Inconsistencies in Estimating the Age of HIV-1 Subtypes Due to Heterotachy

Joel O. Wertheim, Mathieu Fourment, and Sergei L. Kosakovsky Pond

1Department of Pathology, University of California, San Diego
2Department of Medicine, University of California, San Diego
†Present address: Program of Emerging Infectious Diseases, Duke-NUS Graduate Medical School, Singapore

*Corresponding author: E-mail: jwertheim@ucsd.edu.
Associate editor: Sudhir Kumar

Abstract

Rate heterogeneity among lineages is a common feature of molecular evolution, and it has long impeded our ability to accurately estimate the age of evolutionary divergence events. The development of relaxed molecular clocks, which model variable substitution rates among lineages, was intended to rectify this problem. Major subtypes of pandemic HIV-1 group M are thought to exemplify closely related lineages with different substitution rates. Here, we report that inferring the time of most recent common ancestor of all these subtypes in a single phylogeny under a single (relaxed) molecular clock produces significantly different dates for many of the subtypes than does analysis of each subtype on its own. We explore various methods to ameliorate this problem. We conclude that current molecular dating methods are inadequate for dealing with this type of substitution rate variation in HIV-1. Through simulation, we show that heterotachy causes root ages to be overestimated.

Key words: molecular clock, rate variation, HIV-1.
previous studies (Worobey et al. 2008; Wertheim and Worobey 2009). We hypothesize that the differences in tMRCA estimates between the subtype-only and combined HIV-1 M analyses resulted from the inability of the relaxed molecular clock to capture the extensive substitution rate variation likely present among the subtypes.

The discrepancy between subtype-specific and HIV-1 M analyses was not restricted to subtype tMRCA estimates. Ho and Lanfear’s (2010) suggestion of modeling rate variation among lineages by partitioning the molecular clock so that each codon position (first, second, and third) has its own relaxed clock produced a substantially better fit [log10 Bayes factor = 59.4], but the inferred subtype tMRCA estimates changed very little (table 1).

The aforementioned analyses employed a Bayesian skyline plot coalescent prior, which has the fewest demographic constraints (Drummond et al. 2005). Fitting alternate, more-restrictive coalescent models (e.g., constant size, expansion growth, exponential growth, and lognormal growth) produced different tMRCA estimates (supplementary table S1, Supplementary Material online); similar results were reported by Worobey et al. (2008). Inference using the alternate models generally estimated younger tMRCA estimates for the subtypes and the root, except for the expansion growth model. However, using alternate coalescent models did not resolve the discrepancy between the subtype-only and combined analyses. Therefore, we fit a model in which each subtype had its own independent exponential growth rate (for a discussion of this approach, see Ho et al. 2008; Bjork et al. 2011). The tMRCA estimates inferred using this hybrid model, with multiple coalescent priors, were more similar to the combined HIV-1 M analysis under a Bayesian skyline plot than the subtype-only analyses (table 1).

Abecasis et al. (2009) hypothesized that the variable substitution rates among HIV-1 M subtypes may be due to different selective pressures operating on them. When we inferred the tMRCA of the combined HIV-1 M phylogeny using a codon model accounting for differing selection pressures among sites (Goldman and Yang 1994; Suchard and Rambaut 2009), the inferred tMRCA estimates were indistinguishable from those of the more simple general time reversible (GTR) + Γ4 nucleotide model (table 1). This is not surprising, given a recent report that modeling lineage-specific rate variation has a much more pronounced effect, than modeling codon substitutions, on tMRCA estimates (Wertheim and Kosakovsky Pond 2011).

Next, we explored the utility of local molecular clocks (Drummond and Suchard 2010) in resolving the discrepancy between tMRCA estimates. The local clock detected 7.6 (5–10) rate changes throughout the HIV-1 M phylogeny, though none of these changes occurred near the base of the subtypes. A slightly faster mean rate was inferred under the local clock: 1.29 × 10^-2 (1.15 × 10^-3–1.15 × 10^-2) substitutions/site/year compared with 1.16 × 10^-2 (9.81 × 10^-4–1.37 × 10^-3) substitutions/site/year in the combined HIV-1 M analysis using an uncorrelated lognormally distributed clock. Correspondingly, the tMRCA estimates inferred using a local clock were younger (table 1).

To determine the expected effect of substitution rate variation across HIV-1 M subtypes on tMRCA inference, we simulated nucleotide sequences across the combined HIV-1 M phylogeny by applying the subtype-specific mean substitution rates. These simulations yielded subtype tMRCA estimates that were even more inconsistent with the
subtype-only tMRCA's than the combined HIV-1 M analysis (table 1). For example, simulated subtypes A1 and F1 each were 10 years older than in the combined HIV-1 M analysis, whereas subtype D was 13 years younger. These simulations support the hypothesis that if extreme rate variation does exist among HIV-1 subtypes, a single relaxed molecular clock cannot adequately correct for it.

To determine if factors other than heterotachy were responsible for the inconsistent subtype ages, we simulated nucleotide sequence alignments under a single rate (i.e., no heterotachy) across an HIV-1 chronogram containing all subtype sequences analyzed here (578 sequences, see supplementary table S2, Supplementary Material online). The tMRCA of the HIV subtypes was inferred for both subtype-only and combined HIV-1 M data sets. No significant differences between subtype tMRCA's were found, suggesting that heterotachy was responsible for the subtype tMRCA inconsistencies.

Finally, we performed a series of general simulations to assess the effect of heterotachy on tMRCA estimation (fig. 3a). When the substitution rate in a single clade was increased relative to the rest of the phylogeny, the tMRCA inferred for that clade was overestimated; if the rate was decreased, the tMRCA for that clade was underestimated (fig. 3b). The tMRCA's inferred for other clades in the phylogeny experienced the opposite trend. Somewhat unexpectedly, more distantly related clades were more biased than closely related clades (see clade B vs. clades C/D in

**Fig. 2.** Posterior distributions of tMRCA estimates for HIV-1 M subtypes. The probabilities that the subtype-only posteriors (blue) are younger than the combined HIV-1 M posteriors (red) are shown.
fig. 3b). Notably, the root age was biased further back in time in the presence of substantial heterotachy, regardless of whether the substitution rate increased or decreased. Relaxed clock analysis was able to detect rate changes in a single clade, but the magnitude of this change was substantially underestimated (data not shown). When two clades were simulated under a slower substitution rate (halved), the tMRCA of those clades was underestimated, clades were simulated under faster rates (doubled), the opposite trend was observed. These patterns were manifested irrespective of the phylogenetic relationships of the clades.

None of the approaches investigated here were able to resolve discrepancies in tMRCA estimation between the subtype-only and the combined HIV-1 M analyses. The implementations of relaxed molecular clocks studied here appear unable to handle the rate variation present in HIV-1 M. When dating the emergence of HIV subtypes and other recent lineages, it seems prudent to use only the clades of interest, as these inferences have generally been more consistent with the historical record. Moreover, given our inability to account for the discrepancies between the subtype tMRCA and the combined HIV-1 M analysis, we feel less confident in the inferred tMRCA of HIV-1 M. And based on our simulations, one might expect the true tMRCA of HIV-1 M to be younger than previously inferred.

### Methods

HIV-1 M subtype data sets were constructed from a non-overlapping polymerase gene region (HXB2 nucleotide positions 2292–5041) from complete genomes with known years of isolation, downloaded from the Los Alamos National Laboratory HIV Sequence Database (www.hiv.lanl.gov; supplementary table S2, Supplementary Material online). Additional A2, F1, and F2 polymerase sequences (>900 nt) were included to ensure robust substitution rate inference. Recombinants were identified using SCUER (Kosakovsky Pond et al. 2009) and removed from subsequent analyses. Alignment was trivial due to the lack of insertions and deletions and was performed manually in Se-Al v2.0 (tree.bio.ed.ac.uk/software/seal/). To create a computationally tractable HIV-1 M data set containing all five subtypes of interest, a maximum of two sequences per sampling year per subtype was included. In all cases, we ensured that sequences from the basal lineage inferred in the subtype-only analyses were included. This procedure resulted in an alignment containing 215 sequences (supplementary table S2, Supplementary Material online). As a control, we investigated the effect of including one or three sequences per sampling year per subtype on tMRCA inference. This effect was negligible for three sequences; analyses sampling one sequence per subtype was included. In all cases, we ensured that sequences from the basal lineage inferred in the subtype-only analyses were included. This procedure resulted in an alignment containing 215 sequences (supplementary table S2, Supplementary Material online). As a control, we investigated the effect of including one or three sequences per sampling year per subtype on tMRCA inference. This effect was negligible for three sequences; analyses sampling one sequence per subtype had greater variance which trended deeper in time, as would be expected. Data sets are available at www.hiv.lanl.gov/wiki/subtyperates.

Phylogenetic and dating analyses were performed using Bayesian Markov chain Monte Carlo implemented in BEAST v1.6.1 (Drummond and Rambaut 2007). For each analysis, two to four chains of 25 or 50 million generations were run; the first 10% of generations were discarded as burn-in, and chains were combined using LogCombiner. Nucleotide-based analyses were performed using a GTR + $\Gamma_4$ substitution model. Codon-based phylogenetic inference using a GY94 + $\Gamma_4$ substitution model was implemented in BEAGLE (Suchard and Rambaut 2009). Tracer v1.5 was used to check for convergence and adequate mixing (i.e., estimated sample size > 200 for relevant parameters).

### Table 1.

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Subtype-Only</th>
<th>Combined HIV-1 M</th>
<th>Codon Partitions (first, second, and third)</th>
<th>Multiple Coalescent Priors</th>
<th>Codon Model (GY94 + $\Gamma_4$)</th>
<th>Local Clock</th>
<th>HIV-1 M Heterotachy Simulations</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1/A2</td>
<td>1943</td>
<td>1925</td>
<td>1927</td>
<td>1924</td>
<td>1926</td>
<td>1934</td>
<td>1910</td>
</tr>
<tr>
<td></td>
<td>1959</td>
<td>1946</td>
<td>1948</td>
<td>1945</td>
<td>1946</td>
<td>1952</td>
<td>1936</td>
</tr>
<tr>
<td></td>
<td>1943</td>
<td>1939</td>
<td>1940</td>
<td>1946</td>
<td>1938</td>
<td>1947</td>
<td>1953</td>
</tr>
<tr>
<td></td>
<td>1934</td>
<td>1946</td>
<td>1947</td>
<td>1947</td>
<td>1952</td>
<td>1952</td>
<td>1953</td>
</tr>
<tr>
<td></td>
<td>1964</td>
<td>1935</td>
<td>1938</td>
<td>1936</td>
<td>1936</td>
<td>1943</td>
<td>1925</td>
</tr>
<tr>
<td></td>
<td>—</td>
<td>1904</td>
<td>1906</td>
<td>1902</td>
<td>1903</td>
<td>1915</td>
<td>1900</td>
</tr>
</tbody>
</table>
HIV-1 heterotachy simulations were performed using the maximum clade credibility tree (chronogram) from the HIV-1 M combined analysis using mean GTR + G4 parameters inferred from the BEAST analysis. Single-rate simulations were performed on a 578-taxon maximum likelihood phylogeny inferred using PhyML v3.0 (Guindon et al. 2010); the chronogram was optimized using TipDate (Rambaut 2000), implemented in HyPhy (Kosakovsky Pond et al. 2005). Nucleotide sequences were simulated using SeqGen v1.3.2 (Rambaut and Grassly 1997). For both sets of HIV simulations, 20 replicate data sets were analyzed in BEAST, using relaxed (lognormal) and strict clocks.

General simulations were performed on a 40-taxon phylogeny comprised of four identical 10-taxon clades (fig. 3a). Alignments of 1,000 nt were simulated under a Hasegawa-Kishino-Yano + G4 model. The substitution rate of $1 \times 10^{-3}$ sites/year was multiplied by a random sample from a lognormal distribution (mean = 0.01; standard deviation = 0.5). To emulate heterotachy, the mean rate within certain clades was varied. For each rate configuration, 50 replicate data sets were analyzed in BEAST.

**Supplementary Material**

Supplementary tables S1 and S2 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgment**

We thank N. Lance Hepler for his assistance with computational resources. This research was supported in part by the National Institutes of Health (AI47745, AI57167, AI74621, and GM093939), the Joint DMS/NIGMS Mathematical Biology Initiative through grant NSF-0714991, the University of California Universitywide AIDS Research Program (grant number IS02-SD-701), a University of California, San Diego Center for AIDS Research/NIAID Developmental Award to S.L.K.P. (AI36214), and a National Institutes of Health Training Fellowship to J.O.W. (AI43638).

**References**


Drummond AJ, Suchard MA. 2010. Bayesian random local clocks, or one rate to rule them all. BMC Biol. 8:114.


**Fig. 3.** General heterotachy simulations. (a) Chronogram on which simulations were performed. The root of the tree has an age of 80 years before present (ybp), and the four major internal clades have an age of 40 ybp. Clades A, B, and C were permitted to experience increases and decreases in mean substitution rates. (b) tMRCA for the root [R] and internal clades (A, B, C, and D) as the mean rate in clade A varies. (c) tMRCA for the internal clades when multiple clades were simulated under a slower ($0.5 \times$) or faster ($2.0 \times$) substitution rate.


