Reconstructing the evolutionary history of human, simian, and prosimian immunodeficiency viruses

Mahan Ghafari[1,3], Peter Simmonds[3], Aris Katzourakis[3,*]

[1]Big Data Institute, Nuffield Department of Medicine, University of Oxford, UK
[2]Peter Medawar Building. Nuffield Department of Medicine, University of Oxford, UK
[3]Department of Biology, University of Oxford, UK
[*]Corresponding author: aris.katzourakis@biology.ox.ac.uk

Summary

While paleovirological evidence supports the idea that the ancestors of modern simian (SIV) and prosimian (pSIV) immunodeficiency viruses have been evolving in nonhuman primates for millions of years, standard molecular clock methods calibrated using contemporary sequences significantly underestimate their times of origin. This discrepancy is attributed to the time-dependent nature of evolutionary rate estimates whereby the rate of virus evolution varies with respect to the time window of its measurement. While biogeographical calibrations may provide better estimates across intermediate timescales, there is currently no unifying framework that can allow inference of lentiviral ages across all relevant timescales. Here we showed that by correcting for such time-dependent rate effects using the recently developed Prisoner-of-War evolutionary model, we can successfully reconstruct the evolutionary history of human, simian, and prosimian immunodeficiency viruses across vastly different timescales and allowed estimates of the age of lineages that gave rise to HIV-1 and HIV-2. The model also predicted that the most recent common ancestor of SIV and pSIV lived around 21 million years ago, suggesting that the most likely explanation for the origins of primate lentiviruses is a terrestrial transfer of lentiviruses between African and Malagasy primates during the last episode of colonisation of Madagascar. Infections have entered via a land bridge in a group of mammals or through a nonprimate vector or transfer mediated by an aerial vector species. We also found that the most recent common ancestor of SIVs lived nearly a million years ago and that some SIV lineages codiverged with their hosts for several hundreds of thousands of years. The predicted long evolutionary timescales of SIVs and potential for close virus/host co-adaptation is consistent with their reduced or minimal pathogenicity in their native hosts, contrasting with the very recent evolutionary origins of highly pathogenic HIV strains in humans.

Introduction

The discovery of the human immunodeficiency viruses types 1 and 2 (HIV-1, HIV-2) in the 1980s [1-3] led to the identification of a larger and diverse group of primate lentivirus known as simian immunodeficiency viruses (SIVs). These close relatives of HIV are distributed in various non-human apes and Old World monkey species in sub-Saharan Africa including chimpanzees, gorillas, and sooty mangabeys. Subsequent to their discovery, there were extensive investigations into the origins of HIV and identification of potential cross-species transmission events from SIVs to humans (see ref. [4] for a recent review). These studies had revealed unequivocally that SIVcpz transmissions from chimpanzee subspecies, *Pan troglodytes troglodytes*, and SIVsmm transmission from sooty mangabeys, *Cercocebus torquatus atys*, were responsible for the emergence of most or all of the HIV-1 and HIV-2 strains respectively in humans [5-7].

Molecular epidemiological studies based on extant SIV and HIV sequences showed a very high rate of nucleotide sequence change over time that predicted a very recent evolutionary origin for primate lentiviruses [4, 8]. Using a molecular clock rate estimate of around 10^{-3} substitutions per nucleotide per year (s/n/y) for HIV and extrapolation of that across the entire primate lentivirus phylogeny led us to estimate that the most recent common ancestor of the entire group lived no earlier than a couple of centuries ago. More advanced molecular clock methods based on relaxed clocks which, to a certain degree, account for rate variation across different taxa, would push the common ancestor of extant SIV sequences in chimpanzee back to a few hundred to 2,000 years before present (BP) [9]. However, all of these estimates based on molecular clock analysis stand in stark contrast to paleovirological evidence from endogenous viral elements (EVEs) that show lentiviruses have infiltrated their hosts' germlines over millions of years. These examples include the rabbit endogenous lentivirus type K which became endogenised in European rabbits about 12 million years ago and two prosimian endogenous lentiviruses, fat-tailed dwarf lemur (pSIVdfl) and grey mouse lemur (pSIVgml), that independently integrated into the lemur's genome nearly 4 million years ago [10, 11]. The remarkable genetic similarity of these integrated virus elements to currently circulating lentiviruses suggests that exogenous viruses have evolved at a much slower rate over time than predicted by standard molecular clock models.

Molecular clock predictions of a recent evolutionary origin for HIV-1 and HIV-2 were originally supported by their initially restricted distribution in Central / West Africa - if either had widely infected humans much earlier than the late nineteenth century, they would have been exported to the Americas and northern Europe via slave trade routes between the sixteenth and nineteenth century and likely persisted in those populations until the present. However, predictions of relatively recent origins for the more divergent strains of SIV and other primate lentiviruses are contradicted by the results of an analysis of strains of SIV infecting four simian species on the island of Bioko. The fauna of this island has been separated from the mainland for around 10,000 years following the rise in sea levels at the end of the last ice age [12]. Therefore, SIV strains infecting these monkeys must have diverged from their mainland relatives for at least this period. However, SIV strains showed much less sequence divergence than would have been expected by extrapolation of short-term substitution rates. Indeed, using the island separation as a biogeographic calibration point, the authors were able to push the estimate of the time to the most recent common ancestor (TMRCA) of SIV strains infecting monkeys and apes in Africa back to nearly 76,000 years BP, with a corresponding clock rate estimate of approximately 10^{-5} s/n/y over this period. However, they also noted that their estimate for the origins of SIV is still likely an underestimate as they found evidence of substitution saturation in their phylogenetic analysis [12].

What the Bioko Island study also highlighted was the failure of standard molecular clock methods to capture the time-dependent changes in the evolutionary rate estimates for SIV across different timescales. While the biogeographic calibration point allowed for the estimation of the divergence of some SIV strains infecting monkeys across intermediate timescales (c.a. tens of thousands of years), the same rate cannot be used to estimate the short-term evolutionary changes in SIV or HIV because the virus is evolving at a much faster rate over shorter timescales (c.a. hundred years or less). The inferred rate from the biographic calibration point is also not appropriate to estimate the origins of SIV over longer timescales because the substitution rate of the virus is likely even lower as we go further back in time. However, so far, no molecular clock method has been able to explain these changes in the substitution rate estimates for SIV across different timescales under a unified evolutionary framework. The time-dependency of evolutionary rate estimates, however, is not unique to primate lentiviruses and

have also been reported in other retroviruses [13, 14]. A time-dependent evolutionary rate creates huge difficulties in reconstructing virus evolutionary timescales, creating a substantial underestimation of the timescale for deeper evolutionary histories of a virus if a shallow calibration point is chosen for the analysis. Conversely, the relatively slow rate estimate based on paeleovirological and biographic calibration points, such as from the Bioko study, creates an unrealistically expanded timescale for the more recent emergence of HIV-1 and other lentiviruses estimated from recent sampling.

Through developing a unified framework for understanding the variation of evolutionary rates in primate lentiviruses over time, we can shed light on their evolutionary history and better understand widespread lentivirus-primate interactions which is particularly relevant for clinically important pathogens such as HIV-1. In this study, we use a mechanistic evolutionary model that takes into account the time-dependent rate effects in viruses [15] to reconstruct the evolutionary history of SIV and estimate the age of SIV lineages that gave rise to HIV-1 and HIV-2. We also estimate the time to the most recent common ancestor of SIV and the pSIV lentiviruses infecting lemurs and discuss how this evolutionary reconstruction may explain the origins of primate lentiviruses.

Results

We inferred phylogenies for 24 species of SIV and every human-to-human transmissible lineage of HIV (HIV-1 groups M, N, and O and HIV-2 groups A and B) and estimated their TMRCA using the prisoner of war (PoW) model which produces time-dependent rates [15]. We also compared the virus phylogenies against their host (see **Methods**) to identify cross-species transmission events and assessed the degree of lentivirus-host codivergence.

In agreement with previous studies [16, 17], we documented a substantial degree of cross-host species transmission of SIV lineages, particularly between SIVs infecting chimpanzee, gorilla, and sooty mangabeys (**Figure 1**). While there was a tendency for more closely related species to exchange SIV strains, there was also evidence of longer term codivergence between some SIV lineages and their hosts [18]. For instance, we found that SIVs infecting talapoin, red-tailed guenon, greater spot-nosed monkey, and moustached guenon may have codiverged with their hosts for more than five hundred thousand years (see **Figure 1**).

We also compared the estimated TMRCA of HIV and SIV strains using the PoW model against standard molecular clock models that either use contemporary HIV/SIV sequences that are less than 20-30 years apart (shallow calibration) [9] or a biogeographic calibration point based on the Bioko Island separation at 10,000 years BP (deep calibration) [12] for rate estimations. We used extant HIV and SIV sequences from the *pol* locus (**Figure 2A**) and SIV sequences from the Bioko Island and other regions of the world that only cover the integrase region (**Figure 2B**). We found that, over shallow timescales of 50 – 300 years, the TMRCA estimates for HIV-1 and HIV-2 groups based on the PoW model and a standard molecular clock method using shallow calibrations [9] were largely compatible with each other, with TMRCA of HIV-1 group M estimated as 279 (95% highest posterior density (HPD): 183 – 426) years BP using the PoW model (**Figure 2A**). However, the TMRCA estimates for SIV lineages diverged significantly depending on whether shallow (**Figure 2A**) or deep (**Figure 2B**) calibrations were used for rate estimation using standard molecular clock models.

For instance, while the TMRCA estimate for the root of SIV lineages in sooty mangabey and chimpanzee are ~200 – 2,000 years BP using the shallow calibration, they are estimated to be approximately 51,500 (95% HPD: 20,600 – 127,300) and 80,200 (15,400 – 166,600) years old, respectively, using the PoW model. Similarly, the standard molecular clock model which used the deep calibration, estimated the SIV root to be ~76,800 years BP while the PoW model estimated the root to be almost 16 times older at 1.2 (95% HPD: 0.3 – 2.1) million years BP.

One of the main implications of having a much older TMRCA estimate for SIV lineages is that it also extends the number of years between the SIVsmm and SIVcpz TMRCAs and the five transmissible HIV strains. We found that the TMRCA of SIVsmm and HIV-2 group A clade and the group B clade were 22,700 (95% HPD: 5,900 – 50,300) and 7,200 (2,140 – 18,200) years BP, respectively. The TMRCA of SIVcpz and the HIV-1 lineages was 11,100 (3,800 – 27,400) years for HIV-1 group O, 3,800 (1,200 – 8,200) years for HIV-1 group M, 690 (230 – 1,700) years for HIV-1 group N. However, the TMRCA of SIV in gorilla (SIVgor) and HIV-1 group O was much younger, 980 (330 – 1,980) years BP, suggesting that it is more likely for gorilla to be the intermediate host between chimpanzee and humans. We also note that all of these TMRCA estimates could underestimate the actual timing of introduction of viral lineages into humans if genetic diversity has since been lost or is not fully sampled and could overestimate it if the

evolutionary rate at the stem branch connecting SIV to HIV lineages were accelerated relative to the expected evolutionary rate of SIV.

The discovery of pSIV strains in strepsirrhine primates on the island of Madagascar provides another opportunity to estimate timescales for the evolutionary history of primate lentiviruses. Madagascar has been isolated from Africa for nearly 160 million years but the existence of lentiviruses infecting lemurs on the island may have arisen from a number of scenarios (see ref. [11] for a detailed discussion of possible transmission routes):

(1) host-virus cospeciation dating back to 85 million years ago when the strepsirrhine-haplorrhine split occurred

(2) transfer of lentiviruses between Malagasy strepsirrhines and African haplorrhines during the last episode of colonisation of Madagascar via a land bridge by a group of mammals or a nonprimate vector nearly 14 million years ago

(3) transfer of lentiviruses via an aerial vector species

While scenario (2) implies there has been an association between haplorrhine and strepsirrhine lentivirus lineages with their hosts for at least 14 million years when the most recent terrestrial mammal invasion occurred [19], scenario (3) suggests there have been two independent spillover events that may have occurred more recently.

To evaluate the likelihood of these three evolutionary scenarios, we used the PoW model to estimate the TMRCA of pSIV/SIV split. In order to estimate this, we needed to account for the fact that the SIV and pSIV samples are not contemporaneous as the pSIV insertion into *M. murinus* germline took place nearly 4.2 million years ago [10] (see **Methods**). Our analysis based on the *pol* locus suggested that the most recent common ancestor of SIV and pSIV lived 21.1 (95% HPD: 16.9 - 26.8) million years BP (**Figure 3**). The ~21 million years estimate for pSIV/SIV split is still much younger than the ~85 million years of possible host-virus cospeciation as proposed in scenario (1). Therefore, it does not support the ancestral codivergence as a likely scenario for the evolution of primate lentiviruses. Instead, the timing of the split is closely aligned with the time to the most recent colonisation of Madagascar via a land bridge or a nonprimate vector around 14 million years ago as described in scenario (2). It is also compatible with scenario (3) whereby an aerial vector such as bats transferred the lentivirus between

Madagascar and the mainland, but this scenario has an extra level of complexity as it requires two independent spillover events which may be less likely to have taken place.

Discussion

Evolutionary reconstructions using the Prisoner of War model provided a unifying framework for understanding the evolutionary history of human, simian, and prosimian immunodeficiency viruses across vast timescales. It also provided evidence, for the first time, that the most recent common ancestor of SIVs existed nearly one million years ago. It successfully reproduced the TMRCA estimates for five human-to-human transmissible strains of HIV and four strains of SIV on the island of Bioko without the need for any external calibrations, giving further credence to its accuracy.

We also answered one of the longstanding questions on the evolutionary history of primate lentiviruses. By finding the TMRCA of SIV and pSIV to be nearly 21 million years old, we were able to rule out ancestral codivergence of primate lentiviruses into haplorrhine and strepsirrhine lineages as a likely model of primate lentivirus evolution. Instead, we found the land-bridge hypothesis whereby a terrestrial mammal or a nonprimate vector transferred lentiviruses between haplorrhine and strepsirrhine in mainland Africa and Madagascar to be a more plausible scenario given that the timing of the most recent colonisation of Madagascar nearly overlaps with our estimate for the TMRCA of SIV and pSIV.

These findings can also offer new insights into the evolution of ancient innate immune defence mechanisms in simian and prosimian hosts against lentivirus infections which could lead to new HIV treatments or vaccines. The tens to hundreds of thousands of years of codivergence between many SIV strains and their simian hosts estimated by the PoW model may provide the opportunity for the host to develop effective immune defence mechanisms that slow or block retroviral replication to non-harmful levels or to modify immune responses to minimise virus-induced immunopathology. Similarities in set point viral loads of minimally pathogenic SIVmm in sooty mangabeys with those of highly pathogenic HIV-1 in humans (or indeed of SIVsmm in the Asian rhesus macaques) [31] may indeed reflect differences in degrees of virus-host adaptation arising from their distinct evolutionary histories and host associations [18].

While our TMRCA estimates for HIV-1 group M, HIV-2 groups A and B were largely compatible with estimates based on a standard molecular clock method, the median TMRCAs were ~50-100 years older. This suggests time-dependent rate effects not only influences SIV but also HIV divergence time estimates over shallower timescales of ~50-300 years. This conclusion is supported by the results of a recent study investigating the evolutionary processes that influence the substitution rate of HIV-1 group M at different timescales. This provided evidence for the existence of site saturation as a result of rapid reversions of sequences to the consensus which can lead to time-dependent rate estimates in HIV [20]. Further investigations into the possibility of a longer evolutionary history of HIV within humans may be warranted.

The thousand-year-old TMRCA estimate for the HIV-1/SIVcpz and HIV-2/SIVsmm splits also imply that humans may have occasionally been exposed to SIVs over much longer timescales. The fact that such potential encounters never led to nascent HIVs becoming established much earlier is likely due to changes in human social behaviour, increased mobility, urbanisation, and various other factors that have only emerged around or after late 19th century [7, 21].

Our findings also exposed the fundamental flaws of conventional methods for rate estimation based on fixed calibration points and their erroneous inference of the SIV evolutionary history. Such calibration techniques do not fully capture time-dependent rate effects across different timescales that may distort inferred evolutionary timescales of primate lentiviruses. We showed that a calibration based on extant SIVs is not reliable for inferring their longer evolutionary trajectory; nor biogeographic or paleovirological calibrations can be used to infer more recent evolutionary histories of SIVs and HIVs.

The problem with using standard substitution models for estimating TMRCA of SIV is twofold. On one hand, using shallow calibration points yield a relatively fast substitution rate of ~10^{-3} s/n/y and result in the underestimation of deeper splits between SIV lineages. On the other hand, using Bioko Island separation to fix the TMRCA of SIV drill lineages in the mainland and island to 10,000 years ago which would correspond to a substitution rate of ~10^{-5} s/n/y may result in the underestimation of more recent splits between lineages. Furthermore, even using the island separation as a calibration point would still underestimate even deeper evolutionary events. However, the PoW model circumvent this by accounting for the time-dependency of substitution rate estimates in a continuous fashion from shallow to intermediate to deep evolutionary timescales without the need for any external calibrations on internal nodes.

The PoW model has previously been applied to reconstruct the phylogeny of foamy viruses which display strong cospeciation with their host and has successfully recovered their deep evolutionary history [15]. The model was also used to reconstruct the evolutionary origins of the sarbecovirus lineage and was found to be in remarkable concordance with signatures of a selection of human genomic datasets that indicate an arms race with corona-like viruses dating back to 25,000 years BP [22], providing an external comparator for this methodology. A greatly expanded timescale for the origins of genotypes of hepatitis C virus to a period before the out-of-Africa migration of modern humans may resolve the otherwise unexplained distinct regional distributions of individual genotypes in different parts of sub-Saharan Africa and South and South Eastern Asia [22]. Given that we have now successfully reconstructed the evolutionary history of primate lentiviruses using this method, we can look much more widely at other viruses, re-evaluate the timescales of their deeper evolution and gain insights into host relationships that are key to understanding their ability to cause disease.

Methods

Sequence data

We generated two sets of alignments. One was for the *pol* locus which included one pSIV and 324 SIV and HIV samples from 24 lineages of SIV, 3 groups of HIV-1 (M, N, and O), and 2 groups of HIV-2 (A and B) of which 134 were time-stamped. The other had 324 SIV samples of integrase gene from 24 lineages of SIV of which 266 were time-stamped and included all the samples taken from the island of Bioko (see ref. [12]). We downloaded all genome sequences from GenBank in 2021 with sampling dates obtained from the Los Alamos National Laboratory HIV sequence database (http://hiv.lanl.gov/content/index) and prepared codon-aligned sequences using MUSCLE [23]. To improve the accuracy of phylogenetic inference, we excluded recombinant samples using BootScanning in the RDP4 package [24] and removed ambiguously aligned regions using GBlocks [25].

The reason for choosing the *pol* locus was because, unlike *env* and *gag* loci, it tends to be functionally more restricted which made the alignment of significantly divergent sequences more reliable. Secondly, the sequenced SIV samples from the island of Bioko only cover the integrase region of *pol*. Therefore, we wanted to investigate whether there is any noticeable difference between the estimated divergence times using the *pol* locus and integrase gene alignments. The other reason was that the pSIV sample only covers the *pol* region [11].

Molecular clock analyses and phylogeny reconstruction

We first examined the temporal signal of the time-stamped samples for each data set using TempEst [26] by plotting the root-to-tip divergence against sampling time based on a maximum likelihood tree which was constructed using IQTREE [27] under a general time-reversible model with a discrete gamma distribution (GTR+ Γ_4). We found that the HIV samples had the strongest clock signal (**Figure S1A**), while SIV samples from the *pol* data set only showed a very weak signal (**Figure S1B**) with possibly wider rate variation and slower rate compared to the HIV samples. The integrase gene data set did not have enough temporal signal to reliably infer a clock rate (**Figure S1C**).

Then we applied a standard strict clock under GTR+ Γ_4 substitution model using BEAST v.1.10 [28] to infer the short-term substitution rate for both HIV and SIV samples from the *pol* data set (**Figure S2**). We first inferred the substitution rate for the HIV samples, 1.1 (95%HPD: 0.9 – 1.3) x 10^{-3} s/n/y, using the CTMC rate prior in BEAST. In the absence of a strong temporal signal for the time-stamped SIV samples, we used a normal rate prior, centred around the median posterior rate for HIV, with mean 1.1×10^{-3} and standard deviation 2×10^{-4} to infer the substitution rate. We found the SIV substitution rate to be slightly lower than that of HIV with a wider posterior distribution, 8.5 (95%HPD: 5.7 – 12) x 10^{-4} s/n/y (see **Figure S2**). To estimate the divergence times, we applied the PoW model which is a mechanistic evolutionary method for estimating virus substitution rates over long evolutionary timescales (method outlined in ref. [15]). To do this, we first constructed ultrametric distance trees for all HIV and SIV samples in both data sets using a standard HKY substitution model and then used the posterior rate distributions of HIV and SIV from the analysis above to rescale the ultrametric distance trees using the PoW transformation and estimated the TMRCA for pSIV, SIV, and HIV phylogenies.

We note that the standard PoW model only allows for the reconstruction of phylogenies based on their short-term substitution rates. However, since the pSIV sample is from a germline insertion that took place nearly 4.2 million years ago, we also needed to account for the extra amount of divergence that pSIV would have accumulated since its insertion until present if it were to evolve at a rate that is similar to that of extant SIV lineages. This allowed us to correctly estimate the total amount of divergence since the most recent common ancestor of pSIV and SIV using the PoW model.

Figures

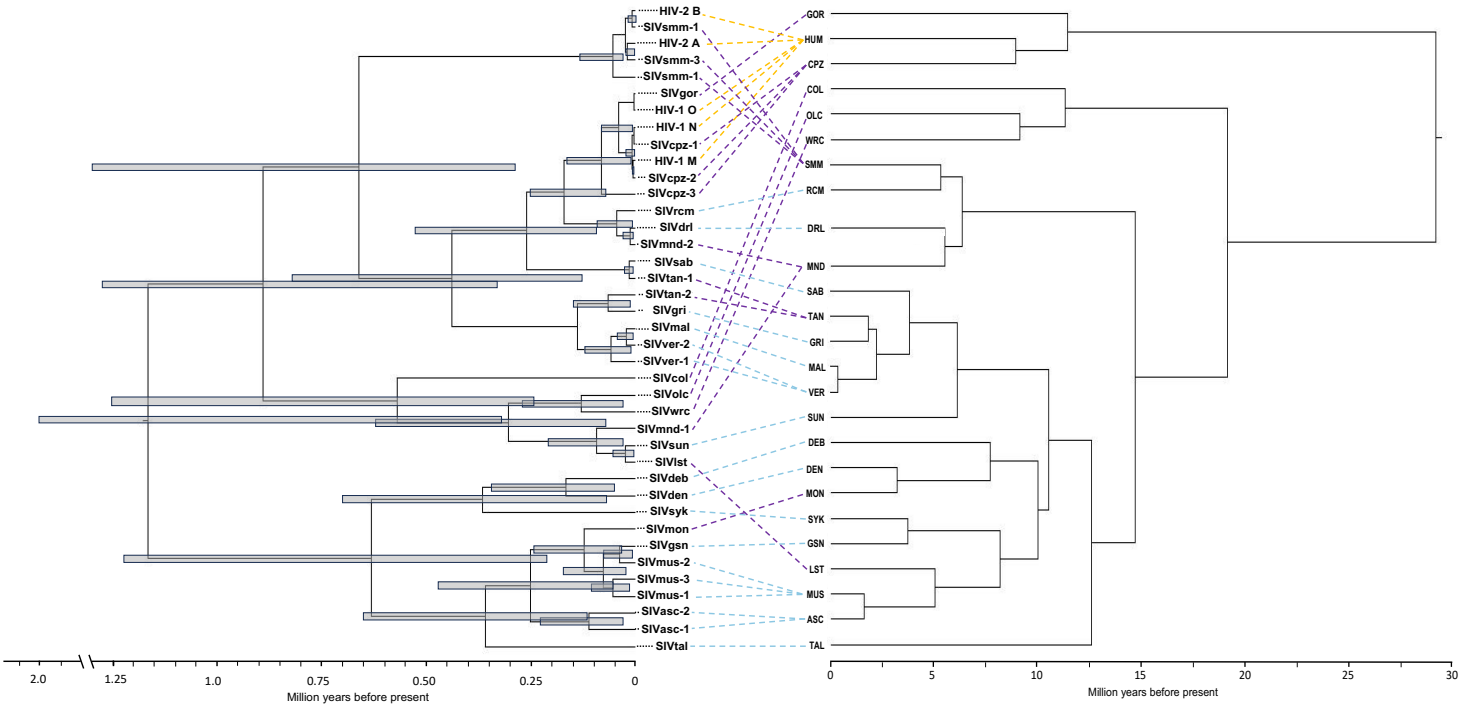
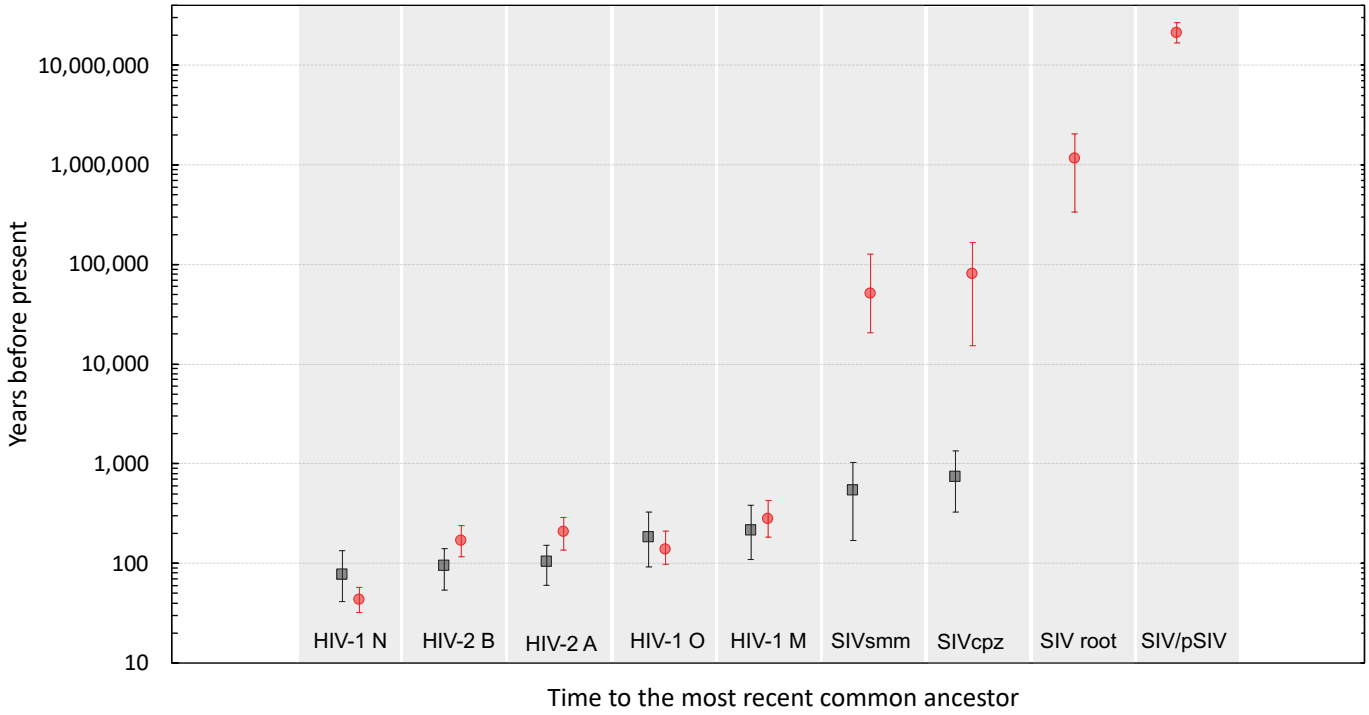


Figure 1: **Tanglegram of rooted phyloegentic trees for SIV and HIV (left) with their simian and human hosts (right).** The phylogenetic association between the PoW-transformed time tree of SIV and HIV lineages (left) is compared to that of their hosts (right, previously published in [29]). Shaded grey rectangles show the 95% highest posterior density. The tranglegram was constructed using TreeMap v3.0 [30] which maximises the congruence between the host and virus phylogenies. Dashed lines are coloured based on whether they show a potential cross-species transmission (purple) or virus-host codivergence (blue). The lines connecting different HIV strains with their human host are highlighted in yellow. Species abbreviations for the SIV lineages are cpz (chimpanzee), gor (gorilla), mnd (mandrill), drl (drill), rcm (red-capped mangabey), sab (sabaeus monkey), tan (tantalus monkey), ver (vervet monkey), mal (malbrouck monkey), gri (grivet monkey), mac (*rhesus macaque*), mne (pig-tailed macaque), smm (sootey mangabey), lst (L'hoest's monkey), sun (sun-tailed monkey), wrc (western red colobus), olc (olive colobus), col (*colobus guereza*), mus (moustached guenon), gsn (greater spot-nosed monkey), mon (mona monkey), asc (red-tailed guenon), tal (northern talapoin), deb (de brazza's monkey), den (dent's mona monkey), and syk (sykes's monkey).

(A)



(B)

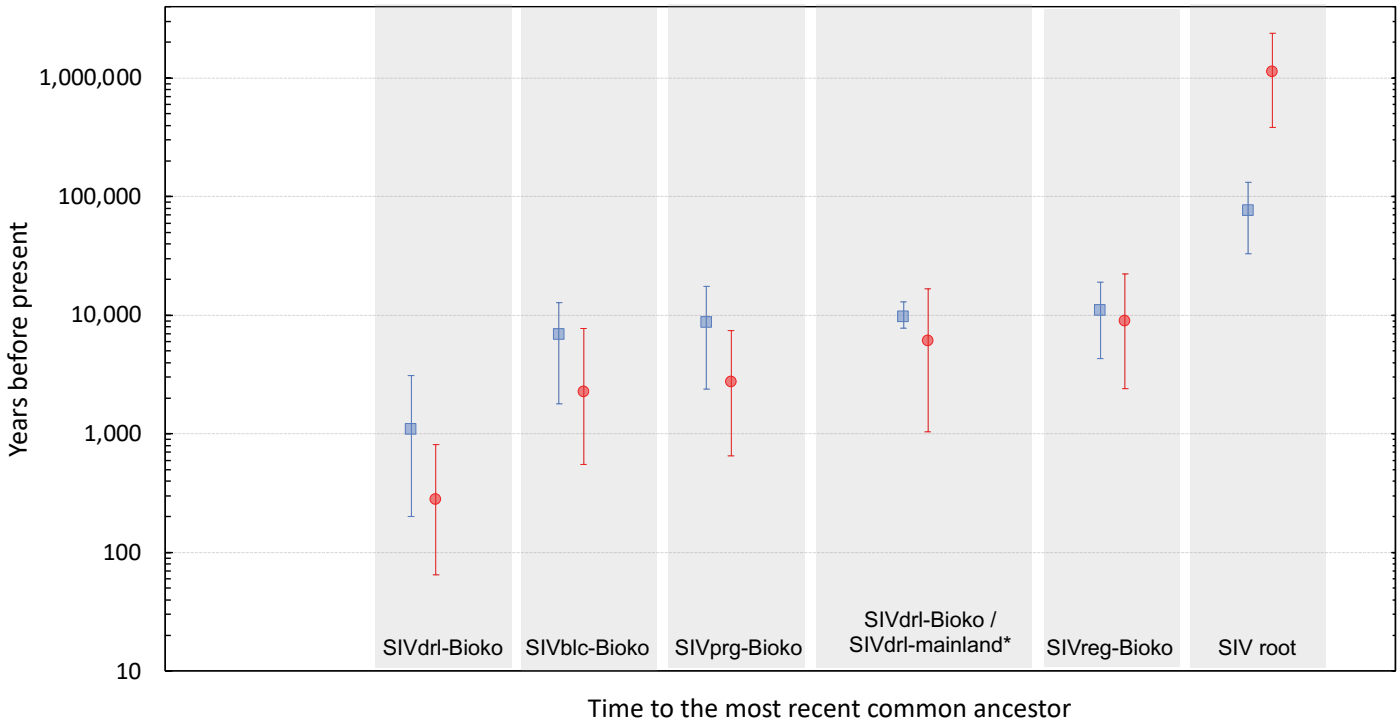


Figure 2: **Estimated divergence times for HIV, SIV, and pSIV strains. (A)** Estimated time to the most recent common ancestor (TMRCA) for HIV-1 and HIV-2 strains, SIV root, pSIV/SIV split, and SIV in chimpanzee (SIVcpz) and sooty mangabeys (SIVsmm) using a standard relaxed clock model (black) [9] and PoW-transformed time tree based on the *pol* dataset (red). **(B)** Estimated TMRCA for the SIV root and various SIV lineages in the island of Bioko using the biogeographic calibration point of island separation (~10,000 years ago) and a standard relaxed clock model (blue)

[12] and PoW model (red) for the integrase gene data set. Vertical lines around each point in both graphs represents the 95% highest posterior distribution (HPD). *The TMRCA estimate for SIV drill in the mainland (SIVdrl-mainland) and SIV drill Bioko (SIVdrl-Bioko) based on the standard molecular clock model (blue) is calibrated such that it matches the separation of isolation of Bioko Island from mainland ~10,000 years ago. Species abbreviations for the SIV lineages are drill (drl), blc (black colobus), preuss's guenon (prg), and red-eared guenon (reg).

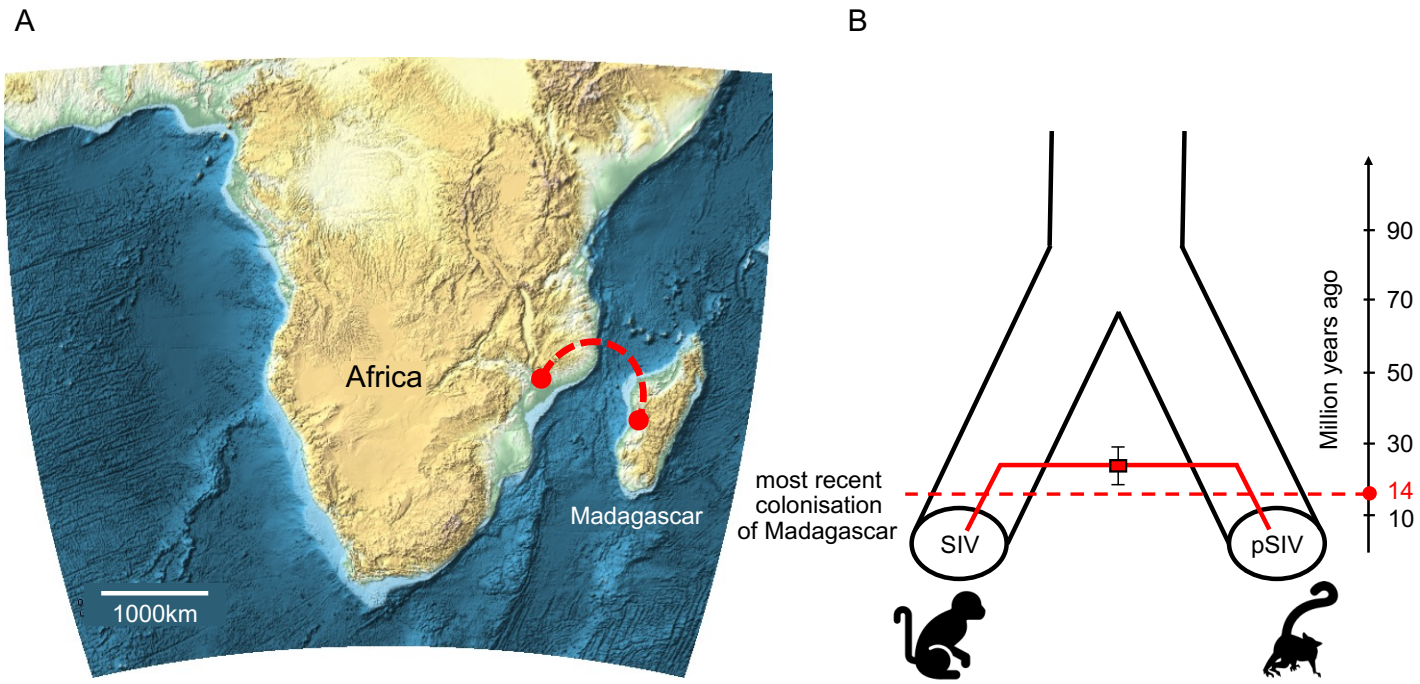
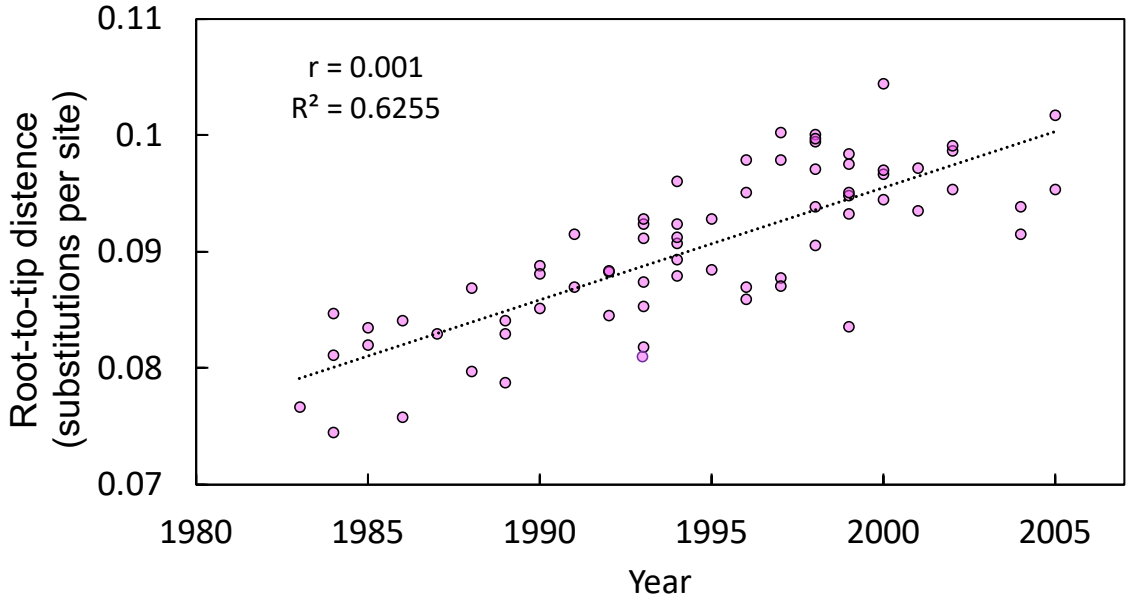


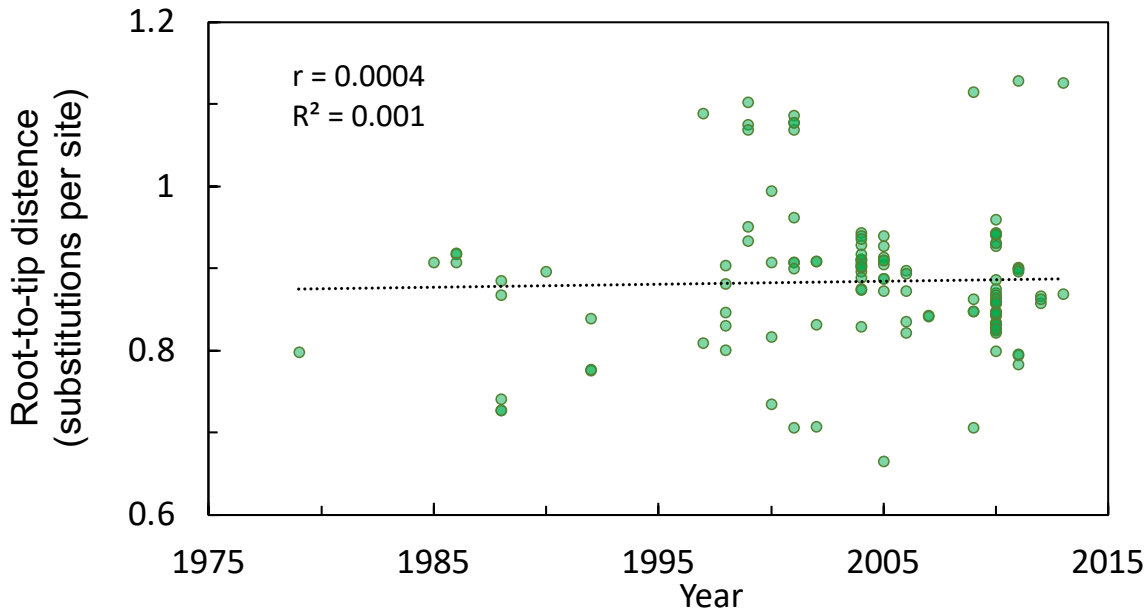
Figure 3: **Evolutionary origins of SIV and pSIV. (A)** Map showing the geographical location of Madagascar relative to Africa and **(B)** the evolutionary timeline of simians and prosimians hosts (black cylindrical tree) along with the SIV/pSIV estimated divergence time according to the PoW model (red solid line). The estimated time to the most recent common ancestor of pSIV and SIV according to the PoW model is 21.1 (95% HPD: 16.9 - 26.8) million years before present which is close to time when the most recent colonisation of Madagascar by primates among other terrestrial mammals took place 14 million years ago (red dashed line).

Supplementary Figures

(A)



(B)



(C)

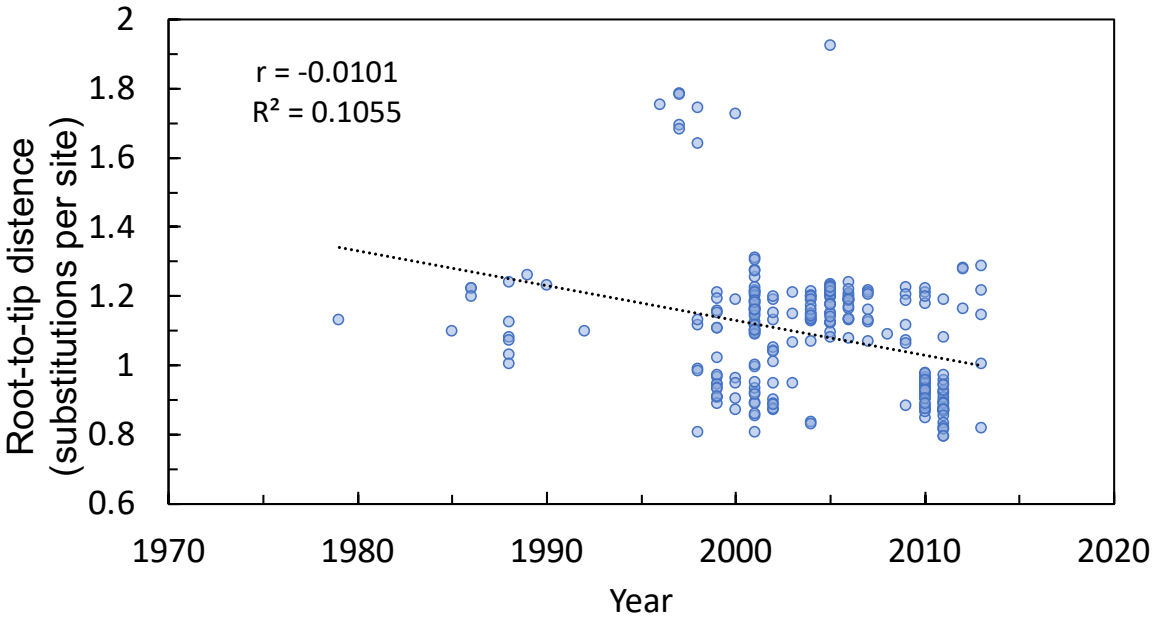


Figure S1: **Evaluating temporal signal in the time-stamped data**. Root-to-tip distance as a function of sampling time for (A) dated HIV-1 group M samples and SIV samples for the (B) *pol* locus and (C) integrase gene data sets. The plots are based on maximum likelihood tree reconstructions with a root position that maximises the residual mean squared for the regression of root-to-tip divergence and sampling time (black dotted line). In each panel, the regression line is shown in black dotted line along with the corresponding slope, r, and coefficient of determination, R^2.

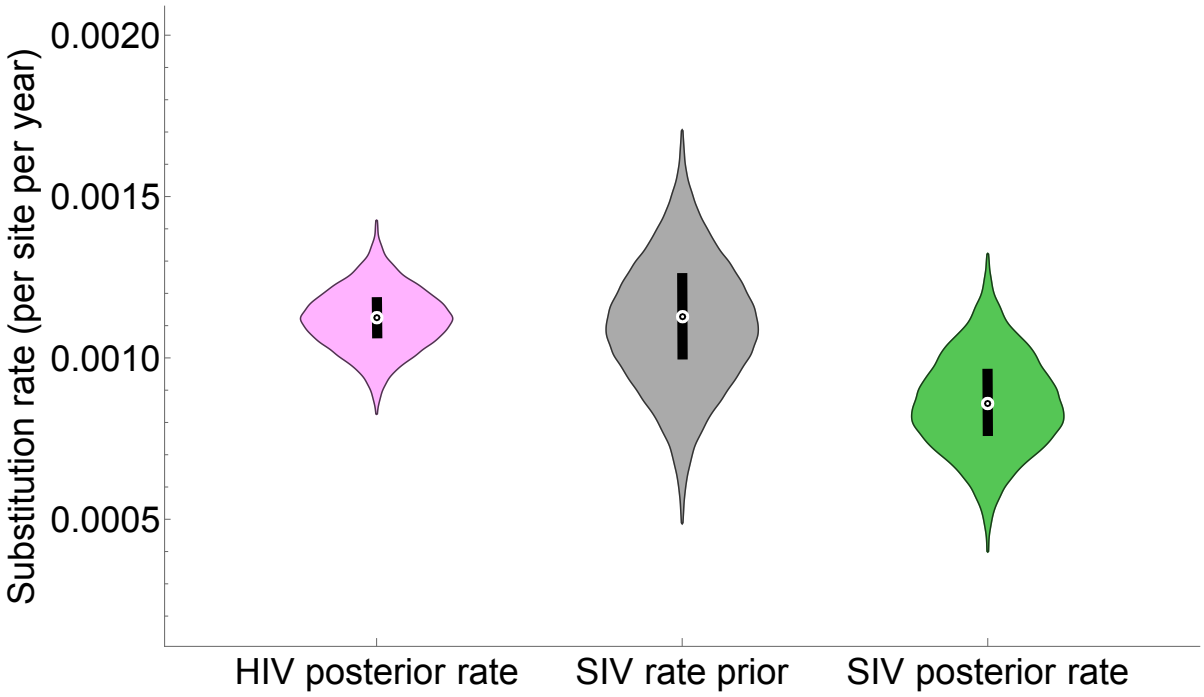


Figure S2: **Prior and posterior substitution rates for HIV and SIV**. Inferred posterior rate distributions for time-stamped HIV and SIV samples obtained from the *pol* locus data set using BEAST v.1.10 [28]. SIV rate prior (normal distribution with mean centred around the median posterior rate for HIV, with mean 1.1×10^{-3} and standard deviation 2×10^{-4} which is subsequently used in the Bayesian inference of SIV posterior rate using BEAST. White circles and vertical black lines represent the median and interquartile range, respectively, for each distribution.

References

1. F. Barresinoussi, et al., *Isolation of a T-Lymphotropic Retrovirus from a Patient at Risk for Acquired Immune-Deficiency Syndrome (AIDS).* Science, 1983. **220**(4599): p. 868-871.
2. R. C. Gallo, et al., *Frequent Detection and Isolation of Cytopathic Retroviruses (Htlv-Iii) from Patients with Aids and at Risk for Aids.* Science, 1984. **224**(4648): p. 500-503.
3. F. Clavel, et al., *Molecular-Cloning and Polymorphism of the Human Immune-Deficiency Virus Type-2.* Nature, 1986. **324**(6098): p. 691-695.
4. P. M. Sharp and B.H. Hahn, *Origins of HIV and the AIDS Pandemic.* Cold Spring Harbor Perspectives in Medicine, 2011. **1**(1): p. a006841.
5. V. M. Hirsch, et al., *An African Primate Lentivirus (SIVsm) Closely Related to HIV-2.* Nature, 1989. **339**(6223): p. 389-392.
6. T. Huet, et al., *Genetic Organization of a Chimpanzee Lentivirus Related to HIV-1.* Nature, 1990. **345**(6273): p. 356-359.
7. B. H. Hahn, et al., *AIDS as a zoonosis: Scientific and public health implications.* Science, 2000. **287**(5453): p. 607-614.
8. B. Korber, et al., *Timing the ancestor of the HIV-1 pandemic strains.* Science, 2000. **288**(5472): p. 1789-1796.
9. J. O. Wertheim and M. Worobey, *Dating the Age of the SIV Lineages That Gave Rise to HIV-1 and HIV-2.* Plos Computational Biology, 2009. **5**(5): p. e1000377.
10. C. Gilbert, et al., *Parallel Germline Infiltration of a Lentivirus in Two Malagasy Lemurs.* Plos Genetics, 2009. **5**(3): p. e1000425.
11. R. J. Gifford, et al., *A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution.* Proceedings of the National Academy of Sciences of the United States of America, 2008. **105**(51): p. 20362-20367.
12. M. Worobey, et al., *Island Biogeography Reveals the Deep History of SIV.* Science, 2010. **329**(5998): p. 1487-1487.
13. P. Aiewsakun and A. Katzourakis, *Time-Dependent Rate Phenomenon in Viruses.* Journal of Virology, 2016. **90**(16): p. 7184-7195.
14. S. Y. W. Ho, et al., *Time dependency of molecular rate estimates and systematic overestimation of recent divergence times.* Molecular Biology and Evolution, 2005. **22**(7): p. 1561-1568.
15. M. Ghafari, et al., *A mechanistic evolutionary model explains the time-dependent pattern of substitution rates in viruses.* Current Biology, 2021. **31**(21): p. 4689-4696.e5.
16. M. A. Charleston and D.L. Robertson, *Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny.* Systematic Biology, 2002. **51**(3): p. 528-535.
17. S. M. Bell and T. Bedford, *Modern-day SIV viral diversity generated by extensive recombination and cross-species transmission.* Plos Pathogens, 2017. **13**(7): p. e1006466.
18. P. M. Sharp, et al., *Origins and evolution of AIDS viruses: estimating the time-scale.* Biochemical Society Transactions, 2000. **28**: p. 275-282.
19. Tattersall, I., *Historical biogeography of the strepsirrhine primates of Madagascar.* Folia Primatol (Basel), 2006. **77**: p. 477-487.

20. V. Druelle and R.A. Neher, *Reversions to consensus are positively selected in HIV-1 and bias substitution rate estimates.* . bioRxiv, 2022. https://doi.org/10.1101/2022.02.13.480259.
21. J. D. de Sousa, et al., *High GUD Incidence in the Early 20(th) Century Created a Particularly Permissive Time Window for the Origin and Initial Spread of Epidemic HIV Strains.* Plos One, 2010. **5**(3): p. e9936.
22. Y. Souilmi, et al., *An ancient coronavirus-like epidemic drove adaptation in East Asians from 25,000 to 5,000 years ago.* Curr Biol, 2021. **31**: p. 1-11.
23. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 2004. **32**(5): p. 1792-1797.
24. D. P. Martin, et al., *RDP4: Detection and analysis of recombination patterns in virus genomes.* Virus Evolution, 2015. **1**(1): p. vev003
25. Castresana, J., *Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.* Molecular Biology and Evolution, 2000. **17**(4): p. 540-552.
26. A. Rambaut, et al., *Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen).* Virus Evolution, 2016. **2**(1).
27. L. T. Nguyen, et al., *IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies.* Molecular Biology and Evolution, 2015. **32**: p. 268-274.
28. M. A. Suchard, et al., *Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10.* Virus Evolution, 2018. **4**(1).
29. S. Alvarez-Carretero, et al., *A species-level timeline of mammal evolution integrating phylogenomic data.* Nature, 2022. **602**(7896): p. 263-267.
30. M. A. Charleston. TreeMap 2011. https://sites.google.com/site/cophylogeny/home.
31. L. E. Pereira and A. A. Ansari, *A case for innate immune effector mechanisms as contributors to disease resistance in SIV-infected sooty mangabeys*. Current HIV Research, 2009. **7**(1): p. 12-22.