VIRUS EVOLUTION

Virus Evolution, 2017, 3(2): vex029

doi: 10.1093/ve/vex029
Research article

The substitution rate of HIV-1 subtypes: a genomic approach

Juan Ángel Patiño-Galindo and Fernando González-Candelas*,†

Unidad Mixta Infección y Salud Pública FISABIO-Salud Pública/Universitat de València, Institute for Integrative Systems Biology (I2SysBio), CIBERESP, c/Catedratico Jose Beltran, 2, 46980 Paterna, Valencia, Spain

*Corresponding author: E-mail: fernando.gonzalez@uv.es

†http://orcid.org/0000-0002-0879-5798

Abstract

HIV-1M causes most infections in the AIDS pandemic. Its genetic diversity is defined by nine pure subtypes and more than sixty recombinant forms. We have performed a comparative analysis of the evolutionary rate of five pure subtypes (A1, B, C, D, and G) and two circulating recombinant forms (CRF01_AE and CRF02 AG) using data obtained from nearly complete genome coding sequences. Times to the most recent common ancestor (tMRCA) and substitution rates of these HIV genomes, and their genomic partitions, were estimated by Bayesian coalescent analyses. Genomic substitution rate estimates were compared between the HIV-1 datasets analyzed by means of randomization tests. Significant differences in the rate of evolution were found between subtypes, with subtypes C and A1 and CRF01_AE displaying the highest rates. On the other hand, CRF02_AG and subtype D were the slowest evolving types. Using a different molecular clock model for each genomic partition led to more precise tMRCA estimates than when linking the same clock along the HIV genome. Overall, the earliest tMRCA corresponded to subtype A1 (median = 1941, 95% HPD = 1943–55), whereas the most recent tMRCA corresponded to subtype G and CRF01_AE subset 3 (median = 1971, 95% HPD = 1967–75 and median = 1972, 95% HPD = 1970–75, respectively). These results suggest that both biological and epidemiological differences among HIV-1M subtypes are reflected in their evolutionary dynamics. The estimates obtained for tMRCAs and substitution rates provide information that can be used as prior distributions in future Bayesian coalescent analyses of specific HIV-1 subtypes/CRFs and genes.

Key words: HIV-1; BEAST; tMRCA; substitution rate; relaxed molecular clock; Bayesian skyline plot.

1. Background

HIV is a retrovirus of the genus *Lentivirus* and is characterized by a very high genetic diversity. There exist two types of HIV, HIV-1 and HIV-2. The former causes the AIDS pandemics and comprises four phylogenetically distinct groups: M, N, O, and P. Groups N and O are found almost exclusively in West-Central Africa (Hahn et al. 2000). Only two strains from group P have been reported so far, both in Cameroon (Plantier et al. 2009; Vallari et al. 2011). HIV-1 group M is the main driver of the HIV pandemics. Within this group, there exist nine subtypes (denoted as A, B, C, D, F, G, H, J, and K) and at least sixty-one circulating recombinant forms (Kuiken et al. 2012).

High mutation and substitution rates favor the genetic diversity of HIV. These result from three main causes: (1) polymerization errors of the reverse transcriptase (Roberts et al. 1988), (2) genetic recombination that produces viral chimeras (Temin 1993), and (3) an explosive within-host proliferation and a large, and still growing, number of infected persons that lead to very large population sizes (Pennings et al. 2014). Although some of these factors may facilitate a fast pace of evolutionary change

(Moya et al., 2004), other factors act in the opposite direction. For instance, Simon-Loriere et al. (2013) showed that gene overlapping, which affects all genes in the HIV genome, is negatively correlated to the rate of evolution due to a reduction in the number of synonymous substitutions, although it would be less relevant in cases of terminal gene overlaps, which are the predominant type in HIV. Genetic bottlenecks during transmission also act slowing the pace of evolution in this virus, because many mutations accumulated within a host are lost after transmission. Because adaptive changes at the within-host level are lost or reverted after transmission, higher intra-host than inter-host substitution rates of HIV-1 are regularly reported (Alizon & Fraser 2013; Duchêne et al. 2014; Lin et al. 2015). The speed of spread of HIV in an epidemic also influences its substitution rate (Maljkovic Berry et al. 2007). Hence, differences in selective pressures, mutation rates, replication capacity, and/or epidemic dynamics may explain differences in substitution rates among HIV-1 subtypes.

There are important differences in the prevalence of the different subtypes around the world. Subtype C is the most prevalent group of HIV-1, occurring mainly in Africa (which presents the highest diversity of HIV-1) and Asia, accounting for almost 50 per cent of the infections. Subtype B is the main HIV-1 group in Western and Central Europe, the Americas, and Australia. It is also common in different countries of Southeast Asia, Northern Africa, the Middle East, and among South African and Russian men who have sex with men. Subtypes A, D, F, G, H, J, and K display their highest prevalence in Sub-Saharan Africa. It is important to mention the increasing prevalence of circulating recombinant forms, especially CRF01_AE and CRF02_AG, which cause most of the infections in South-East Asia and Western Africa, respectively (Buonaguro et al. 2007).

Differences among subtypes in the intensity of selection have also been reported (Choisy et al. 2004). Its importance on the differential pace at which HIV-1M subtypes evolve has been addressed by analyses of partial genes (Abecasis et al. 2009; Wertheim et al. 2012), thus ignoring the differences in mutation rates and/or selective constraints that are known to exist between genomic regions (Geller et al. 2015).

Here, we present a comparative analysis of the evolution of the main HIV-1 subtypes using Bayesian coalescent reconstructions. The primary goal of our study was to compare the substitution rates of HIV-1 subtypes from a genomic perspective, by using near-full viral coding sequences, which should be more informative for the inference of the substitution rates and diversification dates than the individual genomic regions used so far.

2. Materials and methods

2.1. Datasets

Full coding–region sequences (CDS) were retrieved from the Los Alamos HIV Sequence Database, LANL (http://www.hiv.lanl.gov/), on October 2015. Independent datasets were obtained for subtypes A1, B, C, D and G, and the CRF01_AE and CRF02_AG circulating recombinant forms. Although subtype F1 was initially considered, it was excluded from the study due to the low number of sequences retrieved. The criteria for the selection of these sequences were (1) removal of problematic sequences (defined in LANL as sequences with a high proportion of non-ACTG characters or stop codons, presenting hypermutations, deletions or being either contaminants, synthetic constructs or reverse complements); (2) only one sequence per patient was used; and (3) sequences with large deletions or undetermined regions (>5% of the sequence length) were excluded. We also removed sequences without a known sampling date. In order to exclude recombinant or incorrectly subtyped sequences, the retained sequences were re-subtyped with the Comet HIV-1 (http://comet.retrovirology.lu) and REGAv3 HIV-1 subtyping tools (Pineda-Peña et al. 2013). All the sequences were also analyzed with five recombination detection methods implemented using the RDP4 software, RDP, Geneconv, Bootscan, Maxchi, and Chimaera (Smith 1992; Padidam et al. 1999; Martin & Rybicki 2000; Posada 2002; Martin et al. 2005, 2015). Sequences in which at least one method suggested recombination, with a P value <0.05, were considered for exclusion. In order to remove redundant sequences, alignments of the concatenated sequences were processed with CD-HIT (Huang et al. 2010) using a similarity threshold at 0.98. One sequence from each of the clusters found at this level was retained for further analysis.

Independent alignments of the non-overlapping regions from all genes were obtained, including the region spanning from *vpr* to *vpu*, using MAFFT version 7 ('auto' strategy; Katoh & Standley 2013). Subsequently, regions of poor homology ('gappy' sites) were trimmed with trimAl (Capella-Gutiérrez et al. 2009). The final alignment lengths were *gag*—1,295 nt, *pol*—2,746 nt, *vif*—464 nt, *vpr*-to-*vpu*—743 nt, *env*—2,316 nt, and *nef*—609 nt. Consequently, up to 8,173 nt of the 8,627 nt spanning the HIV-1 CDS were analyzed.

Due to the high number of B, C, and CRF01_AE sequences that fulfilled the selection criteria and the computational limitations associated with the analysis of large genomic datasets, three different subsets (each with $n = 100$ sequences) for each of these HIV-1 groups were generated by random sampling with replacement from the original data. These subsets also allowed to check the robustness of the estimates obtained for these subtypes. To reduce uncertainty in the estimates of evolutionary parameters (Wilkinson et al. 2015), eleven early subtype B sequences, corresponding to samples obtained between 1978 and 1983 (Worobey et al. 2016), were also included in all the subtype B subsets.

2.2. Molecular clock signal analysis

We checked the clock-likeness of each dataset by performing linear regression analyses between the parameters 'root-to-tip divergence' and 'sampling date' with TempEST (Rambaut et al. 2016). For each subtype and CRF, a tree reconstructed with Fasttree2.1 (Price et al. 2010) was used as input, and the root was chosen as the branch that maximized the coefficient of correlation (R), under the assumption of a strict molecular clock.

2.3. Evolutionary analyses

Times to the most recent common ancestor (tMRCA) and genomic substitution rate estimates of each HIV-1 subtype and CRF were obtained by independent Bayesian Markov Chain Monte Carlo (MCMC) coalescent analyses, as implemented in BEAST v1.8.1 (Drummond & Rambaut 2007). Initially, the same partition tree and clock models were applied to all gene regions. All the analyses were performed with the HKY+Γ (four categories) substitution model, combined with either an uncorrelated lognormal relaxed or the strict molecular clock model and three different demographic models (Bayesian Skyline Plot, and exponential or logistic demographic change). The best demographic model was chosen as that with the lowest Akaike's information criterion (AIC) through Markov chain Monte Carlo (MCMC) (AICM) (Baele et al. 2012). We repeated the coalescent analyses using the GTR+Γ model, obtaining identical tMRCA and substitution rate results (data not shown).

For each viral group, we also estimated the tMRCA and substitution rate of each gene partition by repeating the BEAST analyses by assigning a different molecular clock model to each gene.

At least two independent runs of BEAST were performed for each alignment, with MCMC chain lengths ranging between 30×10^6 and 20×10^7 states. Convergence of the estimated parameters was confirmed with Tracer (http://tree.bio.ed.ac.uk/software/tracer/) checking that effective sample sizes were larger than 200 for all the parameters.

Because substitutions in external branches may include recent, deleterious mutations leading to overestimates of the actual substitution rates, we compared the genomic substitution rates in internal and external branches for each subtype/CRF, using a Perl script (available upon request), to estimate the substitution rate (mean substitution rate) parameter independently for internal and external branches. As the estimates for internal and external branches were almost identical (data not shown), tMRCA and substitution rate estimates reported in this work correspond to values estimated from both external and internal branches.

2.4. Pairwise comparisons of genomic substitution rates

We tested whether the genomic substitution rate distributions estimated from BEAST were significantly different among subtypes/CRFs comparing pairwise posterior distributions by means of randomization tests, in which P-values were calculated by counting the number of times that one substitution rate was lower than the other, considering 5,000 tree states chosen randomly (with replacement) of each subtype/CRF. The obtained value (υ) divided by 5,000 (number of comparisons) was considered as the probability that the compared values belong to different distributions (Abecasis et al. 2009). P-values were obtained as $P = 1 - \upsilon$, and were adjusted with the false discovery rate method (Benjamini & Hochberg 1995). These comparisons were performed using an in-house R script available upon request (R Core Team 2014).

3. Results

3.1. Datasets

We initially retrieved 2,399 HIV-1M sequences from LANL. After applying the filtering criteria detailed above, 862 different sequences were kept for further analysis: 96 were subtype A1, 248 subtype B, 234 subtype C, 45 subtype D, 32 subtype G, 177 CRF01_AE, and 30 CRF02_AG. As mentioned in the Materials and methods section, three different random subsets (each of $n = 100$) were obtained for each of the subtypes B and C and CRF01_AE. Information on HIV-1 subtype/CRF, country of origin, sampling year, and accession number of the sequences used in each dataset is provided in Supplementary File S1.

3.2. Molecular clock signal analyses

The clock-like signal present in the analyzed datasets was evaluated by calculating the correlation coefficients (R) between the root-to-tip divergence and sampling date. R ranged between 0.50 (CRF02_AG) and 0.90 (subtype G) (Table 1). The possible existence of over-dispersion of the HIV-1 molecular clock, which could be a major limitation for our comparisons, was rejected by ensuring that plots produced in the linear regression analyses of root-to-tip divergence versus sampling date for the concatenates did not present large dispersed clouds of points around the regression line. Residual mean squared values, which estimate the variance of the rates, were lower than 2×10^{-4} for all the subtypes (Table 1).

Table 1. Molecular clock signal of each HIV-1 dataset analyzed: correlation coefficient (R) and residual mean squared (RMS) value obtained in the root-to-tip divergence versus sampling date correlation analyses.

Dataset	R	RMS
A1	0.65	8.50E-05
B-1	0.88	7.67E-05
B-2	0.88	8.62E-05
B-3	0.82	1.07E-04
C-1	0.62	7.80E-05
C-2	0.55	6.10E-05
C-3	0.59	6.30E-05
D	0.77	2.00E-05
G	0.90	1.40E-05
01_AE-1	0.89	2.50E-05
01_AE-2	0.82	3.00E-05
01_AE-3	0.86	4.60E-05
02_AG	0.50	1.00E-04

3.3. Substitution rate and tMRCA estimates and comparisons

Genomic substitution rate estimates of each HIV-1 subtype and CRF were obtained by Bayesian MCMC coalescent analyses, as implemented in BEAST. The best-fitting demographic and molecular clock models for each HIV-1 subtype/CRF are shown in Table 2, and dated phylogenetic trees obtained from the near full CDS of each dataset under the best-fitting demographic and clock model are shown in Supplementary File S1.

Genomic HIV-1 substitution rates ranged between $1.3 \cdot \times 10^{-3}$ substitutions/site/year (s/s/y) (95% HPD = 0.7–$1.8 \cdot \times 10^{-3}$ s/s/y) for CRF02_AG and $3.5 \cdot \times 10^{-3}$ s/s/y (95% HPD = 2.9–$4.2 \cdot \times 10^{-3}$ s/s/y) for subtype C dataset 2 (Fig. 1A; Table 3). Randomization tests revealed significant inter-subtype differences, with subtypes A and C and CRF01_AE displaying significantly higher substitution rates than CRF02_AG and subtypes D and G. Importantly, no significant intra-subtype differences were found between the random subsets of subtypes B and C and CRF01_AE (Fig. 1B), although no convergence was attained for subset B-3.

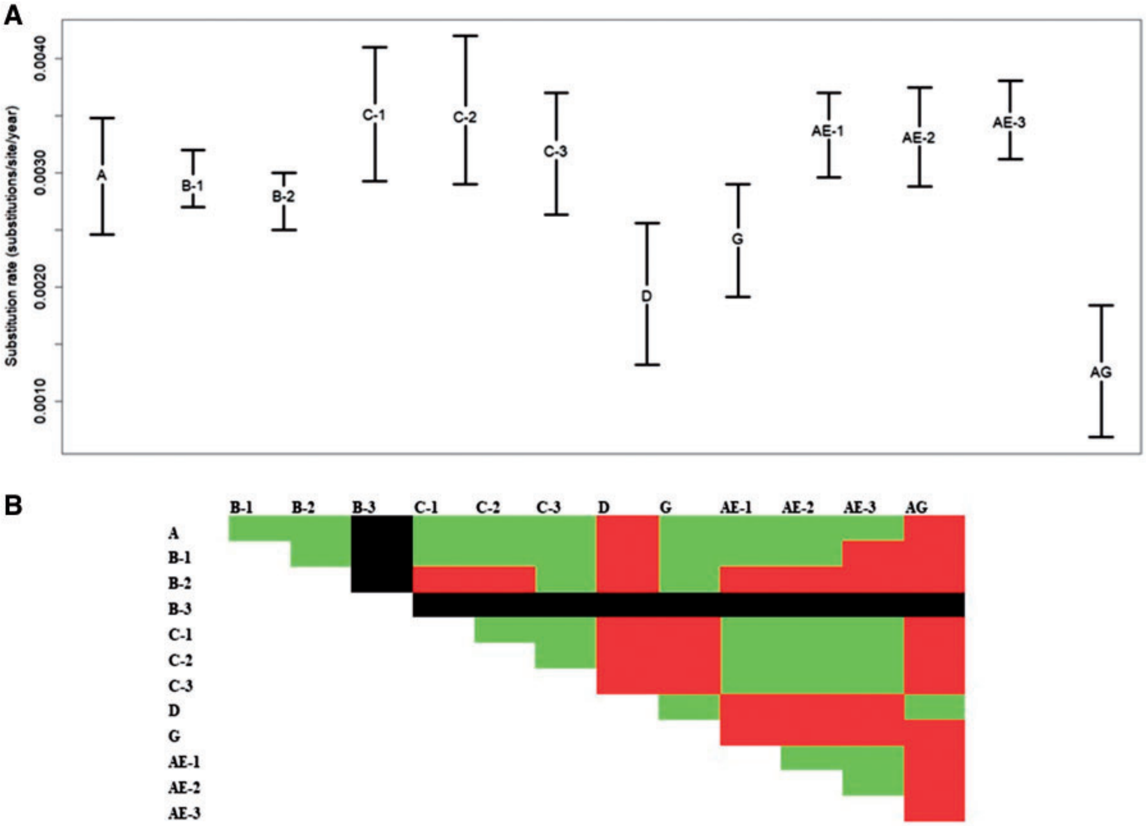
Bayesian coalescent analyses were also performed, unlinking the molecular clock models of the different genomic partitions. Median tMRCAs estimated from this approach were very similar to those obtained when the same clock model was used (largest difference = 6 years, for CRF02_AG). However, 95% HPDs were more precise (narrower) than when applying the same molecular clock model along the whole CDS (Table 3). The 95% HPDs of tMRCAs estimated unlinking the molecular clock models were narrower than 15 years for all datasets, with the only exceptions of CRF02_AG (27 years). However, when the same clock model was applied to the whole CDS, 95% HPDs narrower than 15 years were obtained only for subtypes B (B-1 and B-2 datasets), C (all three datasets), G, and CRF1_AE (all three datasets). Regarding the tMRCA estimates obtained from the different subsets of subtypes B and C and CRF01_AE, the largest difference between medians of the same HIV-1 variant was found for the random subsets C-1 and C-3 (3 years).

Table 2. Akaike's Information Criterion values (AICM) obtained with the three demographic models (under a relaxed molecular clock model) and the strict lock model for each HIV-1 subtype/CRF. Values in brackets represent standard deviation. The best fitting-model is highlighted in black.

	A1	B[a]	C[a]	D	G	CRF01_AE[a]	CRF02_AG
BSP	273,384.5 (0.4)	333,487.0 (0.1)	**323,525.1 (2.3)**	**155,726.5 (0.2)**	**123,944.8 (0.2)**	**230,898.6 (0.6)**	**110,222.6 (0.2)**
Expo	273,353.9 (0.8)	**333,463.2 (0.7)**	323,525.9 (1.5)	155,744.4 (0.2)	123,948.2 (0.2)	230,907.8 (0.6)	110,224.7 (0.2)
Logistic	**273,338.5 (0.5)**	333,466.4 (0.9)	323,528.4 (0.7)	155,733.4 (0.2)	123,946.7 (0.4)	230,925.9 (0.6)	110,224.4 (0.2)
Strict[b]	273,612.1 (0.3)	333,785.6 (0.3)	323,748.2 (0.2)	155,882.3 (0.1)	123,992.4 (0.1)	231,026.6 (0.7)	110,283.0 (0.2)

[a] For subtypes B, C, and CRF01_AE, only the sub-dataset with highest molecular clock signal was subjected to model comparison.

[b] AICM value obtained using the best-fitting demographic model.



	B-1	B-2	B-3	C-1	C-2	C-3	D	G	AE-1	AE-2	AE-3	AG
A	🟩	🟩	⬛	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟥
B-1		🟩	⬛	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟥	🟥
B-2			⬛	🟥	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥
B-3				⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛
C-1					🟩	🟩	🟥	🟥	🟩	🟩	🟩	🟥
C-2						🟩	🟥	🟥	🟩	🟩	🟩	🟥
C-3							🟥	🟥	🟩	🟩	🟩	🟥
D								🟩	🟥	🟥	🟥	🟩
G									🟥	🟥	🟥	🟥
AE-1										🟩	🟩	🟥
AE-2											🟩	🟥
AE-3												🟥

Figure 1. Comparison of the genomic substitution rates and tMRCA estimates of HIV-1 subtypes. (A) Plots of the median and 95% HPD intervals for the substitution rate (mean substitution rate parameter) as obtained with BEAST with the best-fitting demographic and molecular clock models. (B) Pairwise comparisons of the posterior distributions estimated for the substitution rate of each HIV-1 subtype, as obtained with a randomization test. Red: significantly different intervals (P value < 0.05 after FDR correction). Green: not significantly different intervals. Black: not calculated (no convergence in the B-3 dataset).

Overall, the earliest tMRCA corresponded to subtype A1 (median = 1941, 95% HPD = 1943–1955), whereas the most recent tMRCAs corresponded to subtype G and CRF01_AE subset 3 (median = 1971, 95% HPD = 1967–1975 and median = 1972, 95% HPD = 1970–1975, respectively).

In all cases, the substitution rates of the 5′ half of HIV-1 genome (*gag*, *pol*, and *vif*) were lower than those of the 3′ half (*vpr*-to-*vpu*, *env*, and *nef*). Specifically, *pol* presented the lowest substitution rate and *env* the highest in all HIV-1 subtypes/CRFs (Table 3).

4. Discussion

We have estimated and compared the genomic substitution rates of different HIV-1 subtypes (A1, B, C, D, and G) and CRFs (CRF01_AE and CRF02_AG). To obtain representative datasets of the publicly available genomes for each HIV-1 variant, we included sequences from the most complete geographical, temporal, and genetic range as possible and removed epidemiologically related variants, including those obtained from the same patient.

Our analyses were performed using tip-dates of heterochronous samples as the only calibration method. They have revealed differences in the substitution rates between the analyzed subtypes and CRFs. Subtypes C and A1 and CRF01_AE presented the fastest substitution rates among the studied HIV-1 datasets, with CRF02_AG and subtype D being the slowest evolving groups.

As expected, substitution rate estimates for the different subtypes and CRFs differed from previous analyses working only with partial *pol* and/or *env* regions (Abecasis et al. 2009; Wertheim et al. 2012). Abecasis et al. (2009) analyzed partial *pol*

Table 3. Estimates (median and 95% HPD lower and upper limit) for the tMRCA, and for the substitution rates ($\times 10e^{-3}$) of each genomic partition of the HIV-1 datasets analyzed, as obtained using the best-fitting demographic and molecular clock models.

Dataset (time span)	tMRCA (unlinked clock model)	tMRCA (linked clock model)	Rate (CDS)	Rate (*gag*)	Rate (*pol*)	Rate (*vif*)	Rate (*vpr*-to-*vpu*)	Rate (*env*)	Rate (*nef*)
A1 (1985–2011)	1949 (1943–55)	1953 (1944–60)	3.0 (2.5–3.5)	2.3 (2.0–2.6)	1.5 (1.4–1.7)	2.5 (2.1–2.9)	2.7 (2.4–3.1)	4.0 (3.6–4.4)	3.9 (3.4–4.4)
B-1 (1978–2014)	1955 (1951–8)	1957 (1952–62)	2.9 (2.7–3.2)	2.5 (2.3–2.7)	1.6 (1.4–1.7)	2.5 (2.2–2.8)	3.0 (2.6–3.4)	3.8 (3.6–4.1)	3.3 (3.0–3.7)
B-2 (1978–2014)	1954 (1947–59)	1957 (1952–63)	2.8 (2.5–3.0)	1.9 (1.7–2.1)	1.3 (1.2–1.5)	2.2 (2.0–2.5)	2.9 (2.6–3.2)	3.4 (3.1–3.7)	3.6 (3.2–3.9)
B-3 (1978–2014)	No convergence	No convergence	No convergence	No convergence	No convergence	No convergence	No convergence	No convergence	No convergence
C-1 (1989–2011)	1965 (1960–9)	1965 (1958–70)	3.5 (2.9–4.1)	3.5 (3.0–4.0)	1.7 (1.5–1.9)	3.0 (2.5–3.5)	3.5 (3.0–4.0)	4.8 (4.1–5.3)	3.8 (3.3–4.3)
C-2 (1989–2012)	1964 (1959–69)	1965 (1957–71)	3.5 (2.9–4.2)	3.1 (2.7–3.6)	1.7 (1.5–1.9)	2.5 (2.1–3.0)	3.4 (2.9–3.9)	5.0 (4.3–5.7)	4.0 (3.3–4.6)
C-3 (1989–2011)	1961 (1955–66)	1962 (1955–69)	3.2 (2.6–3.7)	3.0 (2.6–3.4)	1.6 (1.4–1.8)	2.3 (1.9–2.7)	3.0 (2.6–3.5)	4.6 (4.0–5.3)	3.5 (3.0–4.1)
D (1983–2011)	1953 (1946–60)	1956 (1927–61)	1.9 (1.3–2.6)	1.9 (1.5-2.2)	1.2 (1.0–1.4)	1.9 (1.5–2.3)	2.8 (2.2–3.5)	3.4 (2.8–4.1)	2.8 (2.3–3.4)
G (1992–2014)	1971 (1967–75)	1969 (1961–74)	2.4 (1.9–2.9)	2.4 (2.0–2.7)	1.6 (1.3–1.8)	2.6 (2.0–3.1)	2.6 (2.2–3.1)	4.2 (3.7–4.8)	3.1 (2.6–3.6)
AE-1 (1990–2012)	1971 (1969–74)	1970 (1967–74)	3.4 (3.0–3.7)	3.1 (2.7–3.4)	1.4 (1.3–1.6)	3.3 (2.9–3.8)	4.0 (3.5–4.4)	4.6 (4.1–5.0)	4.2 (3.7–4.8)
AE-2 (1990–2012)	1971 (1967–74)	1971 (1965–75)	3.3 (2.9–3.8)	3.0 (2.7–3.5)	1.5 (1.3–1.7)	3.4 (2.9–3.9)	4.0 (3.5–4.6)	4.2 (3.7–4.6)	4.3 (3.8–4.8)
AE-3 (1990–2011)	1972 (1970–5)	1972 (1968–75)	3.5 (3.1–3.8)	3.2 (2.9–3.6)	1.5 (1.3–1.6)	4.0 (3.5–4.6)	4.3 (3.7–4.8)	4.6 (4.2–5.0)	4.3 (3.8–4.9)
AG (1991–2012)	1954 (1939–65)	1948 (1913–69)	1.3 (0.7–1.8)	1.2 (0.9–1.5)	0.8 (0.6–1.1)	1.5 (1.0–1.9)	1.6 (1.2–2.1)	2.1 (1.6–2.7)	2.0 (1.4–2.5)

and *env* sequences of up to 799 and 931 nt, respectively, and found that for these two genes subtype G and CRF02_AG had the highest substitution rate and subtype D had the lowest rate. On the other hand, Wertheim et al. (2012) analyzed complete *pol* sequences and found subtype B to be evolving faster than subtypes D and C. These incongruences between different studies can be explained by the different genomic regions analyzed, focusing on the substitution rates of short genomic regions, but ignoring differences in selective constraints or mutation rates that exist along the whole HIV CDS (Geller et al. 2015). This could explain why the genomic substitution rate of the subsets from subtype B were very similar to that estimated recently from complete coding regions by Worobey et al. (2016). The discrepancies between works could also be explained by differences in the size of the analyzed datasets. Although larger datasets have been analyzed previously for CFR02_AG and subtype G (Abecasis et al. 2009), and for subtype D (Wertheim et al. 2012), we have analyzed larger datasets than previous works for the remaining HIV-1 groups.

All the HIV-1 subtypes and CRFs analyzed displayed a similar pattern in the estimated substitution rates along their genomes: genes *gag*, *pol*, and *vif* presented lower rates than the *vpr*-to-*vpu* segment and *env* and *nef* genes, with *pol* and *env* presenting the slowest and fastest substitution rates, respectively. Li et al. (2015) found higher levels of amino acid diversity in the proteins encoded by *tat*, *rev*, *vpu*, *env*, and *nef* genes, and associated their higher levels of variability to different factors. Firstly, a higher variability might be associated with the presence of CD4 T-cell and antibody epitopes, which would favor diversifying selective pressures. Secondly, these proteins were found to present higher numbers of HIV–human associations, which may lead these proteins to present a higher structural flexibility.

Using nearly complete genome coding regions, the 95% HPD intervals obtained for the tMRCA of each subtype and gene were in most cases in agreement with previous estimates (Abecasis et al. 2009; Gray et al. 2009; Yebra et al. 2016). However, tMRCA estimates for subtypes B and C were discordant with respect to those obtained previously (Gilbert et al. 2007; Faria et al. 2014; Worobey et al. 2016). These works estimated the tMRCA of subtype B to have occurred in the 60 s. Faria et al. (2014) also estimated the tMRCA of subtype C to have occurred in the late 30 s. The most plausible reasons for such discrepancies is that the aforementioned studies analyzed older sequences, obtained from samples existing in the geographical locations from which these subtypes initially diversified, such as Kinshasa (Democratic Republic of the Congo) and the Caribbean. Indeed, the tMRCA of our subtype C datasets are more similar to that obtained by Wilkinson et al. (2015) for the origin of the southern African epidemic (median: year 1960), the geographic region from which most of our sequences were obtained. It is also noteworthy that, although the tMRCA that we estimated for subtype B differed from that reported by Worobey et al (2016), the genomic substitution rates (as well as the rates for *gag*, *pol*, and *env*) that they reported are very similar to our estimates. This highlights that, although the tMRCA that we obtained for subtype B may not represent the actual diversification date of this HIV-1 group, the substitution rate estimates that we have obtained are robust.

tMRCA estimated for CRF02_AG presented the broadest 95% HPD among all the HIV subtypes/CRFs analyzed in this work, probably because it was also the dataset with the lowest molecular clock signal. However, the median tMRCA obtained for this CRF was only 8 years older than that estimated by Yebra et al. (2016) in West Africa (median tMRCA between 1962 and 1963 as estimated from PR and gp41, respectively). These tMRCA estimates obtained for CRF02_AG suggest an earlier origin than that of HIV-1 subtype G, which was supposed to be one of its parental subtypes. These estimates support previous results revealing that, indeed, CRF02_AG is the parent of subtype G, which is not an actual pure subtype although it remains classified as one (Abecasis et al. 2007).

Overall, analyzing nearly complete coding regions has produced more precise tMRCA and substitution rate estimates than previous analyses (Abecasis et al. 2009; Gray et al. 2009; Wertheim et al. 2012), especially when different molecular clock models were used for the different gene partitions comprising the CDS. However, it is noteworthy that for some HIV-1 groups (subtypes D and G and CRF02_AG) the number of available genomes was much lower than for the others. This work aimed at obtaining the most representative CDS datasets as possible, and the analyzed datasets represent the currently available genomes in public databases for each HIV-1 subtype/CRF. However, it is possible that the limited number of complete CDS sequences analyzed for subtypes D and G and CRF02_AG may have affected our estimates. Despite this potential caveat, estimating tMRCA and substitution rates from independent datasets for each genome partition would lack the statistical power that confers the information present in the different genome partitions in BEAST. Furthermore, it might introduce another bias, arising from sampling differences between genomic regions from the same HIV subtype/CRF.

Another bias in the estimates may result from the time dependence of the substitution rates. Rates can be very different when analyzing viral datasets from different timescales, with shorter timescales associated with higher rates (Duchêne et al. 2014; Aiewsakun & Katzourakis 2015, 2016). Meyer et al. (2015) assessed the effect of time dependence of the substitution rate of influenza during the 2009 pandemic outbreak, and found that at least 9 months of temporal divergence were necessary to obtain precise estimates for long-term values. In our work, we have used datasets with similar sampling times, which ranged between 21 and 31 years. For this reason, such phenomenon should not introduce a bias in our comparisons.

In conclusion, we have estimated and compared the tMRCAs and genomic substitution rates of the main HIV-1M subtypes and CRFs from a genomic perspective, using the longest possible non-overlapping coding regions. The results obtained show that substitution rates differ significantly among HIV-1 subtypes and CRFs and that the accuracy of the estimated evolutionary parameters increases when independent molecular clock models are applied to each genomic partition. The results obtained provide information that can be used as prior distributions in future Bayesian coalescent analyses of specific HIV-1 subtypes/CRFs and genes, given that the substitution rates of HIV-1 vary among subtypes/CRFs and genomic regions.

Supplementary data

Supplementary data are available at *Virus Evolution* online.

Acknowledgements

This work was supported by Ministerio de Economía y Competitividad (MINECO), Spanish Government (project BFU2014-58656-R), and Generalitat Valenciana (Prometeo/2016/122). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Funding

J.A.P.G. was recipient of a fellowship (FPU-AP2010-0561) from Ministerio de Educación, Cultura y Deporte, Spanish Government. The other authors have declared that they have no conflict of interest.

References

Abecasis, A. B. et al. (2007) 'Recombination Confounds the Early Evolutionary History of Human Immunodeficiency Virus Type 1: Subtype G is A Circulating Recombinant Form', *Journal of Virology*, 81/16: 8543–51.

——, Vandamme, A.-M., and Lemey, P. (2009) 'Quantifying Differences in the Tempo of Human Immunodeficiency Virus Type 1 Subtype Evolution', *Journal of Virology*, 83/24: 12917–24.

Aiewsakun, P., and Katzourakis, A. (2015) 'Time Dependency of Foamy Virus Evolutionary Rate Estimates', *BMC Evolutionary Biology*, 15, 119.

——, and —— (2016) 'Time-Dependent Rate Phenomenon in Viruses', *Journal of Virology*, 90/16: 7184–95.

Alizon, S., and Fraser, C. (2013) 'Within-host and Between-host Evolutionary Rates Across the HIV-1 Genome', *Retrovirology*, 10: 49.

Baele, G. et al. (2012) 'Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty', *Molecular Biology and Evolution*, 29/9: 2157–67.

Benjamini, Y., and Hochberg, Y. (1995) 'Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society B*, 57/1: 289–300.

Buonaguro, L., Tornesello, M. L., and Buonaguro, F. M. (2007) 'Human Immunodeficiency Virus Type 1 Subtype Distribution in the Worldwide Epidemic: Pathogenetic and Therapeutic Implications', *Journal of Virology*, 81/19: 10209–19.

Capella-Gutiérrez, S., Silla-Martínez, J. M., and Gabaldón, T. (2009) 'trimAl: A Tool for Automated Alignment Trimming in Large-scale Phylogenetic Analyses', *Bioinformatics (Oxford, England)*, 25/15: 1972–3.

Choisy, M. et al. (2004) 'Comparative Study of Adaptive Molecular Evolution in Different Human Immunodeficiency Virus Groups and Subtypes', *Journal of Virology*, 78/4: 1962–70.

Drummond, A. J., and Rambaut, A. (2007) 'BEAST: Bayesian Evolutionary Analysis by Sampling Trees', *BMC Evolutionary Biology*, *7*, 214.

Duchêne, S., Holmes, E. C., and Ho, S. Y. W. (2014) 'Analyses of Evolutionary Dynamics in Viruses are Hindered by a Time-dependent Bias in Rate Estimates', *Proceedings. Biological Sciences/the Royal Society*, 281/1786: 20140732.

Faria, N. R. et al. (2014) 'HIV Epidemiology. The Early Spread and Epidemic Ignition of HIV-1 in Human Populations', *Science*, 346/6205: 56–61.

Geller, R. et al. (2015) 'The External Domains of the HIV-1 Envelope are a Mutational Cold Spot', *Nature Communications*, 6: 8571.

Gilbert, M. T. P. et al. (2007) 'The Emergence of HIV/AIDS in the Americas and Beyond', *Proceedings of the National Academy of Sciences of the United States of America*, 104/47: 18566–70.

Gray, R. R. et al. (2009) 'Spatial Phylodynamics of HIV-1 Epidemic Emergence in East Africa', *AIDS (London, England)*, 23/14: F9–F17.

Hahn, B. H. et al. (2000) 'AIDS as a Zoonosis: Scientific and Public Health Implications', *Science (New York, N.Y.)*, 287/5453: 607–14.

Huang, Y. et al. (2010) 'CD-HIT Suite: A Web Server for Clustering and Comparing Biological Sequences', *Bioinformatics (Oxford, England)*, 26/5: 680–2.

Katoh, K., and Standley, D. M. (2013) 'MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability', *Molecular Biology and Evolution*, 30/4: 772–80.

Kuiken, C. L. et al. 2012. *HIV Sequence Compendium 2012*, New Mexico: Theoretical Biology and Niophysics Group, Los Alamos National Laboratory.

Li, G. et al. (2015) 'An Integrated Map of HIV Genome-wide Variation from a Population Perspective', *Retrovirology*, 12: 18.

Lin, Y.-Y. et al. (2015) 'New Insights into the Evolutionary Rate of Hepatitis B Virus at Different Biological Scales', *Journal of Virology*, 89/7: 3512–22.

Maljkovic Berry, I. et al. (2007) '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', *Journal of Virology*, 81/19: 10625–35.

Martin, D., and Rybicki, E. (2000) 'RDP: Detection of Recombination Amongst Aligned Sequences', *Bioinformatics*, 16/6: 562–3.

Martin, D. P. et al. (2005) 'A Modified Bootscan Algorithm for Automated Identification of Recombinant Sequences and Recombination Breakpoints', *AIDS Research and Human Retroviruses*, 21/1: 98–102.

—— et al. (2015) 'RDP4: Detection and Analysis of Recombination Patterns in Virus Genomes', *Virus Evolution*, 1/1: vev003.

Meyer, A. G. et al. (2015) 'Time Dependence of Evolutionary Metrics During the 2009 Pandemic Influenza Virus Outbreak', *Virus Evolution*, 1/1: vev006.

Moya, A., Holmes, E. C., and González-Candelas, F. (2004) 'The Population Genetics and Evolutionary Epidemiology of RNA Viruses', *Nature Reviews Microbiology*, 2/4: 279–88.

Padidam, M., Sawyer, S., and Fauquet, C. M. (1999) 'Possible Emergence of New Geminiviruses by Frequent Recombination', *Virology*, 265/2: 218–25.

Pennings, P. S., Kryazhimskiy, S., and Wakeley, J. (2014) 'Loss and Recovery of Genetic Diversity in Adapting Populations of HIV', *PLoS Genetics*, 10/1: e1004000.

Pineda-Peña, A.-C. et al. (2013) 'Automated Subtyping of HIV-1 Genetic Sequences for Clinical and Surveillance Purposes: Performance Evaluation of the New REGA Version 3 and Seven Other Tools', *Infection, Genetics and Evolution*, 19: 337–48.

Plantier, J.-C. et al. (2009) 'A New Human Immunodeficiency Virus Derived from Gorillas', *Nature Medicine*, 15/8: 871–2.

Posada, D. (2002) 'Evaluation of Methods for Detecting Recombination from DNA Sequences: Empirical Data', *Molecular Biology and Evolution*, 19/5: 708–17.

Price, M. N., Dehal, P. S., and Arkin, A. P. (2010) 'FastTree 2—Approximately Maximum-Likelihood Trees for Large Alignments', *PloS One*, 5/3: e9490.

R Core Team, 2014. *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing). http://www.R-project.org/

Rambaut, A. et al. (2016) 'Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (Formerly Path-O-Gen)', *Virus Evolution*, 2/1: vew007.

Roberts, J. D., Bebenek, K., and Kunkel, T. A. (1988) 'The Accuracy of Reverse Transcriptase from HIV-1', *Science (New York, N.Y.)*, 242/4882: 1171–3.

Simon-Loriere, E., Holmes, E. C., and Pagán, I. (2013) 'The Effect of Gene Overlapping on the Rate of RNA Virus Evolution', *Molecular Biology and Evolution*, 30/8: 1916–28.

Smith, J. M. (1992) 'Analyzing the Mosaic Structure of Genes', *Journal of Molecular Evolution*, 34/2: 126–9.

Temin, H. M. (1993) 'Retrovirus Variation and Reverse Transcription: Abnormal Strand Transfers Result in Retrovirus Genetic Variation', *Proceedings of the National Academy of Sciences of the United States of America*, 90/15: 6900–3.

Vallari, A. et al. (2011) 'Confirmation of Putative HIV-1 Group P in Cameroon', *Journal of Virology*, 85/3: 1403–7.

Wertheim, J. O., Fourment, M., and Kosakovsky Pond, S. L. (2012) 'Inconsistencies in Estimating the Age of HIV-1 Subtypes due to Heterotachy', *Molecular Biology and Evolution*, 29/2: 451–6.

Wilkinson, E., Engelbrecht, S., and de Oliveira, T. (2015) 'History and Origin of the HIV-1 Subtype C Epidemic in South Africa and the Greater Southern African Region', *Scientific Reports*, 5: 16897.

Worobey, M. et al. (2016) '1970s and "Patient 0" HIV-1 Genomes Illuminate Early HIV/AIDS History in North America', *Nature*, 539/7627: 98–101.p

Yebra, G., Kalish, M. L., and Leigh Brown, A. J. (2016) 'Reconstructing the HIV-1 CRF02_AG and CRF06_cpx Epidemics in Burkina Faso and West Africa Using Early Samples', *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 46/1: 209–18.