RECONSTRUCTING THE EVOLUTIONARY HISTORY OF RNA VIRUSES USING RELAXED MOLECULAR CLOCKS

by

Joel Okrent Wertheim

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ECOLOGY AND EVOLUTIONARY BIOLOGY

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2009

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the dissertation committee, we certify that we have read the dissertation prepared by Joel O. Wertheim entitled Reconstructing the Evolutionary History of RNA Viruses using Relaxed Molecular Clocks and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

_____Date: 8/28/09
Dr. Michael Worobey

_____Date: 8/28/09
Dr. James K. Collins

_____Date: 8/28/09
Dr. Nancy A. Moran

_____Date: 8/28/09
Dr. Michael J. Sanderson

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____Date: 8/28/09
Dissertation Director: Dr. Michael Worobey

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at the University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: Joel O. Wertheim

ACKNOWLEDGMENTS

I would not have been able to complete this dissertation without the support and assistance from others. First, I want to express my gratitude to my advisor, Michael Worobey, for his guidance during my graduate career. He helped focus my research and the work is immeasurably better off for it. I am also grateful for the input from my dissertation committee: Nancy Moran for her insights into evolutionary theory, Jim Collins for helping me see viruses as complete "organisms," and Michael Sanderson for providing unparalleled insights in the nature of phylogenetic inference. I feel fortunate to have been able to interact with each of them.

I am greatly indebted to the Department of Ecology and Evolutionary Biology, which fosters an excellent academic environment. In particular, I would like to thank Matt Herron, Travis Wheeler, Karen Cranston, Tammy Haselkorn, Erin Kelleher, Gaelen Burke, and Kevin Vogel for stimulating conversations about research and science. In addition, Adam Bjork provided invaluable mentorship and helped me figure out how to turn an idea into a published piece of research. I would also like to thank Howard Ochman for his insights into the art of scientific writing.

Finally, I would like to thank my mom, dad, and brother Ira for their support and genuine interest in my work over these past five years. And, to my wife Betsy, whose friendship, conversation, love, patience, knowledge of grammar, and statistical expertise helped me every step of the way, thank you.

DEDICATION

To my wife Betsy for her love and support.

TABLE OF CONTENTS

ABSTRACT

Teasing apart the evolutionary forces responsible for biological phenomena is difficult in the absence of a detailed evolutionary history, especially if this history is lacking a temporal component. RNA viruses, due to their rapid rate of molecular and phenotypic evolution, provide a unique biological system in which to study the temporal aspects of evolutionary processes. These types of studies are possible because of relaxed molecular clock dating techniques, which allow the rate of evolution to vary across a phylogenetic tree. The primary focus of the research presented here concerns the age of the simian immunodeficiency virus (SIV), the primate precursor to HIV. SIV has long been thought to be an ancient infection in non-human African primates, and it has been hypothesized that codivergence with its primate hosts has shaped the SIV phylogeny and resulted in a virus capable of apathogenic infection. The codivergence theory was tested by comparing the phylogeny of a group of monkeys thought to be exemplary of SIV-host codivergence to the phylogeny of their SIVs (Appendix A). These phylogenies were incongruent, suggesting that SIV may have infected these monkeys after their common ancestor speciated. The codivergence theory was investigated further by estimating the time of most recent common ancestor for the SIV lineages that directly gave rise to HIV, found in sooty mangabeys and chimpanzees (Appendix B). The temporal estimates suggest that these SIV lineages are only of hundreds of years old, much younger than expected under the codivergence hypothesis. Next, the same dating techniques were employed to elucidate the evolutionary history of an emerging RNA virus of shrimp, Taura syndrome virus (Appendix C). This analysis provided phylogenetic confirmation

that Taura syndrome virus emerged out of the Americas and spread rapidly around the world. Finally, because all of these studies utilized relaxed molecular clocks, a simulation study was performed to test the hypothesis that relaxed molecular clocks provide higher quality phylogenetic inference compared with traditional time-free phylogenetic inference (Appendix D). This simulation found no difference in the overall quality of phylogenetic inference between these methods.

CHAPTER ONE: INTRODUCTION

An explanation of the problem and a review of the literature. The ability to infer a phylogenetic tree has become an indispensable tool for investigating the evolutionary forces shaping the biology of life on earth. A phylogenetic tree is a hypothesis of the relationships among taxa or genes. Whether or not the actual topology of the tree is of genuine interest depends on the biological question being asked.

The first, and arguably most important, feature apparent in an organismal phylogenetic tree is the hierarchical clustering of taxa into clades, groups in which all members are more closely related to each other than they are to any external taxon. Phylogenetic trees can also be used to test evolutionary hypotheses. The topology of a phylogenetic tree is integral to determining if selection for one phenotype is affected by the presence of a second phenotype (Felsenstein 1985). In addition, comparisons between phylogenetic trees can be used to draw inference about coevolution between different organisms (Mitter and Brooks 1983; Lanyon 1992). Phylogenetics can also be important in the detection of natural selection at the level of genes (Goldman and Yang 1994; Kosakovsky Pond and Frost 2005), even though the precise relationships among taxa may be only a secondary consideration in this instance.

In fact, the shape of the phylogeny is only part of the explanation when answering many important biological questions, such as inferring the age of divergence events. Age estimates are important for determining the pace of and forces behind major ancient radiations, such as the Cambrian explosion (Wray et al. 1996) and the diversification of

mammals (Kumar and Hedges 1998; Bininda-Emonds et al. 2007), and more recent adaptive radiations like the Hawaiian silverswords (Baldwin and Sanderson 1998). In addition, an appreciation of the time between evolutionary splits provides insight into the tempo of evolutionary change, such as the transition from single-cellular to multi-cellular life (Herron et al. 2009), the creation of the adaptive immune system (Janeway 2005), or the adaptations that make humans different from chimpanzees and bonobos (Wilson and Sarich 1969).

Generally, the types of dating estimates described above are possible because of the existence of a molecular clock: the gradual accumulation of substitutions in both amino acid (Zuckerkandl and Pauling 1962) and nucleotide (Miyata et al. 1980) sequences over time. However, the molecular clock does not 'tick' at a uniform rate (Ohta and Kimura 1971; Kumar 2005). Instead, the rate of molecular evolution can vary dramatically across a phylogeny. Although the best way to model this rate variation across lineages is still being investigated (Sanderson 1997; Thorne et al. 1998; Drummond et al. 2006), it is clear already that these so-called 'relaxed' molecular clock methods have expanded the groups of organisms that can be analyzed with dating techniques.

Although many of the early studies using phylogenetic and molecular clock methods relied on macro-organisms, RNA viruses have become an important system in which phylogenetic methods have been utilized and tested. In a remarkable study, researchers were able to reconstruct a known patient-to-patient transmission network using phylogenetic trees inferred from HIV nucleotide sequences (Leitner et al. 1996);

this work demonstrated that a variety of phylogenetic inference methods can recover a known topology using real nucleotide sequences, as opposed to simulated sequence data (Huelsenbeck and Hillis 1993). RNA viruses also allow for the calibration of the molecular clock via a method independent of fossil and geological inference in which the years when viruses were isolated provide sufficient temporal reference and replace the need for *a priori* age estimates on internal nodes (Drummond et al. 2003a). Furthermore, RNA viruses have been used to confirm the accuracy and internal consistency of this calibration technique by incorporating archival viral samples [e.g., influenza A virus (Fitch et al. 1991) and HIV (Korber et al. 2000; Worobey et al. 2008)] and predicting the age of these samples based on the rest of the phylogeny.

Finally, RNA viruses are important to the general study of evolution. Owing to their potential for pathogenicity, they are a potent selective force on the organisms they infect; in fact, RNA virus-related selection may be one of the dominant forces shaping the human genome (Worobey et al. 2007). Regardless of whether or not they qualify as living organisms (see Hegde et al. 2009; Koonin et al. 2009; Moreira and Lopez-Garcia 2009), RNA viruses evolve under many of the same constraints as cellular organisms, and the lessons learned from studying RNA viruses should not be discounted. Furthermore, RNA viruses represent the extreme of many important aspects of genomic evolution. With genomes ranging in size on the order of kilobases, they contain some of the most compact genomes known; in many cases, this has produced overlapping open reading frames and simplified regulation of gene expression. Moreover, the high mutation rate of RNA viruses allows the study of evolution in real-time and makes them

an excellent system for studies of drift and selection in the wild and laboratory settings (Drummond et al. 2003b). Therefore, RNA viruses are ideal for studying the temporal aspects of evolutionary phenomena using phylogenetic inference.

An explanation of the dissertation format. Here, I examine the evolutionary history of emerging RNA viruses using relaxed molecular clocks. In Appendix A, I address the question of codivergence of African green monkeys and their simian immunodeficiency viruses (SIVs). A comparison between their phylogenies suggests that SIV, contrary to popular belief, may not have been infecting the African green monkeys for millions of years (Wertheim and Worobey 2007). Given this evidence suggesting that SIV may be relatively young, I then estimate the time of most recent common ancestor (tMRCA) of the two SIV lineages that gave rise to HIV-1 and HIV-2, presented in Appendix B (Wertheim and Worobey 2009). This analysis indicates that the current diversity of these SIV populations is likely only hundreds of years old, which calls into question the current theory of SIV coevolution towards avirulence over millions of years. In Appendix C, I use similar relaxed molecular clock dating techniques to estimate the time of most recent common ancestor of an RNA virus shrimp pathogen, Taura syndrome virus, which appears to have recently spread across the globe (Wertheim et al. 2009). Finally, much of this work was completed using phylogenetic inference that incorporates a relaxed molecular clock. Therefore, I investigate the quality of relaxed molecular clock phylogenetic inference, compared to the more traditional method of time-free phylogenetic inference, using a simulation study (Appendix D).

CHAPTER TWO: PRESENT STUDY

The background, methods, results, and conclusions of this study are presented in the papers appended to this dissertation. The following is a summary of the most important findings in these papers.

Appendix A examines the SIV-host codivergence hypothesis by comparing the phylogeny of African green monkey mitochondrial genomes to that of their SIV, SIVagm, genomes. This work presents the first well-resolved African green monkey phylogeny. I find that the African green monkey and SIVagm phylogenies are incongruent; the branching order in the SIVagm phylogeny does not, as previously suggested, provide evidence for codivergence between SIVagm and their primate hosts. Furthermore, the SIVagm phylogeny corresponds to the geographic distribution of African green monkeys across sub-Saharan Africa, suggesting a more recent west-to-east pattern of viral spread. Finally, using a relaxed molecular clock, the age of the African green monkey clade is estimated to be around three million years old.

Appendix B also investigates the SIV-host codivergence hypothesis by determining the tMRCA for the SIV lineages that gave rise to HIV-1 and HIV-2 in humans: SIVcpz in chimpanzees and SIVsm in sooty mangabeys, respectively. Using a relaxed molecular clock, the ages of SIVcpz and SIVsm are estimated to be 1492 (1266–1685) and 1809 (1729–1875), respectively. If these tMRCAs represent the time SIV has been infecting these host primates, then these estimates would suggest that SIV was present for only hundreds of years before giving rise to HIV. Notably, the rate of

evolution in SIVsm is indistinguishable from the rate of HIV-2 evolution, indicating that SIV possesses sufficient clock-like signal to estimate tMRCAs. This relaxed clock analysis also provides the first estimate for the tMRCA of HIV-1 group N, a small but distinct clade of HIV-1 found in Cameroon, at 1963 (1948–1977). In concert with the results presented in Appendix A, these findings suggest that the current theory of SIV avirulence arising as a result of millions of years of coevolution between virus and host may need to be reconsidered.

Appendix C describes the emergence of Taura syndrome virus, a highly virulent pathogen of paneid shrimp in aquaculture. Using relaxed molecular clock phylogenetic inference, I find that the most recent common ancestor of Taura syndrome virus around 1991 (1988–1993) in the Americas. The virus then spread rapidly around the world. Taura syndrome's appearance in aquaculture in new countries operates under a quick fuse model; the virus is almost always identified within one year of entering a new country or geographic region. Of note, the Taura syndrome virus phylogeny confirms many previously hypothesized paths of transmission of the virus around the world, based on epidemiological data.

Appendix D addresses the hypothesis that the relaxed molecular clock inference method used in Appendices B and C can produce more accurate and precise phylogenetic inference than traditional time-free phylogenetic inference in a Bayesian framework. I perform a simulation study whose results suggest that there is no qualitative difference between relaxed molecular clock and time-free phylogenetic inference. Although relaxed molecular clock inference models are generally more accurate than time-free inference

(i.e., relaxed clock inference models found the true-tree more often), this tendency is likely because these inference models are less precise (i.e., relaxed clock inference models sampled more unique trees). This finding conforms to the predictions of the bias-variance trade-off: adding more parameters may increase the fit of a model at the expense of explanatory power. This simulation study also finds that phylogenetic inference assuming a strict molecular clock performs extremely poorly on all sequence datasets not generated under a strict molecular clock.

REFERENCES

Baldwin BG, Sanderson MJ (1998) Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proceedings of the National Academy of Sciences of the United States of America 95(16): 9402–9406.

Bininda-Emonds OR, Cardillo M, Jones KE, MacPhee RD, Beck RM et al. (2007) The delayed rise of present-day mammals. Nature 446(7135): 507–512.

Drummond AJ, Pybus OG, Rambaut A (2003a) Inference of viral evolutionary rates from molecular sequences. Advances in Parasitology 54: 331–358.

Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biology 4(5): e88.

Drummond AJ, Pybus OG, Rambaut A, Forsberg R, Rodrigo AG (2003b) Measurably evolving populations. Trends in Ecology and Evolution 18(9): 481–488.

Felsenstein J (1985) Phylogenies and the comparative method. American Naturalist 125: 1–15.

Fitch WM, Leiter JM, Li XQ, Palese P (1991) Positive Darwinian evolution in human influenza A viruses. Proceedings of the National Academy of Sciences of the United States of America 88(10): 4270–4274.

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution 11(5): 725–736.

Hegde NR, Maddur MS, Kaveri SV, Bayry J (2009) Reasons to include viruses in the tree of life. Nature Reviews Microbiology 7(8): 615; author reply 615.

Herron MD, Hackett JD, Aylward FO, Michod RE (2009) Triassic origin and early radiation of multicellular volvocine algae. Proceedings of the National Academy of Sciences of the United States of America 106(9): 3254–3258.

Huelsenbeck JP, Hillis DM (1993) Success of phylogenetic methods in the four-taxon case. Systematic Biology 42(3): 247–264.

Janeway C (2005) Immunobiology : the immune system in health and disease. New York: Garland Science. pp. 665–682.

Koonin EV, Senkevich TG, Dolja VV (2009) Compelling reasons why viruses are relevant for the origin of cells. Nature Reviews Microbiology 7(8): 615; author reply 615.

Korber B, Muldoon M, Theiler J, Gao F, Gupta R et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. Science 288(5472): 1789–1796.

Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular Biology and Evolution 22(5): 1208–1222.

Kumar S (2005) Molecular clocks: four decades of evolution. Nature Reviews 6(8): 654–662.

Kumar S, Hedges SB (1998) A molecular timescale for vertebrate evolution. Nature 392(6679): 917–920.

Lanyon SM (1992) Interspecific brood parasitism in blackbirds (Icterinae): a phylogenetic perspective. Science 255(5040): 77–79.

Leitner T, Escanilla D, Franzen C, Uhlen M, Albert J (1996) Accurate reconstruction of a known HIV-1 transmission history by phylogenetic tree analysis. Proceedings of the National Academy of Sciences of the United States of America 93(20): 10864–10869.

Mitter C, Brooks DR (1983) Phylogenetic aspects of coevolution. In: Futuyma DJ, Slatkin M, editors. Coevolution. Sinauer Associates, Sunderland, Massachusetts.

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proceedings of the National Academy of Sciences of the United States of America 77(12): 7328–7332.

Moreira D, Lopez-Garcia P (2009) Ten reasons to exclude viruses from the tree of life. Nature Reviews Microbiology 7(4): 306–311.

Ohta T, Kimura M (1971) On the constancy of evolutionary rate of cistron. Journal of Molecular Evolution 1: 18–25.

Sanderson MJ (1997) A nonparametric approach to estimating divergence times in the absence of rate constancy Molecular Biology and Evolution 14(12): 1218–1231.

Thorne JL, Kishino H, Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. Molecular Biology and Evolution 15(12): 1647–1657.

Wertheim JO, Worobey M (2007) A challenge to the ancient origin of SIVagm based on African green monkey mitochondrial genomes. PLoS Pathogens 3(7): e95.

Wertheim JO, Worobey M (2009) Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. PLoS Computational Biology 5(5): e1000377.

Wertheim JO, Tang KF, Navarro SA, Lightner DV (2009) A quick fuse and the emergence of Taura syndrome virus. Virology 390(2): 324–329.

Wilson AC, Sarich VM (1969) A molecular time scale for human evolution. Proceedings of the National Academy of Sciences of the United States of America 63(4): 1088–1093.

Worobey M, Bjork A, Wertheim JO (2007) Point, counterpoint: the evolution of pathogenic viruses and their human hosts. Annual Review of Ecology, Evolution, and Systematics 38: 515–540.

Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K et al. (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455(7213): 661–664.

Wray GA, Levinton JS, Sharpiro LH (1996) Molecular evidence for deep precambrian divergences among metazoan phyla. Science 274(5287): 568–573.

Zuckerkandl E, Pauling LB (1962) Molecular disease, evolution, and genetic heterogeneity. In: Kasha MA, Pullman B, editors. Horizons in Biochemistry. New York: Academic Press. pp. 189–225.

APPENDIX A: A CHALLENGE TO THE ANCIENT ORIGIN OF SIV_{AGM} BASED ON AFRICAN GREEN MONKEY MITOCHONDRIAL GENOMES

Published: ***PLoS Pathogens*** **(2007) 3(7): e95**

Co-author: Michael Worobey

ABSTRACT

While the circumstances surrounding the origin and spread of HIV are becoming clearer, the particulars of the origin of simian immunodeficiency virus (SIV) are still unknown. Specifically, the age of SIV, whether it is an ancient or recent infection, has not been resolved. Although many instances of cross-species transmission of SIV have been documented, the similarity between the African green monkey (AGM) and SIVagm phylogenies has long been held as suggestive of ancient codivergence between SIVs and their primate hosts. Here, we present well-resolved phylogenies based on full-length AGM mitochondrial genomes and seven previously published SIVagm genomes; these allowed us to perform the first rigorous phylogenetic test to our knowledge of the hypothesis that SIVagm codiverged with the AGMs. Using the Shimodaira–Hasegawa test, we show that the AGM mitochondrial genomes and SIVagm did not evolve along the same topology. Furthermore, we demonstrate that the SIVagm topology can be explained by a pattern of west-to-east transmission of the virus across existing AGM geographic ranges. Using a relaxed molecular clock, we also provide a date for the most recent common ancestor of the AGMs at approximately 3 million years ago. This study substantially weakens the theory of ancient SIV infection followed by codivergence with its primate hosts.

AUTHOR SUMMARY

Elucidating the factors that influence the emergence of viral pathogens is of great importance to the study of infectious disease. HIV is understood to have originated from simian immunodeficiency viruses (SIVs) infecting nonhuman African primates, but the length of time the virus has been present in these apes and monkeys is not known. These infected primates do not normally develop immunodeficiency, and understanding the age of SIV might help explain why. It has been suggested that some of these monkeys have been infected for millions of years, because many closely related monkey species are infected with closely related viruses. One of the most prominent examples of this relationship is between the African green monkeys and their SIVs. In this study, we compared viral phylogenetic trees to those of their hosts' mitochondrial genomes and found that they do not support the theory of ancient infection followed by codivergence. Our results suggest that SIV did not infect these monkeys until after speciation and subsequently swept across their geographical ranges. If this infection is relatively recent, then avirulence may have evolved over a shorter time frame than previously suggested. This finding could have implications for the future trajectory of HIV disease severity.

INTRODUCTION

More than 30 nonhuman primate species in sub-Saharan Africa are naturally infected with simian immunodeficiency virus (SIV) [1]; however, the evolutionary forces shaping SIV diversity remain unclear. One of the most important unanswered questions regarding SIV evolution is whether it is an ancient infection that has been codiverging with its primate hosts for millions of years, or whether the virus may have arrived more recently and swept across already established primate lineages. Codivergence of viruses with their hosts has been inferred in other cases [2,3], including other retroviruses [4,5], where a close match between the host and viral phylogenetic trees suggests an ancient association. Furthermore, recent genomic analysis suggests that endogenous lentiviruses may have been infecting mammals for the last 7 million years [6]. Although it now seems clear that the overall pattern of the SIV and host phylogenies cannot be reconciled with a simple history of codivergence [7], certain groups of SIVs and their hosts seem to suggest a shared evolutionary history.

Among the SIV taxa, perhaps the best candidate for codivergence is the African green monkey (AGM) clade and their viruses, SIVagm. The AGM genus, *Chlorocebus*, consists of four species (*C. aethiops*, *C. pygerythrus*, *C. sabaeus*, and *C. tantalus*), each with its own corresponding SIV lineage (SIVgri, SIVver, SIVsab, and SIVtan) [8–11]. The monkeys are geographically distributed across sub-Saharan Africa, with *C. sabaeus* in West Africa, *C. tantalus* in central Africa, *C. aethiops* (grivet) in northeastern Africa, and *C. pygerythrus* (vervet) ranging from East to southern Africa [12]. Studies using mitochondrial 12s rRNA have demonstrated monophyly among most AGM species (i.e.,

each individual shares a common ancestor more recently with every member of its own species than with any other AGM species) [13,14]. However, 12s analysis provides very low statistical support for the branching order among the AGM taxa, and these studies were unable to resolve whether *C. pygerythrus* from Tanzania and South Africa are monophyletic or paraphyletic.

On the face of it, the fact that this monophyletic clade of primates is infected by SIVs that also form a monophyletic clade provides compelling evidence of codivergence; however, a degree of caution is warranted whenever such inferences are made. An alternative mechanism by which pathogen and host topologies could resemble each other is preferential host-switching [7]. This model proposes that viruses are more likely to be transmitted between hosts with less phylogenetic distance separating them. This will lead to a viral phylogeny that is similar to the host tree, even in the absence of shared history.

There is ample evidence demonstrating that SIV can switch hosts, with many examples of natural cross-species transmission of SIV among primates. SIVagm has been transmitted to the closely related patas monkey [15] and the more distantly related yellow and chacma baboons [16,17]. Furthermore, two distinct viral lineages infecting chimpanzees (and possibly gorillas) [18,19] and sooty mangabeys [20] have been introduced into the human population at least 11 times, giving rise to HIV [21]. In captivity, SIVagm has been transmitted to the African white-crowned mangabey [22], and SIV from sooty mangabeys has been transmitted to several macaque species [23,24]. The relationships among SIVs are further complicated because many viruses, such as those infecting chimpanzees, sabaeus monkeys, mandrills, and Dent's Mona monkeys,

represent recombinant lineages whose origins must have involved cross-species transmissions of SIV [25–28].

Nevertheless, additional evidence in favor of AGM–SIVagm codivergence has been put forward. The codivergence hypothesis predicts not only that the AGM species will share closely related SIVs, but also that the branching order within the virus clade and monkey clade should match. Such congruence has been reported from an analysis of the AGM CD4 gene [29], which suggested phylogenetic congruence between this nuclear marker and the SIVagm env gene. However, the trees inferred for both virus and host genes were not well supported. Another study involving a nuclear gene, CCR5, which codes for a coreceptor SIV uses to gain entry into host cells, concluded that coevolution between SIV and AGMs had occurred, implying an ancient infection [30].

More generally, the fact that primates naturally infected with SIV do not normally develop immunodeficiency seems to indicate a lengthy host–virus association. Prevalence of the virus in adult AGMs has been documented in excess of 70% [31,32]. Despite continuous viral replication, which can reach titers comparable to those found in humans infected with HIV [33,34], immunodeficiency has only been observed once in an AGM that was co-infected with another retrovirus, STLV-I [35]. On the other hand, SIVagm is lethal when transmitted to non-African host monkeys such as the pigtailed macaque [36,37]. The low virulence observed in the natural host (AGMs), however, does not necessarily indicate millions of years of evolution in response to SIV infection. Fossil evidence and genetic diversity studies propose that the AGM clade is on the order of millions of years old [38,39], whereas molecular clock calculations have inferred a date

of the most recent common ancestor (MRCA) of SIVagm at only hundreds or thousands of years old [40]. Estimates of such a recent origin of SIVagm cannot be dismissed simply on the basis of the observation that SIVagm is relatively benign in its natural hosts.

The purpose of this study was to perform a rigorous phylogenetic test of the hypothesis of ancient codivergence between the AGMs and their SIVs. To do so, we sequenced complete AGM mitochondrial genomes, an approach that has produced what is, to our knowledge, the first statistically well-resolved AGM phylogeny. In comparing this phylogeny to ones inferred from SIVagm genomes, we found that the viral genome topology and host mitochondrial topology were incongruent and therefore provided no support for an ancient infection followed by codivergence.

RESULTS

SIVagm Phylogenies. We constructed the maximum likelihood (ML) SIVagm phylogeny using four previously published SIVagm genomes, one from each named species. The inferred phylogeny placed SIVgri (grivet) and SIVver (vervet) together with high bootstrap support (Figure 1A). Midpoint rooting indicated that SIVsab was the most basal taxon. To test the robustness of our ML topology, we performed the Shimodaira–Hasegawa test (SH-test) [41] on all three possible unrooted SIVagm topologies. Using this conservative test, we were able to reject both alternative SIVagm unrooted topologies ($p < 0.05$) (Table 1).

To ensure that this pattern of SIVgri and SIVver forming a monophyletic clade was consistent for a larger sample of SIVagm strains, we also constructed a phylogeny using all seven available SIVagm genomes plus the complete env gene from two SIVver taxa that were isolated from *C. pygerythrus* in South Africa. We decided to include these subgenomic sequences because the complete SIVver genomes were all isolated from *C. pygerythrus* from East Africa, and we desired a better geographic representation of SIVver samples. Using this dataset, we recovered the same species-specific topology, with SIVgri and SIVver clustering together with strong support and SIVtan falling basal when rooted with SIVsab (Figure 1B). All SIVver taxa form a monophyletic clade.

AGM Nuclear Loci. To determine if available sequence data were sufficient to infer the branching order among the AGM species, we constructed phylogenies using the CD4 and CCR5 genes. Although available 12s rRNA data have proven useful for

differentiating AGM species, they were not sufficient for resolving the phylogeny with statistical confidence. Furthermore, additional nuclear gene data have accumulated recently but have not yet been subjected to phylogenetic analysis. Despite earlier studies with fewer sequences, which seemed to determine the AGM topology, our results with the most complete alignments of nuclear gene sequences indicated that coding nuclear loci do not sufficiently resolve the AGM phylogeny. According to the CD4 topology, AGM species are not reciprocally monophyletic (Figure 2A). There is low bootstrap support across the entire CD4 tree. We were also unable to resolve the branching order using CCR5 (Figure 2B). All AGM species for which more than one CCR5 allele was analyzed exhibited paraphyly. Moreover, the only CCR5 allele from *C. tantalus* is identical to one of the C. sabaeus alleles, implying that CCR5 is not useful in distinguishing AGM species, let alone their phylogenetic relationships.

Mitochondrial Phylogenies. To generate a sequence alignment likely to have sufficient phylogenetic signal to resolve the AGM phylogeny with a high degree of confidence, we sequenced complete mitochondrial genomes—an approach that has yielded robust phylogenies for other primates [42]—for *C. sabaeus*, *C. tantalus*, and *C. pygerythrus* from Tanzania and South Africa. Using an ML framework, we constructed a phylogeny comprised of these four genomes plus the previously published *C. aethiops* and *C. sabaeus* mitochondrial genomes. We inferred a single best topology that placed *C. aethiops* and *C. tantalus* together with high bootstrap support (Figure 3); however, we were unable to resolve the phylogenetic relationship between the two *C. pygerythrus*

taxa. While the ML tree indicated these two taxa are paraphyletic, with the taxon from South Africa branching off before the one from Tanzania, there is low bootstrap support for this inference. Of interest, the corrected genetic distance (GTR + Γ_4) between the two *C. pygerythrus* taxa was greater than that between *C. tantalus* and *C. aethiops*. Both midpoint and outgroup rooting using additional mitochondrial genomes (Figure 4) placed *C. sabaeus* as the most basal AGM taxon. A topology identical to the one inferred via ML in PAUP* was also inferred in a Bayesian framework. This tree placed *C. aethiops* and *C. tantalus* together with a posterior probability of 1.0.

We then compared the unrooted AGM topologies using the SH-test, which rejected all AGM mitochondrial topologies that do not place *C. tantalus* with *C. aethiops* and *C. sabaeus* with *C. pygerythrus* (Table 1). However, we were unable to reject either of the alternate arrangements within *C. pygerythrus*, in which the Tanzanian *C. pygerythrus* branched first before the taxon from South Africa, or where the two *C. pygerythrus* taxa formed a monophyletic clade.

Dating the AGM Origin and Radiation. Using a penalized likelihood approach [43], we estimated the age of both the AGM MRCA and the subsequent radiation of *C. aethiops*, *C. tantalus*, and *C. pygerythrus* (Figure 4). This analysis was performed using trees generated under ML and Bayesian Markov-chain Monte Carlo (MCMC) frameworks. According to the ML analysis, the AGM lineages shared an MRCA 2.81 ± 0.35 million years ago (MYA); *C. aethiops*, *C. tantalus*, and *C. pygerythrus* shared an MRCA 1.48 ± 0.16 MYA. Estimates from the MCMC analysis did not differ

considerably, placing the AGM common ancestor at 2.76 ± 0.23 MYA and the radiation at 1.59 ± 0.14 MYA. Because uniform branching order among the *C. aethiops*, *C. tantalus*, and two *C. pygerythrus* lineages was not observed in the ML analysis, we were unable to estimate the date of divergence events among these species. Our dates of other divergence events among the catarrhine species did not differ appreciably from those presented by Raaum et al. [42] and are therefore not reported here.

Test of Phylogenetic Congruence. As an explicit test for host–viral phylogenetic congruence, which, if present, would be a strong indication of codivergence, we used a series of SH-tests to determine if the SIVagm and AGM mitochondrial phylogenies were significantly different from each other (Table 1). First, we compared the ML SIVagm topology to the SIVagm topology that corresponded to the ML AGM mitochondrial topology (labeled footnote b in Table 1). The SH-test on the SIVagm genomes rejected this alternate topology ($p < 0.05$). Hence, there is convincing evidence that the SIVagm genomes did not evolve along the same topology as the AGM mitochondrial genomes. Of note, the test also rejected the alternate SIVagm topologies when the first 3,500 bases, the recombinant region in SIVsab, were included ($p < 0.05$); thus, these results are not affected by the recombinant origin of SIVsab.

Finally, we compared the ML AGM mitochondrial topology to the three AGM mitochondrial topologies that corresponded to the ML SIVagm topology (labeled footnote c in Table 1). We made these three comparisons because of the ambiguity that exists in the branching order of the two *C. pygerythrus* taxa. All three of these topologies

were rejected by the SH-test on the AGM mitochondrial dataset ($p < 0.05$). In other words, we can confidently reject the hypothesis that the AGM mitochondrial genomes evolved along the same topology as the SIVagm genomes.

DISCUSSION

Our results present a significant challenge to the ancient origin of SIVagm followed by codivergence with their AGM hosts. Using an ML framework, we inferred robust phylogenies from the AGM mitochondrial genomes and SIVagm genomes. Although *C. sabaeus* are the most basal taxa for both the mitochondrial genome and SIVagm trees (according to midpoint rooting methods), the other taxa do not share the same topology. As in previous studies on AGM taxonomy, we were unable to determine if *C. pygerythrus* from Tanzania and South Africa form a monophyletic clade, even though they are infected by the same SIVagm. Given the genetic distance between them, if *C. pygerythrus* is monophyletic, then it exhibits greater genetic diversity than is observed between *C. tantalus* and *C. aethiops*. In any case, using the SH-test we can confidently state that the virus did not evolve along the topology of the host mitochondrial genomes, and conversely, that the mitochondrial genomes did not evolve along the viral topology.

These results demonstrate the usefulness of complete mitochondrial genomes in resolving recent primate divergence events, even those that occurred within a short span of time as with *C. pygerythrus*, *C. tantalus*, and *C. aethiops*. We also present the first date for the diversification of AGMs that accounts for the rate variation across the catarrhine phylogeny. The dating of the AGM MRCA at 2.81 ± 0.35 MYA indicates that if SIVagm did codiverge with its hosts (which seems unlikely given our findings), it must have infected the AGM common ancestor nearly 3 MYA.

Without evidence for a shared history, a model other than codivergence is needed to explain the observed pattern of SIVagm infection. A preferential host-switching model [7], whereby viral transmission occurred over already established primate host ranges and favored the cross-species transmission of SIV from an initial AGM population to others, is a strong candidate. When the SIVagm phylogeny is mapped onto the distribution of AGMs in Africa, a geographic pattern of west-to-east transmission emerges (Figure 5). SIVsab is likely the most basal of the SIVagm taxa [1], and its host, *C. sabaeus*, has the westernmost geographic distribution. SIVtan branches off next, and the range of *C. tantalus* begins at the Volta River, just east of the current *C. sabaeus* range, and continues into central Africa [12]. Finally, SIVgri and SIVver are the most derived SIVagm taxa and infect monkeys inhabiting the easternmost part of the continent. Since SIVagm is predominantly a sexually transmitted virus [32], and AGM species are known to hybridize in the wild [12], sexual encounters between AGM species may have facilitated SIV transmission at the edges of AGM ranges and the subsequent geographic spread of the virus.

While geography and behavior likely provided ample opportunity for the transmission of an initial SIVagm variant from one AGM species to the next, these factors alone cannot explain why the various AGM species would have acquired SIV only from other AGM species, rather than other infected primate species. We speculate that intrinsic immunity factors, such as the APOBEC proteins [44–46] and TRIM5α [47], may have played a role in this context. These proteins have been shown to prevent initiation of retroviral infection via a variety of mechanisms [48]. Specifically, these

factors may have prevented the more distantly related *Cercopithecus* monkeys from becoming infected with SIVagm and similarly blocked the introduction of SIV from non-AGM species into the AGMs. In this light, the infection of the patas monkey with SIVsab might be an important clue, since this monkey is very closely related to the AGMs [49] and has been observed engaging in aggressive behavior with *C. sabaeus* [15]. Although the findings presented here on the age of SIV suggest that the virus was not a relevant force in the ancient evolution of these proteins, intrinsic immunity factors may have been crucial in shaping the distribution of SIV across the range of the African primates it infects.

It is important to bear in mind that mitochondrial genomes, despite their length and phylogenetic information, represent only a single maternally inherited genetic locus. Regions of the AGM nuclear genome may have evolved along different evolutionary trajectories, some of which might be congruent with the viral evolutionary history. Furthermore, the AGM mitochondria may have experienced incomplete lineage sorting during the speciation events, which could obscure the species tree [50,51]. Nevertheless, our results represent the first and only statistically robust species-level AGM phylogeny to our knowledge, and this phylogeny unequivocally disagrees with the viral phylogeny. While there are examples of incongruence among mitochondrial and nuclear markers in the African guenons [52], it is unlikely that other population level phenomenon such as introgression would occur in the mitochondria of the AGMs. The strong philopatry observed in AGM females, coupled with a dominance hierarchy that discourages breeding with migrant females, would decrease the likelihood of reproductive success of

a female immigrant and therefore the probability of mitochondrial introgression [49]. In future studies, the inclusion of additional AGM mitochondrial genomes and other informative nuclear loci would be useful in determining if any of these population level phenomena have obscured the evolutionary history of the AGM. Nevertheless, we are confident that this study poses a significant challenge to the theory of ancient infection and codivergence.

Given the conflicting AGM mitochondrial and SIVagm topologies presented here, the case for codivergence between AGMs and their SIVs is limited to the observations that (1) *C. sabaeus* and SIVsab are the basal taxa in the mitochondrial and SIV phylogenies, respectively, and (2) SIVagm forms a monophyletic group. However, the fact that *C. sabaeus* is basal in the mitochondrial phylogeny can hardly be used to argue in favor of codivergence when the remainder of the host phylogeny differs significantly from the viral one. In light of our findings, the ancient codivergence model is, to us, a less parsimonious explanation of the observed patterns than a preferential host-switching model with a relatively recent origin of SIVagm. In the absence of evidence in favor of AGM–SIVagm codivergence, we are left to wonder about the case for codivergence in other African monkeys infected with SIV.

A recent ancestry of SIVagm calls into question the conclusions put forth by Kuhmann et al. [30] regarding the coevolution of SIVagm with host protein CCR5. Our analysis of the CCR5 locus suggests that there are no unique species-specific differences among the alleles that would suggest coevolution. Furthermore, the study by Kuhmann et al. did not perform a formal test for selection and assumed that a higher proportion of

nonsynonymous-to-synonymous substitutions was evidence of positive selection; however, the ratio they observed (approximately two nonsynonymous changes for every one synonymous change) is consistent with purifying selection, as nonsynonymous mutations are more frequent by chance alone.

If SIVagm is not the result of an ancient infection, then its avirulence in its natural hosts may have evolved over a much shorter time frame than implied by the ancient codivergence model. Competition experiments by Ariën et al. [53] between matched pairs of HIV samples from 2002–2003 and the late 1980s suggest that the virus may be attenuating in the human population. The authors proposed that this loss of replicative fitness by HIV might be due to its adaptation to the human immune system coupled with repeated bottlenecks resulting from human-to-human transmission. Their data suggest that evidence of reduced virulence could be perceived in relatively short periods of time. Precise dating of the original SIV infection in the AGMs may help us better appreciate the evolutionary time frame in which such change is possible in the viral lineage.

MATERIALS AND METHODS

Mitochondrial genome amplification and sequencing. DNA extracts from *C. pygerythrus* from Tanzania (CAE9649), *C. pygerythrus* from South Africa (V389), *C. sabaeus* (Letta), and *C. tantalus* (Bébé) were provided by A. C. van der Kuyl. Mitochondrial genomes were amplified using three PCR primer sets whose products ranged from 5 to 8 kilobases, which were designed based on the method developed by Raaum et al. [42]. Reactions were performed using the TripleMaster PCR system with an annealing temperature of 52 °C with an extension time of 9 min for the first ten cycles, which was extended 15 s for each additional cycle. Each reaction was run for 35 cycles. PCR products were purified using QIAquick PCR purification kits (Qiagen, http://www.qiagen.com/), and the templates were sequenced using internal primers. Regions that proved difficult to sequence from original template were re-amplified using internal PCR primers and then sequenced using those primers. All primer sequences are shown in Table S1. PCR reactions were confirmed using a 0.8% agarose gel stained with SYBR Safe (Invitrogen, http://www.invitrogen.com/). DNA sequencing was performed by the Genomic Analysis and Technology Core Facility (University of Arizona, Tucson, Arizona, United States) using an automated sequencer (Applied Biosystems 3730XL DNA Analyzer, http://www.appliedbiosystems.com/) until each base had been sequenced at least twice. Contigs were then assembled using Sequencher version 4.2 (Gene Codes Corporation, http://www.genecodes.com/).

Each of the four mitochondrial genomes was completely sequenced, except for *C. sabaeus*, for which a 200–base-pair region proved problematic, possibly due to secondary

structure of the template. In addition, the mitochondrial genome of *C. tantalus* exhibited a repeat structure within its control region that, while unusual, is not unprecedented [54]. A 115–base-pair region was repeated as many as three instances in some sequencing reactions, whereas in others it appeared only a single time. PCR amplifications of the *C. tantalus* control region indicated that multiple forms of this region existed. Unfortunately, due to degradation of our original template, we were unable to determine whether this repeat structure was due to PCR error or actual heterogeneity in the sample. Nevertheless, this region was excluded from our analysis, because the repeats had no homologous region in any other mitochondrial genome and were therefore phylogenetically uninformative.

Phylogenetic analyses. CD4 and CCR5 sequences were downloaded from GenBank. CD4 genes labeled as Barbados were classified as *C. sabaeus* based on recent genetic testing using cytochrome b sequence analysis [55]. All redundant CCR5 sequences were removed, except for those that were isolated from different species. Each of these datasets was aligned by hand using Se-Al [56]. ML phylogenetic trees for these two loci were inferred using a heuristic search in PAUP* version 4.0b10 [57]. The models of nucleotide substitution, Kimura81 + Inv for CD4 and HKY + Inv for CCR5, were identified by ModelTest version 3.7 [58]. Bootstrap support was assessed using 1,000 and 100 replicates for the CD4 and CCR5 topologies, respectively, using the ML nucleotide substitution parameters estimated from the ML phylogeny.

The four AGM mitochondrial genomes sequenced here and the previously published *C. aethiops* and *C. sabaeus* mitochondrial genomes were aligned by hand using Se-Al, except for the variable D-loop regions, which were aligned using CLUSTAL X [59]. A single phylogenetic tree was inferred using an exhaustive search with ML parameters inferred under a GTR + Γ_4 nucleotide substitution model in PAUP*. Bootstrap support was assessed in an ML framework whereby the nucleotide substitution parameters were reestimated for each replicate and a heuristic search was performed; this was done for 1,000 replicates.

In addition, a phylogenetic tree was inferred with a GTR + Γ_4 nucleotide substitution model in a Bayesian framework using MrBayes version 3.0 [60]. Two independent runs were performed, each using 1 million steps with four chains sampling every 100 steps. The first 10% of the trees were removed and posterior probabilities were calculated from these post-burnin trees.

SIVagm genomes were obtained from the HIV Sequence Database at Los Alamos National Laboratory (LANL, http://hiv.lanl.gov/content/hiv-db/mainp age.html). In the initial analysis, the four genomes were aligned using CLUSTAL X. We excluded the first 3,500 bases of all SIVagm genomes from our analyses, because SIVsab is a known recombinant in the 3′ part of this region, and its phylogenetic placement is ambiguous in the 5′ section of this region [25]. The sequences were aligned using CLUSTAL X, and an exhaustive search inferred a single phylogenetic tree using ML parameters estimated under a GTR + Γ_4 model in PAUP*. In the secondary analysis on all seven published SIVagm genomes and the env genes of SIVver from South Africa, the sequences were

also obtained from the LANL database. A single phylogenetic tree was found using a heuristic search with ML parameters inferred under a GTR + Γ_4 model in PAUP*. Bootstrap support was assessed in an ML framework whereby each nucleotide substitution parameter was reestimated for each replicate and a heuristic search was performed; this was done for 1,000 replicates for the four-taxa tree and for 100 replicates for the nine-taxa tree.

The SH-test was performed in PAUP* on the unrooted bifurcating topologies for the six AGM mitochondrial genomes, in which the *C. sabaeus* taxa are monophyletic, and the four initial SIVagm genomes. The test parameters were estimated using a GTR + Γ_4 model with 1,000 RELL replicates.

Molecular clock. Molecular clock analysis was carried out using the r8s software developed by Sanderson [61]. In order to estimate the divergence dates of and within the AGMs, we included other complete mitochondrial genomes from the Old World monkeys *Colobus guereza*, *Macaca sylvanus*, *Papio hamadryas*, and *Trachypithecus obscurus*; lesser and great apes *Hylobates lar*, *Gorilla gorilla*, *Homo sapiens*, *Pan paniscus*, *Pan troglodytes*, *Pongo pygmaeus pygmaeus*, and *Pongo pygmaeus abelii*; and a New World monkey, *Cebus albifrons*, which was used as an outgroup to root the phylogeny. An alignment of these mitochondrial genomes was obtained using CLUSTAL X. The two variable D-loop regions were removed from further analysis due to their poor sequence conservation.

Our analysis closely followed that of Raaum et al. [42], who first estimated divergence dates using many of the same primate mitochondrial genomes. We used a semiparametric approach with a penalized likelihood method in which the rate of evolution along each branch is allowed to vary, but a roughness penalty prevents the rate from varying too much from branch to branch [61]. An optimal smoothing parameter was chosen by cross-validation analysis. The non-clocklike behavior of this dataset was not unexpected given the decrease in the rate of evolution observed in apes [62,63]. We based our divergence estimates on three fossil-derived calibration points identified by Raaum et al.: the 6-MYA split between *Pan* and *Homo*, the 14-MYA split between the Asian great apes (Pongo) and the African great apes, and the 23-MYA split between hominoids and the Old World monkeys. These fossil-derived dates were entered into r8s as point estimates, rather than intervals, because r8s does not work well with narrow calibration windows.

To estimate confidence intervals for the age of the AGM clade and the radiation of *C. aethiops*, *C. tantalus*, and *C. pygerythrus*, we used ML branch lengths estimated from 100 nonparametric bootstrap replicate trees in PAUP* and 100 trees from a Bayesian MCMC run. Bootstrap trees in PAUP* were obtained using GTR + Γ_4 parameters estimated from an ML tree. Trees from the MCMC run were sampled every 9,000 trees after the first 100,000 burnin trees. In both cases, every tree supported the identical topology for all taxa except *C. aethiops*, *C. tantalus*, and the two *C. pygerythrus*. The Bayesian analysis did, however, place *C. aethiops* and *C. tantalus* together 100% of the time, which is consistent with our previous phylogenetic analysis on the AGM

mitochondrial genomes. We provide estimates of error as two standard deviations from the mean age of the estimated node for each of these datasets (ML and MCMC); these estimates are conservative, as they do not capture the uncertainty in the fossil record.

SUPPORTING INFORMATION

Accession Numbers. The GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) accession numbers for the AGM mitochondrial genomes sequenced in this study are EF597500–EF59750. Accession numbers for other genes and genomes are as follows: previously published AGM mitochondrial genomes (AY863426 and DQ069713), CCR5 (AB015944, AF035221, AF035222, AF035223, AF081577, AF105286, AF162006, AF162007, AF162016, AF162017, AF162020, AF162022, AF162023, AF162025, AF162026, AF162030, AF162031, AF252552, U83324, and U83325), CD4 (AF001221–AF001228, D86589, and X73322), initial SIVagm genomes (M66437, L40990, U58991, and U04005), additional SIVagm genes and genomes (BD092095, M30931, M29975, AF015905, and AF015906), and additional primate mitochondrial genomes (AY863427, NC_002764, Y18001, AY863425, X99256, NC_001645, NC_001807, NC_001644, NC_001643, NC_001646, NC_002083, and AJ309866).

ACKNOWLEDGMENTS

The authors would like to thank Marcia Kalish for fruitful discussions, Michael Sanderson for recommendations on dating techniques, Adam Bjork for comments on the manuscript, and Antoinette van der Kuyl for supplying the African green monkey DNA samples.

AUTHOR CONTRIBUTIONS

JOW and MW conceived of and designed the experiments, performed the phylogenetic analysis, and wrote the paper. JOW performed the experiments.

REFERENCES

1. Bibollet-Ruche F, Bailes E, Gao F, Pourrut X, Barlow KL, et al. (2004) New simian immunodeficiency virus infecting De Brazza's monkeys (*Cercopithecus neglectus*): Evidence for a Cercopithecus monkey virus clade. J Virol 78: 7748–7762.

2. McGeoch DJ, Cook S (1994) Molecular phylogeny of the alphaherpesvirinae subfamily and a proposed evolutionary timescale. J Mol Biol 238: 9–22.

3. Morzunov SP, Rowe JE, Ksiazek TG, Peters CJ, St Jeor SC, et al. (1998) Genetic analysis of the diversity and origin of hantaviruses in *Peromyscus leucopus* mice in North America. J Virol 72: 57–64.

4. Dimcheff DE, Drovetski SV, Krishnan M, Mindell DP (2000) Cospeciation and horizontal transmission of avian sarcoma and leukosis virus gag genes in galliform birds. J Virol 74: 3984–3995.

5. Switzer WM, Salemi M, Shanmugam V, Gao F, Cong ME, et al. (2005) Ancient co-speciation of simian foamy viruses and primates. Nature 434: 376–380.

6. Katzourakis A, Tristem M, Pybus OG, Gifford RJ (2007) Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A 104: 6261–6265.

7. Charleston MA, Robertson DL (2002) Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. Syst Biol 51: 528–535.

8. Ohta Y, Masuda T, Tsujimoto H, Ishikawa K, Kodama T, et al. (1988) Isolation of simian immunodeficiency virus from African green monkeys and seroepidemiologic survey of the virus in various non-human primates. Int J Cancer 41: 115–122.

9. Allan JS, Kanda P, Kennedy RC, Cobb EK, Anthony M, et al. (1990) Isolation and characterization of simian immunodeficiency viruses from two subspecies of African green monkeys. AIDS Res Hum Retroviruses 6: 275–285.

10. Allan JS, Short M, Taylor ME, Su S, Hirsch VM, et al. (1991) Species-specific diversity among simian immunodeficiency viruses from African green monkeys. J Virol 65: 2816–2828.

11. Muller MC, Saksena NK, Nerrienet E, Chappey C, Herve VM, et al. (1993) Simian immunodeficiency viruses from central and western Africa: Evidence for a new species-specific lentivirus in tantalus monkeys. J Virol 67: 1227–1235.

12. Lernould JM (1988) Classification and geographical distribution of guenons: A review. In: Gautier-Hion A, Boulière F, Gautier JP, Kingdon J, editors. A primate radiation: Evolutionary biology of the African guenons. pp. 54–78.

13. van der Kuyl AC, Kuiken CL, Dekker JT, Goudsmit J (1995) Phylogeny of African monkeys based upon mitochondrial 12S rRNA sequences. J Mol Evol 40: 173–180.

14. van der Kuyl AC, van Gennep DR, Dekker JT, Goudsmit J (2000) Routine DNA analysis based on 12S rRNA gene sequencing as a tool in the management of captive primates. J Med Primatol 29: 307–315.

15. Bibollet-Ruche F, Galat-Luong A, Cuny G, Sarni-Manchado P, Galat G, et al. (1996) Simian immunodeficiency virus infection in a patas monkey (*Erythrocebus patas*): Evidence for cross-species transmission from African green monkeys (*Cercopithecus aethiops sabaeus*) in the wild. J Gen Virol 77(Pt 4): 773–781.

16. Jin MJ, Rogers J, Phillips-Conroy JE, Allan JS, Desrosiers RC, et al. (1994) Infection of a yellow baboon with simian immunodeficiency virus from African green monkeys: Evidence for cross-species transmission in the wild. J Virol 68: 8454–8460.

17. van Rensburg EJ, Engelbrecht S, Mwenda J, Laten JD, Robson BA, et al. (1998) Simian immunodeficiency viruses (SIVs) from eastern and southern Africa: Detection of a SIVagm variant from a chacma baboon. J Gen Virol 79(Pt 7): 1809–1814.

18. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, et al. (1999) Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. Nature 397: 436–441.

19. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, et al. (2006) Human immunodeficiency viruses: SIV infection in wild gorillas. Nature 444: 164.

20. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR (1989) An African primate lentivirus (SIVsm) closely related to HIV-2. Nature 339: 389–392.

21. Damond F, Worobey M, Campa P, Farfara I, Colin G, et al. (2004) Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. AIDS Res Hum Retroviruses 20: 666–672.

22. Tomonaga K, Katahira J, Fukasawa M, Hassan MA, Kawamura M, et al. (1993) Isolation and characterization of simian immunodeficiency virus from African white-crowned mangabey monkeys (*Cercocebus torquatus lunulatus*). Arch Virol 129: 77–92.

23. Daniel MD, Letvin NL, King NW, Kannagi M, Sehgal PK, et al. (1985) Isolation of T-cell tropic HTLV-III-like retrovirus from macaques. Science 228: 1201–1204.

24. Murphey-Corb M, Martin LN, Rangan SR, Baskin GB, Gormus BJ, et al. (1986) Isolation of an HTLV-III-related retrovirus from macaques with simian AIDS and its possible origin in asymptomatic mangabeys. Nature 321: 435–437.

25. Jin MJ, Hui H, Robertson DL, Muller MC, Barre-Sinoussi F, et al. (1994) Mosaic genome structure of simian immunodeficiency virus from west African green monkeys. EMBO J 13: 2935–2947.

26. Souquiere S, Bibollet-Ruche F, Robertson DL, Makuwa M, Apetrei C, et al. (2001) Wild *Mandrillus sphinx* are carriers of two types of lentivirus. J Virol 75: 7086–7096.

27. Bailes E, Gao F, Bibollet-Ruche F, Courgnaud V, Peeters M, et al. (2003) Hybrid origin of SIV in chimpanzees. Science 300: 1713.

28. Dazza MC, Ekwalanga M, Nende M, Shamamba KB, Bitshi P, et al. (2005) Characterization of a novel vpu-harboring simian immunodeficiency virus from a Dent's Mona monkey (*Cercopithecus mona denti*). J Virol 79: 8560–8571.

29. Fomsgaard A, Muller-Trutwin MC, Diop O, Hansen J, Mathiot C, et al. (1997) Relation between phylogeny of African green monkey CD4 genes and their respective simian immunodeficiency virus genes. J Med Primatol 26: 120–128.

30. Kuhmann SE, Madani N, Diop OM, Platt EJ, Morvan J, et al. (2001) Frequent substitution polymorphisms in African green monkey CCR5 cluster at critical sites for infections by simian immunodeficiency virus SIVagm, implying ancient virus-host coevolution. J Virol 75: 8449–8460.

31. Hendry RM, Wells MA, Phelan MA, Schneider AL, Epstein JS, et al. (1986) Antibodies to simian immunodeficiency virus in African green monkeys in Africa in 1957–62. Lancet 2: 455.

32. Phillips-Conroy JE, Jolly CJ, Petros B, Allan JS, Desrosiers RC (1994) Sexual transmission of SIVagm in wild grivet monkeys. J Med Primatol 23: 1–7.

33. Muller-Trutwin MC, Corbet S, Tavares MD, Herve VM, Nerrienet E, et al. (1996) The evolutionary rate of nonpathogenic simian immunodeficiency virus (SIVagm) is in agreement with a rapid and continuous replication in vivo. Virology 223: 89–102.

34. Broussard SR, Staprans SI, White R, Whitehead EM, Feinberg MB, et al. (2001) Simian immunodeficiency virus replicates to high levels in naturally infected African green monkeys without inducing immunologic or neurologic disease. J Virol 75: 2262–2275.

35. Traina-Dorge V, Blanchard J, Martin L, Murphey-Corb M (1992) Immunodeficiency and lymphoproliferative disease in an African green monkey dually infected with SIV and STLV-I. AIDS Res Hum Retroviruses 8: 97–100.

36. Gravell M, London WT, Hamilton RS, Stone G, Monzon M (1989) Infection of macaque monkeys with simian immunodeficiency virus from African green monkeys: Virulence and activation of latent infection. J Med Primatol 18: 247–254.

37. Goldstein S, Ourmanov I, Brown CR, Plishka R, Buckler-White A, et al. (2005) Plateau levels of viremia correlate with the degree of CD4+-T-cell loss in simian immunodeficiency virus SIVagm-infected pigtailed macaques: Variable pathogenicity of natural SIVagm isolates. J Virol 79: 5153–5162.

38. Leakey M (1988) Fossil evidence for the evolution of the guenons. In: Gautier-Hion A, Boulière F, Gautier JP, Kingdon J, editors. A primate radiation: Evolutionary biology of the African guenons. pp. 7–12.

39. Shimada MK, Terao K, Shotake T (2002) Mitochondrial sequence diversity within a subspecies of savanna monkeys (*Cercopithecus aethiops*) is similar to that between subspecies. J Hered 93: 9–18.

40. Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, et al. (2000) Origins and evolution of AIDS viruses: Estimating the time-scale. Biochem Soc Trans 28: 275–282.

41. Shimodaira H, Hasegawa M (1999) Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol 16: 1114–1116.

42. Raaum RL, Sterner KN, Noviello CM, Stewart CB, Disotell TR (2005) Catarrhine primate divergence dates estimated from complete mitochondrial genomes: Concordance with fossil and nuclear DNA evidence. J Hum Evol 48: 237–257.

43. Sanderson MJ (2002) Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. Mol Biol Evol 19: 101–109.

44. Bogerd HP, Doehle BP, Wiegand HL, Cullen BR (2004) A single amino acid difference in the host APOBEC3G protein controls the primate species specificity of HIV type 1 virion infectivity factor. Proc Natl Acad Sci U S A 101: 3770–3774.

45. Mangeat B, Turelli P, Liao S, Trono D (2004) A single amino acid determinant governs the species-specific sensitivity of APOBEC3G to Vif action. J Biol Chem 279: 14481–14483.

46. Schrofelbauer B, Chen D, Landau NR (2004) A single amino acid of APOBEC3G controls its species-specific interaction with virion infectivity factor (Vif). Proc Natl Acad Sci U S A 101: 3927–3932.

47. Stremlau M, Owens CM, Perron MJ, Kiessling M, Autissier P, et al. (2004) The cytoplasmic body component TRIM5alpha restricts HIV-1 infection in Old World monkeys. Nature 427: 848–853.

48. Bieniasz PD (2004) Intrinsic immunity: A front-line defense against viral attack. Nat Immunol 5: 1109–1115.

49. Tosi AJ, Melnick DJ, Disotell TR (2004) Sex chromosome phylogenetics indicate a single transition to terrestriality in the guenons (tribe Cercopithecini). J Hum Evol 46: 223–237.

50. Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5: 568–583.

51. Hoelzer GA, Wallman J, Melnick DJ (1998) The effects of social structure, geographical structure, and population size on the evolution of mitochondrial DNA: II. Molecular clocks and the lineage sorting period. J Mol Evol 47: 21–31.

52. Disotell TR, Raaum RL (2002) Molecular timescale and gene tree incongruence in the guenons. In: Glenn ME, Cords M, editors. The guenons: Diversity and adaptation in African monkeys. pp. 27–36.

53. Ariën KK, Troyer RM, Gali Y, Colebunders RL, Arts EJ, et al. (2005) Replicative fitness of historical and recent HIV-1 isolates suggests HIV-1 attenuation over time. AIDS 19: 1555–1564.

54. Wilkinson GS, Mayer F, Kerth G, Petri B (1997) Evolution of repeated sequence arrays in the D-loop region of bat mitochondrial DNA. Genetics 146: 1035–1048.

55. Pandrea I, Apetrei C, Dufour J, Dillon N, Barbercheck J, et al. (2006) Simian immunodeficiency virus SIVagm.sab infection of Caribbean African green monkeys: A new model for the study of SIV pathogenesis in natural hosts. J Virol 80: 4858–4867.

56. Rambaut A (1996) Se-Al: Sequence alignment editor. Available: http://evolve.zoo.ox.ac.uk/. Accessed 31 May 2007.

57. Swofford DL (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4. Sunderland (Massachusetts): Sinauer Associates.

58. Posada D, Crandall KA (1998) MODELTEST: Testing the model of DNA substitution. Bioinformatics 14: 817–818.

59. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882.

60. Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572–1574.

61. Sanderson MJ (2006) r8s version 1.71. Analysis of rates ("r8s") of evolution. Available: http://ginger.ucdavis.edu/r8s/. Accessed 31 May 2007.

62. Yi S, Ellsworth DL, Li WH (2002) Slow molecular clocks in Old World monkeys, apes, and humans. Mol Biol Evol 19: 2191–2198.

63. Elango N, Thomas JW, Yi SV (2006) Variable molecular clocks in hominoids. Proc Natl Acad Sci U S A 103: 1370–1375.

64. Beer BE, Bailes E, Goeken R, Dapolito G, Coulibaly C, et al. (1999) Simian immunodeficiency virus (SIV) from sun-tailed monkeys (*Cercopithecus solatus*): Evidence for host-dependant evolution of SIV within the *C. lhoesti* superspecies. J Virol 79: 7734–7744.

Table 1. Shimodaira-Hasegawa Test on SIVagm and AGM Mitochondrial Phylogenies

Taxa	Genome Topologies	−lnL	SH-Test *p*-Value
SIVagm	(((SIVgri,SIVver),SIVtan),SIVsab)[a]	26272.6009	—
	((SIVgri,(SIVver,SIVtan)),SIVsab)	26303.6219	0.001
	(((SIVgri,SIVtan),SIVver),SIVsab)[b]	26306.0266	<0.001
AGM[d]	((((*aethiops*,*tantalus*),*pygerythrus*_TZ),*pygerythrus*_SA),*sabaeus*)[a]	32212.1891	—
	((((*aethiops*,*pygerythrus*_TZ),*tantalus*),*pygerythrus*_SA),*sabaeus*)	32230.0744	0.009
	(((*aethiops*,*pygerythrus*_TZ),(*pygerythrus*_SA,*tantalus*)),*sabaeus*)	32232.948	0.003
	((((*aethiops*,*pygerythrus*_TZ),*pygerythrus*_SA),*tantalus*),*sabaeus*)[c]	32232.3349	0.003
	(((*aethiops*,(*pygerythrus*_TZ,*tantalus*)),*pygerythrus*_SA),*sabaeus*)	32229.5687	0.007
	(((*aethiops*,*tantalus*),(*pygerythrus*_SA,*pygerythrus*_TZ)),*sabaeus*)	32214.3564	0.672
	((*aethiops*,((*pygerythrus*_SA,*pygerythrus*_TZ),*tantalus*)),*sabaeus*)	32229.3796	0.005
	((*aethiops*,((*pygerythrus*_SA,*tantalus*),*pygerythrus*_TZ)),*sabaeus*)	32233.2247	0.001
	((*aethiops*,(*pygerythrus*_SA,(*pygerythrus*_TZ,*tantalus*))),*sabaeus*)	32232.7567	0.001
	(((*aethiops*,(*pygerythrus*_SA,*pygerythrus*_TZ)),*tantalus*),*sabaeus*)[c]	32228.9714	0.005
	((((*aethiops*,*tantalus*),*pygerythrus*_SA),*pygerythrus*_TZ),*sabaeus*)	32213.8369	0.619
	(((*aethiops*,(*pygerythrus*_SA,*tantalus*)),*pygerythrus*_TZ),*sabaeus*)	32227.9097	0.004
	((((*aethiops*,*pygerythrus*_SA),*tantalus*),*pygerythrus*_TZ),*sabaeus*)	32228.1915	0.005
	((((*aethiops*,*pygerythrus*_SA),*pygerythrus*_TZ),*tantalus*),*sabaeus*)[c]	32233.0693	0.001
	(((*aethiops*,*pygerythrus*_SA),(*pygerythrus*_TZ,*tantalus*)),*sabaeus*)	32233.0598	0.001

All trees are shown rooted with SIVsab or both *C. sabaeus* taxa.
[a]ML topology.
[b]SIVagm genome topology that corresponds to the ML AGM mitochondrial topology.
[c]AGM mitochondrial topologies that correspond to the ML SIVagm genome topology.
[d]All AGM topologies contain two *C. sabaeus* taxa that are monophyletic.
SA, South Africa; TZ, Tanzania.
doi:10.1371/journal.ppat.0030095.t001

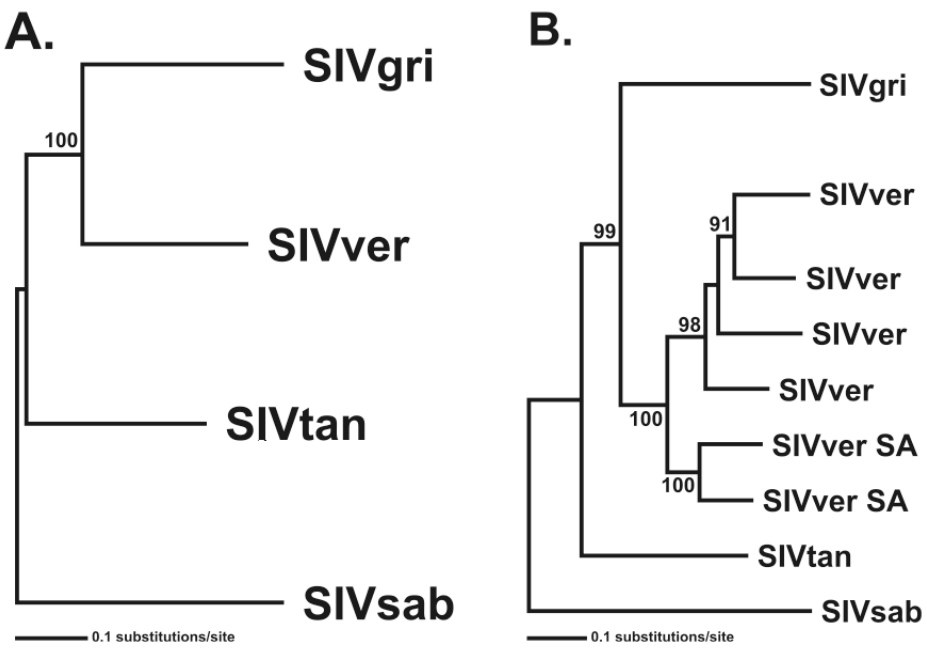


Figure 1. Phylogenetic Relationships among SIVagm Genomes

(A) SIVagm genomes and (B) SIVagm genomes plus two SIVver *env* genes from South Africa (SA). Both trees are shown with SIVsab as an outgroup, although midpoint rooting produces the same rooting pattern. ML nonparametric bootstrap support values (>50%) are shown on nodes.
doi:10.1371/journal.ppat.0030095.g001

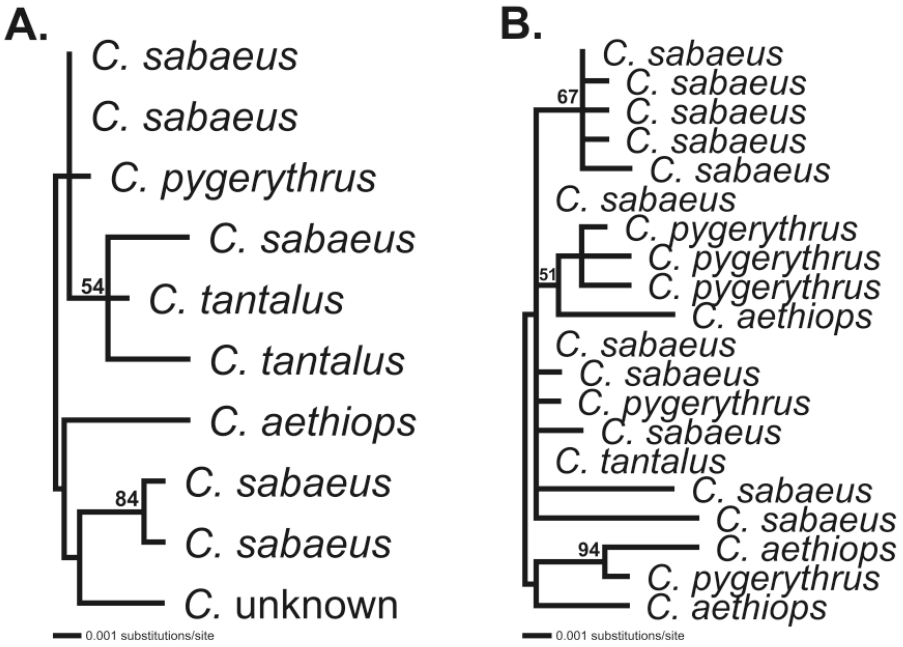


Figure 2. Phylogenetic Relationships among AGM Nuclear Loci

(A) *CD4* phylogeny and (B) *CCR5* phylogeny. Both trees are midpoint rooted. ML nonparametric bootstrap support values (>50) are shown on nodes. "*C.* unknown" in (A) refers to a taxon with no published species-specific information.

doi:10.1371/journal.ppat.0030095.g002

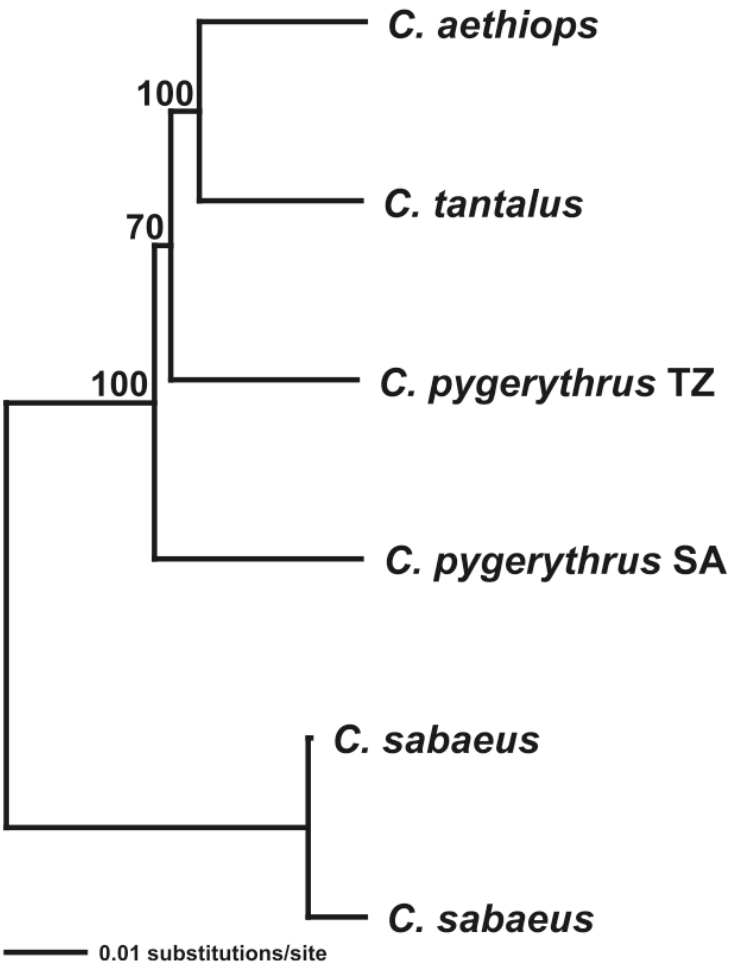


Figure 3. Phylogenetic Relationships among AGM Mitochondrial Genomes

ML tree is shown with *C. sabaeus* taxa as an outgroup, although midpoint rooting produced the same rooting pattern. *C. pygerythrus* were isolated from Tanzania (TZ) and South Africa (SA). ML non-parametric bootstrap support values (>50%) are shown on nodes.
doi:10.1371/journal.ppat.0030095.g003

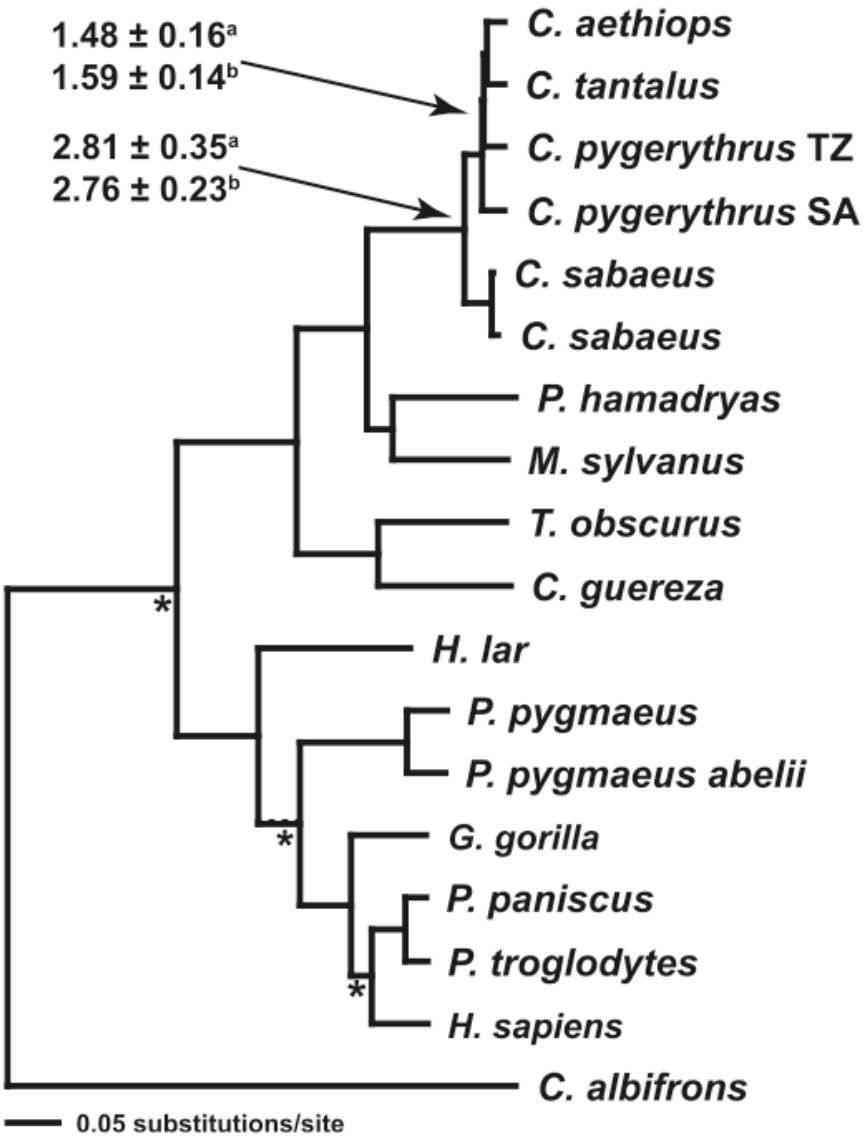


Figure 4. Relaxed Molecular Clock Analysis of Catarrhine Taxa with Estimated AGM Divergence Dates

ML tree is rooted using *C. albifrons* as an outgroup, with mean MRCA date for the AGM taxa and 95% confidence intervals estimated from 100 replicate trees from (a) ML bootstrap analysis and (b) Bayesian MCMC analysis. Asterisks designate nodes as being fossil-constrained as defined by Raaum et al. [42].

doi:10.1371/journal.ppat.0030095.g004

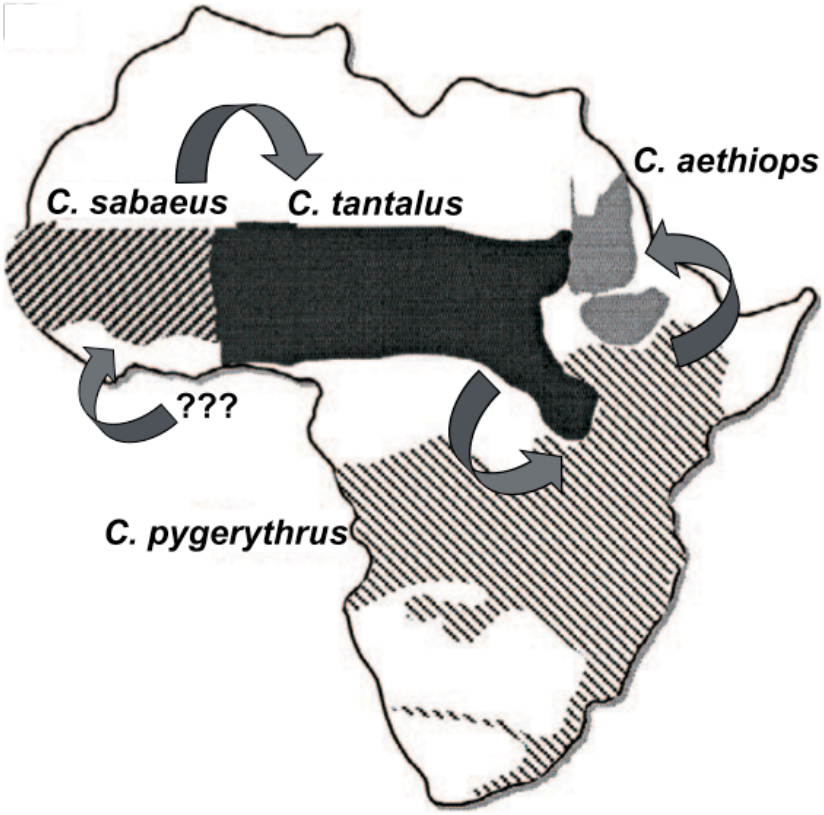


Figure 5. Hypothesized SIVagm Transmission Pattern across sub-Saharan Africa

AGM distributions across the African continent are depicted. According to SIVagm phylogenetic analysis, *C. sabaeus* was the first AGM to be infected with SIV, although the source of this infection is unknown. The arrows depict a possible route of transmission of the virus across already established AGM ranges. It should be noted that the inferred SIVagm phylogeny does not distinguish between the depicted route of transmission and a route in which *C. tantalus* first infected *C. aethiops,* which in turn infected *C. pygerythrus*. This figure is modified from Beer et al. [64], which utilized the range map from Lernould [12].
doi:10.1371/journal.ppat.0030095.g005

APPENDIX B: DATING THE AGE OF THE SIV LINEAGES THAT GAVE RISE TO HIV-1 AND HIV-2

Published: ***PLoS Computational Biology*** **(2009) 5(5): e1000377**

Co-author: Michael Worobey

ABSTRACT

Great strides have been made in understanding the evolutionary history of simian immunodeficiency virus (SIV) and the zoonoses that gave rise to HIV-1 and HIV-2. What remains unknown is how long these SIVs had been circulating in non-human primates before the transmissions to humans. Here, we use relaxed molecular clock dating techniques to estimate the time of most recent common ancestor for the SIVs infecting chimpanzees and sooty mangabeys, the reservoirs of HIV-1 and HIV-2, respectively. The date of the most recent common ancestor of SIV in chimpanzees is estimated to be 1492 (1266–1685), and the date in sooty mangabeys is estimated to be 1809 (1729–1875). Notably, we demonstrate that SIV sequences sampled from sooty mangabeys possess sufficient clock-like signal to calibrate a molecular clock; despite the differences in host biology and viral dynamics, the rate of evolution of SIV in sooty mangabeys is indistinguishable from that of its human counterpart, HIV-2. We also estimate the ages of the HIV-2 human-to-human transmissible lineages and provide the first age estimate for HIV-1 group N at 1963 (1948–1977). Comparisons between the SIV most recent common ancestor dates and those of the HIV lineages suggest a difference on the order of only hundreds of years. Our results suggest either that SIV is a surprisingly young lentiviral lineage or that SIV and, perhaps, HIV dating estimates are seriously compromised by unaccounted-for biases.

AUTHOR SUMMARY

HIV/AIDS continues to be a major health problem worldwide. An understanding of the evolution of HIV in humans may be greatly improved by detailed knowledge of its predecessor, simian immunodeficiency virus (SIV), in non-human primates. While HIV causes AIDS in humans, SIV generally produces a benign infection in its natural hosts. This avirulence is often attributed to coevolution between the virus and its host, possibly due to codivergence over millions of years. Here, we provide a temporal reference for evolution of SIV in its natural primate hosts. Using state-of-the-art molecular clock dating techniques, we estimate the time of most recent common ancestor for SIV in sooty mangabeys and chimpanzees at 1809 (1729–1875) and 1492 (1266–1685), respectively. These ages indicate that SIV may have infected these natural hosts for only hundreds of years before giving rise to HIV. This short duration suggests that viral–host coevolution over millions of years is not a likely explanation for the widespread avirulence of SIV. Finally, despite differences between SIV and HIV in host biology and viral pathogenicity, we have found clear and direct evidence that SIV evolves at a rapid rate in its natural hosts, an evolutionary rate that is indistinguishable from that of HIV in humans.

INTRODUCTION

HIV/AIDS is the result of at least eleven cross-species transmission events of simian immunodeficiency virus (SIV) from non-human African primates to humans. Three transmissions of SIVcpz from the central African chimpanzee subspecies (*Pan troglodytes troglodytes*) gave rise to HIV-1 groups M, N and O [1], and the other eight SIVsm transmissions from sooty mangabeys (*Cercocebus torquatus atys*) gave rise to HIV-2 groups A through H [2],[3]. All three HIV-1 groups, plus HIV-2 groups A and B, have established human-to-human transmission chains, with HIV-1 group M causing pandemic HIV/AIDS. The six other HIV-2 lineages do not appear to be transmissible among humans [2].

Determining when the virus jumped into humans has been a priority for HIV researchers. By analyzing viral sequences obtained over several decades and calibrating a molecular clock based on observed nucleotide changes, a reliable rate of sequence evolution can be inferred. Korber et al. used this method to estimate the time of most recent common ancestor (tMRCA) for HIV-1 group M at 1931 (1915–1941) [4]; this estimate has recently been pushed back slightly to 1908 (1884–1924) [5]. The tMRCA of HIV-1 group O was estimated to be 1920 (1890–1940) [6]. Both HIV-1 group M and O dates were inferred using a relaxed molecular clock, which allows the rate of evolution to vary along different branches of the tree. HIV-2 group A and B tMRCAs were estimated to be 1940 (1924–1956) and 1945 (1931–1959), respectively [7]. These dates were estimated using a strict molecular clock, (i.e., a single, constant evolutionary rate along all branches). No estimate currently exists for the tMRCA of HIV-1 group N.

There has also been success in locating the populations of chimpanzees and sooty mangabeys whose SIVs are the direct ancestors of the transmissible HIV lineages (i.e., the SIVs that lie basal to HIV-1 and HIV-2 on the SIV/HIV phylogeny). Extensive non-invasive fecal sampling of wild chimpanzees pointed to the origin of HIV-1 group M in southeastern Cameroon and HIV-1 group N in south central Cameroon [8]. Although a reciprocally monophyletic clade of SIVcpz has been found in the eastern chimpanzee subspecies (*Pan troglodytes schweinfurthii*), virus from this group does not appear to have jumped successfully into humans [9]. Surprisingly, the SIV lineage that falls immediately basal to HIV-1 group O was found in gorillas, suggesting that they might have been an intermediate host between chimpanzees and humans [10]. Similar fecal analysis in sooty mangabeys indicated that HIV-2 groups A and B were likely transmitted to humans in Côte d'Ivoire [11].

Despite these findings, an important question about the origins of SIV/HIV remains unanswered: How long have these primate hosts been infected with SIV? Answering this question would help determine the length of time SIV was in sooty mangabeys and chimpanzees before giving rise to the transmissible HIV lineages. It might also shed light on the tMRCA of the dozens of other SIV lineages.

Determining the age of SIV would provide perspective on the spread of the virus among African primate species and the subsequent zoonoses. Knowing the age may also have implications for the evolution of pathogenicity and virulence in HIV. AIDS-like symptoms have rarely been observed in non-human African primates infected with SIV [12],[13]. Historically, this lack of disease was attributed to the codivergence and

coevolution of SIV and their primate hosts over millions of years [14] (we use the term codivergence instead of cospeciation, because codivergence considers phylogenetic congruence irrespective of species classification, whereas cospeciation implies that SIVs infecting different primates can be classified as species complexes). Although there is significant correspondence between the SIV and host phylogenies, detailed analysis of this relationship suggested that a preferential host switching model, in which cross-species transmissions of SIV are more likely to occur between closely related primates, could account for this correspondence [15]. Furthermore, subsequent analysis of SIV infecting various African green monkey species, thought to be exemplary of codivergence, demonstrated a lack of evidence for host-virus codivergence [16]. In addition, the codivergence hypothesis does not account for the observation that SIV is geographically confined and naturally infects only African primates. Finally, even with biologically unrealistic assumptions about a molecular clock, Sharp et al. were unable to push the tMRCA of all SIV beyond 2500 years [17]. If it were demonstrated that SIV has evolved in a clock-like manner, then we might be able to accurately determine the age of SIV.

Here, we use relaxed molecular clock phylogenetic inference to determine the tMRCA of SIVsm/HIV-2 and SIVcpz/HIV-1. We also provide, to our knowledge, the first estimate of the age of HIV-1 group N. Taken together, these dates suggest that SIV may indeed be a relatively young viral clade and that its transmission into humans is a natural process.

RESULTS

Dating SIVsm/HIV-2 tMRCAs. We inferred phylogenies for SIVsm/HIV-2 *gag*, *pol*, and *env* loci under a relaxed molecular clock in a Bayesian Markov chain Monte Carlo (BMCMC) framework (Figure 1A–C). In each tree there was very high posterior support for monophyly in HIV-2 group A, HIV-2 group B, and the major SIVsm clades identified by Apetrei et al. [18]. The position of the root, determined by the BMCMC analysis, was also highly supported in each of the three trees. Phylogenetic inference using the three loci produces different topologies, which was expected given the observation of recombination by Apetrei et al. in their initial analysis of these loci in SIVsm [18].

The tMRCA estimates for the root of the SIVsm/HIV-2 trees differed as well (Table 1). The *pol* locus had the oldest root, putting the tMRCA of SIVsm/HIV-2 at 1686 (1525–1811). Estimates from *gag* and *env* were significantly younger, placing the SIVsm/HIV-2 tMRCA at 1809 (1798–1875) and 1861 (1788–1915), respectively. Although *gag* was older than *env*, this difference was not significant. The *pol* results also indicated older dates than *gag* and *env* for the tMRCA of HIV-2 groups A and B, although these differences were not significant. With the exception of the *env* tMRCA estimate for HIV-2 group A, all three genes suggested a slightly older origin of both HIV-2 groups A and B than previously reported by Lemey et al. [7]. There were no discernable differences in the tMRCA estimates of these three genes for the major SIVsm clades (Table 2), although there were significant differences in the age of deeper SIVsm

coalescent events among SIVsm groups 1, 2, 3, 4, and 7 [$P(gag<pol) = 0.003$; $P(env<pol)<0.001$)].

Dating SIVcpz/HIV-1 tMRCAs. We inferred phylogenies for SIVcpz/HIV-1 *gag*, *pol*, and *env* loci under a relaxed molecular clock in a BMCMC framework (Figure 1D–F). There was very high posterior support for monophyly in each of the three HIV-1 lineages as well as for the position of the root. The three loci produced different topologies, which is not surprising given the recombinant history of HIV-1 group N [1],[19].

Like in the previous SIVsm/HIV-2 analyses, the three loci produced variable tMRCA estimates for the root and the major HIV-1 lineages (Table 1). Again, *pol* had the oldest dates, with the SIVcpz/HIV-1 tMRCA at 1265 (658–1679). In contrast, *gag* had the youngest date at 1618 (1471–1746), and *env* produced an intermediate date at 1492 (1266–1685). tMRCA estimates from *gag* and *env* for both HIV-1 group M and HIV-1 group O agreed with previous estimates from Worobey et al. [5] and Lemey et al. [6], although the *pol* tMRCA dates were nearly twice as old. There was good agreement between *gag* and *env* when dating HIV-1 group N, placing the tMRCAs at 1966 (1953–1977) and 1963 (1948–1977), respectively. We also performed additional phylogenetic inference to ensure that we captured the deepest available HIV-1 group N lineages in our analyses (Figure S1).

Resolving the discrepancy among tMRCA estimates. There were significant discrepancies among the tMRCA estimates from *gag*, *pol*, and *env* in both the

SIVsm/HIV-2 and SIVcpz/HIV-1 analyses. We initially thought that this discordance was due to recombination among the loci. If recombination were responsible, the different tMRCA estimates would actually represent different times of coalescence. When examining the phylogenies, however, we found very little evidence for this scenario. There were highly similar patterns of diversity in the SIVsm clades and in the HIV-1 group M sub-types among the three loci. An explanation of recombination would necessitate selective sweeps in *gag* and *env*, which would then go on to recreate the ancestral diversity seen in the *pol* phylogeny. For example, HIV-1 group M would have a mean tMRCA around 1795, and, approximately 100 years later, part of the genome would have experienced a selective sweep that gave rise to the same pattern of sub-type diversity (Table 1).

We then explored the possibility that some of these analyses were biased. That this discrepancy among tMRCA estimates was most pronounced in the older nodes indicated a loss of signal due to this bias deeper in the phylogeny. We examined the demographic parameters (e.g. population size or growth rate) from the three loci in the SIVsm/HIV-2 and SIVcpz/HIV-1 analyses. There were significant differences in these parameter estimates from *gag* to *pol* and from *env* to *pol* ($P<0.05$). Even though these genes evolved along different topologies, their demographic history, and therefore the demographic parameters inferred from them, should be the same. We hypothesized that some genes lack sufficient demographic signal to draw accurate inference about tMRCAs and that allowing the three loci to combine their demographic signal might homogenize their tMRCA estimates.

To test this hypothesis, we compared a partition analysis where the concatenated genes shared a single demographic scenario to analyses where that scenario was inferred for each gene independently. This analysis was performed separately for SIVsm/HIV-2 and SIVcpz/HIV-1 under two different coalescent scenarios: constant population size and exponential growth. In all cases, allowing the three loci to share demographic information homogenized the tMRCA estimates such that there were no longer significant differences in the age of the root among the phylogenies. Among the SIVsm/HIV-2 loci, *gag* tMRCA estimates change the least between the partition analysis and the analyses where demographic parameters were inferred for each gene independently (Table 3). Since tMRCA estimates from *gag* are the most robust to combining demographic parameters, these dates should be taken as the best estimates. Among the SIVcpz/HIV-1 loci, *env* produced the most stable tMRCA estimates, changing as little as 0.01% under the exponential growth model (Table 3). This finding suggests that *env* provided the best tMRCA estimates for SIVcpz/HIV-1.

Given the different selective regimes that these loci experienced, it is unlikely that the differences in the tMRCA estimates among the three loci were due entirely to variable demographic signal. Nevertheless, accounting for this variation in demographic signal appears to have resolved the majority of the discrepancy among the tMRCA estimates. In addition, even if the differences among the tMRCA estimates were real, and due to recombination, all three loci suggest root ages that are of the same order of magnitude. Therefore, although we discuss these results with reference to what appear to be the most robust loci (*gag* for SIVsm/HIV-2 and *env* for SIVcpz/HIV-1), we would be able to draw

the same general conclusions from any of the three loci. Finally, we emphasize that although the tMRCA estimates presented include the mean of the posterior distribution, this mean estimate is meaningful only in context of the 95% highest probability density (HPD).

Clock-like signal in SIV. We next sought to determine if the dates we obtained were the result of clock-like signal within SIVsm or whether SIVsm had no clock-signal and we were inadvertently extrapolating HIV-2 rates across the entire tree. We compared the date estimates from *gag*, *pol*, and *env* to analyses where all non-SIVsm sequences were excluded. For all three genes, there were no significant differences in the tMRCAs between the full and SIVsm-only datasets in any of the clades measured, including all SIVsm (Table 2). Furthermore, there was no significant difference between the SIVsm *gag* substitution rate we estimated of 1.38×10^{-3} ($1.03\times10^{-3} - 1.73\times10^{-3}$) substitutions/site/year and the HIV-2 group A substitution rate of 1.22×10^{-3} substitutions/site/year estimated by Lemey et al. [7]. This similarity indicates that SIVsm does indeed have sufficient clock-like signal to date tMRCAs, and it does not appear to evolve at a different rate than HIV-2 group A, despite differences in host biology and pathogenicity.

BMCMC analysis of an alignment containing only SIVcpz did not provide meaningful date estimates, as the tMRCA estimates from these runs were indistinguishable from the prior distribution of tMRCA estimates. Therefore, the tMRCA date we inferred for SIVcpz may have incorporated HIV-1 rates that could be biasing this

estimate. However, our previous analysis of SIVsm/HIV-2 suggested that HIV-2 rates did not appreciably affect SIVsm tMRCA estimates.

Coalescent scenarios in SIVsm. In a population of constant size, the most basal lineages are consistently lost due to normal coalescent processes; the age of the root is expected to be approximately two times the effective population size [20]. However, if the population is expanding exponentially, the basal lineages will be maintained until a carrying capacity is reached. The BMCMC method used here provides a convenient framework in which to test whether a constant size or exponential growth model better describes the dynamics of a population: If the 95% HPD of the exponential growth rate excludes zero, then a constant population size can be strongly rejected. To determine if exponential growth explains the SIVsm population dynamics better than a constant population size, we looked at the exponential growth rate in alignments containing only SIVsm sequences.

Exponential growth rate 95% HPDs from *gag* and *env* in the SIVsm analysis excluded zero; however, the growth rate 95% HPD from *pol* did not exclude zero. Nevertheless, the exponential growth rate 95% HPD estimated in partition analysis (combining demographic signal from all three loci) rejected a constant population size. Thus, it seems probable that *pol* failed to reject a constant population size because it simply lacked sufficient demographic signal. Therefore, it is likely that SIVsm has not been evolving at a constant population size for the past 200 years.

Maximum age of HIV-1 groups M and N. As a result of the discovery of SIVcpz lineages that are very closely related to HIV-1 groups M and N [8], we were able to investigate when HIV-1 groups M and N shared a most recent common ancestor (MRCA) with an SIVcpz lineage. Prior to our study, there existed one estimate of this date for HIV-1 group M and SIVcpz at 1675 (1590–1761) [21]; however, this date was obtained using only two SIVcpz sequences, neither of which lies directly basal to HIV-1 group M. Our *env* analysis suggested that HIV-1 group M and the SIVcpz sequence that lies immediately basal to it shared an MRCA in 1853 (1799–1904), and HIV-1 group N and its sister SIVcpz shared a MRCA in 1921 (1885–1955). These dates represent the maximum age for the introduction of HIV-1 groups M and N into humans.

Time before zoonoses. We determined the number of years between the SIVsm and SIVcpz tMRCAs and those of the five transmissible HIV lineages (Table 4). If the SIVsm and SIVcpz tMRCAs represent the time SIV has been infecting each host, then this estimate would tell us the number of years that SIV was present in sooty mangabeys and chimpanzees before jumping into humans and giving rise to the transmissible lineages of HIV. We note, however, that a tMRCA estimate will tend to post-date the actual introduction of viral lineages into a new host if genetic diversity has since been lost or is not fully sampled. We believe such comparisons still provide useful information as long as this caveat is recognized. The times between the root of the SIVsm/HIV-2 tree and the base of the HIV-2 group A clade and the group B clade were 122.8 (57.2–199.9) and 126.2 (59.2–203.7) years, respectively. The time between the SIVcpz root and the

HIV-1 lineages was 402.8 (231.0–601.4) years for HIV-1 group M, 471.6 (291.6–693.2) years for HIV-1 group N, and 413.5 (247.1–621.3) years for HIV-1 group O. These estimates are from the *gag* locus for SIVsm/HIV-2 and from the *env* locus for SIVcpz/HIV-1; partition analyses indicated that these genes were the most reliable for each clade. Ninety-five percent HPD intervals are larger for these estimates than for other single clades because the age estimates for any two clades are not perfectly correlated.

DISCUSSION

The findings presented in this study indicate that the tMRCA of SIV in sooty mangabeys and chimpanzees is 1809 (1729–1875) and 1492 (1266–1685), respectively, assuming the relaxed molecular clock is unbiased. In addition, our results suggest that the time between the MRCA of SIVsm and SIVcpz and the MRCA of the human-to-human transmissible HIV lineages may be only hundreds of years. We present the tMRCA for all five of these HIV lineages, though we note that previous age estimates for HIV-1 groups M and O were based on larger datasets [5],[6]. We estimate the tMRCA for HIV-2 group A to be 1932 (1906–1955) and HIV-2 group B to be 1935 (1907–1961); these estimates were generated by incorporating a more biologically plausible model of rate variation among lineages, compared with the strict molecular clock used to obtain the previous HIV-2 tMRCA estimates [7].

In addition, we present the first date, to our knowledge, for the tMRCA of HIV-1 group N at 1963 (1948–1977). This date suggests that HIV-1 group N is the youngest transmissible HIV lineage and the only lineage to have originated in the second half of the twentieth century (though the possibility of a deeper history cannot be excluded given the sparse sampling). Taken together with the previous tMRCA estimates for HIV-1 groups M and O (circa 1900s and 1920s, respectively) and our updated HIV-2 group A and B dates (circa 1930s), it appears that SIV has given rise to transmissible HIV lineages throughout the twentieth century. The dispersed timing of these transmissions to humans implies that no single external factor is needed to explain the cross-species transmission of HIV. This observation is consistent with both of the two prevailing views of the origin

of the HIV epidemics. The first is the bushmeat hypothesis [22], whereby SIV is transmitted to humans during the slaughter or butchering of infected primates. The second is that the growth of sub-Saharan African cities allowed for these nascent lineages to gain a foothold [5],[7]. According to the second hypothesis, SIV may have been jumping into humans since it first infected chimpanzees and sooty mangabeys. A change in human ecology then may have altered the evolutionary dynamics, whereby a virus that historically may have only infected a few individuals and then died out now has the potential to become an epidemic lineage. It does not seem farfetched to venture that SIV will continue to be transmitted to humans well into the twenty-first century.

There are several arguments suggesting that SIV has been present in sooty mangabeys and chimpanzees longer than our results indicate. First, coalescent processes or selective sweeps might have removed the deeper lineages from the phylogeny. While we cannot discount the latter, our finding that the SIVsm population has not evolved under a constant size suggests that deep SIVsm lineages may still be present. It remains unclear whether coalescent processes may have removed deep SIVcpz branches. Nevertheless, the full SIV/HIV tree suggests that there is a relatively short period of time between the MRCAs of SIVsm and SIVcpz and the branches that lead to SIVs that infect other primates (Figure 2). A second argument is that our sampling was not thorough enough, and deep SIV branches were not included. While possible, other studies that included additional non-dated SIVsm and SIVcpz sequences did not uncover additional deeper branches [8],[18],[23]. Thirdly, it has been suggested previously that SIV may lack the clock-like signal necessary to draw inference about tMRCAs. As a part of this

study, we demonstrated that one major SIV lineage evolves in a clock-like fashion and at a rate indistinguishable from HIV. SIVsm sequences sampled over 30 years contain enough information to calibrate the molecular clock and date the tMRCA of an SIV clade. While we were able to use the SIVsm rate to date the tMRCA of SIVsm/HIV-2, this dating was not possible for SIVcpz. This difference is likely because we had far fewer SIVcpz sequences that were sampled over a relatively small window of time. Lastly, it is possible that our relaxed-clock models are biased and therefore unable to accurately date SIV coalescent events. We cannot dismiss this possibility, but the accuracy of these methods has been previously confirmed by other studies predicting the year of sampling of older HIV isolates from 1959 and 1960 [4],[5]. Furthermore, our analyses recovered HIV tMRCA estimates that are in line with those previously inferred for the age of the HIV clades. Conversely, if one were to accept the HIV dates, one would need to provide a compelling reason not to accept the tMRCA estimates for SIVsm and SIVcpz as well. If the SIV tMRCAs are not correct, then we would need to determine what would be biasing their estimates, because it might also be affecting the HIV tMRCAs and those of other RNA viruses.

The young ages of SIVsm and SIVcpz suggest that the entire SIV phylogeny may be relatively young (Figure 2). Even if SIV was present in sooty mangabeys and chimpanzees prior to the coalescence of their current diversity, we have identified divergence events deep in the SIV phylogeny that are on the order of hundreds of years old. The case of SIVsm is particularly compelling in this context since SIVsm sequences alone returned such a young date. It is difficult to reconcile these ages with an SIV

phylogeny that is millions of years old. It seems more reasonable that the SIV phylogeny is on the order of thousands or tens of thousands of years old. While it had previously been suggested that a young-looking phylogeny could actually be the result of codivergence over millions of years, this argument was partly predicated on the assumption that SIV did not have a reliable clock-like signal [17]. In light of our findings, this argument is no longer tenable. What is still needed, however, is a reliable estimate of the age of the entire SIV phylogeny.

SIV is not the only virus once thought to be ancient whose phylogeny may be better explained by the preferential host switching model. Hantaviruses infect a wide array of rodent and insectivore species. Their phylogeny was thought to be the result of an ancient infection followed by codivergence, but recent evidence suggests that the virus and host phylogenies are too dissimilar to suggest codivergence [24]. Furthermore, the molecular clock in hantaviruses suggests a tMRCA orders of magnitude younger than that of their hosts [25]. In addition, the similarity of the Arenavirus phylogeny to that of its host may also be the result of preferential host switching [26]. Furthermore, it has been proposed that feline immunodeficiency virus, a lentivirus whose lack of associated disease in natural feline hosts was thought to be the result of an ancient infection, codiverged and coevolved with its feline hosts [27],[28]; however, in light of the possible young age of SIV, it may be worth taking a more detailed look at the relationship between the feline immunodeficiency virus and feline phylogenies.

Given the ages of the SIV clades presented here, it seems unlikely that SIV evolved apathogenicity over millions of years of coevolution and codivergence with its

primate host species. It is still possible that SIV evolved avirulence in its natural hosts. If SIV were highly pathogenic when it first infected sooty mangabeys and chimpanzees, then it might have decreased in virulence over a remarkably short period of time, possibly on the order of hundreds of years. There remains the distinct possibility, however, that SIV was rarely pathogenic in its natural hosts and the low level of disease associated with SIV infection is actually the ancestral phenotype. The theory of ancient coevolution towards apathogenicity appears less plausible, given the recent discovery that SIVcpz is pathogenic in wild populations of the eastern chimpanzee subspecies (Rudicell RS, Jones JH, Pusey AE, Terio KA, Estes JD, Raphael J, Lonsdorf EV, Wilson ML, Keele BF, and Hahn BH. (2009) SIVcpz is pathogenic in its natural host. Oral Abstract. 16th Conference on Retroviruses and Opportunistic Infections). Future work distinguishing between these two alternative theories on SIV apathogencity is needed.

A young age for SIV contrasts with other ancient retroviruses. The simian foamy virus appears to be at least 30 million years old, based on congruence between the viral and primate host phylogenies [29]. Furthermore, lentiviruses, the viral group to which SIV belongs, are also millions of years old, based on the presence of defective endongenous lentiviruses in rabbits and lemurs [30]–[32]. None of these findings, however, preclude the possibility that SIV is a much younger lentiviral clade.

Finally, it is possible that SIV itself is much older than the tMRCA of the extant lineages. Dating the tMRCA of influenza A viruses infecting avian hosts suggested that deep viral lineages were constantly lost, which resulted in younger than expected tMRCA estimates for subtypes [33]. A similar process of lineage birth and death may have

occurred among SIV, in which SIVs infecting particular primate species would occasionally go extinct and later be replaced by a new species-specific SIV. This process would involve the loss of deep SIV lineages with replacement by younger ones. This extinction and reinfection would be analogous to the loss of deep branches due to the coalescent. If such a phenomenon operated across the entire SIV tree, it could mask the ancient age of the virus. Combined with a preferential host switching mechanism, a macro-evolutionary process such as this could account for a young tMRCA for an ancient virus whose phylogeny is similar to that of its host.

METHODS

Sequences and alignments. SIVsm/HIV-2 (*gag*, *pol*, and *env* sequences) and SIVcpz/HIV-1 (non-recombinant full-length genome sequences) with sampling dates were obtained from the Los Alamos National Laboratory HIV sequence database (http://hiv.lanl.gov/content/index) (Table 5). The majority of the SIVsm sequences (>85%) were sampled from infected sooty mangabeys in US primate centers between 1975 and 2005 [18],[34]. Dated sequences from macaques infected with SIV from sooty mangabeys were also included. We excluded HIV-1 group M subtype G, as this lineage is likely of recombinant origin [35]. To prevent sampling bias from HIV-1 group M lineages, only two sequences, selected randomly, of each subtype from each year were included in the alignment. Sections of the SIVcpz/HIV-1 genomes that correspond to the *gag*, *pol*, and *env* regions used for the SIVsm/HIV-2 analyses were designated (Table 5).

To improve the accuracy of phylogenetic inference, we excluded (i) recombinant regions, determined using BootScanning in the RDP package [36],[37], (ii) multiple sequences from single individuals, (iii) sequences containing frame-shift mutations, and (iv) ambiguously aligned regions. Sequences containing frame-shifts were removed to accommodate codon-partitioning models in our phylogenetic analyses. Alignments were performed using Clustal X [38] and manually cleaned in Se-al (http://tree.bio.ed.ac.uk/software/seal/). SIVsm/HIV-2 and SIVcpz/HIV-1 alignments are provided as supporting information (Datasets S1, S2, S3, S4, S5, S6).

Relaxed molecular clock analyses. To infer the tMRCA for the major SIVsm/HIV-2 and SIVcpz/HIV-1 lineages, we employed a BMCMC approach implemented in BEAST v1.4.7 [39],[40]. Initially, each of the three loci for both SIVsm/HIV-2 and SIVcpz/HIV-1 datasets was analyzed independently. Uninformative priors (i.e., tree priors) were placed on all internal nodes whose tMRCAs were estimated. We tested the appropriateness of GTR + Γ_4 and SRD06 models; the latter allows for different K and Γ values for the third codon position [41]. Three different coalescent tree priors were investigated: constant population size, exponential growth, and Bayesian skyline plot. We compared the six different model combinations for each locus using Bayes factor in Tracer v1.4 (http://beast.bio.ed.ac.uk/Tracer). The Bayes factor provided strong support for SRD06 over GTR + Γ_4 (Bayes factor>20), but there was not support for one coalescent scenario over any of the others. For SIVcpz, an exponential coalescent model produced substantially younger ages; this observation is not surprising given that a single exponential growth rate for SIVcpz and the HIV-1 group M pandemic lineage would likely underestimate the age of SIVcpz. The date estimates from the Bayesian skyline plot runs were used for both SIVsm and SIVcpz analyses because this model places the fewest constraints on the data [42]. XML input files for the SIVsm/HIV-2 *gag* and SIVcpz/HIV-1 *env* Bayesian skyline plot BMCMC runs are provided as supporting information (Datasets S7 and S8). Additional XML input files are available from the authors upon request.

Two BMCMC runs of 50 million generations were performed for each analysis to ensure convergence of parameter estimates. Tracer was used to check for convergence

and mixing (estimated sample size>200). Trees were annotated using the maximum clade credibility tree. All analyses were performed using an uncorrelated lognormal relaxed molecular clock [39]. Each analysis was also run without data to better appreciate how the prior may be affecting the tMRCA estimates. The complete SIV/HIV phylogeny was constructed using a heuristic search in a maximum likelihood framework using a GTR + Γ_4 model in PAUP* v4.1 [43]. Topological support was assessed using non-parametric bootstrapping (100 replicates using a heuristic search in a maximum likelihood framework). We used *env* instead of the entire SIV genome because many SIV lineages are of recombinant origin [44]. The *env* alignment used to construct this phylogeny was obtained from the curated Los Alamos National Laboratory sequence database.

To determine which locus's tMRCA estimates may be affected by a lack of demographic signal, we performed partition analyses. First, we pruned the SIVsm/HIV-2 dataset to contain only those sequences that were found in all three genes from the same individual and sampling year. In BEAST, we analyzed these reduced datasets assuming constant population size and exponential growth. We then concatenated all *gag*, *pol*, and *env* alignments, and each locus was partitioned to allow it to have its own tree topology, substitution model, relaxed clock model, and tMRCA estimates; they shared only the coalescent demographic parameter(s). This analysis was performed assuming constant population size or exponential growth. Partitioning was not possible for Bayesian skyline plot, as this model's demographic estimates are topology-dependant. The same protocol was used with the SIVcpz/HIV-1 dataset.

All date estimates provided are mean values with 95% HPD. Comparisons of tMRCA estimates among BMCMC runs (e.g., among loci and SIV/HIV versus SIV-only) were performed by asking how many times the estimate from one run was greater than the estimate from another run. This value was taken as the probability (P) that the two runs were different.

ACKNOWLEDGMENTS

We thank Philippe Lemey for suggestions on methodology and providing HIV-2 sampling dates, Adam Bjork for comments on the manuscript, and Michael J. Sanderson for providing computational resources. We also thank Andrew Rambaut for advice on BEAST and insightful discussions regarding the deep history of SIV.

AUTHOR CONTRIBUTIONS

Conceived and designed the experiments: JOW MW. Performed the experiments: JOW. Analyzed the data: JOW. Wrote the paper: JOW MW.

REFERENCES

1. Gao F, Bailes E, Robertson DL, Chen Y, Rodenburg CM, et al. (1999) Origin of HIV-1 in the chimpanzee *Pan troglodytes troglodytes*. Nature 397: 436–441.

2. Damond F, Worobey M, Campa P, Farfara I, Colin G, et al. (2004) Identification of a highly divergent HIV type 2 and proposal for a change in HIV type 2 classification. AIDS Res Hum Retroviruses 20: 666–672.

3. Hirsch VM, Olmsted RA, Murphey-Corb M, Purcell RH, Johnson PR (1989) An African primate lentivirus (SIVsm) closely related to HIV-2. Nature 339: 389–392.

4. Korber B, Muldoon M, Theiler J, Gao F, Gupta R, et al. (2000) Timing the ancestor of the HIV-1 pandemic strains. Science 288: 1789–1796.

5. Worobey M, Gemmel M, Teuwen DE, Haselkorn T, Kunstman K, et al. (2008) Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455: 661–664.

6. Lemey P, Pybus OG, Rambaut A, Drummond AJ, Robertson DL, et al. (2004) The molecular population genetics of HIV-1 group O. Genetics 167: 1059–1068.

7. Lemey P, Pybus OG, Wang B, Saksena NK, Salemi M, et al. (2003) Tracing the origin and history of the HIV-2 epidemic. Proc Natl Acad Sci U S A 100: 6588–6592.

8. Keele BF, Van Heuverswyn F, Li Y, Bailes E, Takehisa J, et al. (2006) Chimpanzee reservoirs of pandemic and nonpandemic HIV-1. Science 313: 523–526.

9. Santiago ML, Rodenburg CM, Kamenya S, Bibollet-Ruche F, Gao F, et al. (2002) SIVcpz in wild chimpanzees. Science 295: 465.

10. Van Heuverswyn F, Li Y, Neel C, Bailes E, Keele BF, et al. (2006) Human immunodeficiency viruses: SIV infection in wild gorillas. Nature 444: 164.

11. Santiago ML, Range F, Keele BF, Li Y, Bailes E, et al. (2005) Simian immunodeficiency virus infection in free-ranging sooty mangabeys (*Cercocebus atys atys*) from the Tai Forest, Cote d'Ivoire: implications for the origin of epidemic human immunodeficiency virus type 2. J Virol 79: 12515–12527.

12. Ling B, Apetrei C, Pandrea I, Veazey RS, Lackner AA, et al. (2004) Classic AIDS in a sooty mangabey after an 18-year natural infection. J Virol 78: 8902–8908.

13. Traina-Dorge V, Blanchard J, Martin L, Murphey-Corb M (1992) Immunodeficiency and lymphoproliferative disease in an African green monkey dually infected with SIV and STLV-I. AIDS Res Hum Retroviruses 8: 97–100.

14. Muller MC, Saksena NK, Nerrienet E, Chappey C, Herve VM, et al. (1993) Simian immunodeficiency viruses from central and western Africa: evidence for a new species-specific lentivirus in tantalus monkeys. J Virol 67: 1227–1235.

15. Charleston MA, Robertson DL (2002) Preferential host switching by primate lentiviruses can account for phylogenetic similarity with the primate phylogeny. Syst Biol 51: 528–535.

16. Wertheim JO, Worobey M (2007) A challenge to the ancient origin of SIVagm based on African green monkey mitochondrial genomes. PLoS Pathog 3: e95. doi:10.1371/journal.ppat.0030095.

17. Sharp PM, Bailes E, Gao F, Beer BE, Hirsch VM, et al. (2000) Origins and evolution of AIDS viruses: estimating the time-scale. Biochem Soc Trans 28: 275–282.

18. Apetrei C, Kaur A, Lerche NW, Metzger M, Pandrea I, et al. (2005) Molecular epidemiology of simian immunodeficiency virus SIVsm in U.S. primate centers unravels the origin of SIVmac and SIVstm. J Virol 79: 8991–9005.

19. Simon F, Mauclere P, Roques P, Loussert-Ajaka I, Muller-Trutwin MC, et al. (1998) Identification of a new human immunodeficiency virus type 1 distinct from group M and group O. Nat Med 4: 1032–1037.

20. Hudson R (1990) Gene genealogies and the coalescent process. In: Futuyama D, Antonovics J, editors. Oxford Surveys in Evolutionary Biology. New York: Oxford University Press. pp. 1–44.

21. Salemi M, Strimmer K, Hall WW, Duffy M, Delaporte E, et al. (2001) Dating the common ancestor of SIVcpz and HIV-1 group M and the origin of HIV-1 subtypes using a new method to uncover clock-like molecular evolution. Faseb J 15: 276–278.

22. Peeters M, Courgnaud V, Abela B, Auzel P, Pourrut X, et al. (2002) Risk to human health from a plethora of simian immunodeficiency viruses in primate bushmeat. Emerg Infect Dis 8: 451–457.

23. Ling B, Telfer P, Reed P, Robertson DL, Marx PA (2004) A link between SIVsm in sooty mangabeys (SM) in wild-living monkeys in Sierra Leone and SIVsm in an American-based SM colony. AIDS Res Hum Retroviruses 20: 1348–1351.

24. Ramsden C, Holmes EC, Charleston MA (2009) Hantavirus evolution in relation to its rodent and insectivore hosts: no evidence for codivergence. Mol Biol Evol 26: 143–153.

25. Ramsden C, Melo FL, Figueiredo LM, Holmes EC, Zanotto PM (2008) High rates of molecular evolution in hantaviruses. Mol Biol Evol 25: 1488–1492.

26. Jackson AP, Charleston MA (2004) A cophylogenetic perspective of RNA-virus evolution. Mol Biol Evol 21: 45–57.

27. Pecon-Slattery J, Troyer JL, Johnson WE, O'Brien SJ (2008) Evolution of feline immunodeficiency virus in Felidae: implications for human health and wildlife ecology. Vet Immunol Immunopathol 123: 32–44.

28. Troyer JL, Vandewoude S, Pecon-Slattery J, McIntosh C, Franklin S, et al. (2008) FIV cross-species transmission: an evolutionary prospective. Vet Immunol Immunopathol 123: 159–166.

29. Switzer WM, Salemi M, Shanmugam V, Gao F, Cong ME, et al. (2005) Ancient co-speciation of simian foamy viruses and primates. Nature 434: 376–380.

30. Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, et al. (2008) A transitional endogenous lentivirus from the genome of a basal primate and implications for lentivirus evolution. Proc Natl Acad Sci U S A 105: 20362–20367.

31. Katzourakis A, Tristem M, Pybus OG, Gifford RJ (2007) Discovery and analysis of the first endogenous lentivirus. Proc Natl Acad Sci U S A 104: 6261–6265.

32. van der Loo W, Abrantes J, Esteves PJ (2009) Sharing of endogenous lentiviral gene fragments among leporid lineages separated for more than 12 million years. J Virol 83: 2386–2388.

33. Chen R, Holmes EC (2006) Avian influenza virus exhibits rapid evolutionary dynamics. Mol Biol Evol 23: 2336–2341.

34. Apetrei C, Gautam R, Sumpter B, Carter AC, Gaufin T, et al. (2007) Virus subtype-specific features of natural simian immunodeficiency virus SIVsmm infection in sooty mangabeys. J Virol 81: 7913–7923.

35. Abecasis AB, Lemey P, Vidal N, de Oliveira T, Peeters M, et al. (2007) Recombination confounds the early evolutionary history of human immunodeficiency virus type 1: subtype G is a circulating recombinant form. J Virol 81: 8543–8551.

36. Martin DP, Posada D, Crandall KA, Williamson C (2005) A modified bootscan algorithm for automated identification of recombinant sequences and recombination breakpoints. AIDS Res Hum Retroviruses 21: 98–102.

37. Martin DP, Williamson C, Posada D (2005) RDP2: recombination detection and analysis from sequence alignments. Bioinformatics 21: 260-262.

38. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876–4882.

39. Drummond AJ, Ho SY, Phillips MJ, Rambaut A (2006) Relaxed phylogenetics and dating with confidence. PLoS Biol 4: e88. doi:10.1371/journal.pbio.0040088.

40. Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 7: 214.

41. Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23: 7–9.

42. Drummond AJ, Rambaut A, Shapiro B, Pybus OG (2005) Bayesian coalescent inference of past population dynamics from molecular sequences. Mol Biol Evol 22: 1185–1192.

43. Swofford D (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4. Sunderland (Massachusetts): Sinauer Associates.

44. Salemi M, De Oliveira T, Courgnaud V, Moulton V, Holland B, et al. (2003) Mosaic genomes of the six major primate lentivirus lineages revealed by phylogenetic analyses. J Virol 77: 7202–7213.

Table 1. tMRCA dates of SIVsm/HIV-2 and SIVcpz/HIV-1 clades and comparisons between loci.

Clade	*gag* tMRCA (95% HPD)	*pol* tMRCA (95% HPD)	*env* tMRCA (95% HPD)	P(*gag*<pol)	P(gag<*env*)	P(*env*<*pol*)
SIVsm	1809 (1729–1875)	1686 (1525–1811)	1861 (1788–1915)	0.050	0.870	0.012
HIV-2 A	1932 (1906–1955)	1905 (1857–1949)	1942 (1921–1959)	0.150	0.720	0.060
HIV-2 B	1935 (1907–1961)	1914 (1868–1955)	1937 (1914–1958)	0.210	0.522	0.171
SIVcpz	1618 (1471–1746)	1265 (658–1679)	1492 (1266–1685)	0.051	0.159	0.212
HIV-1 M	1912 (1887–1935)	1795 (1627–1900)	1894 (1857–1927)	0.005	0.218	0.033
HIV-1 N	1966 (1953–1977)	1932 (1876–1968)	1963 (1948–1977)	0.032	0.416	0.053
HIV-1 O	1942 (1922–1958)	1827 (1680–1917)	1905 (1866–1938)	<0.001	0.024	0.062

doi:10.1371/journal.pcbi.1000377.t001

Table 2. tMRCA dates and comparisons between SIVsm/HIV-2 and SIVsm-only analyses.

Locus	Clade	SIVsm/HIV-2 tMRCA (95% HPD)	SIVsm-only tMRCA (95% HPD)	P (SIVsm/HIV-2>SIVsm-only)
gag	SIVsm	1814 (1731–1878)	1745 (1586–1879)	0.777
	SIVsm-1	1959 (1935–1977)	1973 (1964–1981)	0.112
	SIVsm-2	1966 (1951–1978)	1969 (1955–1980)	0.350
	SIVsm-3	1962 (1944–1977)	1958 (1929–1978)	0.590
	SIVsm-4	1980 (1974–1985)	1982 (1977–1986)	0.316
	SIVsm-5	1981 (1968–1993)	1982 (1972–1991)	0.485
	SIVsm-7	1967 (1958–1975)	1966 (1954–1976)	0.526
pol	SIVsm	1697 (1543–1815)	1461 (971–1835)	0.825
	SIVsm-1	1971 (1960–1980)	1973 (1962–1983)	0.362
	SIVsm-2	1975 (1966–1984)	1976 (1967–1985)	0.431
	SIVsm-3	1973 (1956–1986)	1974 (1956–1985)	0.464
	SIVsm-4	1975 (1968–1982)	1976 (1969–1983)	0.418
	SIVsm-5	1972 (1947–1989)	1973 (1950–1990)	0.446
	SIVsm-7	1971 (1964–1976)	1971 (1963–1976)	0.497
env	SIVsm	1880 (1844–1916)	1894 (1852–1927)	0.293
	SIVsm-1	1968 (1958–1977)	1965 (1951–1977)	0.645
	SIVsm-2	1982 (1974–1988)	1979 (1971–1986)	0.694
	SIVsm-3	1972 (1960–1983)	1969 (1955–1981)	0.599
	SIVsm-4	1967 (1958–1975)	1966 (1955–1975)	0.553
	SIVsm-5	1982 (1969–1994)	1977 (1963–1989)	0.714
	SIVsm-7	1968 (1960–1974)	1967 (1957–1974)	0.558

doi:10.1371/journal.pcbi.1000377.t002

Table 3. Percent change of root tMRCA estimates from single-gene to partition analysis.

Taxa	Coalescent Model	Change from Single-Gene to Partition Analysis		
		gag	*pol*	*Env*
SIVsm/HIV-2	Constant	−6.8%	11.0%	−28.5%
	Exponential	−14.5%	42.4%	−38.7%
SIVcpz/HIV-1	Constant	−12.2%	5.1%	0.01%
	Exponential	−13.2%	4.0%	2.5%

doi:10.1371/journal.pcbi.1000377.t003

Table 4. Years between SIVsm and SIVcpz roots and the tMRCA of HIV lineages.

Clade	*gag* years (95% HPD)	*pol* years (95% HPD)	*env* years (95% HPD)
SIVsm root to HIV-2 A	122.8 (57.2–199.9)	218.6 (86.1–369.6)	80.8 (27.7–151.1)
SIVsm root to HIV-2 B	126.2 (59.2–203.7)	227.5 (101.3–379.4)	76.1 (25.1–148.0)
SIVcpz root to HIV-1 M	293.8 (176.2–424.6)	529.4 (207.6–979.8)	402.8 (231.0–601.4)
SIVcpz root to HIV-1 N	347.2 (224.6–489.0)	666.0 (290.1–1225.6)	471.6 (291.6–693.2)
SIVcpz root to HIV-1 O	323.4 (202.0–458.2)	561.5 (229.0–1042.2)	413.5 (247.1–621.3)

doi:10.1371/journal.pcbi.1000377.t004

Table 5. SIV and HIV alignments used in BMCMC analyses.

Taxa	Locus	Number of sequences	Length of sequences (nucleotides)	Range of sampling
SIVsm/HIV-2	*gag*	189	477	1975–2005
	pol	155	612	1975–2005
	env	181	438	1975–2005
SIVsm only	*gag*	166	477	1975–2004
	pol	134	612	1975–2005
	env	155	438	1975–2005
SIVcpz/HIV-1	*gag*	178	666	1983–2005
	pol	179	801	1983–2005
	env	178	582	1983–2005

doi:10.1371/journal.pcbi.1000377.t005

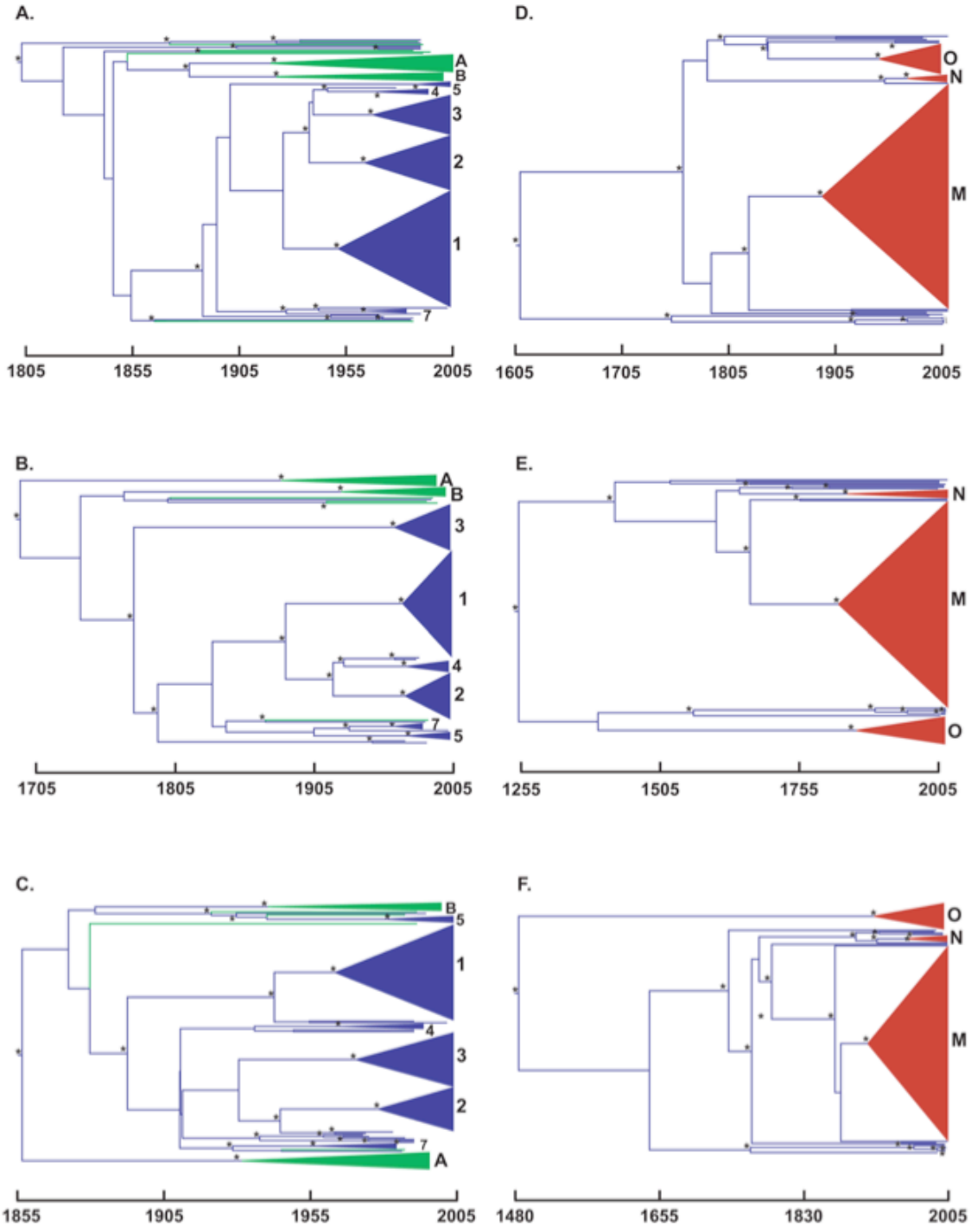


Figure 1. Maximum clade credibility trees. SIVsm/HIV-2 loci (A) *gag*, (B) *pol*, and (C) *env* and SIVcpz/HIV-1 loci (D) *gag*, (E) *pol*, and (F) *env* are depicted. SIV lineages are blue, HIV-2 lineages are green, and HIV-1 lineages are red. Major SIVsm clades are designated by number and the HIV clades are designated by group. Nodes with posterior probability >0.9 are indicated with an asterisk.
doi:10.1371/journal.pcbi.1000377.g001

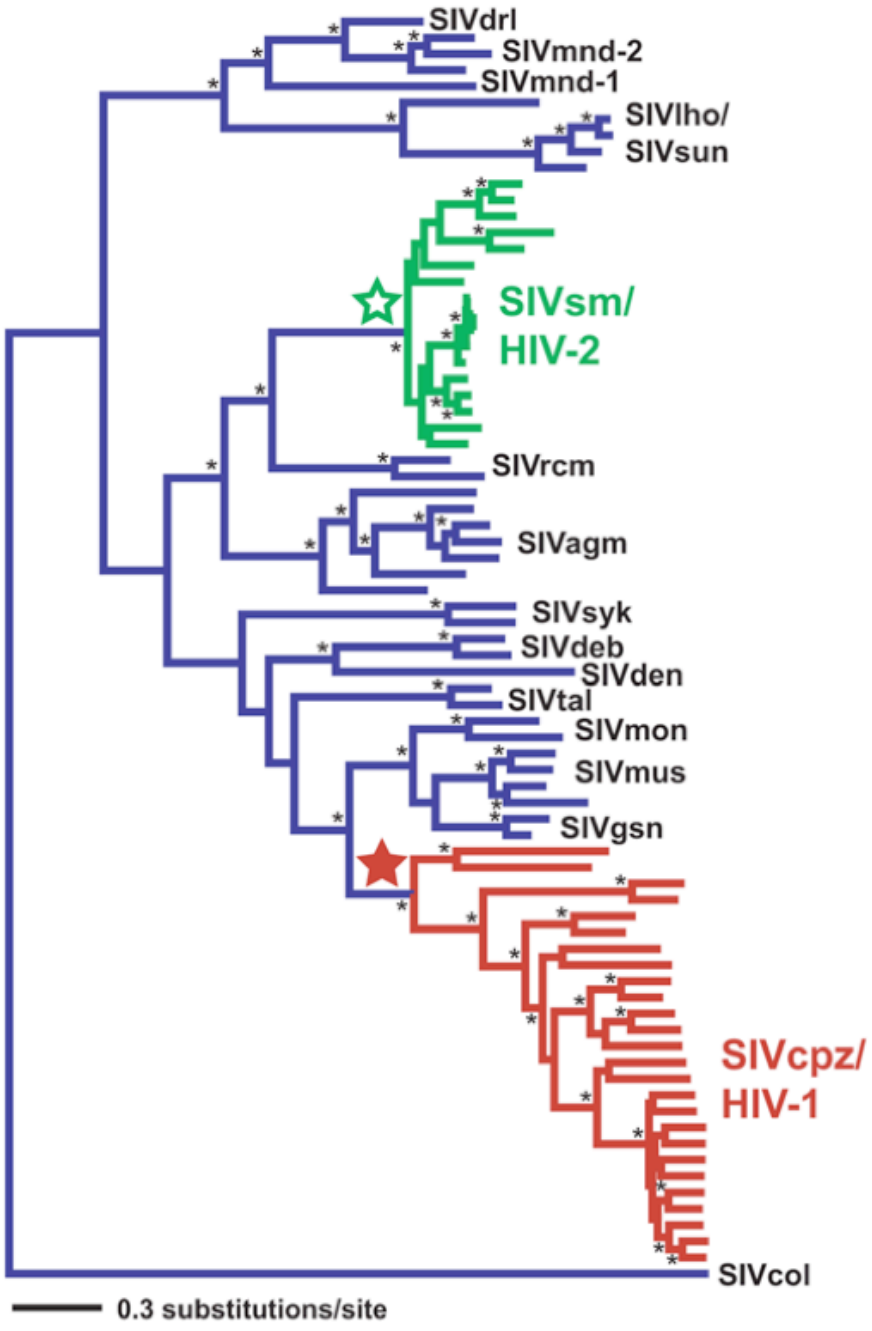


Figure 2. Maximum likelihood phylogeny of SIV/HIV *env* locus. SIVsm/HIV-2 lineages are green and their MRCA is designated with an open green star. SIVcpz/HIV-1 lineages are red and their MRCA is designated with a closed red star. All other SIV lineages are blue. Tree is midpoint rooted. Nodes with bootstrap support >80% are indicated with an asterisk.

doi:10.1371/journal.pcbi.1000377.g002

APPENDIX C: A QUICK FUSE AND THE EMERGENCE OF TAURA SYNDROME VIRUS

Published: ***Virology*** **(2009) 390(2): 324-329**

Co-Authors: Kathy F.J. Tang, Solangel A. Navarro, Donald V. Lightner

ELSEVIER LICENSE
TERMS AND CONDITIONS
Jul 14, 2009

This is a License Agreement between Joel O Wertheim ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard,Langford Lane Kidlington,Oxford,OX5 1GB,UK
Registered Company Number	1982084
Customer name	Joel O Wertheim
Customer address	1041 E. Lowell St. Tucson, AZ 85721
License Number	2227890282415
License date	Jul 14, 2009
Licensed content publisher	Elsevier
Licensed content publication	Virology
Licensed content title	A quick fuse and the emergence of Taura syndrome virus
Licensed content author	Joel O. Wertheim, Kathy F.J. Tang, Solangel A. Navarro and Donald V. Lightner
Licensed content date	1 August 2009
Volume number	390
Issue number	2
Pages	6
Type of Use	Thesis / Dissertation
Portion	Full article
Format	Print
You are an author of the Elsevier article	Yes
Are you translating?	No
Order Reference Number	
Expected publication date	Sep 2009
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
Value added tax 0.0%	0.00 USD
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

ABSTRACT

Over the last two decades, Taura syndrome virus (TSV) has emerged as a major pathogen in penaeid shrimp aquaculture and has caused substantial economic loss. The disease was first discovered in Ecuador in 1991, and the virus is now globally distributed with the greatest concentration of infections in the Americas and Southeast Asia. To determine the evolutionary history of this virus, we constructed a phylogeny containing 83 TSV isolates from 16 countries sampled over a 16-year period. This phylogeny was inferred using a relaxed molecular clock in a Bayesian Markov chain Monte Carlo framework. We found phylogenetic evidence that the TSV epidemic did indeed originate in the Americas sometime around 1991 (1988–1993). We estimated the TSV nucleotide substitution rate at 2.37×10^{-3} (1.98×10^{-3} to 2.82×10^{-3}) substitutions/site/year within capsid gene 2. In addition, the phylogeny was able to independently corroborate many of the suspected routes of TSV transmission around the world. Finally, we asked whether TSV emergence in new geographic locations operates under a quick fuse (i.e. rapid appearance of widespread disease). Using a relaxed molecular clock, we determined that TSV is almost always discovered within a year of entering a new region. This suggests that current monitoring programs are effective at detecting novel TSV outbreaks.

INTRODUCTION

Taura syndrome virus (TSV) is a major pathogen of the Pacific white shrimp *Penaeus vannamei*, one of the most important aquaculture shrimp species. TSV outbreaks in aquaculture facilities can decimate shrimp populations with a mortality rate ranging from 40 to 100%. Taura syndrome was first recognized in Ecuador in 1991 and reported in 1992 (Jimenez, 1992). The disease was then seen in Colombia in 1993, Honduras in 1994, and Mexico in 1995. By the end of the decade it had spread to Southeast Asia. As of 2009, TSV has been isolated on five continents (Cheng et al., 2003, Côté et al., 2008, Do et al., 2006, Hasson et al., 1999, Nielsen et al., 2005, Tang and Lightner, 2005 and Tu et al., 1999).

TSV is a small, non-enveloped, icosahedral virus containing a single-stranded positive-sense RNA genome (Bonami et al., 1997 and Mari et al., 2002) and is a member of the family Dicistroviridae (Mayo, 2005). Its genome contains two open reading frames. The first open reading frame encodes non-structural proteins including helicase, protease, and an RNA-dependent RNA polymerase, and the second encodes three capsid proteins: CP1 (40 kDa), CP2 (55 kDa), and CP3 (24 kDa) (Mari et al., 2002). CP2 has previously been used to differentiate among TSV isolates because it possesses greater genetic variation than other TSV capsid genes (Tang and Lightner, 2005).

Although phylogenetic trees have been constructed for small groups of TSV isolates (Côté et al., 2008, Mari et al., 2002 and Robles-Sikisaka et al., 2001Robles-Sikisaka et al., 2002 and Tang and Lightner, 2005), a phylogenetic perspective on the origin and global diversity of TSV is still lacking. Furthermore, epidemiological evidence

suggests that TSV has recently expanded out of the Americas to the rest of the world causing significant disease-associated loss as it has spread. This model assumes a quick fuse, in which TSV would be detected shortly after entering a new geographic area due to the almost immediate appearance of widespread disease in shrimp. The alternate scenario is one in which TSV would be present for years, slowly building up numbers of infected shrimp and other wild invertebrates in a given geographic area, before widespread disease is observed. A quick fuse would resemble outbreaks in humans like Ebola virus and SARS (Biek et al., 2006 and Hon et al., 2008), whereas a slow fuse epidemic would be reminiscent of HIV in sub-Saharan Africa and the Americas (Gilbert et al., 2007 and Worobey et al., 2008).

Here we infer a phylogenetic tree for 83 CP2 nucleotide sequences obtained from TSV isolates collected over a 16-year period. We identify several main lineages of TSV. Interestingly, the phylogeny confirms many avenues of transmission of TSV around the world, which were previously hypothesized based on epidemiological data. Finally, this phylogenetic analysis employed a relaxed molecular clock, which allowed us to perform an independent phylogenetic test of quick-versus-slow fuse hypotheses in TSV.

RESULTS

The global emergence of TSV

To determine the evolutionary relationships among the TSV isolates, we inferred a phylogeny assuming a relaxed molecular clock in a Bayesian Markov chain Monte Carlo (BMCMC) framework. This analysis found support for four major TSV lineages (Mexico, Southeast Asia, Belize/Nicaragua, and Venezuela/Aruba) as well as the position of the root of the tree (Fig. 1). The root was located among sequences from the Americas, including Ecuador, Columbia, Honduras, Mexico, and the USA; this placement suggests that all 83 of the TSV isolates analyzed here have their origins in the Americas. The most basal lineage on the tree led to a cluster of strains isolated in Columbia, Ecuador, and the USA. This finding corresponds with the first recognition and description of Taura syndrome in Ecuador in 1991 (Hasson et al., 1995, Jimenez, 1992 and Lightner et al., 1995).

The first major lineage identified consisted of isolates collected in Mexico. Thirteen of the fourteen Mexican isolates used in this study formed a highly supported group (posterior probability = 0.97). Therefore, the majority of Mexican TSV outbreaks were likely the result of a single introduction into the country. The remaining Mexican isolate (MX/95c) may be a separate introduction into Mexico, but there was little support for branching order in this part of the phylogeny. The existence of TSV isolates from Texas (US/96), Hawaii (US/07), and Eritrea (ER/04) interspersed among the Mexican isolates suggests that Mexican TSV was the progenitor of these three outbreaks. The single Mexican isolate (MX/95c) that did not fall within this group had a nucleotide

sequence that was identical to two other strains found in Hawaii (US/94) and Honduras (HO/03).

The second major TSV lineage was from Southeast Asia, possibly originating from Honduras, due to the basal position of a Honduran isolate (HO/98). The existence of Southeast Asian TSV lineage is consistent with previous phylogenetic studies (Nielsen et al., 2005 and Tang and Lightner, 2005). The clustering of all Southeast Asian isolates into a single group had moderate support (posterior probability = 0.78), but the inclusion of HO/98 resulted in notably higher support (posterior probability = 1.0). Within the Southeast Asian lineage, TSV isolated in China, Taiwan, and Thailand were dispersed among each other, suggesting multiple transmission events between these countries. TSV strains from Indonesia may be the result of a single introduction into that country, because all the Indonesian TSV were descendants of a single recent common ancestor (posterior probability = 1.0); however, Indonesian TSV apparently had been subsequently transmitted back to the rest of Southeast Asia because Indonesian TSV was paraphyletic; strains from China and Taiwan were found among the Indonesian strains.

The third major TSV lineage identified here was from Belize and Nicaragua. Isolates from these countries clustered with high support (posterior probability = 1.0), in agreement with previous phylogenetic evidence of a Belize-specific TSV lineage (Tang and Lightner, 2005). But it is not clear which of these countries' shrimp were infected first, because the placement of isolates from Belize at a basal position was not well supported. Nevertheless, it appears that the outbreak in Belize, first noted in 2001, is one prolonged epidemic that has yet to be eradicated from local farms. Isolates from

Nicaragua were paraphyletic because their grouping also included an isolate collected in Saudi Arabia. The Saudi Arabian isolate (SA/07) was identical to a Nicaraguan strain (NI/06). Given this fact and high support for the Saudi Arabian strain sharing a most recent common ancestor with Nicaraguan strains (posterior probability = 1.0), it is highly probable that Saudi Arabia acquired TSV from Nicaragua.

Finally, a fourth phylogenetically distinct TSV, group comprised of Venezuelan and Aruban isolates, was inferred (posterior probability = 1.0). While this lineage had previously been described by Côté et al. (2008), we found that this group also contained isolates from Ecuador. The Venezuelan/Aruban TSV lineage was separated from the other TSV isolates by a long internal branch, suggesting a distinct epidemic.

Introductions into the USA

There were at least five separate introductions of the virus into the USA, according to the TSV phylogeny (Fig. 1). This interpretation assumes that TSV did not originate in the USA, which seems reasonable given that it was not seen in the USA until 1994 and that most isolates from the USA were not closely related to each other. Of the six USA isolates sequenced, only two were closely related: Hawaii in 1995 (US/95) and Texas in 1998 (US/98). Two separate introductions into the USA, Texas (US/96) and Hawaii (US/07), fell within the main Mexican lineage. Another USA lineage, isolated in Texas (US/04) fell within a TSV group from Southeast Asia. The TSV isolate from Hawaii (US/94), whose sequence was identical to Mexican and Honduran isolates, fell within an unresolved portion deep in the phylogeny. It is likely that this lineage arose

from TSV in the Americas, although the precise geographical location is unknown. Overall, TSV from the USA has generally been relegated to sporadic, epidemiologically unlinked appearances on the phylogeny.

A quick fuse on the TSV epidemic

By calibrating a molecular clock using TSV sequences sampled between 1993 and 2008, we were able to estimate the time of most recent common ancestor (tMRCA) for many TSV clades. According to the BMCMC analysis, all of the current TSV lineages shared a most recent common ancestor in 1991 (1988–1993), the same year as the initial identification of Taura syndrome in Ecuador. The tMRCA of the TSV lineages were also closely temporally linked with their discoveries (Table 1). The tMRCA of the Mexican lineage was estimated to be 1994 (1993–1995), one year before the earliest isolate from this group. Isolates from Southeast Asia had a tMRCA in 1998 (1997–1999), the same year this virus was first identified in this part of the world (Tu et al., 1999). Indonesian TSV isolates had a tMRCA in 2000 (1999–2002), shortly before TSV was discovered in this country in 2002. Furthermore, the tMRCA of isolates from Belize and Nicaragua was estimated to be in 2000 (1998–2001), one year before the first Belizean isolate was collected. And TSV was first seen in Nicaragua in 2005, one year after the estimated tMRCA at 2004 (2003–2005). Also, the Saudi Arabian isolate was phylogenetically nested among the Nicaraguan strains. Although the Saudi Arabian isolate was identical to one of the Nicaraguan strains, we were still able to infer that it shared a most recent common ancestor with these strains in 2006 (2005–2006). In

addition, a more recent cluster of Venezuelan isolates had a tMRCA in 2004 (2003–2005), one year before the first Venezuela case was detected. We note that while the mean tMRCA estimates presented here often predate the discovery of TSV by approximately one year, the 95% HPD (highest probability density) of these estimates always includes the year of discovery. Therefore, we cannot distinguish between the possibility that TSV is detected a year after initial infection or immediately upon arrival. Regardless, TSV does not appear to survive undetected for a prolonged period of time in any geographic region, suggesting a quick fuse model.

Evolutionary rate of TSV

We estimated the TSV substitution rate within CP2 at 2.37×10^{-3} (1.98×10^{-3} to 2.82×10^{-3}) substitutions/site/year. Across the phylogeny, however, the molecular clock does not "tick" uniformly. The BMCMC approach used here allowed us to estimate the clock-likeness of TSV by modeling a lognormal distribution of evolutionary rates along branches (Drummond et al., 2006). If, in the relaxed clock BMCMC analysis, the 95% HPD of the standard deviation of the lognormal distribution of rates excludes zero, then a strict molecular clock can be rejected in favor of a relaxed molecular clock. The standard deviation of the lognormal distribution estimated here was 0.80 (0.60–1.02), indicating that our relaxed molecular clock model was appropriate. If the TSV phylogeny were inferred under a strict molecular clock, an apparently flawed assumption, the tMRCA of the root would have been 1979 (1973–1984), substantially older than the estimate of 1991 (1988–1993) inferred under a relaxed molecular clock (Table 2).

TSV population demographics

As a part of the BMCMC analysis, we were also able to infer the demographic history of TSV. We tested whether or not TSV has been evolving at a constant population size for the past 18 years. The BMCMC approach utilized here allows for an explicit test for constant versus expanding population size. When the BMCMC analysis is performed assuming an exponentially expanding population, a mean and 95% HPD for the exponential growth rate is estimated. If the 95% HPD of the exponential growth rate does not include zero, then a constant population size can be rejected. In fact, this scenario was observed for TSV under an exponential growth model. The exponential growth rate was estimated to be 0.18 (0.10–0.26), suggesting that TSV has not evolved under a constant population size since 1991. The root tMRCA mean and 95% HPD estimates inferred under these different coalescent models varied, and the most restrictive models produced the oldest tMRCAs (Table 2). Here, we report the mean and 95% HPD ages from the Bayesian skyline plot (BSP), because this model places the fewest prior constraints on the data and allows the population size to change over time in an unconstrained fashion (Drummond et al., 2005).

DISCUSSION

We were able to reconstruct the evolutionary history of the global emergence of TSV using only CP2 sequences from 83 TSV isolates collected over a 16-year period. The TSV phylogeny inferred here suggests that the current epidemic affecting penaeid shrimp aquaculture has its origin in the Americas. The tMRCA of the current isolates was 1991 (1988–1993), the same time Taura syndrome was first described in Ecuador (Jimenez, 1992). We also identified major TSV lineages and estimated their tMRCAs. In every case, the relaxed molecular clock analysis indicated that the tMRCA was at or right before TSV was discovered in these regions. These results support the hypothesis that TSV epidemics have a short fuse and spread rapidly; this virus does not appear to lay low for many years in shrimp aquaculture or wild invertebrate populations.

The TSV transmission routes deduced from the inferred phylogeny are remarkably similar to many commonly held theories about the spread of TSV around the world (the following discussion was assembled from published work, when available, and the archives of the Aquaculture Pathology Lab at the University of Arizona) (Fig. 1). The first route corroborated by the phylogeny was from Sonora, Mexico to Eritrea. The earliest known Mexican outbreaks occurred in Sonora (Hasson et al., 1999), and producers in Sonora were known to ship broodstock to other areas, including Eritrea. The Eritrean TSV isolate (ER/04) is nested within the main Mexican TSV lineage. Another isolate that the phylogeny suggests is of Mexican origin is from Hawaii (US/07); this outbreak was thought to be the result of contamination when a Hawaiian shrimp farm stored and sold imported frozen Mexican shrimp at a road-side lunch stand for tourists

that was located adjacent to its hatchery and farm ponds. This same pattern of storing and selling imported shrimp (in another case, from Ecuador) was also suspected in the 1995 TSV outbreak in Hawaii. One of the affected farms in Hawaii operated a road-side lunch stand which was run by a company that also owned a TSV-infected farm in Ecuador; the Hawaiian isolate (US/95) is closely related to TSV from Ecuador (EC/94). In addition, the Saudi Arabian TSV outbreak was thought to be the result of imported shrimp from Nicaragua, and the TSV phylogeny depicts the Saudi isolate (SA/07) nested within the Nicaraguan isolates (NI/05 and NI/06). Also, the Venezuelan and Aruban isolates are closely related, and the infected shrimp were from two farms owned by the same company in which broodstock and post-larvae were frequently shipped between the two facilities. Furthermore, the introduction of TSV into Southeast Asia was thought to be the result of the transport of infected shrimp from the Gulf of Fonseca in Honduras to Taiwan (Lien et al., 2002 and Yu and Song, 2000). The TSV phylogeny shows that all of Southeast Asian TSV is closely related to two Honduran isolates (HO/94 and HO/98), with the Taiwanese isolate (TW/99) as the basal Southeast Asian lineage; however, the position of TW/99 is not well supported. Finally, the position of US/04 from Texas nested within the Southeast Asian group is not unexpected given that this outbreak occurred in a Texan shrimp farm that was located on the same bay as a processing plant that had imported *P. vannamei* from China a few weeks before the outbreak.

In the USA, where disease monitoring has been more rigorous, TSV outbreaks have been sporadic. Sequence data are available only for six cases in Texas and Hawaii, although there have been additional USA outbreaks. Only two of the outbreaks analyzed

here resulted from closely related TSV isolates: Hawaii (US/95) and Texas (US/98) (Fig. 1). It is unlikely, however, that these isolates are epidemiologically linked; they were sampled three years apart, and US/95 is likely the result of imported Ecuadorian shrimp. The other four USA TSV isolates included in our phylogenetic analysis are all distantly related to each other and likely the result of independent introductions from around the world. The phylogenetic distance separating the majority of TSV isolates from the USA indicates that the monitoring and containment protocols employed by USA shrimp aquaculture are highly effective at detecting outbreaks and preventing further spread.

The phylogenetic analysis also provided insight into the rate of TSV evolution. The TSV substitution rate of 2.37×10^{-3} (1.98×10^{-3} to 2.82×10^{-3}) substitutions/site/year is similar to rapidly evolving RNA viruses such as swine vesicular disease virus (3.4×10^{-3} substitutions/site/year), human enterovirus 71 (3.4×10^{-3} substitutions/site/year), and human immunodeficiency virus type-1 (2.5×10^{-3} substitutions/site/year) (Jenkins et al., 2002). The high rate of substitution accounts for the rapidly increasing diversity of TSV associated with its spread across the globe. This increase in diversity, in turn, may result in the need for perpetual development of new TSV RT-PCR primers and for a continued search to find new populations of TSV-resistant *P. vannamei*.

Although the phylogeny inferred here proved useful at resolving the evolutionary relationships among many of the TSV isolates, it was unable to resolve the branching order at the base of the tree. This polytomy might be due to the rapid diversification experienced by TSV during the early portion of the epidemic. Alternatively, it might be

due to the short length of the CP2 sequence used here. Full-length TSV genome sequences may prove useful at resolving these basal branching patterns, which would help provide phylogenetic evidence for a country of origin for the TSV epidemic.

MATERIALS AND METHODS

Virus samples

Sixty-two samples of penaeid shrimp infected with TSV collected between 1993 and 2008 from 16 countries were obtained from the archives of diagnostic samples at the Aquaculture Pathology Lab at University of Arizona (GenBank Accession numbers: FJ876460, FJ876461, FJ876462, FJ876463, FJ876464, FJ876465, FJ876466, FJ876467, FJ876468, FJ876469, FJ876470, FJ876471, FJ876472, FJ876473, FJ876474, FJ876475, FJ876476, FJ876477, FJ876478, FJ876479, FJ876480, FJ876481, FJ876482, FJ876483, FJ876484, FJ876485, FJ876486, FJ876487, FJ876488, FJ876489, FJ876490, FJ876491, FJ876492, FJ876493, FJ876494, FJ876495, FJ876496, FJ876497, FJ876498, FJ876499, FJ876500, FJ876501, FJ876502, FJ876503, FJ876504, FJ876505, FJ876506, FJ876507, FJ876508, FJ876509, FJ876510, FJ876511, FJ876512, FJ876513, FJ876514, FJ876515, FJ876516, FJ876517, FJ876518, FJ876519, FJ876520, FJ876521 and FJ876522). Another 21 CP2 sequences previously reported in the GenBank were included in the analysis (GenBank Accession numbers: AF277675, AF277378, AF406789, AY755587, AY755588, AY755589, AY755590, AY755591, AY755592, AY755593, AY755594, AY755595, AY755596, AY755597, AY755598, AY755599, AY755600, AY755601 and AY755602, AY355311, and DQ104696). Collection information, identification codes and accession numbers for all sequences are provided as Supplementary material (Table S1). Nucleotide positions given here are in reference to the previously published complete genome sequence of a Hawaiian isolate: GenBank Accession number AF277675.

Purification of total RNA, TSV CP2 RT-PCR, and sequencing

Total RNA was extracted, from either pleopods or gills of shrimp samples, with a High Pure RNA tissue kit (Roche Biochemical, Indianapolis, IN). RT-PCR was performed with a SuperScript one-step RT-PCR system with Platinum Taq DNA polymerase (Invitrogen, Carlsbad, CA). The CP2 region of the TSV genome (1303 nucleotides in length) was amplified with primers 55P1 (nt 7901–7920, 5′-GGC GTA GTG AGT AAT GTA GC-3′) and 55P2 (nt 9184–9203, 5′-CTT CAG TGA CCA CGG TAT AG-3′). The RT-PCR profile was 30 min at 55 °C, followed by 40 cycles of 30 s at 94 °C, 30 s at 55 °C, and 90 s at 68 °C. An aliquot of amplified product was visualized in a 1% agarose gel containing ethidium bromide. DNA sequencing was performed by the Genomic Analysis and Technology Core facility (University of Arizona, Tucson, AZ) using Sanger sequencing (Applied Biosystems 3730 DNA Analyzer).

BMCMC analysis

Phylogenetic inference and tMRCA estimation was performed in a BMCMC framework in BEAST v1.4.8 (Drummond et al., 2006 and Drummond and Rambaut, 2007). We compared two nucleotide substitution models (GTR + Γ_4 and SRD06) (Shapiro et al., 2006) and four demographic scenarios (constant population size, exponential growth, expansion growth, and BSP) using Bayes factor in Tracer v1.4 (http://beast.bio.ed.ac.uk/Tracer). SRD06 performed better than GTR + Γ_4 (Bayes factor > 20), although there were no appreciable differences in the tMRCA estimates under these two substitution models (Table 2). We found no difference among the performance

under the four demographic scenarios using Bayes factor. When the translated amino acid sequences were analyzed under a Blosum62 substitution model in BEAST, the tMRCA estimates were also not appreciably different from the nucleotide-based analyses (Table 2).

For each analysis, two BMCMC runs of 25 million generations were performed using an uncorrelated lognormal relaxed molecular clock. Tracer was used to check for convergence of parameter estimations and proper mixing (estimated sampled size > 200). The position of the root of the tree was determined as a byproduct of the BMCMC. The posterior trees from the SRD06/BSP analysis were annotated using the Maximum Clade Credibility tree. BMCMC analysis was also performed without sequence data to better understand how the prior probabilities were affecting the posterior distribution of the tMRCA estimates. This analysis determined that the priors were not biasing our tMRCA estimates.

ACKNOWLEDGMENTS

We thank Michael J. Sanderson for providing computational resources. This work was supported by the NSF-IGERT (Integrative Graduate Education and Research Traineeship) in Comparative Genomics at the University of Arizona and by grants from the USDA, CSREES US Marine Shrimp Farming Program, and the National Fisheries Institute.

REFERENCES

Biek, R., Walsh, P.D., Leroy ,E.M., Real, L.A. 2006. Recent common ancestry of Ebola Zaire virus found in a bat reservoir. PLoS Pathog. 2, e90. doi:10.1371/journal.ppat.0020090.

Bonami, J.R., Hasson, K.W., Mari, J., Poulos, B.T., Lightner, D.V., 1997. Taura syndrome of marine penaeid shrimp: characterization of the viral agent. J. Gen. Virol. 78, 13–19.

Cheng, X.Z., Gong, Y.Q., Kong, F.D. Wang, J.M., Zhou, B.H., Zheng, Z., Huang, Y.C., 2003. Detection of Taura syndrome virus in imported parent Litopenaeus vannamei by using reverse transcription polymerase chain rection. Chin. J. Animal Quarant. 20, 21–22. (In Chinese).

Côté, I., Navarro, S.A., Tang, K.F.J., Lightner, D.V., 2008. Taura syndrome virus from Venezuela is a new genetic variant. Aquaculture 284, 62–67.

Do, J.W., Cha, S.J., Lee, N.S., Kim, Y.C., Kim, J.W., Kim, J.D., Park, J.W., 2006. Taura syndrome virus from *Penaeus vannamei* shrimp cultured in Korea. Dis. Aquat. Org. 70, 171–174.

Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Mol. Biol. Evol. 22**,** 1185–1192.

Drummond, A.J., Ho, S.Y., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4**,** e88. doi:10.1371/journal.pbio.0040088.

Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7**,** 214.

Gilbert, M.T., Rambaut, A., Wlasiuk, G., Spira, T.J., Pitchenik, A.E., Worobey, M., 2007. The emergence of HIV/AIDS in the Americas and beyond. Proc. Natl. Acad. Sci. USA 104, 18566–18570.

Hasson, K.W., Lightner, D.V., Poulos, B.T., Redman, R.M, White, B.L., Brock, J.A., Bonami, J.R., 1995. Taura syndrome in *Penaeus vannamei*: demonstration of a viral etiology. Dis. Aquat. Organ. 23, 115–126.

Hasson, K.W., Lightner, D.V., Mari, J., Bonami, J.R., Poulos, B.T., Mohney, L.L., Redman, R.M., Brock, J.A., 1999. The geographic distribution of Taura syndrome virus (TSV) in the Americas: determination by histopathology and in situ hybridization using TSV-specific cDNA probes. Aquaculture 171, 13–26.

Hon, C.C., Lam, T.Y., Shi, Z.L., Drummond, A.J., Yip, C.W., Zeng, F., Lam, P.Y., Leung, F.C., 2008. Evidence of the recombinant origin of a bat severe acute respiratory syndrome (SARS)-like coronavirus and its implications on the direct ancestor of SARS coronavirus. J. Virol. 82, 1819–1826.

Jenkins, G.M., Rambaut, A., Pybus, O.G., Holmes, E.C., 2002. Rates of molecular evolution in RNA viruses: a quantitative phylogenetic analysis. J. Mol. Evol. 54**,** 156–65.

Jimenez, R., 1992. Sindrome de Taura (Resumen). In: Acuacultura del Ecuador. Camara Nacional de Aquacultura, Guayaquil, Ecuador, pp. 1–16.

Lien, T.W., Hsiung, H.C., Huang, C.C., Song, Y. 2002. Genomic similarity of Taura syndrome virus (TSV) between Taiwan and Western Hemisphere isolates. Fish Pathology 37, 71–75.

Lightner, D.V., Redman, R.M., Hasson, K.W., Pantoja, C.R., 1995. Taura syndrome in *Penaeus vannamei* (*Crustacea: Decapoda*): gross signs, histopathology and ultrastructure. Dis. Aquat. Organ. 21, 53–59.

Mari, J., Poulos, B.T., Lightner, D.V., Bonami, J.R., 2002. Shrimp Taura syndrome virus: genomic characterization and similarity with members of the genus Cricket paralysis-like viruses. J. Gen. Virol. 83 915–926.

Mayo, M.A., 2005. Changes to virus taxonomy 2004. Arch. Virol. 150: 189-198.

Nielsen, L., Sang-oum, W., Cheevadhanarak, S., Flegel, T.W., 2005. Taura syndrome virus (TSV) in Thailand and its relationship to TSV in China and the Americas. Dis. Aquat. Org. 63, 101–106.

Robles-Sikisaka, R., Garcia, D.K., Klimpel, K.R., Dhar, A.K., 2001. Nucleotide sequence of 3'-end of the genome of Taura syndrome virus of shrimp suggests that it is related to insect picornaviruses. Arch. Virol. 146, 941–952.

Robles-Sikisaka, R., Hasson, K.W., Garcia, D.K., Brovont, K.E., Cleveland, K.D., Klimpel, K.R., Dhar, A.K., 2002. Genetic variation and immunohistochemical differences among geographic isolates of Taura syndrome virus of penaeid shrimp. J. Gen. Virol. 83, 3123–3130.

Shapiro, B., Rambaut, A., Drummond, A.J., 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23**,** 7–9.

Tang, K.F.J., Lightner, D.V., 2005. Phylogenetic analysis of Taura syndrome virus isolates collected between 1993 and 2004 and virulence comparison between two isolates representing different genetic variants. Virus Res. 112, 69–76.

Tu, C., Huang, H.T., Chuang, S.H., Hsu, J.P., Kuo, S.T., Li, N.J., Hsu, T.L., Li, M.C., Lin, S.Y., 1999. Taura syndrome in Pacific white shrimp *Penaeus vannamei* cultured in Taiwan. Dis. Aquat. Org. 38,159–161.

Worobey, M., Gemmel, M., Teuwen, D.E., Haselkorn, T., Kunstman, K., Bunce, M., Muyembe, J.J., Kabongo, J.M., Kalengayi, R.M., Van Marck, E., Gilbert, M.T., Wolinsky, S.M. 2008. Direct evidence of extensive diversity of HIV-1 in Kinshasa by 1960. Nature 455, 661–664.

Yu, C.I. and Song, Y.L. 2000. Outbreaks of Taura syndrome in pacific white shrimp *Penaeus vannamei* cultured in Taiwan. Fish Pathology 35, 21–24.

Table 1
Year of Taura syndrome discovery and estimates of mean tMRCA dates and 95% HPDs for TSV lineages.

Lineage	Year of discovery	tMRCA date (95% HPD)
All TSV (root of phylogeny)	1991	1991 (1988–1993)
Aruba		2004 (2003–2005)
Belize/Nicaragua/Saudi Arabia		2000 (1998–2001)
Belize	2001	2000 (1999–2001)
Ecuador	1991	1991 (1989–1993)
Indonesia	2002	2000 (1999–2002)
Mexico (excluding MX/95c)	1995	1994 (1993–1995)
Nicaragua/Saudi Arabia (NI/06 and SA/07)		2006 (2005–2006)
Nicaragua	2005	2004 (2003–2005)
Southeast Asia	1998	1998 (1997–1999)
Southeast Asia/Honduras		1997 (1996–1998)
Venezuela/Aruba		2003 (2002–2005)
Venezuela	2005	2004 (2003–2005)

Table 2
Mean tMRCA and 95% HPD for the TSV root under various substitution, molecular clock, and coalescent models.

Substitution/clock/coalescent models	tMRCA (95% HPD)
SRD06/relaxed/BSP	1991 (1988–1993)
GTR + Γ_4/relaxed/BSP	1991 (1988–1993)
Blosum62/relaxed/BSP	1992 (1990–1993)
SRD06/relaxed/constant	1980 (1956–1992)
SRD06/relaxed/expansion	1987 (1976–1992)
SRD06/relaxed/exponential	1989 (1986–1992)
SRD06/strict/BSP	1979 (1973–1984)

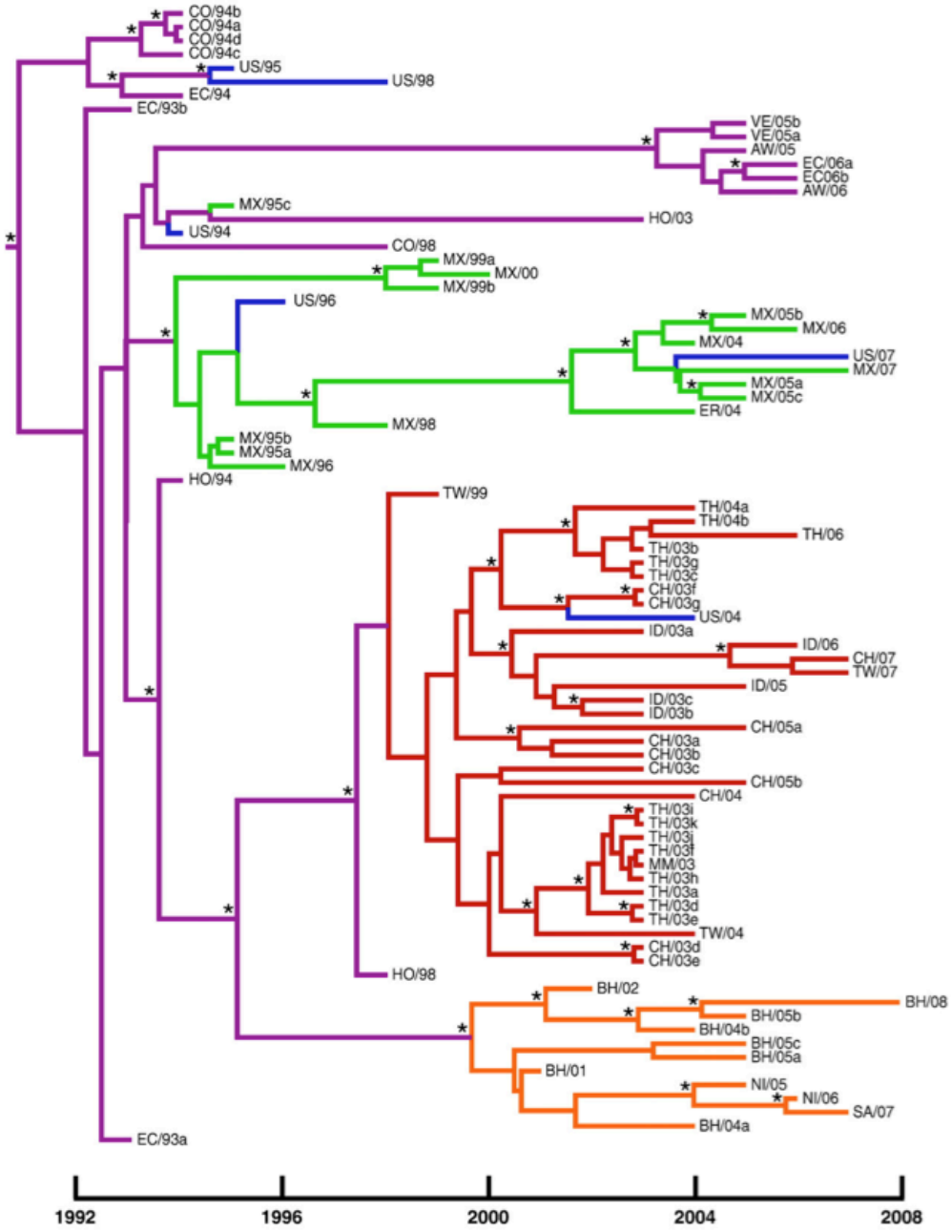


Fig. 1. Maximum clade credibility tree depicting evolutionary relationships among the TSV isolates. Lineages from China (CH), Indonesia (ID), Myanmar (MM), Taiwan (TW), and Thailand (TH) are red. Lineages from Belize (BH), Nicaragua (NI), and Saudi Arabia (SA) are orange. Lineages from Mexico (MX) and Eritrea (ER) are green. Lineages from USA (US) are blue. Lineages from Aruba (AW), Colombia (CO), Ecuador (EC), Honduras (HO), and Venezuela (VE) are purple. Nodes with posterior probability greater than 0.9 are indicated with an asterisk.

APPENDIX D: RELAXED MOLECULAR CLOCKS, THE BIAS-VARIANCE TRADE-OFF, AND THE QUALITY OF PHYLOGENETIC INFERENCE

In Press: ***Systematic Biology***

Co-authors: Michael J. Sanderson, Michael Worobey, Adam Bjork

ABSTRACT

Because a constant rate of DNA sequence evolution cannot be assumed to be ubiquitous, relaxed molecular clock inference models have proven useful when estimating rates and divergence dates. Furthermore, it has been recently suggested that using relaxed molecular clocks may provide superior accuracy and precision in phylogenetic inference compared to traditional time-free methods that do not incorporate a molecular clock. We perform a simulation study to determine if assuming a relaxed molecular clock does indeed improve the quality of phylogenetic inference. We analyze sequence data simulated under various rate distributions using relaxed-clock, strict-clock, and time-free Bayesian phylogenetic inference models. Our results indicate that no difference exists in the quality of phylogenetic inference between assuming a relaxed molecular clock and making no assumption about the clock-likeness of sequence evolution. This pattern is likely due to the bias-variance trade-off inherent in this type of phylogenetic inference. We also compared the quality of inference between Bayesian and maximum likelihood time-free inference models and found them to be qualitatively similar.

INTRODUCTION

The concept of a molecular clock has played a central role in evolutionary biology since its introduction nearly half a century ago by Zuckerkandl and Pauling (1962). Despite its auspicious beginnings, however, the concept of a universal, strict molecular clock has fallen out of favor (Li, 1993; Ayala, 1997; Bromham and Penny, 2003; Kumar, 2005). It is now widely recognized that nucleotide and amino acid substitutions do not generally accumulate at a constant and universal rate even across closely related lineages. Instead, the molecular clock fluctuates. So-called relaxed molecular clock inference models lie on a continuum between strict-clock inference models, which assume a constant evolutionary rate across lineages, and time-free inference models, which do not incorporate evolutionary rates across lineages at all.

Relaxed molecular clocks were introduced by Sanderson (1997, 2002) and Thorne et al. (1998) to estimate the time to most recent common ancestor (tMRCA) in the absence of rate constancy. Their models assumed that the sequences evolve with an inherent temporal component, even though this clock does not tick uniformly across the entire phylogeny or through time. Sanderson's method relied upon semi-parametric penalized likelihood estimation, whereas Thorne et al. embedded the problem of rate estimation in a Bayesian Markov chain Monte Carlo (BMCMC) framework; an expectation of auto-correlation of rates along closely related branches is a feature of both methods. More recently, developments in BMCMC relaxed-clock phylogenetic inference models have allowed uncorrelated rates to be sampled from a variety of distributions, including exponential and lognormal (Drummond et al., 2006). These rate distributions

differ in their assumptions of where on the phylogeny changes in the evolutionary rates occur: at internal nodes (exponential) or along branches (lognormal). Drummond et al. (2006) modeled uncorrelated rates, because their phylogenetic analysis suggested that auto-correlation of rates is not predominant c. While testing these new relaxed-clock inference models, Drummond et al. (2006) put forth the intriguing proposition that incorporation of relaxed molecular clocks might improve the topological accuracy and precision of phylogenetic inference. If true, relaxed molecular clock inference models should supersede traditional time-free phylogenetic analyses, whether or not estimations of substitution rates or tMRCA are desired.

Correctly modeling nucleotide substitution parameters generally increases the probability of inferring the correct phylogenetic tree. This pattern has been demonstrated for the classic four-taxon tree using simulated sequence data (Gaut and Lewis, 1995) as well as for real sequence data (Sullivan and Swofford, 1997). These observations have led to the development and implementation of more realistic models of molecular sequence data, including unequal base frequencies (Felsenstein, 1981), rate heterogeneity (Yang, 1993), and codon position partitioning (Shapiro et al., 2006), along with computational tools designed to determine the appropriate model for a given data set (Posada and Crandall, 1998). Furthermore, seminal work by Huelsenbeck and Hillis (1993) explicitly examined the ability of inference models that assumed a strict molecular clock to reconstruct a tree from sequence data that clearly violated this assumption. They found that, although this model correctly inferred phylogenies for clock-like data, it fared extremely poorly on non clock-like data. Therefore, it seems reasonable to expect that if

one can correctly model the rate of evolution along the branches of a tree, one should better be able to correctly infer the topology of that tree.

Statistical theory, however, does not necessarily support this supposition because of the bias-variance trade-off (Burnham et al., 2002). Bias reflects the ability of a model to accurately predict the data, whereas variance refers to the sensitivity of the model to the sampled data. As variance increases, the precision of the estimate deceases. A model that under-fits the data, because it has fewer parameters, is generally highly biased but has low variance. A low-parameter model may not be realistic, but it might be useful when encountering new data. Increasing the number of parameters may well increase the fit of the model to the data, but this comes at the expense of a decrease in both explanatory power and the precision of estimates. Theoretically, the best model is one with an intermediate number of parameters that simultaneously minimizes bias and variance. The question remains whether, in practice, modeling rate variation among branches can improve phylogenetic inference.

Drummond et al. (2006) set out to answer this question by testing the quality of relaxed-clock, strict-clock, and time-free inference models in a variety of taxa, including bacteria, yeast, and mammals. They inferred a 'true tree' from large sequence datasets, broke these datasets into sub-regions, and compared the inferred phylogenies for each of the sub-regions to the 'true tree.' Their results suggested that relaxed clocks provide more accurate and precise phylogenetic inference; however, their analyses had several limitations. First, their data sets contained relatively few (eight or nine) taxa and their 'true trees' were highly asymmetrical. Second, given the nature of coalescent processes

and horizontal gene transfer, their ‘true tree’ was likely the incorrect tree for many sub-regions (Ochman et al., 2000; Edwards et al., 2007). Finally, their conclusion regarding the superiority of relaxed molecular clocks was not accompanied by statistical analyses. In many cases, the differences in accuracy and precision among the clock models were slight or non-existent.

Here, we study whether or not the assumption of a relaxed molecular clock significantly improves the quality of phylogenetic inference. We simulated sequence data under relaxed-clock and strict-clock scenarios and inferred phylogenies under the assumptions of various clock models. Our findings shed light on the bias-variance trade-off in phylogenetic inference, find little evidence in support of Drummond et al.’s (2006) conclusions, and suggest that additional metrics beyond accuracy and precision are needed to determine whether relaxed clocks improve the quality of phylogenetic topological reconstructions.

METHODS

Sequence Simulation

We constructed 800 sequence alignments that conformed to several models of sequence evolution (Fig. 1). First, we used APE (Paradis et al., 2004) to simulate 200 ultrametric trees ranging in size from 5 to 50 taxa, in 5-taxon intervals (i.e., 20 trees per interval). Individual branch lengths are the product of the time elapsed between nodes and the rate of evolution along a branch. The study by Drummond et al. (2006) explicitly recognized that all sequences evolve with an inherent temporal component. Therefore, we manipulated only the rate component along each branch, by sampling from distributions comprised of 10,000 'rates'. Specifically, four rate distributions (exponential, lognormal, strict, and uniform) were separately applied to each of the 200 tree topologies, and each branch was assigned its own randomly selected number (Fig. 2). These trees are available as supplemental online data files (http://www.sysbio.oxfordjournals.org). The exponential (mean and standard deviation equal to 0.01) and lognormal (mean equal to 0.01 and variance equal to 0.5) distributions represent relaxed-clock models of sequence evolution. The shapes of these rate distributions were based on previous simulations by Drummond et al. (2006). The strict distribution, representing a strict molecular clock, was defined by a single value (1). The uniform distribution (range from 0.0001 to 1.0) is also a relaxed clock model, which essentially minimized the model's information about rates among all possible probability distributions, but retained the biologically relevant assumption that all sequences evolve over time. We emphasize that this uniform

distribution is not intended to reflect the assumptions made by the time-free phylogenetic inference model.

After the heights (i.e., time from root to tip) of these 800 phylogenies were standardized in TreeEdit (Rambaut and Charleston, 2002), we proceeded to generate sequence data for each tree using Seq-Gen v1.5.3 (Rambaut and Grassly, 1997). Each sequence generated was 1,000 bases in length and was evolved according to an HKY + Γ_4 ($\kappa = 2$; $\alpha = 1$) substitution matrix. To incorporate variable root height into the data, each tree's root height was scaled by a random integer (1–30) in Seq-Gen. This scaling created alignments with uncorrected pairwise distances consistent with biologically relevant sequence data used in studies of molecular evolution (approximately 2 to 40 percent maximum pairwise distance).

Phylogenetic Analysis

Each of these 800 alignments was analyzed using four different molecular clock models utilizing BMCMC phylogenetic inference (two relaxed clock models, one strict clock model, and one time-free model where no estimation of rates is performed). The two relaxed-clock inference analyses and strict-clock inference analysis were performed using BEAST v1.4.6 (Drummond and Rambaut, 2007) under an HKY + Γ_4 substitution model. Uninformative priors were assigned for both kappa and alpha. Each analysis was performed for 30,000,000 generations and the first 10% were removed as a burnin. For each run, 9000 post-burnin trees were sampled. Convergence of the BMCMC was confirmed using Tracer v1.4 (Rambaut and Drummond, 2007). If the effective sample

size (ESS) for a given parameter was <100, the analysis was rerun for up to 100,000,000 generations until the ESS values for all parameters were > 100. Root height ESS values of <100 were not addressed, as the subsequent analyses were performed on unrooted trees (see below). BEAST infers the position of the root as a byproduct of its rate estimation analysis. Twenty-nine percent of the BEAST analyses needed to be rerun. BEAUti templates, the input files for BEAST, for each inference model are available as online Appendices 1-3 (http://www.sysbio.oxfordjournals.org).

Time-free phylogenetic analysis [i.e., what Drummond et al. (2006) referred to as the unrooted-Felsenstein model] was performed using MrBayes v3.1 (Ronquist and Huelsenbeck, 2003) under an HKY + Γ_4 substitution model. Time-free analysis is not an available feature of BEAST. Each MrBayes analysis was performed for 1,000,000 generations and the first 10% were removed as burnin. If ESS values for a given parameter were <100, the analysis was rerun for up to 3,000,000 generations until sufficient ESS values were achieved. Generations compute several-fold faster in BEAST, making a direct comparison of run times difficult. Thirty-eight percent of the MrBayes analyses needed to be rerun. For each run, 9000 post-burnin trees were sampled. The MrBayes block template is available as online Appendix 4 (http://www.sysbio.oxfordjournals.org).

We also compared the overall quality of maximum likelihood (ML) time-free inference methods to the aforementioned Bayesian inference methods. The 800 ML trees were inferred in PAUP* v4.1 (Swofford, 2002) under an HKY + Γ_4 substitution model with a heuristic search utilizing the subtree pruning regrafting branch swapping

algorithm. We also performed non-parametric bootstrapping (100 replicates) on all 800 sequence alignments.

Metrics of Phylogenetic Inference Quality

To compare the BEAST and MrBayes analyses, we unrooted all post-burnin trees using PAUP*. Measurements of accuracy and precision of the phylogenetic analyses were performed using TreeLogAnalyser (part of the BEAST package). First, the 95% credible set of trees from each analysis was identified. If the true tree, the topology generated in APE, was found in that credible set, then the analysis was categorized as accurate. The number of trees in the 95% credible set was used to quantify precision. We also used a third metric, the Robinson-Foulds tree-to-tree distance (Robinson and Foulds, 1981), which calculates the number of nodes separating two trees. We determined the distance between the true tree topologies and each of the post-burnin topologies sampled (for the Bayesian analyses) and the ML tree or the bootstrap replicates (for the ML analyses). These values were scaled by the theoretical maximum Robinson-Foulds tree-to-tree distance to standardize across topologies with varying taxon number. The mean of these values was used as an indicator of the overall distance of the sampled trees from the true tree.

We used these three metrics (accuracy, precision, and Robinson-Foulds tree-to-tree distance) to compare the performance of each of the molecular clock models of phylogenetic inference on sequences generated under all four rate distributions. Accuracy, a binary outcome, was assessed using logistic regression. Precision, given its

non-normalizable distribution, was partitioned into quintiles and analyzed using ordinal logistic regression. Robinson-Foulds tree-to-tree distance data were analyzed using multiple linear regression. We chose to analyze the data using regression analyses so that we could adjust for taxon number, Seq-Gen scaling factor, and Colless's Imbalance as fixed effects. Colless's Imbalance (Colless, 1995) is a measurement of topological asymmetry and was calculated using Mesquite (Maddison and Maddison, 2007). We also treated the 200 tree topologies as a random effect in the regression analyses. All statistical analyses were performed in Stata v9.2 (StataCorp, 2005). For each statistical analysis, significance was assessed with $\alpha = 0.05$. Since we performed a simulation study, and our power to detect significant differences was dependent on the length of the simulation, we also employed an additional relevance cut-off. We discounted differences in mean Robinson-Foulds tree-to-tree distances whose β-coefficients were <1%. Any difference smaller than this would not actually result in a different final tree topology and would therefore not be biologically meaningful. This second cut-off was employed only for the strict-clock inference model, in which the variance was so low that small differences, β-coefficient < 1%, were significantly different.

RESULTS

To determine if incorporating a relaxed molecular clock improved the quality of phylogenetic inference, we analyzed sequences simulated under a variety of rate distributions and constructed phylogenies assuming relaxed molecular clocks, a strict molecular clock, and time-free inference.

Accuracy of Inference Methods

The first metric we used to assess the quality of phylogenetic inference was accuracy (i.e., whether or not the true tree was recovered in the 95% credible set). Analyses using relaxed molecular clock inference models consistently were the most accurate (Table 1), though the differences in accuracy were significant only if the sequences had been simulated under an exponential or lognormal relaxed molecular clock (i.e., darker colored circles on the targets; Fig. 3).

Analysis using a strict-clock inference model resulted in significantly poorer accuracy if the sequences were evolved under an exponential, lognormal, or uniform relaxed molecular clock distributions of rates (i.e., the circles for strict inference models are lighter in Fig. 3a, b, d); however, when sequences were evolved under a strict clock, there were no significant differences in accuracy among the four inference models (Fig. 3c). There was not a pattern of increased accuracy of inference models when analyzing sequence data that fit the assumptions of that inference model. In general, relaxed-clock inference models were the most accurate, followed by the time-free model, whereas the strict clock inference model was consistently the least accurate.

Precision of Inference Methods

The precision estimates of the inference models (i.e., the number of distinct topologies sampled in the 95% credible set) appear to show the opposite trend of accuracy (Table 1; Fig 3). Relaxed-clock inference models were the least precise in every case, with the exponential relaxed clock faring the worst under every rate distribution except exponential. The strict-clock inference model was almost always the most precise (Table 1). When analyzing sequences generated under exponential, lognormal, and uniform rate distributions, the strict-clock inference model sampled significantly fewer trees than the other three inference models (i.e., the strict inference model has the smallest circles on the targets in Fig. 3a, b, d). There were no significant differences in precision among the four inference models, when sequences were evolved under a strict clock (Fig. 3c). The time-free inference model generally resulted in intermediate precision, sampling significantly fewer trees than the relaxed-clock inference models when the rates were generated under non-strict distributions. Similar to accuracy, there was not a pattern of greater precision of inference models when analyzing sequence data that fit the assumptions of that inference model.

Robinson-Foulds Tree-to-Tree Distance

Relaxed-clock models are the most accurate, but the least precise, of the inference models tested here. But these results still do not answer the question, which inference model provides the highest quality of phylogenetic inference? We found that a third metric, the Robinson-Foulds tree-to-tree distance (i.e., the number of nodes that separate

the sampled trees from the true tree) best encapsulates the relative quality of phylogenetic inference (Table 1). For exponential, lognormal, and uniform rate distributions, strict-clock inference found topologies that were significantly more distant from the true tree than those of the other inference models (i.e., strict-clock circles are the farthest from the center of the target in Fig. 3a, b, d). When sequences were simulated under a strict molecular clock, all four inference models sampled trees with indistinguishable Robinson-Foulds distances (i.e., all four circles are equidistant from the center of the target in Fig. 3c). Among exponential, lognormal, and time-free inference models, there were no significant differences in the observed Robinson-Foulds tree-to-tree distance measurements (i.e., circles are equidistant from the center of the target; Fig. 3). Relaxed molecular clocks fared no better or worse than the time-free inference model. Although informative, neither accuracy nor precision completely summarized the quality of phylogenetic inference. Robinson-Foulds tree-to-tree distance, however, was the most revealing metric of phylogenetic inference quality because it was informed by both accuracy and precision.

In addition, we tested for interaction between the inference models and three fixed effects (i.e., number of taxa, maximum pairwise distance, and Colless's Imbalance) using Robinson-Foulds distance as an outcome. As taxon number increased, strict clock inference performed increasingly worse than lognormal and exponential relaxed-clock inference methods ($p < 0.05$) and marginally worse than time-free inference ($p = 0.08$). Strict clock inference also performed worse than the other three inference methods as the maximum pairwise distance among the taxa increased ($p < 0.001$). There were no

significant interactions between Colless's Imbalance and the inference model. In general, the more complex the sequence data, the worse strict clock inference performed.

Maximum Likelihood Inference Quality

Our dataset provided us the opportunity to explore how the quality of ML inference compares to Bayesian approaches. We measured the Robinson-Foulds tree-to-tree distance from the true tree to the tree inferred under time-free ML phylogenetic inference. This mean distance (for sequences simulated under each of the four rate distributions) was always smaller than the mean Robinson-Foulds distance from the true tree to the 9000 post-burnin Bayesian topologies ($p < 0.001$) (Table 1). We note, however, that non-parametric bootstrapping is commonly used to assess confidence in the ML topology. Therefore, we also calculated the mean Robison-Foulds distance between the true tree and the bootstrap replicates (Table 1). For sequences simulated under an exponential rate distribution, the ML bootstrap trees were significantly closer to the true tree than the posterior trees from all four Bayesian inference methods ($p < 0.05$). For sequences simulated under a lognormal rate distribution, there were no significant differences among the ML bootstrap trees and the Bayesian trees according to our β-coefficient criterion (see methods). Surprisingly, for sequences evolved under a strict rate distribution, ML bootstrap trees were significantly farther from the true tree than trees inferred under all four Bayesian inference methods ($p < 0.001$). Finally, for sequences evolved under a uniform rate distribution, ML bootstrap trees were better than strict clock inference ($p < 0.001$) but similar to the other Bayesian inference methods.

DISCUSSION

When comparing relaxed molecular clock and time-free methods of Bayesian phylogenetic inference, a trade-off exists between accuracy and precision in our simulation study. Both of these methods sample trees with indistinguishable Robinson-Foulds tree-to-tree distances from the true tree, but their levels of accuracy and precision are model-dependent (Fig. 3). The Robinson-Foulds tree-to-tree distance measurements do not change among these three clock models; as accuracy increases, precision must decrease, and vice-versa. Therefore the quality of the trees sampled when a relaxed molecular clock is assumed is no different than when no assumption is made about a molecular clock. However, if a strict molecular clock is assumed for non-strict clock sequence data, this trade-off is not discernible. Inference under a strict clock on non-strict clock sequence data has extremely high precision, but its accuracy is so poor that the Robinson-Foulds tree-to-tree distance measurements are significantly worse than if the clock was relaxed or was not assumed at all.

These results support the existence of a bias-variance trade-off in topological inference when incorporating a relaxed molecular clock. Relaxing the clock, by adding rate parameters, increases the probability of finding the true tree (accuracy/bias), but it comes at the expense of sampling many more trees (precision/variance). Not making an assumption about a molecular clock (i.e., the time-free inference model) decreases variance (better precision) but biases the analyses (less accurate). Time-free inference appears to under-fit the data, but relaxed molecular clock inference may tend to over-fit (i.e, over-parameterize) the data. In contrast, when a strict molecular clock is violated, the

analysis is so highly biased that the true answer is rarely recovered when using a strict-clock inference model. Collectively, these patterns indicate that assuming a relaxed molecular clock does not improve the quality of phylogenetic inference over a time-free inference model because of a trade-off between bias and variance. We note that over-parameterization does not necessarily mean increasing the total number of parameters in the inference model. Relaxed-clock inference models technically have fewer parameters than time-free models; however, relaxed-clock inference models parameterize rates. Our analysis suggests that including information about rates does not improve topological inference and is therefore an over-parameterization. Nonetheless, unreasonable assumptions, such as a strict molecular clock when multiple evolutionary rates exist, can severely decrease the quality of phylogenetic inference and should be avoided unless there is strong evidence that the sequences in question evolved under a single evolutionary rate.

Our findings contradict those reported by Drummond et al. (2006). Whereas they found an increase in both accuracy and precision of relaxed molecular clock phylogenetic inference compared to the time-free model, we found a trade-off between these metrics suggesting no difference in inference quality. This discrepancy might be due to the decision by Drummond et al. (2006) to remove the least precise 10% of runs from their comparisons. This might have led to artifactually improved precision estimates by relaxed clock methods, which we found to be the least precise.

This study casts doubt on the claim that relaxed molecular clock inference results in improved topological reconstruction. However, one important difference between the

Drummond et al. (2006) study and ours is that they used real sequence data, whereas we looked at simulated sequenced data. There are two possible explanations for our differing results. First, they may have failed to detect the bias-variance trade-off in their analysis. An alternative explanation may be that there are important differences between real and simulated sequence data, and relaxed-clock inference models may actually be superior when analyzing real sequence data [e.g., (Liu et al., 2008)]. Future work will be required to distinguish between these two possibilities.

There does appear to be a relationship between the underlying distribution of rates and the ability of an inference model to reconstruct high quality trees as measured by Robinson-Foulds tree-to-tree distance. Specifically, all inference models (Bayesian and ML) performed best on sequences simulated under a strict rate distribution, followed by lognormal and uniform; inference methods always performed the worst on sequences simulated under an exponential rate distribution (Table 1).

The single ML time-free topology was strikingly closer to the true tree than the posterior distribution of Bayesian trees; however, comparisons between the bootstrapped ML trees and the Bayesian posterior distribution of trees appeared to be qualitatively similar. This finding is in concordance with previous studies that have compared ML and Bayesian phylogenetic inference methods on empirical and simulated data (Alfaro et al., 2003; Cummings et al., 2003; Douady et al., 2003; Erixon et al., 2003; Mar et al., 2005). Nevertheless, there certainly appear to be instances where ML analysis is preferable to Bayesian inference (and vice-versa). Our findings suggest that a systematic exploration of the conditions (beyond rate distribution) that favor ML or Bayesian topological inference

should be undertaken. Our findings also support the notion that the Robinson-Foulds tree-to-tree distance is a highly useful metric for gauging the overall quality of phylogenetic inference.

ACKNOWLEDGMENTS

We thank Darren Boss for assistance with the high-throughput analyses, Simon Ho for guidance in sequence simulation, Betsy C. Wertheim for advice on statistical methods, and Andrew Rambaut for helpful discussion. We also thank the associate editor and two reviewers for their helpful comments on this manuscript. Funding was provided by the Department of Ecology and Evolutionary Biology and BIO5 at the University of Arizona, the David and Lucile Packard Foundation, and a National Institutes of Health Institutional Research and Academic Career Development Award.

REFERENCES

Alfaro, M. E., S. Zoller, and F. Lutzoni. 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol. Biol. Evol. 20:255–266.

Ayala, F. J. 1997. Vagaries of the molecular clock. Proc. Natl. Acad. Sci. USA. 94:7776–7783.

Bromham, L., and D. Penny. 2003. The modern molecular clock. Nat. Rev. Genet. 4:216–224.

Burnham, K. P., D. R. Anderson, and K. P. Burnham. 2002. Model selection and multimodel inference: a practical information-theoretic approach, 2nd edition. Springer, New York.

Colless, D. H. 1995. Relative symmetry of cladograms and phenograms: an experimental study. Syst. Biol. 44:102–108.

Cummings, M. P., S. A. Handley, D. S. Myers, D. L. Reed, A. Rokas, and K. Winka. 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52:477–487.

Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, and E. J. Douzery. 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. Mol. Biol. Evol. 20:248–254.

Drummond, A. J., S. Y. Ho, M. J. Phillips, and A. Rambaut. 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4:e88.

Drummond, A. J., and A. Rambaut. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104:5936–5941.

Erixon, P., B. Svennblad, T. Britton, and B. Oxelman. 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52:665–673.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17:368–376.

Gaut, B. S., and P. O. Lewis. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. 12:152–162.

Huelsenbeck, J. P., and D. M. Hillis. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

Kumar, S. 2005. Molecular clocks: four decades of evolution. Nat. Rev. Genet. 6:654–662.

Li, W. H. 1993. So, what about the molecular clock hypothesis? Curr. Opin. Genet. Dev. 3:896–901.

Liu, W., M. Worobey, Y. Li, B. F. Keele, F. Bibollet–Ruche, Y. Guo, P. A. Goepfert, M. L. Santiago, J. B. Ndjango, C. Neel, S. L. Clifford, C. Sanz, S. Kamenya, M. L. Wilson, A. E. Pusey, N. Gross-Camp, C. Boesch, V. Smith, K. Zamma, M. A. Huffman, J. C. Mitani, D. P. Watts, M. Peeters, G. M. Shaw, W. M. Switzer, P. M. Sharp, and B. H. Hahn. 2008. Molecular ecology and natural history of simian foamy virus infection in wild-living chimpanzees. PLoS Pathog. 4:e1000097.

Maddison, W. P., and D. R. Maddison. 2007. Mesquite: a modular system for evolutionary analysis. Version 2.0 Available from: http://mesquiteproject.org.

Mar, J. C., T. J. Harlow, and M. A. Ragan. 2005. Bayesian and maximum likelihood phylogenetic analyses of protein sequence data under relative branch-length differences and model violation. BMC Evol. Biol. 5:8.

Ochman, H., J. G. Lawrence, and E. A. Groisman. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature. 405:299–304.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 20:289–290.

Posada, D., and K. A. Crandall. 1998. MODELTEST: testing the model of DNA substitution. Bioinformatics. 14:817–818.

Rambaut, A., and M. A. Charleston. 2002. TreeEdit: phylogenetic tree editor v1.0 alpha 10.

Rambaut, A., and A. J. Drummond. 2007. Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer.

Rambaut, A., and N. C. Grassly. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.

Robinson, D. F., and L. R. Foulds. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Sanderson, M. J. 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy Mol. Biol. Evol. 14:1218–1231.

Sanderson, M. J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.

Shapiro, B., A. Rambaut, and A. J. Drummond. 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23:7-9.

StataCorp. 2005. Stata Statistical Software: Release 9. College Station, TX: StataCorp LP.

Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mamm. Evol. 4:77–86.

Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), version 4. Sunderland (Massachusetts): Sinauer Associates.

Thorne, J. L., H. Kishino, and I. S. Painter. 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15:1647–1657.

Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. 10:1396–1401.

Zuckerkandl, E., and L. B. Pauling. 1962. Molecular disease, evolution, and genetic heterogeneity. Pages 189–225 in Horizons in biochemistry. (M. A. Kasha, and B. Pullman, eds.). Academic Press, New York.

TABLE 1. Performance of inference models on sequence simulated under various rate distributions

Metric	Inference model	Rate distribution			
		Exponential	Lognormal	Strict	Uniform
Accuracy[a] (%)	Exponential	57.0	81.5	85.0	68.5
	Lognormal	55.0	76.5	84.5	66.5
	Strict	19.0	56.5	84.0	41.0
	Time-free	49.0	74.0	83.5	64.5
Precision[b]	Exponential	3944	2148	1396	2858
	Lognormal	4035	1881	1097	2718
	Strict	2866	1721	1076	2065
	Time-free	3782	1738	1032	2469
RF distance[c] (%)	Exponential	21.5	12.8	10.0	15.1
	Lognormal	21.6	12.3	9.2	15.1
	Strict	28.8	14.1	9.1	17.4
	Time-free	22.3	12.6	9.6	15.4
	ML time-free[d]	15.5	9.0	7.0	11.3
	ML time-free bootstrap	20.3	13.2	10.4	15.4

Notes: [a]Percentage of the runs in which the true tree was recovered in the 95% credible set.
[b]Mean number of trees in the 95% credible set.
[c]Mean Robinson–Foulds (RF) tree-to-tree distance between the true tree and sampled trees expressed as a percentage of the maximum possible distance.
[d]ML time-free RF tree-to-tree distance is always significantly closer ($P < 0.001$) to the true tree than the Bayesian inference methods (see text for details).

Figure 1. Flow-chart of Bayesian inference simulation study. Software packages used at each step are noted in parentheses.

Figure 2. Distributions sampled to model evolutionary rates among branches according to (a) exponential, (b) lognormal, (c) strict, and (d) uniform distributions.

Figure 3. Summary of Bayesian phylogenetic inference quality. Accuracy, precision, and Robinson-Foulds tree-to-tree distance of exponential, lognormal, strict, and time-free inference models on sequence data evolved under (a) exponential, (b) lognormal, (c) strict, and (d) uniform rate distributions. For a given rate distribution (i.e., target), the darker the circle, the more accurate the inference model on sequences evolved under that rate distribution. Smaller circles indicate better precision. The distance from the center of each circle to the middle of its target represents the Robinson-Foulds distance of the sampled trees from the true tree. Within each target, differences in darkness, size, and distance from the center represent significance at $\alpha = 0.05$.

Figure 1.

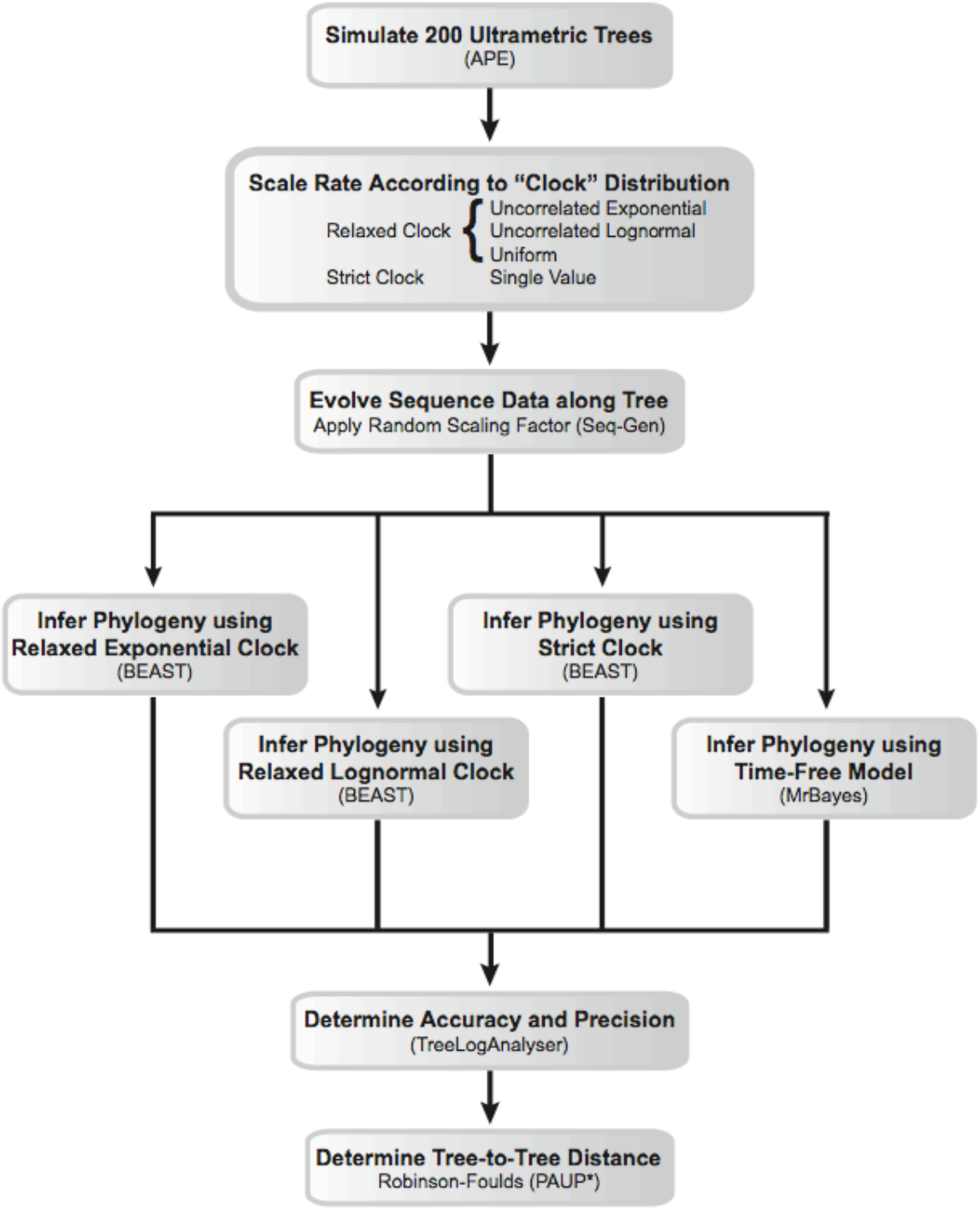

Simulate 200 Ultrametric Trees
(APE)
Scale Rate According to "Clock" Distribution
Relaxed Clock
Uncorrelated Exponential
Uncorrelated Lognormal
Uniform
Strict Clock
Single Value
Evolve Sequence Data along Tree
Apply Random Scaling Factor (Seq-Gen)
Infer Phylogeny using Relaxed Exponential Clock
(BEAST)
Infer Phylogeny using Strict Clock
(BEAST)
Infer Phylogeny using Relaxed Lognormal Clock
(BEAST)
Infer Phylogeny using Time-Free Model
(MrBayes)
Determine Accuracy and Precision
(TreeLogAnalyser)
Determine Tree-to-Tree Distance
Robinson-Foulds (PAUP*)

Figure 2.

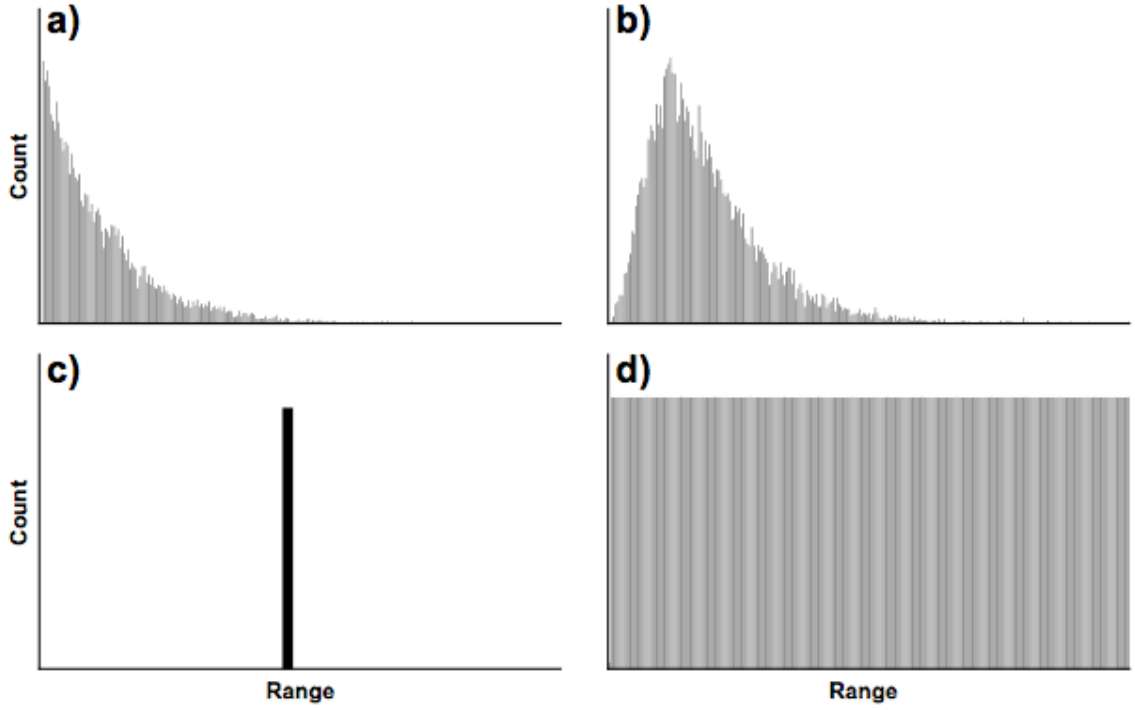
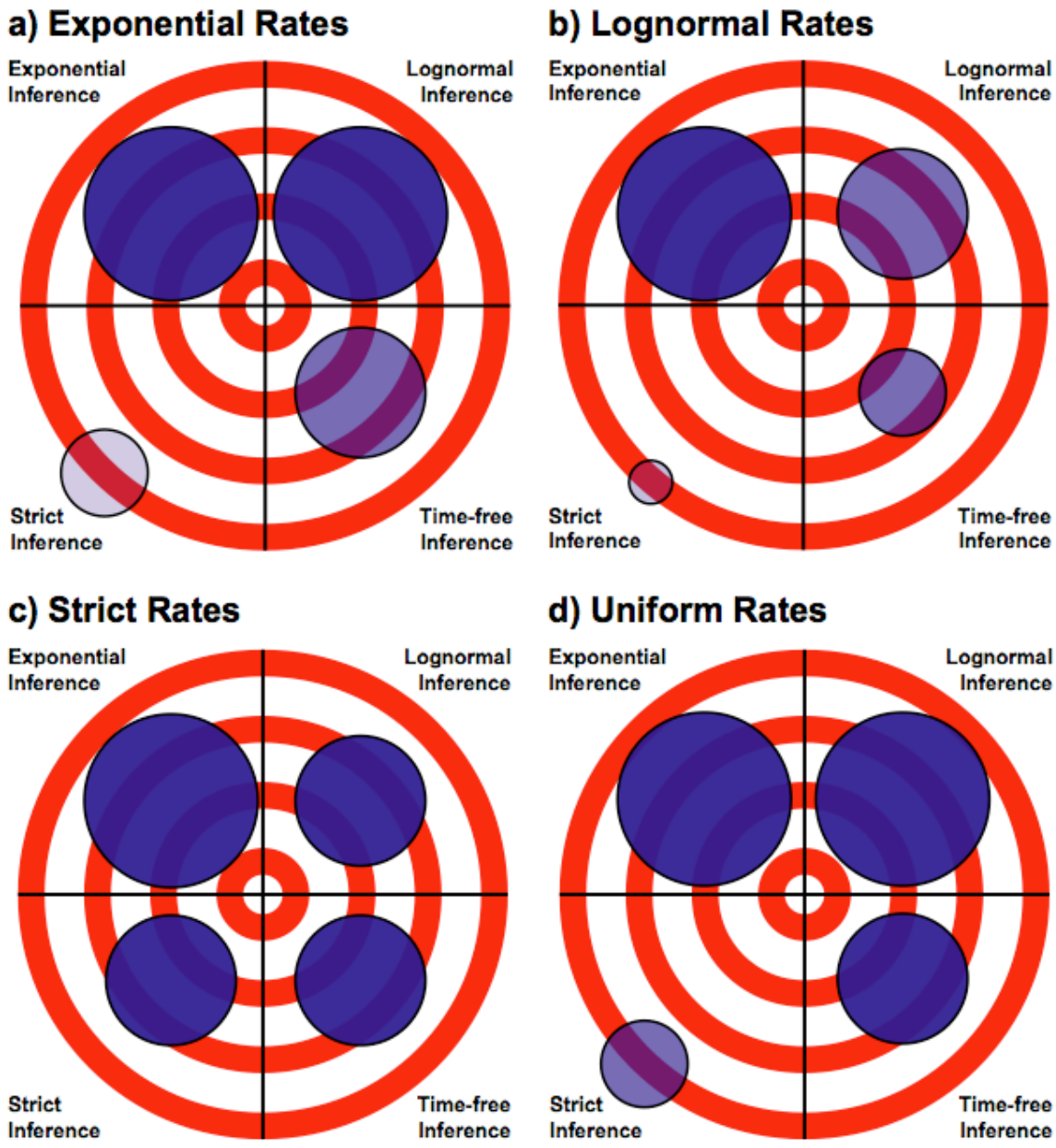


Figure 3.


a) Exponential Rates
Exponential Inference
Lognormal Inference
Strict Inference
Time-free Inference
b) Lognormal Rates
Exponential Inference
Lognormal Inference
Strict Inference
Time-free Inference
c) Strict Rates
Exponential Inference
Lognormal Inference
Strict Inference
Time-free Inference
d) Uniform Rates
Exponential Inference
Lognormal Inference
Strict Inference
Time-free Inference