Virus Evolution, 2016, 2(1): vew002

doi: 10.1093/ve/vew002
Research article

Coevolutionary Analysis Identifies Protein–Protein Interaction Sites between HIV-1 Reverse Transcriptase and Integrase

Madara Hetti Arachchilage[1] and Helen Piontkivska[1,2,*]

[1]Department of Biological Sciences, Kent State University, Kent, OH 44242, USA and [2]School of Biomedical Sciences, Kent State University, Kent, OH 44242, USA

*Corresponding author: E-mail: opiontki@kent.edu

Abstract

The replication of human immunodeficiency virus-1 (HIV-1) requires reverse transcription of the viral RNA genome and integration of newly synthesized pro-viral DNA into the host genome. This is mediated by the viral proteins reverse transcriptase (RT) and integrase (IN). The formation and stabilization of the pre-integration complex (PIC), which is an essential step for reverse transcription, nuclear import, chromatin targeting, and subsequent integration, involves direct and indirect modes of interaction between RT and IN proteins. While epitope-based treatments targeting IN–viral DNA and IN–RT complexes appear to be a promising combination for an anti-HIV treatment, the mechanisms of IN-RT interactions within the PIC are not well understood due to the transient nature of the protein complex and the intrinsic flexibility of its components. Here, we identify potentially interacting regions between the IN and RT proteins within the PIC through the coevolutionary analysis of amino acid sequences of the two proteins. Our results show that specific regions in the two proteins have strong coevolutionary signatures, suggesting that these regions either experience direct and prolonged interactions between them that require high affinity and/or specificity or that the regions are involved in interactions mediated by dynamic conformational changes and, hence, may involve both direct and indirect interactions. Other regions were found to exhibit weak, but positive correlations, implying interactions that are likely transient and/or have low affinity. We identified a series of specific regions of potential interactions between the IN and RT proteins (e.g., specific peptide regions within the C-terminal domain of IN were identified as potentially interacting with the Connection domain of RT). Coevolutionary analysis can serve as an important step in predicting potential interactions, thus informing experimental studies. These studies can be integrated with structural data to gain a better understanding of the mechanisms of HIV protein interactions.

Key words: HIV-1 integrase; HIV-1 reverse transcriptase; pre-integration complex; protein–protein interaction; molecular coevolution.

1. Introduction

Human immunodeficiency virus (HIV)/acquired immunodeficiency syndrome (AIDS) remains a major public health issue worldwide (World Health Organization 2014), with the number of people living with HIV as high as 35.3 million and 2.3 million new infections occurring annually across the world (UNAIDS 2013). However, the development of highly active antiretroviral therapy treatment since the introduction of combination antiretroviral therapy has substantially decreased the morbidity and mortality rate of HIV patients (Brady et al. 2010; Le Douce et al. 2012). There are more than twenty-eight approved drugs targeting six different proteins at different steps in the viral life cycle

(Engelman and Cherepanov 2012; Zhang and Crumpacker 2013), including those that target reverse transcriptase (RT), integrase (IN), and protease (Arts and Hazuda 2012).

Although the long-term suppression of HIV-1 replication in many patients is achieved using strict adherence to highly active antiretroviral therapy (Koletar et al. 2004; Peters and Conway 2011), the high mutation rate of the HIV-1 virus often leads to the emergence of drug resistance, which is one of the major setbacks that prevents the successful anti-HIV drug/vaccine therapy in all HIV patients (Mehellou and De Clercq 2010; Peters and Conway 2011). Even though combination antiretroviral therapy suppresses viral replication and prevents the emergence of drug resistance better than antiretroviral monotherapies (Maenza and Flexner 1998; Arts and Hazuda 2012), drug-resistant strains continue to re-emerge (Luber 2005; Engelman and Cherepanov 2012). Thus, the need for development of more effective treatments remains a major challenge (Deeks et al. 2012), including development of an effective preventive vaccine (Lewin et al. 2011; International AIDS Vaccine Initiative 2012).

RT and IN proteins play central roles in the HIV-1 life cycle. The replication of HIV-1 upon successful penetration of the host cell requires the reverse transcription of viral RNA genome and the integration of newly synthesized pro-viral DNA into the host genome by RT and IN for reverse transcription of the RNA genome and integration, respectively (Chakraborty et al. 2013). HIV-1 RT converts single-stranded RNA into double-stranded DNA in the reverse transcription complex. After completion of reverse transcription, this reverse transcription complex will be associated into a pre-integration complex (PIC), where it comprised viral DNA, HIV-1 RT, IN, Vpr, matrix, and host cellular components. Afterward, this PIC is transported into the host nucleus where viral DNA is integrated into the host genome (Sarafianos et al. 2009). IN-mediated integration is a three-step process. In the three-end processing, HIV-1 IN removes the terminal GT dinucleotide from both ends of viral DNA, which occurs after reverse transcription in PIC in the cytoplasm. Afterward, when PIC is transported to the nucleus, HIV-1 IN catalyses strand transfer where viral DNA is integrated into the host genome. Finally, the disintegration step occurs where viral excision has taken place (Chiu and Davies 2004).

Studies have shown that in HIV-1 RT and IN, enzymes physically interact and inhibit each other, suggesting the existence of functional interactions between RT and IN proteins (e.g., Tasara et al. 2001; Oz, Avidan and Hizi 2002; Oz Gleenberg et al. 2005; Zawahir and Neamati 2006; Oz Gleenberg, Goldgur and Hizi 2007a; Oz Gleenberg et al. 2007b; Herschhorn, Oz-Gleenberg and Hizi 2008; Warren et al. 2009), including the stimulation of initiation mode of RT by full-length intact IN (Hehl et al. 2004). IN-RT protein complex is a key part of the PIC, which is involved in several steps of retrovirus replication, notably in reverse transcription, nuclear import, chromatin targeting and integration (Sarafianos et al. 2009; Levy et al. 2013; Ruff et al. 2014). The mechanism of protein–protein interactions of the IN–RT protein complex within PIC is not well understood. This is mainly due to the transient nature of this protein–protein complex, the dynamics of composition, and the intrinsic flexibility of its components such as IN (Nooren and Thornton 2003; Huang, Grant, and Richards 2011; Levy et al. 2013; Ruff et al. 2014). Structural analysis is further complicated because to date no full-length crystal structure has been published (Neamati and Wang 2011; Krishnan and Engelman 2012). Recently, the crystal structure of a full-length IN of prototype foamy virus (PFV) bound to viral DNA has been described. This PFV IN structure is arguably useful for the study of such protein–protein interactions, although the reliability of such analysis is unclear given the relatively low resolution of crystal structures and low sequence similarity between HIV-1 IN and PFV IN (Neamati and Wang 2011).

In this study, we therefore use a coevolutionary analysis as an alternative approach to identify potentially interacting regions. Because functional and physical interactions are typically reflected in coevolutionary signals in gene sequences (Clark, Alani and Aquadro 2012; de Juan, Pazos and Valencia 2013), evolutionary rate covariation is expected to be elevated between interacting proteins (de Juan, Pazos and Valencia 2013). Here, we use a modified evolutionary rate covariation method to identify potential interacting/coevolving regions ('interacting hot-spots') between RT and IN proteins. This is the first study to use overlapping epitope regions to narrow down the list of residues involved in interaction interface(s) between two proteins. Our results show that experimentally identified interacting regions between IN and RT closely correspond to computationally identified coevolving regions.

2. Methods

2.1 HIV-1 genomic sequences and sequence alignment

All available HIV-1 Pol protein coding DNA sequences were retrieved from the Los Alamos HIV database in July 2014 (4,407 sequences, one sequence per patient) (see Supplementary Table S1). The nucleotide sequence alignments of IN and RT sequences were aligned according to respective amino acid alignment and were extracted using the Gene Cutter program in Los Alamos HIV Database (Gene Cutter Tool - HIV LANL Database). Ambiguously aligned regions were removed with 90 per cent coverage cut-off (i.e., only sites shared by at least 90 per cent of sequences were included in the further study) using MEGA6 (Tamura et al. 2013).

2.2 IN-RT epitope clusters/non-epitope segments

The list of antibody (Ab), T-Helper, and best-defined cytotoxic T-lymphocytes (CTLs) epitopes was obtained for HIV-1 RT and IN regions from the HIV Immunology database (http://www.hiv.lanl.gov/content/immunology/tables/optimal_ctl_summary.html) (Supplementary Table S2). The best-defined HIV CTL/CD8 + epitopes in the HXB2 reference genome are identified on the basis of the classification as described here (Llano et al. 2013). The epitopes were mapped onto the RT and IN genes (Table 1). Because of the frequent observation of the tendency of HIV-specific CTL epitopes to cluster in immuno-dominant regions of proteins (Goulder et al. 1997), the epitopes were grouped. A cluster of epitopes that are overlapping (that includes Ab and/or T-Helper and/or best-defined CTL epitopes) is defined as one epitope cluster. The protein fragments between two epitope clusters are grouped as non-epitope segments. Because overlapping epitope peptides could promote the development of a stronger immune response than one that could be elicited using single isolated epitopes (Yusim et al. 2002), our approach allows us to focus on regions that can be expected to be important to the immune response. To avoid stochastic errors associated with a small fragment size, segments shorter than six amino acids were excluded from the analyses.

Table 1. Epitope clusters and non-epitope segments in HIV-1 RT and HIV-1 IN used in the study (residue coordinates are given per HXB2 amino acid coordinates).

HIV-1 IN				HIV-1 RT			
Epitope cluster/ non-epitope segment[a]	Start position	End position	Fragment size (in aa)	Epitope cluster/ non-epitope segment[a]	Start position	End position	Fragment size (in aa)
IN-EP1	1	8	8	RT-EP1	1	13	13
IN-NE1	9	15	7	RT-EP2	18	26	9
IN-EP2	16	43	28	RT-NE1	27	32	6
IN-NE2	44	65	22	RT-EP3	33	53	21
IN-EP3	66	93	28	RT-NE2	54	72	19
IN-EP4	96	121	26	RT-EP4	73	82	10
IN-EP5	123	132	10	RT-NE3	83	92	10
IN-EP6	135	143	9	RT-EP5	93	115	23
IN-NE3	144	164	21	RT-EP6	118	135	18
IN-EP7	165	234	70	RT-EP7	137	187	51
IN-NE4	235	241	7	RT-NE4	188	194	7
IN-EP8	242	271	30	RT-EP8	195	210	16
IN-NE5	272	288	17	RT-NE5	211	243	33
				RT-EP9	244	318	75
				RT-NE6	319	332	14
				RT-EP10	333	350	18
				RT-EP11	354	366	13
				RT-NE7	367	374	8
				RT-EP12	375	401	27
				RT-NE8	402	410	9
				RT-EP13	411	457	47
				RT-NE9	458	494	37
				RT-EP14	495	505	11
				RT-NE10	506	519	14
				RT-EP15	520	544	25
				RT-NE11	545	552	8
				RT-EP16	553	560	8

[a]A cluster of best-defined CTL, T-Helper, or Ab epitopes that are overlapping is defined as one epitope cluster

2.3 Random subset sampling and estimation of evolutionary rate of epitope clusters/non-epitope segments

For amino acid sequences corresponding to each of the epitope clusters (Eight IN and sixteen RT epitope clusters, respectively) and non-epitope segments (five IN and eleven RT non-epitope segments, respectively), the branch lengths were estimated on the tree topology, which was generated for the full-length Pol gene by using the maximum-likelihood method with the Dayhoff substitution model, taking into account rate heterogeneity and proportions of invariant sites (Dayhoff + G + I, five class parameters for gamma distribution). We estimated the underlying tree topology for each sample for full-length nucleotide sequences of the Pol gene by using the maximum-likelihood method with the Jukes–Cantor model, taking into account rate heterogeneity and properties of invariant sites (JC + G + I, five class parameters for gamma distribution). All analyses were performed with the Computing Core MEGA version six (MEGA-CC six) as automated and iterative data analysis (Kumar et al. 2012).

Rather than examining the entire dataset, a resampling technique allows the removal of sequence bias towards specific HIV-1 subtypes in the global sequence data set (Yonezawa et al. 2013). For example, the majority of the sequences available in the database belong to subtype B (~39%). Performing resampling is critical for reducing sampling bias when studying coevolutionary relationships between genomic regions. Therefore, the analysis was performed on 1,000 samples of 500 sequences, which were drawn from a complete set of 4,407 sequences using the simple random sampling method in R program (R Development Core Team) for all the partitioned epitope clusters and non-epitope segments. Although the treatment status was unavailable for many sequences (it was not annotated explicitly as either drug treated or drug naïve), we used the Surveillance Drug Resistance Mutation Worksheet 2014 at HIV Drug Resistance Database (http://hivdb.stanford.edu/pages/surveillance.html) to determine whether an individual sequence harboured drug resistance mutation. On average, about 10 per cent of sequences in each sample carried one or more resistance mutations (mean of fifty-five sequences, with the interquartile range from 50 to 59 sequences). This indicates that potential influence of resistance mutations can be expected to be approximately the same (if any) across all samples.

2.4 Estimation of evolutionary rate covariation

The resulting branch lengths were used to calculate the Pearson correlation coefficient (r) between all possible pairs of interacting epitope clusters and non-epitope segments to identify potential interacting regions. In other words, we expected branch lengths to approximate similarities in substitution patterns due to correlated changes (if any exist due to interactions)

(e.g., Li and Rodrigo 2009). The Pearson *r* values were calculated between each epitope region in IN against all epitope regions in RT for each sequence sample to determine the extent of such similarities. The combined correlation coefficient for each epitope pair across all samples was then calculated using the Pearson correlation coefficient scores, which were statistically significant at the 0.01 significance level. Because the correlation coefficients are not additive, combined correlation scores are computed on Fisher Z-transformed Pearson correlation (*r*) values (Garcia 2012). All statistical analyses were performed in R 3.2.0 and SAS 9.3 (SAS Institute, Cary, NC). The combined correlation coefficients were calculated only for peptide pairs that had significant correlation in at least 2/3rd of the total samples (i.e., 750 samples out of 1,000). It should be noted that alternative non-parametric approaches have been developed to detect coevolution between sites; for example, the approach based on mutual information, although its power depends on the extent of sequence conservation and structural constraints, among other limitations (e.g.,Fodor and Aldrich 2004; Patel, Garde and Stormo 2015). In this study, a parametric approach was chosen because of its greater sensitivity in detecting weak signals of coevolution (Codoner and Fares 2008).

2.5 Phylogenetically independent sister pairs

To remove the potential influence of a shared phylogenetic history (e.g., Sato et al. 2005) on detected interactions, we repeated the analysis using phylogenetically independent comparisons of pairs of sequences (Felsenstein 1985; Piontkivska and Hughes 2004). One hundred tree topologies generated for the full-length Pol gene were used to select sister pairs of sequences that had at least 50 per cent bootstrap support for the respective internal branches. Corresponding branch length values for these sister pairs (that are phylogenetically and statistically independent) were used to calculate the Pearson correlation coefficients (*r*) between all possible pairs of interacting epitope clusters and non-epitope segments. The combined correlation coefficient for each epitope pair was then calculated using the Pearson correlation coefficient scores which were statistically significant at the 0.01 significance level.

2.6 Branch length randomization

To evaluate whether detected strong correlations can be attributed to background coevolution noise and/or differences in segment sizes, we used a randomization approach, with randomly permutated branch length values, to evaluate the extent of such background noise due to shared phylogenetic history and stochastic noises. For each pair of IN-RT regions, corresponding branch lengths of IN were randomly permuted 1,000 times, and correlation coefficients were computed. The combined correlation coefficient for each epitope pair was then calculated using the mean values of correlation scores, which were significant at the 0.01 level.

2.7 Sliding window analysis

Since some of the top coevolving regions were fairly large in size (up to seventy-five residues), we performed a sliding window analysis, using thirty amino acid long peptides with a step size of fifteen residues, which allowed us to further narrow down the range of residues that show a coevolutionary signal. There were thirty-seven and nineteen sliding windows examined in RT and IN, respectively (some sliding windows spanned several epitope or non-epitope regions). Similar to previous steps, 1,000 tree topologies generated for the full-length Pol gene were used to estimate corresponding branch length values to calculate the Pearson correlation coefficients (*r*) between all possible pairs of sliding windows. The combined correlation coefficient for each sliding window pair was then calculated using *r* scores statistically significant at the 0.01 significance level, and the top coevolving sliding window pairs were identified as those with the combined correlation coefficient values at or above 0.5.

3. Results and discussion

To identify potential interacting regions between HIV-1 IN and RT, we estimated the degree of coevolution of epitope and non-epitope regions between two proteins by calculating combined correlation coefficients of evolutionary rates of all possible region pairs. Although phylogenetic tree-based coevolution analysis has been primarily used with whole protein sequences to identify interacting proteins, phylogenetic trees derived from protein domains have also been used to identify interacting domains, under the assumption that interacting domains show stronger coevolution signals than non-interacting domains in the same protein (Jothi et al. 2006; Kann et al. 2009; Dib and Carbone 2012; de Juan, Pazos and Valencia 2013). Here, we analysed coevolving regions between two proteins, where a region is defined as a cluster of overlapping epitopes (Table 1, Fig. 1) located next to each other.

3.1 Identification of coevolved sites between HIV-1 RT and IN

The combined correlation coefficient scores for all the IN-RT epitope/non-epitope cluster pairs are listed in Supplementary Table S3. The combined correlation coefficients ranged from 0.161 to 0.736 ($P = 0.01$) with a median value of 0.390 and an upper quartile at 0.457 (Supplementary Fig. S1). Here, we primarily focused on the top 15 per cent of interactions that had combined correlation coefficient of at least 0.5 (Table 2). We expect that these interacting regions are more likely to experience direct and/or prolonged (as opposed to transient) interactions than the regions that have correlation values below 0.5. The latter category can be expected to experience more transient and/or indirect interactions, resulting in lower correlation coefficients. However, it is possible that our stringent cut-off of 0.5 for the combined correlation coefficients may have missed some important interactions. For example, when interferences from other proteins/substances act on the IN-RT protein complex, this could be interpreted as weak coevolution signal because external interferences are not weighted. Further, some interactions may be occurring as a result of dynamic conformational changes where both direct and indirect interactions play a role at the same sites (e.g., Seckler et al. 2009; Ruff et al. 2014), thereby the coevolutionary signal is unable to distinguish between different types of interactions.

We identified 37 out of 244 epitope/non-epitope cluster pairs as the most likely regions to interact (Fig. 1, Table 2). As listed in Table 2, while interactions are distributed across all the domains of RT and IN, the majority of likely interactions is observed between the regions in the C-terminal end of the Finger-Palm domain, as well as in the Thumb, Connection, and RNaseH domains in RT protein and the regions in the N-terminal and catalytic core domains (NTD and CCD, respectively) and the latter end of C-terminal domain (CTD) in IN protein (see Fig. 1 for details). Notably, out of thirty-seven IN-RT cluster pairs listed in Table 2, twenty-eight pairs have some corroborating

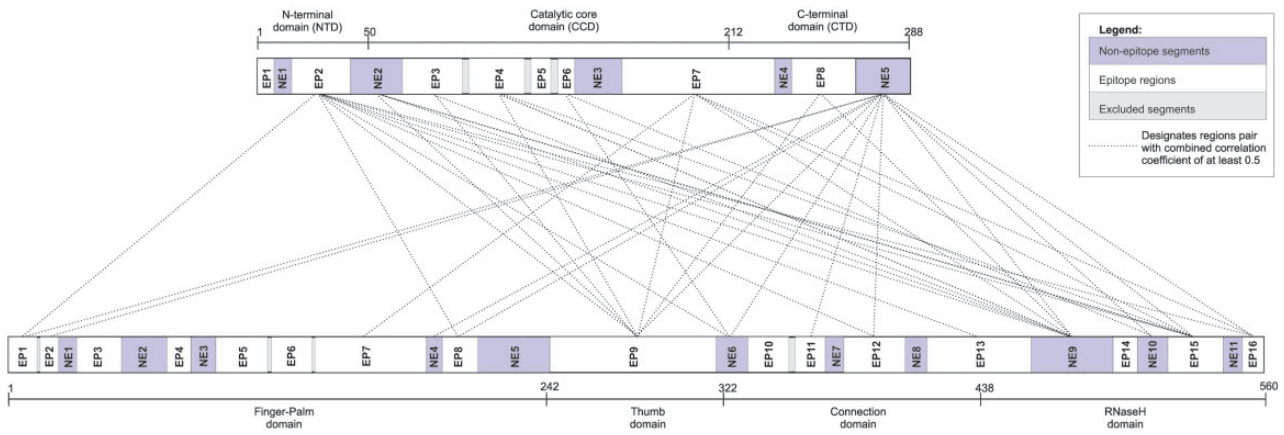

Figure 1. The distribution of the top coevolving region pairs in major domains of HIV-1 IN and RT proteins. Dotted lines represent potential interactions among thirty-seven region pairs that have the combined correlation coefficient of at least 0.5 or higher.

evidence from experimental studies of IN-RT interactions, which we will discuss below.

3.2 Distribution of coevolving region pairs reveals most frequent likely interacting domains between HIV-1 RT and IN

It is observed that the majority of likely coevolving region pairs (i.e., those with combined correlation coefficients at or above cut-off of 0.5) is located in the Finger-Palm domain and CTDs of RT and IN, respectively (Table 2, also Supplementary Fig. S2). Also, there is a significant number of coevolving pairs between the Connection domain of RT that appear to interact with regions within the CCD of IN. Likewise, identified coevolving pairs indicated interactions between regions in the Finger-Palm and the Connection domains of RT and the regions in the NTD of IN, as well as regions in the Finger-Palm domain and CTD of RT and IN, respectively (Fig. 1). Because the majority of the coevolving pairs identified above have combined correlation coefficients of 0.5 or higher, we expect that these domains are likely involved in direct (and/or prolonged) interactions. We should note that a substantial number of coevolving pairs with somewhat weak correlations is also located within the Finger-Palm domain and CCD of RT and IN, respectively (in other words, the number of coevolving pairs in the Finger-Palm and CCD are much higher at the threshold value of 0.4 relative to 0.5) (Supplementary Fig. S2 and Supplementary Fig. S3). This could be attributed to the fact that these two domains play an important role in functional interaction between IN and RT, with some segments likely involved in direct interactions (i.e., those with higher combined correlation coefficients), while others (i.e., those with lower combined correlation coefficients) may be playing a supporting role and/or be interacting through one or more intermediates.

3.3 Comparison of identified coevolved epitope/non-epitope regions that corresponds to experimentally known interacting regions between HIV-1 RT and IN

The experimental studies so far have shown that the C-terminal and Catalytic Core domains of IN, but not the N-terminal zinc-binding domain, were able to bind to RT (Hehl et al. 2004; Oz Gleenberg et al. 2005; Oz Gleenberg et al. 2007b; Wilkinson et al. 2009), although the specific details of binding and interaction interfaces remain unknown. For example, Hehl et al. (2004) reported that the carboxy-terminal domain of IN alone could interact with the Finger-Palm domain and the carboxy-terminal half of the Connection subdomain of RT (residues 1–242 and 387–422 of HXB2 IN, respectively) (Hehl et al. 2004). However, in the same study, it was suggested that other regions of IN may also be required for interaction with RT either directly or indirectly (Hehl et al. 2004). Furthermore, Oz Gleenberg et al. (2005) showed that a twenty amino acid-long peptide (residues 166–185, HXB2R p51) derived from the DNA polymerase active site of HIV-1 RT, which is located in the Palm sub domain, interacts with the CCD of IN and inhibits its disintegration activity (Oz Gleenberg et al. 2005). Several other non-inhibitory twenty amino acid-long RT-derived peptides that directly interact with full-length IN were also reported (Oz Gleenberg et al. 2005). The same group also described a twenty amino-acid-long peptide (residues 46–65, HXB2R p31), derived mostly from the IN-CTD that binds to RT and inhibits DNA polymerase activity and several other non-inhibitory binding peptides (Table 2 lists specific studies available for individual segments) (Oz Gleenberg et al. 2007b). However, the major limitation of these studies is that peptides rather than whole protein complexes were used, leaving the possibility that folded structures differ between the studied peptides and intact protein complexes *in vivo*.

The results of coevolutionary analysis showed that coevolving region pairs were not limited to the CCD and CTD but instead were distributed over the three domains of IN. This suggests that some protein regions in the NTD of IN can also interact with RT, for example, to increase the affinity of binding and/or efficiency of inhibition. Alternatively, these regions might be involved with structural stability of the IN-RT protein complex and might not be binding directly. It should be noted that coevolution analysis can provide a signal of co-functionality, either due to a direct physical binding or due to an indirect functional relationship without an explicit way to distinguish between these two types of interactions (Atchley et al. 2000; Xu et al. 2013).

Approximately 75 per cent of our identified coevolving region pairs correspond to experimentally shown interacting regions of RT and IN (twenty-eight out of thirty-seven). Table 2 lists both computationally predicted region pairs as well as any experimental evidence in support of such interactions. For example, it was shown that the most active peptide, which inhibited all IN activities (i.e., three-end processing, strand

Table 2. Top coevolving IN-RT epitope/non-epitope segment pairs, with corresponding structural domains of the two proteins and references to experimental studies of interactions.

RT epitope clusters/ non-epitope segment	Relevant RT domain	IN Epitope clusters/ non-epitope segments	Relevant IN domain	Combined correlation coefficient	Lower CI at 95% confidence level	Upper CI at 95% confidence level	Sample size[a]	Identified by analysis of phylogenetically independent sister pairs[b]	If top half of sliding windows in RT[c]	If top half of sliding windows in IN[c]	Experimental evidence (if available)
RT-EP1	Finger-Palm	IN-NE5	CTD	0.622	0.224	0.843	997				
RT-EP1	Finger-Palm	IN-EP2	NTD	0.499	0.293	0.660	999		+	+	
RT-EP11	Connection	IN-NE5	CTD	0.505	0.153	0.743	950		+	+	(Oz Gleenberg et al. 2005; Zawahir and Neamati 2006)
RT-EP12	Connection	IN-NE5	CTD	0.585	0.066	0.855	965			+	(Oz Gleenberg et al. 2005;
RT-EP12	Connection	IN-EP2	NTD	0.529	0.145	0.774	941	Yes	+	+	Zawahir and Neamati 2006)
RT-EP13	Connection and RNaseH	IN-EP2	NTD	0.531	0.264	0.722	956	Yes	+	+	(Oz Gleenberg et al. 2005; Zawahir and Neamati 2006)
RT-EP15	RNaseH	IN-EP2	NTD	0.532	0.373	0.661	1,000	Yes		+	
RT-EP15	RNaseH	IN-NE5	CTD	0.522	0.236	0.725	999		+	+	
RT-EP15	RNaseH	IN-EP4	CCD	0.503	0.322	0.648	1,000	Yes	+	+	
RT-EP15	RNaseH	IN-EP7	CCD and CTD	0.495	0.295	0.653	1,000	Yes	+		(Oz Gleenberg et al. 2007b)
RT-EP15	RNaseH	IN-NE2	NTD and CCD	0.495	0.328	0.632	1,000		+	+	(Oz Gleenberg, Goldgur and Hizi 2007a)
RT-EP16	RNaseH	IN-NE5	CTD	0.601	0.131	0.851	996		+	+	
RT-EP16	RNaseH	IN-EP2	NTD	0.544	0.309	0.716	1,000		+	+	
RT-EP2	Finger-Palm	IN-NE5	CTD	0.507	0.059	0.785	801		+	+	
RT-EP7	Finger-Palm	IN-EP7	CCD and CTD	0.498	0.209	0.707	994	Yes	+	+	(Oz Gleenberg et al. 2005; Zawahir and Neamati 2006; (Oz Gleenberg, Goldgur and Hizi 2007a)
RT-EP8	Finger-Palm	IN-NE5	CTD	0.658	0.135	0.894	997			+	(Oz Gleenberg et al. 2005;
RT-EP8	Finger-Palm	IN-EP2	NTD	0.555	0.273	0.749	994			+	Zawahir and Neamati 2006)
RT-EP9	Thumb	IN-EP2	NTD	0.598	0.466	0.705	1,000	Yes		+	(Oz Gleenberg et al. 2005)
RT-EP9	Thumb	IN-EP7	CCD and CTD	0.576	0.377	0.724	1,000	Yes	+	+	
RT-EP9	Thumb	IN-EP4	CCD	0.552	0.400	0.675	1,000	Yes	+	+	
RT-EP9	Thumb	IN-EP8	CTD	0.542	0.335	0.700	1,000	Yes	+		
RT-EP9	Thumb	IN-NE5	CTD	0.530	0.257	0.725	998		+	+	
RT-EP9	Thumb	IN-EP3	CCD	0.529	0.394	0.642	1,000	Yes	+	+	
RT-EP9	Thumb	IN-NE2	NTD and CCD	0.526	0.389	0.640	1,000	Yes	+		
RT-NE10	RNaseH	IN-NE5	CTD	0.708	0.039	0.939	991		+	+	
RT-NE10	RNaseH	IN-EP6	CCD	0.527	0.065	0.803	760		+	+	(Oz Gleenberg et al. 2007b)
RT-NE4	Finger-Palm	IN-NE5	CTD	0.518	0.026	0.808	992		+	+	(Oz Gleenberg et al. 2005; Zawahir and Neamati 2006)
RT-NE6	Thumb and Connection	IN-NE5	CTD	0.736	0.283	0.921	998			+	(Oz Gleenberg et al. 2005)
RT-NE6	Thumb and Connection	IN-EP6	CCD	0.540	0.132	0.792	781		+	+	
RT-NE6	Thumb and Connection	IN-EP2	NTD	0.520	0.242	0.719	999		+	+	
RT-NE9	RNaseH	IN-NE5	CTD	0.604	0.281	0.804	998		+	+	Oz Gleenberg et al. (2005)
RT-NE9	RNaseH	IN-EP2	NTD	0.585	0.428	0.708	1,000	Yes	+	+	
RT-NE9	RNaseH	IN-EP4	CCD	0.565	0.358	0.718	1,000	Yes	+	+	

(continued)

Table 2. Continued

RT epitope clusters/ non-epitope segment	Relevant RT domain	IN Epitope clusters/ non-epitope segments	Relevant IN domain	Combined correlation coefficient	Lower CI at 95% confidence level	Upper CI at 95% confidence level	Sample size[a]	Identified by analysis of phylogenetically independent sister pairs[b]	If top half of sliding windows in RT[c]	If top half of sliding windows in IN[c]	Experimental evidence (if available)
RT-NE9	RNaseH	IN-EP7	CCD and CTD	0.545	0.338	0.702	1,000	Yes	+		
RT-NE9	RNaseH	IN-EP3	CCD	0.541	0.378	0.671	1,000	Yes	+	+	
RT-NE9	RNaseH	IN-EP8	CTD	0.508	0.255	0.696	998	Yes	+		
RT-NE9	RNaseH	IN-NE2	NTD and CCD	0.497	0.318	0.642	1,000	Yes	+	+	

For each pair of regions, combined correlation coefficient (with confidence intervals) is given. Only top 15 per cent of pairs that have the combined correlation coefficient at or above 0.5 threshold are listed.
[a]Sample size refers to the number of samples (out of 1,000) that had a significant Pearson correlation value in at least two-third of the total samples (i.e., 750 samples out of 1,000).
[b]IN/RT regions that are identified by analysis of phylogenetic independent sequence pairs.
[c]'+' designates whether the specific region has been identified as part of the top half of coevolving sliding windows by the number of coevolving regions (e.g., sliding windows in RT that coevolve with 5–8 IN regions; see Section 2 for further details).

transfer, and disintegration), came from a DNA-polymerase active-site (amino acid 166–185) (Oz Gleenberg Goldgur and Hizi 2007a), which has been identified in our findings as well. Further, the surface loop (residues 141–148 of RT) that is disordered has been shown to be important for substrate binding and catalysis (Oz Gleenberg, Goldgur and Hizi 2007a), in agreement with our identification of the positions 135–143 as one of the coevolving regions. On the other hand, it is possible that interactions between individual residues may have been missed because of the limitations of the coevolutionary approach, such as the interactions between Gly-149 of IN and Ile-178 of RT that were shown to interact with each other through molecular docking analysis (Oz Gleenberg Goldgur and Hizi 2007a).

Since the size of protein segments analysed here varied from 7 to 70 and 6 to 75 residues in IN and RT, respectively (with median sizes of twenty one and sixteen residues, respectively), our results allow us to further narrow down the size and location of peptides that are potentially interacting in the IN-RT protein complex (Fig. 1, Supplementary Fig. S4). Our results suggest that two domains of IN, including several regions in the NTD and the latter end of the C-terminal, interact with the full-length RT protein. The results also reveal strong coevolutionary signals implying existence of interactions between the Catalytic Core domain of IN and several domains of RT, including the C-terminal end of the Finger-Palm domain, as well as with the Thumb, Connection, and RNaseH domains.

By examining the number of potential interactions for each region, several segments can be identified as the top candidates due to their large number of likely interactions (Table 2). Specifically, within IN, IN-EP2 and IN-NE5 regions, which are located within N- and C-terminals, respectively, were found to exhibit strong correlation coefficients with ten and twelve RT regions. Likewise, within RT, RT-EP9, RT-NE9, and RT-EP15 regions, all mapped within the RNaseH domain had strong correlation coefficients with seven, seven, and five IN regions, respectively. Identification of such broadly interacting regions can help to focus future experimental studies by allowing the further narrowing down of the list of the most likely interacting positions to only about 16 per cent and 25 per cent of total positions in IN and RT, respectively (45 out of 283, and 137 out of 550 amino acid positions). Furthermore, our results may also provide a basis for excluding specific positions from consideration in experimental studies. Specifically, in RT over 35 per cent and in IN over 15 per cent of positions (194 out of 550 and 43 out of 283 positions, respectively) can be excluded as likely not interacting ones.

One of the concerns often raised in coevolutionary studies is whether an underlying phylogenetic history influences detected relationships (e.g., Sato et al. 2005). Thus, to remove the potential influence of a shared phylogenetic history, we repeated our analysis using phylogenetically independent comparisons of pairs of sequences (Felsenstein 1985); in other words, sister pairs. Combined correlation coefficients for each epitope pair were computed using only the identified sister pairs (see Section 2 for details), and the top 15 per cent (thirty-seven) scoring region pairs were examined as the most likely interacting and coevolving candidates (See Supplementary Table S4). Notably, about half of the identified coevolving region pairs were the same as those identified in our initial analysis (eighteen out of thirty-seven region pairs; see Table 2 for details). Further, of those eighteen region pairs, when the number of interactors of individual regions was considered, RT regions mapped within RNaseH domain, namely, RT-EP9, RT-NE9, and RT-EP15, were found to have strong correlation coefficients with

six, six, and three IN regions, respectively, consistent with the above results. Likewise, the IN-EP2 region was also found to interact with five RT regions, in agreement with prior findings. Thus, the agreement between two approaches allows us to further narrow down the set of residues as the most likely interacting pairs for future experimental confirmation.

We also performed a sliding window analysis, in an attempt to further narrow down the range of interacting residues from large coevolving regions. A total of eighty-eight sliding window pairs were identified as the top coevolving sliding window pairs (i.e., those with the combined correlation coefficients of 0.5 or above), which in turn comprised nineteen and eight sliding windows from RT and IN, respectively. Eighteen out of nineteen and eight out of eight of top coevolving sliding windows were co-located with the top coevolving regions identified in Table 2 and corresponded to fifty out of eighty-eight sliding window pairs. Of the remaining thirty-eight sliding window pairs that were not already represented in Table 2, only one pair included a novel segment from RT. We further examined the total number of coevolving regions linked to each sliding window (ranging from one to eight for RT, and three to eighteen in IN) and identified those in the top half by the number of coevolving regions (i.e., those with five to eight, and ten to eighteen regions in RT and IN, respectively) (Table 2). Thus, the results of the sliding window analysis were mostly consistent with prior findings and may offer additional insights into the strength of coevolutionary relationships among individual regions. This approach also can provide additional information for the design of future experimental validation or structural studies, for example, region pairs with a large number of coevolving partners may be expected to be more flexible than those with a small number of partners (Seckler et al. 2009).

By using a randomization approach, we examined the possibility that strong correlations identified among multiple region pairs could be attributed to background noise instead of functional/structural constraints. The results showed that for randomized datasets, the maximum value of the upper confidence interval of the combined correlation coefficient does not exceed approximately 0.063, which is significantly smaller than the smallest combined correlation coefficient values of approximately 0.495 and approximately 0.298, which were derived from the empirical (non-randomized) datasets ($P < 0.001$, per non-parametric two-sample rank tests) (Table 2, see also Supplementary Table S4 and Fig. S5). This indicates that while it is possible that some of the detected correlations may be attributable to random noise, the reported values for the top 15 per cent region pairs cannot be attributed to random correlations.

While our study provides evidence about potentially interacting regions between IN and RT proteins, there are also known limitations of this approach. Because we used a stringent combined correlation coefficient threshold of 0.5, we may have missed potential interactions. Supplementary Fig. S3 shows a broader range of potentially interacting regions (those with combined correlation coefficient value cut-off at 0.4). Further, the lower correlation value may also be attributed to interference from other proteins/substances that play a role in the IN-RT protein complex. For example, the coevolutionary signal of two interacting regions may appear weaker and, hence, result in a relatively low score when those regions interact with multiple partners (e.g., Jothi et al. 2006). Another challenge is not being able to recognize all the interacting proteins involved in a complex. For example, both IN and RT proteins are inhibited by Vpr protein, either together or possibly separately (BouHamdan et al. 2000; Gleenberg, Herschhorn and Hizi 2007). Thus, interactions with Vpr could be mediating other interactions within the IN-RT complex (e.g., Gleenberg, Herschhorn and Hizi 2007). Although Vpr peptides could not be included in this pairwise-based approach, this question should be addressed in future studies. There is definitely a need to develop a suitable method to account for multiple partners.

Overall, the identification of coevolving residues is a promising approach to predict potential targets for multi-epitope- or adjuvant-based treatments. The inclusion of conserved functionally important epitopes from different genomic regions holds promise in eliciting a strong immune response (e.g., Grimm and Ackerman 2013), and our results help to narrow down the list of such functionally important candidates. Likewise, these functionally and/or structurally important regions can be used to design novel protein inhibitors that will target the IN-RT complex. These findings can further inform validation studies, for example, using site-directed mutagenesis, and/or be integrated with structural data and physicochemical amino acid properties to gain a better understanding of the mechanism of protein interactions between RT and IN proteins of HIV-1. For proteins with poorly resolved three-dimensional structures, like IN, coevolution analysis offers important clues to enable a better understanding of likely functional interactions and to identify specific segments and/or residues involved in interaction interfaces.

4. Conclusion

Coevolutionary analysis can provide important predictions that can be integrated with structural data to gain a better understanding of the mechanisms of protein interaction between RT and IN, including interactions in transient protein–protein complexes such as the IN-RT complex within PIC. For example, our results identified multiple interactions occurring in parts of the NTD and CTD of IN. Likewise, in RT, many interactions appear to involve the Connection and RNaseH domains. Thus, these results highlight sets of residues likely involved in interaction interface(s) of HIV-1 RT and IN protein complex through coevolutionary signatures, although experimental validation of specific interactions is needed.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

This work was partially supported by NIH NIGMS grant GM86782-01A1 to H.P. We would like to thank our fellow lab members Reeba Paul, Mary Halpin, Yi Wei and Kendall Furbee for their insightful comments.

Conflict of interest: None declared.

References

Arts, E. J., and Hazuda, D. J. (2012) 'HIV-1 Antiretroviral Drug Therapy', *Cold Spring Harbor Perspectives in Medicine*, 2: a007161.

Atchley, W. R., et al. (2000) 'Correlations Among Amino Acid Sites in Bhlh Protein Domains: An Information Theoretic Analysis', *Molecular Biology and Evolution*, 17: 164–78.

BouHamdan, M., et al. (2000) 'Inhibition of Hiv-1 Replication and Infectivity by Expression of a Fusion Protein, Vpr-Anti-

Integrase Single-Chain Variable Fragment (SFv): Intravirion Molecular Therapies', *Journal of Human Virology*, 3: 6–15.

Brady, M. T., et al. (2010) 'Declines in Mortality Rates and Changes in Causes of Death in Hiv-1-Infected Children During the HAART Era', *Journal of Acquired Immune Deficiency Syndromes*, 53: 86–94.

Chakraborty, A., et al. (2013) 'Biochemical Interactions between HIV-1 Integrase and Reverse Transcriptase', *FEBS Letters*, 587: 425–9.

Chiu, T. K., and Davies, D. R. (2004) 'Structure and Function of HIV-1 Integrase', *Current Topics in Medicinal Chemistry*, 4: 965–77.

Clark, N. L., Alani, E., and Aquadro, C. F. (2012) 'Evolutionary Rate Covariation Reveals Shared Functionality and Coexpression of Genes', *Genome Research*, 22: 714–20.

Codoner, F. M., and Fares, M. A. (2008) 'Why Should We Care About Molecular Coevolution?', *Evolutionary Bioinformatics Online*, 4: 29–38.

de Juan, D., Pazos, F., and Valencia, A. (2013) 'Emerging Methods in Protein Co-Evolution', *Nature Reviews Genetics*, 14: 249–61.

Deeks, S. G., et al. (2012) 'Towards an HIV Cure: A Global Scientific Strategy', *Nature Reviews Immunology*, 12: 607–14.

Dib, L., and Carbone, A. (2012) 'Protein Fragments: Functional and Structural Roles of Their Coevolution Networks', *PLoS One*, 7: e48124.

Engelman, A., and Cherepanov, P. (2012) 'The Structural Biology of HIV-1: Mechanistic and Therapeutic Insights', *Nature Reviews Microbiology*, 10: 279–90.

Felsenstein, J. (1985) 'Phylogenies and the Comparative Method', *The American Naturalist*, 125: 1–15.

Fodor, A. A., and Aldrich, R. W. (2004) 'Influence of Conservation on Calculations of Amino Acid Covariance in Multiple Sequence Alignments', *Proteins*, 56: 211–21.

Garcia, E. (2012) 'The Self-Weighting Model', *Communications in Statistics-Theory and Methods*, 41: 1421–7.

Gene Cutter Tool–HIV LANL Database. <http://www.hiv.lanl.gov/content/sequence/GENE_CUTTER/cutter.html> accessed 14 July 2014.

Gleenberg, I. O., Herschhorn, A., and Hizi, A. (2007) 'Inhibition of the Activities of Reverse Transcriptase and Integrase of Human Immunodeficiency Virus Type-1 by Peptides Derived from the Homologous Viral Protein R (Vpr)', *Journal of Molecular Biology*, 369: 1230–43.

Goulder, P. J., et al. (1997) 'Patterns of Immunodominance in HIV-1-Specific Cytotoxic T Lymphocyte Responses in Two Human Histocompatibility Leukocyte Antigens (HLA)-Identical Siblings with HLA-A*0201 are Influenced by Epitope Mutation', *Journal of Experimental Medicine*, 185: 1423–33.

Grimm, S. K., and Ackerman, M. E. (2013) 'Vaccine design: Emerging Concepts and Renewed Optimism', *Current Opinion in Biotechnology*, 24: 1078–88.

Hehl, E. A., et al. (2004) 'Interaction between Human Immunodeficiency Virus Type 1 Reverse Transcriptase and Integrase Proteins', *Journal of Virology*, 78: 5056–67.

Herschhorn, A., Oz-Gleenberg, I., and Hizi, A. (2008) 'Quantitative Analysis of the Interactions between HIV-1 Integrase and Retroviral Reverse Transcriptases', *Biochemical Journal*, 412: 163–70.

Huang, M., Grant, G. H., and Richards, W. G. (2011) 'Binding Modes of Diketo-Acid Inhibitors of HIV-1 Integrase: A Comparative Molecular Dynamics Simulation Study', *Journal of Molecular Graphics and Modelling*, 29: 956–64.

International AIDS Vaccine Initiative (IAVI). (2012) *AIDS Vaccines: Exploring the Potential Cost/Benefit 2012* <http://www.iavi.org/Information-Center/Publications/Documents/Costs%20Impact%20Brief.pdf> accessed 25 July 2014.

Jothi, R., et al. (2006) 'Co-evolutionary Analysis of Domains in Interacting Proteins Reveals Insights Into Domain-Domain Interactions Mediating Protein-Protein Interactions', *Journal of Molecular Biology*, 362: 861–75.

Kann, M. G., et al. (2009) 'Correlated Evolution of Interacting Proteins: Looking Behind the Mirrortree', *Journal of Molecular Biology*, 385: 91–8.

Koletar, S. L., et al. (2004) 'Long-Term Follow-Up of HIV-Infected Individuals Who Have Significant Increases in CD4+ Cell Counts During Antiretroviral Therapy', *Clinical Infectious Diseases*, 39: 1500–06.

Krishnan, L., and Engelman, A. (2012) 'Retroviral Integrase Proteins and HIV-1 DNA Integration', *Journal of Biological Chemistry*, 287: 40858–66.

Kumar, S., et al. (2012) 'MEGA-CC: Computing Core of Molecular Evolutionary Genetics Analysis Program for Automated and Iterative Data Analysis', *Bioinformatics*, 28: 2685–6.

Le Douce, V., et al. (2012) 'Achieving a Cure for HIV Infection: Do We Have Reasons to Be Optimistic?', *Journal of Antimicrobial Chemotherapy*, 67: 1063–74.

Levy, N., et al. (2013) 'Structural and Functional Studies of the HIV-1 Pre-Integration Complex', *Retrovirology*, 10(Suppl. 1): P76.http://www.retrovirology.com/content/10/S1/P76/

Lewin, S. R., et al. (2011) 'Finding a Cure for HIV: Will It Ever Be Achievable?', *Journal of the International AIDS Society*, 14: 4.

Li, W. L., and Rodrigo, A. G. (2009) 'Covariation of Branch Lengths in Phylogenies of Functionally Related Genes', *PLoS One*, 4: e8487.

Llano, A. et al. (2013) 'Best-Characterized HIV-1 CTL Epitopes: The 2013 Update', in K., Yusim et al. (eds.) *HIV Molecular Immunology 2013*, pp. 3–25. Los Alamos, NM: Los Alamos National Laboratory, Theoretical Biology and Biophysics.

Luber, A. D. (2005) 'Genetic Barriers to Resistance and Impact on Clinical Response', *Journal of the International AIDS Society*, 7: 69.

Maenza, J., and Flexner, C. (1998) 'Combination Antiretroviral Therapy for HIV Infection', *American Family Physician*, 57: 2789–98.

Mehellou, Y., and De Clercq, E. (2010) 'Twenty-Six Years of Anti-HIV Drug Discovery: Where Do We Stand and Where Do We Go?', *Journal of Medicinal Chemistry*, 53: 521–38.

Neamati, N., and Wang B.(ed.) (2011) *HIV-1 integrase: mechanism and inhibitor design*. John Wiley & Sons, Hoboken, NJ.

Nooren, I. M., and Thornton, J. M. (2003) 'Structural Characterisation and Functional Significance of Transient Protein-Protein Interactions', *Journal of Molecular Biology*, 325: 991–1018.

Oz, I., Avidan, O., and Hizi, A. (2002) 'Inhibition of the Integrases of Human Immunodeficiency Viruses Type 1 and Type 2 by Reverse Transcriptases', *Biochemical Journal*, 361(Pt 3): 557–66.

Oz Gleenberg, I., et al. (2005) 'Peptides Derived From the Reverse Transcriptase of Human Immunodeficiency Virus Type 1 as Novel Inhibitors of the Viral Integrase', *Journal of Biological Chemistry*, 280: 21987–96.

——, Goldgur, Y., Hizi, A. (2007a) 'Ile178 of HIV-1 Reverse Transcriptase is Critical for Inhibiting the Viral Integrase', *Biochemical and Biophysical Research Communication*, 364: 48–52.

——, et al. (2007b) 'Inhibition of Human Immunodeficiency Virus Type-1 Reverse Transcriptase by a Novel Peptide Derived From the Viral Integrase', *Archives of Biochemistry and Biophysics*, 458: 202–12.

Patel, R., Garde, C., and Stormo, G. (2015) 'Determination of Specificity Influencing Residues for Key Transcription Factor Families', *Quantitative Biology*, 3: 115–23.

Peters, B. S., and Conway K. (2011) 'Therapy for HIV: Past, Present, and Future', *Advances in Dental Research*, 23: 23–7.

Piontkivska, H., and Hughes, A. L. (2004) 'Between-Host Evolution of Cytotoxic T-Lymphocyte Epitopes in Human Immunodeficiency Virus Type 1: An Approach Based on Phylogenetically Independent Comparisons', *Journal of Virology*, 78: 11758–65.

R Core Team (2015). *R: A Language and Environment for Statistical Computing, version 3.1.1*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/> accessed 30 March 2015.

Ruff, M., et al. (2014) 'Structural and Functional Studies of HIV-1 Pre-Integration Complexes', *BMC Infectious Diseases*, 14(Suppl. 2): O9.

Sarafianos, S. G., et al. (2009) 'Structure and Function of HIV-1 Reverse Transcriptase: Molecular Mechanisms of Polymerization and Inhibition', *Journal of Molecular Biology*, 385: 693–713.

SAS Institute Inc. (2011) *SAS® 9.3 System Options: Reference, Second Edition*, Cary, NC: SAS Institute Inc.

Sato, T., et al. (2005) 'The Inference of Protein-Protein Interactions by Co-Evolutionary Analysis is Improved by Excluding the Information About the Phylogenetic Relationships', *Bioinformatics*, 21: 3482–9.

Seckler, J. M., et al. (2009) 'Solution Structural Dynamics of HIV-1 Reverse Transcriptase Heterodimer', *Biochemistry*, 48: 7646–55.

Tamura, K., et al. (2013) '"MEGA6: Molecular Evolutionary Genetics Analysis version 6.0.", *Mol Biol Evol*, 30: 2725–2729.

Tasara, T., et al. (2001) 'HIV-1 Reverse Transcriptase and Integrase Enzymes Physically Interact and Inhibit Each Other', *FEBS Letters*, 507: 39–44.

UNAIDS (2013). Global Report on AIDS Epidemic. UNAIDS, November 2013.<http://www.unaids.org/sites/default/files/media_asset/UNAIDS_Global_Report_2013_en_1.pdf> accessed 23 Jul 2014.

Warren, K., et al. (2009) 'Reverse Transcriptase and Cellular Factors: Regulators of HIV-1 Reverse Transcription', *Viruses*, 1: 873–94.

Wilkinson, T. A., et al. (2009) 'Identifying and Characterizing a Functional HIV-1 Reverse Transcriptase-Binding Site on Integrase', *Journal of Biological Chemistry*, 284: 7931–9.

World Health Organization (2014). *World Health Statistics*. Geneva: World Health Organization.<http://www.who.int/gho/publications/world_health_statistics/2014/en/> accessed 23 Jul 2014.

Xu, H., et al. (2013) 'Identifying Coevolution between Amino Acid Residues in Protein Families: Advances in the Improvement and Evaluation of Correlated Mutation Algorithms', *Current Bioinformatics*, 8: 148–60.

Yonezawa, K., et al. (2013) 'Resampling Nucleotide Sequences with Closest-Neighbor Trimming and its Comparison To Other Methods', *PLoS One*, 8: e57684.

Yusim, K., et al. (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', *Journal of Virology*, 76: 8757–68.

Zawahir, Z., and Neamati, N. (2006) 'Inhibition of HIV-1 Integrase Activity by Synthetic Peptides Derived From the HIV-1 HXB2 Pol Region of the Viral Genome', *Bioorganic and Medicinal Chemistry Letters*, 16: 5199–202.

Zhang, J., and Crumpacker, C. (2013) 'Eradication of HIV and Cure of AIDS, Now and How?', *Frontiers in Immunology*, 4: 337.