Title: Rare HIV-1 subtype J genomes and a new H/U/CRF02_AG recombinant genome suggests an ancient origin of HIV-1 in Angola

Running title: Full-length genomes of HIV-1 from Angola

Inês Bártolo[1*], Rita Calado[1*], Pedro Borrego[1], Thomas Leitner[2] and Nuno Taveira[1,3§]

[1] Research Institute for Medicines (iMed.ULisboa), Faculty of Pharmacy, Universidade de Lisboa, Portugal.

[2]Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA;

[3]Centro de Investigação Interdisciplinar Egas Moniz (CiiEM), Instituto Superior de Ciências da Saúde Egas Moniz, Monte de Caparica, Portugal.

* These authors have contributed equally to the work.

[§]Corresponding author: Nuno Taveira

Instituto de Investigação do Medicamento (iMed.ULisboa)

Faculdade de Farmácia, Universidade de Lisboa

Av. Prof. Gama Pinto, 1649-003 Lisboa, Portugal.

Phone/Fax: +351217934212

ntaveira@ff.ul.pt

Keywords: HIV, Epidemiology, Phylogenetics

Abstract

Angola has an extremely diverse HIV-1 epidemic fueled in part by the frequent interchange of people with the Democratic Republic of Congo (RDC) and Republic of Congo (RC). Characterization of HIV-1 strains circulating in Angola should help to better understand the origin of HIV-1 subtypes and recombinant forms and their transmission dynamics. In this study we characterize the first near full-length HIV-1 genomic sequences from HIV-1 infected individuals from Angola.

Samples were obtained in 1993 from three HIV-1 infected patients living in Cabinda, Angola. Near full-length genomic sequences were obtained from virus isolates. Maximum likelihood phylogenetic tree inference and analyses of potential recombination patterns were performed to evaluate the sequence classifications and origins. Phylogenetic and recombination analyses revealed that one virus was a pure subtype J, another mostly subtype J with a small uncertain region, and the final virus was classified as a H/U/CRF02_AG recombinant. Consistent with their epidemiological data, the subtype J sequences were more closely related to each other than to other J sequences previously published. Based on the *env* gene, taxa from Angola occur throughout the global subtype J phylogeny.

HIV-1 subtypes J and H are present in Angola at low levels since at least 1993. Low transmission efficiency and/or high recombination potential may explain their limited epidemic success in Angola and worldwide. The high diversity of rare subtypes in Angola suggests that Angola was part of the early establishment of the HIV-1 pandemic.

According to the World Health organization (WHO) there are currently about 35 million people living with HIV/AIDS worldwide and it is estimated that approximately 39 million people have died of AIDS since 1981 [1]. HIV can be divided into two types: HIV-1, which is responsible for the worldwide AIDS pandemic, and HIV-2, which is less prevalent and virulent [2]. Based on extreme genetic diversity, HIV-1 variants can be classified into four major groups: M (main), O (outlier), N (non-M, non-O), and more recently P [3]. The M group, which is responsible for the pandemic, includes nine different subtypes (A-D, F-H, J and K), more than 70 circulating recombinant forms (CRFs), and many unique recombinant forms (URFs) [4]. Groups O, N and P have infected a relatively small number of patients worldwide [5].

In sub-Saharan Africa almost all subtypes, CRFs, and URFs circulate; the highest genetic diversity has been observed in West Central Africa where the HIV-1 pandemic is believed to have originated [5-7]. On a global perspective, the most prevalent HIV-1 subtypes are C (50%), A (12%), B (11%), followed by CRF02_AG (8%), G (5%), CRF01_AE (5%) and D (2%). The remaining subtypes and recombinant strains represent less than 1% of HIV-1 infections [5, 6, 8]. Currently there are only three full-length genomes of subtype J and four of subtype H available in the Los Alamos HIV sequence database [4]. Full-length subtype J genomes are from Sweden and Cameroon (GenBank accession numbers AF082394, AF082395, GU237072) whereas subtype H genomes are from Belgium, United Kingdom and Central African Republic (AF190127, AF190128, FJ711703, AF005496). In addition, subtype J appears as fragments in several CRFs and shorter sequences [4].

Angola is a South-western African country surrounded by Namibia, Zambia, Democratic Republic of Congo (DRC), and Republic of Congo (RC). According to the UNAIDS report, HIV/AIDS prevalence among adults in Angola was 2.4% in 2014 [9].

Despite this low prevalence, all subtypes except B and many CRFs and URFs have been detected in Angola, in particular in the Provinces of Luanda and Cabinda where most studies have been done. This high diversity is a direct consequence of the long-standing presence of HIV-1 in the country and the frequent interchange of people with the DRC and RC [10-14]. In this study we describe the first near full-length HIV-1 genomic sequences from HIV-1 infected individuals from Cabinda, a province of Angola which is an exclave surrounded by DRC in the south and east, and by the RC in the north.

Blood samples from HIV-1 infected individuals were collected in 1993 from Hospital Distrital de Cabinda, Cabinda, Angola. Samples were collected anonymously with oral consent. The study was approved by the ethics committees of the participating institutions. Virus isolates were obtained using the co-cultivation method as described previously [15]. All three isolates used the CCR5 co-receptor [16]. Viral genomic RNA was extracted from cell culture supernatant and RT-PCR was performed using Titan One Tube RT-PCR System (Roche Diagnostic Systems). All amplifications were performed using the Expand Long Template PCR system (Roche Diagnostic Systems) according to the manufacturer's instructions. New primers were designed to amplify the full-length genomes (Supplementary material, Table 1). Amplified DNA fragments were purified using JETQUICK Gel Extraction Spin Kit (Genomed) and sequenced. The nucleotide sequence data was deposited in GenBank with accession numbers KU310618-20. Genomic sequences were aligned with a set of reference sequences representative of all HIV-1 group M subtypes obtained from the Los Alamos HIV Sequence Database[17]using Clustal X 2.1 (http://www.clustal.org/clustal2/) and the alignment was manually edited with GeneDoc (http://iubio.bio.indiana.edu/soft/molbio/ibmpc/genedoc-readme.html). Maximum likelihood (ML) trees were inferred with program PhyML using the Seaview software (http://pbil.univ-lyon1.fr/software/seaview.html) using the best-fit substitution

model identified by Modeltest v3.7 using the Akaike information criterion. To find the ML tree an iterative heuristic method combining nearest neighbor interchange and subtree pruning and regrafting tree rearrangement methods was used. The reliability of the obtained tree was estimated with the approximate likelihood-ratio test (aLRT). Potential recombination patterns of our new near full-length sequences were analyzed by bootscanning using SimPlot with a sliding window of 500 bp advanced in 100 bp increments (http://sray.med.som.jhmi.edu/SCRoftware/simplot/). For each window 100 bootstrap replicates were generated. Potential recombination breakpoints between subtypes were considered when the percentage of permuted trees for a given subtype was above 70%. We also used RIP, jpHMM and the branching index (BI) to confirm the bootscanning results [18-21].

An overall genome-wide tree analysis of 93AOHDC250 and 93AOHDC253 showed that they clustered with subtype J reference sequences and were more closely related to each other than to the other J isolates, which is consistent with their epidemiological data (Figure 1). Note that this tree is not a true phylogeny as it cannot depict the evolutionary history of all recombinant sequences; we used it here merely to investigate the overall sequence-based similarity for our classification purpose.

Bootscanning analysis showed no evidence of recombination in isolate 93AOHDC253 (Figure 2A). Isolate 93AOHDC250 displayed an untypable region between positions 1450 and 2050 corresponding to the end of *gag* gene, the protease (PR) region, and the first 192 nucleotides of reverse transcriptase (RT), which was further confirmed by phylogenetic analysis (Figure 2B). BI results confirmed that 93AOHDC253 was a pure subtype J and that 93AOHDC250 had a genomic region with no known subtype in this *gag/pol* region (Figure 5).

Comparing our new J sequences to existing J sequences in the Los Alamos HIV database revealed that subtype J is quite diverse (Figure 3). In the region where most J sequence fragments exist (HXB2 positions 7041-7358), 93AOHDC250 and 93AOHDC253 cluster together with a previous sequence from Angola, 93AOHDC247 [10]. Interestingly, Angolan J sequences occur throughout the phylogeny of this *env* region, suggesting that the J epidemic in Angola is either the origin of subtype J or that there has been a lot of influx of subtype J from other geographic regions. The great diversity of subtype J has been previously noted in analyses of J-containing CRF11 and CRF13 genomes [22, 23].

The overall tree analysis showed that 93AOHDC251 clustered with subtype H reference sequences (Figure 1). Bootscanning analysis indicated an unclassifiable (U) region between positions 4936 and 5167 (corresponding to *vpr* gene), and a region between positions 8151 and 8888 (corresponding to the *nef* gene and 3´LTR) appeared to cluster between subtypes A1 and G (Figure 2C). BI analyses again confirmed these results, and suggested that the *nef*/LTR region was below the subtype-defining threshold for subtype G (Figure 5). Upon closer inspection using RIP, comparing this region to subtypes A1, G and CRF02, it became clear that this region was in fact derived from CRF02; 93AOHDC251 was closer to CRF02 than either A1 or G in all parts of this region, which included a CRF02 A/G breakpoint (Figure 4). Together, these analyses led to a classification of this genome sequence as the first H/U/CRF02_AG recombinant.

In this report we describe the first three HIV-1 genomic sequences from Angola, a country that with the DRC and RC had a crucial role in the early dissemination of the HIV-1 epidemic [7, 10]. Sequences were derived from isolates obtained in 1993 from patients living in Cabinda, a province in the North that borders both DRC and RC.

One isolate was a pure subtype J, another a J with an at this point unclear segment, and one was an H-based recombinant. The unclear region in 93AOHDC250 may be a divergent J segment or a segment of an as yet undiscovered subtype. As mentioned above, subtype J has previously been described as very diverse and may hide further sub-subtypes or recombinants of sub-subtype nature [22].

Despite being in circulation in Cabinda since 1993, as of February 2015 only 27 J and 73 H sequences from Angola had been deposited in the Los Alamos HIV Sequence Database (3.1 and 8.3% of respective subtype sequences from Angola). Globally, pure subtypes J and H are also very rare which suggests low biological fitness, low transmissibility, high recombination potential, and unsuccessful introductions into high-risk populations. However, the isolates described herein replicate to normal levels in different cell types suggesting that their biological fitness is no different from other subtypes [16]. Interestingly, we have found that our J and H isolates are much more sensitive (about nine times) to the CCR5-antagonist TAK-779 when compared with isolates from other subtypes that are more prevalent in Angola and Portugal [16]. This suggests poor binding to the main HIV co-receptor (CCR5) which may indicate transmissibility problems of these subtypes. On the other hand, J and H subtypes have been found in many recombinants such as CRF04_cpx, CRF06_cpx, CRF11_cpx, CRF13_cpx, CRF18_cpx, CRF27_cpx, and CRF49_cpx[17]. Thus, the rarity of pure subtype J and H, together with the large HIV-1 diversity in Angola and neighboring countries, which suggests they have been around for long times, suggests that these subtypes are less fit to transmit or establish infection by themselves. Similar to how latent virus may survive within a host through recombination with non-latent immune escaping plasma virus, subtype J and H virus may do better if they recombine with more fit subtypes [24].

HIV-1 subtypes J and H are present in Angola at low levels since at least 1993. The high diversity among Angolan subtype J *env* sequences and the fact that the rare subtype H has recombined in Angola together suggest that Angola is either the origin of subtype J or, more complicated, that there has been a lot of influx of subtype J from other geographic regions. Low transmission efficiency and/or high recombination potential may explain their limited epidemic success in Angola and worldwide.

Competing interests

The authors have no commercial or other type of association that might pose a conflict of interest.

Acknowledgements and funding

Financial support for this research was provided by the Fundação para a Ciência e a Tecnologia (FCT), Portugal (project PTDC/SAU-EPI/122400/2010), part of the EDCTP2 program supported by the European Union. Rita Calado is supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal (grant number SFRH/BD/70715/2010). Inês Bártolo is supported by Fundação para a Ciência e a Tecnologia (FCT), Portugal (grant number SFRH/BPD/76225/2011). Thomas Leitner was supported by the National Institutes of Health (NIH), USA (grant number R01AI087520).

Sequence data

Sequences have been assigned with GenBank accession numbers KU310618, KU310619 and KU310620

Authors' contributions

Conceived and designed the experiments: IB, TL and NT. Performed the experiments: IB, PB, TL and RC. Analyzed the data: IB, TL, RC and NT. Wrote the paper: IB, RC, TL and NT. The final text was read and approved for submission by all authors.

References

1. WHO. *Fact Sheets HIV/AIDS* July 2015 2015.

2. Santoro MM, Perno CF. HIV-1 Genetic Variability and Clinical Implications. *ISRN Microbiol.* 2013;2013:481314.

3. Mourez T, Simon F, Plantier JC. Non-M variants of human immunodeficiency virus type 1. *Clin Microbiol Rev.* Jul 2013;26(3):448-461.

4. Foley B, Leitner T, Apetrei C, Hahn B, Mizrachi I, Mullins J, Rambaut A, Wolinsky S, and Korber B. *HIV Sequence Compendium 2015*: Theoretical Biology and Biophysics Group, Los Alamos National Laboratory; 2015.

5. Peeters M, Jung M, Ayouba A. The origin and molecular epidemiology of HIV. *Expert Rev Anti Infect Ther.* Sep 2013;11(9):885-896.

6. Buonaguro L, Tornesello ML, Buonaguro FM. Human immunodeficiency virus type 1 subtype distribution in the worldwide epidemic: pathogenetic and therapeutic implications. *J Virol.* Oct 2007;81(19):10209-10219.

7. Faria NR, Rambaut A, Suchard MA, et al. HIV epidemiology. The early spread and epidemic ignition of HIV-1 in human populations. *Science.* Oct 3 2014;346(6205):56-61.

8. Abecasis AB, Wensing AM, Paraskevis D, et al. HIV-1 subtype distribution and its demographic determinants in newly diagnosed patients in Europe suggest highly compartmentalized epidemics. *Retrovirology.* 2013;10:7.

9. UNAIDS. Epidemiological fact sheets on HIV and AIDS. HIV and AIDS estimates (2013). Angola. http://www.unaids.org/sites/default/files/epidocuments/AGO.pdf. Accessed January 2016, 2015.

10. Bartolo I, Rocha C, Bartolomeu J, et al. Highly divergent subtypes and new recombinant forms prevail in the HIV/AIDS epidemic in Angola: new insights into the origins of the AIDS pandemic. *Infect Genet Evol.* Jul 2009;9(4):672-682.

11. Abecasis A, Paraskevis D, Epalanga M, et al. HIV-1 genetic variants circulation in the North of Angola. *Infect Genet Evol.* Apr 2005;5(3):231-237.

12. Bartolo I, Epalanga M, Bartolomeu J, et al. High genetic diversity of human immunodeficiency virus type 1 in Angola. *AIDS Res Hum Retroviruses.* Apr 2005;21(4):306-310.

13. Bartolo I, Rocha C, Bartolomeu J, et al. Antiretroviral drug resistance surveillance among treatment-naive human immunodeficiency virus type 1-infected individuals in Angola: evidence for low level of transmitted drug resistance. *Antimicrob Agents Chemother.* Jul 2009;53(7):3156-3158.

14. Bartolo I, Zakovic S, Martin F, et al. HIV-1 diversity, transmission dynamics and primary drug resistance in Angola. *PLoS One.* 2014;9(12):e113626.

15. Cavaco-Silva P, Taveira NC, Rosado L, et al. Virological and molecular demonstration of human immunodeficiency virus type 2 vertical transmission. *J Virol.* Apr 1998;72(4):3418-3422.

16. Borrego P, Calado R, Marcelino JM, et al. Baseline susceptibility of primary HIV-2 to entry inhibitors. *Antivir Ther.* 2012;17(3):565-570.

17. Los Alamos Sequence Database. http://www.hiv.lanl.gov/. Accessed January, 12th 2016.

18. Schultz AK, Zhang M, Bulla I, et al. jpHMM: improving the reliability of recombination prediction in HIV-1. *Nucleic Acids Res.* Jul 2009;37(Web Server issue):W647-651.

19. Zhang M, Schultz AK, Calef C, et al. jpHMM at GOBICS: a web server to detect genomic recombinations in HIV-1. *Nucleic Acids Res.* Jul 1 2006;34(Web Server issue):W463-465.

20. Hraber P, Kuiken C, Waugh M, Geer S, Bruno WJ, Leitner T. Classification of hepatitis C virus and human immunodeficiency virus-1 sequences with the branching index. *J Gen Virol.* Sep 2008;89(Pt 9):2098-2107.

21. Wilbe K, Salminen M, Laukkanen T, et al. Characterization of novel recombinant HIV-1 genomes using the branching index. *Virology.* Nov 10 2003;316(1):116-125.

22. Zhang M, Wilbe K, Wolfe ND, Gaschen B, Carr JK, Leitner T. HIV type 1 CRF13_cpx revisited: identification of a new sequence from Cameroon and signal for subsubtype J2. *AIDS research and human retroviruses.* Nov 2005;21(11):955-960.

23. Wilbe K, Casper C, Albert J, Leitner T. Identification of two CRF11-cpx genomes and two preliminary representatives of a new circulating recombinant form (CRF13-cpx) of HIV type 1 in Cameroon. *AIDS Res Hum Retroviruses.* Aug 10 2002;18(12):849-856.

24. Immonen TT, Conway JM, Romero-Severson EO, Perelson AS, Leitner T. Recombination Enhances HIV-1 Envelope Diversity by Facilitating the Survival

of Latent Genomic Fragments in the Plasma Virus Population. *PLoS computational biology*. Dec 2015;11(12):e1004625.

Reprint requests should be directed to Nuno Taveira (ntaveira@ff.ul.pt)

Figure legends

Figure 1

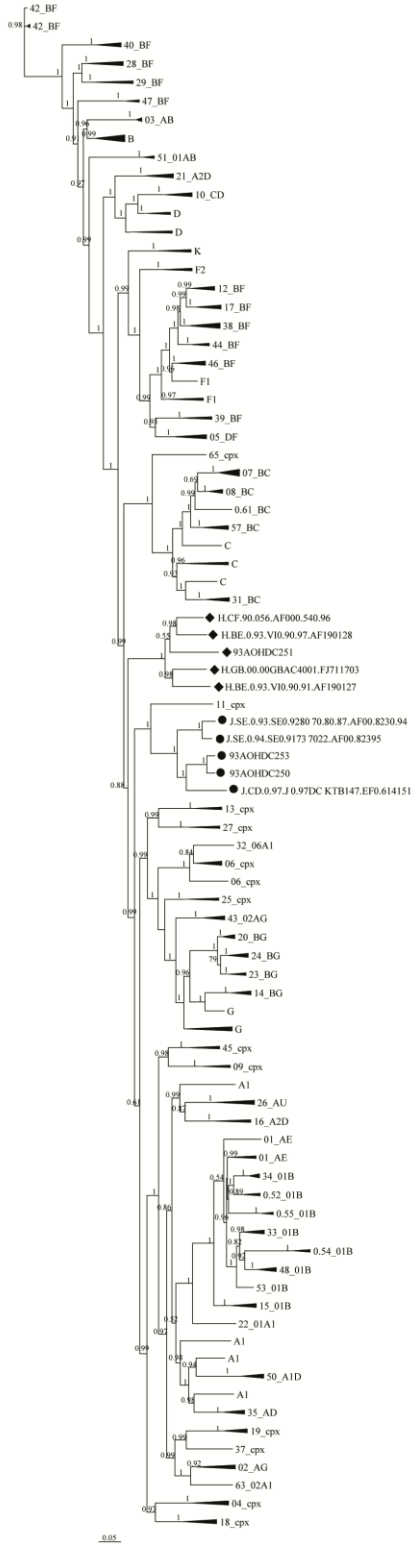

Figure 1 – Overall tree classification of HIV-1 near full-length genomes of isolates 93AOHDC250, 93AOHDC251 and 93AOHDC253. ML trees were constructed with our sequences (highlighted by colored dots) and 136 reference sequences representative

of all HIV-1 group M subtypes. Cladistics support is indicated by aLRT values. The scale is in units of substitutions per site.

Figure 2

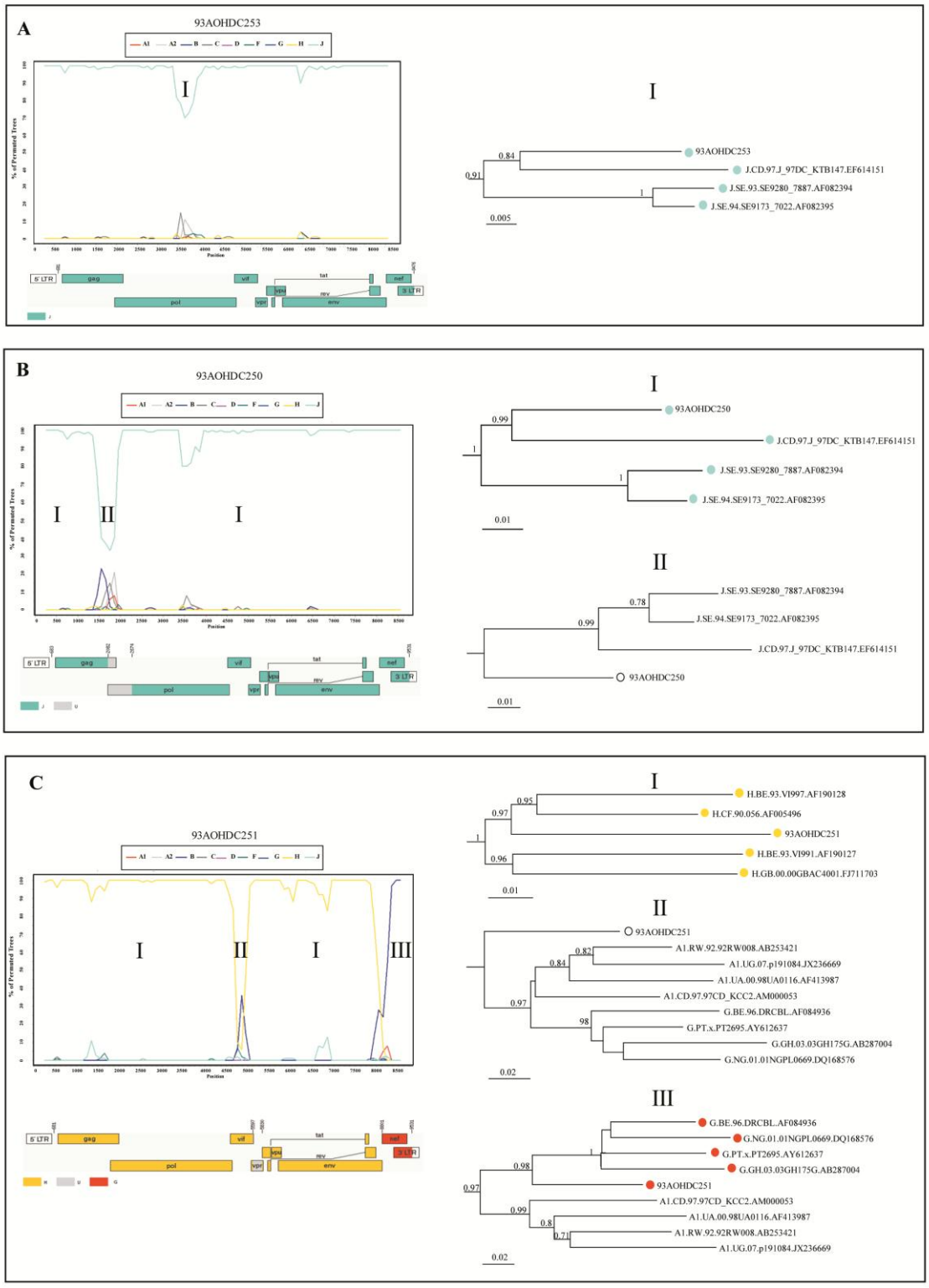

Figure 2 – Genomic segment analysis of near full-length genomes. Panels show bootscanning and tree analyses of isolates 93AOHDC253 (A), 93AOHDC250 (B) and 93AOHDC251 (C). Tree analysis was done for regions suggested by bootscanning to

belong to different subtypes. A schematic genome map of the subtype composition of the isolates was produced based on the results obtained by bootscanning.

Figure 3

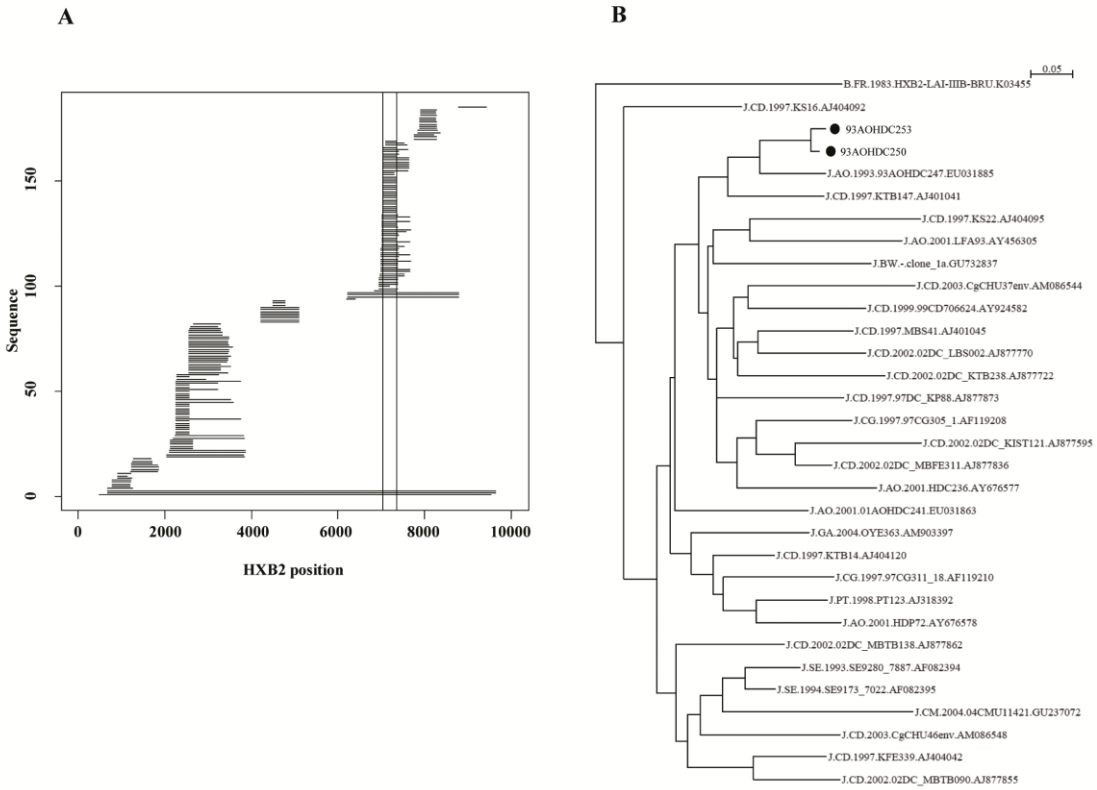

Figure 3 – Subtype J *env* fragment comparison. At the time this report was written 185 subtype J sequence fragments were available in the LANL HIV database. Their HXB2 coordinates covered various parts of the HIV-1 genome (A). Previous to our new

sequences, only 3 full subtype J genomes have been described (long horizontal lines at bottom of graph). The most sequenced region was a part of *env* (HXB2 coordinates 7041 to 7358), indicated by the vertical lines in the graph. A ML phylogeny of this region revealed that subtype J is very diverse (B). Our new sequences are highlighted by black dots.

Figure 4

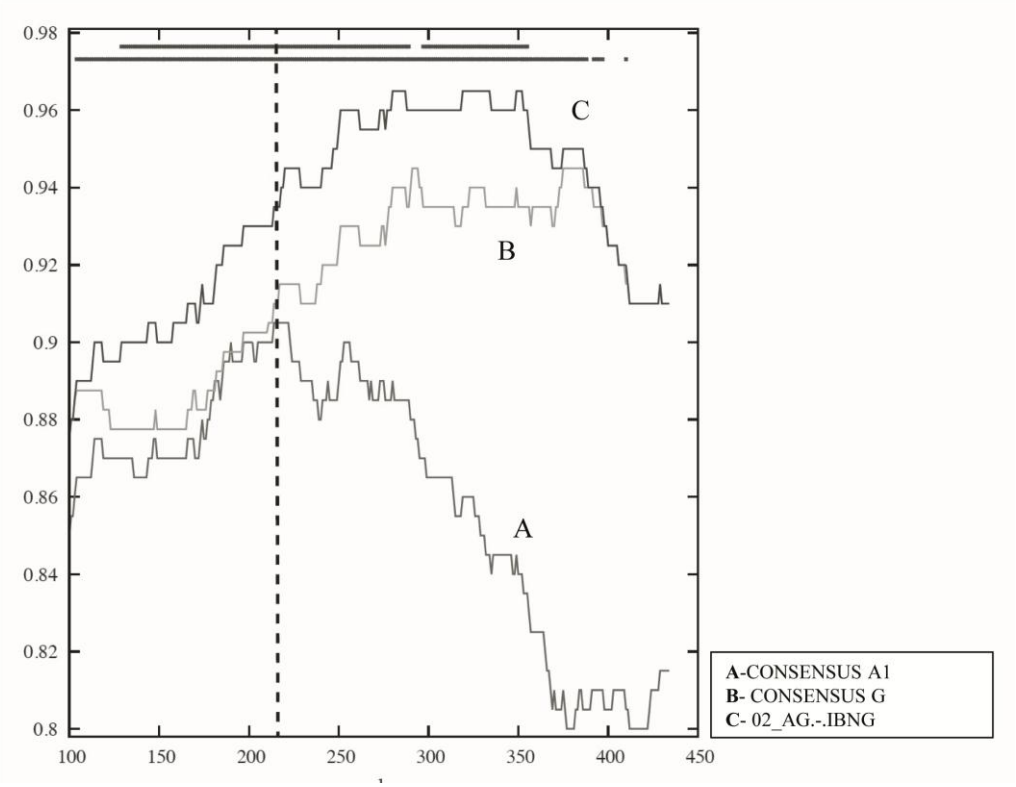

Figure 4 – RIP analysis of segment III of 93AOHDC251. A sliding window analysis (window size 200 nt, step 1 nt; significance threshold 0.90) comparing 93AOHDC251 to consensus subtype sequences A1 and G, and CRF02 prototype sequence IBNG showed that this region was closer to CRF02 in both the A1 and G parts. The CRF02 A1/G breakpoint is indicated by a bashed vertical line. The top bar indicates statistical significance, and the lower bar best match.

Figure 5

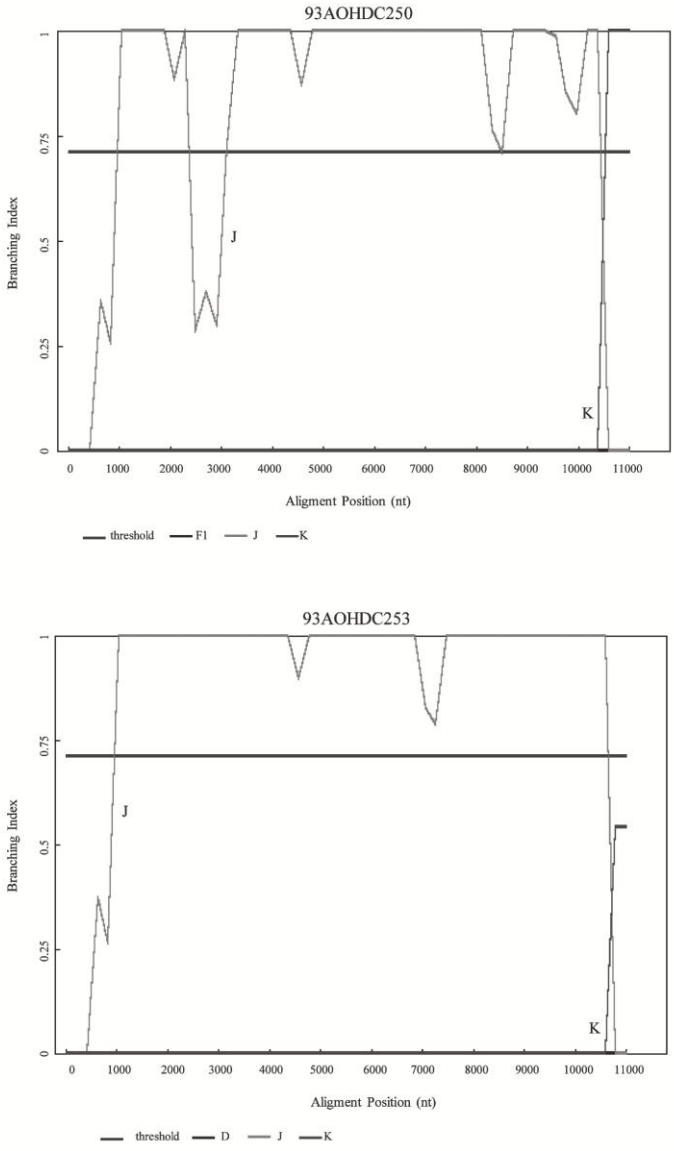

Figure 5 - Branching index (BI) analysis of 93AOHDC250 and 93AOHDC253.

Supplementary material

Table 1- Primers used for polymerase chain reaction amplification of HIV-1 near full-length genomes

NAME	HXB2 POSITION	REGION	SEQUENCE (5´-3´)
JA153	683-707	p17	CTCTCGACGCAGGACTCGGCTTGCT
P17-894F	894-913	p17	ATGGGCAAGCAGGGAGCTGG
P17-913R	913-894	p17	CCAGCTCCCTGCTTGCCCAT
P24-1300F	1300-1324	p24	ATACCCATGTTT(A/T)CAGCATTATCAG
JA155 (R)	1324-1300	p24	CTGATAATGCTGAAAACATGGGTAT
P24-1818F	1818-1838	p24	AGAAGAAATGATGACAGCATG
P24-1838R	1838-1818	p24	CATGCTGTCATCATTTCTTCT
IBPR1.1	2008-2030	gag NC	AAAAGGGCTGTTGGAAATGTGG
IBPR2.2	2733-2712	RT	GCAAATACTGGAGT(A/G)TT(G/A)TATG
IBPR3.1	2119-2140	gag P1-P6	AGGCCAGGGAATTT(T/C)C(T/C)TCAGA
IB2621PR4	2591-2621	PR	AATGCTTTTATTTT(C/T)TCTTCTGTCAATGGC
IBRT1	2487-2505	RT	CCTACACCTGTCAACATAA
IBRT2	3649-3630	RT	TGTTTTACATCATTAGTGTG
IBRT3	2542-2560	RT	TAAATTTTCCAATTAGTCC
IBRT4	3579-3560	RT	TAAATTTGATATGTCCATTG
RT3482FI	3482-3509	RT	AGAACCAGTACATGGRGTATATTATGA
IB3626RT2	3626-3593	RT	TCCGTTAA(C/T)TGT(C/T)TTACATCATTAGTGTG(A/G)GCA
RT3864F	3864-3881	RT	CAACAAATCA(A/G)AAGACTG
P15-3881R	3881-3864	p15	CAGTCTT(T/C)TGATTTGTTG
P15-4176F	4176-4194	p15	GGAGGAAATGAACAAGTAG
P15-4194R	4194-4176	p15	CTACTTGTTCATTTCCTCC
P31-4658F	4658-4675	p31	CAATCCCCAAAGTCAAGG
P31-4675R	4675-4658	p31	CCTTGACTTTGGGGATTG
Vif5041F	5041-5059	vif	ATGGAAAACAGATGGCAGG
Vif5059R	5059-5041	vif	CCTGCCATCTGTTTTCCAT
Vif5461F	5461-5476	vif	AAGGTAGGATC(T/C)(T/C)TAC
Vif5476R	5476-5461	vif	GTA(A/G)(A/G)GATCCTACCTT
PBENV1	5968-5986	REV	CTATGGCAGGAAGAAGCGG
REV5986R	5986-5968	REV	CCGCTTCTTCCTGCCATAG
ENV6223RI	6223-6203	REV	CCACTGTCTTCTGCTCTTTC
PBENV2	6203-6223	REV	GAAAGAGCAGAAGAYAGTGGC
ENV8637FI	8637-8657	gp41	CAGGAACTAAAGAATAGTGC
PBENV4	8797A8817	NEF	TTTTGACCACTTGCCHCCCAT
PBENV3	9036-9016	NEF	AGTCATTGGTCTTARAGGTAC
NEF9532RI	9532-9511	3´LTR	GCGAAAAGCRGCTGCTTATAT