univ. of California San Francisco and San Francisco General Hospital, San Francisco.

Gruber, A. R., S. Findeiss, S. Washietl, I. L. Hofacker, and P. F. Stadler. 2010. "Rnaz 2.0: Improved Noncoding Rna Detection." *Pac Symp Biocomput* 15:69-79. doi: 9789814295291_0009 [pii].

Guil, S., and M. Esteller. 2012. "Cis-acting noncoding RNAs: friends and foes." *Nat Struct Mol Biol* 19 (11):1068-75. doi: 10.1038/nsmb.2428.

Hajdin, C. E., S. Bellaousov, W. Huggins, C. W. Leonard, D. H. Mathews, and K. M. Weeks. 2013. "Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots." *Proc Natl Acad Sci U S A* 110 (14):5498-503. doi: 10.1073/pnas.1219988110.

Han, J., Y. Lee, K. H. Yeom, J. W. Nam, I. Heo, J. K. Rhee, S. Y. Sohn, Y. Cho, B. T. Zhang, and V. N. Kim. 2006. "Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex." *Cell* 125 (5):887-901. doi: 10.1016/j.cell.2006.03.043.

Harrison, G. P., M. S. Mayo, E. Hunter, and A. M. Lever. 1998. "Pausing of reverse transcriptase on retroviral RNA templates is influenced by secondary structures both 5' and 3' of the catalytic site." *Nucleic Acids Res* 26 (14):3433-42.

Hartzog, G. A., and J. A. Martens. 2009. "ncRNA transcription makes its mark." *EMBO J* 28 (12):1679-80. doi: 10.1038/emboj.2009.136.

Hemelaar, J., E. Gouws, P. D. Ghys, and S. Osmanov. 2006. "Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004." *AIDS* 20 (16):W13-23. doi: 10.1097/01.aids.0000247564.73009.bc 00002030-200610240-00024 [pii].

Hirsch, V. M., R. A. Olmsted, M. Murphey-Corb, R. H. Purcell, and P. R. Johnson. 1989. "An African primate lentivirus (SIVsm) closely related to HIV-2." *Nature* 339 (6223):389-92. doi: 10.1038/339389a0.

Ho, D. D. 1997. "Perspectives series: host/pathogen interactions. Dynamics of HIV-1 replication in vivo." *J Clin Invest* 99 (11):2565-7. doi: 10.1172/JCI119443.

Hobert, O. 2008. "Gene regulation by transcription factors and microRNAs." *Science* 319 (5871):1785-6. doi: 10.1126/science.1151651.

Hofacker, I. L., and P. F. Stadler. 2006. "Memory efficient folding algorithms for circular RNA secondary structures." *Bioinformatics* 22 (10):1172-6. doi: btl023 [pii] 10.1093/bioinformatics/btl023.

Hoffman, B. E., and P. J. Grabowski. 1992. "U1 snRNP targets an essential splicing factor, U2AF65, to the 3' splice site by a network of interactions spanning the exon." *Genes Dev* 6 (12B):2554-68.

Hu, D. J., T. J. Dondero, M. A. Rayfield, J. R. George, G. Schochetman, H. W. Jaffe, C. C. Luo, M. L. Kalish, B. G. Weniger, C. P. Pau, C. A. Schable, and J. W. Curran. 1996. "The emerging genetic diversity of HIV. The importance of global surveillance for diagnostics, research, and prevention." *JAMA* 275 (3):210-6.

Hung, M., P. Patel, S. Davis, and S. R. Green. 1998. "Importance of ribosomal frameshifting for human immunodeficiency virus type 1 particle assembly and replication." *J Virol* 72 (6):4819-24.

Huynen, M., R. Gutell, and D. Konings. 1997. "Assessing the reliability of RNA folding using statistical mechanics." *Journal of Molecular Biology* 267 (5):1104-1112.

Jacks, T., H. D. Madhani, F. R. Masiarz, and H. E. Varmus. 1988. "Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region." *Cell* 55 (3):447-58.

Jacks, T., M. D. Power, F. R. Masiarz, P. A. Luciw, P. J. Barr, and H. E. Varmus. 1988. "Characterization of ribosomal frameshifting in HIV-1 gag-pol expression." *Nature* 331 (6153):280-3. doi: 10.1038/331280a0.

Jacks, T., K. Townsley, H. E. Varmus, and J. Majors. 1987. "Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus gag-related polyproteins." *Proc Natl Acad Sci U S A* 84 (12):4298-302.

Jacks, T., and H. E. Varmus. 1985. "Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting." *Science* 230 (4731):1237-42.

Judson, H.F. 1979. *The eighth day of creation: makers of the revolution in biology*: Simon and Schuster.

Kaikkonen, M. U., M. T. Lam, and C. K. Glass. 2011. "Non-coding RNAs as regulators of gene expression and epigenetics." *Cardiovasc Res* 90 (3):430-40. doi: 10.1093/cvr/cvr097.

Kalish, M. L., A. Baldwin, S. Raktham, C. Wasi, C. C. Luo, G. Schochetman, T. D. Mastro, N. Young, S. Vanichseni, H. Rubsamen-Waigmann, and et al. 1995. "The evolving molecular epidemiology of HIV-1 envelope subtypes in injecting drug users in Bangkok, Thailand: implications for HIV vaccine trials." *AIDS* 9 (8):851-7.

Karn, Jonathan. 2000. Tat, a novel regulator of HIV transcription and latency. *In HIV Sequence Compendium*.

Kawashima, Y., K. Pfafferott, J. Frater, P. Matthews, R. Payne, M. Addo, H. Gatanaga, M. Fujiwara, A. Hachiya, H. Koizumi, N. Kuse, S. Oka, A. Duda, A. Prendergast, H. Crawford, A. Leslie, Z. Brumme, C. Brumme, T. Allen, C. Brander, R. Kaslow, J. Tang,

E. Hunter, S. Allen, J. Mulenga, S. Branch, T. Roach, M. John, S. Mallal, A. Ogwu, R. Shapiro, J. G. Prado, S. Fidler, J. Weber, O. G. Pybus, P. Klenerman, T. Ndung'u, R. Phillips, D. Heckerman, P. R. Harrigan, B. D. Walker, M. Takiguchi, and P. Goulder. 2009. "Adaptation of HIV-1 to human leukocyte antigen class I." *Nature* 458 (7238):641-5. doi: 10.1038/nature07746.

Kearney, M., F. Maldarelli, W. Shao, J. B. Margolick, E. S. Daar, J. W. Mellors, V. Rao, J. M. Coffin, and S. Palmer. 2009. "Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals." *J Virol* 83 (6):2715-27. doi: 10.1128/JVI.01960-08.

Keen, N. J., M. J. Churcher, and J. Karn. 1997. "Transfer of Tat and release of TAR RNA during the activation of the human immunodeficiency virus type-1 transcription elongation complex." *EMBO J* 16 (17):5260-72. doi: 10.1093/emboj/16.17.5260.

Klase, Z., R. Winograd, J. Davis, L. Carpio, R. Hildreth, M. Heydarian, S. Fu, T. McCaffrey, E. Meiri, M. Ayash-Rashkovsky, S. Gilad, Z. Bentwich, and F. Kashanchi. 2009. "HIV-1 TAR miRNA protects against apoptosis by altering cellular gene expression." *Retrovirology* 6:18. doi: 10.1186/1742-4690-6-18.

Klasens, B. I., A. T. Das, and B. Berkhout. 1998. "Inhibition of polyadenylation by stable RNA secondary structure." *Nucleic Acids Res* 26 (8):1870-6.

Klasens, B. I., M. Thiesen, A. Virtanen, and B. Berkhout. 1999. "The ability of the HIV-1 AAUAAA signal to bind polyadenylation factors is controlled by local RNA structure." *Nucleic Acids Res* 27 (2):446-54.

Klimkait, T., K. Strebel, M. D. Hoggan, M. A. Martin, and J. M. Orenstein. 1990. "The human immunodeficiency virus type 1-specific protein vpu is required for efficient virus maturation and release." *J Virol* 64 (2):621-9.

Knoepfel, S. A., and B. Berkhout. 2013. "On the role of four small hairpins in the HIV-1 RNA genome." *RNA Biol* 10 (4):540-52. doi: 10.4161/rna.24133.

Kollmus, H., A. Honigman, A. Panet, and H. Hauser. 1994. "The sequences of and distance between two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency virus type 1 and human T-cell leukemia virus type II in vivo." *J Virol* 68 (9):6087-91.

Korber, B., B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir, and V. Detours. 2001. "Evolutionary and immunological implications of contemporary HIV-1 variation." *Br Med Bull* 58:19-42.

Korber, B., M. Muldoon, J. Theiler, F. Gao, R. Gupta, A. Lapedes, B. H. Hahn, S. Wolinsky, and T. Bhattacharya. 2000. "Timing the ancestor of the HIV-1 pandemic strains." *Science* 288 (5472):1789-96.

Kruger, K., P. J. Grabowski, A. J. Zaug, J. Sands, D. E. Gottschling, and T. R. Cech. 1982. "Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena." *Cell* 31 (1):147-57.

Kuiken, C., T. Leitner, B. H. Hahn, J. I. Mullins, S. Wolinsky, B. Foley, C. Apetrei, I. Mizrachi, A. Rambaut, and B. Korber. 2012. *HIV Sequence Compendium 2012*: T-6 Group, MS K710, Los Alamos National Laboratory, NM 87544, U.S.A. LA-UR-12-24653.

Lavery, R., and A. Pullman. 1984. "A new theoretical index of biochemical reactivity combining steric and electrostatic factors. An application to yeast tRNAPhe." *Biophys Chem* 19 (2):171-81.

Lee, R. C., R. L. Feinbaum, and V. Ambros. 1993. "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14." *Cell* 75 (5):843-54.

Lever, A., H. Gottlinger, W. Haseltine, and J. Sodroski. 1989. "Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions." *J Virol* 63 (9):4085-7.

Levin, J. G., M. Mitra, A. Mascarenhas, and K. Musier-Forsyth. 2010. "Role of HIV-1 nucleocapsid protein in HIV-1 reverse transcription." *RNA Biol* 7 (6):754-74.

Levy, J. A., A. D. Hoffman, S. M. Kramer, J. A. Landis, J. M. Shimabukuro, and L. S. Oshiro. 1984. "Isolation of lymphocytopathic retroviruses from San Francisco patients with AIDS." *Science* 225 (4664):840-2.

Li, Y., R. Uenishi, S. Hase, H. Liao, X. J. Li, T. Tsuchiura, K. K. Tee, O. G. Pybus, and Y. Takebe. 2010. "Explosive HIV-1 subtype B' epidemics in Asia driven by geographic and risk group founder events." *Virology* 402 (2):223-7. doi: 10.1016/j.virol.2010.03.048.

Los-Alamos-HIV-Sequence-Database. "www.hiv.lanl.gov." www.hiv.lanl.gov.

Low, J. T., and K. M. Weeks. 2010. "SHAPE-directed RNA secondary structure prediction." *Methods* 52 (2):150-8. doi: S1046-2023(10)00161-1 [pii]

10.1016/j.ymeth.2010.06.007.

Lucks, J. B., S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. 2011. "Multiplexed RNA structure characterization with

selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)." *Proc Natl Acad Sci U S A* 108 (27):11063-8. doi: 10.1073/pnas.1106501108.

Malim, M. H., and M. Emerman. 2008. "HIV-1 accessory proteins--ensuring viral survival in a hostile environment." *Cell Host Microbe* 3 (6):388-98. doi: 10.1016/j.chom.2008.04.008.

Mansky, L. M., and H. M. Temin. 1995. "Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase." *J Virol* 69 (8):5087-94.

Manzourolajdad, A., Y. Wang, T. I. Shaw, and R. L. Malmberg. 2013. "Information-theoretic uncertainty of SCFG-modeled folding space of the non-coding RNA." *J Theor Biol* 318:140-63. doi: 10.1016/j.jtbi.2012.10.023.

Matera, A. G., R. M. Terns, and M. P. Terns. 2007. "Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs." *Nat Rev Mol Cell Biol* 8 (3):209-20. doi: 10.1038/nrm2124.

Mathews, D. H., M. D. Disney, J. L. Childs, S. J. Schroeder, M. Zuker, and D. H. Turner. 2004. "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure." *Proc Natl Acad Sci U S A* 101 (19):7287-92. doi: 10.1073/pnas.0401799101 [pii].

Matlin, A. J., F. Clark, and C. W. Smith. 2005. "Understanding alternative splicing: towards a cellular code." *Nat Rev Mol Cell Biol* 6 (5):386-98. doi: 10.1038/nrm1645.

Mattick, J. S., and I. V. Makunin. 2006. "Non-coding RNA." *Hum Mol Genet* 15 Spec No 1:R17-29. doi: 10.1093/hmg/ddl046.

McCutchan, F. E. 2006. "Global epidemiology of HIV." *J Med Virol* 78 Suppl 1:S7-S12. doi: 10.1002/jmv.20599.

Moore, M. D., and W. S. Hu. 2009. "HIV-1 RNA dimerization: It takes two to tango." *AIDS Rev* 11 (2):91-102.

Mortimer, S. A., J. S. Johnson, and K. M. Weeks. 2009. "Quantitative analysis of RNA solvent accessibility by N-silylation of guanosine." *Biochemistry* 48 (10):2109-14. doi: 10.1021/bi801939g.

Moumen, A., L. Polomack, T. Unge, M. Veron, H. Buc, and M. Negroni. 2003. "Evidence for a mechanism of recombination during reverse transcription dependent on the structure of the acceptor RNA." *J Biol Chem* 278 (18):15973-82. doi: 10.1074/jbc.M212306200 M212306200 [pii].

Mouzakis, K. D., A. L. Lang, K. A. Vander Meulen, P. D. Easterday, and S. E. Butcher. 2013. "HIV-1 frameshift efficiency is primarily determined by the stability of base pairs positioned at the mRNA entrance channel of the ribosome." *Nucleic Acids Res* 41 (3):1901-13. doi: 10.1093/nar/gks1254.

Mueller, Nancy, Ben Berkhout, and Atze Das. 2013. "The RNA structure of the major splice donor site controls HIV-1 splicing." *Retrovirology* 10 (Suppl 1):P62.

Mujeeb, A., T. G. Parslow, A. Zarrinpar, C. Das, and T. L. James. 1999. "NMR structure of the mature dimer initiation complex of HIV-1 genomic RNA." *FEBS Lett* 458 (3):387-92.

Ndung'u, T., and R. A. Weiss. 2012. "On HIV diversity." *AIDS* 26 (10):1255-60. doi: 10.1097/QAD.0b013e32835461b5.

Negroni, M., and H. Buc. 2001. "Mechanisms of retroviral recombination." *Annu Rev Genet* 35:275-302. doi: 10.1146/annurev.genet.35.102401.090551.

Nixon, P. L., and D. P. Giedroc. 2000. "Energetics of a strongly pH dependent RNA tertiary structure in a frameshifting pseudoknot." *J Mol Biol* 296 (2):659-71. doi: 10.1006/jmbi.1999.3464.

Nixon, P. L., A. Rangan, Y. G. Kim, A. Rich, D. W. Hoffman, M. Hennig, and D. P. Giedroc. 2002. "Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot." *J Mol Biol* 322 (3):621-33.

Nussinov, R., and A. B. Jacobson. 1980. "Fast algorithm for predicting the secondary structure of single-stranded RNA." *Proc Natl Acad Sci U S A* 77 (11):6309-13.

Ocwieja, K. E., S. Sherrill-Mix, R. Mukherjee, R. Custers-Allen, P. David, M. Brown, S. Wang, D. R. Link, J. Olson, K. Travers, E. Schadt, and F. D. Bushman. 2012. "Dynamic regulation of HIV-1 mRNA populations analyzed by single-molecule enrichment and long-read sequencing." *Nucleic Acids Res* 40 (20):10345-55. doi: 10.1093/nar/gks753.

Osmanov, S., C. Pattou, N. Walker, B. Schwardlander, and J. Esparza. 2002. "Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000." *J Acquir Immune Defic Syndr* 29 (2):184-90.

Park, J., and C. D. Morrow. 1991. "Overexpression of the gag-pol precursor from human immunodeficiency virus type 1 proviral genomes results in efficient proteolytic processing in the absence of virion production." *J Virol* 65 (9):5111-7.

Parkin, N. T., M. Chamorro, and H. E. Varmus. 1992. "Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo." *J Virol* 66 (8):5147-51.

Parthasarathi, S., A. Varela-Echavarria, Y. Ron, B. D. Preston, and J. P. Dougherty. 1995. "Genetic rearrangements occurring during a single cycle of murine leukemia virus vector replication: characterization and implications." *J Virol* 69 (12):7991-8000.

Peeters, M., and P. M. Sharp. 2000. "Genetic diversity of HIV-1: the moving target." *AIDS* 14 Suppl 3:S129-40.

Peeters, M., C. Toure-Kane, and J. N. Nkengasong. 2003. "Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials." *AIDS* 17 (18):2547-60. doi: 10.1097/01.aids.0000096895.73209.89.

Peleg, O., E. N. Trifonov, and A. Bolshoy. 2003. "Hidden messages in the nef gene of human immunodeficiency virus type 1 suggest a novel RNA secondary structure." *Nucleic Acids Res* 31 (14):4192-200.

Peliska, J. A., and S. J. Benkovic. 1992. "Mechanism of DNA strand transfer reactions catalyzed by HIV-1 reverse transcriptase." *Science* 258 (5085):1112-8.

Pelletier, J., and N. Sonenberg. 1988. "Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA." *Nature* 334 (6180):320-5. doi: 10.1038/334320a0.

Plant, E. P., and J. D. Dinman. 2005. "Torsional restraint: a new twist on frameshifting pseudoknots." *Nucleic Acids Res* 33 (6):1825-33. doi: 10.1093/nar/gki329.

Plantier, J. C., M. Leoz, J. E. Dickerson, F. De Oliveira, F. Cordonnier, V. Lemee, F. Damond, D. L. Robertson, and F. Simon. 2009. "A new human immunodeficiency virus derived from gorillas." *Nat Med* 15 (8):871-2. doi: 10.1038/nm.2016.

Pollard, V. W., and M. H. Malim. 1998. "The HIV-1 Rev protein." *Annu Rev Microbiol* 52:491-532. doi: 10.1146/annurev.micro.52.1.491.

Pollom, E., K. K. Dang, E. L. Potter, R. J. Gorelick, C. L. Burch, K. M. Weeks, and R. Swanstrom. 2013. "Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs." *PLoS Pathog* 9 (4):e1003294. doi: 10.1371/journal.ppat.1003294.

Ponting, C. P., P. L. Oliver, and W. Reik. 2009. "Evolution and functions of long noncoding RNAs." *Cell* 136 (4):629-41. doi: 10.1016/j.cell.2009.02.006.

Popovic, M., M. G. Sarngadharan, E. Read, and R. C. Gallo. 1984. "Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS." *Science* 224 (4648):497-500.

Purcell, D. F., and M. A. Martin. 1993. "Alternative splicing of human immunodeficiency virus type 1 mRNA modulates viral protein expression, replication, and infectivity." *J Virol* 67 (11):6365-78.

Qu, Z., and D. L. Adelson. 2012. "Evolutionary conservation and functional roles of ncRNA." *Front Genet* 3:205. doi: 10.3389/fgene.2012.00205.

Rambaut, A., D. Posada, K. A. Crandall, and E. C. Holmes. 2004. "The causes and consequences of HIV evolution." *Nat Rev Genet* 5 (1):52-61. doi: 10.1038/nrg1246 nrg1246 [pii].

Reuter, J. S., and D. H. Mathews. 2010. "RNAstructure: software for RNA secondary structure prediction and analysis." *BMC Bioinformatics* 11:129. doi: 10.1186/1471-2105-11-129.

Rich, Alexander. 1961. *The transfer of information between the nucleic acids*. Edited by D. Rudnick, *Molecular and Cellular Synthesis*. New York: Ronald Press.

Rich, Alexander, and David R. Davies. 1956. "A NEW TWO STRANDED HELICAL STRUCTURE: POLYADENYLIC ACID AND POLYURIDYLIC ACID." *J Am Chem Soc* 78 (14):3548-3549. doi: 10.1021/ja01595a086.

Robberson, B. L., G. J. Cote, and S. M. Berget. 1990. "Exon definition may facilitate splice site selection in RNAs with multiple exons." *Mol Cell Biol* 10 (1):84-94.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber. 2000. "HIV-1 nomenclature proposal." *Science* 288 (5463):55-6.

Rodriguez-Alvarado, G., and M. J. Roossinck. 1997. "Structural analysis of a necrogenic strain of cucumber mosaic cucumovirus satellite RNA in planta." *Virology* 236 (1):155-66. doi: 10.1006/viro.1997.8731.

Roy, S., U. Delling, C. H. Chen, C. A. Rosen, and N. Sonenberg. 1990. "A bulge structure in HIV-1 TAR RNA is required for Tat binding and Tat-mediated trans-activation." *Genes Dev* 4 (8):1365-73.

Sakuragi, J., S. Sakuragi, M. Ohishi, and T. Shioda. 2010. "Direct correlation between genome dimerization and recombination efficiency of HIV-1." *Microbes Infect* 12 (12-13):1002-11. doi: 10.1016/j.micinf.2010.06.012.

Sanjuan, R., and A. V. Borderia. 2011. "Interplay between RNA structure and protein evolution in HIV-1." *Mol Biol Evol* 28 (4):1333-8. doi: 10.1093/molbev/msq329.

Schroeder, S. J. 2009. "Advances in RNA structure prediction from sequence: new tools for generating hypotheses about viral RNA structure-function relationships." *J Virol* 83 (13):6326-34. doi: JVI.00251-09 [pii] 10.1128/JVI.00251-09.

Sharp, P. M., and B. H. Hahn. 2011. "Origins of HIV and the AIDS pandemic." *Cold Spring Harb Perspect Med* 1 (1):a006841. doi: 10.1101/cshperspect.a006841.

Shaw, T. I., and M. Zhang. 2013. "HIV N-linked glycosylation site analyzer and its further usage in anchored alignment." *Nucleic Acids Res* 41 (Web Server issue):W454-8. doi: 10.1093/nar/gkt472.

Shehu-Xhilaga, M., S. M. Crowe, and J. Mak. 2001. "Maintenance of the Gag/Gag-Pol ratio is important for human immunodeficiency virus type 1 RNA dimerization and viral infectivity." *J Virol* 75 (4):1834-41. doi: 10.1128/JVI.75.4.1834-1841.2001.

Si, Z., M. Cayabyab, and J. Sodroski. 2001. "Envelope glycoprotein determinants of neutralization resistance in a simian-human immunodeficiency virus (SHIV-HXBc2P 3.2) derived by passage in monkeys." *J Virol* 75 (9):4208-18. doi: 10.1128/JVI.75.9.4208-4218.2001.

Snoeck, J., J. Fellay, I. Bartha, D. C. Douek, and A. Telenti. 2011. "Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints." *Retrovirology* 8:87. doi: 10.1186/1742-4690-8-87.

Solomatin, S. V., M. Greenfeld, S. Chu, and D. Herschlag. 2010. "Multiple native states reveal persistent ruggedness of an RNA folding landscape." *Nature* 463 (7281):681-4. doi: 10.1038/nature08717.

Soto-Ramirez, L. E., B. Renjifo, M. F. McLane, R. Marlink, C. O'Hara, R. Sutthent, C. Wasi, P. Vithayasai, V. Vithayasai, C. Apichartpiyakul, P. Auewarakul, V. Pena Cruz, D. S. Chui, R. Osathanondh, K. Mayer, T. H. Lee, and M. Essex. 1996. "HIV-1 Langerhans' cell tropism associated with heterosexual transmission of HIV." *Science* 271 (5253):1291-3.

Stanley, P., H. Schachter, and N. Taniguchi. 2009. "N-Glycans." In *Essentials of Glycobiology*, edited by A. Varki, R. D. Cummings, J. D. Esko, H. H. Freeze, P. Stanley, C. R. Bertozzi, G. W. Hart and M. E. Etzler. Cold Spring Harbor (NY).

Stephenson, J. D., H. Li, J. C. Kenyon, M. Symmons, D. Klenerman, and A. M. Lever. 2013. "Three-dimensional RNA structure of the major HIV-1 packaging signal region." *Structure* 21 (6):951-62. doi: 10.1016/j.str.2013.04.008.

Stern, S., D. Moazed, and H. F. Noller. 1988. "Structural analysis of RNA using chemical and enzymatic probing monitored by primer extension." *Methods Enzymol* 164:481-9.

Stoltzfus, C. M. 2009. "Chapter 1. Regulation of HIV-1 alternative RNA splicing and its role in virus replication." *Adv Virus Res* 74:1-40. doi: 10.1016/S0065-3527(09)74001-1.

Strappe, P. M., J. Greatorex, J. Thomas, P. Biswas, E. McCann, and A. M. Lever. 2003. "The packaging signal of simian immunodeficiency virus is upstream of the major splice donor at a distance from the RNA cap site similar to that of human immunodeficiency virus types 1 and 2." *J Gen Virol* 84 (Pt 9):2423-30.

Sukosd, Z., B. Knudsen, J. Kjems, and C. Pedersen. 2012. "PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data." *Bioinformatics*. doi: 10.1093/bioinformatics/bts488.

Sukosd, Z., B. Knudsen, M. Vaerum, J. Kjems, and E. S. Andersen. 2011. "Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars." *BMC Bioinformatics* 12:103. doi: 1471-2105-12-103 [pii] 10.1186/1471-2105-12-103.

Sukosd, Z., M. S. Swenson, J. Kjems, and C. E. Heitsch. 2013. "Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions." *Nucleic Acids Res* 41 (5):2807-16. doi: 10.1093/nar/gks1283.

Svoboda, P., and A. Di Cara. 2006. "Hairpin RNA: a secondary structure of primary importance." *Cell Mol Life Sci* 63 (7-8):901-8. doi: 10.1007/s00018-005-5558-5.

Taft, R. J., K. C. Pang, T. R. Mercer, M. Dinger, and J. S. Mattick. 2010. "Non-coding RNAs: regulators of disease." *J Pathol* 220 (2):126-39. doi: 10.1002/path.2638.

Tazi, J., N. Bakkour, V. Marchand, L. Ayadi, A. Aboufirassi, and C. Branlant. 2010. "Alternative splicing: regulation of HIV-1 multiplication as a target for therapeutic action." *FEBS J* 277 (4):867-76. doi: 10.1111/j.1742-4658.2009.07522.x.

Tersmette, M., R. A. Gruters, F. de Wolf, R. E. de Goede, J. M. Lange, P. T. Schellekens, J. Goudsmit, H. G. Huisman, and F. Miedema. 1989. "Evidence for a role of virulent human immunodeficiency virus (HIV) variants in the pathogenesis of acquired immunodeficiency syndrome: studies on sequential HIV isolates." *J Virol* 63 (5):2118-25.

Thompson, M. A., J. A. Aberg, J. F. Hoy, A. Telenti, C. Benson, P. Cahn, J. J. Eron, H. F. Gunthard, S. M. Hammer, P. Reiss, D. D. Richman, G. Rizzardini, D. L. Thomas, D. M. Jacobsen, and P. A. Volberding. 2012. "Antiretroviral treatment of adult HIV infection: 2012 recommendations of the International Antiviral Society-USA panel." *JAMA* 308 (4):387-402. doi: 10.1001/jama.2012.7961.

Tournoud, M., R. Ecochard, L. Kuhn, and A. Coutsoudis. 2008. "Diversity of risk of mother-to-child HIV-1 transmission according to feeding practices, CD4 cell count, and haemoglobin concentration in a South African cohort." *Trop Med Int Health* 13 (3):310-8. doi: 10.1111/j.1365-3156.2008.02004.x.

Tullius, T. D., and J. A. Greenbaum. 2005. "Mapping nucleic acid structure by hydroxyl radical cleavage." *Curr Opin Chem Biol* 9 (2):127-34. doi: 10.1016/j.cbpa.2005.02.009.

UNAIDS. 2010. *Global Report: UNAIDS Report on the Global AIDS Epidemic: 2010*: UN Joint Programme on HIV/AIDS (UNAIDS).

UNAIDS. 2012. UNAIDS Report on the Global AIDS Epidemic. UNAIDS.

van der Kuyl, A. C., and B. Berkhout. 2012. "The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus." *Retrovirology* 9:92. doi: 10.1186/1742-4690-9-92.

van Hemert, F. J., A. C. van der Kuyl, and B. Berkhout. 2013. "The A-nucleotide preference of HIV-1 in the context of its structured RNA genome." *RNA Biol* 10 (2):211-5. doi: 10.4161/rna.22896.

Wain, L. V., E. Bailes, F. Bibollet-Ruche, J. M. Decker, B. F. Keele, F. Van Heuverswyn, Y. Li, J. Takehisa, E. M. Ngole, G. M. Shaw, M. Peeters, B. H. Hahn, and P. M. Sharp. 2007. "Adaptation of HIV-1 to its human host." *Mol Biol Evol* 24 (8):1853-60. doi: 10.1093/molbev/msm110.

Wang, Q., I. Barr, F. Guo, and C. Lee. 2008. "Evidence of a novel RNA secondary structure in the coding region of HIV-1 pol gene." *RNA* 14 (12):2478-88. doi: 10.1261/rna.1252608.

Wang, Z., and C. B. Burge. 2008. "Splicing regulation: from a parts list of regulatory elements to an integrated splicing code." *RNA* 14 (5):802-13. doi: 10.1261/rna.876308.

Washietl, S., I. L. Hofacker, M. Lukasser, A. Huttenhofer, and P. F. Stadler. 2005. "Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome." *Nat Biotechnol* 23 (11):1383-90. doi: nbt1144 [pii] 10.1038/nbt1144.

Washietl, S., I. L. Hofacker, P. F. Stadler, and M. Kellis. 2012. "RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction." *Nucleic Acids Res* 40 (10):4261-72. doi: 10.1093/nar/gks009.

Washietl, S., S. Will, D. A. Hendrix, L. A. Goff, J. L. Rinn, B. Berger, and M. Kellis. 2012. "Computational analysis of noncoding RNAs." *Wiley Interdiscip Rev RNA* 3 (6):759-78. doi: 10.1002/wrna.1134.

Watts, J. M., K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess, Jr., R. Swanstrom, C. L. Burch, and K. M. Weeks. 2009. "Architecture and secondary structure of an entire HIV-1 RNA genome." *Nature* 460 (7256):711-6. doi: nature08237 [pii]

10.1038/nature08237.

Weeks, K. M. 2010. "Advances in RNA structure analysis by chemical probing." *Curr Opin Struct Biol* 20 (3):295-304. doi: 10.1016/j.sbi.2010.04.001.

Whisnant, A. W., H. P. Bogerd, O. Flores, P. Ho, J. G. Powers, N. Sharova, M. Stevenson, C. H. Chen, and B. R. Cullen. 2013. "In-depth analysis of the interaction of HIV-1 with cellular microRNA biogenesis and effector mechanisms." *MBio* 4 (2):e000193. doi: 10.1128/mBio.00193-13.

Wianny, F., and M. Zernicka-Goetz. 2000. "Specific interference with gene function by double-stranded RNA in early mouse development." *Nat Cell Biol* 2 (2):70-5. doi: 10.1038/35000016.

Wilkinson, K. A., R. J. Gorelick, S. M. Vasa, N. Guex, A. Rein, D. H. Mathews, M. C. Giddings, and K. M. Weeks. 2008. "High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states." *PLoS Biol* 6 (4):e96. doi: 10.1371/journal.pbio.0060096.

Wilkinson, K. A., E. J. Merino, and K. M. Weeks. 2006. "Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution." *Nat Protoc* 1 (3):1610-6. doi: 10.1038/nprot.2006.249.

Willey, R. L., F. Maldarelli, M. A. Martin, and K. Strebel. 1992. "Human immunodeficiency virus type 1 Vpu protein induces rapid degradation of CD4." *J Virol* 66 (12):7193-200.

Worobey, M., M. Gemmel, D. E. Teuwen, T. Haselkorn, K. Kunstman, M. Bunce, J. J. Muyembe, J. M. Kabongo, R. M. Kalengayi, E. Van Marck, M. T. Gilbert, and S. M. Wolinsky. 2008. "Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960." *Nature* 455 (7213):661-4. doi: 10.1038/nature07390.

Wu, W., B. M. Blumberg, P. J. Fay, and R. A. Bambara. 1995. "Strand transfer mediated by human immunodeficiency virus reverse transcriptase in vitro is promoted by pausing and results in misincorporation." *J Biol Chem* 270 (1):325-32.

Wyatt, R., P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski. 1998. "The antigenic structure of the HIV gp120 envelope glycoprotein." *Nature* 393 (6686):705-11. doi: 10.1038/31514.

Xia, T., J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. 1998. "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs." *Biochemistry* 37 (42):14719-35. doi: 10.1021/bi9809425.

Yu, C. H., M. H. Noteborn, C. W. Pleij, and R. C. Olsthoorn. 2011. "Stem-loop structures can effectively substitute for an RNA pseudoknot in -1 ribosomal frameshifting." *Nucleic Acids Res* 39 (20):8952-9. doi: 10.1093/nar/gkr579.

Yu, G., Y. Li, J. Li, L. Diao, X. Yan, P. Lin, Q. He, Y. Wang, X. Fu, F. Yang, and Q. Long. 2009. "Genetic diversity and drug resistance of HIV type 1 circulating recombinant Form_BC among drug users in Guangdong Province." *AIDS Res Hum Retroviruses* 25 (9):869-75. doi: 10.1089/aid.2008.0312.

Yusim, K., C. Kesmir, B. Gaschen, M. M. Addo, M. Altfeld, S. Brunak, A. Chigaev, V. Detours, and B. T. Korber. 2002. "Clustering patterns of cytotoxic T-lymphocyte epitopes in human immunodeficiency virus type 1 (HIV-1) proteins reveal imprints of immune evasion on HIV-1 global variation." *J Virol* 76 (17):8757-68.

Zhang, H., R. J. Pomerantz, G. Dornadula, and Y. Sun. 2000. "Human immunodeficiency virus type 1 Vif protein is an integral component of an mRNP complex of viral RNA and could be involved in the viral RNA folding and packaging process." *J Virol* 74 (18):8252-61.

Zhang, M., B. Foley, A. K. Schultz, J. P. Macke, I. Bulla, M. Stanke, B. Morgenstern, B. Korber, and T. Leitner. 2010. "The role of recombination in the emergence of a complex and dynamic HIV epidemic." *Retrovirology* 7:25. doi: 10.1186/1742-4690-7-25.

Zhang, M., B. Gaschen, W. Blay, B. Foley, N. Haigwood, C. Kuiken, and B. Korber. 2004. "Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin." *Glycobiology* 14 (12):1229-46. doi: 10.1093/glycob/cwh106.

Zhuang, J., A. E. Jetzt, G. Sun, H. Yu, G. Klarmann, Y. Ron, B. D. Preston, and J. P. Dougherty. 2002. "Human immunodeficiency virus type 1 recombination: rate, fidelity, and putative hot spots." *J Virol* 76 (22):11273-82.

Znosko, B. M., S. B. Silvestri, H. Volkman, B. Boswell, and M. J. Serra. 2002. "Thermodynamic parameters for an expanded nearest-neighbor model for the formation of RNA duplexes with single nucleotide bulges." *Biochemistry* 41 (33):10406-17.

Zuker, M. 1986. "RNA folding prediction: the continued need for interaction between biologists and mathematicians." *Lectures Math. Life Sci* 17.

Zuker, M. 1989. "On finding all suboptimal foldings of an RNA molecule." *Science* 244 (4900):48-52.

CHAPTER 2

HIV N-LINKED GLYCOSYLATION SITE ANALYZER AND ITS FURTHER USAGE IN ANCHORED ALIGNMENT[1]

---

[1] Timothy I. Shaw and Ming Zhang. 2013. *Nucleic Acids Research*. (Web Server issue):W454-8.

Reprinted here with permission from the publisher.

Abstract

N-linked glycosylation is a posttranslational modification that has significantly contributed to the rapid evolution of HIV-1. In particular, enrichment of N-linked glycosylation sites can be found within Envelope variable loops, regions that play an essential role in HIV pathogenesis and immunogenicity. The web server described here, the HIV N-linked Glycosylation Site Analyzer, was developed to facilitate study of HIV diversity by tracking gp120 N-linked glycosylation sites. This server provides an automated platform for mapping and comparing variable loop N-linked glycosylation sites across populations of HIV-1 sequences. Furthermore, this server allows for refinement of HIV-1 sequence alignment by using N-linked glycosylation sites in variable loops as alignment anchors. Availability of this web server solves one of the difficult problems in HIV gp120 alignment and analysis imposed by the extraordinary HIV-1 diversity. The HIV N-linked Glycosylation Site Analyzer web server is available at ***http://hivtools.publichealth.uga.edu/N-Glyco/***.

Introduction

Strategic placement and loss and gain of N-linked glycosylation sites are one of the most important evolutionary mechanisms adopted by HIV-1 to generate its extraordinary sequence diversity (Zhang et al. 2004). A typical N-linked glycosylation site contains an amino acid pattern of N-X-[S or T] (Marshall 1974), with X being any amino acid except Proline (Gavel and von Heijne 1990). Highly glycosylated regions are referred to as immunologically silent faces (Moore and Sodroski 1996), reducing antigenicity and restricting access to chemokine receptors. Changes in N-linked glycosylation sites in HIV-1 can induce conformational changes in Envelope gp120, diminishing binding of many gp120-specific antibodies (Si, Cayabyab, and Sodroski 2001). Comparison between neutralization-sensitive and neutralization-resistant HIV-1 strains shows a higher number of glycosylation sites associated with the resistant clusters (Kulkarni et al. 2009). Changes in N-linked glycosylation sites have also been linked to both disease stage and co-receptor usage. Leal et al. reported an increase in N-linked glycosylated sites during late stages of HIV-1 infection (Leal et al. 2012). Evaluation of co-receptor usage has demonstrated a tendency for higher mutation rates, higher net positive charges and fewer glycosylation sites within HIV-1 strains with CXCR4 co-receptor usage (Lin et al. 2012).

In HIV-1, N-linked glycosylation sites are enriched within the variable loops, which contain multiple neutralizing antibody binding sites (Wyatt et al. 1998). Changes of N-linked glycosylation sites within variable loops, as well as changes of lengths of variable loops imposed by frequent indels (insertion and deletions), are highly favored in HIV-1 (Wyatt et al. 1998, Zhang et al. 2004). Both changes are important measures of HIV-1 diversity (Korber et al. 2001). Although immunologically and evolutionarily important, HIV-1 variable loops are notoriously difficult to analyze due to extraordinary viral diversity in these regions (Abecasis 2007). As a

result, variable loops are typically excluded from phylogenetic analyses (Korber et al. 2001, Kulkarni et al. 2009, Leal et al. 2012), leading to frequent underestimation of HIV-1 diversity in immunologically important genomic regions.

To address the importance of N-linked glycosylation sites in HIV-1 and problems in analyzing variable loops as described above, we present the development of the HIV N-linked Glycosylation Site Analyzer, available at http://hivtools.publichealth.uga.edu/N-Glyco/. This server provides an automated platform for mapping and comparing N-linked glycosylation sites within variable loops among populations of HIV-1 sequences. Furthermore, considering the functional importance and conserved patterns of N-linked glycosylation sites, we have implemented in this server a feature that optimizes HIV-1 sequence alignment using N-linked glycosylation sites in variable loops as alignment anchors. As a result, our N-linked Glycosylation Site Analyzer serves as a valuable gateway for exploring HIV-1 diversity in immunologically important genomic regions, contributing to an improved understanding of host-virus interaction and enhanced viral vaccine strain selection.

Material and Method

Two key features distinguish our HIV-1 N-linked Glycosylation Site Analyzer from other HIV-1 sequence analysis tools and servers. First, through an automated pipeline, changes at N-linked glycosylation sites within each variable loop region, as well as loop lengths, can be easily tracked and compared among populations of HIV sequences. Second, the server optimizes HIV-1 sequence alignment by using the N-linked glycosylation site as the alignment anchor. Implementation of both features has been written in Java. The web server interface is implemented through HTML and Bootstrap JavaScript. Visualization methods are available for all results (see details in 'Server Output' section below).

Algorithm

In the N-linked Glycosylation Site Comparison program (N-Glyco Site Compare), input sequences are automatically aligned with the HIV-1 reference strain HXB2 as recommended by Korber et al. (Accession number: K03455. http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html). Through implementation of the HIV alignment algorithm as described by Gaschen B. et al (Gaschen et al. 2001), the variable loops V1-V5 are identified and clipped based on genomic coordinates defined in HIV Sequence Compendium 2012 (Kuiken et al. 2012). Within each variable loop region, the N-linked glycosylation sites, whose pattern is N-X-[S or T] (Marshall 1974), are identified by pattern matching of asparagine followed by any amino acid except Proline, followed by either a serine or threonine. In the case of continuous N-linked glycosylation sites (e.g., NNST), only the first N-linked glycosylation site is counted because two continuous N-linked glycosylation sites would induce steric occlusion. An exception exists for NNST in which the second glycosylated asparagine is counted because N-X-T are more frequently glycosylated than N-X-S (Kasturi et al. 1995), and oligosaccharyltransferase has a higher affinity for N-X-T than N-X-S (Gerber et al. 2013).

In the program 'V Loop Alignment', we optimize V loop region alignments by using N-linked glycosylation sites within V loops as alignment anchors. The 'V Loop Alignment' program accepts input for both aligned and unaligned sequences. The HXB2 sequence (Accession number: K03455) is used as the reference in the alignment procedure; therefore, HXB2 is automatically added to the input sequences when absent from the user input. For unaligned sequences, they are initially aligned through a HMMER-generated HIV profile (Gaschen et al. 2001). The aligned sequences, from direct user input or HMMER derived alignment, will then be refined based on a

heuristic approach for manual curation of HIV-1 alignments (Gnanakaran et al. 2011, Kulkarni et al. 2009, Zhang et al. 2010). For each variable loop region, the input sequence with the highest number of N-linked glycosylation sites for that region is identified, and its N-linked glycosylation sites are used as alignment anchors for all input sequences. This process continues through each variable loop region. The N-linked glycosylation sites for the rest of the input sequences are then aligned to these anchors based on a greedy algorithm, mapping each N-linked glycosylation site to its closest available anchor. Shannon Entropy (Shannon 1997b) is used to evaluate the N-linked Glycosylation anchored alignment (Equation 2.1). In equation 2.1, i represents the position in the V loop where the N-linked Glycosylation site is present, and p is the proportion of N-linked Glycosylation at each position. Lower entropy indicates a higher ordering of N-Glycosylation sites. Table 2.1 shows example of N-linked Glycosylation anchor alignment improving the HIV LANL 2010 reference sequence alignment (Los-Alamos-HIV-Sequence-Database)with the lowering of the Shannon Entropy.

$$\sum p_i log p_i \quad \text{[Eq. 2.1]}$$

Table 2.1 Shannon Entropy's Decrease after N-linked Glycosylation Alignment Improvement of the 2010 HIV LANL Reference Sequence Alignment

VLoops Region	Entropy Before Improvement	Entropy After Alignment
V1	4.69518	1.66266
V2	2.60315	1.45905
V3	0.04873	0.04873
V4	3.82823	1.43705
V5	2.08315	0.87835

Results

N-linked Glycosylation Site Comparison Program

Input

The N-Glyco Site Comparison program is designed to compare groups of HIV sequences for variation and changes in N-linked glycosylation patterns. The comparison groups are those sequences under different conditions, for instance, sequences at different time points, of different subtypes and associated with different risk factors. The N-Glyco Site Comparison program reads in two sets of FASTA sequences, namely query and background, respectively, and compares their N-linked glycosylation site frequency and variable loop lengths. Three options are provided for selecting the background sequences: (i) No background, which allows N-glycosylation site analysis to be performed in one single sequence or one set of sequences (i.e. in the query set); (ii) Using the most recent HIV-1 M group reference sequence set as the background. The reference sequences were obtained from the Los Alamos HIV Sequence Database group; and (iii) User-defined background sequences, which bestow flexibility in performing user-defined comparisons. The input of the N-Glyco Site Comparison program can be either aligned or unaligned gp120 sequences. Both nucleotide and protein sequences are acceptable as input.

Output

Output from the N-Glyco Site Comparison program highlights N-linked glycosylation sites through a graphical histogram spanning across HXB2 Envelope positioning (Korber 1998) (Fig. 2.1a). Loop length and frequency of N-glycosylation site distribution within each variable loop (V1-V5) are compared and depicted in a boxplot between comparison groups (Fig. 2.1b and 2.1c). Furthermore, a two-sided Wilcoxon test with 1000 replicates of Monte Carlo resampling is

provided for comparison statistics. The N-Glycosylation site and V loop mapping for each sequence are provided. Visual representation for the N-Glycosylation mapping is described in further detail in the section below ('N-linked Glycosylation Site Alignment Program – Output' section).

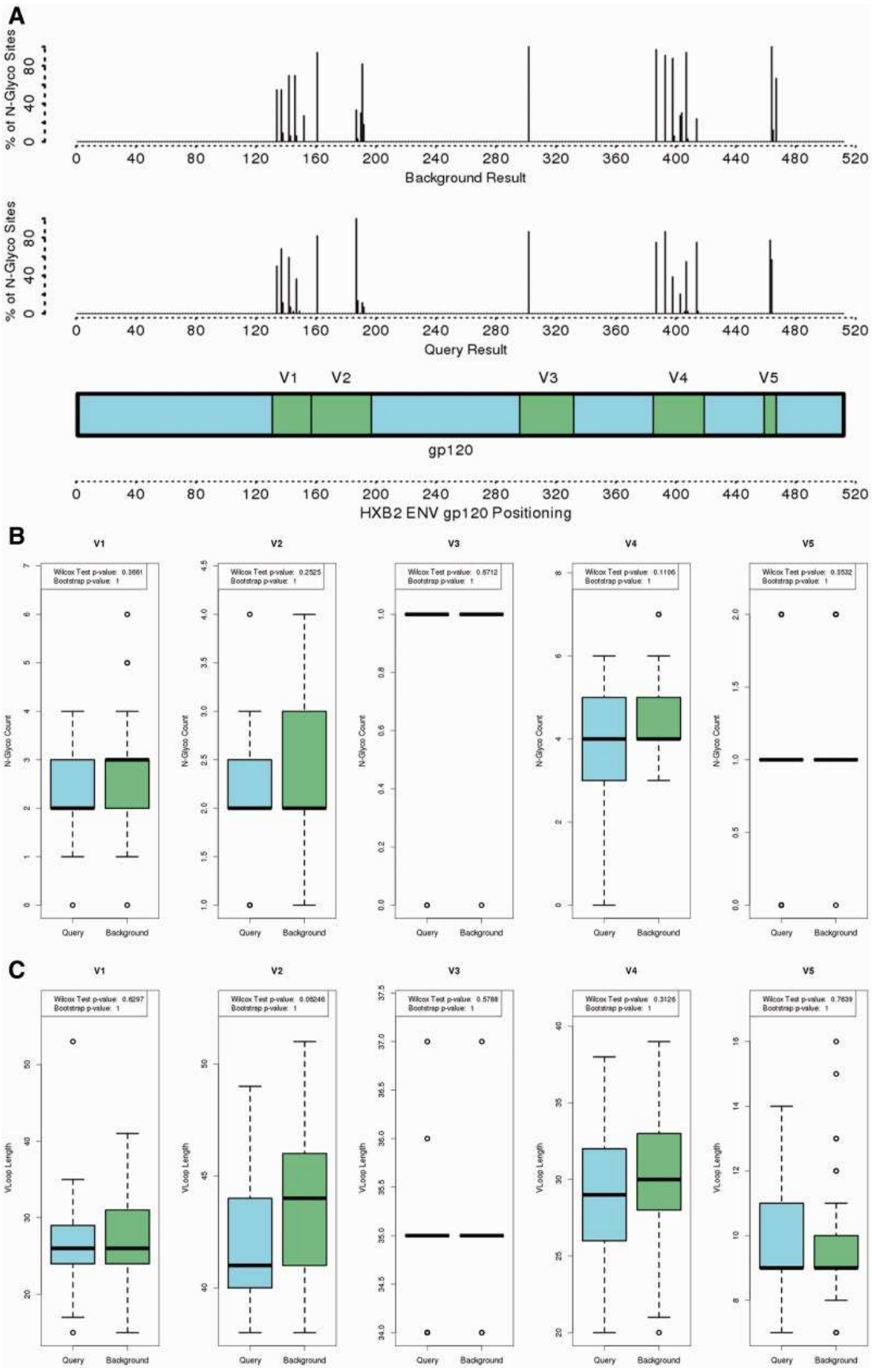

A
% of N-Glyco Sites
Background Result
Query Result
V1
V2
V3
V4
V5
gp120
HXB2 ENV gp120 Positioning
B
N-Glyco Count
Query
Background
C
VLoop Length

Figure 2.1. An example output of N-linked Glycosylation Site Comparison program

*(A) Location of identified N-linked glycosylation sites within the variable loops (V1-V5) in terms of HXB2 numbering (**http://www.hiv.lanl.gov/content/sequence/HIV/REVIEWS/HXB2.html**). Y-axis: Percentage of sequences with N-linked glycosylation site at each alignment position. X-axis: HXB2-based gp120 sequence positions. (B and C) The distribution of number of N-linked glycosylation sites and lengths of variable loops. P-value is calculated in two-sided Wilcoxon test. The bootstrap P-value is calculated by 1000 times of Monte Carlo resampling.*

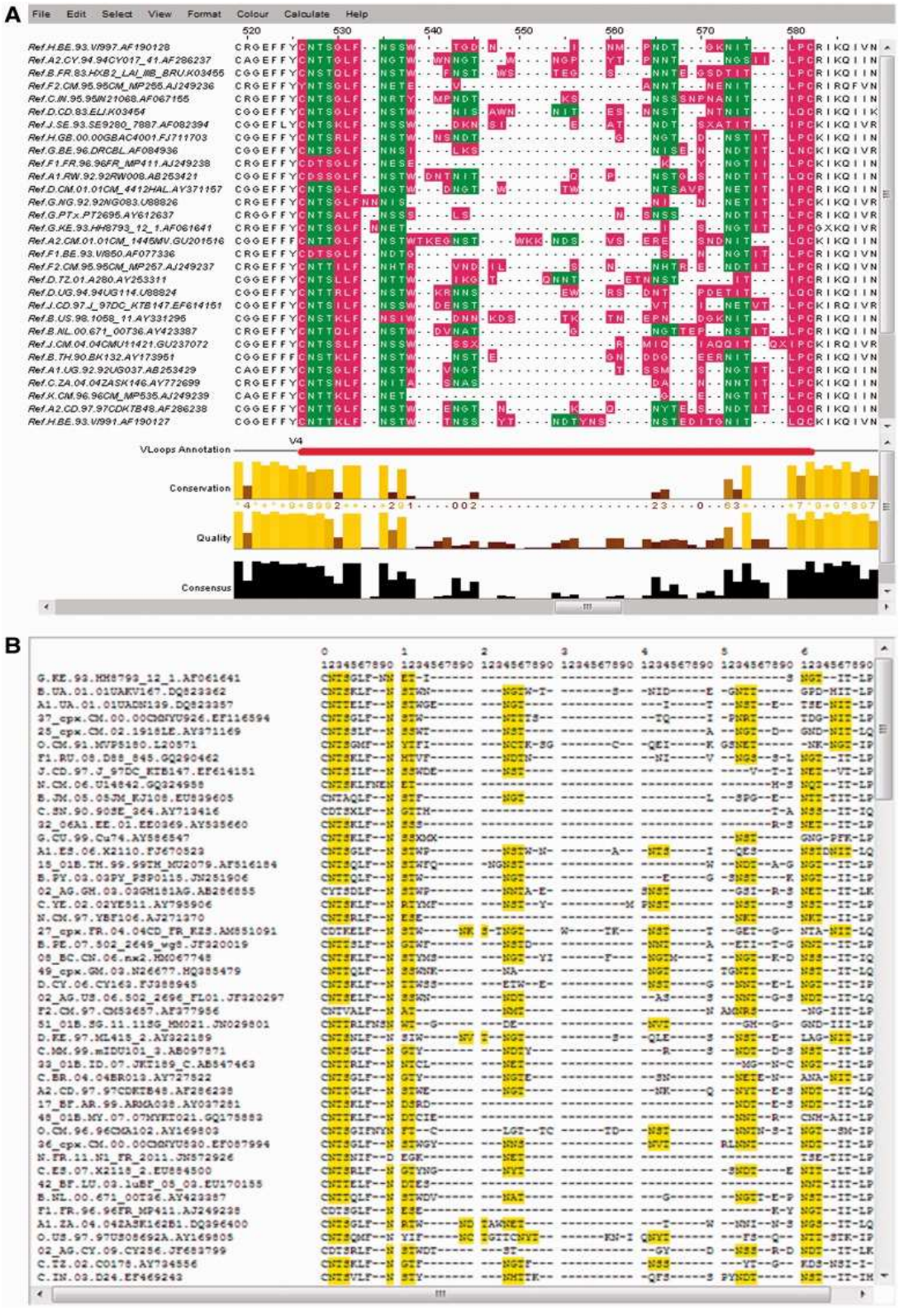


Figure 2.2. An example output of N-linked Glycosylation Site Alignment program.

(A) A Jalview-based alignment editor depicts an optimized alignment using N-linked glycosylation sites as alignment anchors. From the V loop Alignment program, V loop regions are highlighted as pink and each N-linked glycosylation sites are highlighted in green. An additional V loop annotation track is added underneath the alignment. (B) An HTML view of the optimized alignment with the N-linked glycosylation sites are highlighted in yellow.

N-linked Glycosylation Site Alignment Program

Input

The N-linked Glycosylation Site Alignment program ('V Loop Alignment' program) reads in one FASTA input file, regardless of whether aligned or not. Both nucleotide and protein sequences are acceptable. The N-glycosylation sites within the variable loops are used as alignment anchors to optimize sequence alignments as described in the Algorithm Section.

Output

Two mechanisms are used for visualizing the N-linked glycosylation optimized alignment: (i) Jalview, a Java-based alignment editor that provides extensive functionality for alignment visualization and editing (Waterhouse et al. 2009, Clamp et al. 2004). The V loop regions are highlighted as pink and N-linked glycosylation sites are highlighted in green; an additional V loop annotation track is added underneath the alignment. (**Fig. 2.2a**); and (ii) an HTML-based visualization of the alignment with N-linked glycosylation sites highlighted in yellow (**Fig. 2.2b**). The nucleotide and protein version of the alignment are downloadable. Also available in the downloadable results are the annotation for the location of N-linked glycosylation site and V loop region for each sequence.

Conclusion

Our HIV-1 N-linked Glycosylation Site Analyzer provides an automated platform to map and compare patterns of N-linked glycosylation sites between populations of HIV-1 sequences. In addition, to address the problem of improper variable loop region alignment that causes underestimation of HIV-1 diversity, we have developed an algorithm for performing anchored alignment based on N-linked glycosylation sites. The toolset and analysis pipeline described here can be extended to understanding diversity and N-linked glycosylation patterns in other viruses. Our web server provides an important gateway to track N-linked glycosylation site patterns within HIV-1 populations, thus improving our capability to better understand viral diversity under changing contexts of antigenic structures and transmission mechanisms.

Funding

NIH [R03AI104258]; University of Georgia Research Fund [10793GR002] and University of Georgia Research Foundation Award [1021RX064536]. We would also like to acknowledge the support from the ARCS foundation for TIS. Funding for open access charge: University of Georgia [UGA10793GR002].

Acknowledgements

The authors thank Ms Tess Z. Griffin and anonymous reviewers for help and comments on the manuscript. We would also like to thank members of the Ming Zhang HIV Lab and various HIV research groups for their extensive testing of our web server.

References

Abecasis, A., A.-M. Vandamme, and P. Lemey. 2007. Sequence alignment in HIV computational analysis. In *HIV sequence compendium 2006/2007*, edited by Theoretical Biology and Biophysics Group. Los Alamos, NM: Los Alamos National Laboratory.

Clamp, M., J. Cuff, S. M. Searle, and G. J. Barton. 2004. "The Jalview Java alignment editor." *Bioinformatics* 20 (3):426-7. doi: 10.1093/bioinformatics/btg430.

Gaschen, B., C. Kuiken, B. Korber, and B. Foley. 2001. "Retrieval and on-the-fly alignment of sequence fragments from the HIV database." *Bioinformatics* 17 (5):415-8.

Gavel, Y., and G. von Heijne. 1990. "Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering." *Protein Eng* 3 (5):433-42.

Gerber, S., C. Lizak, G. Michaud, M. Bucher, T. Darbre, M. Aebi, J. L. Reymond, and K. P. Locher. 2013. "Mechanism of Bacterial Oligosaccharyltransferase: IN VITRO QUANTIFICATION OF SEQUON BINDING AND CATALYSIS." *J Biol Chem* 288 (13):8849-61. doi: 10.1074/jbc.M112.445940.

Gnanakaran, S., T. Bhattacharya, M. Daniels, B. F. Keele, P. T. Hraber, A. S. Lapedes, T. Shen, B. Gaschen, M. Krishnamoorthy, H. Li, J. M. Decker, J. F. Salazar-Gonzalez, S. Wang, C. Jiang, F. Gao, R. Swanstrom, J. A. Anderson, L. H. Ping, M. S. Cohen, M. Markowitz, P. A. Goepfert, M. S. Saag, J. J. Eron, C. B. Hicks, W. A. Blattner, G. D. Tomaras, M. Asmal, N. L. Letvin, P. B. Gilbert, A. C. Decamp, C. A. Magaret, W. R. Schief, Y. E. Ban, M. Zhang, K. A. Soderberg, J. G. Sodroski, B. F. Haynes, G. M. Shaw, B. H. Hahn, and B. Korber. 2011. "Recurrent signature patterns in HIV-1 B clade envelope

glycoproteins associated with either early or chronic infections." *PLoS Pathog* 7 (9):e1002209. doi: 10.1371/journal.ppat.1002209.

Kasturi, L., J. R. Eshleman, W. H. Wunner, and S. H. Shakin-Eshleman. 1995. "The hydroxy amino acid in an Asn-X-Ser/Thr sequon can influence N-linked core glycosylation efficiency and the level of expression of a cell surface glycoprotein." *J Biol Chem* 270 (24):14756-61.

Korber, B., B. Foley, C. Kuiken, S. Pillai, and J. Sodroski. 1998. "Numbering Positions in HIV Relative to HXB2CG." In *Human Retroviruses and AIDS 1998*, edited by C. Kuiken Korber, B. Foley, B. Hahn, F. McCutchan, J. Mellors, and J. and Sodroski. Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.

Korber, B., B. Gaschen, K. Yusim, R. Thakallapally, C. Kesmir, and V. Detours. 2001. "Evolutionary and immunological implications of contemporary HIV-1 variation." *Br Med Bull* 58:19-42.

Kuiken, C., T. Leitner, B. H. Hahn, J. I. Mullins, S. Wolinsky, B. Foley, C. Apetrei, I. Mizrachi, A. Rambaut, and B. Korber. 2012. *HIV Sequence Compendium 2012*: T-6 Group, MS K710, Los Alamos National Laboratory, NM 87544, U.S.A. LA-UR-12-24653.

Kulkarni, S. S., A. Lapedes, H. Tang, S. Gnanakaran, M. G. Daniels, M. Zhang, T. Bhattacharya, M. Li, V. R. Polonis, F. E. McCutchan, L. Morris, D. Ellenberger, S. T. Butera, R. C. Bollinger, B. T. Korber, R. S. Paranjape, and D. C. Montefiori. 2009. "Highly complex neutralization determinants on a monophyletic lineage of newly transmitted subtype C HIV-1 Env clones from India." *Virology* 385 (2):505-20. doi: 10.1016/j.virol.2008.12.032.

Leal, E., J. Casseb, M. Hendry, M. P. Busch, and R. S. Diaz. 2012. "Relaxation of adaptive evolution during the HIV-1 infection owing to reduction of CD4+ T cell counts." *PLoS One* 7 (6):e39776. doi: 10.1371/journal.pone.0039776.

Lin, N. H., C. Becerril, F. Giguel, V. Novitsky, S. Moyo, J. Makhema, M. Essex, S. Lockman, D. R. Kuritzkes, and M. Sagar. 2012. "Env sequence determinants in CXCR4-using human immunodeficiency virus type-1 subtype C." *Virology* 433 (2):296-307. doi: 10.1016/j.virol.2012.08.013.

Los-Alamos-HIV-Sequence-Database. "www.hiv.lanl.gov." www.hiv.lanl.gov.

Marshall, R. D. 1974. "The nature and metabolism of the carbohydrate-peptide linkages of glycoproteins." *Biochem Soc Symp* (40):17-26.

Moore, J. P., and J. Sodroski. 1996. "Antibody cross-competition analysis of the human immunodeficiency virus type 1 gp120 exterior envelope glycoprotein." *J Virol* 70 (3):1863-72.

Shannon, C. E. 1997b. "The mathematical theory of communication. 1963." *MD Comput* 14 (4):306-17.

Si, Z., M. Cayabyab, and J. Sodroski. 2001. "Envelope glycoprotein determinants of neutralization resistance in a simian-human immunodeficiency virus (SHIV-HXBc2P 3.2) derived by passage in monkeys." *J Virol* 75 (9):4208-18. doi: 10.1128/JVI.75.9.4208-4218.2001.

Waterhouse, A. M., J. B. Procter, D. M. Martin, M. Clamp, and G. J. Barton. 2009. "Jalview Version 2--a multiple sequence alignment editor and analysis workbench." *Bioinformatics* 25 (9):1189-91. doi: 10.1093/bioinformatics/btp033.

Wyatt, R., P. D. Kwong, E. Desjardins, R. W. Sweet, J. Robinson, W. A. Hendrickson, and J. G. Sodroski. 1998. "The antigenic structure of the HIV gp120 envelope glycoprotein." *Nature* 393 (6686):705-11. doi: 10.1038/31514.

Zhang, M., B. Foley, A. K. Schultz, J. P. Macke, I. Bulla, M. Stanke, B. Morgenstern, B. Korber, and T. Leitner. 2010. "The role of recombination in the emergence of a complex and dynamic HIV epidemic." *Retrovirology* 7:25. doi: 10.1186/1742-4690-7-25.

Zhang, M., B. Gaschen, W. Blay, B. Foley, N. Haigwood, C. Kuiken, and B. Korber. 2004. "Tracking global patterns of N-linked glycosylation site variation in highly variable viral glycoproteins: HIV, SIV, and HCV envelopes and influenza hemagglutinin." *Glycobiology* 14 (12):1229-46. doi: 10.1093/glycob/cwh106.

CHAPTER 3

SYSTEMATIC STUDY OF HIV GENOTYPING ERRORS: A NEGLECTED ISSUE IN HIV RESEARCH AND EPIDEMIC SURVEILLANCE[1]

---

[1] Timothy I. Shaw and Ming Zhang. Submitted to *Journal of Acquired Immune Deficiency Syndromes, 8/1/2013.*

Abstract

With rapid growth of identified HIV strains, HIV global surveillance has becoming more challenging. A knowledge gap remains in the genotyping quality of existing HIV sequences that have laid the foundation of HIV research and epidemic surveillance.

Methods: We systematically examined the genotyping quality of worldwide HIV sequences from four risk groups: heterosexual, MSM, IDU and mother-to-child group. Our evaluation pipeline encompasses phylogenetics, genetic distance measure, and subtype profile comparison.

Results: 5.28% sequences were found to be mis-genotyped. An uneven genotyping quality among the risk groups was found across different geo-regions and genomic region. Our results suggested a close monitoring of HIV-1 genotyping quality is needed for the heterosexual group in West Central Africa, the IDU group in South America, and the IDU and mother-to-child groups in China.

Conclusions: HIV global surveillance requires accurate genotyping information for establishing reliable and duplicable case definition. Genotyping quality evaluation for the publicly available HIV sequences had persistently been a neglected research and epidemic problem. The results presented here not only fill this knowledge gap but can also be used as a background reference for global and regional HIV surveillance and antiviral therapy development.

Keywords: HIV-1; mis-genotyping; HIV epidemics; risk factor groups.

Background

Accurate and reliable genotyping is pivotal for establishing disease case definition and for understanding an epidemic disease. HIV-1 M group is classified into 9 subtypes (Robertson et al. 2000) and over 50 recombinant clades (Los-Alamos-HIV-Sequence-Database), that lay the fundamental basis for HIV genotyping. With the advancement of sequencing biotechnology, large-scale genomic sequencing has improved genotyping strategies revealing extensive genetic diversity in HIV-1 (Zhang et al. 2010, Taveira 2012, UNAIDS 2012). As a result, traditional classification of HIV subtypes has been challenged (Paraskevis et al. 2001, Abecasis et al. 2007). Subtype I, which is now re-classified as a recombinant clade, exemplifies our evolving knowledge for HIV diversity and HIV classification modification (Paraskevis et al. 2001).

The constant influx of newly identified HIV strains has made us put less emphasis on retrospective sequences, which has contributed significantly in shaping all aspects of contemporary HIV research. Retrospective HIV sequences, despite their critical importance, have long been plagued as being erroneously genotyped due to inconsistent genotyping criteria utilized by different studies. Genotyping quality is further affected by the mis-use of genotyping programs or mis-interpretation of genotyping results. In our preliminary survey of B and C sequences sampled from Asian countries, we surprisingly found that over 50% of B and C subtype fragment sequences were mis-genotyped (Zhang et al. 2010). The consequence of neglecting HIV genotyping quality can be severe, as proper genotyping is critical to many aspects of HIV research and epidemic surveillance.

Here we present a comprehensive overview of genotyping quality evaluation for all publicly available HIV sequences, which were further stratified by geographic region and transmission route. Previous studies have revealed a possible association between HIV-1 clades

and transmission routes. CRF01_AE was reported to be associated with heterosexual transmission (Gao et al. 1996); and subtype C was found to be more closely associated with vaginal shedding of the viral particle than subtype A or D (John-Stewart et al. 2005). The stratified results presented here are fundamental to inferring the impact of mis-genotyping on tracking HIV epidemics and surveillance, and consequently the influence on regional vaccine development.

Methods

All sequences involved in this study were retrieved from GenBank and the Los Alamos HIV Sequence Database (Los-Alamos-HIV-Sequence-Database). To ensure proper genotyping quality, all sequences shorter than 400 nucleotides were filtered out. Only one sequence per patient was selected to avoid sequence redundancy.

The resulting sequences were subjected to genotype analysis using jpHMM and phylogenetics, a standard method for HIV genotyping. The jpHMM (jumping profile hidden Markov model) is a method based on genetic distance and subtype profile (Schultz et al. 2006, Zhang, Schultz, et al. 2006). One prominent feature that distinguishes jpHMM from other HIV genotyping programs is its capability of locating recombinant breakpoints with statistically supported uncertainty region estimates (Schultz et al. 2009). We had extensively tested the performance of jpHMM program (Zhang, Schultz, et al. 2006) before we confidently applied it in this study. The sequences with conflicting genotype information between their original GenBank/publication assignment and the jpHMM prediction were further examined by phylogenetic analyses. For each recombinant sequence, the phylogenetic analyses were performed in individual pure subtype genomic regions with less than 1% uncertainty region estimate predicted by jpHMM. PHYLIP program version 3.69 (J 1989) was applied for the

phylogenetic analysis, using F84 neighbor-joining method supported by 100 iterations of nonparametric bootstrapping. In brief, DNADIST and NEIGHBOR were used for constructing the phylogenetic trees, and SEQBOOT, DNADIST, NEIGHBOR, and CONSENSE were applied to assess reliability of clade clustering. The HIV subtype references (Los-Alamos-HIV-Sequence-Database) were used to provide reference for correct clade clustering in each constructed tree.

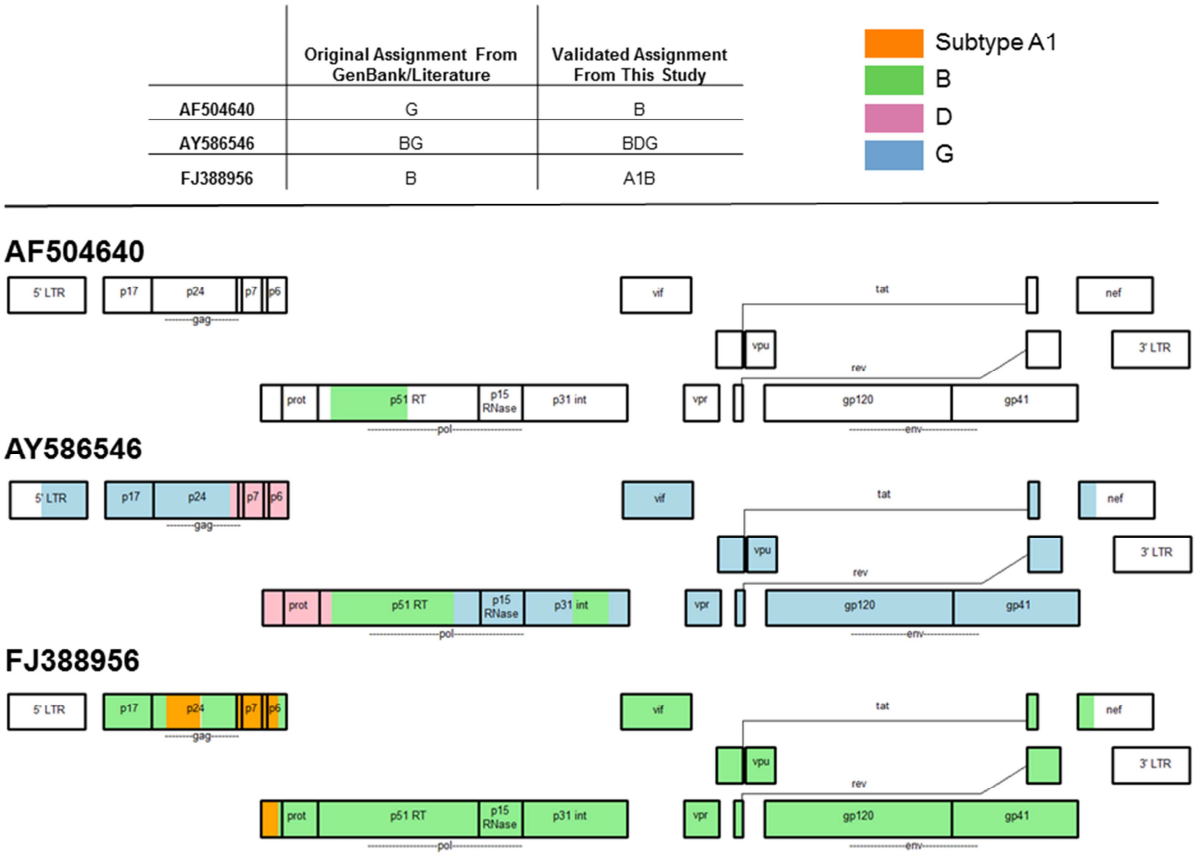


Figure 3.1. HIV-1 mis-genotyped cases in full-length genomes and fragment sequences.

Mis-genotyping is defined as a scenario in which a sequence's original genotyping assignment by GenBank/publication is inconsistent with results of jpHMM and the phylogenetic analyses, as described in the Method section. AF504640 is an incomplete genome example in which the

original G assignment was found to be of subtype B. In full-length genomes AY586546 and FJ388956, our results identified more complex genomic composition than their original subtype assignment; extra parental subtype was found in each sequence.

Results

A total of 459,881 sequences sampled worldwide were retrieved initially from GenBank and the Los Alamos HIV Sequence Database. The sequences were filtered to avoid sequence redundancy and to ensure sequences with adequate length for the downstream analysis. The resulting sequences were subjected to genotyping analyses by jpHMM and phylogenetic method as described in the Methods section. 2596 sequences with unique patient identification numbers were found to be mis-genotyped. Here we define mis-genotype as a scenario in which the original GenBank/publication genotype assignment is inconsistent with our jpHMM and phylogenetic results.

Figure 3.1 demonstrates three examples of mis-genotyped sequences. Incomplete genome sequence AF504640 was genotyped as subtype G in its original publication/GenBank assignment (GenBank-AF504640). We corrected its genotype as subtype B through jpHMM and further validation by phylogenetic analysis. Full genome sequence AY586546 was re-genotyped as BDG, compared to its original BG assignment (GenBank-AY586546). Full genome sequence FJ388956 was found to be of A1B recombinant, rather than a B subtype sequence (GenBank-FJ388956 , Kousiappa, Van De Vijver, and Kostrikis 2009).

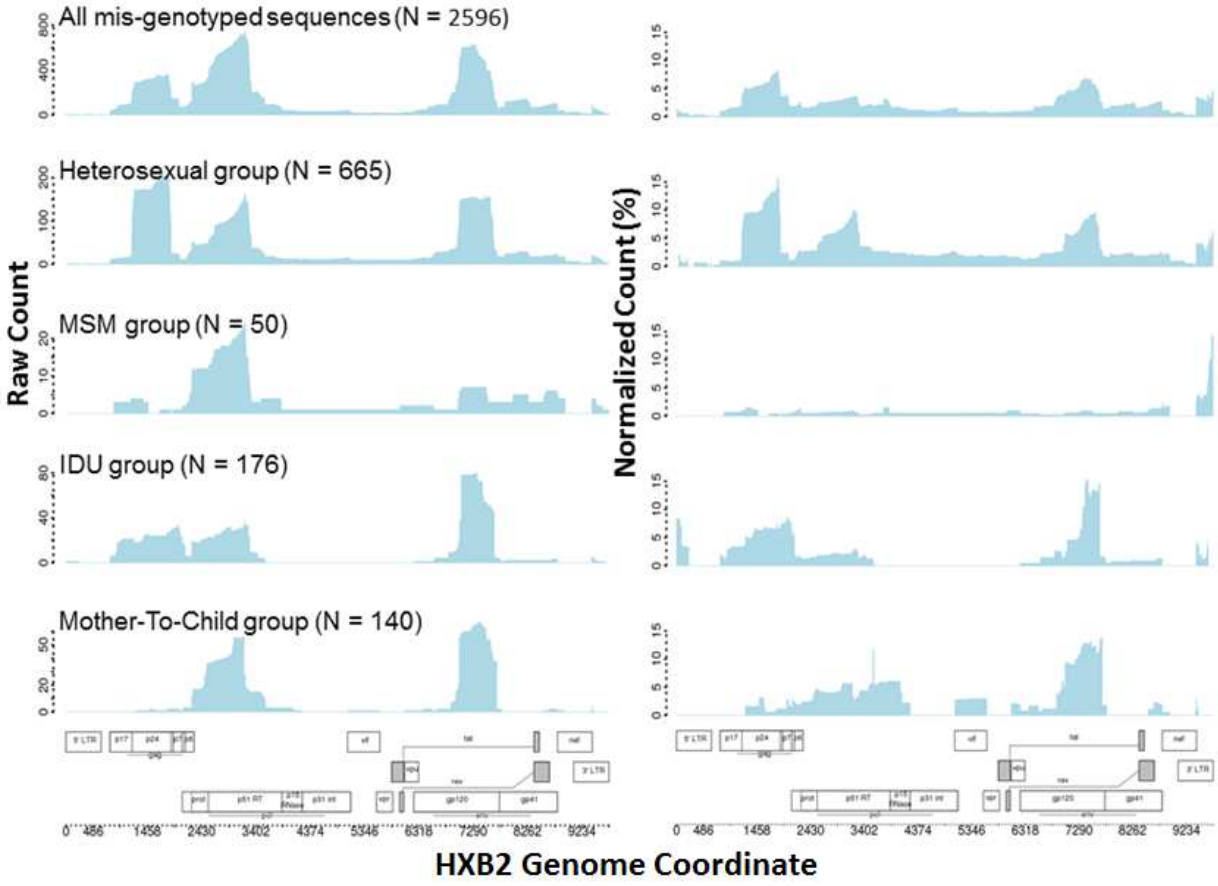


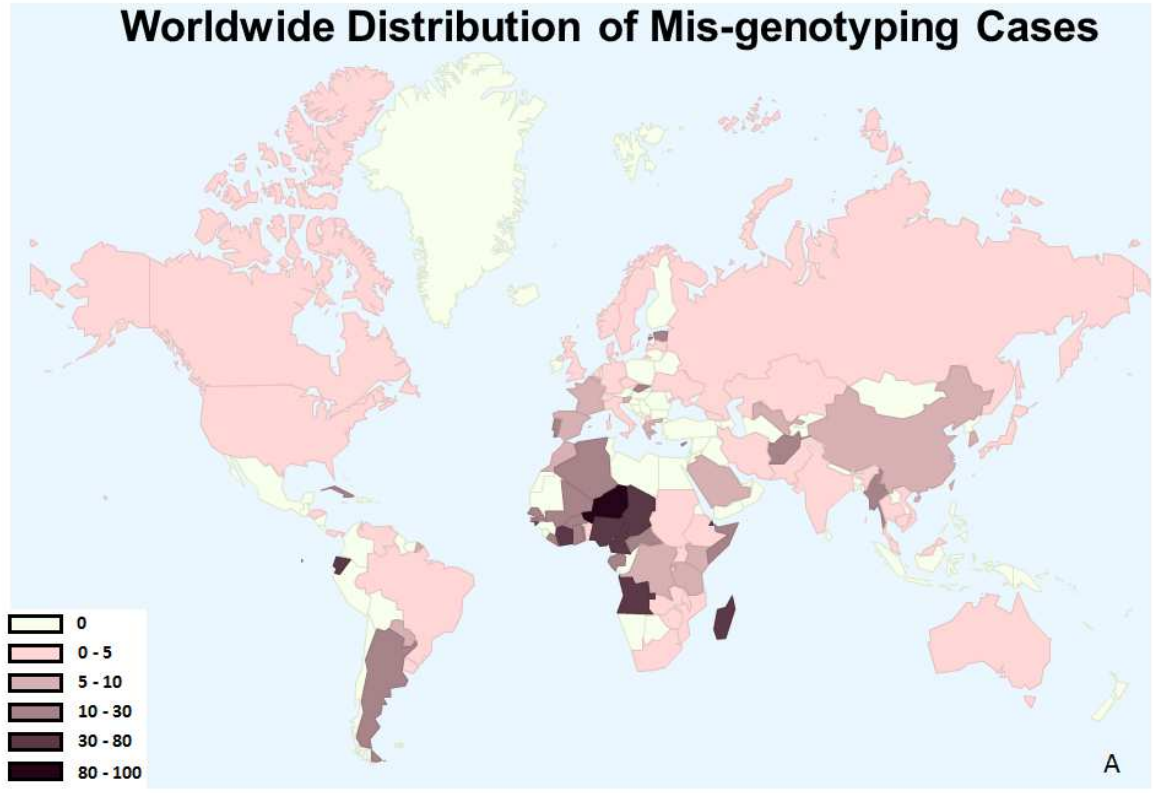
Figure 3.2. Comparison of four risk factor groups in terms of the mis-genotyping quality across HIV-1 genome.

Out of 2596 mis-genotyped sequences, 1031 sequences that have adequate transmission information were classified into four risk factor groups according to their derived routes of transmission. The sequences have been filtered to remove the sampling bias caused by multiple sequences per patient redundancy. Further, to avoid the bias imposed by unequal sampling across the genome, the counts of mis-genotyped sequences were normalized by the total number of sampled sequences in the same genomic region and route of transmission compared. The count of mis-genotyped sequences in each risk factor group is depicted here. Left: sequence

number based on raw count. Right: Normalized count. The reference genome shown at the bottom is based on the HXB2 strain (Accession number: K03455). Overall the mis-genotyping percentage is 3-6% across the genome, peaked at gag and env, followed by pol.

To estimate the impact of mis-genotyping on the HIV global epidemic, we further analyzed the identified mis-genotyped sequences stratified by risk factor and sampling geographic region. There were 1230 sequences with risk factor information available. The sequences associated with four risk factors were examined in terms of genotyping quality across the viral genome. The four risk factors are the most common routes of HIV transmission that include, heterosexual transmission, men who have sex with men (MSM), intravenous drug use, and children from mother-to-child transmission (the data of the pairing mothers was inadequate). To avoid biases imposed by unequal sampling across the genome, the counts of mis-genotyped sequences were normalized by the total number of sampled sequences in the same genomic region and route of transmission compared. Figure 3.2 outlines the count of mis-genotyped sequences in each risk factor group, based on the raw count (left) and normalized count (right), respectively. Overall the mis-genotyping percentage was 3-6% across the genome, primarily occurring on frequently sampled/sequenced region of *gag*, *pol, and env* (Fig. 3.2 top right panel). Similar genomic peak distribution was found across all risk groups (Fig. 3.2 left panels). We also observed from the normalized data that the MSM group had the least mis-genotyping cases (Fig. 3.2 the 3rd panel on right). This result is consistent with another observation in which the B subtype was found to be the most frequently mis-genotyped subtype within the MSM group. With MSM primarily occurring in Western countries, this is an indirect indication of higher genotyping capability within Western countries.

In the stratified analysis of mis-genotyped sequences by sampling geographic region, all sequences were selected by one sequence per patient and normalized by the total number of available HIV-1 sequences in the same geographic region. The prevalence of mis-subtype sequences in all countries is presented in Figure 3.3A. Further stratification results by transmission risk factor in each geo-region were summarized in Table 3.1. Overall, worldwide sampled sequences contained 5.28% genotyping errors (2596 sequences out of 49,175 sequences with unique patient ID). The geo-regions that topped the mis-genotyping list were sub-Saharan Africa, South America, East Asia, and Western Europe, where sequences were more frequently sequenced due to high prevalence of HIV epidemic in these regions. West Central Africa contained the highest density of mis-genotyped sequences within the heterosexual group. Frequent mis-genotyping within the mother-to-child group was observed in East Africa, Cuba, Argentina, China, and Portugal. Argentina, China, Taiwan, Myanmar, Greece, Portugal, and Uzbekistan topped the mis-genotyping countries within the IDU group (Fig. 3.3B).


Worldwide Distribution of Mis-genotyping Cases
0
0 - 5
5 - 10
10 - 30
30 - 80
80 - 100
A

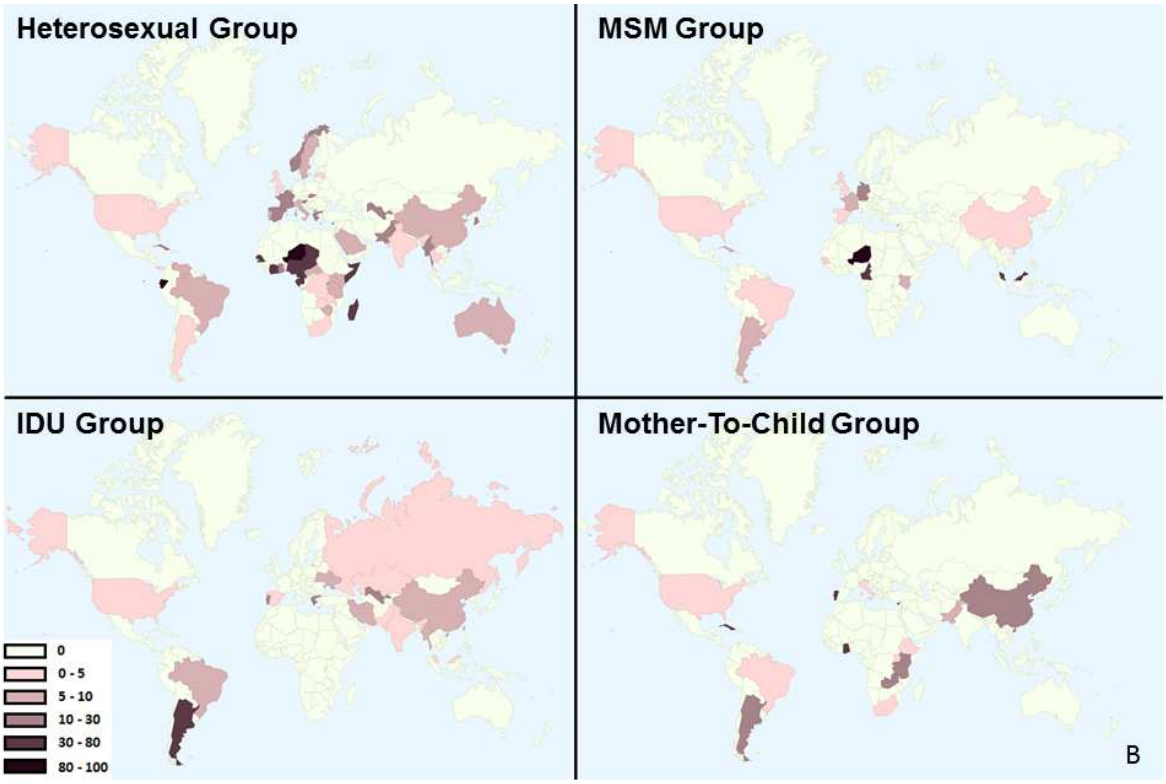
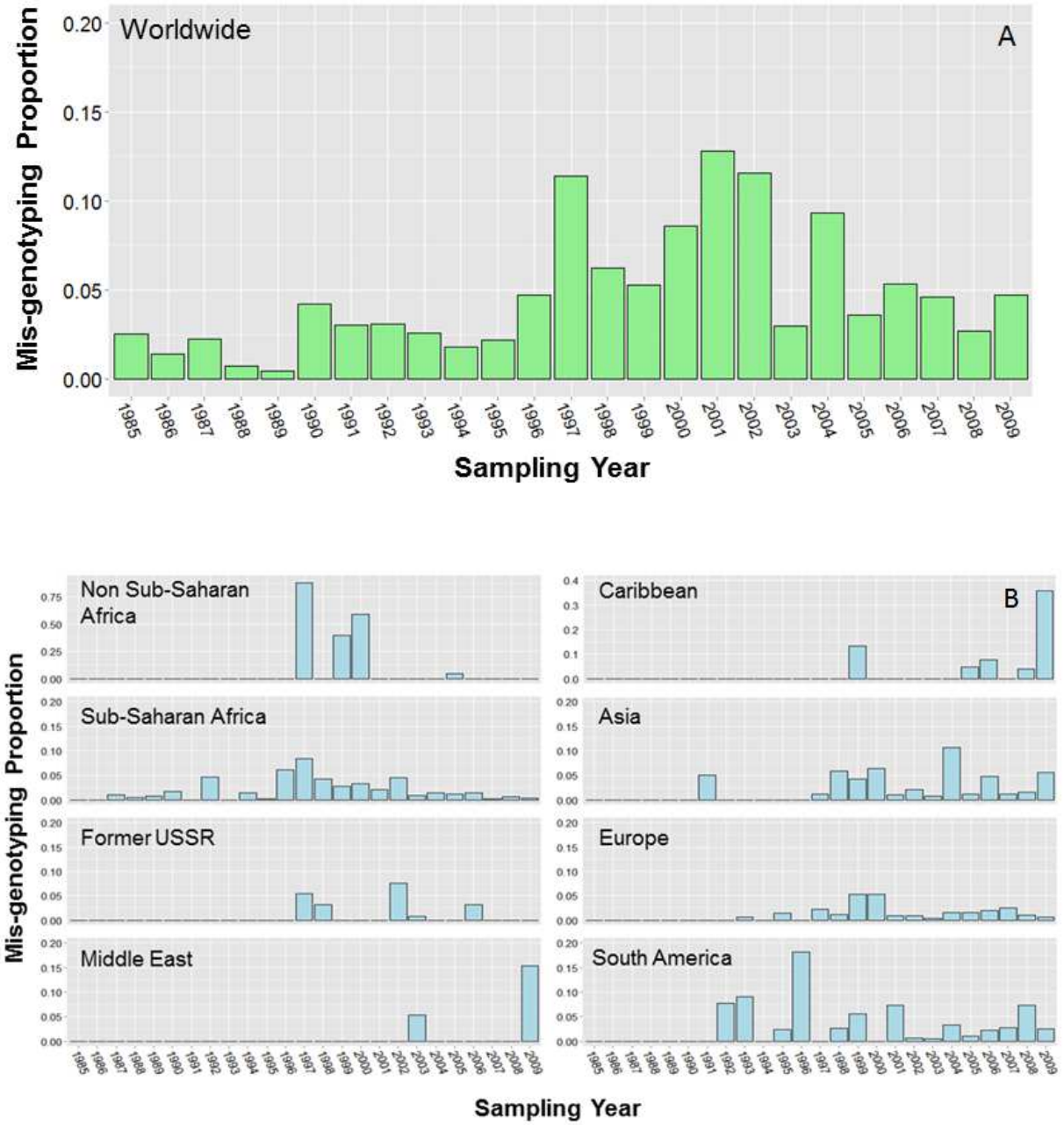

Heterosexual Group
MSM Group
IDU Group
Mother-To-Child Group
0
0 - 5
5 - 10
10 - 30
30 - 80
80 - 100
B

Figure 3.3. Distribution of HIV-1 mis-genotyped cases.

All sequences were normalized by patient and geographic regions as described in the Method section. (A) Worldwide overview of genotyping quality. (B) Genotyping quality in four risk groups. Color scale: percentage of mis-genotyping in each geographic region.



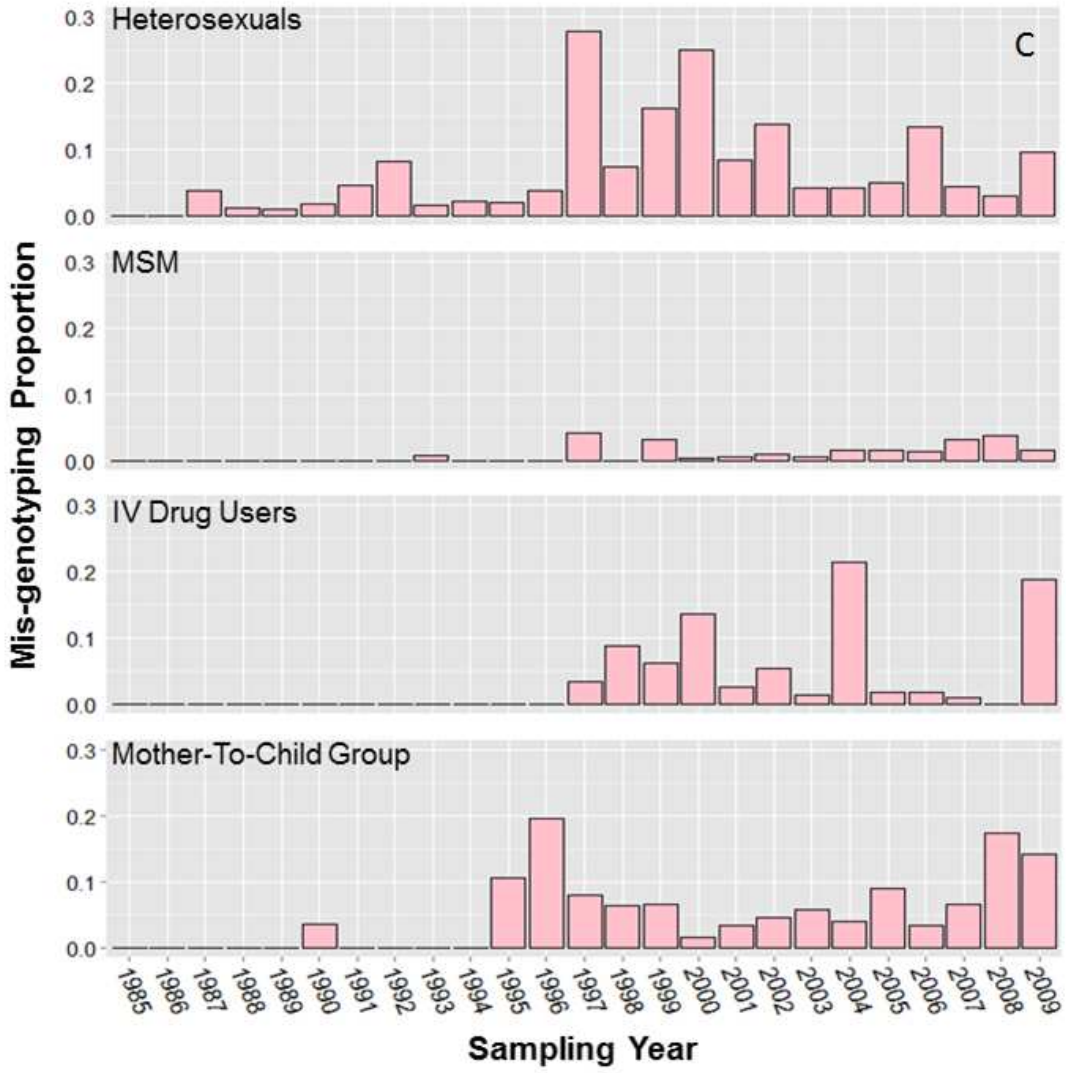


Figure 3.4. Genotyping quality over sampling years.

All sequences were normalized by patient and geographic region as described in the Method section. (A) Worldwide overview. (B) Genotyping quality in various geo-regions. (C) Genotyping quality in four risk groups.

Table 3.1. Genotyping quality evaluation in four risk factor groups from various geo-regions.

All sequences were selected based on one sequence per patient to avoid sampling redundancy. Two numbers are provided for each stratified level. Numbers inside the parenthesis: percentage of mis-genotyping cases; numbers outside of the parenthesis: raw count.

	Heterosexual	**MSM**	**IV Drug Users**	**Mother To Child**
North America	3 (1.00%)	1 (0.10%)	1 (0.40%)	1 (0.50%)
Former USSR	2 (2.10%)	0 (0.00%)	19 (3.50%)	0 (0.00%)
Caribbean	12 (6.20%)	6 (8.70%)	0 (0.00%)	6 (18.20%)
Europe	83 (8.80%)	18 (0.80%)	38 (3.20%)	16 (11.50%)
South America	26 (6.60%)	9 (3.20%)	20(14.80%)	39 (9.20%)
Sub-Saharan Africa	308 (9.10%)	6 (5.10%)	0 (0.00%)	76 (7.80%)
Oceania	1 (9.10%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Non sub-Saharan Africa	188 (75.20%)	1 (100%)	0 (0.00%)	0 (0.00%)
Asia	38 (4.20%)	9 (2.90%)	96 (5.30%)	2 (1.20%)
Central America	1 (0.80%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Middle-East	3 (4.10%)	0 (0.00%)	2 (4.30%)	0 (0.00%)

We also assessed the impact of mis-genotyping on HIV epidemics from 1985 to 2009 (Fig. 3.4). The global mis-genotyping percentage was approximately 4-5% over all time except the 1997-2002 period showing more mis-genotyped cases (Fig. 3.4A). The increase of mis-genotyped cases during 1997-2002 was likely attributed to more sequences being identified in South America and non sub-Saharan Africa regions (Fig. 3.4B), and primarily from the heterosexual group (Fig. 3.4C). For the heterosexual group, mis-genotyped cases sampled during 1997-2002 were mainly of A1 subtype from Africa, in particular from Niger, Senegal, Gabon, and Cameroon. Mis-genotyping peaks for the IDU group during 1998, 2000, and 2004 were mostly from China. For the mother-to-child group, the highest mis-genotyping portion was observed in 2008 and was mostly related to subtype B from Argentina. South America was also found to contain a high frequency of mis-genotyped B subtype from Brazil during 1992 and 1996. A mis-subtyping peak was noted in the Caribbean region during 1999, primarily associated with D subtype identified from Cuba.

Discussion

Our study represents a comprehensive assessment of the worldwide genotyping quality for publicly available HIV-1 sequences. Despite the importance of accurate genotyping in HIV global surveillance and epidemic track, the reliability and validity of the genotyping quality provided by individual research groups are often unchecked. Inconsistent genotyping criteria and misinterpretation by submission groups can lead to the persistent presence of mis-genotyped sequence entries. Genotyping is further compounded by our evolving knowledge of HIV. The genotyping information of sequences identified in the early stage of HIV/AIDS study (1980s-

1990s) may not have been updated after new subtypes have been identified/revised as documented in the most recent HIV nomenclature (Robertson et al. 2000).

In this study we investigated the genotyping information of all published HIV-1 sequences and compared them with results derived from our stand-alone genotyping pipeline, which consisted of phylogenetic analysis, distance measures and profile analysis imbedded in the jpHMM program. Mis-genotyping is defined as a scenario in which a sequence's original genotyping assignment within the GenBank/publication was inconsistent with the consensus of phylogenetic and jpHMM analysis. We assessed the degree of the mis-genotyping across the virus genome and further stratify sequences by risk factor, geographical region, and sampling year. All sequences were normalized to avoid sampling bias. We observed an uneven genotyping quality in all examined groups. Across the HIV-1 genome, overall mis-genotyping percentage was 3-6%, *p24, gp120,* and *pol* RT are regions most frequently mis-genotyped ([more] *p24* > *gp120* > RT [less]). This order differed in individual stratified risk factor groups, indicating a varied degree of genome complexity for each risk factor group. Of note, the mother-to-child group contains a high mis-genotyped percentage in *gp120* (10% cases were from *gp120* regions outside of V1V2), suggesting a knowledge gap may exist pertaining to co-receptor and N-glycosylation site changes in *gp120*.

Across the world, we found 5.28% mis-genotyped HIV-1 sequences. Most mis-genotyped cases are of non-B subtypes, but were mis-genotyped as other subtypes in their original GenBank and literature submission. The geographic region most influenced by mis-genotyping was West Central Africa, where the heterosexual group was more frequently mis-genotyped than the other three risk groups. Other risk groups requiring closer monitoring of genotyping quality includes: IDU group in South America, and IDU and mother-to-child groups

from China. It is also noted that some geographic regions, such as Cuba, with its low prevalence of HIV epidemic (Eduard J. Beck 2006) may lead to inexperience genotyping and a higher mis-genotyping rate.

Given all genotyping quality evaluation results, we propose the following reasons that could cause mis-genotyping. First, inconsistent genotyping standard being used prior to 2000-2002 would likely explain elevated occurrence of mis-genotyping prior to 2000. A sharp decrease of mis-genotyping was observed after the establishment of a consistent HIV nomenclature in 2000. Second, evolving HIV epidemics provided an increased opportunity for complex HIV strains to develop, making accurate genotyping more challenging. Third, limitation on genotyping capability would explain observed genotyping quality variations across different geographical regions. In West Central Africa, both old and contemporary HIV strains co-circulate (Zhang et al. 2010). As existing genotyping programs rely heavily on contemporary sequences as genotyping references, the old strains may be difficult to genotype correctly. This also raises a possibility that true HIV epidemics may have been underestimated.

Our study suggested the necessity of close monitoring of the genotyping quality in HIV epidemic research, especially in West Africa, Asia, and Caribbean. Further, results from this study would provide a more robust basis for case definition, prevention and control, additionally, for better understanding of clade differences in pathogenesis, disease progression, and drug resistance.

Conclusion

Consistent genotyping quality control is critical for monitoring HIV epidemics. The results presented here not only fill the knowledge gap in HIV genotyping quality control, but can also be used as reference information for improving global and regional HIV surveillance. Finally, the

approach described in this manuscript outlines a genotyping quality control pipeline that can be easily applied to other infectious disease epidemic studies.

Acknowledgement

We thank Ms. Gretchen Parrott for manuscript editing. Sources of support: University of Georgia Research Fund 10793GR002 and University of Georgia Faculty Research Grants in the Sciences Award 1021RX064536.

References

Abecasis, A. B., P. Lemey, N. Vidal, T. de Oliveira, M. Peeters, R. Camacho, B. Shapiro, A. Rambaut, and A. M. Vandamme. 2007. "Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form." *Journal of virology* 81 (16):8543-51. doi: 10.1128/JVI.00463-07.

Eduard J. Beck, Nicholas Mays, Alan W. whiteside, Jose M Zuniga. 2006. *The HIV pandemic : local and global implications*. United States: Oxford University Press.

Gao, F., D. L. Robertson, S. G. Morrison, H. Hui, S. Craig, J. Decker, P. N. Fultz, M. Girard, G. M. Shaw, B. H. Hahn, and P. M. Sharp. 1996. "The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is caused by an intersubtype (A/E) recombinant of African origin." *J Virol* 70 (10):7013-29.

GenBank-AF504640. http://www.ncbi.nlm.nih.gov/nuccore/AF504640.

GenBank-AY586546. http://www.ncbi.nlm.nih.gov/nuccore/AY586546.

GenBank-FJ388956. http://www.ncbi.nlm.nih.gov/nuccore/FJ388956.

J, Felsenstein. 1989. "PHYLIP - Phylogeny Inference Package (Version 3.2)." *Cladistics* (5):164-166.

John-Stewart, G. C., R. W. Nduati, C. M. Rousseau, D. A. Mbori-Ngacha, B. A. Richardson, S. Rainwater, D. D. Panteleeff, and J. Overbaugh. 2005. "Subtype C Is associated with increased vaginal shedding of HIV-1." *J Infect Dis* 192 (3):492-6. doi: 10.1086/431514.

Kousiappa, I., D. A. Van De Vijver, and L. G. Kostrikis. 2009. "Near full-length genetic analysis of HIV sequences derived from Cyprus: evidence of a highly polyphyletic and evolving infection." *AIDS Res Hum Retroviruses* 25 (8):727-40. doi: 10.1089/aid.2008.0239.

Los-Alamos-HIV-Sequence-Database. "www.hiv.lanl.gov." www.hiv.lanl.gov.

Paraskevis, D., M. Magiorkinis, A. M. Vandamme, L. G. Kostrikis, and A. Hatzakis. 2001. "Re-analysis of human immunodeficiency virus type 1 isolates from Cyprus and Greece, initially designated 'subtype I', reveals a unique complex A/G/H/K/? mosaic pattern." *J Gen Virol* 82 (Pt 3):575-80.

Robertson, D. L., J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber. 2000. "HIV-1 nomenclature proposal." *Science* 288 (5463):55-6.

Schultz, A. K., M. Zhang, I. Bulla, T. Leitner, B. Korber, B. Morgenstern, and M. Stanke. 2009. "jpHMM: improving the reliability of recombination prediction in HIV-1." *Nucleic Acids Res* 37 (Web Server issue):W647-51. doi: 10.1093/nar/gkp371.

Schultz, A. K., M. Zhang, T. Leitner, C. Kuiken, B. Korber, B. Morgenstern, and M. Stanke. 2006. "A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes." *BMC Bioinformatics* 7:265. doi: 1471-2105-7-265 [pii] 10.1186/1471-2105-7-265.

Taveira, Ines Bartolo and Nuno. 2012. *HIV-1 Diversity and Its Implications in Diagnosis, Transmission, Disease Progression, and Antiretroviral Therapy*. Edited by Mahmut Caliskan, *Genetic Diversity in Microorganisms* InTech.

UNAIDS. 2012. UNAIDS Report on the Global AIDS Epidemic. UNAIDS.

Zhang, M., B. Foley, A. K. Schultz, J. P. Macke, I. Bulla, M. Stanke, B. Morgenstern, B. Korber, and T. Leitner. 2010. "The role of recombination in the emergence of a complex and dynamic HIV epidemic." *Retrovirology* 7:25. doi: 10.1186/1742-4690-7-25.

Zhang, M., A. K. Schultz, C. Calef, C. Kuiken, T. Leitner, B. Korber, B. Morgenstern, and M. Stanke. 2006. "jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1." *Nucleic Acids Res* 34 (Web Server issue):W463-5.

CHAPTER 4

ANALYZING MODULAR RNA STRUCTURE REVEALS LOW GLOBAL STRUCTURAL ENTROPY IN MICRORNA SEQUENCE[1]

[1] Timothy I. Shaw, Amir Manzour, Yingfeng Wang, Russel L Malmberg, Liming Cai. 2011. *Journal of Bioinformatics and Computational Biology*. 09:283.

Reprinted here with permission from the publisher.

Abstract

Secondary structure remains the most exploitable feature for non-coding RNA (ncRNA) gene finding in genomes. However, methods based on secondary structure prediction may generate superfluous numbers of candidates for validation and have yet to deliver the desired performance that can complement experimental efforts in ncRNA gene finding. This paper investigates a novel method, unpaired structural entropy (USE) as a measurement for the structure fold stability of ncRNAs. USE proves to be effective in identifying from the genome background a class of ncRNAs, such as precursor microRNAs (pre-miRNAs) that contains a long stem hairpin loop. USE correlates well and performs better than other measures on pre-miRNAs, including the previously formulated structural entropy. As a support vector machine classifier, USE outperforms existing pre-miRNA classifiers. A long stem hairpin loop is common for a number of other functional RNAs including introns splicing hairpins loops and intrinsic termination hairpin loops. We believe USE can be further applied in developing ab initio prediction programs for a larger class of ncRNAs.

Keywords: non coding RNA; microRNA; Shannon Entropy.

Introduction

Non-coding RNAs (ncRNAs) carry out many critical functions in living cells (Griffiths-Jones 2007). As more functional roles by ncRNAs are being discovered, there is a rapid growth of interest in developing bioinformatics methods that may effectively identify ncRNAs genes from genomic backgrounds. Unlike their protein coding counterparts, ncRNAs do not possess strong statistical signals (e.g., open reading frames), making ncRNA identification a computational challenge. For example, programs based on relevant sequential features, such as base composition, are often limited to certain classes of organisms or specific families of ncRNAs(Schattner 2003). On the other hand, most transcribed single-strand ncRNAs can potentially fold; their secondary structure is the most exploitable feature for a truly successful ncRNA prediction methods. Indeed, such a potential has energized the use of secondary structure prediction methods for ncRNA finding(Pedersen et al. 2006), (Rivas and Eddy 2001), (Washietl et al. 2007). Nonetheless, prediction results have generated a rather unclear picture; in particular, RNAz(Washietl, Hofacker, and Stadler 2005) and EvoFold (Pedersen et al. 2006) generated 30,000 and 48,000 predicted candidates, respectively, on the human and other vertebrate genomes; the overlap of these two sets of candidates is disappointingly small (7.2%) (Washietl et al. 2007) indicating a low-sensitivity (for at least one of the programs). In addition, the validation rates for the predicted structural RNAs are low, possibly attributable to high false-positive rates because of the rare low expression levels in known ncRNAs.

The underperformance of structure prediction based methods could be due to the fact that functional RNA may admit alternative structures and random sequences may fold and be detected (e. g. 11 million hairpin loops were found within the human genome) (Bentwich et al. 2005). This suggests that, in addition to the minimum free energy (Clote et al. 2005), other rules

governing secondary structure folding may be needed for effective ncRNA finding. Fold stability may be one such characteristic as it could help understand the differences between alternative folds of real ncRNAs and between folds of real ncRNAs and of random sequences. Based on the partition functions (McCaskill 1990) that define a thermodynamic energy ensemble of RNA secondary structures, the fold stability of a given RNA sequence can be measured using Shannon's entropy (McCaskill 1990, Dirks et al. 2007, Shannon 1997a) over various random variables that define base pairings of the sequence. It turns out that some ncRNAs, typically precursor microRNAs (pre-miRNA), have entropy significantly lower than that of their randomly shuffled counterparts, while others do not (Freyhult, Gardner, and Moulton 2005). Independent studies on other measures, such as average free energy (Bonnet et al. 2004), self-containment (Lee and Kim 2008), compactness (Loong and Mishra 2007), and thermodynamic entropy (Zuker 2009), appears to confirm that precursor miRNAs possess much higher fold stability than other kinds of ncRNAs and such structural characteristics may be exploited to discriminate miRNAs from the genome background.

MicroRNAs are endogenous small noncoding RNAs that function as a regulator and anti-viral defenders within animals, plants, and viruses (Zhang, Pan, et al. 2006, Scaria et al. 2006, Ambros 2004). The transcript containing the microRNA is processed into a short 22nucleotide mature microRNA and incorporated into a RNA Induced Silencing Complex (RISC). The RISC can then forms a duplex on the target mRNA or viral sequence enabling endonucleolytic cleavage, degradation and translational repression. Within precursor miRNAs, the secondary structure requirement for RNase III drosha and dicer processing consists of a long stem loop with low fold stability entropy (Lee et al. 2003, Zeng and Cullen 2004). Since such a structural feature is shared by other functional RNAs such as snRNA, introns splicing hairpins loops and

intrinsic termination hairpin loops (Smith et al. 2000, Gusarov and Nudler 1999), it is of acute interest to develop structure stability based methods that can effectively detect such ncRNAs from genomes, especially given the recent discoveries of important roles played by miRNAs(Ambros 2004).

In this paper, we present some preliminary results toward developing an ab initio ncRNA prediction framework based on structure stability of long hairpin stem loops. We propose a novel objective function called Unpaired Structural Entropy (USE), which captures the structural variability for a given sequence. The USE measure was found to be effective in distinguishing miRNAs from its genomic background as well as other ncRNAs. Through the USE objective function, we were able to create a single feature classifier to distinguish miRNAs with a sensitivity of 85% and specificity of 90%, an improvement upon all existing multi-feature miRNA classifiers including the previously investigated structural entropy. Finally, we included the USE along with existing RNA measurements to further improve the performance of an SVM classifier.

Methods

Although the minimal free energy is generally chosen as the predicted structure for a given RNA sequence, many alternative structures also exist. The probability for each of these structures to occur can be calculated through the Boltzmann Partition Function (Zuker 2009) and thus it is possible to calculate the likelihood for base pairings between nucleotides. In this work, we introduce a novel method USE which measures the structure's stability through computing the entropy of the non-pairing probabilities.

Structural Variation

Here we consider structural variation to be the amount of potential ways into which a sequence may fold. Higher structural variation implies a higher number of potential foldings. Figure 4.1 displays two different NUPACK7 RNA folds, Figure 4.1a shows a predicted folding for mir-32 while figure 4.1b shows the folding for a dinucleotide shuffled mir-32. Although both of them possess a folded structure, the coloring scheme shows more confidence in the pairings for mir-32 than the shuffled sequence.

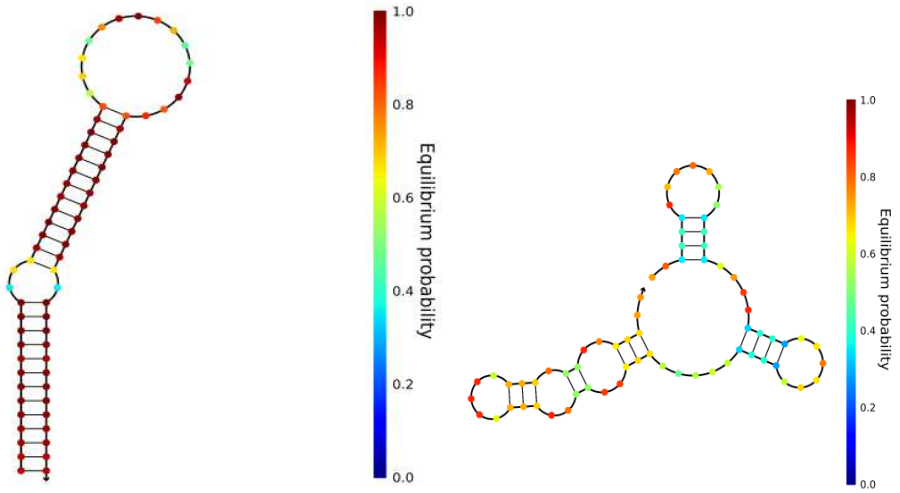


Figure 4.1. Structures predicted by NUPACK for pre-mir-32 (Left). Structures predicted by NUPACK for dinucleotide shuffling of pre-mir-32 (Right). The equilibrium probability was computed by NUPACK.

Base Pairing Probabilities

RNA can exist in an ensemble of structures, and the distribution of these structures can be captured by a Boltzmann distribution. The Boltzmann distribution can allow computation of the

partition function (Z) for each substructure. The Partition Function algorithm has been implemented by McCaskill (McCaskill 1990) and it calculates the base-pairing probability distribution based on the free energies for each structure within the structural ensemble space **Ω**. Let s_α be a structure with free energy G_α. Assume that the molar gas constant R=8.31451Jmol-1K-1, and the temperature T, then the partition function is defined in equation 1.

$$Z = \sum_{s_\alpha \in \Omega} e^{-G_\alpha / RT}. \quad (1)$$

The probability for each structure to occur is the following

$$p(s_\alpha) = \frac{1}{Z} e^{-G_\alpha / RT} \quad (2)$$

The term δ_{ij}^{α} denotes the occurrence of pairing between nucleotides i and j in s_α. Hence, the probability base pairing probability P_{ij} is as follows:

$$p_{ij} = \sum_{s_\alpha \in \Omega} p(s_\alpha) \delta_{ij}^{\alpha} \quad (3)$$

Where p_{i0} corresponds to the non-pairing probability of nucleotide at position i:

$$p_{i0} = 1 - \sum_{j=1}^{N} p_{ij} \quad (4)$$

with *N* being the length of the sequence.

Unpaired Structural Entropy

Shannon Entropy is one of the most fundamental and basic concepts in the field of Information theory; it measures the amount of uncertainty of values taken by a random variable. It also measures the amount of diversity that exists within a set of quantities. In this work, we propose

the Unpaired Structural Entropy (USE) which computes the entropy of the non-pairing probabilities of the nucleotides that are normalized across the sequence:

$$USE(S) = \sum_{i=1}^{N} -\frac{p_{i0}}{L_0} log \frac{p_{i0}}{L_0} \, , where \; L_0 = \sum_{i=1}^{N} p_{i0} \quad (6)$$

Previous attempts have been made to capture the structural variability of a sequence through the entropy of its base pairing probabilities. Huynen et al.(Huynen, Gutell, and Konings 1997) have defined the positional entropy (Q) as follows which has been traditionally used in previous research(Freyhult, Gardner, and Moulton 2005):

$$Q(S) = \frac{1}{N} \sum_{i=1}^{N} E(n_i) \, , where \; E(n_i) = -\sum_{j=0}^{N} p_{ij} \log p_{ij} \quad (7)$$

The relationship between Q and USE is as follows:

$$Q(S) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=0}^{N} p_{ij} \log p_{ij} = -\frac{1}{N} \sum_{i=1}^{N} p_{i0} \log p_{i0} - \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} p_{ij} \log p_{ij}$$

$$= \frac{L_0}{N} USE(S) + \log(L_0) + \frac{2L_1}{N} \varphi(S) + \log(L_1)$$

$$where \; \varphi(S) = -\sum_{i=1}^{N} \sum_{j>i}^{N} \frac{p_{ij}}{L_1} \log \frac{p_{ij}}{L_1} \log \frac{p_{ij}}{L_1}, L_1 = \sum_{i=1}^{N} \sum_{j>i}^{N} p_{ij} \quad (8)$$

In Q(S), the entropy of pairing possibilities is calculated for each nucleotide. These individual nucleotide entropies are then averaged across the sequence. To calculate USE, only the non-pairing probabilities of nucleotides are taken into consideration. These probabilities are

normalized across the sequence before their entropy is calculated; therefore, unlike Q(S), USE(S) is a global computation of the structural entropy. In this study, we will show that USE is more successful than Q as well as other structural features in capturing the stability of long stems structures, in particular pre-miRNAs.

Results

A number of tests and analyses were conducted to examine the capabilities of the USE measure in identifying structures with low variability. (1) The USE score was studied and compared across different RNA families. The USE score distributions for miRNAs were significantly lower than other RNA families. (2) We then examined the USE score's ability to distinguish miRNAs from their genomic background. (3) Finally, the performance of the USE score as a classifying feature in distinguishing miRNA from pseudo-miRNA was assessed. The USE score was shown to be highly effective in classifying miRNA from pseudo-miRNA.

USE across RNA Families

To investigate the utility of USE, we compared the entropy across different RNA families using the same sequences evaluated by Freyhult et al.(Freyhult, Gardner, and Moulton 2005) USE quartile boxplot of different ncRNA families can be viewed from Figure 4.2a. The USE score for miRNA was generally much lower than of the other RNA families. The intron sequences possessed the next lowest entropy values. Based on the Youden Index, a cutoff of 0.82 the USE score maximized the Youden index separating microRNA USE calculation to that of intron sequences. The graph showed that roughly around 0.82 we have 75% of the miRNA lower than this value and 75% of the intron sequences higher than this value. Across the ncRNAs, the order of USE values from smallest to largest: miRNA, SRP, intron, snRNA, tRNA, riboswitch, RNase, telomerase, snoRNA, shuffled, tmRNA and mRNA. In addition from Figure 4.2b, we included

the Q calculation across different RNA families. This feature also showed that miRNA tends to be lower than of the other RNA families; however, if we compared the two graphs, Q as an objective function cannot distinguish miRNA as well from the other families of ncRNA than that of the USE score.

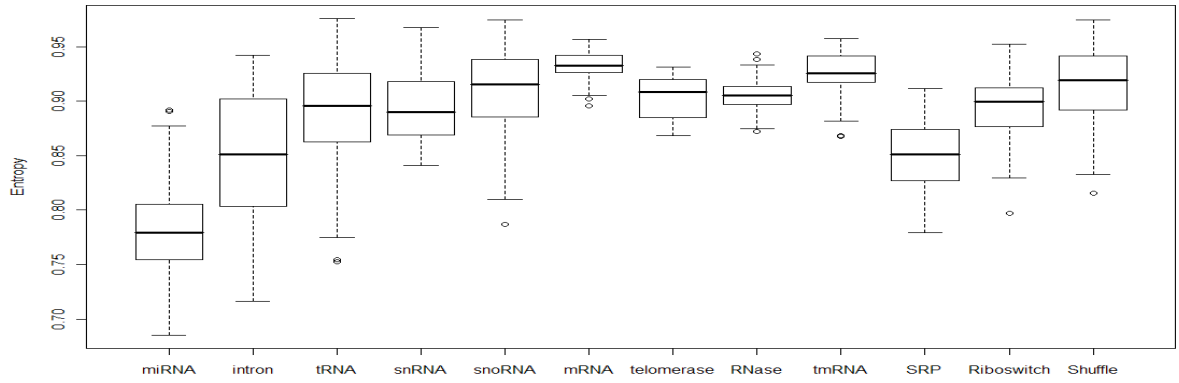


Figure 4.2A. A box plot of USE distribution across various ncRNAs.

Box and whisker plots displaying distribution of USE score through quartiles across various ncRNAs. From the graph, the low entropic feature calculation separates miRNA from the other ncRNAs.

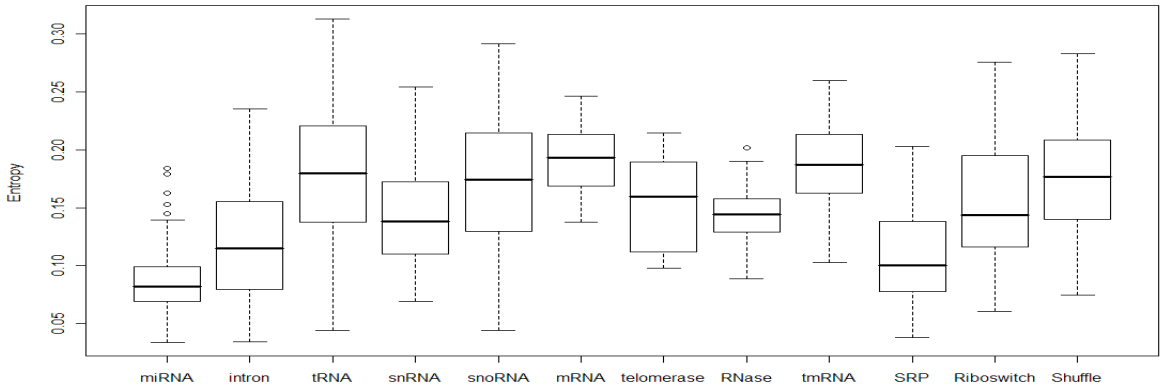


Figure 4.2B. A box plot showing the Q distribution across various ncRNA.

We also observe that miRNA tends to have a lower entropy; however compared to USE, Q has a harder time distinguishing miRNAs from the other families of ncRNAs.

Detecting pre-microRNAs

We explored the ability of the USE function to identify miRNA precursors by sliding a window across a sequence, then calculating a score within the window. The lengths of known precursor microRNAs (pre-miRNA) in humans usually range from 70 to 100nt. Therefore, we evaluated the behavior of the USE score across different window-sizes and sequences surrounding the precursor miRNA. Figure 4.3 shows the USE Score of Sliding Window Scan of 500nt upstream and downstream of a human miRNA (mir-30e) the actual window size was the same as the pre-miRNA. The graph indicates that the USE score succeeds in distinguishing the low entropy for the true miRNA in its real genome context.

Since the length of the pre-miRNA sequence was not always known, we performed sliding window scans of different length for each sequence containing the upstream and downstream of 721 pre-miRNAs. To observe the behavior or USE on surrounding sequences of miRNA, we varied the window size by increasing and decreasing in increments of 5 nt and repeated the same process for all pre-miRNA sequences. Figure 4.4 presented the results of USE scores corresponding to different window lengths and positions, which were averaged across the 721 pre-miRNAs. We showed that for any length of window scan, the lowest average USE values always occurs at the position 0 which correspond to the exact location of the pre-miRNA within the genome.

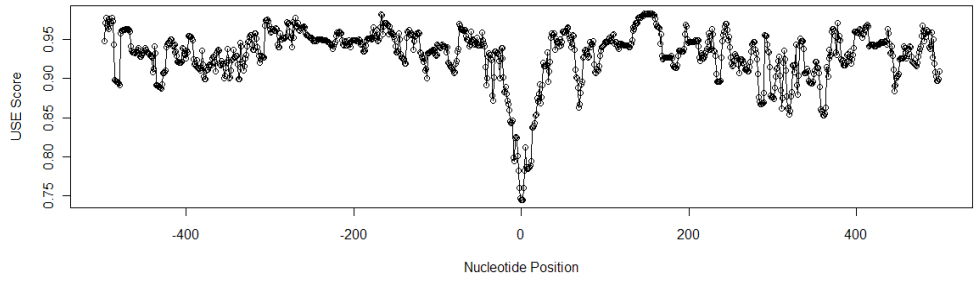


Figure 4.3. Sliding window scan of USE score across 500nt upstream and downstream of has-mir-30e with window size 93nt (the length of the mir-30e sequence).

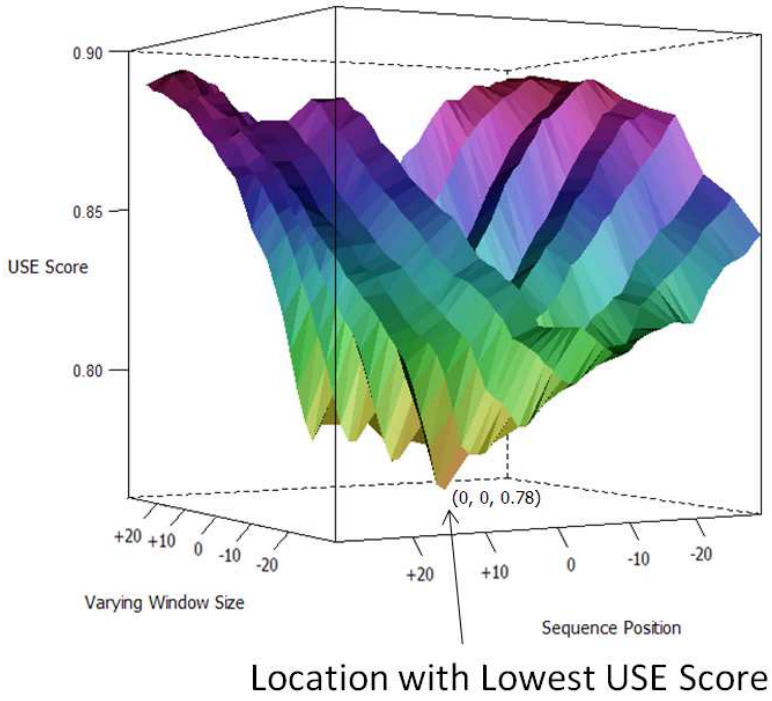


Figure 4.4. Average USE value of all miRNAs.

Any point on the graph corresponded to a specific window length and position and represents the USE score averaged across all 721 Human microRNA sequences. The labels on the

sequence-position axis represented the relative upstream/downstream position from the location of the actual microRNA. The window-size axis represented the amount of increments/decrements of the window-length relative to the length of the actual miRNA. The lowest averaged USE values were aligned in position 0 for all window lengths.

USE Correlation with other RNA measures on Human pre-miRNA

We calculated Pearson correlation coefficients for six variables to evaluate the correlation relationship between USE and various other RNA measurements including: Q, miR-CYK, Self Containment (SC), Length, Minimal Free Energy (MFE), and Structural Ensemble as shown in figure 4.5.

We used the Cocke-Younger Kasami (CYK) algorithm to develop an in house CYK program to perform microRNA gene finding called miR-CYK. CYK in general was used to find the maximum probability alignment of the CFG to the string. Therefore, by defining a Stochastic Context Free Grammar (SCFG) based on the human pre-miRNA structure feature, we used the miR-CYK to score the sequence based on its predicted similarity to the miRNA structure.

Self-Containment (SC) was shown to measure the tendency to retain their structure regardless of the neighboring upstream and downstream sequences. This particular measurement was developed by Kim et al.(Lee and Kim 2008). They took the query sequence and added additional sequences upstream and downstream, and used RNAFold to fold the sequence to examine the structure prior and after the additional sequences. This was done repeatedly to obtain a statistic of the frequency for the structure to retain its shape.

MFE and Ensemble Frequency calculation was based on RNAFold's calculation(Gruber et al. 2008). Previously Bonnet et al.(Bonnet et al. 2004) showed that miRNA compared to other ncRNA tends to have lower MFE, and this might be attributed to the stability of the folding. Ensemble Frequency provided a score of the frequency of the specific structure to occur within the structural space.

Since we were interested in possibly using USE to perform pre-miRNA gene finding, we used the Human pre-miRNA as the sequence to make the comparison. Table 4.1 presented the correlation coefficient for these six measures. USE was shown to be most closely related to Q, which is not surprising since USE and Q were inherently based on the same type of idea. MiR-CYK and SC both possessed decent correlation to USE while Length, MFE, and Ensemble Frequency possessed the weakest correlation with USE.

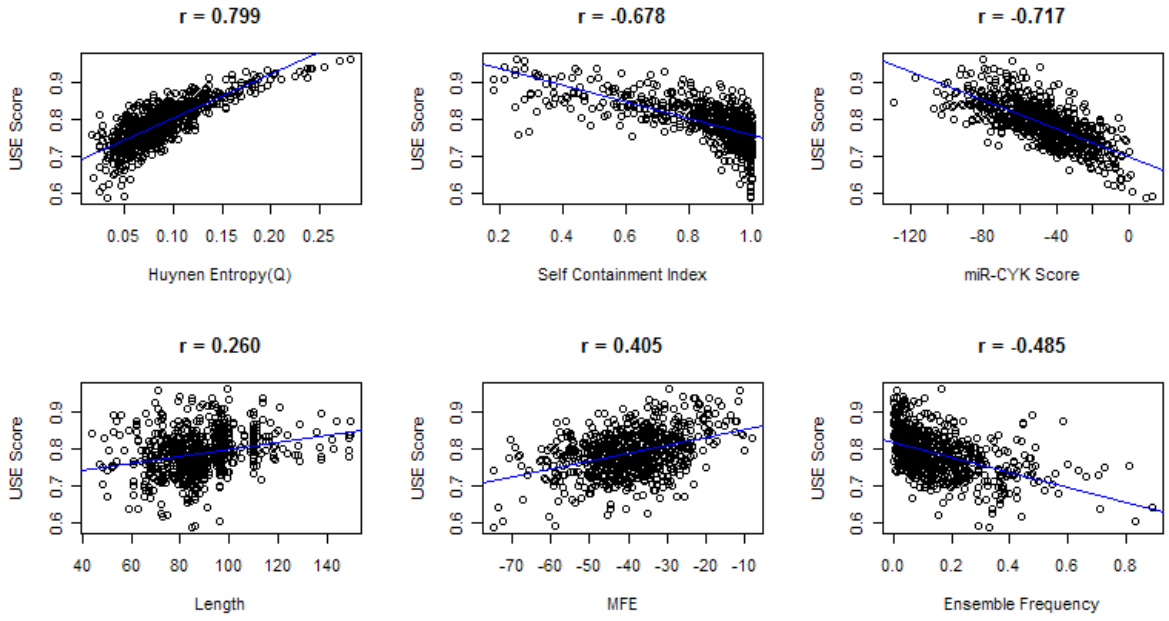


Figure 4.5. Correlation plot of the USE score compared to Q, Self Containment Index, miR-CYK, Length, MFE, and Ensemble Frequency.

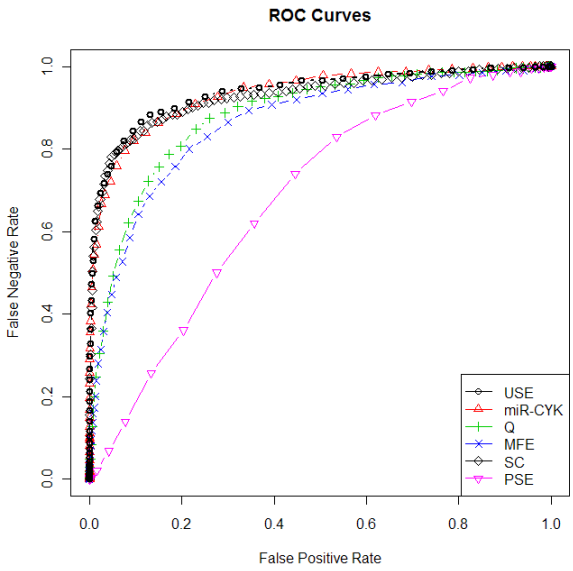


Figure 4.6. ROC plot of for prediction of classifying miRNA through various RNA measurements USE, CYK, Q, MFE, and SC.

Table 4.1. R2 between USE and other RNA Measures.

Measure	Correlation with USE
Q	0.638
SC	0.46
miR-CYK Score	0.514
MFE	0.068
Length	0.164
Frequency in Ensemble	0.235

Classifying Human pre-miRNA via USE

We also evaluated the performance of the USE function as a classifier. To do so, we used 721 Human miRNAs as the positive datasets and 8494 pseudo miRNA as the negative datasets. The Human miRNA sequences were downloaded from miRbase(Griffiths-Jones 2006). The negative set of pseudo-miRNAs, was obtained from Xue et al(Xue et al. 2005) derived from a set of sequences with hairpin loops located within the coding region.

The ROC for various RNA objective functions is plotted in figure 4.6. We also included a Paired Structural Entropy (PSE). The probability of each nucleotide to pair can be defined as (1 – Pi0), and PSE was the same entropy calculation as USE but calculated only on the paired nucleotide region. Figure 4.6 showed that the entropy of the non-pairing probability has more classification power than that of the pairing probability. A true positive was defined as a Human microRNA that was below the cutoff, and a false positive was defined as a Human microRNA that is above the cutoff. A true negative was defined as a pseudo microRNA that was above the cutoff, and a false negative was defined as a pseudo microRNA that was below the cutoff. The graph showed that USE, CYK, and SC's performances were relatively similar.

Comparisons across different microRNA Classifiers

Most existing miRNA prediction programs rely on a machine learning algorithm trained with a variety of features in primary and secondary structure for classification. To assess the power of the USE classifier, we compared its performance to four other SVM microRNA classifiers: TripletSVM (Xue et al. 2005), Virgo(Kumar, Ansari, and Scaria 2009) miRFinder(Huang et al. 2007), and microPred(Batuwita and Palade 2009). Triplet-SVM, an ab initio algorithm uses the local contiguous base-pairing structures as features for the SVM classification(Xue et al. 2005).

Virgo, a viral miRNA detector that was trained on human miRNA sequences(Kumar, Ansari, and Scaria 2009). miRFinder used the pre-miRNA structural characteristic and structural mutation information for the classification(Huang et al. 2007). MicroPred attempted to improve the prediction through effective machine learning techniques(Batuwita and Palade 2009). To allow a valid fair comparison, we used the 8494 pseudo-microRNA as the negative dataset. This particular negative dataset was used in all of the programs. Using a larger negative dataset satisfied a requirement of the miRNA gene finder that it not produce many false positives, one of the primary difficulties in miRNA detection. To evaluate each program's performance, we chose to use Sensitivity, Specificity, Youden Index, and the Mathews Correlation Coefficient which were calculated:

$$\text{Specificity} = TN / (TN + FP)$$

$$\text{Sensitivity} = TP / (TP + FN)$$

$$\text{Youden Index} = \text{Specificity} + \text{Sensitivity} - 1$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \qquad (9)$$

Table 4.2 contained the sensitivities, specificities, and MCC for the above classifiers. Two different cutoffs were chosen for the USE model, to demonstrate the diversity of the sensitivity and specificity over a range of cutoff values. The USE model cutoff 1 was a rough estimation based on the comparison over the distribution of USE score between miRNA and different RNA families (See Figure 4.2a). For cutoff 2, the threshold was more stringent than cutoff 1 in attempt to reduce the number of false positives. The SVM USE Model corresponded to an SVM model that integrated USE score, SC, and CYK, and we trained the SVM using cross

validation. We saw that the MCC for such a classifier was significantly higher than other classifiers.

Table 4.2. Performance comparison across different miRNA gene finding models.

	Sensitivity	Specificity	Youden Index	MCC
SVM with USE Model	0.777	0.974	0.751	0.724
USE Model (Youden)	0.865	0.889	0.754	0.652
USE Model (Cutoff 1)	0.845	0.906	0.751	0.560
USE Model (Cutoff 2)	0.760	0.950	0.71	0.620
miRFinder	0.809	0.906	0.715	0.538
Virgo	0.823	0.712	0.535	0.306
triplet-SVM	0.739	0.914	0.653	0.510
microPred	0.908	0.733	0.641	0.363

Discussion and Conclusion

In this work, a novel objective function called **USE** is presented that measures variability based on unpaired probabilities of nucleotides. USE can be interpreted as a measurement of uncertainty for a nucleotide in a structure to be unpaired (i.e. bulge or loop). MicroRNAs generally possess a stereotypical long-stem structure making them relatively less flexible and

more stable than other ncRNAs. Nucleotide sequence variations within microRNA were shown not to affect the drosha and dicer processing (Diederichs and Haber 2006). In fact, the most important secondary structure determinants for miRNA were found to be a greater than 16bp stem, and a lower number and reduced size of bulges and internal loops. (Zeng and Cullen 2004), (Ritchie, Legendre, and Gautheret 2007). From this we can infer that structures with long stem, fewer bulges and shorter loop tend to be more stable. Figure 4.2 indicates that the USE score can be an acceptable criterion in distinguishing miRNA structures from other ncRNA families. If we compare our miRNA USE score distribution to Q, USE can distinguish more miRNAs than Q, demonstrating the novelty and statistical power of the USE function.

Although RNA families are characterized by their sequence and structure, this does not imply that size is preserved across the RNA family. This is especially observable within pre-miRNAs of different length while having similar hairpin loop structures. Therefore, it is important for a RNA-family detector to be robust against violations of assumptions on the sequence length. Here, the window scan of USE calculation is done for different window lengths in order to evaluate the USE window scan's performance. Figure 4.3 shows that the USE score is not dependent on the window size (i.e. the length of the microRNA to be found), since the USE score is always the lowest at the miRNA position, regardless of the length of the sliding window.

Furthermore, Lee and Kim (Lee and Kim 2008) have demonstrated that miRNAs have a tendency to retain their structure regardless of the neighboring upstream and downstream sequences. Their finding indicates that the sequence containing the miRNA and additional upstream and downstream sequence has as relatively low structural variability as the original miRNA, and they have termed this phenomenon as self-containment (SC). The USE scores of

microRNAs are also observed to have a similar behavior, since they tend to stay relatively low even when upstream and/or downstream sequences are added to the actual microRNA. This suggests a high correlation between the USE Score and the SC index which can also be observed from figure 4.5.

Capturing structural features as well as other RNA measurements has always played a significant role in classifying different RNA sequences. The challenge is to select the features that are specific to a category of ncRNA. The power of such features can be assessed through various machine learning techniques. As we have discussed earlier, the low structural variability of miRNAs distinguishes them from other families of ncRNAs as well as from their background. Table 4.2 is a comparison of the different miRNA classifiers, and our single feature USE classifier's sensitivity, specificity, and MCC outperforms all existing SVM methods. The two cutoffs of the USE classifier demonstrate that a higher specificity can be achieved without sacrificing the sensitivity. Finally, the inclusion of USE together with other known features results in an even better performance, with a higher MCC value, suggesting that a great deal of information is contained in the USE structural feature. For our SVM method we purposely chose a high cutoff to have a stringent specificity, since this is the major difficulty in computational detection of miRNA.

In conclusion, microRNA molecules possess low structural variability compared to other families; USE successfully captures this low global variability, offering a substantial improvement to the current state of miRNA gene finding. Since USE is able to better quantify the structural variability of a sequence of long stems and small bulges, we believe USE can be further applied to develop ab initio prediction programs for a larger class of ncRNAs, or be applied to study stem-loop structures within viral sequences. A limitation of our study is its

dependency on the NUPACK's secondary structural model. Looking to future applications, there is potential for the USE feature to be applied in the prediction and validation of various tertiary models by quantifying the structural variability of a sequence or be applied to the identification of miRNA gene targets. The USE method seems to work well on the long stem loop of pre-miRNAs but not on long stem loops of random sequences or other ncRNAs like snoRNAs. This could indicate some intrinsic nature of other ncRNAs that has yet to be discovered. Such a phenomenon would offer an opportunity for future investigation on techniques for detecting other ncRNAs.

References

Ambros, V. 2004. "The functions of animal microRNAs." *Nature* 431 (7006):350-355. doi: 10.1038/nature02871.

Batuwita, R., and V. Palade. 2009. "microPred: effective classification of pre-miRNAs for human miRNA gene prediction." *Bioinformatics* 25 (8):989-995. doi: 10.1093/bioinformatics/btp107.

Bentwich, I., A. Avniel, Y. Karov, R. Aharonov, S. Gilad, O. Barad, A. Barzilai, P. Einat, U. Einav, E. Meiri, E. Sharon, Y. Spector, and Z. Bentwich. 2005. "Identification of hundreds of conserved and nonconserved human microRNAs." *Nature Genetics* 37 (7):766-770. doi: 10.1038/ng1590.

Bonnet, E., J. Wuyts, P. Rouze, and Y. Van de Peer. 2004. "Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences." *Bioinformatics* 20 (17):2911-2917. doi: 10.1093/bioinformatics/bth374.

Clote, P., F. Ferre, E. Kranakis, and D. Krizanc. 2005. "Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency." *Rna-a Publication of the Rna Society* 11 (5):578-591. doi: 10.1261/rna.7220505.

Diederichs, S., and D. A. Haber. 2006. "Sequence variations of microRNAs in human cancer: Alterations in predicted secondary structure do not affect processing." *Cancer Research* 66 (12):6097-6104. doi: 10.1158/0008-5472.can-06-0537.

Dirks, R. M., J. S. Bois, J. M. Schaeffer, E. Winfree, and N. A. Pierce. 2007. "Thermodynamic analysis of interacting nucleic acid strands." *Siam Review* 49 (1):65-88. doi: 10.1137/060651100.

Freyhult, E., P. P. Gardner, and V. Moulton. 2005. "A comparison of RNA folding measures." *Bmc Bioinformatics* 6. doi: 10.1186/1471-2105-6-241.

Griffiths-Jones, S. 2007. "Annotating noncoding RNA genes." *Annual Review of Genomics and Human Genetics* 8:279-298. doi: 10.1146/annurev.genom.8.080706.092419.

Griffiths-Jones, Sam. 2006. "MiRBase: The MicroRNA sequence database." *Methods in Molecular Biology*:129-138.

Gruber, A. R., R. Lorenz, S. H. Bernhart, R. Neuboock, and I. L. Hofacker. 2008. "The Vienna RNA Websuite." *Nucleic Acids Research* 36:W70-W74. doi: 10.1093/nar/gkn188.

Gusarov, I., and E. Nudler. 1999. "The mechanism of intrinsic transcription termination." *Molecular Cell* 3 (4):495-504.

Huang, T. H., B. Fan, M. F. Rothschild, Z. L. Hu, K. Li, and S. H. Zhao. 2007. "MiRFinder: an improved approach and software implementation for genome-wide fast microRNA precursor scans." *Bmc Bioinformatics* 8. doi: 10.1186/1471-2105-8-341.

Huynen, M., R. Gutell, and D. Konings. 1997. "Assessing the reliability of RNA folding using statistical mechanics." *Journal of Molecular Biology* 267 (5):1104-1112.

Kumar, S., F. A. Ansari, and V. Scaria. 2009. "Prediction of viral microRNA precursors based on human microRNA precursor sequence and structural features." *Virology Journal* 6. doi: 10.1186/1743-422x-6-129.

Lee, Miler T., and Junhyong Kim. 2008. "Self Containment, a Property of Modular RNA Structures, Distinguishes microRNAs." *PLoS Computational Biology* 4 (8). doi: :10.1371/journal.pcbi.1000150.

Lee, Y., C. Ahn, J. J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Radmark, S. Kim, and V. N. Kim. 2003. "The nuclear RNase III Drosha initiates microRNA processing." *Nature* 425 (6956):415-419. doi: 10.1038/nature01957.

Loong, S. N. K., and S. K. Mishra. 2007. "Unique folding of precursor microRNAs: Quantitative evidence and implications for de novo identification." *Rna-a Publication of the Rna Society* 13 (2):170-187. doi: 10.1261/rna.223807.

McCaskill, J. S. 1990. "THE EQUILIBRIUM PARTITION-FUNCTION AND BASE PAIR BINDING PROBABILITIES FOR RNA SECONDARY STRUCTURE." *Biopolymers* 29 (6-7):1105-1119.

Pedersen, J. S., G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. 2006. "Identification and classification of conserved RNA secondary structures in the human genome." *PLoS Computational Biology* 2:251-262. doi: 10.1371/journal.pcbi.0020033.

Ritchie, W., M. Legendre, and D. Gautheret. 2007. "RNA stem-loops: To be or not to be cleaved by RNAse III." *Rna-a Publication of the Rna Society* 13 (4):457-462. doi: 10.1261/rna.366507.

Rivas, Elena, and Sean R. Eddy. 2001. "Noncoding RNA gene detection using comparative sequence analysis." *Bmc Bioinformatics* 2 (8 Cited April 17, 2002):1-19.

Scaria, V., M. Hariharan, S. Maiti, B. Pillai, and S. K. Brahmachari. 2006. "Host-virus interaction: a new role for microRNAs." *Retrovirology* 3:68. doi: 1742-4690-3-68 [pii] 10.1186/1742-4690-3-68.

Schattner, Peter. 2003. "Computational gene-finding for noncoding RNAs." *Non-coding RNAs, J. Barciszewski and V. Erdmann (eds.) Landes Bioscience*.

Shannon, C. E. 1997a. "The mathematical theory of communication (Reprinted)." *M D Computing* 14 (4):306-317.

Smith, N. A., S. P. Singh, M. B. Wang, P. A. Stoutjesdijk, A. G. Green, and P. M. Waterhouse. 2000. "Gene expression - Total silencing by intron-spliced hairpin RNAs." *Nature* 407 (6802):319-320.

Washietl, S., I. L. Hofacker, and P. F. Stadler. 2005. "Fast and reliable prediction of noncoding RNAs." *Proceedings of the National Academy of Sciences of the United States of America* 102 (7):2454-2459. doi: 10.1073/pnas.0409169102.

Washietl, S., J. S. Pedersen, J. O. Korbel, C. Stocsits, A. R. Gruber, J. Hackermuller, J. Hertel, M. Lindemeyer, K. Reiche, A. Tanzer, C. Ucla, C. Wyss, S. E. Antonarakis, F. Denoeud, J. Lagarde, J. Drenkow, P. Kapranov, T. R. Gingeras, R. Guigo, M. Snyder, M. B. Gerstein, A. Reymond, I. L. Hofacker, and P. F. Stadler. 2007. "Structured RNAs in the ENCODE

selected regions of the human genome." *Genome Research* 17 (6):852-864. doi: 10.1101/gr.5650707.

Xue, C. H., F. Li, T. He, G. P. Liu, Y. D. Li, and X. G. Zhang. 2005. "Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine." *Bmc Bioinformatics* 6. doi: 10.1186/1471-2105-6-310.

Zeng, Y., and B. R. Cullen. 2004. "Structural requirements for pre-microRNA binding and nuclear export by Exportin 5." *Nucleic Acids Research* 32 (16):4776-4785. doi: 10.1093/nar/gkh824.

Zhang, B., X. Pan, G. P. Cobb, and T. A. Anderson. 2006. "Plant microRNA: a small regulatory molecule with big impact." *Dev Biol* 289 (1):3-16. doi: S0012-1606(05)00764-5 [pii]

10.1016/j.ydbio.2005.10.036.

Zuker, Michael. 2009. "The entropy of the Boltzman distribution of RNA folding, Benasque Workshop on Computational Methods for RNA Analysis." *Benasque, Spain.*

CHAPTER 5

MODELING HIV-1 RNA SECONDARY STRUCTURES[1]

---

[1] Timothy I. Shaw, Russell L. Malmberg, Ming Zhang. To be submitted to *PLoS Computational Biology*.

Abstract

HIV RNA secondary structure plays a key role in regulating viral replication and alternative splicing. Recent efforts have elucidated the RNA secondary structure of NLM4-3, an HIV-1 B subtype strand through high throughput RNA structure analysis SHAPE technology. Given the importance of RNA secondary structure in the HIV life cycle, HIV sequence diversity can be greatly influenced by its RNA structure architecture. In our study, we evaluated the capacity and accuracy for NLM4-3 and RNA structure prediction programs to be used for RNA secondary structure modeling across HIV strains. We found tremendous HIV RNA secondary structural diversity across different subtypes, and we noticed RNA structural conservation to vary across the HIV genome. The HIV RNA Pasta Folder pipeline was developed which uses NLM4-3 to guide initial folding with further RNA structure refinement through CONTRAFold. For a proof of concept, we used HIV RNA Pasta Folder to investigate the structural variations between B and C subtype for both recombinants and pure subtypes. We found limited number of stems conserved between C and NLM4-3, suggesting we cannot at this time confidently predict non-B subtype sequences. Additional investigation on CRF07-08 China BC subtype recombination finds that the GAG and POL region to be relatively conserved in their RNA structure; however, substantial RNA structure variations was found in VPR-ENV splice donor and acceptors sites as well as in the NEF/LTR region. These results indicate NLM4-3 reference might not be able to accurately capture the RNA structure in certain HIV genomic region, and NLM4-3 reference should be cautiously applied in future studies.

Introduction

The impact of RNA secondary structure is undeniably critical to the human immunodeficiency virus (HIV)'s proliferation and fitness. RNA secondary structure's importance in HIV-1 has been implicated in a number of viral regulatory processes (Peleg et al. 2002, Peleg, Trifonov, and Bolshoy 2003). RNA structure could directly impact HIV sequence diversity through the regulation of HIV recombination (Krzywinski et al. 2009) and alternative splicing (Pollom et al. 2013, Abbink and Berkhout 2008). With HIV-1's ability to evolve at a rapid rate (Ho et al. 1995), a competing evolutionary pressure exists between RNA structure and proteins (Sanjuan and Borderia 2011). Particularly, RNA secondary structure has been shown to be a key feature for epitope identification (Snoeck et al. 2011), and can be correlated to drug resistance associated mutations (Sanjuan and Borderia 2011). HIV can also generate mutations that form RNA base pairs protecting it from RNA interference (Westerhout et al. 2005), opening the prospect for an RNAi-based gene therapy guided by RNA structure (Low et al. 2012). Improving the models for HIV RNA secondary structure can help us understand the driving mechanism impacting HIV life cycle and its sequence diversity.

Through selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) methodology, the full secondary structure of a 9kb NLM4-3 strand B subtype HIV genome was recently resolved (Watts et al. 2009). SHAPE technology allows for high throughput interrogation of local nucleotide flexibility, without the use of thermodynamic information (Low and Weeks 2010, Lucks et al. 2011). This technology examines each nucleotide and produces a reactivity index describing the propensity for a nucleotide to be part of a RNA structure base pairing. The success of this approach was demonstrated in Watts et al. finding that NLM4-3 strand is orthogonal to base pairing probabilities generated by phylogenetic based computational

models (Watts et al. 2009). The RNA secondary structure for NLM4-3 represents the first experimentally supported HIV full length RNA structural reference. Although NLM4-3 RNA structure have already been applied to multiple studies (Snoeck et al. 2011, Sanjuan and Borderia 2011), preliminary screening of multiple HIV and SIV sequences has revealed significant structural variation across HIV-1 subtypes (Pollom et al. 2013, Knoepfel and Berkhout 2011). The appropriateness for NLM4-3 to be extrapolated to other HIV strains was called into question, and in our study we will attempt to examine RNA structure variation between NLM4-3 and HIV strains.

Current computational approaches predict RNA secondary structure through recursive algorithms to generate all potential base-pairing in a given RNA sequence (Nussinov and Jacobson 1980, Eddy 2004). These algorithms utilize either one of two types of structure prediction/parsing parameters: thermodynamic or non-thermodynamic probabilistic (Schroeder 2009). Stochastic context-free grammar (SCFG) can be used to generate RNA structure. SCFG based design provides rules for parsing the RNA sequence and predicts structures through maximizing the weight of the parsing rules such as Crocke-Younger-Kasami (CYK) algorithm (Durbin 1998b). SCFG based on non-thermodynamic approach utilizes probabilities as the weight and hence is equivalent to a maximum-likelihood approach. A thermodynamic approach instead ranks potential structures based on free-energy estimates derived using thermodynamic parameters. RNA Fold algorithms could be further divided into either phylogenetic based approach (Bernhart et al. 2008, Sukosd et al. 2011), de novo based (Hofacker and Stadler 2006, Dirks and Pierce 2003, 2004, Do, Woods, and Batzoglou 2006), and SHAPE reactivity guided folding (Sukosd et al. 2012, Washietl, Hofacker, et al. 2012, Reuter and Mathews 2010). Phylogenetic approaches make use of patterns in nucleotide conservation and compensatory

mutations within an alignment to postulate the RNA structure, while de novo folding directly infers the RNA Fold algorithm from a single sequence. SHAPE guided folding is an extension of either de novo based method or phylogenetic method to limit the exploration of all possible structure. Existing RNA fold algorithm are generally trained using noncoding RNA structure from prokaryotes and eukaryotes (Do, Woods, and Batzoglou 2006, Andronescu et al. 2007, Lu, Gloor, and Mathews 2009). Provided that HIV possesses a nucleotide composition different from other living organism (van der Kuyl and Berkhout 2012), our study will evaluate the accuracy for each RNA fold algorithm to predict RNA structures in HIV.

Table 5.1. RNA Fold Algorithms and Fold Measures.

Program Name	Fold Algorithm	Fold Measure	Thermodynamic Based?	Uses Phylogenetic Information	Note
Mfold	Yes	Minimal Free Energy	Yes	No	Use nearest neighbor Turner's energy parameters
RNAFold	Yes	Minimal Free Energy	Yes	No	Use nearest neighbor Turner's energy parameters
CONTRAFold	Yes	None	No	No	Conditional Training
Nupack	Yes	Base Pair Probability	Yes	No	Use nearest neighbor Turner's energy parameters
USE	No	Entropy of Base Pair Probability	Yes	No	Determines fold certainties
CYK-hairpin	Yes	Sum of Log Probability	No	No	Identifies presence of RNA structure

Materials and Methods

HIV sequences used in this work were aligned to the HXB2 reference sequence, and all positions correspond to HXB2 numbering (Korber 1998). A summary of the programs used within this paper is provided in table 1. These programs are also described in depth in Schroeder's review of RNA structure prediction tools (Schroeder 2009).

Section 1: HIV Sequence Base Pair Conservation

Construction of the "America & Europe" sequence alignment

NLM4-3 is a recombinant strain, having originated from two virus strains of HIV-1: NY5, sampled in America, and LAV, sampled in France (Barre-Sinoussi et al. 2004, Benn et al. 1985, Adachi et al. 1986). To examine RNA structure variation, we extracted all B subtype sequences sampled during 1982-1985 from America, France, and United Kingdom. These sequences were aligned using the gene cutter sequence alignment tool developed at the Los Alamos National Lab (LANL) and HIV N-linked glycosylation site analyzer (Shaw and Zhang 2013) for aligning the variable loop region. Manual alignment curating was performed through Bioedit.

Protein domain junction and Inter-protein linkage HIV element

RNA secondary structures are found to accumulate within the HIV genome's Inter-protein linkage (IPL) and protein domain junction (PDJ). In their 2009 SHAPE study, Watt's et al identified 14 genome elements of inter-protein linkage (IPL) and protein domain junctions (PDJ) based on the Protein Data Bank (PDB) structures: 2GOL, 1A43, 3PHV, 1HMV, 1WJA, 1BIS, 1QMC, 1GC1, and 1SZT. Using this information, we mapped IPL and PDJ location from the PDB structures onto the HXB2 genome using Watt et al's PDB coordinate references (Watts et al. 2009), and converted these positioning to HXB2 positioning.

Calculating Number of Conserved/Non-Conserved Base Pairing

RNA Covariant base pairs are evolutionary indicators for structural conservation across phylogenetically distant HIV sequences. Under mutational pressure, the conserved base pairs are also called covariate base pairs. Covariant base pairing across phylogenetically-distant sequences generally indicate evolutionary importance. In our

text, a mutation that leads to a loss of base pairing will be referred to as "covariate base pairing loss" or "non-conserved base pairing." In order to calculate the covariate base pair loss for each query sequence, genecutter (***http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html***) is used to align sequences with NLM4-3. Alignment in the variable loop region is further improved using V Loop Alignment from HIV N-linked glycosylation site analyzer (Shaw and Zhang 2013). Based on the derived alignment, the RNA structure for each sequence from the alignment can be inferred based on the NLM4-3 RNA structure reference.

Bootstrapping for correlation

Prior to the bootstrap, average number of base pair losses is calculated for all genome elements (IPL and PDJ). For each bootstrap-iteration, a genome element of the same size is randomly selected from the genome; if two genome elements overlap with one another, this element is re-sampled. The average number of non-conserved base pair for each random genome element are calculated and compared to the original average number of non-conserved base pair. The proportion of times the bootstrap of number of base pair loss is less than the initial number of non-conserved base pairs average is the calculated p-value. Therefore, our p-value indicates the probability that an observed base pair loss in the true collection of genome elements is significantly smaller than if the event occurred by chance within the set of elements. A total of 10,000 bootstraps are performed to examine each genome element.

Section 2: HIV RNA Fold Method accuracy comparison

Assessing Folding Accuracy using Rfam HIV Sequences

"Seed" alignment from HIV RNA structure families is downloaded from Rfam 10.0 (Gardner et al. 2009), an RNA family database. HIV RNA families includes: GSL3, SD, SL3, SL4, RRE, DIS, FE, PBS, POL and TAR. This database contains a consensus structure for each RNA structure family, and the consensus structure is used to evaluate the accuracy of the RNA structure prediction program based on sensitivity (Eq. 1), and PPV (Eq. 2) shown below (Do, Woods, and Batzoglou 2006)

$$SEN = \frac{TP}{TP+FN} = \frac{number\ of\ correct\ base\ pairings}{number\ of\ correct + number\ of\ missed\ (base\ pairings)} \quad \text{(Eq. 1)}$$

$$PPV = \frac{TP}{TP+FP} = \frac{number\ of\ correct\ base\ pairings}{number\ of\ correct + number\ of\ incorrect\ (base\ pairing)} \quad \text{(Eq. 2)}$$

To simplify the comparison, we applied the F measure, a harmonic mean of both sensitivity and PPV.

$$F = \frac{2}{\frac{1}{SEN}+\frac{1}{PPV}} \quad \text{(Eq. 3)}$$

Assessing Accuracy using NLM4-3 RNA Structure

RNA structure prediction for long sequences such as HIV is a challenging process (Schroeder 2009). To estimate algorithm performance as a function of sequence length, we used SHAPE's resolved NLM4-3 structure as our reference structure and extracted all sequences with base pairing interactions ranging from 60nt to 300nt with increments of 30nt. We then compared RNA structure prediction programs' result against one another based on the sensitivity measure. A matrix was created and pheatmap was used to cluster the different sequences, pheatmap's default agglomeration method "complete" was used.

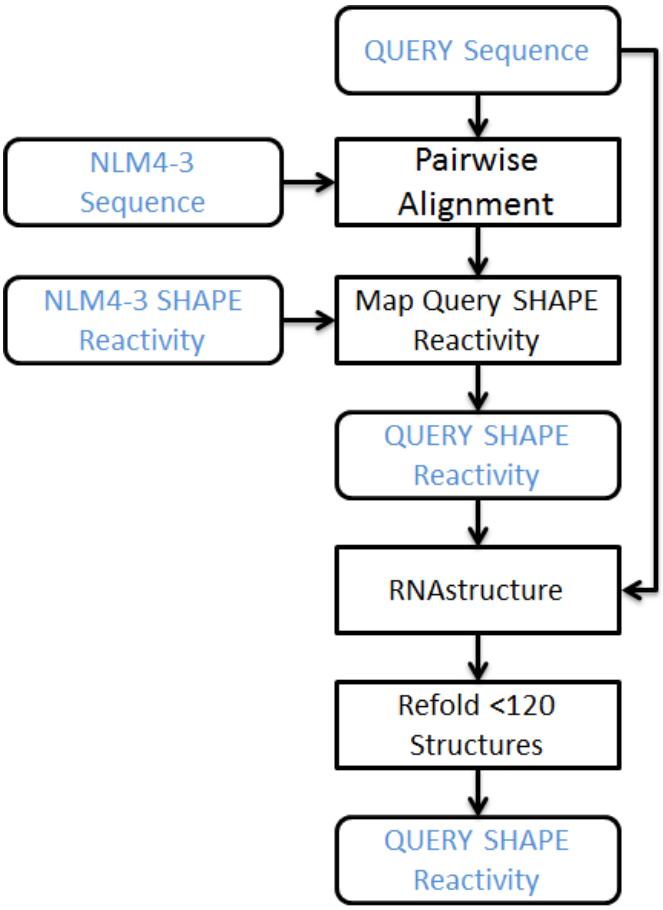


Figure 5.1. Pipeline for HIV RNA folding pipeline.

The circular rectangles represent user input and output. The rectangles represent the step being performed in the pipeline. The NLM4-3 SHAPE reactivity is first extrapolated onto the sequence and used to predict the initial RNA structure. Finally RNA structure with 120 bp was refolded using CONTRAFold.

Section 3: SHAPE based Reference Folding

HIV RNA Pasta Folder

Query sequence was aligned to NLM4-3. Based on the alignment, SHAPE reactivity was extrapolated onto the query sequence. RNAstructure (Reuter and Mathews 2010) was used to

fold the RNA secondary structure with the addition of auxiliary SHAPE reactivity. Parameters m = 1.9 and b = -0.7 were used for RNAstructure for the highest prediction sensitivity (Hajdin et al. 2013). Structures less than 120nt are refolded using CONTRAFold. Figure 5.1 displays the general work flow of our methodology. The choice of using CONTRAFold as the refolding methodology was based on our results showing CONTRAFold having the highest accurate F-measure (harmonic mean of sensitivity and PPV, see results and discussion).

Mountain Plot

The mountain plots (Hogeweg and Hesper 1984) were created through an in house java program and visualized through customized R script and circos plots (Krzywinski et al. 2009). Mountain plot is a graph representation of an RNA structure, showing the depth of the embedded stem loops. Within a mountain plot, the y-axis represents the number of base pairs enclosed at a specific genome nucleotide position, x-axis. For example, assume we have nucleotide sequence, GGGAACCC. The secondary structure for this sequence will have base pairing occur between the first three and last three nucleotides. On a mountain plot, this sequence would be represented by the numbering 1, 2, 3, 3, 3, 3, 2, and 1 along the y-axis. In circos, we added blue arcs to indicate the conserved stems between the compared structures. The stems were initially identified through RNApasta (Malmberg, Shaw, and Cai 2010). Conservation were determined if the stem from query and reference contains +/- 3nt overlapping the same base pairing region.

China BC Recombinant Analysis

All full length CRF07, CRF08, and pure B subtype sequences from China were downloaded from the LANL HIV sequence database. The sequences were aligned with NLM4-3 through gene cutter and further improved by HIV N-linked glycosylation site analyzer (Shaw and

Zhang 2013). RNA structure was directly inferred based on NLM4-3 RNA structure and refolding was performed based on CONTRAFold. RNA structures predicted by each sequences were compared by examining the proportion of discrepant base pairs (see sensitivity calculation). A matrix was created and pheatmap was used to cluster the different sequences, pheatmap's default agglomeration method "complete" was used. In addition, a phylogenetic analysis was performed to examine the clustering between China recombinant B compared to pure B subtype. PHYLIP program version 3.69 (J 1989) was applied for the phylogenetic analysis, using F84 neighbor-joining method supported by 100 iterations of nonparametric bootstrapping.

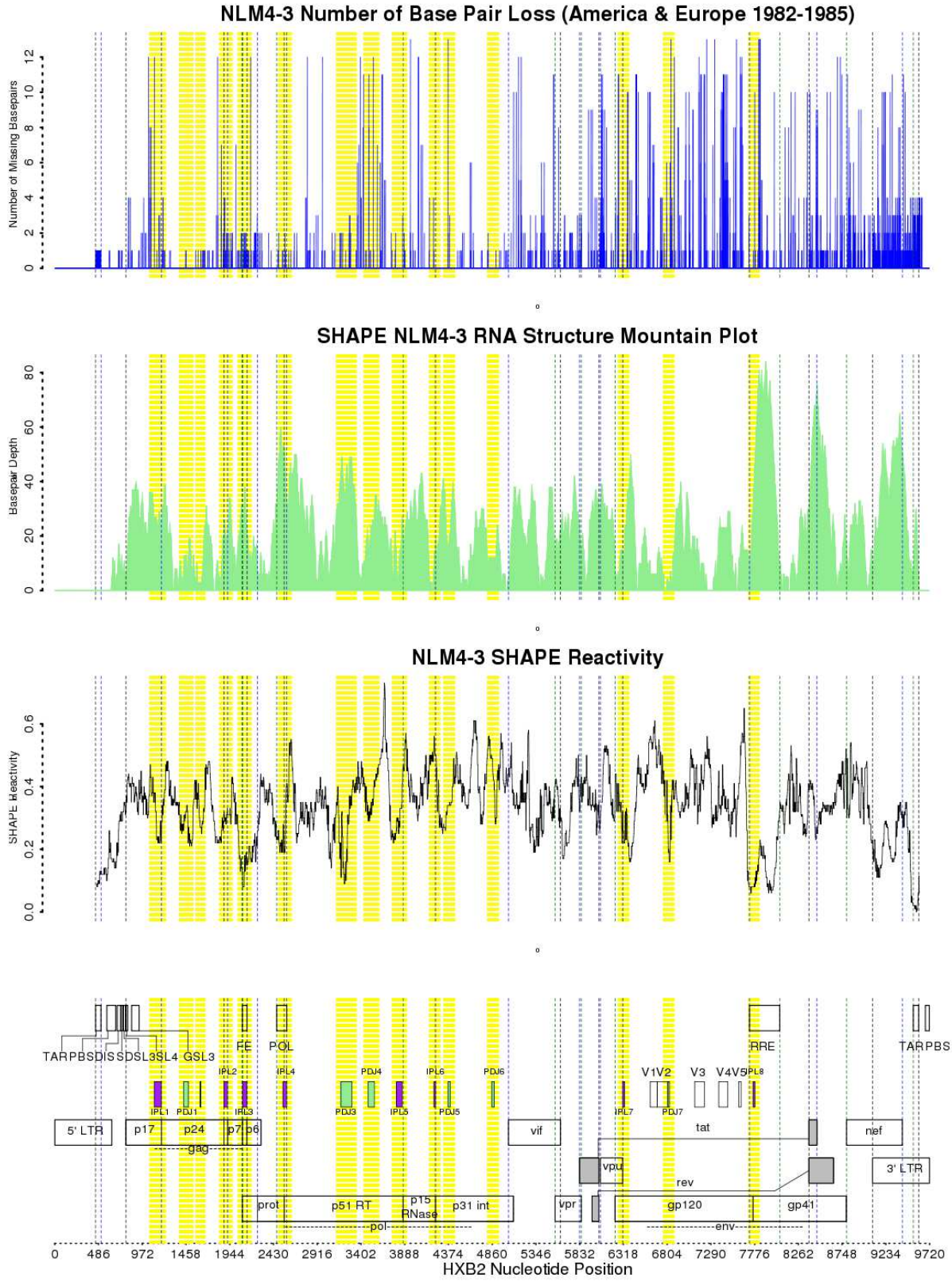

NLM4-3 Number of Base Pair Loss (America & Europe 1982-1985)
Number of Missing Basepairs
0
2
4
6
8
10
12
SHAPE NLM4-3 RNA Structure Mountain Plot
Basepair Depth
0
20
40
60
80
NLM4-3 SHAPE Reactivity
SHAPE Reactivity
0.0
0.2
0.4
0.6
TAR
PBS
DIS
SD
SL3
SL4
GSL3
FE
POL
IPL1
PDJ1
IPL2
IPL3
IPL4
PDJ3
PDJ4
IPL5
IPL6
PDJ5
PDJ6
IPL7
V1V2
PDJ7
V3
V4V5
IPL8
RRE
TAR
PBS
5' LTR
p17
p24
p7
p6
gag
vif
tat
nef
vpu
rev
3' LTR
prot
p51 RT
p15
RNase
p31 int
vpr
gp120
gp41
pol
env
0
486
972
1458
1944
2430
2916
3402
3888
4374
4860
5346
5832
6318
6804
7290
7776
8262
8748
9234
9720
HXB2 Nucleotide Position

Figure 5.2. Sequences similar to the NLM4-3 reference and their sequence base pair lost mapped to HXB2 positioning.

We aligned the America & Europe 1982-1985 base pair lost dataset against SHAPE NLM4-3 structure mountain plot and SHAPE reactivity. Highlighted in yellow are the genome elements, interprotein linker (IPL) and protein domain junction (PDJ). These region have a lower number of base pair lost (i.e. a higher conservation for covariate base pairs) supported by bootstrap p-value of < 0.05.

Results

HIV Sequence Base Pair Conservation

NLM4-3 is a chimeric B subtype HIV strain created based on NY5 and LAV strand from North America and Europe, respectively. HIV being highly diverse, the ability for NLM4-3's RNA structure to be extended to other sequences was put into question. Base pair conservation between NLM4-3 and sequences from other subtypes were compared, specifically in two areas: (1) Sequences derived NLM4-3 based time period and geographic region, and (2) Sequences from different subtypes.

Sequences from similar time period and geographic region as NLM4-3 were downloaded from the LANL database. Multiple sequence alignment was then generated using gene cutter. Each sequence's RNA structure was inferred based on NLM4-3 RNA structure (See material method for *Calculating Number of Conserved/Non-Conserved Base Pairing*). Figure 5.2 displays the NLM4-3 RNA structural architecture, SHAPE reactivity, and the number of non-conserved base pairs. Across the HIV genome, the IPL and PDJ elements (highlighted in yellow in figure 5.2) have higher preservation of covariate base pairs, supported by p-value 0.0068 from

bootstrapping, indicating tendencies for structure motifs to be conserved in these regions. Difference in base pair preservation was found between different protein coding regions (Table 5.2). The viral proteins, GAG and POL, contained the least amount of covariate base pair losses (17% and 9%), while the viral protein ENV showed a slightly higher level of base pair losses (22%). Finally, proteins NEF and VPR were the most variable viral proteins exhibiting close to 40% covariate base pair loss.

The same analysis above was also applied to compare across different subtypes: A1 and A2, B, C, D, F1 and F2, G, H, J, and K subtype. When compared to NLM4-3, B subtype sequences had the lowest level of covariate base pair loss, at 19% (Table 5.3). For non-B subtype sequences, elevated levels of covariate base pair loss were observed, ranging from 30% loss in the D subtype to 37% loss in the J subtype (Table 5.3).

Table 5.2. Percentage of Non-conserved Base Pairs across B subtype America & Europe Sequences.

Protein	% of Covariate base pair losses
GAG	16.9719
POL	9.0383
VIF	15.508
VPR	38.2114
TAT	19.4093

VPU	19.4093
REV	20.5479
ENV	21.7407
NEF	39.8438

Table 5.3. Percentage of Non-conserved Base Pairs between HIV 2010 Reference Subtypes and NLM4-3 Sequence

Subtype	% of Covariate base pair losses
A1	32.2970
A2	32.6158
B	19.2432
C	34.7882
D	29.8001
F1	33.1056
F2	32.6273
G	34.3953
H	31.9525
J	37.2737
K	31.9098

Table 5.4. Rfam reference estimation of Sensitivity, PPV, and F-measure.

	Sensitivity	PPV	F-measure
CONTRAFold	0.85911	0.71011	0.77754
Mfold	0.841565	0.67354	0.74824
Nupack	0.814321	0.63366	0.71272
RNAfold	0.825609	0.6779	0.7445
RNAstructure	0.833488	0.67741	0.74739
RNAstructure + SHAPE	0.855469	0.69371	0.76614
CYK	0.731685	0.49711	0.59201
CYK + SHAPE	0.885276	0.68817	0.77438

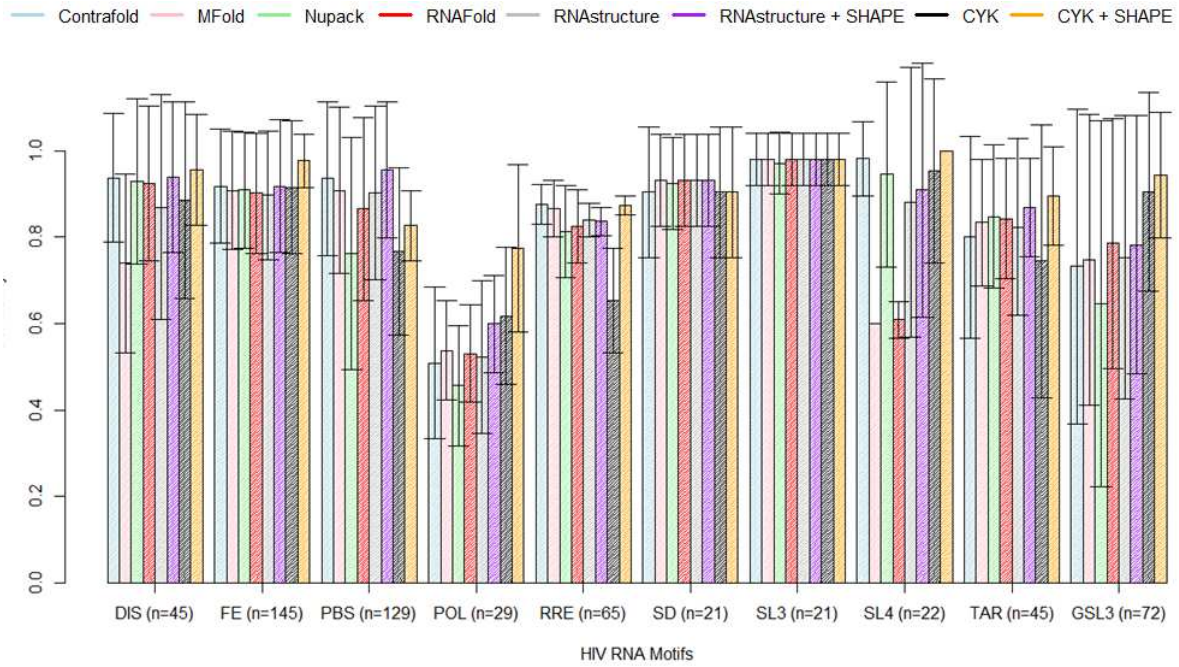


Figure 5.3A. Sensitivity of RNA Fold Algorithm assessed by Rfam 10.0.

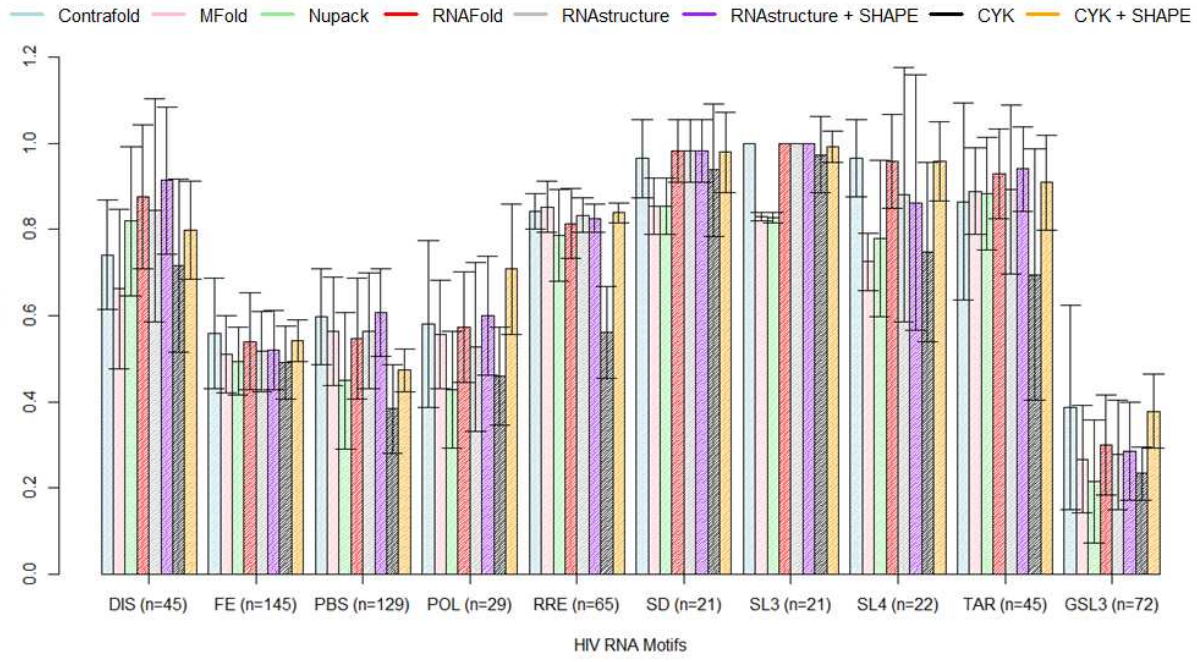


Figure 5.3B. PPV of RNA Fold Algorithm assessed by Rfam 10.0.

Figure 5.3. Comparison of RNA Fold Algorithms using Rfam 10.0 HIV RNA Families sequences.

The figure represent four fold algorithms that were assessed for their sensitivity and PPV across the all HIV Rfam 10.0 SEED dataset. The result indicates non-thermodynamic algorithms have a slight edge over thermodynamic based algorithms.

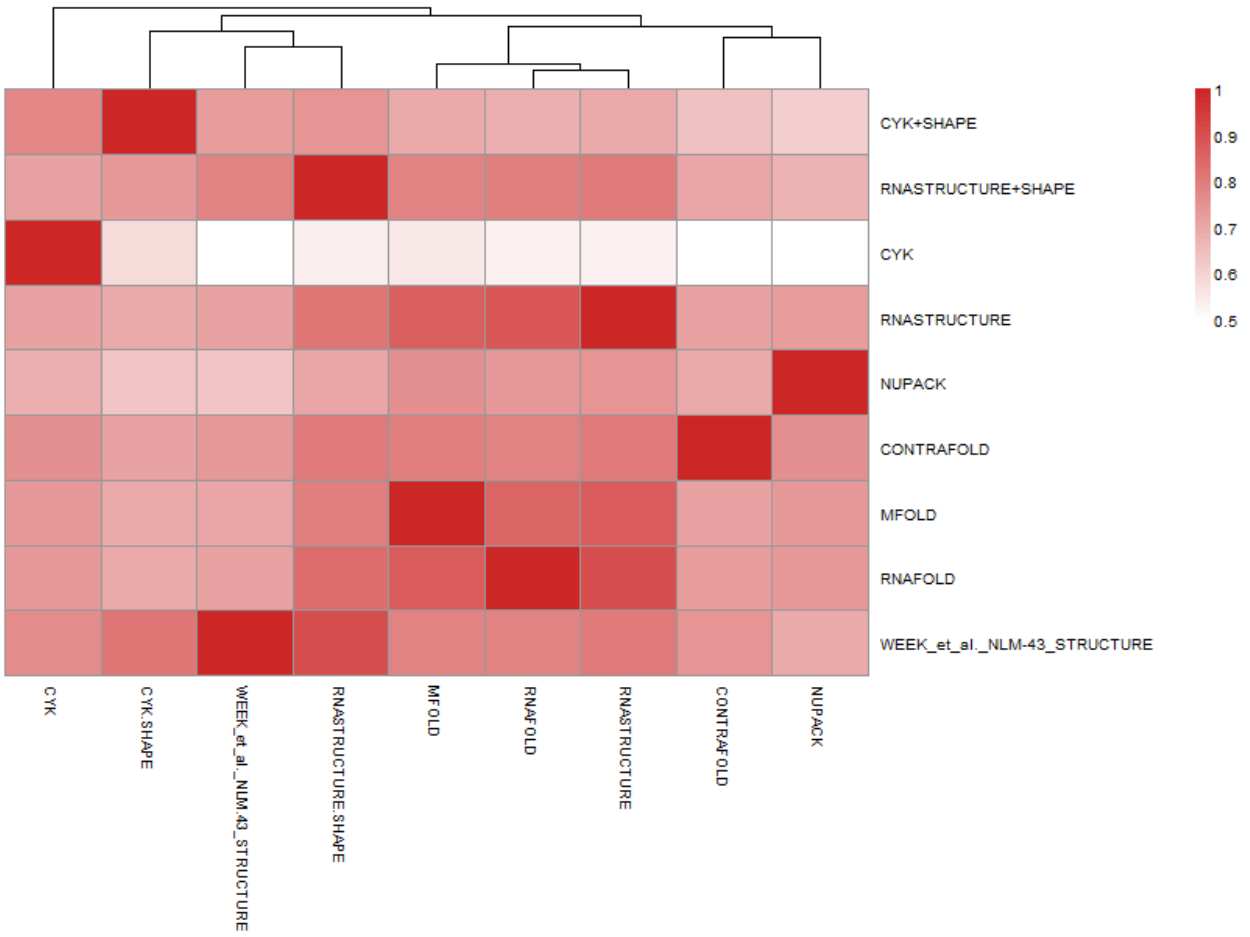


Figure 5.4. Comparison of RNA Fold Algorithms using NLM4-3 SHAPE Reference.

Different base paired sequence size cutoff were used to extract the NLM4-3 SHAPE sequence to evaluate performance of different programs as a function of sequence length ranging from 60nt – 300nt. For a cutoff of window size 120, NLM4-3 solved structure is found to be most similar to SHAPE dependent method of CYK + SHAPE and RNAstructure + SHAPE. Thermodynamic based methods such as RNAstructure, RNAfold, and MFold tend to cluster together.

HIV RNA Fold Method accuracy comparison

Rfam was used as reference to assess different RNA Fold methodologies Based on sensitivity, PPV, and F-measure (see methodology). Overall, non-thermodynamic methods were found to be

higher in both sensitivity and PPV than thermodynamic based methods (Figure 5.3). CONTRAFold had the highest F-measure and the highest PPV value. CYK + SHAPE had the second highest F-measure and the highest sensitivity (Table 5.4). Although CYK had the lowest performance, the addition of SHAPE reactivity into CYK's algorithm significantly increased its performance, producing comparable result as CONTRAFold. In addition to assessing the accuracy of different methods, RNA structural similarities from the structural predictions were compared. Sequences from NLM4-3 with RNA structures ranging from 60nt to 300nt base pair were extracted. Each predicted RNA structures was compared with one another by examining the proportion of discrepant base pairs (it's the same calculation as sensitivity Eq. 1). A matrix comparing every possible combination of RNA prediction program was generated, and clustering was performed using default parameters by pheatmap (Figure 5.4). After 150nt the RNA structure prediction method's clustering became significantly different from the clustering result for NLM4-3 structures less than 120nt. For 120nt (Figure 5.4), the reference NLM4-3 RNA structure is most similar to the predicted structure by RNAstructure + SHAPE with CYK + SHAPE possessing the next most similar predicted structure, as CYK + SHAPE is part of the parental node of the hierarchical cluster.. Methods using thermodynamic based method, RNAstructure, RNAfold and MFold tend to be clustered together. Finally, CONTRAFold and NUPACK were clustered together.

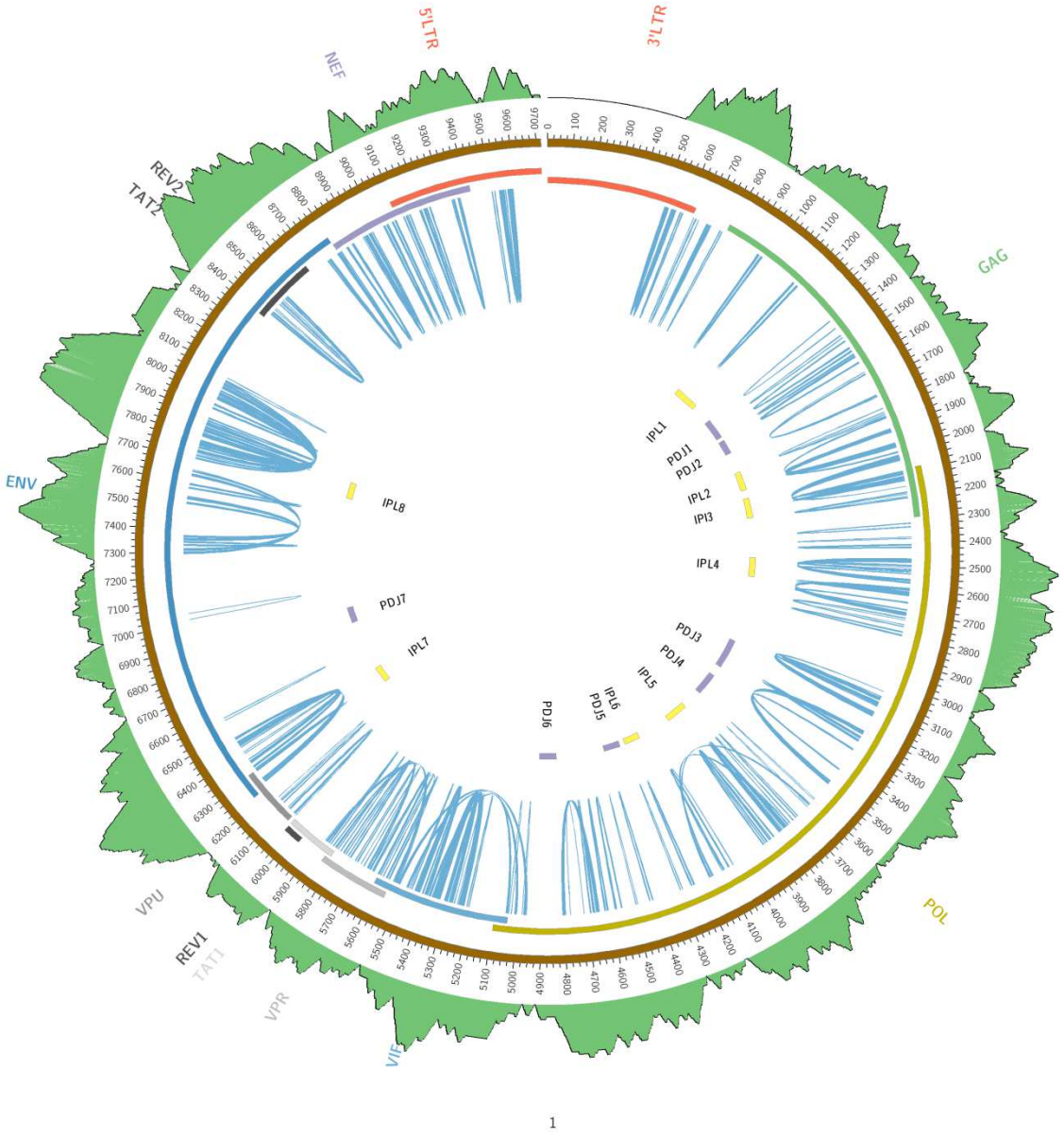


Figure 5.5A. RNA secondary structure for a B subtype HIV sequence AY423387

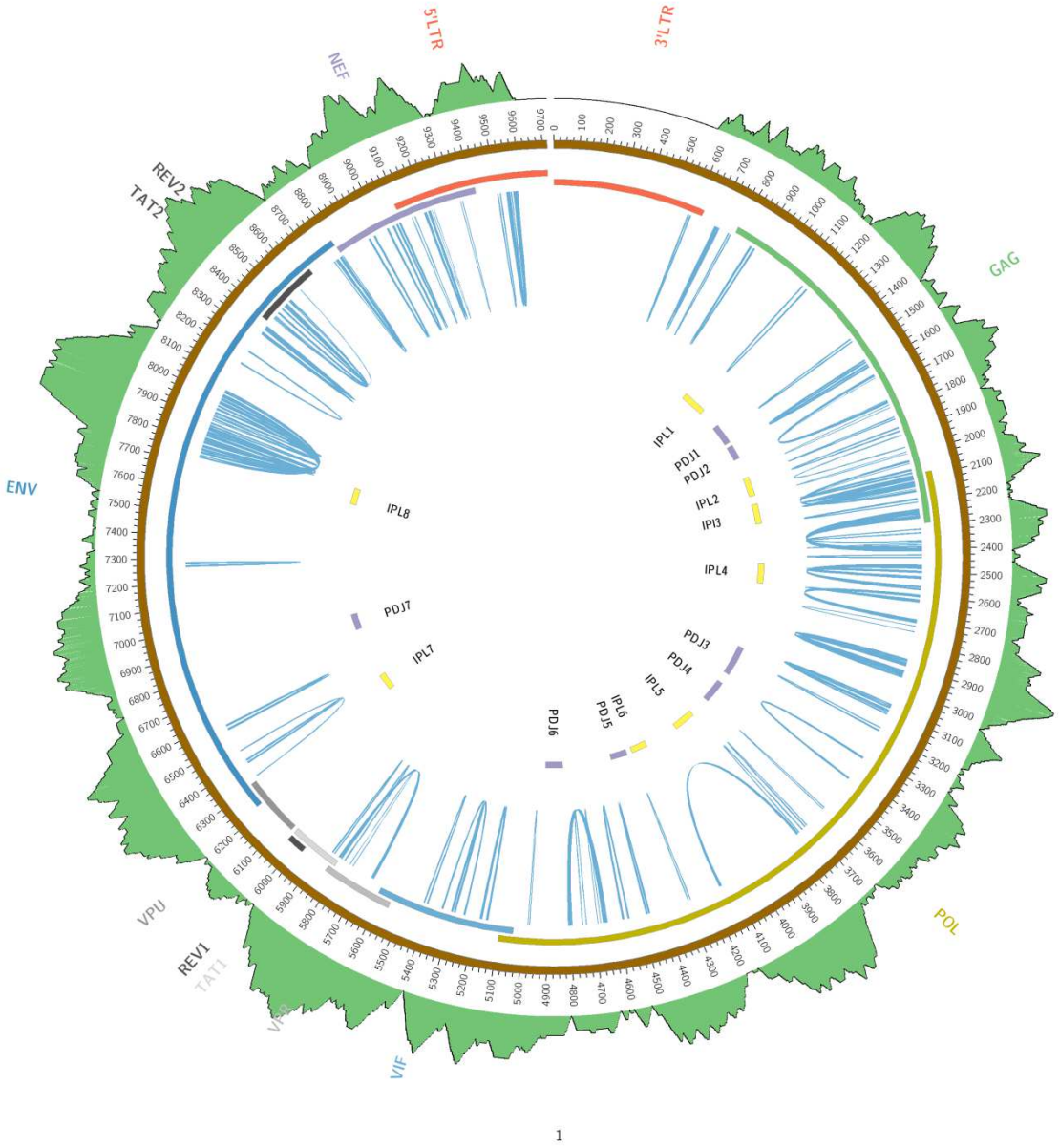


Figure 5.5B. RNA secondary structure for C subtype HIV sequence AY772699

Figure 5.5. Circos plot comparing NLM4-3 to HIV B and C subtype sequence.

Figure 5.5a represents a B subtype sequence from 2000 Netherland and Figure 5.5b represents a C subtype HIV sequence from 2004. The blue curves indicate base pairs that are conserved between NLM4-3 and the query sequence. The mountain plot on the outer ring represents the

base pairing depth for secondary structures in either B or C subtype sequences. Compared to NLM4-3, the B subtype sequence has more conserved stems than C subtype sequences.

HIV RNA Pasta Folder Comparing B and C subtypes

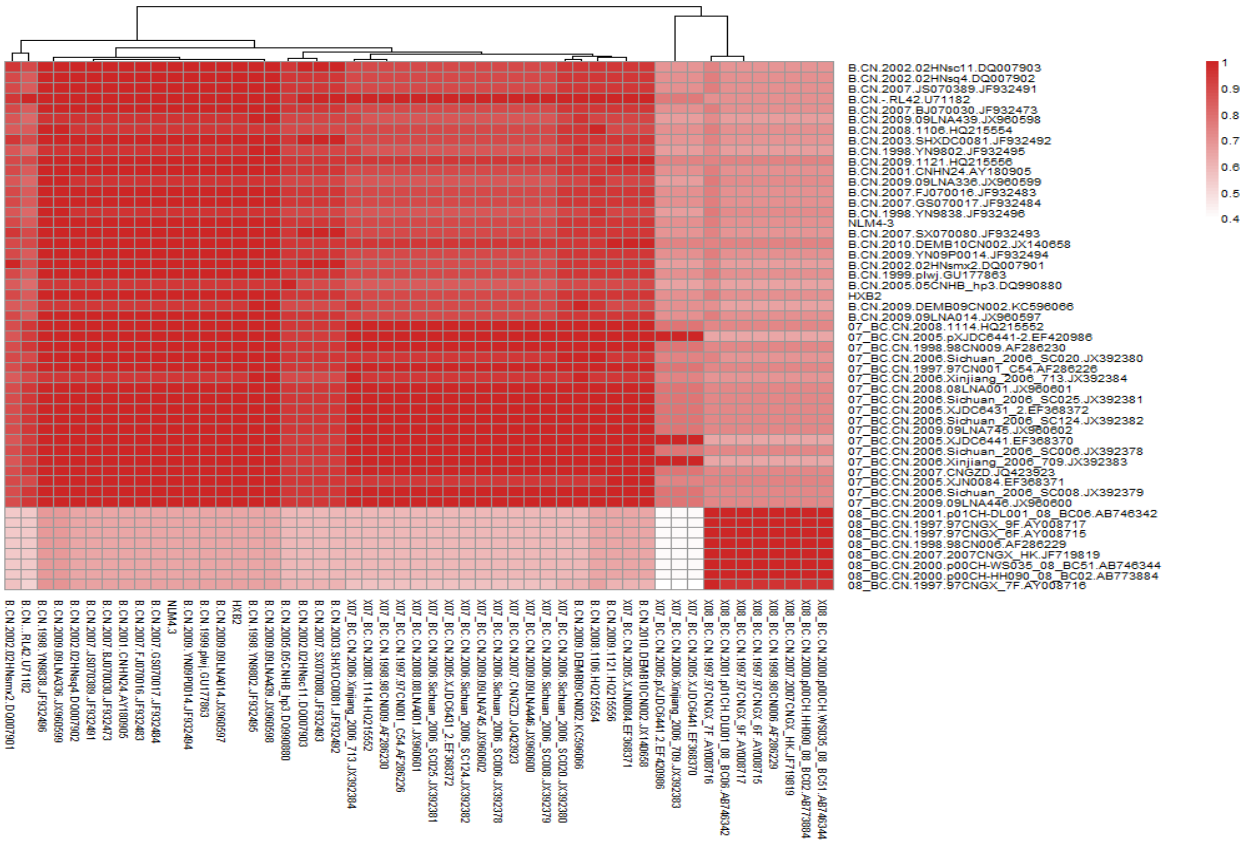
HIV RNA Pasta Folder used NLM4-3's SHAPE reactivity to predict query sequence's structure and further refined its structure using CONTRAFold. We examined the predicted structure using 2010 Reference B and C subtypes. A circos plot was used for visualizing the mountain plot representations of predicted RNA structure (Figure 5.5). Through circos plot, each HIV sequence was compared against NLM4-3, and the conserved stems were represented through blue arcs (Figure 5.5a-b). As indicated by Figure 5.5a and 5.5b, there were more conserved structures between NLM4-3 to the B subtype than NLM4-3 to the C subtype. We did identify conserved RNA structural motif such as RRE, Gag-Pol frameshift element. In the GAG region, numerous stems were conserved between B and C subtype.

Table 5.5. Present the comparison between, CRF07, CRF08, and pure B fragments. *Comparisons were made based on sequence identity and base pair similarity. Only regions that is of China B subtype was compared within this section. The phylogenetic bootstrap confidence is indicated in bracket for 07_BC and 08_BC. The comparison with NLM4-3 was based on RNA structure prediction without CONTRAFold refolding.*

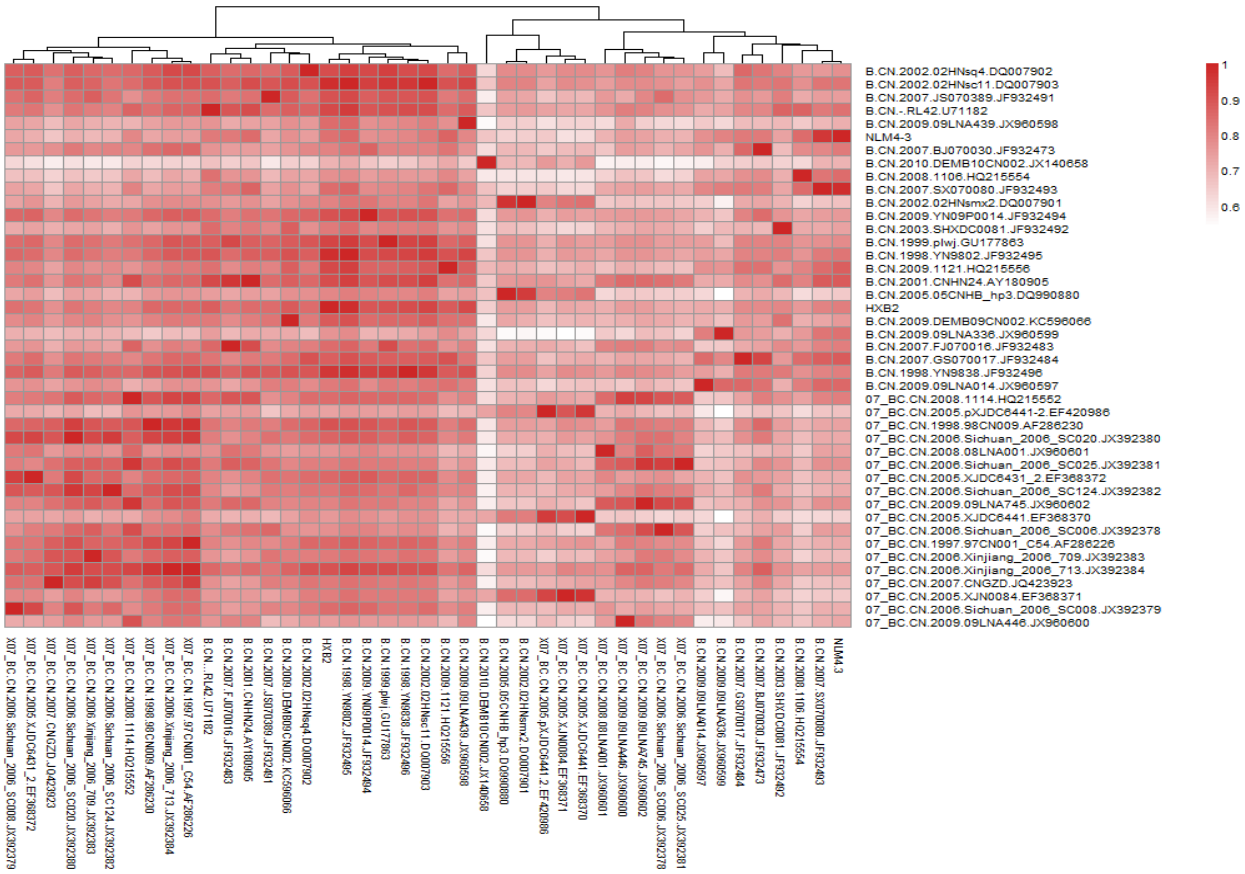
ID	Position	Subtype [Phylogenetic Bootstrap]	% Nucleotide Similarity with NLM43 (Without Refolding)	% BP Similarity with NLM43 (Without Refolding)	% Identity with 07_BC	% BP similarity with 07_BC	% Identity with 08_BC	% BP similarity with 08_BC	% Identity with pure B	% BP similarity with pure B
	1270-1411	07_BC[18%]	0.955	0.928	0.977	0.961	0.947	0.560	0.947	0.869
A	Gag	08_BC[85%]	0.936	0.931	0.947	0.706	0.989	0.995	0.931	0.699
		Pure_B	0.958	0.976	0.948	0.979	0.932	0.624	0.953	0.963
	2081-2621	07_BC[93%]	0.929	0.943	0.952	0.823	-	-	0.932	0.751
B	gag-pol	08_BC[NA]	-	-	-	-	-	-	-	-
		Pure_B	0.953	0.970	0.915	0.755	-	-	0.940	0.796
	3011-3311	07_BC[91%]	0.946	0.962	0.970	0.792	-	-	0.935	0.715
C	pol-rt	08_BC[NA]	-	-	-	-	-	-	-	-
		Pure_B	0.971	0.972	0.935	0.711	-	-	0.957	0.740
	5700-6311	07_BC[100%]	0.872	0.834	0.945	0.719	-	-	0.896	0.622
D	vpr-env	08_BC[NA]	-	-	-	-	-	-	-	-
		Pure_B	0.889	0.864	0.895	0.650	-	-	0.910	0.659
	8797-9059	07_BC[70%]	0.908	0.876	0.920	0.618	0.901	0.547	0.848	0.268
E	Nef	08_BC[97%]	0.897	0.886	0.914	0.549	0.962	0.748	0.850	0.279
		Pure_B	0.885	0.882	0.854	0.242	0.844	0.266	0.854	0.346

	1234-1691	07_BC[NA]	-	-	-	-	-	-	-	-
F	Gag	08_BC[96%]	0.949	0.964	-	-	0.986	0.903	0.946	0.786
		Pure_B	0.962	0.977	-	-	0.946	0.715	0.952	0.814
	2853-3150	07_BC[NA]	-	-	-	-	-	-	-	-
G	pol-rt	08_BC[52%]	0.966	0.990	-	-	0.990	0.924	0.957	0.838
		Pure_B	0.950	0.967	-	-	0.957	0.806	0.943	0.767
	8797-9417	07_BC[NA]	-	-	-	-	-	-	-	-
H	Nef	08_BC[100%]	0.849	0.795	-	-	0.973	0.759	0.790	0.453
		Pure_B	0.839	0.797	-	-	0.829	0.431	0.841	0.553

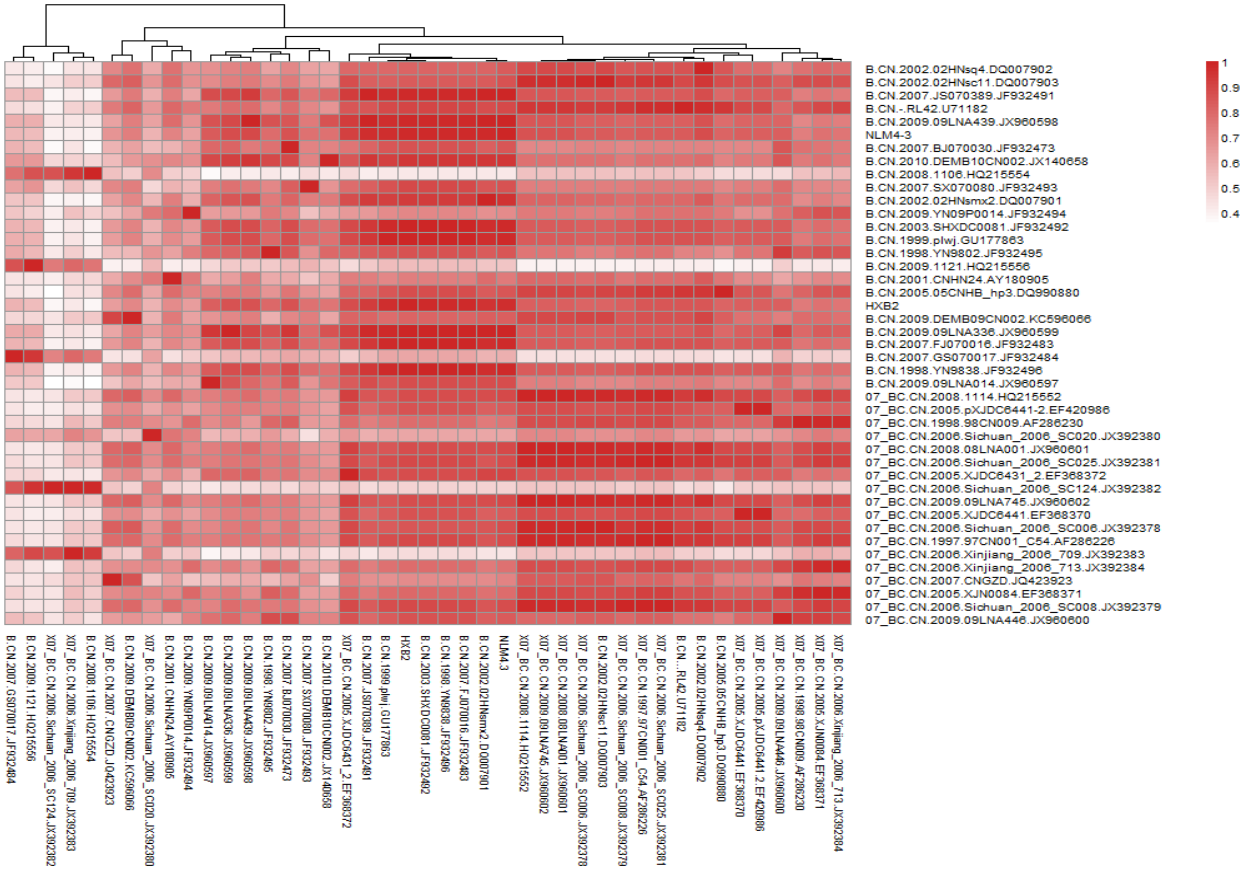
5.6A. HXB2 1270-1411nt



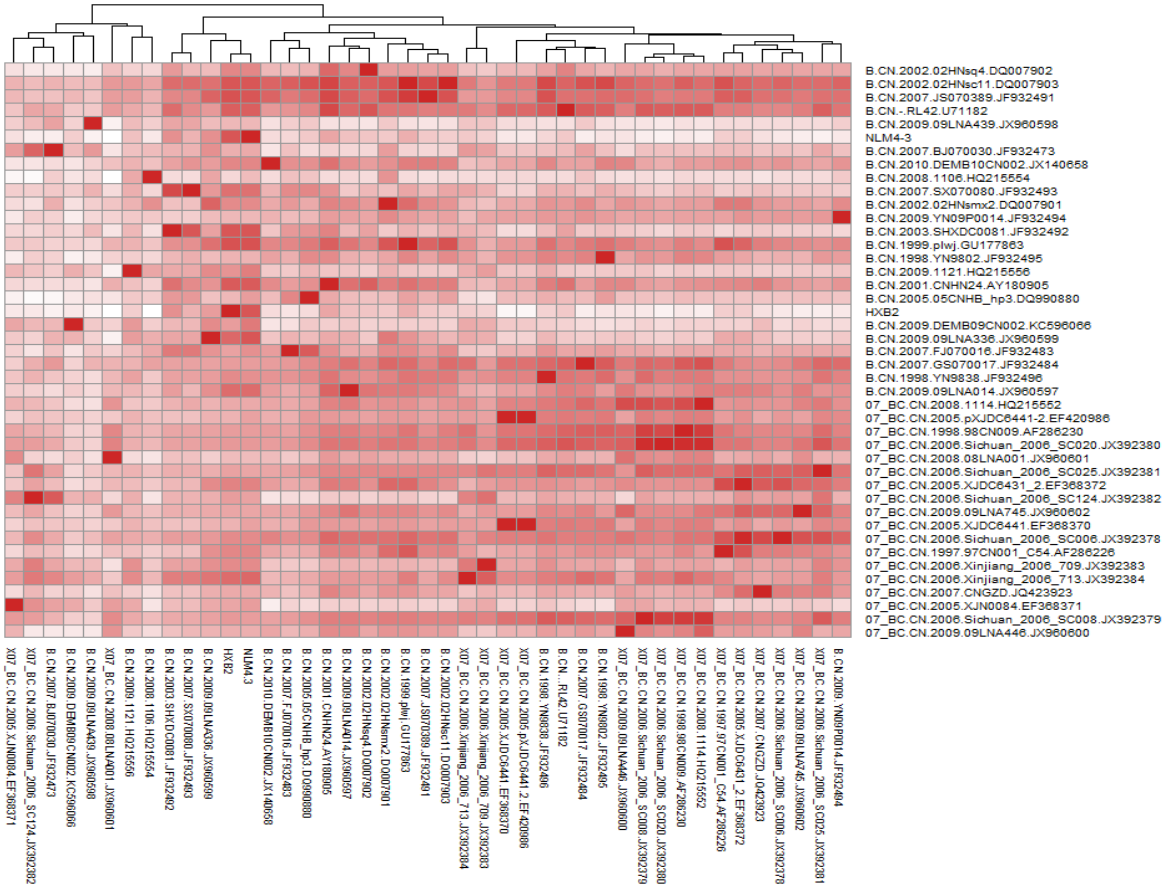
5.6B. HXB2 2081-2621nt



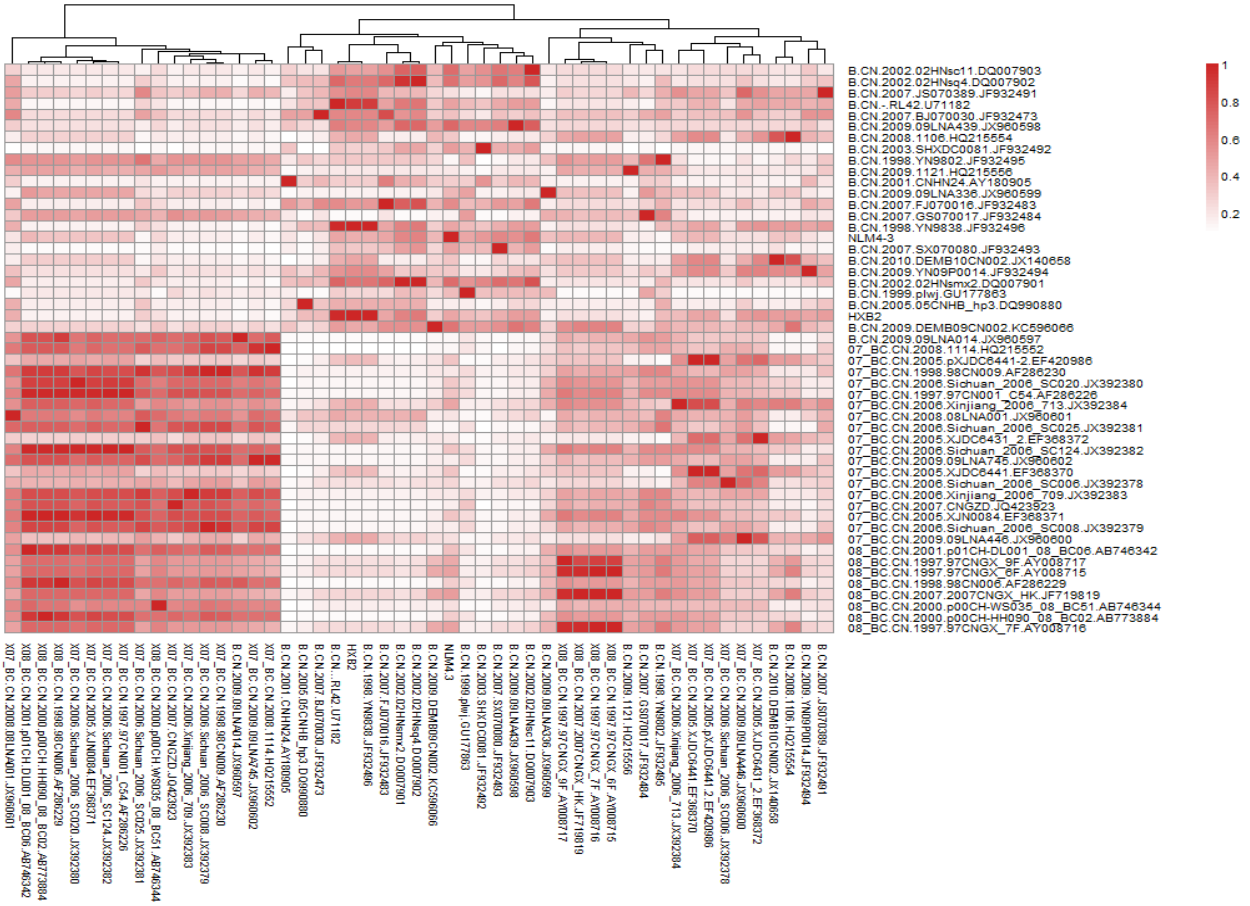
5.6C. HXB2 3011-3311nt



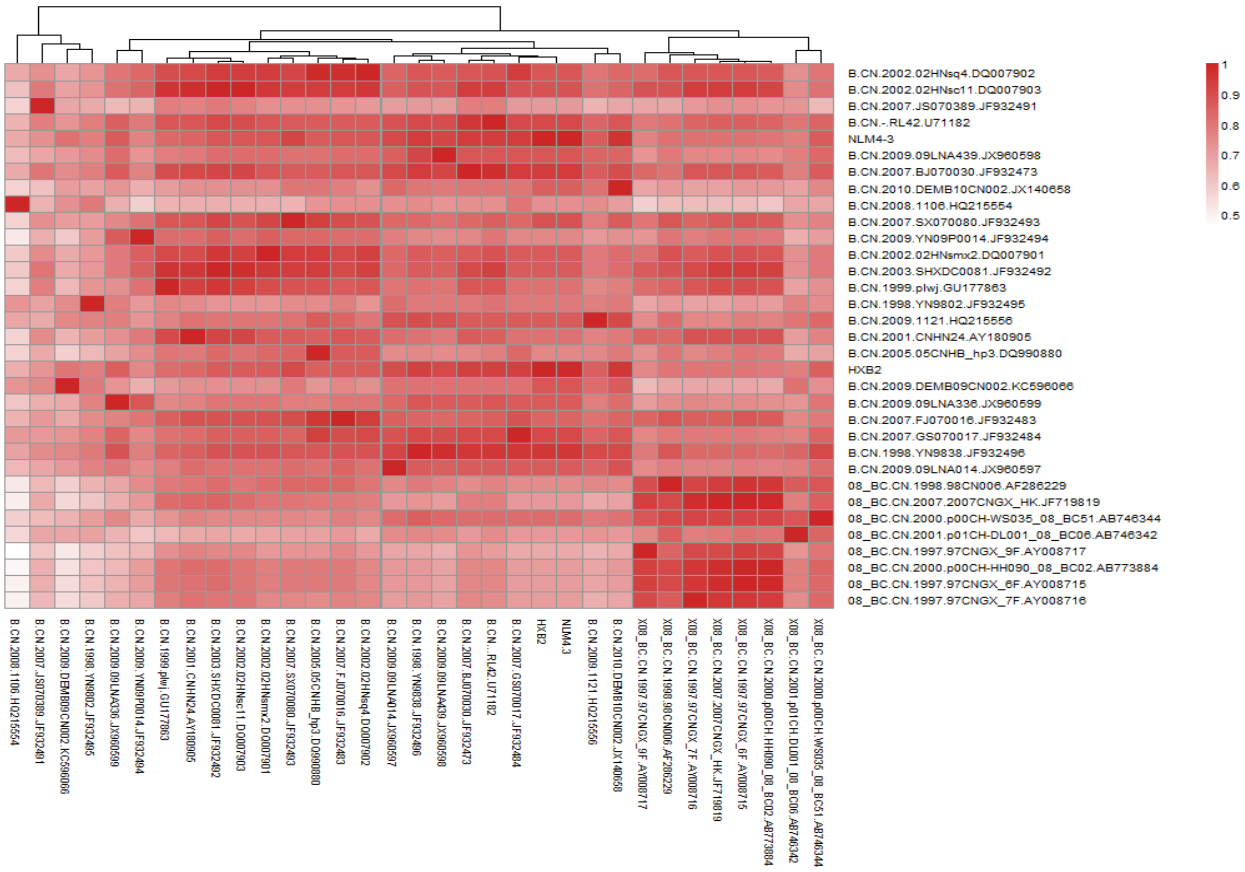
5.6D. HXB2_5700-6311nt



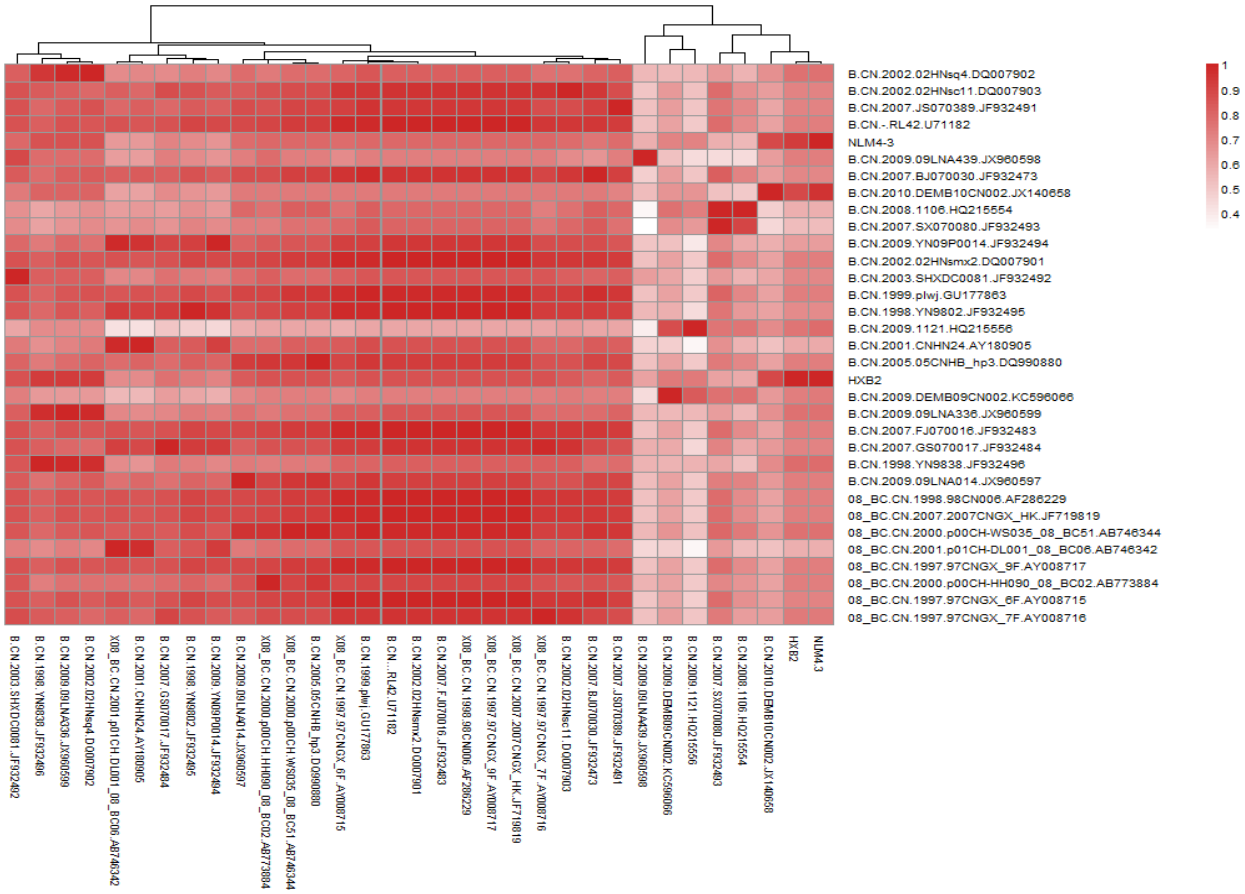
5.6E. HXB2 8797-9059nt



5.6F. HXB2 1234-1691nt



5.6G. HXB2 2853-3150nt



5.6H. HXB2 8797-9417nt

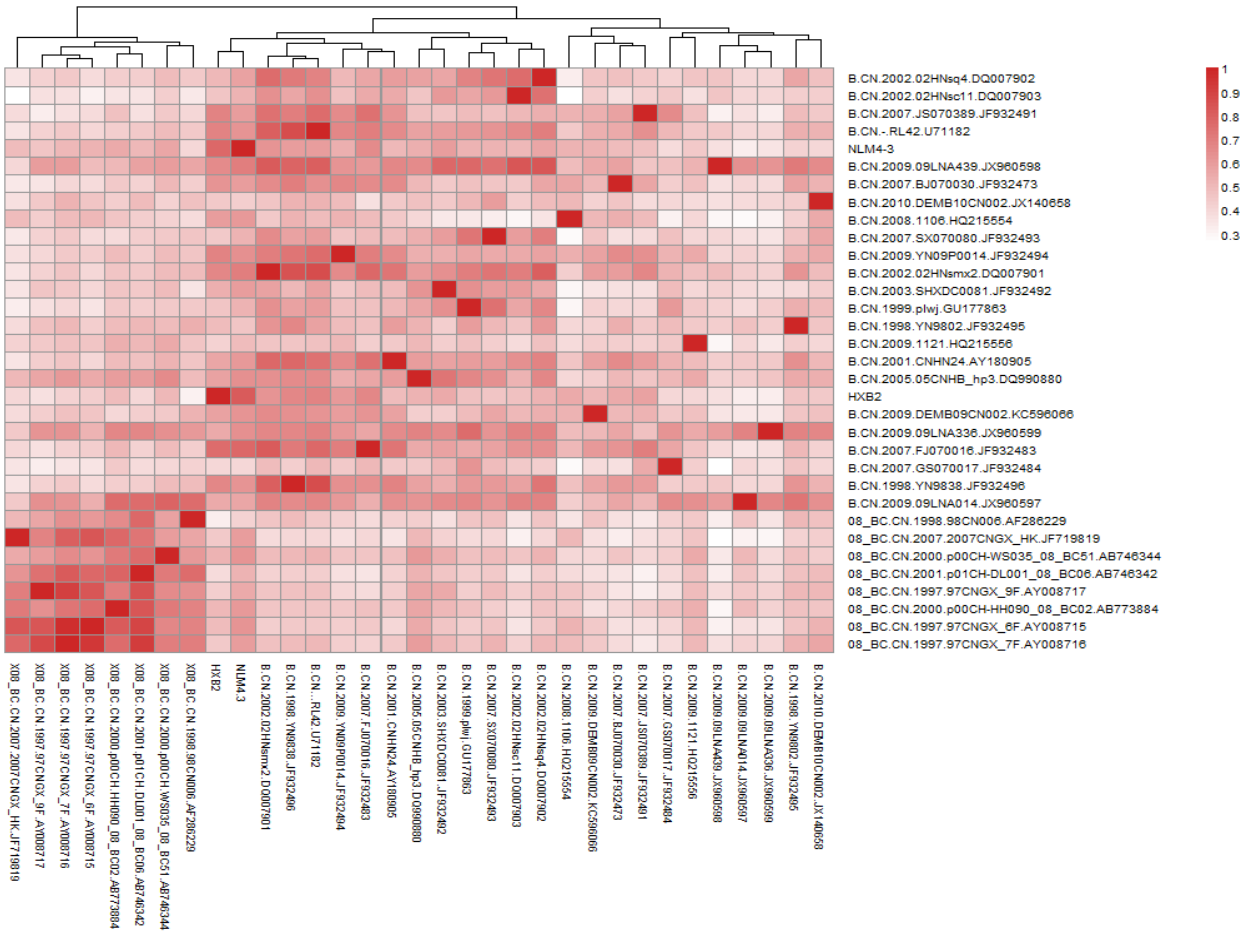


Figure 5.6. Heatmap RNA Structure comparison for China BC Recombinants.

The heatmap shows the comparison of the RNA structures using the sensitivity calculation (See Eq. 1). ***(a)*** *HXB2 1270-1411 covers the gag p24 protein for both CRF07 and CRF08 with their structure base pairs conserved.* ***(b)*** *HXB2 2081-2621 covers the gag-pol region (2081-2621nt) for CRF07.* ***(c)*** *HXB2 3011-3311 covers the pol RT region for CRF07.* ***(d)*** *HXB2_5700-6311 covers the VPR-TAT-REV-VPU-ENV region for CRF07.* ***(e)*** *HXB2 8797-9059 covers the NEF region for both CRF07 and CRF08.* ***(f)*** *HXB2 1234-1691 covers the gag region for CRF08.* ***(g)***

HXB2 2853-3150 covers the pol rt region for CRF07 and CRF08. ***(h)*** *HXB2 8797-9417 covers the NEF region for CRF08.*

RNA Structures in China BC Recombinants

As a case study, RNA secondary structure of B' subtype of BC recombinant (CRF07 and CRF08) sequences were compared against China pure B subtype fragments. The query's RNA structure was determined based on NLM4-3 covariate base pairing and CONTRAFold was used to refold the RNA sequence. To reduce bias of geographic sampling for CRF07-08 recombinants, all pure China B subtype sequences were ensured to possess similar sampling year as CRF07-08. Each B subtype fragment was labeled A through H (Table 5.4). Pairwise comparison of refolded RNA structures for all possible combination was performed, resulting in a matrix of sensitivity values. Finally, hierarchical clustering was used to cluster sequences with similar RNA structures (Figure 5.6). **Fragment A** covered the gag p24 protein (1270-1411nt) for both CRF07 and CRF08. In this region, more structural base pairs were conserved between CRF07 and B subtype than CRF08. Nevertheless, we noticed CRF07 contained a weak phylogenetic clustering. **Fragment B** covered the gag-pol region (2081-2621nt) for CRF07. After refolding, CRF07 shared on average 75% of the base pairing with B subtype fragments. **Fragment C** covered the pol RT region (3011-3311nt) for CRF07. The predicted RNA structure from CRF07 shared on average 71% of the base pairing with B subtype fragments. There were minor variations within CRF07; however, two distinct RNA structures were observed between 3053 to 3140nt. **Fragment D** covered the VPR-TAT-REV-VPU-ENV region (5700-6311nt) for CRF07. CRF07 shared only 72% of base pairing with itself and 65% of base pairing with B subtype. RNA structures within this region were highly variable even within the same subtype. **Fragment E** covered the NEF region (8797-9059nt) for both CRF07 and CRF08. RNA structure within this

region was also highly variable, with only 61%, 75%, and 34% base pair conservation for CRF07, CRF08, and pure B subtype respectively. **Fragment F** covered the gag region (1234-1691nt) for CRF08. RNA structure of CRF08 had less variation than pure B subtype. **Fragment G** covered the pol rt region (2853-3150nt) for CRF07 and CRF08. Fragment G had a higher sequence identity within pure B subtype than fragment C. Phylogenetically, CRF08 was weakly clustered together. There were roughly 81-84% base pairs conserved between CRF08 and pure B subtype. **Fragment H** covered the NEF region (8797-9417nt) for CRF08. There was a higher conservation of base pairs for CRF08 (76%) than pure B subtype (55%). For this region, RNA structure was very different between CRF08 and pure B subtypes.

Discussion

While controlling for NLM4-3 specific HIV subtype, geographical region, and time period, we found distinct regions with high conservation of covariate base pairing. These regions were primarily found in inter-protein linkage (IPL) and protein domain junction (PDJ) genomic elements, supported by a low bootstrapping p-value. These results indicate selective pressure exists for the conserved structural motifs, shaping the overall architecture of the B subtype RNA structure. The result is consistent with previous findings that lower SHAPE values tend to be localized within IPL and PDJ domains elements (Watts et al. 2009).

We tested different structure prediction program on HIV RNA secondary structures from Rfam. CONTRAFold had the highest F-measure and highest PPV value. Although CYK using BJK grammar as described by (Manzourolajdad et al. 2013, Wang et al. 2012) had the worst performance, the CYK algorithm experienced dramatic improvement with the inclusion of NLM4-3 SHAPE reference, making its performance as good as CONTRAFold. The result demonstrated SHAPE's ability to enhance RNA structure prediction. Based on our result, non-

thermodynamic algorithms was generally more accurate than thermodynamic algorithm. Since thermodynamic method is based on predicting the structure at the minimal free energy state, the result is further proof that HIV RNA structures might not be function at the minimal free energy state.

Although SHAPE reactivity was shown to be orthogonal to phylogenetically-based base pairing probabilities, we observed base pair differences between NLM4-3 and other subtypes. 20% base pairs were not preserved between NLM4-3 and B subtype while 32% base pairs were not conserved between NLM4-3 and non-B subtypes. We believe within the B subtype, NLM4-3's SHAPE reactivity can be used as a guide to perform the initial folding of the query sequence. As part of the novel pipeline developed, NLM4-3 SHAPE reactivity was used as auxiliary information to predict the HIV sequence structures. Considering the volatility of HIV sequence diversity, we expect errors to be introduced by the extrapolated SHAPE reactivity and refolding could compensate some of these errors. CONTRAFold was used for refolding, and 120nt was picked as the maximum size for us to confidently refold the sequence. We chose 120nt for the HIV RNA Pasta Fold pipeline because we found NLM4-3 RNA structures of 120nt or less covered 80% of all existing RNA base pairs, and RNA structure program clustering below this cutoff seems to share the most hierarchical structure (Figure 5.4). .

Recently, numerous studies were performed using NLM4-3 RNA structure as a model for other HIV B subtypes (van Hemert, van der Kuyl, and Berkhout 2013, Sanjuan and Borderia 2011, Snoeck et al. 2011). We wanted to examine the possible details of the structure differences of our analysis comparing B and C subtypes. RNA structures of B and C subtypes were predicted using HIV RNA Pasta Folder pipeline, and pairwise comparison between NLM4-3 to B subtype and NLM4-3 to C subtype was performed. The result showed a high coverage of

base pair stems conserved particularly for the B subtype. C subtype experienced a much greater reduction of conserved RNA stems when compared to NLM4-3. This confirmed that NLM4-3 should be able to model most RNA structures in B subtype; however, greater caution should be considered if SHAPE reactivity is applied to non-B subtype. We also put great caution that base pair conservation across the B subtype genome was not uniform; therefore, the NLM4-3's structure would definitely be inadequate for genomic region with high RNA structure variability.

For our case study, we compared RNA structure for eight B subtype fragments in CRF07 and CRF08 to that of pure B Chinese subtypes. For most of the regions, we found greater than 70% base pairs conserved after CONTRAFold refolding (Table 4). Although many of the recombinant fragments have a relatively high 92-95% shared nucleotide identity, the base pair conservation can be much lower in comparison. There are two regions that were found to have high variable RNA structures: fragment D (region between VPR and ENV) and fragment E and H (region in NEF). Within fragment D, several alternate splice acceptor and donors are known to be present in this region. RNA structures have been implicated to regulate the efficiency of the splicing event (Patterson, Yasuhara, and Ruzzo 2002, Buratti and Baralle 2004). The presence of loosely defined secondary structure in 3' splice site have resulted in improvement in splice site prediction indicating 3' splice site is sensitive to presence of structured RNA (Patterson, Yasuhara, and Ruzzo 2002, Buratti and Baralle 2004), and removing the RNA structure at the py-tract increased the use of A3 splice site (Jacquenet et al. 2001). In our result, we found RNA structure to be present within the NLM4-3 A3 py-tract but varied across HIV strains. With no clear conservation of structures between pure B and recombinant B subtypes, we suspect that shifts in RNA structure could potentially alter HIV gene expression profile to adapt to the new immune system. In addition, higher variability within the NEF region was observed

for fragment E and fragment H. CRF07 and CRF08 had a meager 50% conserved base pairing, but pure B subtypes had a much lower 25% of conserved base pair when compared to the CRF recombinant B subtypes. A higher sequence similarity was observed within fragment E than fragment H (contains NEF region overlapping 3'LTR); however, NEF H fragment has a higher conservation of RNA structure than NEF E fragment, indicating the possibility of weakly conserved RNA structures. NEF is known to enable high viral loads for HIV-1 and are found to be modulated throughout different stages of disease progression (Carl et al. 2001) early stop codon or partial deletion of NEF has resulted in slower progression to AIDS (Foster and Garcia 2008, 2007). Sequence diversity in NEF could also be attributed to the frequent CTL response targeting NEF's epitopes region (Deeks and Walker 2007); therefore, the associated fitness cost for changes in RNA structure is probably not high enough as compared to the fitness contribution by particular protein mutations. Although, previous studies did observe conserved RNA structure in the NEF region (Peleg, Trifonov, and Bolshoy 2003, Knoepfel and Berkhout 2013), only 25-50% of base pair conservation was found to be conserved in NEF. NEF RNA structure probably have less of an impact compared to NEF the protein coding region on overall HIV-1 fitness.

Research on HIV evolution, diversity, and viral fitness, would all be greatly aided through the enhanced resolution of HIV RNA structures. Currently, there is limited HIV RNA structure modeling resources. We believe our tool will be able to enhance exploration of RNA structure's impact on recombination, splice junctions, and epitope research particularly for the B subtype. Due to some dependency of the NLM4-3 RNA reference, our algorithm's accuracy would probably decrease for HIV sequences that diverged tremendously from the NLM4-3 strain.

Our HIV modeling pipeline will become more accurate with the availability of additional SHAPE experiment on other HIV strains and subtypes.

Availability

Webserver for HIV RNA Pasta Folder was made available at: hivtools.publichealth.uga.edu/SHAPE_PastaFold/PastaFold.php.

References

Abbink, T. E., and B. Berkhout. 2008. "RNA structure modulates splicing efficiency at the human immunodeficiency virus type 1 major splice donor." *J Virol* 82 (6):3090-8. doi: 10.1128/JVI.01479-07.

Adachi, A., H. E. Gendelman, S. Koenig, T. Folks, R. Willey, A. Rabson, and M. A. Martin. 1986. "Production of acquired immunodeficiency syndrome-associated retrovirus in human and nonhuman cells transfected with an infectious molecular clone." *J Virol* 59 (2):284-91.

Andronescu, M., A. Condon, H. H. Hoos, D. H. Mathews, and K. P. Murphy. 2007. "Efficient parameter estimation for RNA secondary structure prediction." *Bioinformatics* 23 (13):i19-28. doi: 10.1093/bioinformatics/btm223.

Barre-Sinoussi, F., J. C. Chermann, F. Rey, M. T. Nugeyre, S. Chamaret, J. Gruest, C. Dauguet, C. Axler-Blin, F. Vezinet-Brun, C. Rouzioux, W. Rozenbaum, and L. Montagnier. 2004. "Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS)." *Revista De Investigacion Clinica* 56 (2):126-129.

Benn, S., R. Rutledge, T. Folks, J. Gold, L. Baker, J. McCormick, P. Feorino, P. Piot, T. Quinn, and M. Martin. 1985. "Genomic heterogeneity of AIDS retroviral isolates from North America and Zaire." *Science* 230 (4728):949-51.

Bernhart, S. H., I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. 2008. "RNAalifold: improved consensus structure prediction for RNA alignments." *BMC Bioinformatics* 9:474. doi: 1471-2105-9-474 [pii] 10.1186/1471-2105-9-474.

Buratti, E., and F. E. Baralle. 2004. "Influence of RNA secondary structure on the pre-mRNA splicing process." *Mol Cell Biol* 24 (24):10505-14. doi: 10.1128/MCB.24.24.10505-10514.2004.

Carl, S., T. C. Greenough, M. Krumbiegel, M. Greenberg, J. Skowronski, J. L. Sullivan, and F. Kirchhoff. 2001. "Modulation of different human immunodeficiency virus type 1 Nef functions during progression to AIDS." *J Virol* 75 (8):3657-65. doi: 10.1128/JVI.75.8.3657-3665.2001.

Deeks, S. G., and B. D. Walker. 2007. "Human immunodeficiency virus controllers: mechanisms of durable virus control in the absence of antiretroviral therapy." *Immunity* 27 (3):406-16. doi: 10.1016/j.immuni.2007.08.010.

Dirks, R. M., and N. A. Pierce. 2003. "A partition function algorithm for nucleic acid secondary structure including pseudoknots." *J Comput Chem* 24 (13):1664-77. doi: 10.1002/jcc.10296.

Dirks, R. M., and N. A. Pierce. 2004. "An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots." *J Comput Chem* 25 (10):1295-304. doi: 10.1002/jcc.20057.

Do, C. B., D. A. Woods, and S. Batzoglou. 2006. "CONTRAfold: RNA secondary structure prediction without physics-based models." *Bioinformatics* 22 (14):e90-8. doi: 22/14/e90 [pii] 10.1093/bioinformatics/btl246.

Durbin, R., Eddy,S., Krogh,A. and Mitchison,G. 1998b. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge, UK: Cambridge University Press.

Eddy, S. R. 2004. "How do RNA folding algorithms work?" *Nat Biotechnol* 22 (11):1457-8. doi: 10.1038/nbt1104-1457.

Foster, J. L., and J. V. Garcia. 2007. "Role of Nef in HIV-1 replication and pathogenesis." *Adv Pharmacol* 55:389-409. doi: 10.1016/S1054-3589(07)55011-8.

Foster, J. L., and J. V. Garcia. 2008. "HIV-1 Nef: at the crossroads." *Retrovirology* 5:84. doi: 10.1186/1742-4690-5-84.

Gardner, P. P., J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. 2009. "Rfam: updates to the RNA families database." *Nucleic Acids Res* 37 (Database issue):D136-40. doi: gkn766 [pii] 10.1093/nar/gkn766.

Hajdin, C. E., S. Bellaousov, W. Huggins, C. W. Leonard, D. H. Mathews, and K. M. Weeks. 2013. "Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots." *Proc Natl Acad Sci U S A* 110 (14):5498-503. doi: 10.1073/pnas.1219988110.

Ho, D. D., A. U. Neumann, A. S. Perelson, W. Chen, J. M. Leonard, and M. Markowitz. 1995. "Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection." *Nature* 373 (6510):123-6. doi: 10.1038/373123a0.

Hofacker, I. L., and P. F. Stadler. 2006. "Memory efficient folding algorithms for circular RNA secondary structures." *Bioinformatics* 22 (10):1172-6. doi: btl023 [pii] 10.1093/bioinformatics/btl023.

Hogeweg, P., and B. Hesper. 1984. "Energy directed folding of RNA sequences." *Nucleic Acids Res* 12 (1 Pt 1):67-74.

J, Felsenstein. 1989. "PHYLIP - Phylogeny Inference Package (Version 3.2)." *Cladistics* (5):164-166.

Jacquenet, S., D. Ropers, P. S. Bilodeau, L. Damier, A. Mougin, C. M. Stoltzfus, and C. Branlant. 2001. "Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing." *Nucleic Acids Res* 29 (2):464-78.

Knoepfel, S. A., and B. Berkhout. 2013. "On the role of four small hairpins in the HIV-1 RNA genome." *RNA Biol* 10 (4):540-52. doi: 10.4161/rna.24133.

Knoepfel, Stefanie, and Ben Berkhout. 2011. "Phylogenetic screen for important RNA structure motifs in the HIV-1 genome %U http://www.retrovirology.com/content/8/S2/P40." *Retrovirology* 8 %@ 1742-4690 (Suppl 2 %M doi:10.1186/1742-4690-8-S2-P40):P40.

Korber, B., B. Foley, C. Kuiken, S. Pillai, and J. Sodroski. 1998. "Numbering Positions in HIV Relative to HXB2CG." In *Human Retroviruses and AIDS 1998*, edited by C. Kuiken Korber, B. Foley, B. Hahn, F. McCutchan, J. Mellors, and J. and Sodroski. Los Alamos, NM: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory.

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones, and M. A. Marra. 2009. "Circos: an information aesthetic for comparative genomics." *Genome Res* 19 (9):1639-45. doi: 10.1101/gr.092759.109.

Low, J. T., S. A. Knoepfel, J. M. Watts, O. ter Brake, B. Berkhout, and K. M. Weeks. 2012. "SHAPE-directed discovery of potent shRNA inhibitors of HIV-1." *Mol Ther* 20 (4):820-8. doi: 10.1038/mt.2011.299.

Low, J. T., and K. M. Weeks. 2010. "SHAPE-directed RNA secondary structure prediction." *Methods* 52 (2):150-8. doi: S1046-2023(10)00161-1 [pii] 10.1016/j.ymeth.2010.06.007.

Lu, Z. J., J. W. Gloor, and D. H. Mathews. 2009. "Improved RNA secondary structure prediction by maximizing expected pair accuracy." *RNA* 15 (10):1805-13. doi: 10.1261/rna.1643609.

Lucks, J. B., S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin. 2011. "Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-Seq)." *Proc Natl Acad Sci U S A* 108 (27):11063-8. doi: 10.1073/pnas.1106501108.

Malmberg, R. L., T. I. Shaw, and L. Cai. 2010. "RNApasta: a tool for analysis of RNA structural alignments." *Int J Bioinform Res Appl* 6 (6):571-83. doi: EJ4M7207161M7708 [pii].

Manzourolajdad, A., Y. Wang, T. I. Shaw, and R. L. Malmberg. 2013. "Information-theoretic uncertainty of SCFG-modeled folding space of the non-coding RNA." *J Theor Biol* 318:140-63. doi: 10.1016/j.jtbi.2012.10.023.

Nussinov, R., and A. B. Jacobson. 1980. "Fast algorithm for predicting the secondary structure of single-stranded RNA." *Proc Natl Acad Sci U S A* 77 (11):6309-13.

Patterson, D. J., K. Yasuhara, and W. L. Ruzzo. 2002. "Pre-mRNA secondary structure prediction aids splice site prediction." *Pac Symp Biocomput*:223-34.

Peleg, O., S. Brunak, E. N. Trifonov, E. Nevo, and A. Bolshoy. 2002. "RNA secondary structure and squence conservation in C1 region of human immunodeficiency virus type 1 env gene." *AIDS Res Hum Retroviruses* 18 (12):867-78. doi: 10.1089/08892220260190353.

Peleg, O., E. N. Trifonov, and A. Bolshoy. 2003. "Hidden messages in the nef gene of human immunodeficiency virus type 1 suggest a novel RNA secondary structure." *Nucleic Acids Res* 31 (14):4192-200.

Pollom, E., K. K. Dang, E. L. Potter, R. J. Gorelick, C. L. Burch, K. M. Weeks, and R. Swanstrom. 2013. "Comparison of SIV and HIV-1 genomic RNA structures reveals impact of sequence evolution on conserved and non-conserved structural motifs." *PLoS Pathog* 9 (4):e1003294. doi: 10.1371/journal.ppat.1003294.

Reuter, J. S., and D. H. Mathews. 2010. "RNAstructure: software for RNA secondary structure prediction and analysis." *BMC Bioinformatics* 11:129. doi: 10.1186/1471-2105-11-129.

Sanjuan, R., and A. V. Borderia. 2011. "Interplay between RNA structure and protein evolution in HIV-1." *Mol Biol Evol* 28 (4):1333-8. doi: 10.1093/molbev/msq329.

Schroeder, S. J. 2009. "Advances in RNA structure prediction from sequence: new tools for generating hypotheses about viral RNA structure-function relationships." *J Virol* 83 (13):6326-34. doi: JVI.00251-09 [pii] 10.1128/JVI.00251-09.

Shaw, T. I., and M. Zhang. 2013. "HIV N-linked glycosylation site analyzer and its further usage in anchored alignment." *Nucleic Acids Res* 41 (Web Server issue):W454-8. doi: 10.1093/nar/gkt472.

Shaw, T. I., and M. Zhang. 2013. "HIV N-linked glycosylation site analyzer and its further usage in anchored alignment." *Nucleic Acids Res* 41 (Web Server issue):W454-8. doi: 10.1093/nar/gkt472.

Snoeck, J., J. Fellay, I. Bartha, D. C. Douek, and A. Telenti. 2011. "Mapping of positive selection sites in the HIV-1 genome in the context of RNA and protein structural constraints." *Retrovirology* 8:87. doi: 10.1186/1742-4690-8-87.

Sukosd, Z., B. Knudsen, J. Kjems, and C. Pedersen. 2012. "PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data." *Bioinformatics*. doi: 10.1093/bioinformatics/bts488.

Sukosd, Z., B. Knudsen, M. Vaerum, J. Kjems, and E. S. Andersen. 2011. "Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars." *BMC Bioinformatics* 12:103. doi: 1471-2105-12-103 [pii] 10.1186/1471-2105-12-103.

van der Kuyl, A. C., and B. Berkhout. 2012. "The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus." *Retrovirology* 9:92. doi: 10.1186/1742-4690-9-92.

van der Kuyl, A. C., and B. Berkhout. 2012. "The biased nucleotide composition of the HIV genome: a constant factor in a highly variable virus." *Retrovirology* 9:92. doi: 10.1186/1742-4690-9-92.

van Hemert, F. J., A. C. van der Kuyl, and B. Berkhout. 2013. "The A-nucleotide preference of HIV-1 in the context of its structured RNA genome." *RNA Biol* 10 (2):211-5. doi: 10.4161/rna.22896.

Wang, Y., A. Manzour, P. Shareghi, T. I. Shaw, Y. W. Li, R. L. Malmberg, and L. Cai. 2012. "Stable stem enabled Shannon entropies distinguish non-coding RNAs from random backgrounds." *BMC Bioinformatics* 13 Suppl 5:S1. doi: 10.1186/1471-2105-13-S5-S1.

Washietl, S., I. L. Hofacker, P. F. Stadler, and M. Kellis. 2012. "RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction." *Nucleic Acids Res* 40 (10):4261-72. doi: 10.1093/nar/gks009.

Watts, J. M., K. K. Dang, R. J. Gorelick, C. W. Leonard, J. W. Bess, Jr., R. Swanstrom, C. L. Burch, and K. M. Weeks. 2009. "Architecture and secondary structure of an entire HIV-1 RNA genome." *Nature* 460 (7256):711-6. doi: nature08237 [pii] 10.1038/nature08237.

Westerhout, E. M., M. Ooms, M. Vink, A. T. Das, and B. Berkhout. 2005. "HIV-1 can escape from RNA interference by evolving an alternative structure in its RNA genome." *Nucleic Acids Res* 33 (2):796-804. doi: 10.1093/nar/gki220.

CHAPTER 6

RELATIVE RIBOSOMAL FRAMESHIFT EFFICIENCY ACROSS HIV-1 SUBTYPES[1]

[1]Timothy I. Shaw, Russell L. Malmberg, Ming Zhang. To be submitted to *Journal of Virology*.

Abstract

HIV-1's programmed -1 ribosomal frameshift element is a critical regulator for the Gag to Gag-Pol polyprotein production; genetic changes to the frameshift element could alter frameshift efficiency, viral infectivity and viral production. In this study, frameshift efficiency is estimated based on RNA secondary structural features. The frameshift element's RNA upper stem base pairing stability and variability are found to be associated with frameshift efficiency. Frameshift efficiency is found to be predictive for intrasubtype fitness. Our frameshift efficiency model is able to predict 75% of the intrasubtype fitness competition assay results, but failed to predict intersubtype competition assay results. Across HIV subtypes, each subtype has a different frameshift efficiency range. Examining the sequence space for each subtype and their recombination has revealed that the two recombinant subtypes often have an overlapping sequence space and range of frameshift efficiency. Risk factors (Intravenous drug users and MSM) associated with higher multiplicity of infection (MOI) tends to have a higher frameshift efficiency possibly an artifact of competing viral strands. Sequences possessing two protease specific drug resistant mutations tend to have a higher frameshift efficiency, suggesting that the virus might be increasing its frameshift efficiency to compensate for the mutation's fitness cost.

Introduction

HIV-1 is an etiological agent for AIDS, a disease that has posed a significant burden to public health and economic growth worldwide. HIV-1 M group is responsible for majority of the HIV infection cases, and can be divided into nine subtypes: A, B, C, D, F, G, H, J and K. In addition, there are at least 55 circulating recombinant strands that constitute over 20% of the HIV isolates (Osmanov et al. 2002, Peeters and Sharp 2000, Peeters, Toure-Kane, and Nkengasong 2003, Hemelaar et al. 2006). With HIV's rapid mutation rate, the risk of antiretroviral drug resistance is a persistent threat (Ndung'u and Weiss 2012). Given the important role of frameshifting in HIV replication (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009), there is an emerging interests in using frameshift element as a drug target (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009, Hung et al. 1998, Dinman, Ruiz-Echevarria, and Peltz 1998).

Aside from retroviruses, programmed ribosomal frameshifting is ubiquitous across eukaryotes and prokaryotes (Farabaugh 1996, Chamorro, Parkin, and Varmus 1992, Jacks, Madhani, et al. 1988, Jacks, Power, et al. 1988, Jacks et al. 1987, Jacks and Varmus 1985). Comparative analysis of frameshift element sequences has identified a canonical structure within the frameshift sites, suggesting that evolutionary convergence has allowed it to be prevalent across the organism spectrum (Farabaugh 1996). The -1 programmed ribosomal frameshift (PRF) primarily occurs on a slippery heptamer which was first found in Rous Sarcoma Virus (Jacks, Madhani, et al. 1988) and its motif later verified: XXXYYYZ (Farabaugh 1996, Brierley, Digard, and Inglis 1989). The frameshifting capacity was shown to correlate to the mechanical strength of the pseudoknot (Hansen et al. 2007, Green et al. 2008) downstream of the slippery heptamer (ten Dam, Pleij, and Bosch 1990). The proposed mechanism for the frameshift event is driven

by the stability of the RNA structures that hinders the unwinding of the elongating ribosome, causing the ribosome to pause and slip in the -1 nucleotide position (Green et al. 2008, Plant and Dinman 2005).

Although thermodynamic stability from the pseudoknot structure can effectively induce -1 PRF (Nixon and Giedroc 2000, Giedroc, Theimer, and Nixon 2000, Nixon et al. 2002), a stable stem-loop structure can also induce -1 PRF as effectively as a pseudoknot structure (Yu et al. 2011). For example, the frameshift elements in HIV-1 and SIV possess a bipartite stem-loop instead of a pseudoknot, which is sufficient to stimulate gag-pol frameshift (Gaudin et al. 2005, Marcheschi, Staple, and Butcher 2007, Staple and Butcher 2005). HIV-1's frameshift element consists of a lower and an upper stem loop, and the strength of the upper stem base pairing is suspected to be the driver for determining frameshift efficiency (Gaudin et al. 2005, Staple and Butcher 2005).

The HIV-1 Gag-Pol transcript can code for two polyproteins (Parkin, Chamorro, and Varmus 1992). The frameshift mechanism is responsible for maintaining a well-regulated ratio of Gag to Gag-Pol to ensure efficient assembly of infectious virus particles (Park and Morrow 1991). The HIV-1 heptanucleotide is primarily composed of a UUUUUUA slippery sequence, with an 8nt spacer located between the slippery sequence and the RNA stem loop (Mouzakis et al. 2013, Kollmus et al. 1994). The -1 PRF can roughly induce a frameshift 5% of the time in HIV-1, and alteration of its frameshift efficiency can inhibit viral replication (Dulude et al. 2006, Hung et al. 1998, Shehu-Xhilaga, Crowe, and Mak 2001). These results suggest that an effective range of frameshift efficiency exists for individual HIV virus.

Experimentally, frameshift efficiency has been characterized through a rabbit reticulocyte lysate system (Jacks, Power, et al. 1988) or a dual luciferase reporter system (Gareiss and Miller

2009). The dual luciferase construct consists of luciferase reporter genes and a frameshift inducing element. In the dual luciferase system, Renilla luciferase (Rluc) is incorporated upstream of the HIV-1 frameshift region and firefly luciferase (Fluc) is incorporated downstream of the HIV-1 frameshift region. Both frameshift region and Fluc are placed at the -1 frame. In the control luciferase construct, the slippery site is mutated to prevent a frameshift with Fluc being placed at the 0 frame. The percentage of frameshift is calculated based on the ratio of experiment's luminescence (firefly/renilla) divided by the control's luminescence (Mouzakis et al. 2013).

Several studies have proposed RNA structure stability affects frameshift efficiency; however, they lack strong statistical support (Bidou et al. 1997, Mouzakis et al. 2013, Telenti et al. 2002). Telenti et al. (Telenti et al. 2002) has found that a reduced frameshift activity affects viral replication, and reduction in thermodynamic stability reduces frameshift efficiency, but it is modeled by a weak R^2 of 0.3364. Mouzakis et al. (Mouzakis et al. 2013) has indicated that the thermodynamic stability of 3-4bp upper stem-loop base at the mRNA entrance channel of the ribosome can impact its frameshift efficiency. While Mouzakis et al (Mouzakis et al. 2013) have indicated different thermodynamic combination of 3-4bp base stem can be used to model frameshift efficiency, our review has indicated that sequences applied in Mouzakis et al.'s study are often absent in existing HIV sequence database (Los-Alamos-HIV-Sequence-Database).

The frameshift efficiency's impact on virus replication and pathogenicity has been demonstrated in multiple studies (Telenti et al. 2002, Bidou et al. 1997, Mouzakis et al. 2013). The frameshift efficiency ratio was found to be related to HIV replication capabilities (Dulude et al. 2006). As replication capacity is an integral contributor to viral fitness (Prado et al. 2005, De Luca 2006), we hypothesize that frameshift efficiency can influence viral fitness. Using

sequences derived from Africa and North America, Abraha et al. has evaluated ex vivo the pathogenic fitness of HIV subtypes (Abraha et al. 2009); they have revealed that B and D isolates are much more fit than A and C isolates while the A subtype is slightly more fit than the C subtype (Abraha et al. 2009). However, in an alternative study in an Indian population, subtype C is found to be more fit than the A subtype (Rodriguez et al. 2009). These studies indicate that although subtypes can be ranked based on relative fitness, the underlying mechanism driving each subtype's fitness demands further investigation. In our study, we will attempt to evaluate inter and intra-subtype's sequence fitness and their association to frameshift efficiency.

Although, the frameshift element is relatively conserved in HIV-1, variation in nucleotide composition has been observed across HIV subtypes (Baril et al. 2003). Characterizing variations of frameshift element across HIV-1 subtypes will enhance future exploration of using the frameshift element as a drug target or therapeutic vaccine. To this date, there has not yet been a comprehensive analysis of frameshift element RNA structure across HIV-1 subtypes and recombinants. Here, we examined frameshift element and their RNA structure diversity for each HIV subtype. The model developed in the study can help us elucidate the HIV replication mechanism and HIV fitness specifically driven by RNA secondary structure variation.

Methods

Sequence Data: Frameshift Element

All sequences in this study have been retrieved from GenBank and the Los Alamos HIV Sequence Database (Los-Alamos-HIV-Sequence-Database). Genotyping is performed by jpHMM (Schultz et al. 2006, Zhang, Schultz, et al. 2006). Sequences with more than one

subtype are categorized as “Recombinant”, and sequences with only one subtype are categorized as “Pure Subtype”. Recombinant breakpoints are identified based on jpHMM’s uncertainty region assignment.

The HIV frameshift element (Rfam accession RF00480) is predicted using Infernal 2.0 (Nawrocki, Kolbe, and Eddy 2009). The Infernal output is parsed, and sequences containing a UUUUUUA heptanucleotide are extracted. RNAstructure (Reuter and Mathews 2010) is used to predict the RNA secondary structure, and the predicted structure is drawn using VARNA (Darty, Denise, and Ponty 2009). Sequences with recombinant breakpoint overlapping the frameshift element (ambiguous assignment of subtypes) are removed from analysis. To analyze the diversity of frameshift elements, sequences with similar frameshift element are grouped together. All frameshift element groups with less than 10 sequences are filtered out, and the remaining sequences are phylogenetically assessed through PhyML 2.0, a maximum likelihood phylogenetic approach (Zhang et al. 2010). R pheatmap is used to construct the heatmap based on normalizing the frequencies for each subtype and recombinant.

Frameshift element network

A network is constructed to visualize the mutation and sequence variation for each grouped frameshift element (see previous section). Each frameshift element is represented by a node, and a line is used to indicate a single-nucleotide polymorphism between two frameshift elements. For each node, a pie chart is embedded showing the proportion of the subtype covering the frameshift element. The proportion for each subtype is normalized based on the subtype’s percentage occupied in each of the node. The network is analyzed using the betweenness centrality function (Brandes 2001) from Cytoscape 2.8.3 (Kohl, Wiese, and Warscheid 2011), a

global measure of centrality in the network. The graph are generated using Cytoscape 2.8.3 (Kohl, Wiese, and Warscheid 2011).

RNA Structure Analysis

The frameshift element RNA structure free energy base stability is calculated through the method described by Mouzakis et al. (Mouzakis et al. 2013): Turner's RNA base pairing energy parameter (Xia et al. 1998) with an updated GU parameter (Chen et al. 2012). In addition, the HIV RNA structure fold stability are assessed through the Unpaired Structural Entropy (USE) (Shaw et al. 2011). USE calculates the certainty for bases to be unpaired based on a derivation of Shannon Entropy (eq. 1). S corresponds to each given sequence. P_{i0} corresponds to the non-pairing probability of nucleotide at position i, and N corresponds to the length of the sequence. NUPACK is used to calculate the base pair probabilities for the RNA structures (Dirks and Pierce 2004, 2003).

$$USE(S) = -\sum_{i=1}^{N} \frac{p_{i0}}{L_0} log \frac{p_{i0}}{L_0}, where\ L_0 = \sum_{i=1}^{N} p_{i0} \quad [Equation\ 1]$$

Modeling Frameshift Efficiency

The frameshift element efficiency data used to generate our model is gathered from four different papers (Mouzakis et al. 2013, Chen et al. 1997, Bidou et al. 1997, Baril et al. 2003). Baril et al (Baril et al. 2003) have studied the frameshift efficiency based on analyzing frameshift elements across subtypes, while other studies are based on mutagenesis experiments on B subtype sequences. Considering each dataset's differences in viral and cellular factors, the frameshift efficiency from the NL4-3 HIV strand is used to normalize each dataset's frameshift efficiency. A normalization factor is calculated based on adjusting each dataset's NL4-3 frameshift

efficiency to 5% (Jacks, Power, et al. 1988, Mouzakis et al. 2013). Frameshift efficiency is plotted as a function of thermodynamic stability and USE (See Eq 1), and a nonlinear model is fitted to the dataset via nonlinear least squares from R 2.15.1. The Mean Squared Error (MSE) is used to assess the data fitting error. Previously, Mouzakis et al (Mouzakis et al. 2013) have described a linear model transformed by an exponential decay function fitting thermodynamic stability to frameshift efficiency (eq. 2). Mouzakis et al's model finds a strong R^2; however, the R^2 is driven by outlying data points (R^2 decreased significantly after removing outliers). The correlation is evaluated based on Spearman Rank correlation coefficient. Variable "y" is the frameshift efficiency. In equation 2, variable "w" represents either USE or 3bp Free Energy. For equation 3, a nonlinear model is constructed based on 3bp Free Energy and USE (eq. 3). Incorporating both RNA structural features improves the modeling of frameshift efficiency. Variable "x" represent USE and "z" represent 3bp Free Energy. Variables *K1*, *K2* and *K3* are used to offset the exponential fit. β0, β1, β2, β3 represents regression coefficients of the nonlinear model parameters. Since the B subtype is over-represented in the dataset, the parameters are fitted based on the across subtype dataset.

Mouzaki et al's equation:

$$y = \beta 0 + \beta 1 * e^{-(K1*w)} \qquad \textit{[Equation 2]}$$

Our newly developed equation:

$$y = \beta 0 + \beta 1 * e^{-(K2*x)} + \beta 2 * e^{-(K3*z)} \qquad \textit{[Equation 3]}$$

Evaluation of the data fitting is performed through nonparametric bootstrapping. The confidence interval for the MSE is calculated. The confidence interval is obtained through a bootstrap bias-corrected and accelerated (BCa) bootstrap (Efron 1987).

Analyses of all sequences are stratified across geographic region, risk factor, drug resistance, sampling year, and days after infection. Geographic region stratification examines if certain region possess particular clustering of frameshift element. Risk factor analysis compares the frameshift element among different transmission routes: men having sex with men (MSM), heterosexual, IV drug users, and mother to baby. Changes of frameshift efficiency across sampling year and infection time are examined. Drug resistance (DR) mutations for HIV sequences of B subtype are determined using HIV Drug Resistance Database (HIVdb) (Liu and Shafer 2006). The genotypic resistance HIVDB algorithm version number 6.3.1 is applied on protease, reverse transcriptase and integrase genomic regions. Sequences are excluded if they do not overlap the full protein coding gene (protease, reverse transcriptase or integrase). All statistical comparisons are performed using a two-sided Wilcoxon Test.

HIV Fitness and Frameshift Efficiency

Frameshift efficiency is closely related to replication capacity (Dulude et al. 2006). Since HIV fitness is often dependent on HIV's replication capacity (Prado et al. 2005, De Luca 2006), we hypothesize potential relationship between frameshift efficiency and HIV relative fitness. Bell et al. have performed a pairwise competition assay between sequences (Ball et al. 2003) ,estimating the virus's relative fitness based on HIV-1 isolate production (Ball et al. 2003) and Heteroduplex tracking analysis (HTA) to determine the final ratio of the two viruses (Quinones-Mateu et al. 2000). For sequences used in the competition assay, we have filtered out sequences that do not overlap the gag-pol frameshift element. Five B subtype sequence and four C subtype sequences are assessed. The final dataset consist of ten B intrasubtype competitions, six C intrasubtype competitions, and twenty B/C intersubtype competitions. Sequences of higher frameshift efficiency are predicted to be the fitter virus. For cases of tied frameshift efficiency, sequence

with CXCR4 co-receptor is assumed to be fitter than sequences with CCR5 co-receptor. The prediction based on frameshift efficiency is then compared to the HTA competition experiment result. A binomial function is used to estimate the probability for the prediction to be derived from chance.

Results

Ribosomal frameshift efficiency mechanism is hypothesized to be dependent on the stability of the RNA secondary structure. To test this, an exponential decay function is used to model frameshift efficiency data using 3bp stem-loop free energy and unpaired structural entropy (USE) measure. Spearman rank correlation coefficient shows both USE and 3bp free energy correlate to frameshift efficiency, with USE having a strong correlation coefficient than 3bp free energy, -0.7579324 (p-value = 9.567e-09) and -0.4208237 (p-value = 0.006148) respectively. Through a nonlinear least square algorithm, a negative exponential decay function is fitted to 3bp free energy (Figure 6.1a) and USE (Figure 6.1b) with USE method having a better fit across subtypes (Table 6.1). To take advantage of both RNA structure features, an exponential decay predictive model using both features is constructed, and the BCa 95% estimates the confidence interval MSE for the model to be 0.1149 to 0.9293 across subtype. Although the training dataset have a frameshift efficiency ranging between 3-10%, the predicted frameshift efficiency from the database ranged between 3%-8%.

Table 6.1. Shows the mean square error based on the fitted model.

Datasets	3bp Free Energy	USE Function	Combined Function
B Subtype n = 12 (Telenti et al. 2002)	2.814833	1.642365	**0.796689**
B Subtype n = 13 (Mouzakis et al. 2013)	**2.004857**	12.24248	2.450697
B Subtype n = 6 (Bidou et al. 1997)	0.8469377	1.063281	**0.7219714**
Across Subtype n = 10 (Baril et al. 2003)	0.4567413	**0.1304088**	**0.2702647**

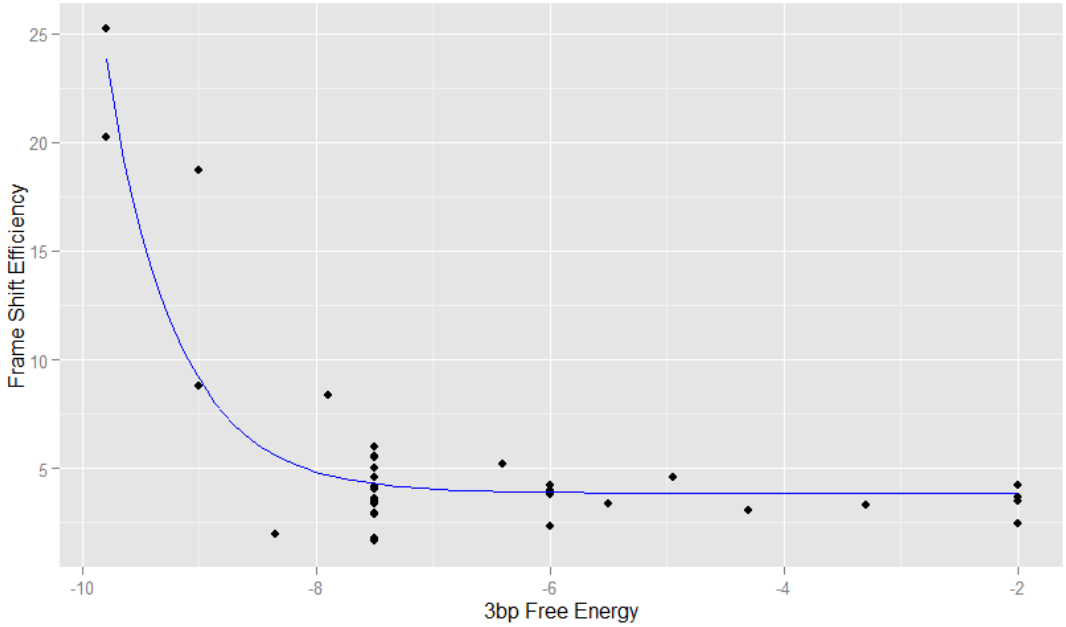


Figure 6.1A. Nonlinear model fitting 3bp free energy and frameshift efficiency.

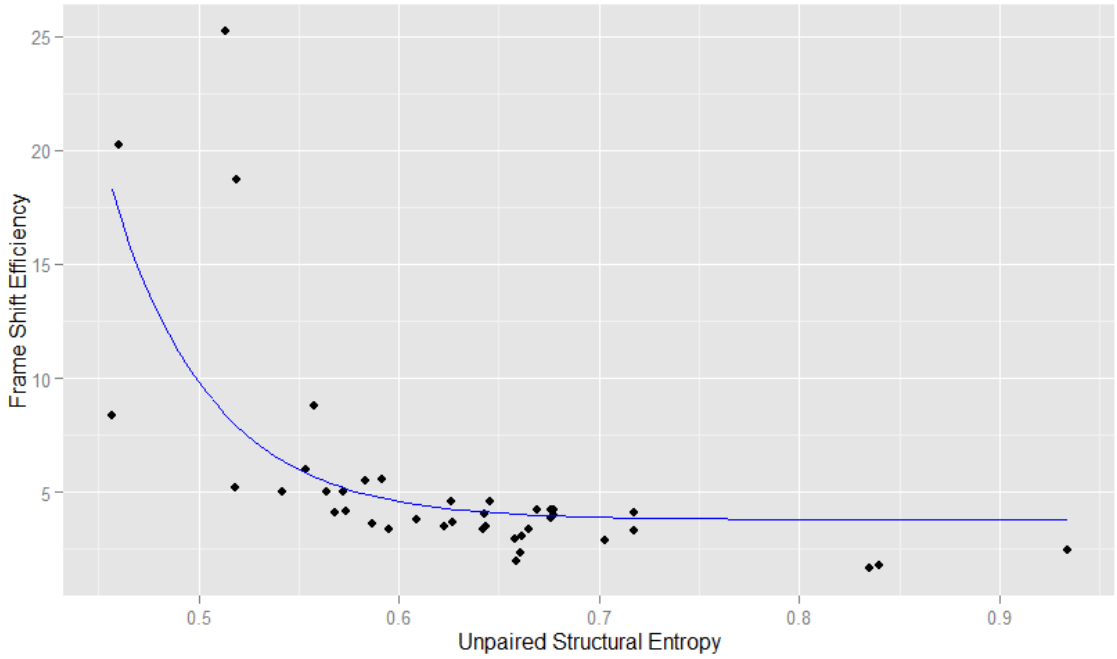


Figure 6.1B. Nonlinear model fitting USE and frameshift efficiency.

Figure 6.1. Evaluation of correlation between RNA measures and frameshift efficiency.

(A) The nonlinear model for 3bp free energy had an overall mean square error of 4.174165 frameshift efficiency. (B) The nonlinear model for unpaired structural entropy had an overall mean square error of 13.59481.

Examining HIV Frameshift Element Sequence Space and its Clustering

We have retrieved 459,881 sequences from open sequence sources (see Method). Infernal 2.0 using Rfam accession RF00480 as reference have identified frameshift element in 32,269 sequences from 6372 patients. Frameshift elements are grouped together based on 100% sequence identity. Infrequently used frameshift elements (< 10 sequences per node) are filtered out from analysis. Frameshift elements are labeled FE1 to FE41 (Figure 6.2a-b), and a maximum likelihood approach is used to generate its phylogenetic clustering. To assist in our

analysis, sequence space is divided into six clusters (C1-C6) (Figure 6.2a). For each cluster, a consensus sequence is generated using Simple Consensus Maker (Los-Alamos-HIV-Sequence-Database). The phylogenetic clustering of the frameshift elements are shown on the left and right panel of figure 6.2a-b and 6.3. On the left phylogenetic tree, branches at the tip are colored based on its 3bp free energy, and on the right phylogenetic tree branch tip are colored based on the 5% frameshift efficiency cutoff. A frameshift element network is shown in figure 6.2c. Multiple covariate base pair mutations are observed at positions 1, 11, 19, and 25 (Figure 6.2b). Three distinct features separate the clusters: (i) Changes of nucleotide base pairing near the base stem region (position 1 and 25) affects the free energy base pair stability of the mRNA ribosomal entry site; (ii) Mutations disrupting base pair close to the hairpin loop result in longer loop regions (position 11); and (iii) Covariate variations in the middle stem region destabilize the stem region (position 19). For sequences in Cluster C1 and some sequences in Cluster C2, the hairpin loop region is longer than sequences from other clusters. Based on the USE measure, longer hairpin loops region can increase base pairing variation for the entire RNA structure. Within the stem regions, position 19 contains covariate mutations between A and G purines for U-A and U-G base pairs. In general, higher free energy is required to break apart a U-A base pair compared to a U-G wobble base pair. Based on figure 6.3, clusters C3 and C4 are of the B subtype and cluster C5 is of the C subtype. Major differences between the B and C subtype are found at the base of the stem. The B subtype has a preference for G-C base pair and C subtype has a preference for G-U wobble base pair (Figure 6.2C).

Figure 6.3a shows the frameshift element's frequency for pure and recombinant subtypes. The distribution of recombinant subtypes is further stratified in figure 6.3b. Distinct subtype associated with clustering based on the free energy stability of 3bp. The B subtype sequences

are often associated with -7.5 and -6.5 ΔG. F1 and F2 subtypes are often associated with a less stable -4.95 ΔG. A base pairing energy of -6.0 ΔG is most frequently observed. An overlapping sequence space between two recombinant subtypes is present in 7 out of the 13 cases (Table 6.2). For the remaining 6 cases, the parental subtype's sequence space is always present in the recombinant frameshift element (Table 6.2). FE1 and FE37 are absent in the pure C subtype sequence space. FE13, FE29, and FE32 are absent in the pure B subtype sequence space. Out of the 41 frameshift elements, FE1 and FE2 have the highest frameshift efficiency and they are predominantly of the B subtype. FE12 is widely covered by the sequence space for pure and recombinant subtypes of CRF01, A1, A2, B, C, D, F1, and G subtypes.

Figure 6.2. Frameshift Element Clusters and Mutation Patterns.

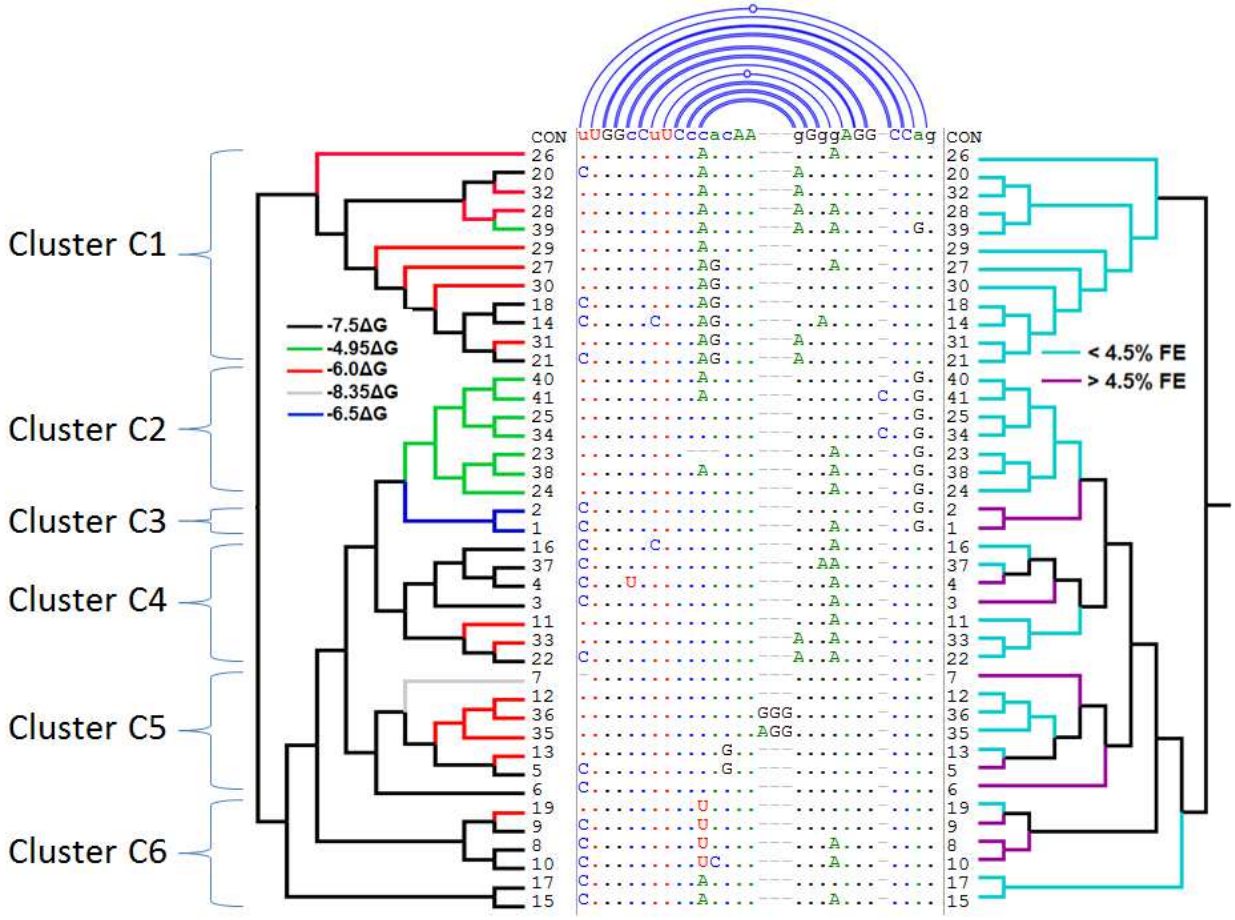


Figure 6.2A. Phylogenetic clustering of frameshift element and their predicted RNA structures.

Forty-one bins of frameshift element are organized based on the phylogenetic clustering. Sequences are separated to six clusters labeled C1-C6. Multiple sequence alignment of the 41 frameshift element is shown at the middle of the figure. The consensus sequence generated from the 41 frameshift element is shown at the top of the alignment. Left and right of the heatmap show the maximum likelihood phylogeny for the top 41 occurring frameshift element within the database, the phylogeny is created based on the frameshift element sequence. Each frameshift element is labeled FE1-FE41. The phylogeny stem coloring on the left is based on the different 3bp function thermodynamic energies at the base stem region. The phylogeny stem coloring on the right is based on predicted frameshift efficiency (less than 5% frameshift efficiency or greater than 5% frameshift efficiency).

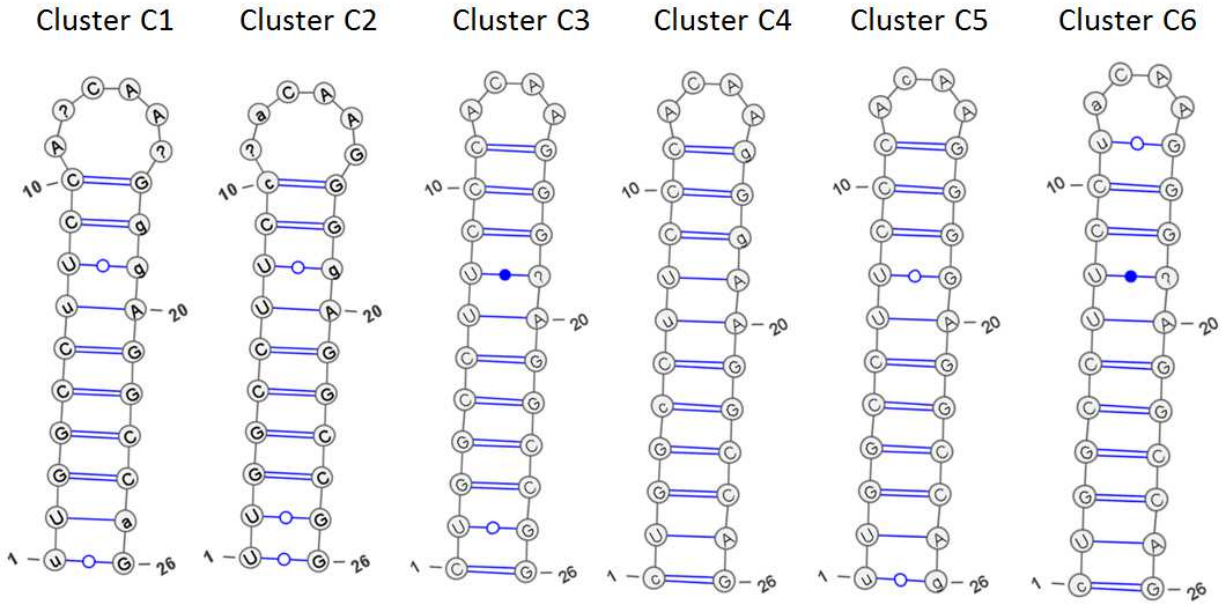


Figure 6.2B. Predicted RNA structure of the frameshift clusters.

Lower case nucleotides indicate an imperfectly conserved nucleotide, and nucleotide listed as "?" indicate no dominant nucleotide. Loop region of the RNA hairpin is longer in Cluster C1 and C2 than loop regions in other clusters. Based on the USE measure longer RNA hairpin could potentially destabilize the RNA structure. The stem regions were fairly conserved, with some portion possessing covariate mutations between A and G purines for U-A and U-G base pairs. Cluster C3 and C4 primary consist of B subtype and Cluster C5 primary consists of C subtype. The major difference between B and C subtypes were found at the base of the stem with C subtype preference for G-U wobble base pair and B subtype preference for a G-C base pair.

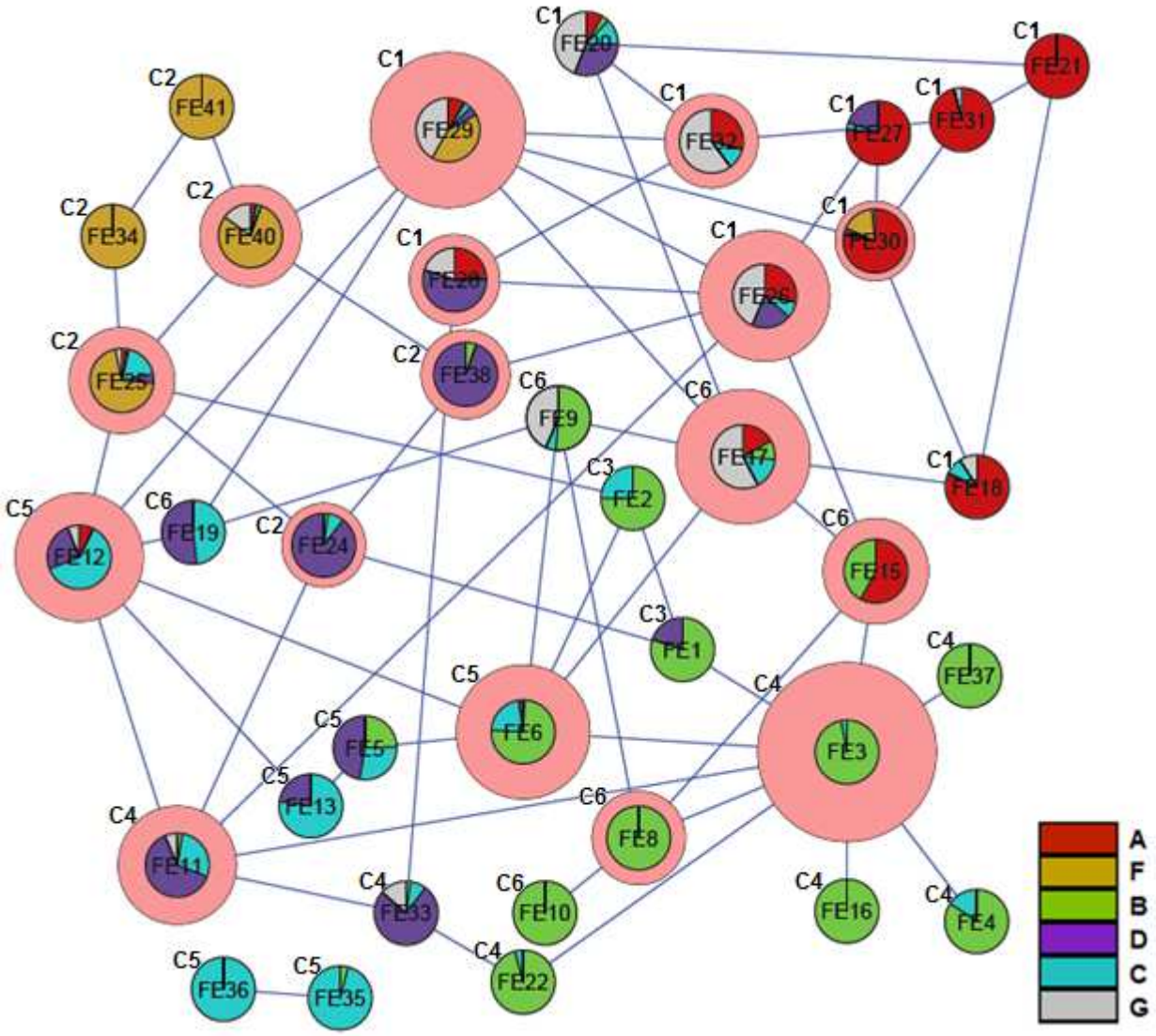


Figure 6.2C. Frameshift Element sequence space and mutational patterns of different HIV Subtypes.

Each node represents a frameshift element F1 through F41. Frameshift element node with one nucleotide mutation difference is connected by an edge. For each frameshift element node, a pie chart is constructed showing the normalized proportion of subtype occurring within each frameshift element. Distinct clustering of HIV subtype can be seen based on the network

structure. The nodes with high centrality measure are indicated by the pink circles. Each node's clustering is indicated at their top left corner (C1-C6).

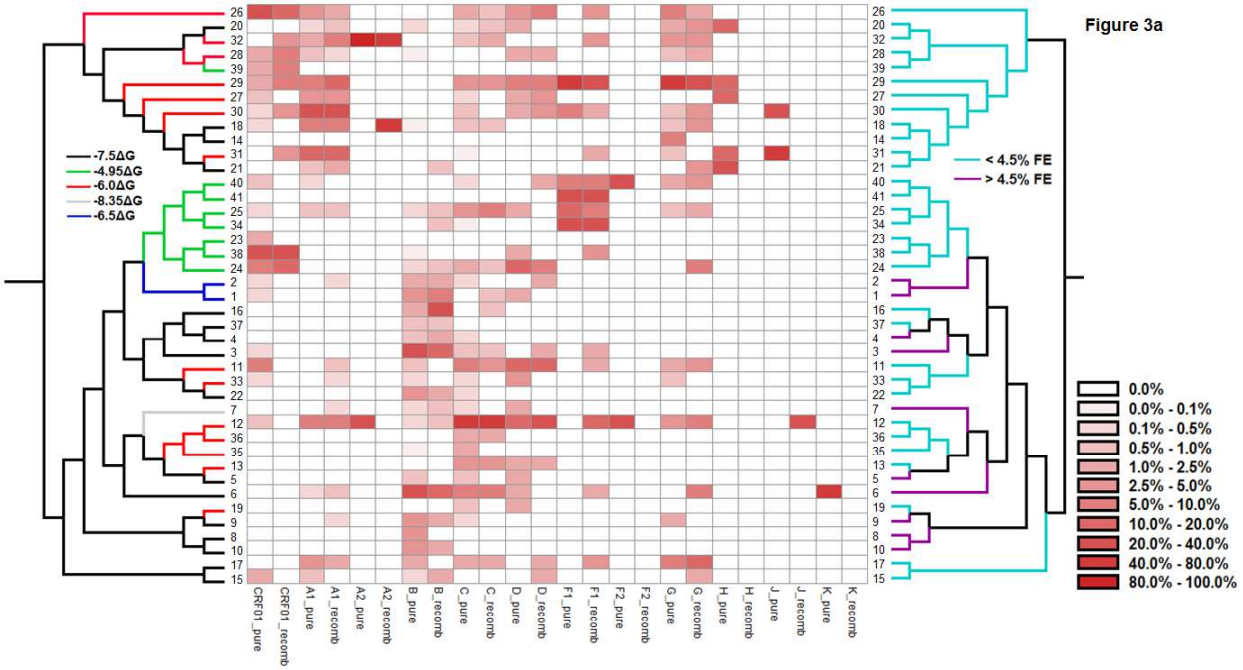


Figure 3a

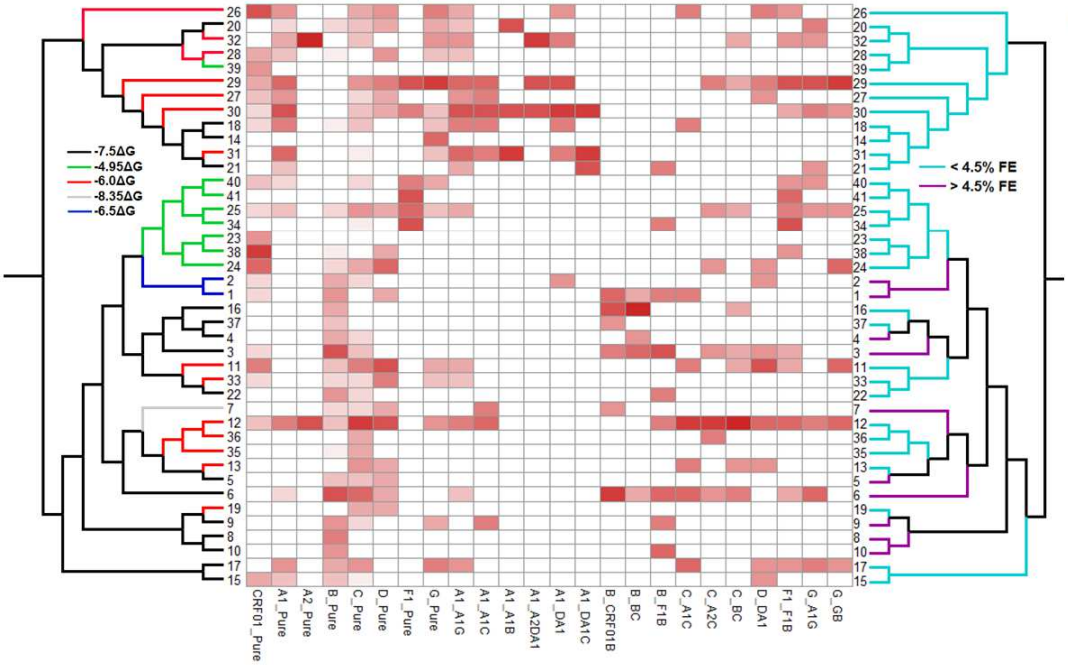


Figure 3b

Figure 6.3. Heatmap of Frameshift Element Distribution across Subtype and Recombinant.

Heatmap shows the relative frequency of frameshift element for pure and recombinant subtypes. Sequences with only one subtype are labeled with "Pure" in its naming. Figure 6.3b contains a more detailed breakdown of subtype occurring within recombinant. The naming convention used for labeling at the bottom of the heatmap is that the frameshift element fragment subtype is indicated left of the underscore and the right of the underscore indicate presence of other subtypes in the sequence.

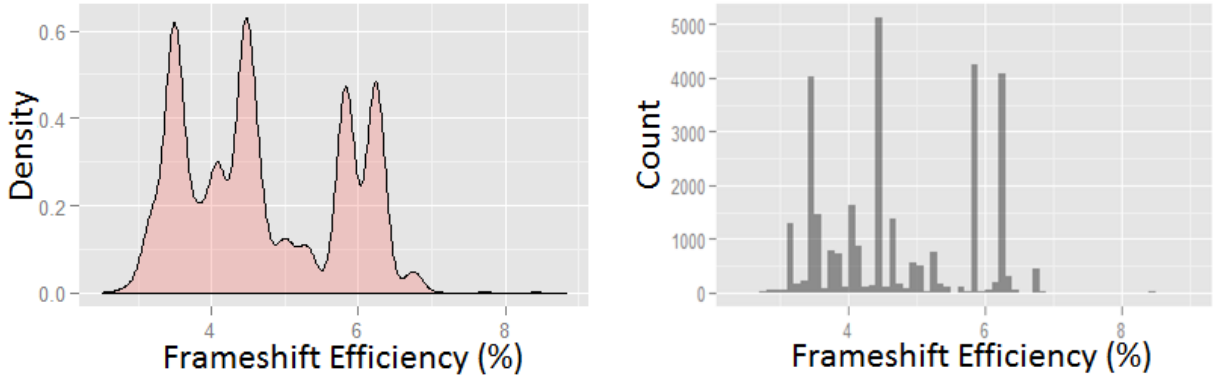


Figure 6.4. Frameshift Efficiency Histogram Density Plot.

Frameshift efficiency histogram and histogram plot shows the distribution of estimated frameshift efficiency within the HIV database. Three distinct peak can be observed by the cutoff of < 4%, 4-5%, and > 5%.

Table 6.2. Comparing Recombinant Frameshift Element Subtypes.

If both subtypes use the same frameshift element then it will appear as "Found in Both". We define parental subtype as the jpHMM annotated subtype that is covering the frameshift element. The majority of the frameshift elements were either from the distribution of parental subtype or

present in both parent and the other subtype. "Found in Neither" represents frameshift elements absent in both subtypes. "Found in Neither" could indicate change of novel mutations or incomplete sampling for a particular subtype.

Recombinant	Found in Both	Found Only in Parent	Found Only in Other	Found in Neither
A1G [A1]	**90.68%**	7.45%	1.86%	0.00%
A1C [A1]	**85.71%**	0.00%	14.29%	0.00%
A1B [A1]	20.00%	**80.00%**	0.00%	0.00%
DA1 [A1]	**79.49%**	17.95%	0.00%	2.56%
BC [B]	18.31%	**81.69%**	0.00%	0.00%
F1B [B]	0.00%	**86.67%**	6.67%	6.67%
A1C [C]	**80.00%**	13.33%	0.00%	6.67%
A2C [C]	**61.90%**	38.10%	0.00%	0.00%
BC [C]	**91.53%**	6.78%	1.69%	0.00%
DA1 [D]	38.10%	**42.86%**	9.52%	9.52%
F1B [F1]	5.80%	**65.22%**	23.19%	5.80%
A1G [G]	**80.95%**	0.00%	19.05%	0.00%
GB [G]	37.50%	**45.83%**	16.67%	0.00%

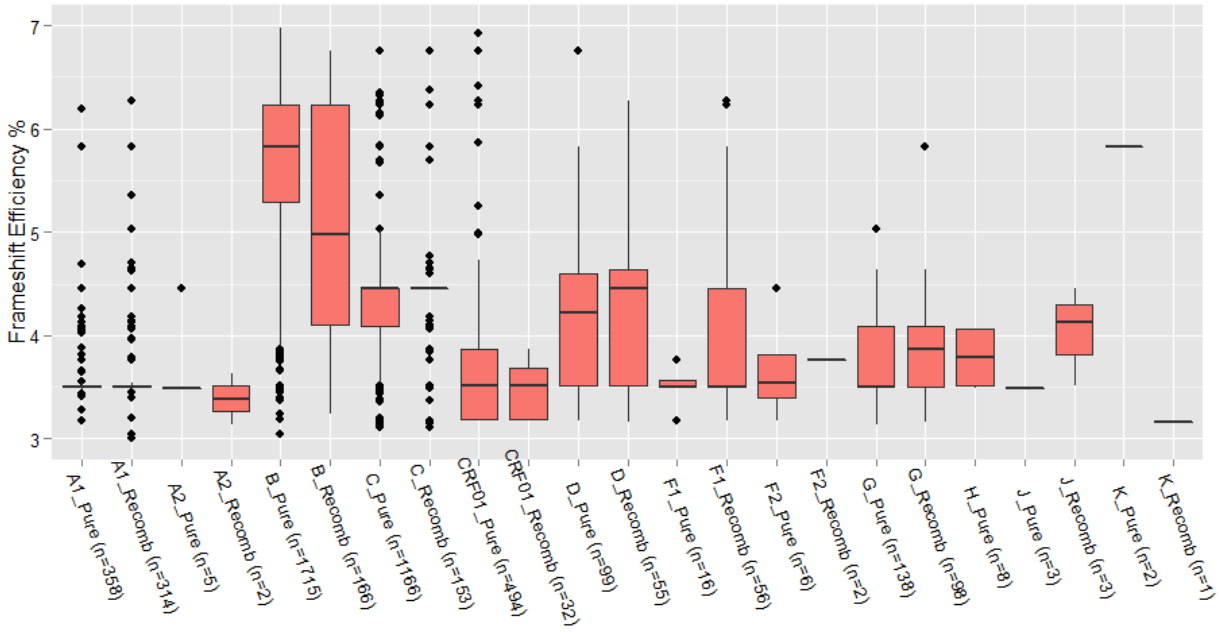


Figure 6.5. Frameshift Efficiency Comparison across Subtypes.

The boxplot compares the estimated relative frameshift efficiency across different subtypes. Notably the B subtype and B subtype recombinants possess the highest frameshift efficiency than other subtypes. After the B subtype, the C subtype was indicated to be the second most efficient group D subtype as the third most efficient. Most of the other subtypes range within 3.5-4.0% of relative frameshift efficiency. For the vast majority of cases, the frameshift efficiency for the recombinants was very similar to that of the pure subtypes.

Replication fitness and frameshift efficiency

Methods to reduce HIV viral fitness could provide clinical benefits. Since frameshift efficiency is closely related to replication capacity and replication fitness, we hypothesize possible relationship between replication fitness and frameshift efficiency. Previously, the fitness of B and C subtypes were calculated using the heteroduplex tracking assay (HTA) (Quinones-Mateu et al. 2000). HTA can quantify the production of HIV-1 isolate during dual virus competition

(Ball et al. 2003). Pairwise competition experiments can reveal the fitter viral variant. In our assessment, we have assumed sequences with higher frameshift efficiency are fitter variants than sequence of lower frameshift efficiency. Comparing prediction from frameshift efficiency's to the competition experiment's result, the intra-subtype HIV fitness via frameshift efficiency is able to correctly predict a total of 12 out of 16 competition assays, with a p-value of 0.011. The model can correctly predict 8 out of 10 competition assays for B subtypes (Table 6.3) and 4 out of 6 competition assays for C subtypes (Table 6.4). Pearson correlation R^2 of 0.2257 is found between the competition assay ratio and the frameshift efficiency ratio. For intersubtype comparisons, the model can only predict 11 out of 20 competition assays (Table 6.5), for a p-value of 0.252.

Table 6.3. B subtype intrasubtype dual infection competition assay result and prediction based on frameshift efficiency.

Virus Strand 'a' (FE)	Virus Strand 'b' (FE)	Experimental Result	Prediction
B3 (6.234638)	B4 * (6.234638)	b is more fit	b is more fit
B3 (6.234638)	B5 (3.773273)	a is more fit	a is more fit
B3 (6.234638)	B7 (4.460998)	b is more fit	a is more fit
B3 (6.234638)	B9 (6.234638)	b is more fit	a is more fit
B4 * (6.234638)	B5 (3.773273)	a is more fit	a is more fit
B4 * (6.234638)	B7 (4.460998)	a is more fit	a is more fit
B4 * (6.234638)	B9 (6.234638)	a is more fit	a is more fit
B5 (3.773273)	B7 (4.460998)	b is more fit	b is more fit
B5 (3.773273)	B9 (6.234638)	b is more fit	b is more fit

B7 (4.460998)	B9 (6.234638)	b is more fit	b is more fit

Table 6.4. C subtype intrasubtype dual infection competition assay result and prediction based on frameshift efficiency.

Virus Strand 'a' (FE)	Virus Strand 'b' (FE)	Experimental Result	Prediction
C2 (5.829682)	C3 (5.829682)	b is more fit	Same fitness
C2 (5.829682)	C5 (4.643323)	a is more fit	a is more fit
C2 (5.829682)	C6 (5.829682)	Same fitness	Same fitness
C3 (5.829682)	C5 (4.643323)	a is more fit	a is more fit
C3 (5.829682)	C6 (5.829682)	a is more fit	Same fitness
C5 (4.643323)	C6 (5.829682)	b is more fit	b is more fit

Table 6.4. B and C subtype intersubtype dual infection competition assay result and prediction based on frameshift efficiency.

Virus Strand 'a' (FE)	Virus Strand 'b' (FE)	Experimental Result	Prediction
B3 (6.234638)	C2 (5.829682)	a is more fit	a is more fit
B4 * (6.234638)	C2 (5.829682)	a is more fit	a is more fit
B5 (3.773273)	C2 (5.829682)	a is more fit	b is more fit
B7 (4.460998)	C2 (5.829682)	a is more fit	b is more fit
B9 (6.234638)	C2 (5.829682)	a is more fit	a is more fit
B3 (6.234638)	C3 (5.829682)	b is more fit	a is more fit
B4 * (6.234638)	C3 (5.829682)	a is more fit	a is more fit

B5 (3.773273)	C3 (5.829682)	a is more fit	b is more fit
B7 (4.460998)	C3 (5.829682)	a is more fit	b is more fit
B9 (6.234638)	C3 (5.829682)	a is more fit	a is more fit
B3 (6.234638)	C5 (4.643323)	b is more fit	a is more fit
B4 * (6.234638)	C5 (4.643323)	a is more fit	a is more fit
B5 (3.773273)	C5 (4.643323)	a is more fit	b is more fit
B7 (4.460998)	C5 (4.643323)	a is more fit	b is more fit
B9 (6.234638)	C5 (4.643323)	a is more fit	a is more fit
B3 (6.234638)	C6 (5.829682)	a is more fit	a is more fit
B4 * (6.234638)	C6 (5.829682)	a is more fit	a is more fit
B5 (3.773273)	C6 (5.829682)	a is more fit	b is more fit
B7 (4.460998)	C6 (5.829682)	b is more fit	b is more fit
B9 (6.234638)	C6 (5.829682)	a is more fit	a is more fit

Stratified Comparison of Relative Frameshift Efficiency

The relative frameshift efficiency is compared across subtypes, geographic regions, risk factors, post infection days, sampling years, drug treatment mutations, and drug treatment statuses. Across all HIV subtypes, pure and recombinant B subtypes have the highest frameshift efficiency (Figure 6.5); recombinant B subtypes are lower than pure B subtypes and higher than non-B subtypes. Frameshift efficiency distribution for recombinant seems to be slightly different to its pure subtype category. The B subtype recombinants have the most diverse frameshift efficiency range compared to existing frameshift element subtypes. The B recombinant subtype's frameshift efficiency is also slightly lower than its pure subtype. For non-B subtypes,

although the Wilcoxon comparison indicates only the G pure/recombinant pair has a statistically different distribution of frameshift efficiency. The recombinant frameshift efficiency in the recombinant is often observed higher than its pure subtype pair, observed in: C, D, F1, and G subtypes. Over the course of a patient's infection from acute to chronic state, the B subtype's frameshift efficiency has been found changing over time, but no example of frameshift efficiency change has been found in C subtype (Figure 6.6).

Across geographical regions, we have observed that C subtype Asia sequences tend to have higher frameshift efficiency than other geographic regions supported by a Wilcoxon two-sided p-value of 2.454e-05 (Figure 6.7). Across risk factors, men having sex with men (MSM)'s frameshift efficiency is often different to heterosexual risk factors: MSM for B subtype in Asia is higher than heterosexual risk factor with a p-value of *0.01422* and MSM for C subtype in Sub-Saharan Africa is lower than heterosexual risk factor with a p-value of *0.003278* (Figure 6.8). In Figure 6.8, IV Drug users in Asia have frameshift efficiency higher than heterosexual risk factors, with a p-value of 0.00662. Across sampling year, we have failed to observe significant changes in frameshift efficiency (Figure 6.9).

The B subtype sequence's drug resistance (DR) mutations are determined for protease, integrase, and reverse transcriptase genes (Liu and Shafer 2006). For each drug resistant mutation annotated by HIVdb, frameshift efficiency for sequences with the drug resistant mutation and sequences without the drug resistant mutation are compared. Out of the three genes, the protease gene contains two DR mutation positions with higher frameshift efficiency (Figure 6.10): 46I (p-value of 0.0338) and 84V (p-value of 0.0497).

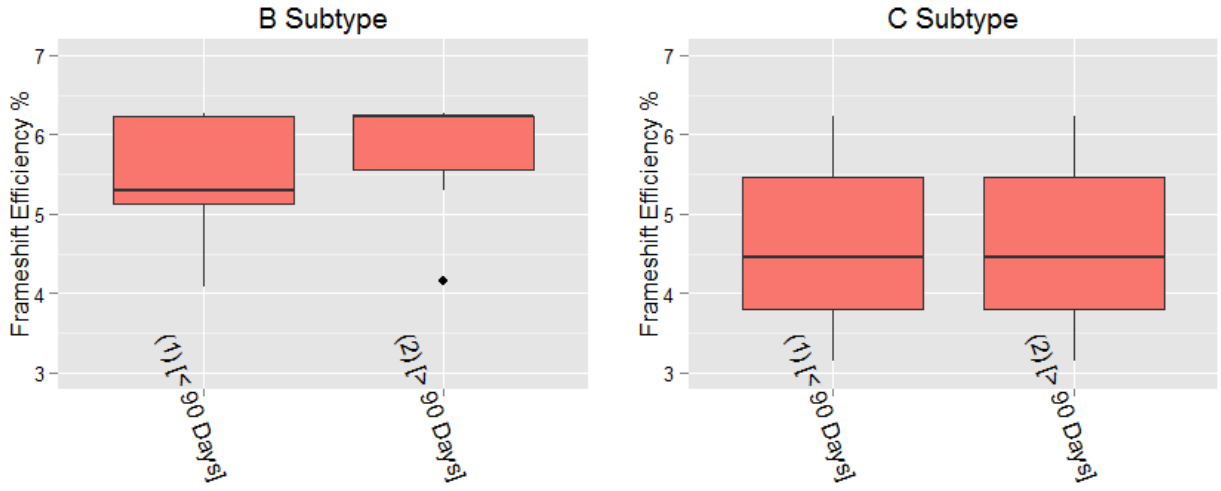


Figure 6.6. Changes in frameshift efficiency over infection time for B and C Subtype.

For each patient, we have compared the frameshift efficiency before 90 days post infection and after 90 days post infection. Four out of the seven B subtype patient's frameshift element has changed over time. The frameshift efficiency medium has gotten higher during chronic stages of the infection. Within C subtype, none of the patient's frameshift efficiency has changed over time.

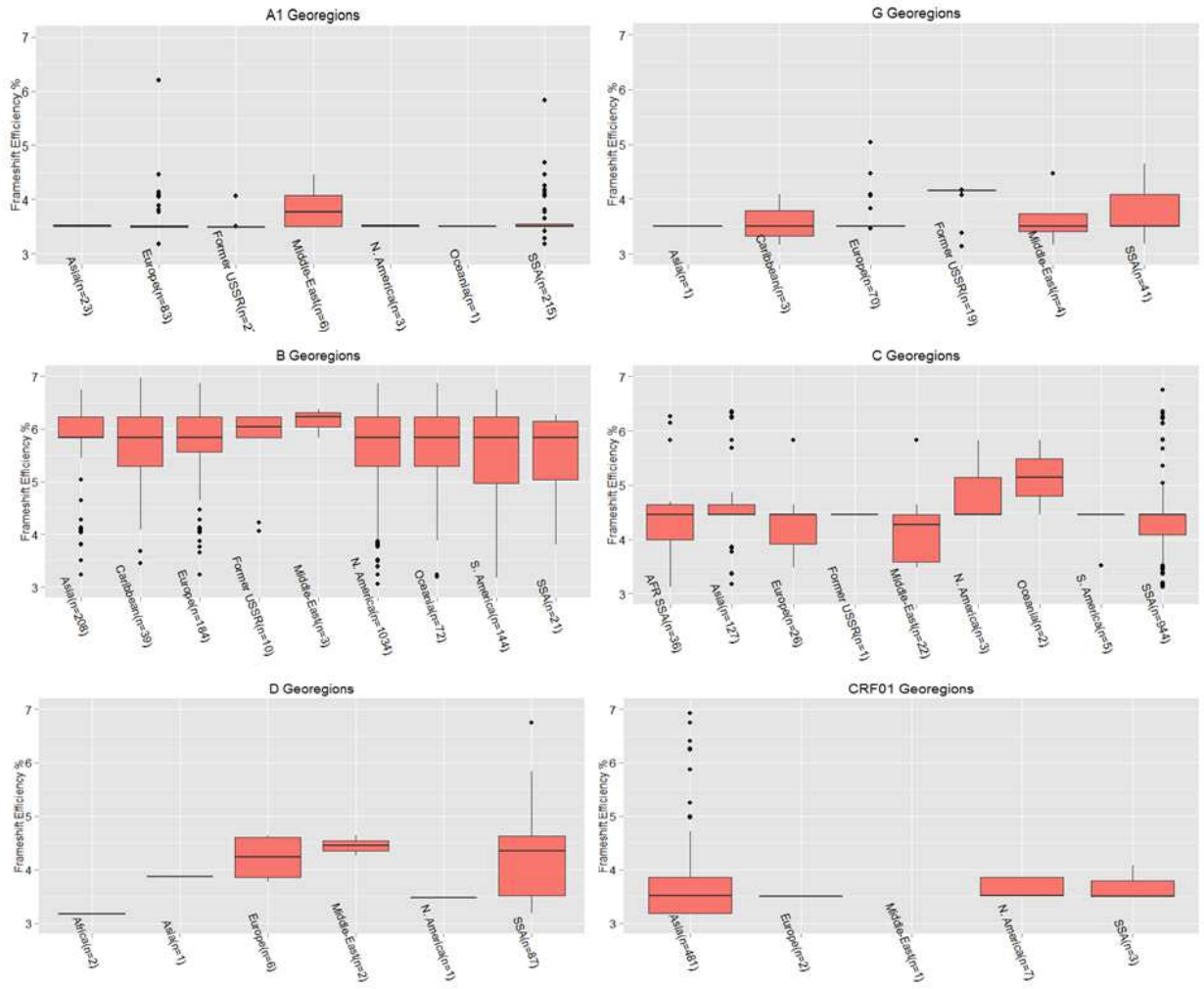


Figure 6.7. Frameshift efficiency across different geographic regions.

SSA is the acronym for Sub-Saharan Africa. For A1, B and C subtypes, the frameshift efficiency tends to be higher in Asia than other geo-regions. The comparison between C subtype Asia and non-Asia region resulted in a Wilcoxon two sided test p-value of 2.454e-05 *when comparing Asia with other Non-Asia region. Comparing Asia and non-Asia region for A1 and B subtype resulted in a p-value of 0.07512 and 0.6019.*

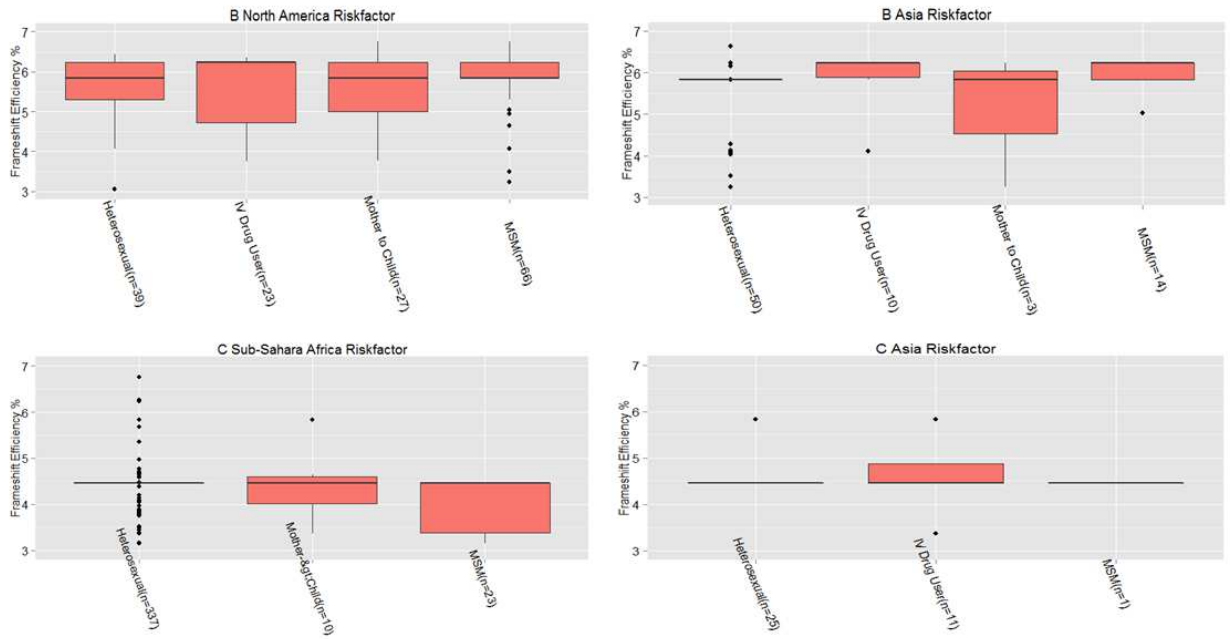


Figure 6.8. Stratification across risk factors.

B and C subtypes were chosen since more risk factor data points were present within these subtypes. Heterosexuals, MSMs, IV Drug Users, and Mother to Child risk factors were more present in database than other factors. In C Subtype Asia, IV Drug Users seemed higher than other risk factors but with a weak p-value support of 0.1008. In B subtype Asia, IV Drug Users and MSMs have higher frameshift efficiency than Heterosexual risk factors with p-value of 0.01422 and 0.00662 respectively. In North America IV Drug Users and MSM's frameshift efficiency appears slightly higher than other risk factors, but with weak statistical support: p-value of 0.9038 and 0.1874 respectively. In C subtype Sub-Saharan Africa, we found MSMs to have lower frameshift efficiency than other risk factors with a p-value of 0.003327.

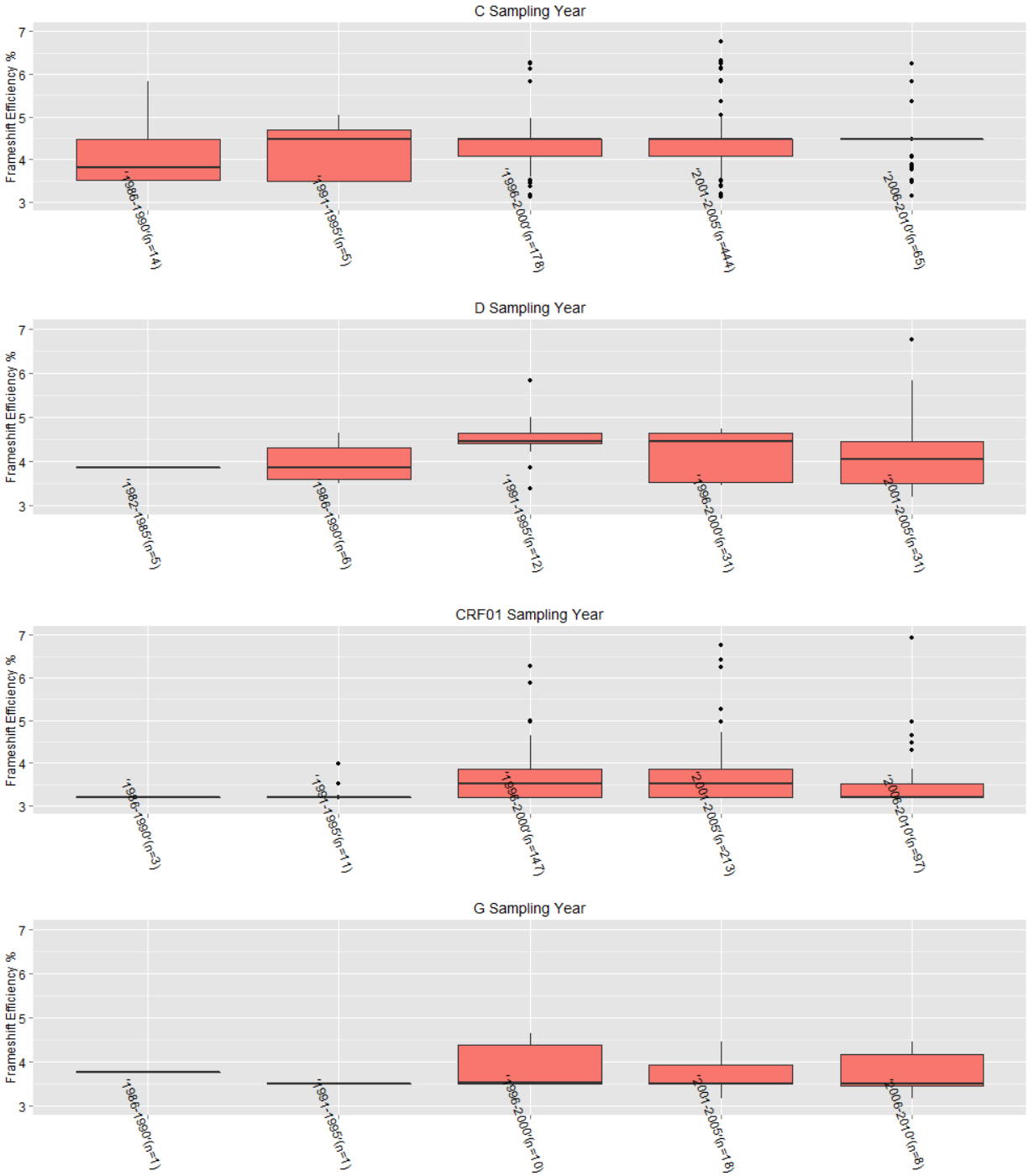


Figure 6.9. Across Sampling Year.

There were no noticeable difference across different time points. The only difference across sampling year is the increase in number of sequences sampled after 1996-2000.

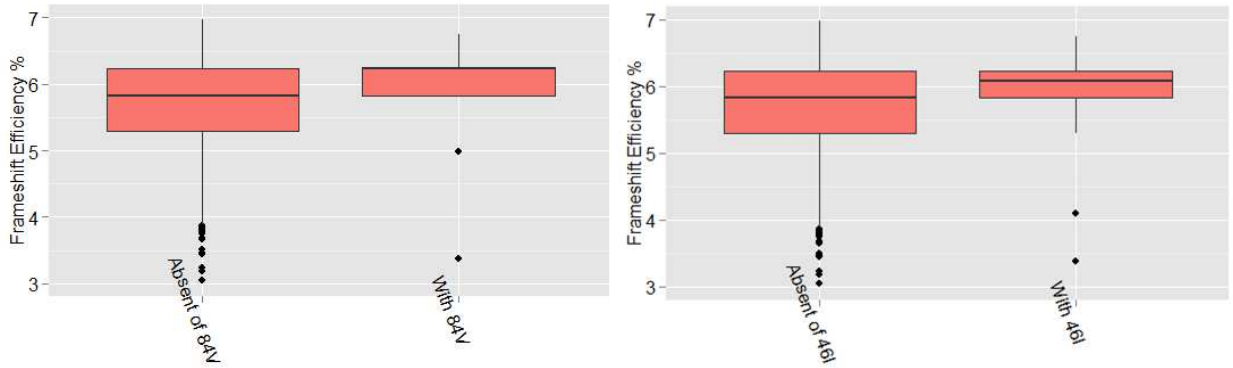


Figure 6.10. Drug resistant positions on protease gene.

Sequences containing two protease drug resistant mutation positions tend to have higher frameshift efficiency. For position 46I, the Wilcoxon two-sided test indicates a p-value of 0.0338. For position 84V, the Wilcoxon two-sided test indicates a p-value of p-value of 0.0497.

Discussion

The frameshift element genomic region contains mechanisms critical to regulating Gag and Pol protein which in turn are important regulators for HIV replication (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009). Given the important role of frameshifting in HIV replication (Brakier-Gingras, Charbonneau, and Butcher 2012, Gareiss and Miller 2009), the efficiency of the frameshift element will impact HIV fitness and its evolutionary diversity. As the frameshift element is varied across different subtype, mutation and recombination would select frameshift element that confers fitness gains. Understanding the genetic changes in the element in HIV subtypes can impact treatment and vaccine design. Here we analyzed the diversity of the frameshift element RNA secondary structure for each subtype, and a frameshift efficiency model was developed based on the RNA secondary structure's stability.

Consistent with previous studies, our result supports the hypothesis that -1 PRF is modulated by the stability of the frameshift element's upper stem region. Based on Spearman Rank correlation, both 3bp free energy measure (Mouzakis et al. 2013) and unpaired structural entropy (Shaw et al. 2011) are correlated with frameshift efficiency. In particular, USE has a stronger correlation to frameshift efficiency than 3bp free energy, and across subtypes, USE's interpolation of frameshift efficiency is also more accurate than 3bp free energy. A combined model featuring USE and 3bp free energy has further improved prediction accuracy for frameshift efficiency, indicating that the -1 PRF is dependent on both base pairing stability and its structural variation (Shaw et al. 2011).

The complete frameshift element sequence space is represented by FE1-FE41. Out of the 41 frameshift element, five frameshift elements are broadly used by pure and recombinant

subtypes, and are central to the frameshift element mutational network: FE6, FE12, FE25, FE29 and FE30 (Figure 6.2c). Out of these five frameshift elements, only FE6 (dominated by B, C and D subtypes) has greater than 5% frameshift efficiency; the result indicates that nodes with higher frameshift efficiency does not necessary translate to frequent recombination. The network indicates (Figure 6.2c) a dramatic alteration of frameshift efficiency can be achieved by a single nucleotide mutation, but certain mutations are often avoided in an individual subtype. One such example is FE25 occupied by the F subtype; although a single mutation on the 1st nucleotide position could change FE25 into FE1. FE1 is absent in F subtype's sequence space (Figure 6.3), and FE1 is a frameshift element with high frameshift efficiency (Figure 6.2c). The results suggest that each subtype's frameshift element's sequence space is restricted to an effective range of frameshift efficiency. For intra-subtype sequences, frameshift efficiency has been found capable of predicting outcome of fitness competition assays. The R^2 between the fitness ratio and frameshift efficiency ratio indicates a relative weak relationship; frameshift efficiency is not the only factor contributing to viral fitness. Virus genotypic variation might also influence viral fitness (van Opijnen and Berkhout 2005); for intersubtype cases, frameshift efficiency is not shown predictive for competition assay result.

The propensity for HIV-1 to recombine is an inherent feature for all lentiviruses (Chen, Powell, and Hu 2006). For each recombinant, their frameshift element's sequence space can often overlap between parent and donor subtypes' sequence space. Previous studies have indicated factors promoting recombination include sequence identity (Zhang and Temin 1994) and viral fitness (Burke 1997). The overlapping of sequence space supports the hypothesis that a shared sequence identity is a determining factor for recombination. Recombination patterns fail to indicate a selective advantage for frameshift element of high frameshift efficiency. Instead,

recombination patterns are limited to those subtypes that share the same effective frameshift efficiency range. However, exceptions to the above rule are observed for sequences recombining B or D subtypes. The recombinant frameshift element sequence space for B and D subtypes often covers only sequence of parental subtypes. We suspect since B and D subtypes are generally fitter than sequences of other subtype, the fitness cost of altering the frameshift element out-competes the fitness benefits associated with attaining a B and D subtype genomic region. We also note that out of all the subtypes, B subtype possesses the broadest frameshift element sequence space and the broadest range of effective frameshift elements.

The frameshift efficiency for B subtype recombinants is lower than for pure B subtypes. Two possible explanations could explain this difference: (i) after the B subtype recombine with a different subtype, the frameshift element then may mutate to optimize its Gag to Gag-Pol ratio potential; and (ii) the majority of the non-B subtype sequence has a lower frameshift efficiency; recombination events are more likely to occur for B subtype sequences with low frameshift efficiency element. Given the frequent observation of shared sequence space across subtypes and recombinant clades (figure 6.3), if explanation (i) is true, then more dissimilarities between pure and recombinant subtypes should be observed. Therefore, the latter explanation (ii) could better describe why B subtype recombinant tends to have lower frameshift efficiency subtypes. Although minor genetic variation can occur over time, these types of mutations are infrequently observed (Figure 6.6).

Across risk factors, IDU is more prevalent in Asia and North America. From our results, we have observed higher frameshift efficiency in Asian strains, and higher frameshift efficiency is also observed for IV drug users in both B and C subtypes. In addition, frameshift efficiency for B subtype HIV sequences from MSM tends to be higher than B subtype sequences from

heterosexuals. Higher multiplicity of infection (MOI) is found for IDU (Bar et al. 2010) and MSM risk factors (Li, Bar, et al. 2010). As higher frameshift efficiencies are found in certain populations of IDU and MSM, we suspect a link between higher MOI to higher frameshift efficiency in the HIV population. When multiple closely related virus strands infect the same cell, a higher frameshift efficiency could provide the HIV sequence a slight advantage over other viral strands. We hypothesize that higher frameshift efficiency is an artifact of fitness competition from multiple HIV strands.

The protease gene contains two drug resistance mutations, 46I and 84V with elevated frameshift efficiency. The 46I mutation reduces the drug susceptibility for antiretroviral drugs targeting protease: ATV, IDV, LPV, NFV, and RTV (Rhee et al. 2003). The 84V mutation reduces drug susceptibility for IDV (Belec et al. 2000), LPV(Santos et al. 2012), DRV (Sterrantino et al. 2012), SQV (Michelet et al. 2001), and TPV (Naeger and Struble 2007). Capel et al, finds that sequences with protease resistance mutation have significantly lower ex vivo replication capacity than naïve viruses (Capel et al. 2012). In addition to those two mutations, we do find higher frameshift efficiency in sequences with drug resistant mutation. Drug resistance mutation are often associated with reduced viral fitness (Harrigan, Bloor, and Larder 1998, Martinez-Picado et al. 1999, Nijhuis et al. 1999, Quiñones-Mateu 2001). We hypothesize that by increasing its frameshift efficiency can counteract the reduced viral fitness from the drug resistance mutation, and this hypothesis definitely deserve further attention in future experiments.

Conclusion

We have observed complex patterns of frameshift element preference across pure and recombinant subtypes. Frameshift elements that are shared between two subtypes have an easier time of recombining. Our model suggests that the length of the loop region of the frameshift

RNA hairpin can potentially influence the stability of the RNA structure with added influence on the capacity for the polymerase ability to induce frameshift. Here, we have performed a comprehensive analysis of the frameshift element of existing HIV sequences, and have found a potential association of frameshift efficiency to HIV's replication fitness. Frameshift efficiency model can potentially be applied as an indicator for viral fitness within a subtype; however, additional experiments on the relationship between frameshift efficiency and viral fitness are needed to reveal the exact amount of contribution that frameshift efficiency has on viral fitness.

Reference

Abraha, A., I. L. Nankya, R. Gibson, K. Demers, D. M. Tebit, E. Johnston, D. Katzenstein, A. Siddiqui, C. Herrera, L. Fischetti, R. J. Shattock, and E. J. Arts. 2009. "CCR5- and CXCR4-tropic subtype C human immunodeficiency virus type 1 isolates have a lower level of pathogenic fitness than other dominant group M subtypes: implications for the epidemic." *J Virol* 83 (11):5592-605. doi: JVI.02051-08 [pii]

10.1128/JVI.02051-08.

Ball, S. C., A. Abraha, K. R. Collins, A. J. Marozsan, H. Baird, M. E. Quinones-Mateu, A. Penn-Nicholson, M. Murray, N. Richard, M. Lobritz, P. A. Zimmerman, T. Kawamura, A. Blauvelt, and E. J. Arts. 2003. "Comparing the ex vivo fitness of CCR5-tropic human immunodeficiency virus type 1 isolates of subtypes B and C." *J Virol* 77 (2):1021-38.

Bar, K. J., H. Li, A. Chamberland, C. Tremblay, J. P. Routy, T. Grayson, C. Sun, S. Wang, G. H. Learn, C. J. Morgan, J. E. Schumacher, B. F. Haynes, B. F. Keele, B. H. Hahn, and G. M. Shaw. 2010. "Wide variation in the multiplicity of HIV-1 infection among injection drug users." *J Virol* 84 (12):6241-7. doi: 10.1128/JVI.00077-10.

Baril, M., D. Dulude, K. Gendron, G. Lemay, and L. Brakier-Gingras. 2003. "Efficiency of a programmed -1 ribosomal frameshift in the different subtypes of the human immunodeficiency virus type 1 group M." *RNA* 9 (10):1246-53.

Belec, L., C. Piketty, A. Si-Mohamed, C. Goujon, M. C. Hallouin, S. Cotigny, L. Weiss, and M. D. Kazatchkine. 2000. "High levels of drug-resistant human immunodeficiency virus variants in patients exhibiting increasing CD4+ T cell counts despite virologic failure of protease inhibitor-containing antiretroviral combination therapy." *J Infect Dis* 181 (5):1808-12. doi: 10.1086/315429.

Bidou, L., G. Stahl, B. Grima, H. Liu, M. Cassan, and J. P. Rousset. 1997. "In vivo HIV-1 frameshifting efficiency is directly related to the stability of the stem-loop stimulatory signal." *RNA* 3 (10):1153-8.

Brakier-Gingras, L., J. Charbonneau, and S. E. Butcher. 2012. "Targeting frameshifting in the human immunodeficiency virus." *Expert Opin Ther Targets* 16 (3):249-58. doi: 10.1517/14728222.2012.665879.

Brandes, Ulrik. 2001. "A Faster Algorithm for Betweenness Centrality." *Journal of Mathematical Sociology* 25:163-177.

Brierley, I., P. Digard, and S. C. Inglis. 1989. "Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot." *Cell* 57 (4):537-47.

Burke, D. S. 1997. "Recombination in HIV: an important viral evolutionary strategy." *Emerg Infect Dis* 3 (3):253-9. doi: 10.3201/eid0303.970301.

Capel, E., G. Martrus, M. Parera, B. Clotet, and M. A. Martinez. 2012. "Evolution of the human immunodeficiency virus type 1 protease: effects on viral replication capacity and protease robustness." *J Gen Virol* 93 (Pt 12):2625-34. doi: 10.1099/vir.0.045492-0.

Chamorro, M., N. Parkin, and H. E. Varmus. 1992. "An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA." *Proc Natl Acad Sci U S A* 89 (2):713-7.

Chen, J. L., A. L. Dishler, S. D. Kennedy, I. Yildirim, B. Liu, D. H. Turner, and M. J. Serra. 2012. "Testing the nearest neighbor model for canonical RNA base pairs: revision of GU parameters." *Biochemistry* 51 (16):3508-22. doi: 10.1021/bi3002709.

Chen, J., D. Powell, and W. S. Hu. 2006. "High frequency of genetic recombination is a common feature of primate lentivirus replication." *J Virol* 80 (19):9651-8. doi: 10.1128/JVI.00936-06.

Chen, Z., A. Luckay, D. L. Sodora, P. Telfer, P. Reed, A. Gettie, J. M. Kanu, R. F. Sadek, J. Yee, D. D. Ho, L. Zhang, and P. A. Marx. 1997. "Human immunodeficiency virus type 2 (HIV-2) seroprevalence and characterization of a distinct HIV-2 genetic subtype from the natural range of simian immunodeficiency virus-infected sooty mangabeys." *J Virol* 71 (5):3953-60.

Darty, K., A. Denise, and Y. Ponty. 2009. "VARNA: Interactive drawing and editing of the RNA secondary structure." *Bioinformatics* 25 (15):1974-5. doi: 10.1093/bioinformatics/btp250.

De Luca, A. 2006. "The impact of resistance on viral fitness and its clinical implications." In *Antiretroviral Resistance in Clinical Practice*, edited by A. M. Geretti. London.

Dinman, J. D., M. J. Ruiz-Echevarria, and S. W. Peltz. 1998. "Translating old drugs into new treatments: ribosomal frameshifting as a target for antiviral agents." *Trends Biotechnol* 16 (4):190-6.

Dirks, R. M., and N. A. Pierce. 2003. "A partition function algorithm for nucleic acid secondary structure including pseudoknots." *J Comput Chem* 24 (13):1664-77. doi: 10.1002/jcc.10296.

Dirks, R. M., and N. A. Pierce. 2004. "An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots." *J Comput Chem* 25 (10):1295-304. doi: 10.1002/jcc.20057.

Dulude, D., Y. A. Berchiche, K. Gendron, L. Brakier-Gingras, and N. Heveker. 2006. "Decreasing the frameshift efficiency translates into an equivalent reduction of the

replication of the human immunodeficiency virus type 1." *Virology* 345 (1):127-36. doi: 10.1016/j.virol.2005.08.048.

Efron, Bradley. 1987. "Better Bootstrap Confidence Intervals." *Journal of the American Statistical Association* 82 (397):171.

Farabaugh, P. J. 1996. "Programmed translational frameshifting." *Microbiol Rev* 60 (1):103-34.

Gareiss, P. C., and B. L. Miller. 2009. "Ribosomal frameshifting: an emerging drug target for HIV." *Curr Opin Investig Drugs* 10 (2):121-8.

Gaudin, C., M. H. Mazauric, M. Traikia, E. Guittet, S. Yoshizawa, and D. Fourmy. 2005. "Structure of the RNA signal essential for translational frameshifting in HIV-1." *J Mol Biol* 349 (5):1024-35. doi: 10.1016/j.jmb.2005.04.045.

Giedroc, D. P., C. A. Theimer, and P. L. Nixon. 2000. "Structure, stability and function of RNA pseudoknots involved in stimulating ribosomal frameshifting." *J Mol Biol* 298 (2):167-85. doi: 10.1006/jmbi.2000.3668.

Green, L., C. H. Kim, C. Bustamante, and I. Tinoco, Jr. 2008. "Characterization of the mechanical unfolding of RNA pseudoknots." *J Mol Biol* 375 (2):511-28. doi: 10.1016/j.jmb.2007.05.058.

Hansen, T. M., S. N. Reihani, L. B. Oddershede, and M. A. Sorensen. 2007. "Correlation between mechanical strength of messenger RNA pseudoknots and ribosomal frameshifting." *Proc Natl Acad Sci U S A* 104 (14):5830-5. doi: 10.1073/pnas.0608668104.

Harrigan, P. R., S. Bloor, and B. A. Larder. 1998. "Relative replicative fitness of zidovudine-resistant human immunodeficiency virus type 1 isolates in vitro." *J Virol* 72 (5):3773-8.

Hemelaar, J., E. Gouws, P. D. Ghys, and S. Osmanov. 2006. "Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004." *AIDS* 20 (16):W13-23. doi: 10.1097/01.aids.0000247564.73009.bc 00002030-200610240-00024 [pii].

Hung, M., P. Patel, S. Davis, and S. R. Green. 1998. "Importance of ribosomal frameshifting for human immunodeficiency virus type 1 particle assembly and replication." *J Virol* 72 (6):4819-24.

Jacks, T., H. D. Madhani, F. R. Masiarz, and H. E. Varmus. 1988. "Signals for ribosomal frameshifting in the Rous sarcoma virus gag-pol region." *Cell* 55 (3):447-58.

Jacks, T., M. D. Power, F. R. Masiarz, P. A. Luciw, P. J. Barr, and H. E. Varmus. 1988. "Characterization of ribosomal frameshifting in HIV-1 gag-pol expression." *Nature* 331 (6153):280-3. doi: 10.1038/331280a0.

Jacks, T., K. Townsley, H. E. Varmus, and J. Majors. 1987. "Two efficient ribosomal frameshifting events are required for synthesis of mouse mammary tumor virus gag-related polyproteins." *Proc Natl Acad Sci U S A* 84 (12):4298-302.

Jacks, T., and H. E. Varmus. 1985. "Expression of the Rous sarcoma virus pol gene by ribosomal frameshifting." *Science* 230 (4731):1237-42.

Kohl, M., S. Wiese, and B. Warscheid. 2011. "Cytoscape: software for visualization and analysis of biological networks." *Methods Mol Biol* 696:291-303. doi: 10.1007/978-1-60761-987-1_18.

Kollmus, H., A. Honigman, A. Panet, and H. Hauser. 1994. "The sequences of and distance between two cis-acting signals determine the efficiency of ribosomal frameshifting in human immunodeficiency virus type 1 and human T-cell leukemia virus type II in vivo." *J Virol* 68 (9):6087-91.

Li, H., K. J. Bar, S. Wang, J. M. Decker, Y. Chen, C. Sun, J. F. Salazar-Gonzalez, M. G. Salazar, G. H. Learn, C. J. Morgan, J. E. Schumacher, P. Hraber, E. E. Giorgi, T. Bhattacharya, B. T. Korber, A. S. Perelson, J. J. Eron, M. S. Cohen, C. B. Hicks, B. F. Haynes, M. Markowitz, B. F. Keele, B. H. Hahn, and G. M. Shaw. 2010. "High Multiplicity Infection by HIV-1 in Men Who Have Sex with Men." *PLoS Pathog* 6 (5):e1000890. doi: 10.1371/journal.ppat.1000890.

Liu, T. F., and R. W. Shafer. 2006. "Web resources for HIV type 1 genotypic-resistance test interpretation." *Clin Infect Dis* 42 (11):1608-18. doi: 10.1086/503914.

Los-Alamos-HIV-Sequence-Database. "www.hiv.lanl.gov." www.hiv.lanl.gov.

Marcheschi, R. J., D. W. Staple, and S. E. Butcher. 2007. "Programmed ribosomal frameshifting in SIV is induced by a highly structured RNA stem-loop." *J Mol Biol* 373 (3):652-63. doi: 10.1016/j.jmb.2007.08.033.

Martinez-Picado, J., A. V. Savara, L. Sutton, and R. T. D'Aquila. 1999. "Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1." *J Virol* 73 (5):3744-52.

Michelet, C., A. Ruffault, V. Sebille, C. Arvieux, P. Jaccard, F. Raffi, C. Bazin, J. M. Chapplain, J. P. Chauvin, E. Dohin, F. Cartier, and E. Bellissant. 2001. "Ritonavir-saquinavir dual protease inhibitor compared to ritonavir alone in human immunodeficiency virus-infected patients." *Antimicrob Agents Chemother* 45 (12):3393-402. doi: 10.1128/AAC.45.12.3393-3402.2001.

Mouzakis, K. D., A. L. Lang, K. A. Vander Meulen, P. D. Easterday, and S. E. Butcher. 2013. "HIV-1 frameshift efficiency is primarily determined by the stability of base pairs

positioned at the mRNA entrance channel of the ribosome." *Nucleic Acids Res* 41 (3):1901-13. doi: 10.1093/nar/gks1254.

Naeger, L. K., and K. A. Struble. 2007. "Food and Drug Administration analysis of tipranavir clinical resistance in HIV-1-infected treatment-experienced patients." *AIDS* 21 (2):179-85. doi: 10.1097/QAD.0b013e3280119213.

Nawrocki, E. P., D. L. Kolbe, and S. R. Eddy. 2009. "Infernal 1.0: inference of RNA alignments." *Bioinformatics* 25 (10):1335-7. doi: 10.1093/bioinformatics/btp157.

Ndung'u, T., and R. A. Weiss. 2012. "On HIV diversity." *AIDS* 26 (10):1255-60. doi: 10.1097/QAD.0b013e32835461b5.

Nijhuis, M., R. Schuurman, D. de Jong, J. Erickson, E. Gustchina, J. Albert, P. Schipper, S. Gulnik, and C. A. Boucher. 1999. "Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy." *AIDS* 13 (17):2349-59.

Nixon, P. L., and D. P. Giedroc. 2000. "Energetics of a strongly pH dependent RNA tertiary structure in a frameshifting pseudoknot." *J Mol Biol* 296 (2):659-71. doi: 10.1006/jmbi.1999.3464.

Nixon, P. L., A. Rangan, Y. G. Kim, A. Rich, D. W. Hoffman, M. Hennig, and D. P. Giedroc. 2002. "Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot." *J Mol Biol* 322 (3):621-33.

Osmanov, S., C. Pattou, N. Walker, B. Schwardlander, and J. Esparza. 2002. "Estimated global distribution and regional spread of HIV-1 genetic subtypes in the year 2000." *J Acquir Immune Defic Syndr* 29 (2):184-90.

Park, J., and C. D. Morrow. 1991. "Overexpression of the gag-pol precursor from human immunodeficiency virus type 1 proviral genomes results in efficient proteolytic processing in the absence of virion production." *J Virol* 65 (9):5111-7.

Parkin, N. T., M. Chamorro, and H. E. Varmus. 1992. "Human immunodeficiency virus type 1 gag-pol frameshifting is dependent on downstream mRNA secondary structure: demonstration by expression in vivo." *J Virol* 66 (8):5147-51.

Peeters, M., and P. M. Sharp. 2000. "Genetic diversity of HIV-1: the moving target." *AIDS* 14 Suppl 3:S129-40.

Peeters, M., C. Toure-Kane, and J. N. Nkengasong. 2003. "Genetic diversity of HIV in Africa: impact on diagnosis, treatment, vaccine development and trials." *AIDS* 17 (18):2547-60. doi: 10.1097/01.aids.0000096895.73209.89.

Plant, E. P., and J. D. Dinman. 2005. "Torsional restraint: a new twist on frameshifting pseudoknots." *Nucleic Acids Res* 33 (6):1825-33. doi: 10.1093/nar/gki329.

Prado, J. G., N. T. Parkin, B. Clotet, L. Ruiz, and J. Martinez-Picado. 2005. "HIV type 1 fitness evolution in antiretroviral-experienced patients with sustained CD4+ T cell counts but persistent virologic failure." *Clin Infect Dis* 41 (5):729-37. doi: 10.1086/432619.

Quiñones-Mateu, M. E., and E. J. Arts. 2001. "HIV-1 fitness: implications for drug resistance, disease progression, and global epidemic evolution." In, ed B. Foley In C. Kuiken, B. Hahn, P. Marx, F. McCutchan, J. Mellors, S. Wolinsky, and B. Korber. Los Alamos National Laboratory, Los Alamos, N.Mex: HIV Sequence Compendium, Theoretical Biology and Biophysics Group.

Quinones-Mateu, M. E., S. C. Ball, A. J. Marozsan, V. S. Torre, J. L. Albright, G. Vanham, G. van Der Groen, R. L. Colebunders, and E. J. Arts. 2000. "A dual infection/competition

assay shows a correlation between ex vivo human immunodeficiency virus type 1 fitness and disease progression." *J Virol* 74 (19):9222-33.

Reuter, J. S., and D. H. Mathews. 2010. "RNAstructure: software for RNA secondary structure prediction and analysis." *BMC Bioinformatics* 11:129. doi: 10.1186/1471-2105-11-129.

Rhee, S. Y., M. J. Gonzales, R. Kantor, B. J. Betts, J. Ravela, and R. W. Shafer. 2003. "Human immunodeficiency virus reverse transcriptase and protease sequence database." *Nucleic Acids Res* 31 (1):298-303.

Rodriguez, M. A., M. Ding, D. Ratner, Y. Chen, S. P. Tripathy, S. S. Kulkarni, R. Chatterjee, P. M. Tarwater, and P. Gupta. 2009. "High replication fitness and transmission efficiency of HIV-1 subtype C from India: Implications for subtype C predominance." *Virology* 385 (2):416-24. doi: 10.1016/j.virol.2008.12.025.

Santos, J. R., J. M. Llibre, A. Imaz, P. Domingo, J. A. Iribarren, A. Marino, C. Miralles, M. J. Galindo, A. Ornelas, S. Moreno, J. M. Schapiro, and B. Clotet. 2012. "Mutations in the protease gene associated with virological failure to lopinavir/ritonavir-containing regimens." *J Antimicrob Chemother* 67 (6):1462-9. doi: 10.1093/jac/dks080.

Schultz, A. K., M. Zhang, T. Leitner, C. Kuiken, B. Korber, B. Morgenstern, and M. Stanke. 2006. "A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes." *BMC Bioinformatics* 7:265. doi: 1471-2105-7-265 [pii] 10.1186/1471-2105-7-265.

Shaw, T. I., A. Manzour, Y. Wang, R. L. Malmberg, and L. Cai. 2011. "Analyzing modular RNA structure reveals low global structural entropy in microrna sequence." *J Bioinform Comput Biol* 9 (2):283-98. doi: S0219720011005495 [pii].

Shehu-Xhilaga, M., S. M. Crowe, and J. Mak. 2001. "Maintenance of the Gag/Gag-Pol ratio is important for human immunodeficiency virus type 1 RNA dimerization and viral infectivity." *J Virol* 75 (4):1834-41. doi: 10.1128/JVI.75.4.1834-1841.2001.

Staple, D. W., and S. E. Butcher. 2005. "Solution structure and thermodynamic investigation of the HIV-1 frameshift inducing element." *J Mol Biol* 349 (5):1011-23. doi: 10.1016/j.jmb.2005.03.038.

Sterrantino, G., M. Zaccarelli, G. Colao, F. Baldanti, S. Di Giambenedetto, T. Carli, F. Maggiolo, and M. Zazzi. 2012. "Genotypic resistance profiles associated with virological failure to darunavir-containing regimens: a cross-sectional analysis." *Infection* 40 (3):311-8. doi: 10.1007/s15010-011-0237-y.

Telenti, A., R. Martinez, M. Munoz, G. Bleiber, G. Greub, D. Sanglard, and S. Peters. 2002. "Analysis of natural variants of the human immunodeficiency virus type 1 gag-pol frameshift stem-loop structure." *J Virol* 76 (15):7868-73.

ten Dam, E. B., C. W. Pleij, and L. Bosch. 1990. "RNA pseudoknots: translational frameshifting and readthrough on viral RNAs." *Virus Genes* 4 (2):121-36.

van Opijnen, T., and B. Berkhout. 2005. "The host environment drives HIV-1 fitness." *Rev Med Virol* 15 (4):219-33. doi: 10.1002/rmv.472.

Xia, T., J. SantaLucia, Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and D. H. Turner. 1998. "Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs." *Biochemistry* 37 (42):14719-35. doi: 10.1021/bi9809425.

Yu, C. H., M. H. Noteborn, C. W. Pleij, and R. C. Olsthoorn. 2011. "Stem-loop structures can effectively substitute for an RNA pseudoknot in -1 ribosomal frameshifting." *Nucleic Acids Res* 39 (20):8952-9. doi: 10.1093/nar/gkr579.

Zhang, J., and H. M. Temin. 1994. "Retrovirus recombination depends on the length of sequence identity and is not error prone." *J Virol* 68 (4):2409-14.

Zhang, M., B. Foley, A. K. Schultz, J. P. Macke, I. Bulla, M. Stanke, B. Morgenstern, B. Korber, and T. Leitner. 2010. "The role of recombination in the emergence of a complex and dynamic HIV epidemic." *Retrovirology* 7:25. doi: 10.1186/1742-4690-7-25.

Zhang, M., A. K. Schultz, C. Calef, C. Kuiken, T. Leitner, B. Korber, B. Morgenstern, and M. Stanke. 2006. "jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1." *Nucleic Acids Res* 34 (Web Server issue):W463-5.

CHAPTER 7

CONCLUSION

Key Findings

RNA molecules play a critical role in cellular processes across all domains of life including HIV. Existing RNA structure research has mainly focused on noncoding RNA structure prediction, and RNA structure prediction accuracy in HIV has not yet been assessed. We have assessed RNA structure prediction accuracy, and SHAPE technology is found to improve the accuracy of existing algorithms. In addition, we find that non-thermodynamic grammar based models performed better than thermodynamic based models. Based on our assessment, we have constructed a structure prediction pipeline that utilizes a combination of thermodynamic and grammar based approaches. We have identified RNA structure differences between non-B and NLM4-3 subtypes, indicating that the NLM4-3 SHAPE reactivity should be sparingly applied to non-B subtype sequences. Our investigation comparing CRF07-08 recombinant subtypes have revealed that RNA structure in the genomic region with splice donor/acceptor and NEF/LTR tends to be highly variable. The availability of our tool will further facilitate RNA structural research in the context of HIV evolution, diversity, and viral fitness.

Frameshift element is capable of regulating the ratio of gag and gag-pol protein products, and the production of gag-pol protein could determine HIV's replication fitness. In our study, we have examined subtype specific variation in frameshift element, and mutations are identified to alter RNA structure's fold stability. A frameshift efficiency nonlinear model is constructed using thermodynamic free energy and the USE measures. The predicted frameshift efficiency is correlated to HIV fitness ratio. By using frameshift efficiency, we are able to correctly predict the fitness comparison between intra-subtype sequences. We have estimated that frameshift

efficiency tends to be higher for B and D subtypes. B and D subtypes are often associated with higher fitness and faster disease progression. Examining the sequence space for each subtype reveals that two recombinant subtypes often have overlapping sequence space. Across subtype study of HIV frameshift diversity will enable future development of oligonucleotide specific drug target as well as therapeutic vaccine designs. Although other environmental variables could also influence the virus's fitness, our analysis indicates that frameshift efficiency is a main driver for its replication fitness.

Aside from HIV RNA structure research, tools are developed to measure HIV diversity. One such tool is the HIV-1 N-linked Glycosylation Site Analyzer. The HIV-1 N-linked Glycosylation Site Analyzer is a webserver with two main functions: (1) Performing mapping and comparison of N-linked Glycosylation sites between populations of HIV-1 sequences; (2) Alignment of variable loop by using N-linked glycosylation sites as anchors. The tool also enables the tracking of N-linked glycosylation site patterns across HIV-1 populations, expanding an understanding of viral diversity under the changing context of antigenic structure and transmission mechanisms.

Accurate subtype annotation of HIV sequences is integral for proper HIV diversity estimation. The jpHMM algorithm is one of the few HIV genotyping programs capable of identifying recombination break points. However, jpHMM executes in approximately 15 minutes per HIV sequence, and a parallelization method is able to annotate the entire HIV database (460,000 sequences) in less than 3 weeks. To verify the jpHMM result, we have constructed an automated HIV phylogenetic analyzer. The algorithm extracts the break point result from jpHMM and constructs a neighbor joining tree with 100 bootstrapping. The algorithm then examines the phylogenetic distance to each reference subtype. The

comprehensive assessment of HIV database finds five percent of the sequences to be incorrectly genotyped. The genotyping error was more prevalent between 1996 through 2002 in Central West Africa. Our results and genotyping pipeline will enable researchers to more efficiently obtain an accurate HIV genotyping definition.

The ability to capture structural features or RNA measurement is a crucial component for the classification and understanding of different RNA structures. We have developed a novel tool called unpaired structural entropy (USE) that is capable of quantifying the structural variability of a sequence. Due to the method's capability to distinguish microRNAs from other ncRNA families, we have applied our function to predict microRNA in the Jamaican Fruit Bat (Shaw et al. 2012). In our dissertation, we have incorporated SHAPE reactivity information into an existing CYK stochastic context free grammar developed by Yingfeng Wang from the RNA Informatics Lab (Wang et al. 2012). The tool is based on the BJK grammar (Knudsen and Hein 1999). Based on our analysis, the SHAPE corrected algorithm outperforms existing models in HIV RNA structures (hivtools.publichealth.uga.edu/SHAPE_PastaFold/PastaFold.php). The availability of our tool enables investigation of HIV RNA structure's involvement in evolution, diversity, and viral fitness. Our model will become more accurate as additional auxiliary chemical probing information is made available. Potential phylogenetic enhancement can also be applied to our model.

Other Future Directions

1. With the discovery of genotyping errors, a comprehensive database with correct subtyping information will benefit the HIV research community.

2. RNA structure stability for the splice donor can control HIV-1 splicing (Mueller, Berkhout, and Das 2013). We hypothesize the existence of an RNA stability range for the splicing event to occur effectively. (Baird et al. 2006)
3. As part of the N-linked glycosylation webserver, since the V3 region can potentially be used to predict co-receptors usage, a tool that can predict co-receptor usage should be added to our webserver tools.

References

Baird, H. A., Y. Gao, R. Galetto, M. Lalonde, R. M. Anthony, V. Giacomoni, M. Abreha, J. J. Destefano, M. Negroni, and E. J. Arts. 2006. "Influence of sequence identity and unique breakpoints on the frequency of intersubtype HIV-1 recombination." *Retrovirology* 3:91. doi: 10.1186/1742-4690-3-91.

Knudsen, B., and J. Hein. 1999. "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history." *Bioinformatics* 15 (6):446-54.

Mueller, Nancy, Ben Berkhout, and Atze Das. 2013. "The RNA structure of the major splice donor site controls HIV-1 splicing." *Retrovirology* 10 (Suppl 1):P62.

Shaw, T. I., A. Srivastava, W. C. Chou, L. Liu, A. Hawkinson, T. C. Glenn, R. Adams, and T. Schountz. 2012. "Transcriptome sequencing and annotation for the Jamaican fruit bat (Artibeus jamaicensis)." *PLoS One* 7 (11):e48472. doi: 10.1371/journal.pone.0048472.

Wang, Y., A. Manzour, P. Shareghi, T. I. Shaw, Y. W. Li, R. L. Malmberg, and L. Cai. 2012. "Stable stem enabled Shannon entropies distinguish non-coding RNAs from random backgrounds." *BMC Bioinformatics* 13 Suppl 5:S1. doi: 10.1186/1471-2105-13-S5-S1.