Journal of Bioinformatics and Computational Biology
Vol. 8, No. 3 (2010) 607–625

DOI: 10.1142/S0219720010004677

Imperial College Press
www.icpress.co.uk

SNPS IN THE HIV-1 TATA BOX AND THE AIDS PANDEMIC

VALENTIN V. SUSLOV

Sector of Evolutionary Bioinformatics, Institute of Cytology and Genetics
10 Lavrentyev Ave., Novosibirsk, 630090, Russia
valya@bionet.nsc.ru

PETR M. PONOMARENKO*, VADIM M. EFIMOV, LUDMILA K. SAVINKOVA, MIKHAIL P. PONOMARENKO and NIKOLAY A. KOLCHANOV

Laboratory of Theoretical Genetics, Institute of Cytology and Genetics
10 Lavrentyev Ave., Novosibirsk, 630090, Russia
**pon@bionet.nsc.ru*

Received 17 October 2009
Revised 7 January 2010
Accepted 16 January 2010

Evolutionary trends have been examined in 146 HIV-1 forms (2662 copies, 2311 isolates) polymorphic for the TATA box using the "DNA sequence→affinity for TBP" regression (TBP is the TATA binding protein). As a result, a statistically significant excess of low-affinity TATA box HIV-1 variants corresponding to a low level of both basal and TAT-dependent expression and, consequently, slow replication of HIV-1 have been detected. A detailed analysis revealed that the excess of slowly replicating HIV-1 is associated with the subtype E-associated TATA box core sequence "CATAAAA". Principal Component Analysis performed on 2662 HIV-1 TATA box copies in 70 countries revealed the presence of two principal components, PC1 (75.7% of the variance) and PC2 (23.3% of the variance). They indicate that each of these countries is specifically associated with one of the following trends in HIV-1 evolution: neutral drift around the normal TATA box; neutral drift around the slowly replicating TATA box core sequence (phylogenetic inertia); an adaptive increase in the frequency of the slowly replicating form.

Keywords: TATA box; TATA binding protein (TBP); TBP/TATA-affinity; prediction; HIV-1; SNP.

1. Introduction

The evolution of HIV-1 has been intensively studied ever since this virus was identified as one of the main causative agents of AIDS. The phylogeny of HIV-1, its relationships with other lentiviruses and the waves of the AIDS pandemic have

*Novosibirsk State University, 2 Pirogova St., Novosibirsk, Russia.

been reconstructed.[1–4] The most likely ancestor of HIV-1 is simian immunodeficiency virus of chimpanzee (SIVcpz), from which independently came dominant group M and minor group N. The involvement of simian immunodeficiency virus of gorilla (SIVgor, minor group) is possible too.[5] A retrospective study of the AIDS history and molecular phylogeny reconstruction[1] demonstrated that the virus (HIV-1 group M) was first transmitted to man in the first decade of the 20th century in Congo and/or Cameroon.[6,7] The initial pesthole of the AIDS epidemic appeared in part due to enhanced migration which followed the division of the German colonies and later the break-up of the colonial system. This break-up was followed by re-emigration of the Haitians from Congo[3] and the emergence of a persistent pesthole in Haiti in the 1960s, whence the pandemic reached the USA and Western European countries.[8] Later, HIV-1 was disseminated both from persistent pestholes and directly from Africa. Repeated import of HIV-1 gave rise to a lot of recombinant forms.[9–11]

HIV-1 is one of the most variable viruses. This is because revertase has a high error rate (∼1 substitution per genome replication), the virus has high reproductive rates (∼10^{10} virions a day per patient), a complex life cycle including a DNA and an RNA stage, and AIDS has a prolonged symptom-free period, which all together promote co-infection and recombination. Group M, which is dominant in the pandemic, includes nine subtypes (subtypes A and F have subsubtypes), dozens of recombinant forms, both circulating (CRFs are the forms isolated from two or more patients) and unique (URF). This phylogeny was inferred from an analysis of protein-coding RNA (first of all, the *env* gene, which encodes the protein gp120; later this phylogeny was improved using the *pol*, *gag*, *vif* genes and others).[5,12,13]

The diversity of non-coding regulatory RNA is given much less consideration. At the DNA stage, HIV-1 transcription is initiated by the only TATA box, which has an imperfect copy in the central part of each 5′- and 4′-half of the long terminal repeats (LTRs).[14,15] Upstream of the TATA box are NF-κB[16] and Sp1[17] binding sites, which modulate transcription; the TATA box is immediately flanked by E boxes,[15] which bind the transcription factor AP-4.[18] Downstream of the transcription start site is the RNA-enhancer TAR (transactivation response RNA). To enable HIV-1 expression, it is required that the TATA-binding protein (TBP) be bound to the TATA box, and this binding initiates the assembly of the preinitiation complex (PIC).[17,19] The viral protein Tat activates this process in at least three ways. First, by binding to the protein complex P-TEFb, which forms part of the PIC and eventually[20,21] allows TBP to carry on without TAFs (TBP-associated factors, which mediate the activation of expression by Sp1.[22,23] Secondly, by enhancing PIC elongation, which it does by binding to the viral RNA of the TAR.[19] Finally, by changing the density of DNA packaging into nucleosomes.[24,25]

In this paper, we used the TATA box as a marker of global evolutionary trends. Not associated with any particular function of HIV-1, but affecting any and all of them through transcription, with a readily recognizable sequence, the TATA box

fits in well with the criterion of being a diagnostic character[a]; however, because their small sizes increase the risk of sequence convergence, molecular phylogeny methods are of little help. We examined 146 HIV-1 forms (2662 copies, 2311 isolates) polymorphic for the TATA box using the "DNA sequence→affinity for TBP" regression[27] and found a statistically significant excess of variants with low affinity for human TBP.

2. Materials and Methods

2.1. *Sampling HIV-1*

HIV-1 sequences were retrieved from Genbank and aligned using Mauve 2.0, a free multiple genome alignment system.[28] 2662 sequence of 2311 HIV-1 isolates aligned with respect to the promoter region gave 146 forms polymorphic for the TATA box (Table 1). The "agatgctgCATATAAgcagctgcttt" variant found in 1567 of 2662 copies (59%), which is five times as frequent as the second-most frequently occurring variant[b], was taken as a normal HIV-1 TATA box sequence, S^0.

2.2. *In silico analysis of TBP/TATA affinity*

The TATA box and their flanking sequences were examined. The total length of the sequences was set at 26 base pairs (bp) based on our previous works demonstrating a good agreement between the *in silico* predicted and experimentally measured affinity of TBP for the TATA box.[27,30,31] For each TATA box sequence $S = \{s_{i-12} \ldots s_{i-1} s_i s_{i+1} \ldots s_{i+13}\}$, the $-\ln[K_{D,TATA}(S)]$ estimate is the affinity for human TBP in natural logarithmic units (ln-unit). This affinity was measured using a precise equilibrium equation for TBP/TATA binding in four subsequent steps[27]: (1) binding of the TBP to non-specific DNA[32]; (2) sliding along DNA[33]; (3) recognition of the TATA box[34] and the assembly of the TBP/TATA complex[35]; (4) stabilization of the TBP/TATA complex by endothermic rearrangement[36] which includes DNA bending[35]:

$$-\mathrm{Ln}[K_{D,TATA}(S)] = 10.90 + 0.15 \times \mathrm{PWM}_{TATA}(S) - 0.20 \times \mathrm{Ln}[K_{D,ssDNA}](S) - 0.23 \times \mathrm{Ln}[K_{D,dsDNA}](S); \quad (1)$$

where 10.90 is the non-specific affinity of TBP for DNA ($\approx 10^{-5}$ mol after Hahn's experiment[32]); $\mathrm{PWM}_{TATA}(S)$ is the maximum of Bucher's criterion[34] of the TATA box with the weight matrix $f_X(i)$ built for the sliding window $(\xi_{-1}, \ldots \xi_{+13})$ of 15 bp

[a] The widely used examples of a diagnostic character in molecular phylogeny are the genes for transferring 18S rRNA.[26]

[b] agatgctgCATAAAAgcagccgcttt: the core "CATAAAA" is associated with HIV-1 subtype E and its CRFs.[29]

Table 1. 146 polymorphic HIV-1 TATA box variants their predicted affinity for TBP.

n	DNA sequence, S	Δ#	n	DNA sequence, S	Δ#
1567	agatgctgCATATAAgcagctgcttt	19.52#	**3**	agatgctgCATATAAgcagc**a**gcttt	−0.03
294	agatgctgCATA**A**AAgcagc**c**gcttt	−0.51*	**2**	agatgctgCATA**C**AAgcagctgcttt	−1.11*
152	agatgctgCATATAAgcagc**c**gcttt	+0.04	**2**	agatgctgCATAT**G**Agcagctgcttt	−1.10*
55	agatgctgCATA**A**AAgcagctgcttt	−0.55*	**2**	agatgctgCATA**A**AAgcagctgctt**c**	−0.55*
51	agatgctgCATATAAgcagctgctt**c**	0	**2**	a**a**atgctgCATA**A**AAgcagc**c**gctt**c**	−0.51*
37	agatgct**a**CATATAAgcagctgcttt	+0.54*	**2**	ag**t**tgctgCATA**A**AAgc**g**gctgcttt	−0.48*
32	agatgctgCATA**A**AAgcagccgctt**c**	−0.51*	**2**	agatgctgC**G**TATAAgcagctgcttt	−0.43*
31	agatgctgCATATAAgcagctgct**c**t	0	**2**	**cagatgct**CATATAAgcagctgcttt	−0.19*
27	a**a**atgctgCATATAAgcagctgcttt	0	**2**	agatgctgCATATAAgca**c**ctgcttt	−0.10
24	aga**a**gctgCATATAAgcagc**c**gctt**c**	+0.04	**2**	a**a**ctgctgCATATAAgcagctgcttt	+0.08
24	agatgctgCATATAAgcagc**c**gctt**c**	+0.04	**2**	a**ct**tgctgCATATAAgcagctgcttt	+0.08
22	aga**a**gctgCATATAAgcagc**c**gcttt	+0.04	**2**	agatgctgCATATAAgcagc**g**gcttt	+0.08
20	aga**a**gctgCATATAAgcagctgcttt	0	**2**	g**ac**tgctgCATATAAgcagctgcttt	+0.08
19	agatg**a**tgCATA**A**AAgcagctgct**c**t	−0.56*	**2**	**cagatg**tgCATATAAgcagctgcttt	−0.06
18	agatgctgCATA**A**AAgcagctgct**c**t	−0.55*	**2**	agatgctgCATATAA**a**cagc**c**gcttt	+0.05
17	agat**c**ctgCATATAAgcagctgcttt	0	**2**	a**a**atgctgCATATAAgcagc**c**gcttt	+0.04
15	agatgctgCATATAAgcagctgc**c**tt	0	**2**	aga**c**gctgCATATAAgcagc**c**gcttt	+0.04
12	agatgctgCATATAAgcagctgc**g**tt	0	**2**	agatgctgCATATAAgcagc**c**gct**c**t	+0.04
11	aga**c**gctgCATATAAgcagctgcttt	0	**2**	agatgctgCATATAAgcagc**c**gctt**a**	+0.04
10	**c**gatgctgCATATAAgcagctgcttt	0	**2**	**g**gatgctgCATATAAgcagc**c**gcttt	+0.04
10	**g**gatgctgCATATAAgcagctgcttt	0	**2**	a**a**atgctgCATATAAAcagctgcttt	+0.01
9	agatgctgCATA**A**AAgcagc**c**gct**c**t	−0.51*	**2**	a**a**atgctgCATATAAgcagctgctt**c**	0
8	aga**a**gctgCATA**A**AAgcagc**c**gctt**c**	−0.51*	**2**	agatgctgCATATAAgcagctgctt**a**	0
7	aga**a**gctgCATA**A**AAgcagc**c**gcttt	−0.51*	**2**	agatgctgCATATAAgcagctg**t**ttt	0
7	agatgctgCATATAAgcagct**a**cttt	−0.01	**2**	agatgctgCATATAAgcag**tgct**ttt	0
7	**t**gatgctgCATATAAgcagctgcttt	0	**2**	agatg**t**tgCATATAAgcagctgcttt	0
5	agat**a**ctgCATATAAgcagctgcttt	+0.11*	**2**	**cag**tgctgCATATAAgcagctgcttt	0
5	agatgctgCATATAAgc**g**gctgcttt	−0.02	**2**	**g**gatgctgCATATAAgcagctgctt**g**	0

Table 1. (*Continued*)

n	DNA sequence, S	$\Delta^{\#}$	n	DNA sequence, S	$\Delta^{\#}$
5	agatgctgCATATAAgcagctgcttg	0	**2**	tgaagctgCATATAAgcagctgcttt	0
4	agatgctgCATATAGgcagctgcttt	−0.40*	**1**	agatgctgCATCTAAgcagctgcttt	−1.59*
4	agatgcggCATATAAgccgctgcttt	−0.20*	**1**	agatgctgCACAAAgcagccgcttt	−1.51*
4	agatcgtgCATATAAgcagctgcttt	−0.08	**1**	agatgctgCATAGAAgcagccgcttc	−1.37*
4	agaagctgCATATAAgcagctgcttc	0	**1**	agatgctgCATTTAAgcagcttcttt	−1.21*
3	agatgccgCATATAAgcagctgcttt	−0.13*	**1**	atgctgCAATAAcagccatccgcttt	−1.17*
1	agatgctgCATTTAAgcagctgcttt	−1.10*	**1**	ggatgttgCATATAAagagctgcttt	+0.10
1	agatgctgCATGTAAgcagctgcttt	−1.05*	**1**	agaagctgCATATAAgcagcggcttc	+0.08
1	agatgctgCACATAAgcagctgctct	−1.01*	**1**	agaatctgCATATAAgcagcgcgctt	+0.08
1	agatgctgCACATAAgcagctgcttt	−1.01*	**1**	agaccgtgCATATAAgcagctgcttc	−0.08
1	agatgctgCATAAAgcagccgcttt	−0.99*	**1**	agatgctgCATATAAgaagatgcctt	−0.08
1	aaattctgGATATTAAccaatgcctt	−0.97*	**1**	agatgctgCATATAAgcacgtgcttt	−0.08
1	ctatgctgCTTATAATcagctgcttt	−0.80*	**1**	agatgctgCATATAAgcaagctgctt	+0.07
1	aaatgctgATAAAAAgagcgctct	−0.60*	**1**	agatgctgCATATAAgcggcggcttt	+0.06
1	agatgctaCATATAAgcagccgcttt	+0.58*	**1**	gaatgccgCATATAAgcagcgcgcgc	−0.05
1	aaatgctgCATAAAAAccaccccctt	−0.57*	**1**	agatgctgCATATAAgcaggctgctt	+0.05
1	agatgatgCATAAAAgcagctactct	−0.56*	**1**	ggatgctgCATATAAgcaggctgctt	+0.05
1	aaatgctgCATAAAAgcagctgcttc	−0.55*	**1**	agaagctgCATATAAgcagccgctcc	+0.04
1	agaagctgCATAAAAgcagctgcttc	−0.55*	**1**	agacgctgCATATAAgcagccgcttc	+0.04
1	aaatgctaCATATAAgcagctgctct	+0.54*	**1**	agatgctgCATATAAgcagccgctt	+0.04
1	aaatgctgCATAAAAgcagccgctct	−0.51*	**1**	agatgctgCATATAAgcagccgtttt	+0.04
1	aaatgctgCATAAAAgcagccgcttt	−0.51*	**1**	agatgttgCATATAAgcagccgcttt	+0.04
1	agaagctgCATAAAAgcagccgcttg	−0.51*	**1**	gaatgctgCATATAAgcagccgcttt	+0.04
1	agatgctgCATAAAAgcagccgcctt	−0.51*	**1**	agatgctgCATATAAgccgccgcttt	+0.03
1	agatgctgCATAAAAgcagccgcttg	−0.51*	**1**	agatgctgCATATAAgccgccgttttt	+0.03
1	ggaagctgCATAAAAgcagccgcttt	−0.51*	**1**	cgatgctgCATATAAgcggctgcttt	−0.02
1	agatgctgCGTATAAgcagctactttt	−0.50*	**1**	agatgctgCATATAAacagctgcttt	+0.01
1	agatactgCATAAAAgcagtccgcttc	−0.48*	**1**	agatgctgCATATAAggagctgcttt	+0.01
1	cagagctgCATAAAAgcagccgcttc	−0.48*	**1**	aaatgctgCATATAAgcagctacttt	−0.01
1	agatgctgCATAAAAgcagcgcttct	−0.46*	**1**	atatgctgCATATAAgcagctgcttt	−0.01
1	aactgctgCATAAAAgcagccgcttc	−0.42*	**1**	agaagctgCATATAAgcagctgcctt	0

Table 1. (*Continued*)

n	DNA sequence, S	$\Delta^{\#}$	**n**	DNA sequence, S	$\Delta^{\#}$
1	agatgctgC**G**TATAAgcagc**c**gcttt	-0.42^{*}	**1**	aga**c**gctgCATATAAgcagctgctt**c**	0
1	agatgctgC**T**TATAAgcagctgcttt	-0.42^{*}	**1**	aga**g**gctgCATATAAgcagctgc**c**tt	0
1	**ca**a**atgctG**ATATAAgcagctgcttt	-0.37^{*}	**1**	aga**g**gctgCATATAAgcagctgcttt	0
1	agat**a**ctgCATATAAgcagc**g**cg**c**tt	$+0.19^{*}$	**1**	agatgctgCATATAAgcagctg**a**ttt	0
1	agatgc**c**gCATATAAgcagc**a**gcttt	-0.17^{*}	**1**	agatgctgCATATAAgcagctgct**cc**	0
1	agatgctgCATATAA**agcagctgc**tt	$+0.16^{*}$	**1**	agatgctgCATATAAgcagctgct**gc**	0
1	**ctc**agctgCATATAAgcagc**c**gcttt	$+0.15^{*}$	**1**	agatgctgCATATAAgcagctgct**g**t	0
1	a**a**atgctgCATATAA**c**cagctgcttt	-0.15^{*}	**1**	agatgctgCATATAAgcagctg**t**ttc	0
1	agatgctgCATATAA**c**cagctgcttt	-0.15^{*}	**1**	agatgctgCATATAAgcag**tgct**t**g**t	0
1	agatgc**c**gCATATAAgcagctgct**c**t	-0.13^{*}	**1**	ag**g**tgctgCATATAAgcagctgcttt	0
1	agatgc**c**gCATATAA**a**cagctgct**c**t	-0.12^{*}	**1**	**ca**atgctgCATATAAgcagctgcttt	0
1	a**at**tgctgCATATAAgcagc**c**gcttt	$+0.12^{*}$	**1**	**c**ga**a**gctgCATATAAgcagctgctt**a**	0
1	**gat**tgctgCATATAAgcagc**c**gcttt	$+0.12^{*}$	**1**	**ga**atgctgCATATAAgcagctgcttt	0
1	**ctca**g**a**tgCATATAAgcagctg**tgcc**	+0.10	**1**	**t**g**g**tgctgCATATAAgcagctgcttt	0

Note: **n**, occurrence; (#) Δ is the affinity, $-\mathrm{Ln}[K_{D,TATA}(S^0)]$, calculated by Eq. (1) for S^0 occurring n=1567 times; Δ is the departure, $-\mathrm{Ln}[K_{D,TATA}(S)]-(-\mathrm{Ln}[K_{D,TATA}(S^0)])$, from the normal affinity for non-S^0sequences; (*) α <0.05; the core sequence of the HIV-1 TATA box is in CAPITALS; differences from S^0 are in **bold**. Statistically significant deviations at α <0.05*: 9 positive ($\Delta > +\delta_{5\%}$, P>0.08) and 54 negative ($\Delta < -\delta_{5\%}$, P>10^{-31}).

in length for all positions of the sequence S and the complementary strand $S^{\$}$:

$$\mathrm{PWM_{TATA}}(\xi_{-1}...\xi_{13}) = \sum_{X\in\{A,T,G,C\}} \sum_{i=-1}^{13} f_X(i) \times \Delta(X = \xi_i); \tag{1a}$$

$-\mathrm{Ln}[K_{D,ssDNA}](S)$ is the average affinity of TBP for strands S and $S^{\$}$ in the window $(\zeta_{-2}, \ldots, \zeta_{+12})$ of 15 bp in length with the given weights,[37] F_{WR}, of WR dinucleotides[38] and the weights, F_{TV}, of TV dinucleotides[38] at position ζ_0 of the maximum of Bucher's criterion[34] of the TATA box:

$$\begin{aligned} -\mathrm{LnK_{D,ssDNA}}(\zeta_{-2}\cdots\zeta_{12}) = 14.5 + \sum_{i=-2}^{11} \{&0.9F_{WR}(i+2) \times \Delta(WR = \zeta_i\zeta_{i+1}) \\ &+ 2.5F_{TV}(i+2) \times \Delta(TV = \zeta_i\zeta_{i+1})\}; \end{aligned} \tag{1b}$$

$-\mathrm{Ln}[K_{D,dsDNA}](S)$ is the average affinity of TBP for double-stranded DNA in the window $(\Psi_{-5}, \ldots \Psi_{+9})$ of 15-bp in length with the minor groove width[39] in angstroms, h, taken from a database[31] and with the given weights,[31] F_{TA}, of TA dinucleotides; this affinity was calculated along the strand given the maximum of Bucher's criterion[34] of the TATA box:

$$\begin{aligned} -\mathrm{LnK_{D,dsDNA}}(\Psi_{-5}...\Psi_9) = -35.1 + 3.4\sum_{i=0}^{2} h(\Psi_I\Psi_{i+1}) - \\ -0.8\sum_{i=-5}^{9} F_{TA}(i+5) \times \Delta(TA = \psi_i\psi_{i+1}); \end{aligned} \tag{1c}$$

0.15 ± 0.05, 0.20 ± 0.09 and 0.23 ± 0.09 are the stoichiometric coefficients of the stages of TBP searching for the TATA box which are proportional to the numbers of DNA atoms involved into each stage, being assessed by the ratio of the TATA box length (15-bp) to that of the examined sequence S length (26-bp) and, also normalized to the total number (3) of these stages,[27] i.e. $(15/26)/3 = 0.19$. This three-step TBP/TATA-binding mechanism in solution has been just confirmed experimentally.[40]

2.3. *The 95% confidence boundary for the TBP/TATA affinity calculated (Eq. 1)*

The 95% confidence boundary for $-\mathrm{Ln}[K_{D,TATA}(S)]$, which is the TBP/TATA affinity estimate obtained using Eq. (1) for the sequence S 26-bp in length, was estimated using all the 3×26=78 single nucleotide substitutions $\{S^{\#}_{i,1}, S^{\#}_{i,2}, S^{\#}_{i,3}\}_{1\le i\le 26}$ that can occur in it. Because these are the smallest possible changes to S, we tested the hypothesis that "H_0: $-\mathrm{Ln}[K_{D,TATA}(S)] = -\mathrm{Ln}[K_{D,TATA}(S^{\#}_{i,k})]$", which is that these changes are too small for Formula (1) to help tell the affinity of the mutant sequences

from the affinity of the normal sequence; here $1 \leq k \leq 3$. Given these assumptions, we heuristically[41] estimated the confidence boundaries $\pm\delta_{95\%}$ by using Student's t-criterion with ν degrees of freedom:

$$\pm\,\delta_{95\%}(S) = t_{\alpha<0.05,\upsilon=78-1=77} \times\sqrt{\frac{\sum_{i=1}^{26}\sum_{k=1}^{3}\left(-Ln[K_{\mathrm{D,TATA}}(S^{\#}_{\mathrm{i,k}})] - \{-Ln[K_{\mathrm{D,TATA}}(S)]\}\right)^2}{3\times 26\times(3\times 26-1)}}. \quad (2)$$

The frequency of 146 HIV-1 TATA box variants in each particular country was calculated using the standard software program package STATISTICA.

3. Results

3.1. *TATA box and 3′-flanking E-box form a composite element*

A composite element is a regulatory unit composed of two overlapping or closely spaced binding sites, which function as a whole.[42] The 26-bp TATA-containing sequences with flanking E-boxes were taken in analysis; the binding of the transcription factor AP-4 to the 3′E-box inhibits basal expression.[18] Measuring the basal expression,[18] we came to the conclusion that TBP affinity is affected by mutations in the TATA box rather than in E-boxes. Upon analyzing the same sequences (Eq. 1), we demonstrated that the decrease in $-\ln[\mathrm{K_{D,TATA}(S)}]$ for E-box mutants correlates well with ($r = 0.884$ at $\alpha < 0.005$) with the decrease in the level of expression observed[18] (Table 2). Therefore, the 3′E-box and TATA box form a composite element, function cooperatively and must coevolve. This is consistent with the Imai-Okamoto's model,[43] which assumes that the proteins AP-4 and TBP interact both directly and involving chromatin proteins.

Table 2. Predicted TBP affinity, $-\ln[\mathrm{K_{D,TATA}(S)}]$, for HIV-1 TATA box variants correlates with the level of induction, $\ln(\mathrm{A/A_{WT}})$, of reporter gene expression.[18]

	Experiment[18]		Prediction, Eq. (1)
Probe name	DNA sequence, S	$\mathrm{Ln(A/A_{WT})}$	$-\mathrm{Ln[K_{D,TATA}(S)]}$
WT	tcagatgctgCATATAAgcagctgct	0.00	19.01
SV40	tcagatgctgCATATTTAgcagctgc	−1.50	19.12
NTATA	tcagatgctgCAgcgAAgcagctgct	−3.56	15.42
TATA 5′ HLH(-)	tTCgatgctgCATATAAgcagctgct	−0.12	19.01
TATA 3′ HLH(-)	tcagatgctgCATATAAgTCgctgct	−1.44	18.94
TATA 5′/3′ HLH(-)	tTCgatgctgCATATAAgTCgctgct	−1.58	18.94
5′/3′ E3 TATA	tCCCGGGcAgGGTATAAgcaCctgct	−1.44	18.36
5′/3′ E3 NTATA	tCCCGGGcAgGGgcgAAgcaCctgct	−3.47	14.82
Linear correlation coefficient, r			0.884
Statistical significance, α			< 0.005

3.2. *Low TBP/TATA affinity excess pin-points the slowly replicating HIV-1 forms*

For the agatgctgCATATAAgcagctgcttt variant found in 1567 out of 2662 HIV-1 copies (59%) and taken as a normal sequence, S^0, $-\mathrm{Ln}[K_{D,TATA}(S^0)]=19.52$ ln-units calculated by Eq. (1) with the confidence boundaries $\pm\delta_{5\%} = 0.10$ calculated by Eq. (2) for statistically significant ($\alpha < 0.05$) mutational deviations, $\Delta = -\mathrm{Ln}[K_{D,TATA}(S^{\#})]-\{-\mathrm{Ln}[K_{D,TATA}(S^0)]\}$, of the remaining 145 TATA box variants, $S^{\#}$'s, 63 gave significant mutational deviations Δ from S^0 (Table 1, asterisked). TBP/TATA affinity estimates were above normal ($\Delta > +\delta_{95\%}$) for nine out of 145 TATA box variants and below normal ($\Delta < -\delta_{95\%}$) for 54 out of 145 variants. This directly implies that the low-affinity TATA boxes outnumber the high-affinity TATA boxes by six to one (54/9=6). According to Struhl,[44] a mutational decrease in TBP affinity for the TATA box in a gene corresponds to a decrease in the level of expression of this gene, which allows us to interpret this obvious excess of low-affinity TATA boxes as a case of the slowly replicating HIV-1 forms.

For a detailed analysis of this excess, we built the histogram Δ (Fig. 1) of commonly accepted 12 equal-sized classes ($12^2 < 145 < 13^2$).

In Fig. 1, there are two peaks above the $1/12\times100\%$ level of equiprobable deviations. The major peak, $\Delta = 0$, corresponds to the normal sequence S^0 with the core sequence "CATATAA". The minor peak, $\Delta = -0.5$, corresponds to the second-most frequent TATA box variant "agatgctgCATAAAAgcagccgcttt" (Table 1, 294 isolates). The minor peak TATA box core sequence "CATAAAA" is associated[5,11,29] with HIV-1 subtype E. Because we found the minor peak, $\Delta = -0.5$, associated

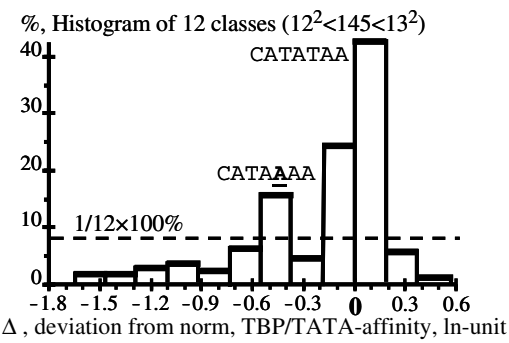


Fig. 1. A histogram of differences between 145 TATA box variants, $S^{\#}$, and the normal TATA box sequence, S^0, ($\Delta = -\ln[K_{D,TATA}(S^{\#})]-\{-\ln[K_{D,TATA}(S^0)]\}$). The histogram is built up from 12 equal-sized classes in accordance with the number of variants, $12^2 < 145 < 13^2$. It has two peaks above the $1/12\times100\%$ level of equiprobable deviations (broken line). The major peak, $\Delta = 0$, corresponds to the normal sequence S^0 with the core sequence "CATATAA". The minor peak, $\Delta = -0.5$, corresponds to the second most frequent TATA box variant "agatgctgCATAAAAgcagccgcttt" (Table 1, 294 isolates). The minor peak TATA box core sequence "CATAAAA" is asscoiated[5,11,29] with HIV-1 subtype E.

with the HIV-1 subtype that is the marker for HIV-1 migration, we wanted to look at the geographic distribution of 2662 HIV-1 TATA box variants: the frequency of occurrence, P_0, of the normal variant, S^0; the frequency of occurrence, P_1, of non-S^0 variants with the affinity as S^0 has; the frequency of occurrence, P_2, of high-affinity sequences at $\alpha < 0.05$; the frequency of occurrence, P_3, of low-affinity sequences at $\alpha < 0.05$ and the subtype E-associated HIV-1 TATA box core sequence "CATAAAA"; and the frequency of occurrence, P_4, of the other 146 TATA box variants (Table 3).

3.3. *Analysis of the frequencies of occurrence of HIV-1 TATA box variants in 70 countries*

The set of frequencies, $\{P_0, P_1, P_2, P_3, P_4\}$, for each of the 70 countries (Table 3) was weighted by the number, n, of the associated sequences (Table 3) and subjected to Principal Component Analysis (PCA) without normalization of the covariance matrix using the standard software program package STATISTICA. This procedure resulted in the observation of two principal components, PC1 (75.7% of the variance) and PC2 (23.3% of the variance). They significantly account for 75.7%+23.3% = 99% of the variance. In Fig. 2, the names of the countries with $n<10$ are typed in plain; those with $10 \leq n < 99$, in italics; those with $n \geq 100$, in bold. In the diagram presented in Fig. 2, all the countries fall within a triangle with the vertices $P_0 = 1$ (to the right below), $P_1 = 1$ (above) and $P_3 = 1$ (to the left below) and the sides $P_0 = 0$ (to the right above), $P_1 = 0$ (below) and $P_3 = 0$ (to the right above).

Europe (excluding France), North America and South America (excluding Cuba), Sub-Saharan West Africa and the former British African colonies lie close to the side $P_3 = 0$ (any TATA box with normal affinity). This placement corresponds to the major peak ($\Delta = 0$) in the histogram (Fig. 1), which suggests that neutral drift around the normal HIV-1 TATA box prevails here.

Caribbean countries (excluding Antilles), France and its former colonies in West, Central and Equatorial Africa (the place of HIV-1 origin) lie close to the sides $P_0 = 0$ and $P_1 = 0$. This placement corresponds to the minor peak ($\Delta = -0.5$) (Fig. 1). Consequently, advantage is being gained by the viruses with slowly replicating TATA box variants. A lot of such variants (see Table 1) may have emerged due to neutral drift around the core sequence "CATAAAA" (phylogenetic inertia) in the place of HIV-1 origin and repeated export to France[45] and Caribbean countries.[3]

Finally, South-East Asian countries distribute quite evenly within the triangle along the line from the vertex $P_3 = 1$ to the side $P_3 = 0$ (Fig. 2, broken line), that is, between two alternatives: "TATA box variants with low affinity at $\alpha < 0.05$ and the subtype E-associated HIV-1 core "CATAAAA" or "any TATA box variant with normal affinity". Because there is heavy migration between these countries, their lying between the two extremes may reflect the current state of some ongoing evolutionary process. For example, it is possible that slowly replicating HIV-1 variants

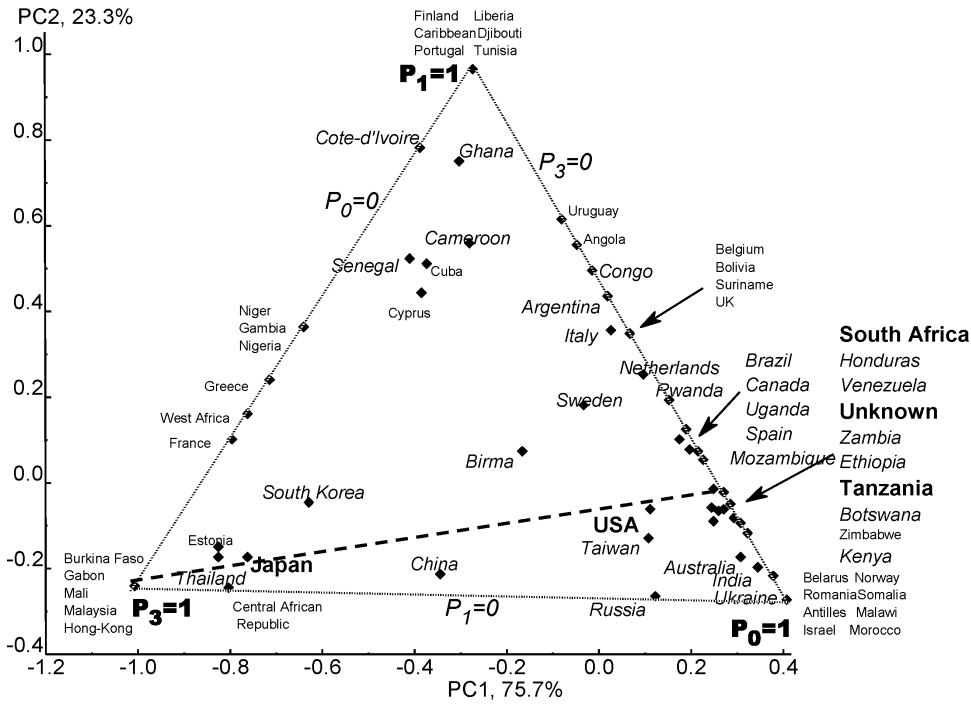


Fig. 2. Analysis of frequencies of 146 TATA box variants across each particular country (Table 2) revealed two principal components: PC1 (75.7% of the variance) and PC2 (23.3% of the variance). P_0, P_1, P_3 - see Table 2. The names of the countries with $n<10$ are typed in plain; those with $10 \leq n < 99$, in italics; those with $n \geq 100$, in bold. All the countries fall within a triangle with the vertices $P_0=1$ (to the right below), $P_1=1$ (above) and $P_3=1$ (to the left below) and the sides $P_0=0$ (to the right above), $P_1=0$ (below) and $P_3=0$ (to the right above). The broken line corresponds to the linear regression for the South-East Asian countries from the vertex $P_3=1$ to the side $P_3=0$.

are gaining advantage over both S^0 and non-S^0 variants – or the reverse may be true. The data contained in GenBank is too general to help in *in silico* identification of where that evolutionary process leads. However, literature data indicate that in South Asia, the prevalence of subtype E and its CRF (slowly replicating HIV-1 variants, as implied by our data) is growing.[5,46,47] What we can do is discuss various scenarios and compare them with literature data on the dynamics of the epidemic caused by HIV-1, primarily in South-East Asian countries.

4. Discussion

Ever since Wilson's classic work,[48] a growing body of molecular evidence strongly suggests that evolution has levels or tiers: microevolution and idioadaptation are associated with changes to the structural part of the genome, while long-term evolutionary trends and aromorphoses are associated with the evolution of regulatory genes (transcription factors, miRNA, etc.) and non-coding regulatory

Table 3. The frequencies of 146 HIV-1 TATA box variants ($P_0 - P_4$) in 70 countries and PCA results.

Country	**n**	P_0	P_1	P_2	P_3	P_4	PC1	PC2
USA	**411**	0.664	0.129	0.092	0.090	0.024	0.112	−0.063
South Africa	**296**	0.774	0.203	0.007	0.003	0.014	0.250	−0.012
Japan	**244**	0.139	0.057	0.000	0.795	0.008	−0.762	−0.174
Tanzania	**142**	0.831	0.148	0.007	0.000	0.014	0.291	−0.080
Zambia	**91**	0.813	0.165	0.000	0.011	0.011	0.271	−0.063
Botswana	**83**	0.783	0.120	0.012	0.012	0.072	0.249	−0.088
South Korea	**83**	0.181	0.169	0.000	0.651	0.000	−0.628	−0.044
Canada	**82**	0.659	0.280	0.024	0.000	0.037	0.174	0.100
Thailand	**82**	0.085	0.049	0.000	0.829	0.037	−0.824	−0.172
Uganda	**78**	0.705	0.282	0.000	0.013	0.000	0.196	0.078
Ghana	**56**	0.071	0.821	0.000	0.107	0.000	−0.305	0.749
Mozambique	**53**	0.736	0.264	0.000	0.000	0.000	0.226	0.056
Italy	**50**	0.440	0.480	0.040	0.000	0.040	0.025	0.355
Spain	**50**	0.720	0.280	0.000	0.000	0.000	0.216	0.075
China	**48**	0.438	0.021	0.021	0.500	0.021	−0.343	−0.215
India	**45**	0.911	0.044	0.000	0.000	0.044	0.345	−0.199
Australia	**42**	0.881	0.071	0.000	0.024	0.024	0.308	−0.173
Brazil	**28**	0.679	0.321	0.000	0.000	0.000	0.187	0.127
Venezuela	**28**	0.821	0.179	0.000	0.000	0.000	0.285	−0.050
Cameroon	**26**	0.154	0.654	0.000	0.154	0.038	−0.283	0.558
Ukraine	**23**	0.957	0.043	0.000	0.000	0.000	−0.343	−0.215
Netherlands	**22**	0.545	0.409	0.000	0.000	0.045	0.096	0.253
Sweden	**22**	0.500	0.364	0.000	0.136	0.000	−0.034	0.183
Argentina	**21**	0.429	0.571	0.000	0.000	0.000	0.017	0.436
Congo	**21**	0.381	0.619	0.000	0.000	0.000	−0.015	0.495
Taiwan	**21**	0.714	0.095	0.000	0.143	0.048	0.107	−0.130
Kenya	**16**	0.875	0.125	0.000	0.000	0.000	0.321	−0.117
Rwanda	**16**	0.625	0.375	0.000	0.000	0.000	0.151	0.193
Honduras	**15**	0.800	0.200	0.000	0.000	0.000	0.270	−0.024
Ethiopia	**14**	0.786	0.143	0.000	0.000	0.071	0.260	−0.066
Cote-d'Ivoire	**13**	0.000	0.846	0.000	0.154	0.000	−0.387	0.781
Birma	**11**	0.455	0.273	0.000	0.273	0.000	−0.165	0.074
Senegal	**11**	0.091	0.636	0.000	0.273	0.000	−0.413	0.525
Russia	**10**	0.800	0.000	0.000	0.200	0.000	0.124	−0.266
Cuba	**8**	0.125	0.625	0.000	0.250	0.000	−0.373	0.510
Malaysia	**8**	0.000	0.000	0.000	1.000	0.000	−1.007	−0.242
Niger	**8**	0.000	0.500	0.000	0.500	0.000	−0.641	0.362
Central African Republic	**7**	0.143	0.000	0.000	0.857	0.000	−0.805	−0.247
Cyprus	**7**	0.143	0.571	0.000	0.286	0.000	−0.387	0.445
France	**7**	0.000	0.286	0.000	0.714	0.000	−0.798	0.103
Uruguay	**7**	0.286	0.714	0.000	0.000	0.000	−0.080	0.613
Zimbabwe	**7**	0.857	0.143	0.000	0.000	0.000	0.309	−0.094
Belarus	**6**	1.000	0.000	0.000	0.000	0.000	0.406	−0.271
Greece	**5**	0.000	0.400	0.000	0.600	0.000	−0.714	0.241
Estonia	**4**	0.000	0.000	0.000	0.750	0.250	−0.826	−0.150
Gambia	**4**	0.000	0.500	0.000	0.500	0.000	−0.641	0.362
Angola	**3**	0.333	0.667	0.000	0.000	0.000	−0.048	0.554

Table 3. (*Continued*)

Country	**n**	P_0	P_1	P_2	P_3	P_4	PC1	PC2
Finland	**3**	0.000	1.000	0.000	0.000	0.000	−0.275	0.967
Liberia	**3**	0.000	1.000	0.000	0.000	0.000	−0.275	0.967
Norway	**3**	1.000	0.000	0.000	0.000	0.000	0.406	−0.271
Romania	**3**	1.000	0.000	0.000	0.000	0.000	0.406	−0.271
Somalia	**3**	1.000	0.000	0.000	0.000	0.000	0.406	−0.271
West Africa	**3**	0.000	0.333	0.000	0.667	0.000	−0.763	0.161
Antilles	**2**	1.000	0.500	0.000	0.000	0.000	0.406	−0.271
Belgium	**2**	0.500	0.500	0.000	0.000	0.000	0.066	0.348
Bolivia	**2**	0.500	0.500	0.000	0.000	0.000	0.066	0.348
Burkina Faso	**2**	0.000	0.000	0.000	1.000	0.000	−1.007	−0.242
Caribbean	**2**	0.000	1.000	0.000	0.000	0.000	−0.275	0.967
Djibouti	**2**	0.000	1.000	0.000	0.000	0.000	−0.275	0.967
Malawi	**2**	1.000	0.000	0.000	0.000	0.000	0.406	−0.271
Nigeria	**2**	0.000	0.500	0.000	0.500	0.000	−0.641	0.362
Portugal	**2**	0.000	1.000	0.000	0.000	0.000	−0.275	0.967
Suriname	**2**	0.500	0.500	0.000	0.000	0.000	0.066	0.348
UK	**2**	0.500	0.500	0.000	0.000	0.000	0.066	0.348
Gabon	**1**	0.000	0.000	0.000	1.000	0.000	−1.007	−0.242
Hong-Kong	**1**	0.000	0.000	0.000	1.000	0.000	−1.007	−0.242
Israel	**1**	1.000	0.000	0.000	0.000	0.000	0.406	−0.271
Mali	**1**	0.000	0.000	0.000	1.000	0.000	−1.007	−0.242
Morocco	**1**	1.000	0.000	0.000	0.000	0.000	0.406	−0.271
Tunisia	**1**	0.000	1.000	0.000	0.000	0.000	−0.275	0.967
Unknown	**281**	0.793	0.168	0.007	0.029	0.004	0.244	−0.058

Note: n, occurrence. Frequencies: P_0, normal S^0; P_1, non-S^0 with affinity as in S^0; P_2, any with high affinity at $\alpha < 0.05$; P_3, any with low affinity at $\alpha < 0.05$ and the subtype E associated HIV-1 TATA box core sequence "CATAAAA"; P_4, the remaining 146 TATA box variants (Table 1); Principal components PC1 and PC2 correspond to 75.7% and 23.3% of the variance.

sequences.[49] Over the past 25 years, the evolution of the structural part of the HIV-1 genome has been studied especially well (primarily because of its importance for vaccine development). Meanwhile, the search of subtypes for adaptive advantages other than being able to survive the immune attack yields inconsistent results.[5,13,46,50,51] One of the few well-substantiated observations is the dynamics of subtypes in South-East Asia: subtypes other than B (for example, C, E and F) either reduced the share of subtype or formed CRFs with it, which is commonly associated with the adaptation of non-B subtypes to heterosexual transmission.[5,46,47]

As HIV-1 progresses through its life cycle, it has to adequately respond to different evolutionary challenges. In a host organism, the virus is supposed to survive the immune attack and, here, selection acts so as to increase diversity.[12] A particular type of transmission (homosexual, heterosexual, whatever) limits diversity to the most adaptive genotypes and selection acts so as to decrease diversity.[52] Finally, at a population level, selection rates life strategies rather than diversity. A low transmissibility is compensated for by a prolonged latent stage of disease,

which favors the emergence of pestholes in loose host populations, where separate small groups of people (ingroups) are totally infected.[12] On the other hand, rapid frequent transmissions take the virus away from the immune attack and thus ease the requirements for generating diversity.[53]

Faced with controversial challenges, the virus has to respond with either specialization or wide adaptation, that is, to make expression regulation more flexible. The latter evolutionary trend should be long-lasting, not associated with the local evolution of separate proteins in a very limited number of proteins that the virus possesses. Homosexual or IDU-associated transmission, rapid and frequent, promotes specialization. Natural heterosexual transmission promotes flexible regulation.[12] The first decade of the AIDS pandemic was caused by rapid transmission within various risk groups, that is, was associated with specialization.[5,9,12,54] We hypothesize that this specialization resulted in the emergence, at the earlier stages of the pandemics, of high-affinity S^0's (the major peak in $\Delta = 0$, Fig. 1). We propose that the side $P_3 = 0$ (any TATA box variant with normal affinity) is where the cases of this specialization distributed by neutral drift between the vertices P_0=1 and P_1=1 lie along (Fig. 2).

In the second decade, the vector of selection changed its direction in some countries,[5,9,47] and this change corresponds to the broken line in Fig. 2, which reveals the only case of correspondence between the evolution of subtypes (subtype E in South-East Asia) and the TATA box variants possessing the core sequence TATAAA. We hypothesize that the observed excess of low-affinity variants reflects an evolutionary trend towards reduction in HIV-1 replication rates, which is consistent with the comparison of the directly measured replication status of "elder" (isolated in the 1980s) and "younger" (isolated in the XXI century) HIV-1 strains[55–57] and the studies of the life strategies of simian and feline immunodeficiency viruses.[4] The presence of TATA box variants other than those in subtype E within the triangle (Fig. 2) in addition to their presence on the vertex $P_3 = 1$ and the sides P_0=0 and $P_1 = 0$ suggests that evolution is going on not only in the core sequence of the TATA box but also in its flanks, thus optimizing all the three TBP/TATA-recognition steps[27,40] and the interactions in the TATA/AP-4 composite elements alike.[18]

In Fig. 2, the vertex $P_3 = 1$ and the side $P_0 = 0$ contain the place of HIV-1 origin, where natural heterosexual transmission in semi-isolated ingroups has had an important role all the way throughout the pandemic.[58,59] We propose that the presence of France on the side $P_0 = 0$ is due to its close relationships with its former African colonies, whose immigrants formed ingroups in the former mother country and retained the natural type of transmission.[45] In our opinion, the evolution of slowly replicating HIV-1 TATA box variants in these countries is most possibly phylogenetic inertia. These ingroups, now not so well-delineated,[60] yield low-affinity variants, which regulate HIV-1 replication rates by way of the TATA/AP-4 composite element, and high-affinity non-S^0 variants, which reduce HIV-1 replication rates, possibly by way of TAR.

Acknowledgments

The authors are grateful to Vladimir Filonenko for translating this paper from Russian into English. This work was supported by grants 08-04-01048 from the Russian Foundation for Fundamental Research, grant 119 of SB RAS, grant 23.29 "Biological diversity" of RAS, RAS Program "Molecular and Cell Biology" and "Origin and evolution of biosphere", grant 2447.2008.4 from Scientific School Program and State contract 10104-37/∏-18/110-327/180608/015.

References

1. Gupta RM, Sahni AK, Jena J, Nema SK, Genomic diversity of human immunodeficiency viruses, *MJAFI* **61**:267–270, 2005.
2. Peeters M, Chaix M-L, Delaporte E, Phylogénie des SIV et des VIH, *Medecine/Sciences* **24**:621–628, 2008 (in French).
3. Holmes ES, When HIV spread afar, *Proc Natl Acad Sci USA*, **104**:18351–18352, 2007.
4. *In vivo Models of HIV Disease and Control*, Friedman H, Specter S, Bendinelli M (eds), Springer Science+Business Media, New York, 2006.
5. Taylor BS, Sobieszczyk ME, McCutchan FE, Hammer SM, The Challenge of HIV-1 Subtype Diversity, *N Engl J Med* **358**:1590–1602, 2008.
6. Keele BF, Van Heuverswyn F, Li Y *et al.*, Chimpanzee reservoirs of pandemic and nonpandemic HIV-1, *Science* **313**:523–526, 2006.
7. Korber B, Muldoon M, Theiler J *et al.*, Timing the ancestor of the HIV-1 pandemic strains, *Science* **288**:1789–1796, 2000.
8. Gilbert MT, Rambaut A, Wlasiuk G *et al.*, The emergence of HIV/AIDS in the Americas and beyond, *Proc Natl Acad Sci USA* **104**:18566–18570, 2007.
9. Requejo HI, Worldwide molecular epidemiology of HIV, *Rev Saude Publica* **40**: 331–345, 2006.
10. McCutchan FE, Global epidemiology of HIV, *J Med Virol* **78**:S7–S12, 2006.
11. Hemelaara J, Gouws E, Ghys PD, Osmanov S, Global and regional distribution of HIV-1 genetic subtypes and recombinants in 2004, *AIDS* **20**:W13–W23, 2006.
12. Supotnitskiy VV, *Micro-Organisms, Toxins, and Epidemics. Vusovskaya. Kniga, Moscow,* Russia, 2005 (in Russian).
13. Kalish ML, Robbins KE, Pieniazek D *et al.*, Recombinant viruses and early global HIV-1 epidemic, *Emerg Infect Dis* **10**:1227–1234, 2004.
14. Olsen HS, Rosen CA, Contribution of the TATA motif to Tat-mediated transcriptional activation of human immunodeficiency virus gene expression, *J Virol* **66**:5594–5597, 1992.
15. Estable MC, Bell B, Merzouki A, Montaner JS, O'Shaughnessy MV, Sadowski IJ, Human immunodeficiency virus type 1 long terminal repeat variants from 42 patients representing all stages of infection display a wide range of sequence polymorphism and transcription activity, *J Virol* **70**:4053–4062, 1996.
16. Nabel G, Baltimore D, An inducible transcription factor activates expression of human immunodeficiency virus in T cells, *Nature* **326**:711–717, 1987.
17. Zhang L, Huang Y, Yuan H, Chen BK, Ip J, Ho DD, Genotypic and phenotypic characterization of long terminal repeat sequences from long-term survivors of human immunodeficiency virus type 1 infection, *J Virol* **71**:5608–5613, 1997.
18. Ou SH, Garcia-Martínez LF, Paulssen EJ, Gaynor RB, Role of flanking E box motifs in human immunodeficiency virus type 1 TATA element function, *J Virol* **68**: 7188–7199, 1994.

19. Brady J, Kashanchi F, Tat gets the "green" light on transcription initiation, *Retrovirology* **2**:69, 2005.
20. Raha T, Cheng SW, Green MR, HIV-1 Tat stimulates transcription complex assembly through recruitment of TBP in the absence of TAFs, *PLoS Biol* **3**(2):e44, 2005.
21. Majello B, Napolitano G, Lania L, Recruitment of the TATA-binding protein to the HIV-1 promoter is a limiting step for Tat transactivation, *AIDS* **12**(15):1957–1964, 1998.
22. Hoey T, Weinzierl RO, Gill G, *et al.*, Molecular cloning and functional analysis of Drosophila TAF110 reveal properties expected of coactivators, *Cell* **72**:247–260, 1993.
23. Gill G, Pascal E, Tseng ZH, Tjian R, A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation, *Proc Natl Acad Sci USA* **91**:192–196, 1994.
24. Deng L, Ammosova T, Pumfery A, Kashanchi F, Nekhai S, HIV-1 Tat interaction with RNA polymerase II C-terminal domain (CTD) and a dynamic association with CDK2 induce CTD phosphorylation and transcription from HIV-1 promoter, *J Biol Chem* **277**:33922–33929, 2002.
25. Lusic M, Marcello A, Cereseto A, Giacca M, Regulation of HIV-1 gene expression by histone acetylation and factor recruitment at the LTR promoter, *EMBO J* **22**: 6550–6561, 2003.
26. JPawlowski J, Montoya-Burgos JI, Fahrni JF, Wüest J, Zaninetti L, Origin of the Mesozoa inferred from 18S rRNA gene sequences, *Mol Biol Evol* **13**:1128–1132, 1996.
27. Ponomarenko PM, Savinkova LK, Drachkova IA, Lysova MV, Arshinova TV, Ponomarenko MP, Kolchanov NA, A step-by-step model of TBP/TATA box binding allows predicting human hereditary diseases by single nucleotide polymorphism, *Dokl Biochem Biophys* **419**:88–92, 2008.
28. Darling AC, Mau B, Blattner F, Perna NT, Mauve: multiple alignment of conserved genomic sequence with rearrangements, *Genome Res* **14**:1394–1403, 2004.
29. Montano MA, Nixon CP, Essex M, Dysregulation through the NF-kappaB enhancer and TATA box of the human immunodeficiency virus type 1 subtype E promoter, *J Virol* **72**:8446–8452, 1998.
30. Savinkova LK, Drachkova IA, Ponomarenko MP, Lysova MV, Arshinova TV, Kolchanov NA, Interaction between the recombinant TATA-binding protein and the TATA-boxes of the mammalian gene promoters, *Ecological Genetics* **V**:44–49, 2007 (in Russian).
31. Ponomarenko MP, Ponomarenko JV, Frolov AS, Podkolodny NL, Savinkova LK, Kolchanov NA, Overton GC, Identification of sequence-dependent features correlating to activity of DNA sites interacting with proteins, *Bioinformatics* **15**:687–703, 1999.
32. Hahn S, Buratowski S, Sharp PA, Guarente L, Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences, *Proc Natl Acad Sci USA* **86**:5718–5722, 1989.
33. Coleman RA, Pugh BF, Evidence for functional binding and stable sliding of the TATA binding protein on nonspecific DNA, *J Biol Chem* **270**:13850–13859, 1995.
34. Bucher P, Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences, *J Mol Biol* **212**:563–578, 1990.
35. Juo ZS, Chiu TK, Leiberman PM, Baikalov I, Berk AJ, Dickerson RE, How proteins recognize the TATA box, *J Mol Biol* **261**:239–254, 1996.
36. Powell R, Parkhurst K, Parkhurst L, Comparison of TATA-binding protein recognition of a variant and consensus DNA promoters, *J Biol Chem* **277**:7776–7784, 2002.

37. Ponomarenko MP, Savinkova LK, Ponomarenko JV, Kel AE, Titov II, Kolchanov NA, Modeling TATA-box sequences in eukaryotic genes, *Mol Biol (Mosk)* **31**:726–732, 1997 (in Russian).
38. IUPAC-IUB commission on biochemical nomenclature (CBN), *J Mol Biol* **55**: 299–310, 1971.
39. Karas H, Knüppel R, Schulz W, Sklenar H, Wingender E, Combining structural analysis of DNA with search routines for the detection of transcription regulatory elements, *Comput Appl Biosci* **12**(5):441–446, 1996.
40. Delgadillo RF, Whittington JE, Parkhurst LK, Parkhurst LJ, The TATA-binding protein core domain in solution variably bends TATA sequences via a three-step binding mechanism, *Biochemistry* **48**:1801–1809, 2009.
41. Ponomarenko PM, Ponomarenko MP, Drachkova IA, Lysova MV, Arshinova TV, Savinkova LK, Kolchanov NA, Prognosis of affinity change of the TATA-binding protein to TATA-boxes upon polymorphisms of the human gene promoter TATA boxes, *Mol Biol (Mosk)* **43**(3):512–520, 2009 (in Russian).
42. Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA, Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL, *Nucleic Acids Res* **26**(1):362–367, 1998.
43. Imai K, Okamoto T, Transcriptional repression of human immunodeficiency virus type 1 by AP-4, *J Biol Chem* **281**:12495–12505, 2006.
44. Struhl K, The yeast his3 promoter contains at least two distinct elements, *Proc Natl Acad Sci USA* **79**:7385–7389 (1982).
45. Bourée P, Lamour P, Bisaro F, Didier E, Etude d'une population d'origine tropicale, VIH positive, dans un centre de réfugiés en France, *Bull Soc Pathol Exot* **88**:24–28, 1995 (in French).
46. Cohen MS, Hellmann N, Levy JA, DeCock K, Lange J, The spread, treatment, and prevention of HIV-1: evolution of a global pandemic, *J Clin Invest* **118**:1244–1254, 2008.
47. Lau KA, Wang B, Saksena NK, Emerging trends of HIV epidemiology in Asia, *AIDS Rev* **9**:218–229, 2007.
48. King MC, Wilson AC, Evolution in two levels in humans and chimpanzees, *Science* **188**:107–116, 1975.
49. Carroll SB, Evolution at Two Levels: On Genes and Form, *PLoS Biol* **3**:e245, 2005.
50. Roques P, Robertson DL, Souquière S *et al.*, Phylogenetic analysis of 49 newly derived HIV-1 group O strains: high viral diversity but no group M-like subtype structure, *Virology* **302**:259–273, 2002.
51. Coutinho RA, Some aspects of the natural history of HIV infection, *Trop Med Int Health* **5**:A22–25, 2000.
52. Herbeck JT, Nickle DC, Learn GH *et al.,* Human immunodeficiency virus type 1 env evolves toward ancestral states upon transmission to a new host, *J Virol* **80**: 1637–1644, 2006.
53. Maljkovic Berry I, Ribeiro R, Kothari M *et al.*, Unequal evolutionary rates in the human immunodeficiency virus type 1 (HIV-1) pandemic: the evolutionary rate of HIV-1 slows down when the epidemic rate increases, *J Virol* **81**:10625–1035, 2007.
54. Couturier E, Damond F, Roques P *et al.*, HIV-1 diversity in France, 1996–1998. The AC 11 laboratory network, *Bull Soc Pathol Exot* **88**:24–28, 1995.
55. Ariën KK, Troyer RM, Gali Y, Colebunders RL, Arts EJ, Vanham G, Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time, *AIDS* **19**:1555–1564, 2005.

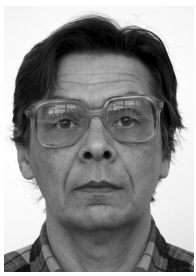
56. Ariën KK, Vanham G, Arts EJ, Is HIV-1 evolving to a less virulent form in humans?, *Nat Rev Microbiol* **5**:141–151, 2007.
57. Quiñones-Mateu ME, Is HIV-1 evolving to a less virulent (pathogenic) virus?, *AIDS* **19**:1689–1690, 2005.
58. Du Guerney J, Migrations et SIDA en Afrique, *Chron CEPED* **Autumn** (15):4–5, 1994 (in French).
59. Amat-Roze JM, Coulaud JP, Charmot G, La géographie de l'infection par les virus de l'immunodéficience humaine (VIH) en Afrique Noire: mise en évidence de facteurs d'épidémisation et de régionalisation, *Bull Soc Pathol Exot* **83**:137–148, 1990 (in French).
60. Vallet S, Legrand-Quillien MC, Roger C *et al.*, HIV-1 genetic diversity in Western Brittany, France, *FEMS Immunol Med Microbiol* **34**:65–71, 2002.

Valentin Suslov received his M.Sc. degree in Genetics from Novosibirsk State University, Russia, in 1997. He is currently Staff Research Scientist in the Sector for Evolutionary Bioinformatics of the Department for Systems Biology at the Institute of Cytology and Genetics, Novosibirsk, Russia.

Petr Ponomarenko received his Bachelor's degree in Physics from Novosibirsk State University, Russia in 2008. He is currently in the Department for Chemical and Biological Physics of Novosibirsk State University, Russia. He received his M.Sc. degree in Physics in July, 2010, following a scholarship from both V.O. Potanin's Foundation and Schlumberger Limited for young scientists.

Vadim Efimov received his M.Sc. degree in Mathematics from Novosibirsk State University, Russia in 1970, and his Ph.D. and Dr.Sci. degree in Ecology from Tomsk State University, Russia in 2000 and 2003, respectively. He is currently Leading Staff Scientist in the Laboratory for Molecular Genetic Systems of the Department for Systems Biology at the Institute of Cytology and Genetics, Novosibirsk, Russia.

Ludmila Savinkova received her M.Sc. degree in Biology from Novosibirsk State University, Russia in 1970, her Ph.D. and Senior Staff Scientist (equivalent to Assistant Professor) degrees in Biology from Institute of Cytology and Genetics, Novosibirsk, Russia in 1984 and 1997, respectively. She is currently Chief of the Sector for Molecular-Genetic Mechanisms of Protein-Nucleic Acid Interactions of the Gene Expression Laboratory at the Institute of Cytology and Genetics, Novosibirsk, Russia.



Mikhail Ponomarenko received his M.Sc. degree in Physics from Novosibirsk State University, Russia in 1985 and his Ph.D. degree in Biology from Institute of Cytology and Genetics, Novosibirsk, Russia in 1994. He is currently Senior Staff Scientist in the Laboratory for Theoretical Genetics of the Department of System Biology at the Institute of Cytology and Genetics, Novosibirsk, Russia. He is Laureate in Physics for USSR National Graduate Students (1985) and D.K Belyaev Prize Laureate in Genetics (1995).

Nikolay Kolchanov received his M.Sc. degree in Biology from Novosibirsk State University, Russia in 1971, his Ph.D., Dr.Sci. and Professor degrees in Genetics from Institute of Cytology and Genetics, Novosibirsk, Russia in 1975, 1989 and 1992, respectively. He is currently Academician of Russian Academy of Sciences, Director of the Institute of Cytology and Genetics, Novosibirsk, Russia, being Chief of the Department for System Biology, A.A. Baev Prize Laureate in Genomics and Genoinformatics (1995).

He is the HUGO member, Vice-President of the Scientific Council for Genetics and Breeding of the Russian Academy of Sciences and the N.I. Vavilov's Society of Geneticists and Breeders (VOGiS) and a Member of the Scientific Council of Molecular Biology and Genetics of the Russian Academy of Sciences. He is also a Member of the Scientific Councils of Supercomputing and Life Sciences of Siberian Branch of the Russian Academy of Sciences, Working Group "Nanotechnology Strategy Development" of the Russia Government Council for Nanotechnology, Russian Foundation for Basic Research, Russian National Human Genome Project, and editorial board of the journals "In Silico Biology", "VOGiS Herald", "Siberian Ecological Journal" and "Siberian Journal on Applied Mathematics".