

Research

Open Access

## Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences

Nobubelo K Ngandu<sup>1</sup>, Konrad Scheffler<sup>2</sup>, Penny Moore<sup>3</sup>, Zenda Woodman<sup>4</sup>, Darren Martin<sup>4</sup> and Cathal Seoighe\*<sup>1</sup>

Address: <sup>1</sup>National Bioinformatics Network Node, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Anzio Road, Observatory, 7925, South Africa, <sup>2</sup>Computer Science Division, Dept of Mathematical Sciences, University of Stellenbosch, Private Bag X1, 7602 Matieland, Stellenbosch, South Africa, <sup>3</sup>National Institute for Communicable Diseases, Private Bag X4, Sandringham, Johannesburg, 2131, South Africa and <sup>4</sup>HIV Diversity and Pathogenesis Group, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences, University of Cape Town, Anzio Road, Observatory, 7925, Cape Town, South Africa

E-mail: Nobubelo K Ngandu - nobubelo@cbio.uct.ac.za; Konrad Scheffler - konrad@cbio.uct.ac.za; Penny Moore - pennym@nicd.ac.za; Zenda Woodman - zenda.woodman@uct.ac.za; Darren Martin - Darrin.Martin@uct.ac.za; Cathal Seoighe\* - cseoighe@gmail.com;

\*Corresponding author

Published: 23 December 2008

Received: 2 December 2008

*Virology Journal* 2008, **5**:160 doi: 10.1186/1743-422X-5-160

Accepted: 23 December 2008

This article is available from: <http://www.virologyj.com/content/5/1/160>

© 2008 Ngandu et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Positive selection pressure acting on protein-coding sequences is usually inferred when the rate of nonsynonymous substitution is greater than the synonymous rate. However, purifying selection acting directly on the nucleotide sequence can lower the synonymous substitution rate. This could result in false inference of positive selection because when synonymous changes at some sites are under purifying selection, the average synonymous rate is an underestimate of the neutral rate of evolution. Even though HIV-1 coding sequences contain a number of regions that function at the nucleotide level, and are thus likely to be affected by purifying selection, studies of positive selection assume that synonymous substitutions can be used to estimate the neutral rate of evolution.

**Results:** We modelled site-to-site variation in the synonymous substitution rate across coding regions of the HIV-1 genome. Synonymous substitution rates were found to vary significantly within and between genes. Surprisingly, regions of the genome that encode proteins in more than one frame had significantly higher synonymous substitution rates than regions coding in a single frame. We found evidence of strong purifying selection pressure affecting synonymous mutations in fourteen regions with known functions. These included an exonic splicing enhancer, the rev-responsive element, the poly-purine tract and a transcription factor binding site. A further five highly conserved regions were located within known functional domains. We also found four conserved regions located in *env* and *vpu* which have not been characterized previously.

**Conclusion:** We provide the coordinates of genomic regions with markedly lower synonymous substitution rates, which are putatively under the influence of strong purifying selection pressure at the nucleotide level as well as regions encoding proteins in more than one frame. These regions should be excluded from studies of positive selection acting on HIV-1 coding regions.

## Background

Several statistical models of codon evolution have been developed and applied to protein-coding sequences from viral and other pathogens [1-4]. The primary application of these models has been the detection of evidence of diversifying selection acting on protein coding DNA sequences. Within maximum likelihood or Bayesian frameworks these models can be used to identify specific sites at which adaptive mutations have occurred. In the context of virus infections this information can be especially useful for identifying immune escape and drug resistance mutations [3, 5, 6].

Positive selection is frequently inferred by comparing the rate of non-synonymous substitutions per non-synonymous site (dN) to the rate of synonymous substitutions per synonymous site (dS). The ratio of these two rates is often represented by the symbol  $\omega$ . Under the assumption that synonymous substitutions are neutral and that the synonymous substitution rate therefore approximates the neutral rate of evolution, diversifying selection can be inferred when  $\omega$  is greater than one. Several methods exist to determine whether there is evidence that  $\omega$  is greater than one at a subset of sites in a protein-coding gene (i.e. the gene is evolving under diversifying selection) and to identify the sites within the gene at which diversifying selection occurs [3, 4, 7-9].

In many of the situations in which this strategy is applied, the assumption that synonymous substitutions are fixed at a constant rate and provide a good estimate of the neutral rate of evolution, may not hold. Kosakovsky Pond & Muse reported that coding sequences from a wide range of taxa, including HIV-1, show strong evidence of variation in the rate of synonymous substitution across coding regions [10]. There are two possible causes of synonymous rate variation. If synonymous substitutions are indeed neutral, variation in the mutation rate can cause the synonymous substitution rate to vary. In such a case, it is possible to include a varying synonymous substitution rate in the codon models of evolution and inference of positive selection from comparison of the local synonymous and nonsynonymous substitution rates remains feasible. However, if the variation in synonymous substitution rate is caused by selection acting to preserve functions that are encoded at the nucleotide level, even a comparison of local nonsynonymous and synonymous substitution rates cannot be used to infer positive selection because the synonymous substitution rate is no longer a valid proxy for the neutral rate of evolution and the standard approach of inferring the action of diversifying selection when  $\omega > 1$  is not valid.

Failure to model variation in synonymous substitution rate will result in an overall underestimate of the neutral rate of evolution. This undermines the validity of the inference of selection, because nonsynonymous substitution rates are compared against a rate which is no longer a good estimate of the neutral rate, and this is likely to result in inference of diversifying selection at a proportion of the sites that are actually evolving neutrally. Indeed, as the number of taxa increases, we expect an ever greater proportion of the neutral sites to be classified as diversifying selection sites in this scenario. Alternatively, if the synonymous substitution rate variation is modeled and selection inferred when dN is greater than the local dS rate then we expect a very high probability of false inference of selection at codons where the synonymous positions happen to be functionally important and conserved, and the nonsynonymous positions are neutral or experience less purifying selection. Thus, in general, in a codon-based method, analysis of selection is unreliable when there is purifying selection acting to preserve functions at the nucleotide level. An example where an elevated  $\omega$  was attributable to purifying selection acting on synonymous sites was reported by Hurst & Pal [11].

Several examples of sequence motifs within protein-coding sequences that are expected to be under purifying selection at the nucleotide level are known in HIV-1, many of which are involved in regulating gene expression. Examples include the 3' long terminal repeat (LTR) region, part of which also encodes the Nef protein [12, 13]. In addition to conserved RNA secondary structures, the LTR contains several regulatory elements, some of which directly interact with cellular transcription factors (e.g. the Ets protein family) [14, 15]. The viral sequence also has an intragenic nuclease hypersensitive region involved in regulating gene expression (also referred to as HS7) in the *pol* gene [16-18] and the rev-responsive element (RRE) in the *env* gene which interacts with the Rev protein to transport unspliced or partially spliced RNA from the nucleus to the cytoplasm of the infected cell [17, 19-21]. Some functionally important regions of the RRE have previously been found to be conserved at the nucleotide sequence level, presumably the result of purifying selection pressure to preserve this function [22]. The inhibitory sequence elements (INS) in gag and the cis-repressive sequence (CRS) in *pol* are examples of negative regulators of transcription [23-26]. If these functional sites are important for viral viability, then we expect them to be preserved by purifying selection.

While it represents a significant challenge for studies of selection acting on the HIV-1 amino acid sequence, the

variability in the synonymous substitution rate may also provide useful information about previously unknown sequence motifs within the coding fraction of the HIV-1 genome that function at the nucleotide level. Although some variability can be explained by a variable mutation rate, the identification of regions of very high conservation that cannot be explained by selection acting on the amino acid sequence or by known motifs that function at the nucleotide level has the potential to highlight novel functions encoded in the HIV-1 genome.

Here we use an existing model of codon sequence evolution [10] to provide the first complete overview of site-to-site variation in synonymous substitution rate across the whole HIV-1 genome and identify selection pressures likely to be driving this variation. This model allows dN and dS to vary independently across sites, ensuring that the estimated dS values reflect selection pressure acting upon the nucleotide sequence and not at the amino acid level. It is worthwhile to distinguish between selective pressure acting at the nucleotide level, affecting both synonymous and nonsynonymous changes, and at the amino acid level, affecting nonsynonymous changes only. Unfortunately, quantifying the relative contributions of nucleotide and amino acid level effects on nonsynonymous changes is highly sensitive to model assumptions. We therefore restricted the analysis to synonymous changes and do not attempt to quantify the nucleotide-level selective pressure on nonsynonymous changes. We took into account recombination breakpoints in order to avoid biased estimates that can result from fitting phylogenetic models that do not take recombination into account [27, 28].

We report patterns of sequence conservation around nucleotide sequence motifs with known functions and identify additional conserved nucleotide elements that do not fall within any currently characterized functional motifs. Finally, we report the locations of all HIV-1 genome regions where we infer that purifying selection acting directly on the nucleotide sequence is likely to cause a substantial reduction in the synonymous substitution rate. These are provided with respect to the HXB2 reference strain, to enable other researchers to mask these regions from their analyses of positive selection acting on HIV-1 genes.

## Methods

### Sequence data

Nucleotide sequence alignments consisting of HIV-1 Group M subtype reference sequences were downloaded from the Los Alamos database <http://www.hiv.lanl.gov> for each gene of the HIV-1 genome [29]. Each alignment

had at least one sequence (total ranging from 32 to 37) from each of the 11 non-recombinant HIV-1 group M subtypes A1, A2, B, C, D, F1, F2, G, H, J and K [Genbank: AB253421, AB253429, AF004885, AF005494, AF005496, AF061641, AF061642, AF067155, AF069670, AF075703, AF077336, AF082394, AF082395, AF084936, AF190127, AF190128, AF286237, AF286238, AF377956, AF484509, AJ249235, AJ249236, AJ249237, AJ249238, AJ249239, AY173951, AY253311, AY331295, AY371157, AY371158, AY423387, AY612637, AY772699, DQ676872, DQ853463, K03454, K03455, U46016, U51190, U52953, U88824, U88826]. All regions encoding amino acids in more than one frame were identified and regions judged by eye to be unreliably aligned, i.e., positions 6544–6595, 6700–6715, 7318–7375 of the *env* gene region, were excluded from the analysis. We used the HIV-1 genome map and sequence annotations available from the Los Alamos database to identify the regions of the genome that encode proteins in a single reading frame (see Table 1) [29].

We identified recombination breakpoints in each alignment using the GARD (Genetic Algorithm for Recombination Detection) algorithm implemented in the HyPhy (Hypothesis testing using Phylogenies) package [30, 31]. Evidence of recombination was detected in all genes except *tat*, *vpr* and *vpu*. GARD outputs both an alignment showing the positions of recombination breakpoints and separate tree topologies for each of the sequence alignment segments bounded by these breakpoints.

### Synonymous substitution rate estimation

Synonymous substitution rates were estimated using a version of the MG94 codon substitution model [1, 10]. We used the Dual Model, which allows dS to vary independently of dN and used three discrete categories for each rate. We ran the selected models using a HyPhy batch script for analysis of selection acting on recombining sequences which we developed previously [32]. This method uses separate tree topologies for each partition of the sequence alignment while keeping the rest of the model parameters fixed across all partitions. We used sliding window plots of mean dS values, calculated over three adjacent codons, to identify regions with low synonymous substitution rates.

We also analyzed an alignment of HIV-1 subtype C *gag* sequences [Genbank: DQ792982–DQ793045] described previously [33] to assess the impact of conservation acting on synonymous substitutions on inference of positive selection. We inferred positive selection using model M2a of Yang and colleagues [34], taking recombination into account [32].

**Table 1: Summary of HIV-1 reference sequence data used**

Gene	Non-overlapping region used	Position on genome	Number of sequences
<i>env</i>	88 – 2154	6313 – 8379	37
<i>gag</i>	1 – 1295	790 – 2084	37
<i>nef</i>	1 – 621	8797 – 9417	32
<i>pol</i>	211 – 2955	2296 – 5040	37
<i>tat</i>	22 – 138	5851 – 5967	37
<i>vif</i>	58 – 519	5098 – 5559	37
<i>vpr</i>	61 – 273	5620 – 5832	37
<i>vpu</i>	1 – 162	6061 – 6222	37
Rev	total overlap		
genome	Includes overlapping sites	790 – 9417	36

The HXB2 genome numbering is used.

### Simulations

We used HyPhy to generate simulated data under a neutral model with trees generated from the original alignments (or the tree from the largest un-recombined region for alignments where recombination was detected). The same sequence alignments used as input in the initial analysis were used and one hundred simulated datasets were generated for each alignment. Each simulated dataset was then analyzed using the Dual Model as described above. For each gene the minimum value of mean dS across all sliding windows of three adjacent codons, in all of the one hundred simulated datasets, was used as a conservative threshold to identify windows of reduced dS in the observed data. This stringent threshold and a less stringent one that included 95% of the values inferred from the simulated data are shown in the sliding window plots.

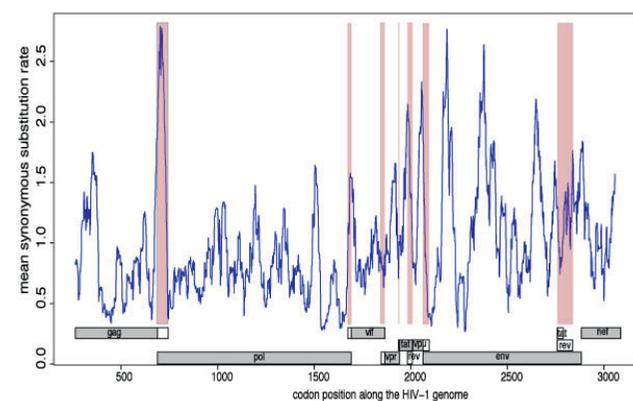
### Functional analysis of a novel nucleotide sequence motif in env

JC53-bl and 293T cells were obtained from Dr George Shaw (University of Alabama, Birmingham, AL) and cultured as described previously [35]. M7-Luc cells (5.25. EGFP.Luc.M7) were kindly provided by Dr Nathaniel Landau through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID. M7-Luc cells were cultured in RPMI-1640 medium containing 2 mM L-glutamine, 25 mM HEPES, 10% heat-inactivated fetal bovine serum (FBS) and 50 µg/ml gentamicin (Sigma), supplemented with 10 µg/ml DEAE-Dextran for infectivity assays. An infectious molecular clone, p81, was obtained through the AIDS Research and Reference Reagent Program, Division of AIDS, NIAID, NIH: p81A-4 (Cat#11440) from Dr. Bruce Chesebro. Mutations were introduced into p81 using the Stratagene QuickChange XL kit. Infectious viral particles were produced by transfection of 293T cells using Fugene reagent (Roche BioSciences), and transfection output was assessed by determination of TCID<sub>50</sub> in JC53-bl cells as described previously [35]. Equivalent numbers of TCID<sub>50</sub> were

used to infect M7-luc cells seeded into 96 well plates at a density of  $7.5 \times 10^5$  cells/ml followed by a washout step 12 hours post-infection. Replication was monitored by measurement of p24 production using the Vironostica HIV-1 antigen Microelisa system (Biomérieux) over 5 days.

### Results

Consistent with previous reports [5, 10], we found evidence of variation in synonymous substitution rates within and across HIV-1 genes (Figure 1). For all genes the Dual Model [10], which allows independent variation of dS and dN had a much better fit to the data than a model with constant dN and dS (referred to as the Constant model in Table 2) or than a model in which only dN varied across sites (the Nonsynonymous model in Table 2). The variance of dS gives an indication of the extent of site-to-site synonymous rate heterogeneity



**Figure 1**  
**HIV-1 genome-wide plot of mean nonsynonymous substitution rates.** A 30 codon sliding window was used. Regions coding for proteins in more than one frame are shaded in pink. The frames that were used in each region are shown in grey rectangles, with frame 1 at the top and frame 3 at the bottom.

**Table 2: AIC model selection index to show how different models fit to the data.**

Gene regions	Akaike Information Criterion index (AIC) per Model		
	Dual	Nonsynonymous	Constant
Gag	23862.95	23963.83	25889.10
Pol	41504.20	41668.99	45642.96
Vif	9794.65	9843.41	10502.37
Vpr	9581.57	9692.29	10869.78
Tat	2834.49	2878.99	3110.23
Vpu	11701.57	11832.41	12938.54
Env	45201.62	45422.72	48658.16
Nef	13984.66	14061.46	14986.75

The Dual Model which allows dS to vary across sites has the lowest AIC (best fit to the data) for all 8 genes.

within the different genomic regions (Table 3). There was significant variation between genes (p-value =  $2 \times 10^{-7}$ , from Levene's test) with the least site-to-site variation in dS observed in *vpu* and the most in *vpr* followed by *nef* and *env* (Figure 2).

We found evidence of strong purifying selection acting directly on the nucleotide sequence at twenty-three sites

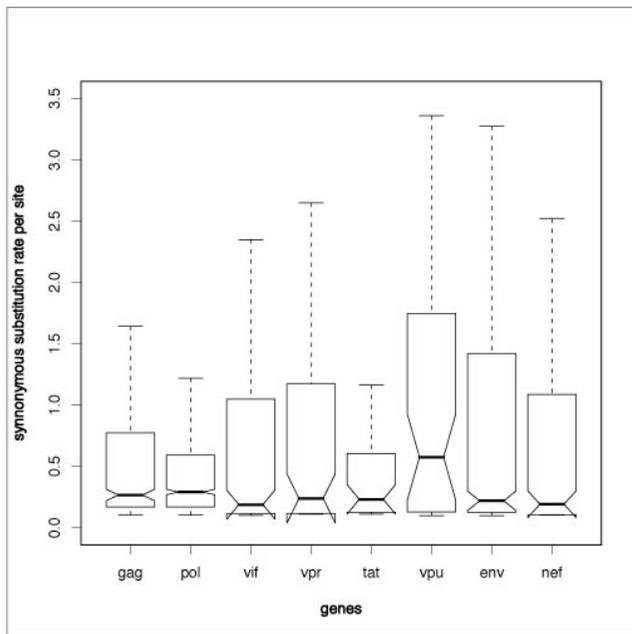
across the HIV-1 genome (Figures 3, 4, 5 and 6). Fourteen of these regions (marked in black in Figures 3 and 4) coincided exactly with well characterized functional motifs while for another five (marked in green in Figure 3), we were able to identify possible functions based on the known functions of the sequence domains in which they were situated. We could not, however, find plausible explanations for high degrees of sequence conservation observed within a twelve-nucleotide region of the *env* gene and three other regions in *vpu* (marked in red in Figure 4). Sequence logos illustrating the conservation in each of these twenty-three significantly conserved regions are shown in Figure 5 (for those with known specific function) and Figure 6 (for those with predicted and unknown functions).

The fourteen regions with known functions included one region consisting of fifteen nucleotides following the *gag* start codon, positions 793–807 of HXB2. This forms the fourth stem loop (sl4, Figure 3a) of the dimerization/encapsidation signal. The encapsidation signal is a four stem-loop structure which stretches from the 5' LTR and interacts with the nucleocapsid protein, promoting formation of genomic RNA and blocking the initiation of transcription [14, 36, 37].

**Table 3: Regions of the HIV-1 sequence which should be considered for exclusion in positive selection analysis studies.**

Gene	dS Variance	Overlapping regions	Highly conserved (most stringent cutoff)	Other sites conserved at 0.05 significance
<i>gag</i>	2.1	2086 – 2295 ( <i>gag/pol</i> )	793 – 807 898 – 903 985 – 996 1309 – 1314	821–823 1036–1038 1810–1812 1831–1833 1969 – 1974 2002–2004 2023 – 2025 2490–2495 2850 – 2858 3252 – 3257 3304 – 3312 3867 – 3881 3966 – 3971
<i>pol</i>	1.8	5041 – 5097 ( <i>pol/vif</i> )	4092 – 4094 4764 – 4790 4864 – 4866 4926 – 4937	3304 – 3312 3867 – 3881 3966 – 3971
<i>vif</i>	2.7	5560 – 5619 ( <i>vif/vpr</i> )	-	-
<i>vpr</i>	3.9	5833 – 5850 ( <i>vpr/tat</i> )	5769 – 5777 5794 – 5805	-
<i>tat</i>	2.9	5968 – 6060 ( <i>tat/rev</i> )	5855 – 5863 5957 – 5968	-
<i>vpu</i>	1.4	6223 – 6312 ( <i>vpu/env</i> )	6101 – 6106 6143 – 6151 6167 – 6178	-
<i>env</i>	3.2	8380 – 8796 ( <i>env/rev</i> )	7656 – 7667 7834 – 7842 8349 – 8354 8376 – 8378	7077 – 7082 7125 – 7130 7629 – 7634
<i>nef</i>	3.6	-	9067 – 9086 9087 – 9093 ( <i>nef/LTR</i> ) 9183 – 9192 ( <i>nef/LTR</i> ) 9391 – 9399 ( <i>nef/LTR</i> )	8869 – 8874 8887 – 8892 9121 – 9126 9235 – 9237

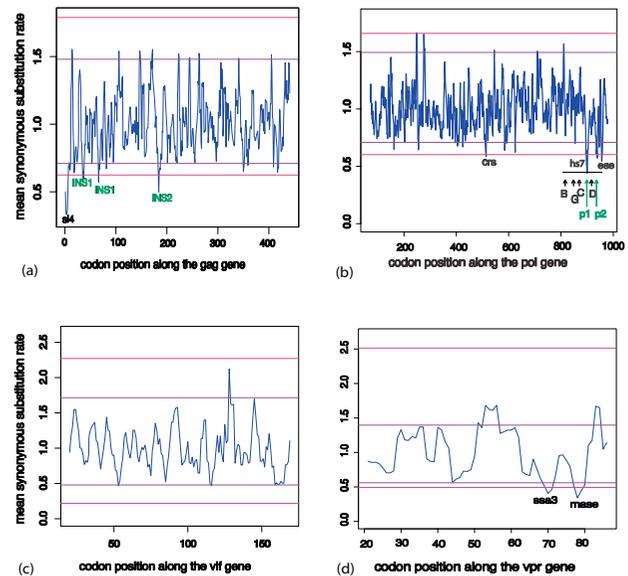
Co-ordinates are adapted from the HXB2 numbering system. Dashes indicate that there are no sites within that category for a particular gene.



**Figure 2**  
**Box-and-whisker plot showing variation of dS values per gene.** The values are from non-overlapping regions of HIV-1 genes.

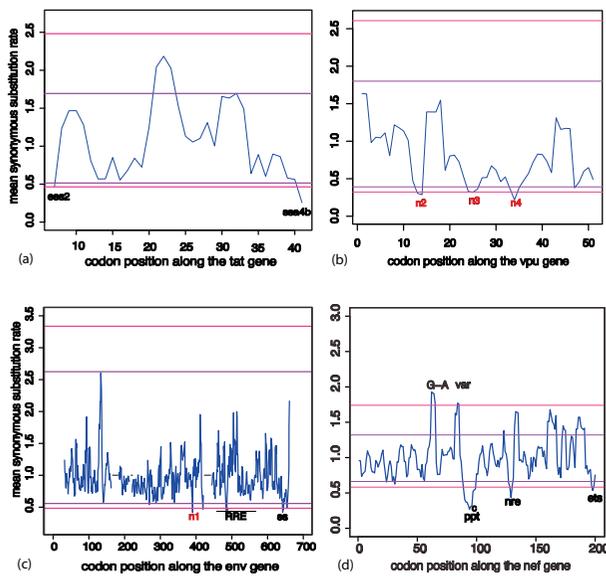
In the *pol* gene, we found two highly conserved regions with known functions. One corresponded to the first three nucleotides (HXB2 coordinate positions 4092–4094) of the 260 nucleotide long cis-repressive sequence (CRS; Figure 3b). The cis-repressive sequence inhibits expression of structural protein mRNAs by preventing their transportation from the nucleus – a process that is reversed by the rev-responsive element (RRE) [24-26]. The other was found at the 3' end of the intragenic nuclease hypersensitive domain (hs7 in Figure 3b). The latter motif, located between positions 4926 and 4937 and labeled "ese" in Figure 3b, is also located within HIV-1 exon 2 and the last six nucleotides, TGGAAA, of this conserved region form a known exonic splicing enhancer (ESE) of HIV mRNAs [38-42].

Two regions in *vpr*, a 3' splice acceptor site [41] and an RNase-V1 cleavage site [43, 44] were also conserved (labeled *ssa3* and *mase* respectively in Figure 3d) at positions 5759–5777 and 5794–5805 respectively. An additional two highly conserved regions were observed within the *tat* gene, one containing an exonic splicing silencer, ESS2 between nucleotide positions 5855 and 5863 and the other at the 3' splice acceptor site A4b located at positions 5957 to 5968 (*ess2* and *ssa4b* respectively in Figure 4a) [41].



**Figure 3**  
**Mean (blue) synonymous substitution rates observed across *gag*, *pol*, *vif* and *vpr* genes.** Mean dS was calculated over sliding windows of three codons. Horizontal lines mark the most stringent (red) and less stringent (purple) significance thresholds. (a) dS across the *gag* gene. 'sl4'; the fourth stem loop of the encapsidation signal, 'INS1'; a motif within the first inhibitory sequence region, 'INS2'; a motif within the second inhibitory sequence region. (b) dS across the *pol* gene. 'crs'; start of the cis-repressive sequence, horizontal dotted line is the nuclease hypersensitive region and sites 'B', 'G', 'C' and 'D' are confirmed transcription factor binding sites known as site-B, GC-box, site-C and site-D respectively. 'pl' and 'p2'; conserved sites within nuclease hypersensitive region. "ese"; exonic splicing enhancer. (c) dS across the *vif* gene. (d) dS across the *vpr* gene. 'ssa3'; 3' splice acceptor site A3, 'rnase'; RNase-V1 cleavage site.

Three of the highly conserved regions with known functions were in the *env* gene and included a nine-nucleotide long motif from position 7834 to 7842 within the RRE. The approximately two hundred nucleotides long RRE element within *env*, is known to interact with the Rev protein and facilitates the transport of late un-spliced and partially-spliced RNAs from the nucleus to the cytoplasm [19-21]. Although the RRE is associated with a long stretch of sequence that forms a well characterized secondary structure with various conserved domains [22], only the nine nucleotides that bind Rev with highest affinity [21, 45] were sufficiently conserved to be detected using our conservative threshold (Figure 4c). Also conserved within the *env* gene were the two splice site regions at the end of the *tat/rev* exon,



**Figure 4**  
**Mean (blue) synonymous substitution rates observed across *tat*, *vpu*, *env* and *nef* genes.** (a) dS across the *tat* gene. 'ess2'; exonic splicing silencer ESS2, 'ssa4b'; 3' splice acceptor site A4b. (b) dS across the *vpu* gene. 'n2', 'n3' and 'n4'; novel conserved sites. (c) dS across the *env* gene. "n1" is the novel conserved site. The black dotted horizontal lines indicate poorly aligned regions that were excluded from the analysis. "rre"; rev-responsive element, \*; the 9 nucleotides (5' GACGGUACA 3') which bind to the Rev protein with highest affinity, "ss"; splice site region for the *tat* and *rev* 3' exons. (d) dS across the *nef* gene. "G-A"; G-to-A hypermutations (see Additional file 1), 'var'; highly variable region, 'ppt'; poly-purine tract, "c"; PPT integrase attachment site, 'nre'; start of the negative repressive sequence, 'ets'; Ets-1 transcription factor binding site.

positions 8349–8354 and 8376–8378, usually referred to as 7a/7b, and 7 ("ss" in Figure 4c; [39, 41, 46, 47]

The last four of the significantly conserved regions with known function were in the *nef* gene and coincided with the poly-purine tract (PPT), integrase attachment site, negative regulatory element (NRE) and Ets-1 transcription factor binding site (Figure 4d) (with HXB2 coordinates 9066–9083, 9084–9091, 9183–9192 and 9391–9399 respectively). The PPT precedes the start of the LTR and is known to associate with the 3' LTR, serving as a primer for the initiation of HIV-1 plus strand DNA replication [12, 48-50]. A previous detailed RT RNase-H binding analysis revealed that priming of the plus strand occurs specifically at the 3' end of the PPT, at the "GGGGG" motif [51-53]. The region adjacent to

Gene	HXB2 coordinates	Sequence Logo and degeneracy	Function
<i>gag</i> Figure 3a 'sl4'	793 - 807		Fourth stem loop of encapsidation signal
<i>pol</i> Figure 3b 'crs'	4092 - 4094		Cis-repressive sequence start site
Figure 3b 'ese'	4926 - 4937		"TGGAAA" is an exonic splicing enhancer
<i>vpr</i> Figure 3d 'ssa3'	5759 - 5777		3' splice acceptor site A3
Figure 3d 'rnase'	5794 - 5805		RNase-V1 cleavage site
<i>tat</i> Figure 4a 'ess2'	5855 - 5863		Exonic Splicing Silencer 2
<i>env</i> Figure 4c '*'	7834 - 7842		Rev binding loop in rev-responsive element
Figure 4c 'ss'	8349 - 8354		Tat/rev 3' exons splice sites 7a, 7b
Figure 4c 'ss'	8376 - 8378		Tat/rev 3' exons splice sites 7
<i>nef</i> Figure 4d 'ppt'	9066 - 9076		PPT (polypurine tract)
Figure 4d 'ppt'	9077 - 9086		PPT, Priming of transcription
Figure 4d 'c'	9084 - 9091		PPT cleavage region
Figure 4d 'nre'	9183 - 9192		3'LTR negative responsive element start sites
Figure 4d 'ets'	9391 - 9399		TF binding region for Ets-1 proteins

**Figure 5**  
**Sequence motifs for the highly conserved regions with known function.** The fourteen regions with known specific functions found to be under strong purifying selection in HIV-1 genes. The range of coordinates on the HIV-1 genome for each motif is given in column 2. Numbers above each logo represent the degeneracy at each nucleotide site.

the 3' end of the PPT was also highly conserved. This region corresponds to the start of the 3'LTR and contains the cleavage site of the PPT by RNase H as well as the start of integrase attachment region for the integration of the viral genome into the genome of the host [48, 53-55]. Two codons in the central region of *nef* were highly variable, one dominated by G-to-A mutations in a sequence context consistent with APOBEC-induced hypermutations when compared to the Group M ancestral sequence [56] (see Additional file 1) and the second had a high rate of synonymous and nonsynonymous substitutions (labeled "G-A" and "var" in Figure 4d respectively).

Gene	HXB2 coordinates	Sequence Logo and degeneracy	Function
Regions with predicted functions			
<i>gag</i> Figure 3a 'INS1'	898 - 903		Within (INS1) first Inhibitory sequence region
Figure3a 'INS1'	985 - 996		Within INS1
Figure 3a 'INS2'	1309 - 1314		Within INS2
<i>pol</i> Figure 3b 'p1'	4764 - 4776		Within nuclease-hypersensitive region
Figure 3b: 'p1'	4777 - 4790		potential binding motif for Oct-1
Figure 3b 'p2'	4864 - 4866		In nuclease-hypersensitive region
Novel regions			
<i>vpu</i> Figure 4b 'n2'	6101 - 6106		novel conserved region
Figure 34b 'n3'	6143 - 6151		novel conserved region
Figure 3b 'n4'	6167 - 6178		novel conserved region
<i>env</i> Figure 4c 'n1'	7656 - 7667		novel region, analyzed in fitness assays

**Figure 6**  
**Sequence motifs for the highly conserved regions with unknown specific function.** The five regions with predicted functions and four regions with unknown functions found to be under strong purifying selection in HIV-1 genes. The range of coordinates on the HIV-1 genome for each motif is given in column 2. Numbers above each logo represent the degeneracy at each nucleotide site.

One of the novel regions, the twelve nucleotide long motif in *env*, upstream of the RRE showed the highest degree of conservation of any region in the *env* gene ('n1' in Figure 4c; positions 7656–7667 of HXB2). Introduction of synonymous point mutations at positions 7, 9 and 12 in this nucleotide motif had no effect on virus output from transfected 293T cells (see Additional file 2 (a)). No significant differences were observed between the wild type and mutated viruses with respect to their infectivity in M7-Luc cells (see Additional file 2(b)). Although the effect of mutations within this region is therefore not clear at present, future work making use of more sensitive competitive replication assays will determine whether these changes have an impact on viral fitness. Functional analyses of the three novel conserved regions in *vpu*, with HXB2 coordinates 6101–6106,

6143–6151 and 6167–6178 (n2, n3, n4 in Figure 4b) are being considered. The protein products of *vpu* and *env* are known to be produced from a bicistronic transcript [46] and the conserved regions in *vpu* may be involved in the control of translation.

In order to assess whether purifying selection is likely to cause false inference of positive selection we used standard methods to detect positive selection in a subtype C *gag* coding sequence alignment described previously [33]. Sites with  $\omega$  significantly greater than one, implying positive selection, overlapped significantly with sites that had lower than average dS values (Fisher's exact test odds ratio = 4.4; p-value = 0.006; Additional file 3). This is consistent with a substantial proportion of the positive selection signals resulting from conservation of the synonymous sites rather than diversifying selection acting on the nonsynonymous sites.

### Discussion

This is the first study to provide a detailed analysis of site-to-site variation in the rate of synonymous substitutions across the HIV-1 genome. In the past, site-to-site variation in dS in HIV-1 has been investigated in a single gene [5, 10] and in another study a single overall synonymous substitution rate for the entire genome was determined for comparison to other viral lineages [57]. We modeled site-to-site synonymous rate variation using a similar approach to a previous study [10], in that case only one HIV-1 gene, *vif*, was considered and sites that encode proteins in multiple reading frames were included. As a consequence, it was not clear whether the observed site-to-site rate variation resulted from variation in the synonymous rate or from selection acting on nonsynonymous substitutions in another reading frame. Here we focused primarily on regions of the HIV genome that encode proteins in a single reading frame and explored functions of nucleotide sequences that have the largest influence on synonymous substitution rate variation.

Previous studies have demonstrated that recombination causes false inference of positive selection. Since recombination affects tree topologies used in fitting phylogenetic models, it is also likely to cause biased estimates of dS. The recent development of methods to account for recombination in selection analyses [32] permitted us to remove recombination as a source of bias in our estimates of synonymous substitution rates.

In addition to the fourteen conserved regions with known functions and the novel sites in *env* and *vpu*, five conserved sites without previously reported specific functions occurred within known functional domains.

These include three short (3–6 bp) motifs in the inhibitory (INS) sequence regions of gag (HXB2 positions 898–903, 985–996 and 1309–1314, Figure 3a). Previous *in-vitro* analyses have shown that short motifs within the approximately two hundred bp long INS regions, are responsible for the actual inhibition of mRNA expression [23, 24, 26]. Although the three motifs we find within this region have not been specifically identified *in-vitro*, a computational study by Wolff et al (2003) showed that the INS sequences have several short functional motifs within them [26]. The conserved sites we identified within INS1 and INS2 could serve the same inhibitory function. Functional assays elucidating the role of these sites in the inhibition of mRNA expression could help to determine the precise mechanisms by which inhibition occurs and whether these sites also play a role. In another previous study which analyzed the RNA secondary structure of the 5' region of HIV-1, these two regions, labeled 'INS1' in Figure 3a, were found to be involved in conserved Watson-Crick base-pairing [58]

The last two of the five conserved regions with unidentified specific functions were within the *pol* HS7. HS7 spans five hundred nucleotides, between positions 4481 and 4982 of the HIV-1 genome, and has an LTR-like regulatory function [16, 17]. Previous studies revealed four domains towards the 3' end of this region (PU box, GC-box, site-C and site-D) that bind to specific transcription factors (TFs) and are also important for viral infectivity [16, 17]. In these studies, the Oct-1, Oct-2, PU.1, Sp1 and Sp3 transcription factors were found to bind to at least one of the four identified sites. The two conserved regions we identified are outside these specific identified functional domains, but one of them at positions 4767–4790 (labeled "p1" in Figure 3b) showed potential binding to Oct-1 using the MATCH tool from the TRANSFAC database [59, 60]. Potential association with a transcription factor was not observed for the three nucleotides (positions 4864–4866) labeled 'p2' and its adjacent sites.

Knowledge of regions of the genome that function at the nucleotide level is important for positive selection analysis. Conserved synonymous sites can cause false detection of positive selection and need to be either excluded from analyses or modeled appropriately. The danger is that some sites may be assumed to be evolving adaptively simply as a result of the purifying selection acting directly on the nucleotide sequence. We found evidence of this in subtype C *gag* sequences where positively selected sites significantly coincided with significantly low dS (Additional file 3). In addition, a study by Hurst & Pal (2001) also showed false detection of positive selection caused by purifying selection pressure acting on synonymous sites.

In many selection, studies the synonymous rate is assumed to be constant. However, negative selection acting on synonymous sites can potentially reduce gene-wide estimates of the synonymous rates below the neutral evolution rate. Comparison of site-specific nonsynonymous substitution rates against this underestimate of the neutral rate is likely to cause a proportion of the selectively neutral nonsynonymous sites to seem as though they are evolving adaptively. We have, however, used a very stringent cutoff to identify the twenty three regions within the HIV genome that have obviously reduced synonymous substitution rates. For a more conservative analysis of selection, all the significantly conserved sites, i.e., including those conserved at 95% confidence (p-value < 0.05, listed in Table 3) should be excluded from analyses along with sites that encode proteins in multiple frames. Surprisingly, we found that the rate of synonymous substitution, was higher, on average, in overlapping gene regions that encode proteins in more than one frame ( $p = 6 \times 10^{-7}$  from Wilcoxon rank sum test; Figure 1); however, lower dS was observed within some genome regions that are translated in multiple reading frames. Analysis of the most diverse sequences within subtype B and C revealed more highly conserved sites across the RRE and INS1 regions at the subtype level (Additional file 4). These putatively functional domains could also be removed in more conservative studies of selection acting on HIV protein sequences.

## Conclusion

We have analyzed sequence variation at the synonymous sites of non-overlapping regions of HIV-1 genes. We found substantial site-to-site variation in the rate of synonymous substitution with evidence of purifying selection pressure within functional domains such as the Rev-responsive element. The majority of conserved sites we identified are within functional regions that are well documented in the literature; however, the total number of sites that function at the nucleotide level is unknown and it is therefore difficult to assess the fraction of the known and novel functional sites that are detectable using our method. In addition to identifying putatively functional sites under purifying selection, these results contribute to the robustness of analyses of positive selection by identifying conserved synonymous sites that can cause false positive inference of selection. The sites presented in Table 3 and Figures 5 and 6 thus form a resource for future studies of selection pressures acting on HIV-1 genes.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

NKN carried out the analysis, interpreted the results and drafted the manuscript. KS provided HyPhy scripts, contributed to the methodology and participated in drafting the manuscript. PM carried out the biochemical fitness assays and participated in writing the manuscript. ZW and DM participated in validating the results and editing the manuscript. CS conceived and supervised the study and participated in writing the manuscript.

## Additional material

### Additional file 1

G-A mutations in a variable region in the nef gene. Mutations observed in reference sequences in comparison to Group M ancestral sequence identified using the hypermut tool available in the Los Alamos database. The highly variable region (labeled "G-A" in Figure 3d) showed G-A mutations and is boxed in red.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1743-422X-5-160-S1.pdf>]

### Additional file 2

Functional analysis of a novel region in the env gene. (a) Production of p24 from transfected wildtype p81 and 3 mutants produced from synonymous mutations introduced in the previously uncharacterized conserved region in env. (b) Comparison of infectivity between wildtype and the mutants.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1743-422X-5-160-S2.pdf>]

### Additional file 3

Evidence of overlap between high omega at a codon and low dS at the synonymous sites. Positively selected sites at which a significantly low dS was observed at the synonymous sites. Positively selected sites are shown in blue vertical lines and sites with low dS are shaded in light blue.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1743-422X-5-160-S3.pdf>]

### Additional file 4

Highly conserved regions observed at the subtype-level. dS across subtypes B and C gag and env genes showing more conserved sites at the subtype sequence level within the INS regions in gag and RRE in env.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1743-422X-5-160-S4.pdf>]

## Acknowledgements

This study was funded by the South African National Bioinformatics Network. NKN was supported by a training grant under the Stanford-South Africa Biomedical Informatics Training Program which is supported by the Fogarty International Center, part of the National Institutes of Health (grant no. 5D43 TW006993).

## References

- Muse SV and Gaut BS: **A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution**

- rates, with application to the chloroplast genome. *Mol Biol Evol* 1994, **11**:715-724.
- Goldman N and Yang Z: **A codon-based model of nucleotide substitution for protein-coding DNA sequences.** *Mol Biol Evol* 1994, **11**:725-736.
- Nielsen R and Yang Z: **Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene.** *Genetics* 1998, **148**:929-936.
- Yang Z, Nielsen R, Goldman N and Pedersen AM: **Codon-substitution models for heterogeneous selection pressure at amino acid sites.** *Genetics* 2000, **155**:431-449.
- Lemey P, Kosakovsky Pond SL, Drummond AJ, Pybus OG, Shapiro B, Barroso H, Taveira N and Rambaut A: **Synonymous substitution rates predict HIV disease progression as a result of underlying replication dynamics.** *PLoS Comput Biol* 2007, **3**:e29.
- Seoighe C, Ketwaroo F, Pillay V, Scheffler K, Wood N, Duffet R, Zvelebil M, Martinson N, McIntyre J, Morris L and Hide W: **A model of directional selection applied to the evolution of drug resistance in HIV-1.** *Mol Biol Evol* 2007, **24**:1025-1031.
- Choisy M, Woelk CH, Guegan JF and Robertson DL: **Comparative study of adaptive molecular evolution in different human immunodeficiency virus groups and subtypes.** *J Virol* 2004, **78**:1962-1970.
- de Oliveira T, Salemi M, Gordon M, Vandamme AM, van Rensburg EJ, Engelbrecht S, Coovadia HM and Cassol S: **Mapping sites of positive selection and amino acid diversification in the HIV genome: an alternative approach to vaccine design?** *Genetics* 2004, **167**:1047-1058.
- Zanotto PM, Kallas EG, de Souza RF and Holmes EC: **Genealogical evidence for positive selection in the nef gene of HIV-1.** *Genetics* 1999, **153**:1077-1089.
- Kosakovsky Pond S and Muse SV: **Site-to-site variation of synonymous substitution rates.** *Mol Biol Evol* 2005, **22**:2375-2385.
- Hurst LD and Pal C: **Evidence for purifying selection acting on silent sites in BRCA1.** *TRENDS in Genetics* 2001, **17**:62-65.
- Quinones-Mateu ME, Mas A, Lain dL, Soriano V, Alcami J, Lederman MM and Domingo E: **LTR and tat variability of HIV-1 isolates from patients with divergent rates of disease progression.** *Virus Res* 1998, **57**:11-20.
- Das AT, Klaver B and Berkhout B: **The 5' and 3' TAR elements of human immunodeficiency virus exert effects at several points in the virus life cycle.** *J Virol* 1998, **72**:9217-9223.
- Wilkinson KA, Gorelick RJ, Vasa SM, Guex N, Rein A, Mathews DH, Giddings MC and Weeks KM: **High-throughput SHAPE analysis reveals structures in HIV-1 genomic RNA strongly conserved across distinct biological states.** *PLoS Biol* 2008, **6**:e96.
- Pereira LA, Bentley K, Peeters A, Churchill MJ and Deacon NJ: **A compilation of cellular transcription factor interactions with the HIV-1 LTR promoter.** *Nucleic Acids Res* 2000, **28**:663-668.
- Goffin V, Demonte D, Vanhulle C, de Walque S, de Launoit Y, Burny A, Collette Y and Van Lint C: **Transcription factor binding sites in the pol gene intragenic regulatory region of HIV-1 are important for virus infectivity.** *Nucleic Acids Res* 2005, **33**:4285-4310.
- Van Lint C, Ghysdael J, Paras P Jr, Burny A and Verdin E: **A transcriptional regulatory element is associated with a nuclease-hypersensitive site in the pol gene of human immunodeficiency virus type 1.** *J Virol* 1994, **68**:2632-2648.
- Verdin E, Becker N, Bex F, Droogmans L and Burny A: **Identification and characterization of an enhancer in the coding region of the genome of human immunodeficiency virus type 1.** *Proc Natl Acad Sci* 1990, **87**:4874-4878.
- Hadzopoulou-Cladaras M, Felber BK, Cladaras C, Athanassopoulos A, Tse A and Pavlakis GN: **The rev (trs/art) protein of human immunodeficiency virus type 1 affects viral mRNA and protein expression via a cis-acting sequence in the env region.** *J Virol* 1989, **63**:1265-1274.
- Renwick SB, Critchley AD, Adams CJ, Kelly SM, Price NC and Stockley PG: **Probing the details of the HIV-1 Rev-Rev-responsive element interaction: effects of modified nucleotides on protein affinity and conformational changes during complex formation.** *Biochem J* 1995, **308**(Pt 2):447-453.
- Peterson RD and Feigon J: **Structural change in Rev responsive element RNA of HIV-1 on binding Rev peptide.** *J Mol Biol* 1996, **264**:863-877.
- Phuphuakrat A and Auewarakul P: **Heterogeneity of HIV-1 Rev response element.** *AIDS Res Hum Retroviruses* 2003, **19**:569-574.
- Schwartz S, Campbell M, Nasioulas G, Harrison J, Felber BK and Pavlakis GN: **Mutational inactivation of an inhibitory sequence**

- in human immunodeficiency virus type I results in Rev-independent gag expression. *J Virol* 1992, **66**:7176–7182.
24. Schneider R, Campbell M, Nasioulas G, Felber BK and Pavlakis GN: **Inactivation of the human immunodeficiency virus type I inhibitory elements allows Rev-independent expression of Gag and Gag/protease and particle formation.** *J Virol* 1997, **71**:4892–4903.
  25. Cochrane AW, Jones KS, Beidas S, Dillon PJ, Skalka AM and Rosen CA: **Identification and characterization of intragenic sequences which repress human immunodeficiency virus structural gene expression.** *J Virol* 1991, **65**:5305–5313.
  26. Wolff H, Brack-Werner R, Neumann M, Werner T and Schneider R: **Integrated functional and bioinformatics approach for the identification and experimental verification of RNA signals: application to HIV-1 INS.** *Nucleic Acids Research* 2003, **31**:2839–2851.
  27. Anisimova M, Nielsen R and Yang Z: **Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites.** *Genetics* 2003, **164**:1229–1236.
  28. Shriner D, Nickle DC, Jensen MA and Mullins JI: **Potential impact of recombination on sitewise approaches for detecting positive natural selection.** *Genet Res* 2003, **81**:115–121.
  29. Leitner T, Foley B, Hahn B, Marx P, McCutchan F, Mellors J, Wolinsky S and Korber B: **HIV Sequence Compendium. Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, NM, LA-UR 2005, 06–0680.**
  30. Kosakovsky Pond SL, Frost SD and Muse SV: **HyPhy: hypothesis testing using phylogenies.** *Bioinformatics* 2005, **21**:676–679.
  31. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH and Frost S: **Automated Phylogenetic Detection of Recombination Using a Genetic Algorithm.** *Mol Biol Evol* 2006, **23**:1891–1901.
  32. Scheffler K, Martin DP and Seoighe C: **Robust inference of positive selection from recombining coding sequences.** *Bioinformatics* 2006, **22**:2493–2499.
  33. Ngandu NG, Bredell H, Gray CM, Williamson C and Seoighe C: **CTL response to HIV type I subtype C is poorly predicted by known epitope motifs.** *AIDS Res Hum Retroviruses* 2007, **23**:1033–1041.
  34. Yang Z, Wong WS and Nielsen R: **Bayes empirical bayes inference of amino acid sites under positive selection.** *Mol Biol Evol* 2005, **22**:1107–1118.
  35. Moore PL, Gray ES, Choge IA, Ranchobe N, Mlisana K, Abdool Karim SS, Williamson C and Morris L: **The c3-v4 region is a major target of autologous neutralizing antibodies in human immunodeficiency virus type I subtype C infection.** *J Virol* 2008, **82**:1860–1869.
  36. Clever J, Sassetti C and Parslow TG: **RNA secondary structure and binding sites for gag gene products in the 5' packaging signal of human immunodeficiency virus type I.** *J Virol* 1995, **69**:2101–2109.
  37. Huthoff H, Das AT, Vink M, Klaver B, Zorgdrager F, Cornelissen M and Berkhout B: **A human immunodeficiency virus type I-infected individual with low viral load harbors a virus variant that exhibits an in vitro RNA dimerization defect.** *J Virol* 2004, **78**:4907–4913.
  38. Exline CM, Feng Z and Stoltzfus CM: **Negative and positive mRNA splicing elements act competitively to regulate human immunodeficiency virus type I vif gene expression.** *J Virol* 2008, **82**:3921–3931.
  39. Schwartz S, Felber BK, Benko DM, Fenyo EM and Pavlakis GN: **Cloning and functional analysis of multiply spliced mRNA species of human immunodeficiency virus type I.** *J Virol* 1990, **64**:2519–2529.
  40. Madsen JM and Stoltzfus CM: **A suboptimal 5' splice site downstream of HIV-1 splice site A1 is required for unspliced viral mRNA accumulation and efficient virus replication.** *Retrovirology* 2006, **3**:10.
  41. Kammler S, Otte M, Hauber I, Kjems J, Hauber J and Schaal H: **The strength of the HIV-1 3' splice sites affects Rev function.** *Retrovirology* 2006, **3**:89.
  42. Krummheuer J, Lenz C, Kammler S, Scheid A and Schaal H: **Influence of the small leader exons 2 and 3 on human immunodeficiency virus type I gene expression.** *Virology* 2001, **286**:276–289.
  43. Jacquenet S, Mereau A, Bilodeau PS, Damier L, Stoltzfus CM and Branlant C: **A second exon splicing silencer within human immunodeficiency virus type I tat exon 2 represses splicing of Tat mRNA and binds protein hnRNP H.** *J Biol Chem* 2001, **276**:40464–40475.
  44. Jacquenet S, Ropers D, Bilodeau PS, Damier L, Mougin A, Stoltzfus CM and Branlant C: **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* 2001, **29**:464–478.
  45. Hung LW, Holbrook EL and Holbrook SR: **The crystal structure of the Rev binding element of HIV-1 reveals novel base pairing and conformational variability.** *Proc Natl Acad Sci USA* 2000, **97**:5107–5112.
  46. Schwartz S, Felber BK, Fenyo EM and Pavlakis GN: **Env and Vpu proteins of human immunodeficiency virus type I are produced from multiple bicistronic mRNAs.** *J Virol* 1990, **64**:5448–5456.
  47. Swanson AK and Stoltzfus CM: **Overlapping cis sites used for splicing of HIV-1 env/nef and rev mRNAs.** *J Biol Chem* 1998, **273**:34551–34557.
  48. Miles LR, Agresta BE, Khan MB, Tang S, Levin JG and Powell MD: **Effect of polypurine tract (PPT) mutations on human immunodeficiency virus type I replication: a virus with a completely randomized PPT retains low infectivity.** *J Virol* 2005, **79**:6859–6867.
  49. Luo GX, Sharmeen L and Taylor J: **Specificities involved in the initiation of retroviral plus-strand DNA.** *J Virol* 1990, **64**:592–597.
  50. Rausch JW and Le Grice SF: **Purine analog substitution of the HIV-1 polypurine tract primer defines regions controlling initiation of plus-strand DNA synthesis.** *Nucleic Acids Res* 2007, **35**:256–268.
  51. Powell MD and Levin JG: **Sequence and structural determinants required for priming of plus-strand DNA synthesis by the human immunodeficiency virus type I polypurine tract.** *J Virol* 1996, **70**:5288–5296.
  52. Pullen KA, Rattray AJ and Champoux JJ: **The sequence features important for plus strand priming by human immunodeficiency virus type I reverse transcriptase.** *J Biol Chem* 1993, **268**:6221–6227.
  53. Rausch JW and Le Grice SF: **'Binding, bending and bonding': polypurine tract-primed initiation of plus-strand DNA synthesis in human immunodeficiency virus.** *Int J Biochem Cell Biol* 2004, **36**:1752–1766.
  54. Masuda T, Kuroda MJ and Harada S: **Specific and independent recognition of U3 and U5 att sites by human immunodeficiency virus type I integrase in vivo.** *J Virol* 1998, **72**:8396–8402.
  55. Brown HE, Chen H and Engelman A: **Structure-based mutagenesis of the human immunodeficiency virus type I DNA attachment site: effects on integration and cDNA synthesis.** *J Virol* 1999, **73**:9011–9020.
  56. Rose PP and Korber BT: **Detecting hypermutations in viral sequences with an emphasis on G → A hypermutation.** *Bioinformatics* 2000, **16**:400–401.
  57. Hanada K, Suzuki Y and Gojobori T: **A large variation in the rates of synonymous substitution for RNA viruses and its relationship to a diversity of viral infection and transmission modes.** *Mol Biol Evol* 2004, **21**:1074–1080.
  58. Paillart JC, Skripkin E, Ehresmann B, Ehresmann C and Marquet R: **In vitro evidence for a long range pseudoknot in the 5'-untranslated and matrix coding regions of HIV-1 genomic RNA.** *J Biol Chem* 2002, **277**:5995–6004.
  59. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV and Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31**:3576–3579.
  60. Matsy V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S and Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31**:374–378.