JVI
Journals.ASM.org

A Case for the Ancient Origin of Coronaviruses

Joel O. Wertheim,[a,b] Daniel K. W. Chu,[c,d] Joseph S. M. Peiris,[c,d] Sergei L. Kosakovsky Pond,[b] Leo L. M. Poon[c,d]

Department of Pathology, University of California, San Diego, California, USA[a]; Department of Medicine, University of California, San Diego, California, USA[b]; Centre for Influenza Research and School of Public Health, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, People's Republic of China[c]; State Key Laboratory of Emerging Infectious Diseases, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, People's Republic of China[d]

Coronaviruses are found in a diverse array of bat and bird species, which are believed to act as natural hosts. Molecular clock dating analyses of coronaviruses suggest that the most recent common ancestor of these viruses existed around 10,000 years ago. This relatively young age is in sharp contrast to the ancient evolutionary history of their putative natural hosts, which began diversifying tens of millions of years ago. Here, we attempted to resolve this discrepancy by applying more realistic evolutionary models that have previously revealed the ancient evolutionary history of other RNA viruses. By explicitly modeling variation in the strength of natural selection over time and thereby improving the modeling of substitution saturation, we found that the time to the most recent ancestor common for all coronaviruses is likely far greater (millions of years) than the previously inferred range.

Coronaviruses (family *Coronaviridae*, subfamily *Coronavirinae*) are important pathogens of birds and mammals. Coronaviruses are positive-sense RNA viruses and are currently classified into four genera: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus*, and *Deltacoronavirus* (1). Alphacoronaviruses and betacoronaviruses are found exclusively in mammals, whereas gammacoronaviruses and deltacoronaviruses primarily infect birds. The identification of severe acute respiratory syndrome (SARS) coronavirus in 2003 (2) prompted an intensive search for novel coronaviruses, resulting in the detection of a number of novel coronaviruses in humans, domesticated animals, and wildlife (3–7). Interestingly, surveillance of coronaviruses in wild animals has led to the discovery of the greatest diversity of coronaviruses in bat and avian species, which suggests that these animals are the natural reservoirs of the viruses (8–10). Indeed, phylogenetic studies of bat and avian coronaviruses suggest an ancient relationship with possible codivergence and coevolution with their hosts. Conversely, many coronaviruses found in bats and other mammals diverged near the tips of coronavirus phylogeny, suggesting that these viruses were the result of recent cross-species transmission events (9, 11).

Molecular clock analysis based on the RNA-dependent RNA polymerase (RdRp) genomic region suggests a time of most recent common ancestor (tMRCA) for the four coronavirus genera of around 10,100 years ago, with a mean rate of 1.3×10^{-4} nucleotide substitutions per site per year (10). This tMRCA estimate is difficult to reconcile with a hypothetical ancient, coevolutionary relationship between coronaviruses and their bat or bird hosts (12, 13). Moreover, a group of genetically related, yet distinct, alphacoronaviruses have been detected in different mouse-eared bats (*Myotis* spp.) on multiple continents. However, these bat species do not migrate long distances, with few traveling farther than hundreds of miles to overwinter sites (Gary F. McCracken, personal communication). Yet, the tMRCA of *Alphacoronavirus* is estimated to be around 200 or 4,400 years ago, on the basis of analyses of helicase (9) and RdRp (10), respectively. The limited interaction among these bat populations suggests a more ancient evolutionary association with alphacoronaviruses (i.e., codivergence or coevolution), which is incompatible with the relatively young viral tMRCAs. Notably, coronaviruses have a unique proofreading mechanism for viral RNA replication (3, 14); because of the exoribonuclease activity of viral nonstructural protein 14 (Nsp14), the mutation rate of coronaviruses has been found to be similar to that of single-stranded DNA viruses ($\sim 1 \times 10^{-5}$ to 1×10^{-6} mutation per site per replication) and well below those measured in other RNA viruses ($\sim 1 \times 10^{-3}$ to 1×10^{-5} mutation per site per replication) (15, 16). For these reasons, we speculate that there is a substantial underestimation of the length of the natural evolutionary history of coronaviruses.

Recent developments in the modeling of the evolution of RNA viruses have revealed that purifying selection can mask the ancient age of viruses that appear to have recent origins, according to a molecular clock (e.g., measles, Ebola, and avian influenza viruses) (17). Strong purifying selection can maintain evidence of sequence homology long after saturation has occurred at synonymous sites; this phenomenon can lead to underestimation of the overall depth of a viral phylogeny. In the absence of strong purifying selection, nucleotide sequences would diverge more quickly, lose detectable homology, and become difficult to align and compare. Here, we asked if similar evolutionary patterns led to underestimation of the tMRCA of the coronavirus lineage. We employed evolutionary models that account for variation in the pressure of natural selection across sites in viral loci and lineages in their phylogenies. Our results indicate that coronaviruses are orders of magnitude older than suggested by previous molecular clock analyses.

MATERIALS AND METHODS

Sequence data sets. Representative coronavirus genomes ($n = 43$; Table 1) from the four genera were selected for this study. For several viral taxa, genomic sequences from multiple sampling years were included, though these dates were not used explicitly in our analyses. To avoid highly re-

Received 27 November 2012 Accepted 8 April 2013
Published ahead of print 17 April 2013
Address correspondence to Joel O. Wertheim, jwertheim@ucsd.edu, or Leo L. M. Poon, llmpoon@hku.hk.

doi:10.1128/JVI.03273-12

TABLE 1 Coronavirus sequences analyzed in this study

Genus and virus name	Host species	Strain	Accession no.	Sampling yr
Alphacoronavirus				
Bat coronavirus 1A[a]	*Miniopterus magnater*	AFCD62	EU420138	2005
Bat coronavirus 1B[a]	*Miniopterus pusillus*	AFCD307	EU420137	2006
Bat coronavirus HKU2	*Rhinolophus sinicus*	HKU2/GD/430/2006	EF203064	2006
Bat coronavirus HKU8	*Miniopterus pusillus*	AFCD77	EU420139	2005
Feline coronavirus[b]	*Felis catus*	UU2	FJ938060	1993
Feline coronavirus[b]	*Felis catus*	UU54	JN183883	2010
Human coronavirus NL63	*Homo sapiens*	NL63/DEN/2009/14	JQ765564	2009
Human coronavirus NL63	*Homo sapiens*	NL63/DEN/2005/1876	JQ765575	2005
Porcine epidemic diarrhea virus	*Sus scrofa*	CH/FJND-3	JQ282909	2011
Porcine epidemic diarrhea virus	*Sus scrofa*	CH/S	JN547228	1986
Transmissible gastroenteritis virus[b]	*Sus scrofa*	H16	FJ755618	1973
Transmissible gastroenteritis virus[b]	*Sus scrofa*	Purdue	DQ811789	1952
Betacoronavirus				
Bat coronavirus HKU5	*Pipistrellus abramus*	TT07f	EF065512	2006
Bat coronavirus HKU9	*Rousettus leschenaulti*	10-1	HM211100	2006
Bat coronavirus/133/2005	*Tylonycteris pachypus*	BtCoV/133/2005	DQ648794	2005
Bat SARS coronavirus[c]	*Rhinolophus pearsoni*	Rp3	DQ071615	2004
Bovine coronavirus[d]	*Bos taurus*	DB2	DQ811784	1983
Bovine coronavirus[d]	*Bos taurus*	E-AH187-TC	FJ938064	2000
Civet SARS coronavirus[c]	*Paguma larvata*	SZ3	AY304486	2003
Human coronavirus HKU1	*Homo sapiens*	Caen1	HM034837	2005
Human coronavirus OC43[d]	*Homo sapiens*	HK04-02	JN129835	2004
Human SARS coronavirus[c]	*Homo sapiens*	HKU-39849 isolate UOB	JQ316196	2003
Mouse hepatitis virus	*Mus musculus*	MHV-MI	AB551247	1994
Mouse hepatitis virus RA59/R13	*Mus musculus*	RA59/R13	FJ647218	2006
Sammbar deer coronavirus	*Cervus unicolor*	US/OH-WD388-TC	FJ425188	1994
Gammacoronavirus				
Duck coronavirus	Duck[f]	DK/CH/HN/ZZ2004	JF705860	2004
Infectious bronchitis virus[e]	*Gallus gallus*	Holte	GU393336	1954
Infectious bronchitis virus[e]	*Gallus gallus*	ck/CH/LHLJ/100902	JF828980	2010
Infectious bronchitis virus[e]	*Gallus gallus*	Conn46	FJ904719	1991
Infectious bronchitis virus[e]	*Gallus gallus*	Mass41	FJ904721	1972
Infectious bronchitis virus[e]	*Gallus gallus*	Massachusetts	GQ504724	1941
Turkey coronavirus[e]	*Meleagris gallopavo*	IN-517	GQ427175	1994
Turkey coronavirus[e]	*Meleagris gallopavo*	TX-1038	GQ427176	1998
Turkey coronavirus[e]	*Meleagris gallopavo*	VA-74	GQ427173	2003
Deltacoronavirus				
Bulbul coronavirus HKU11	*Pycnonotus jocosus*	HKU11-934	FJ376619	2007
Common-moorhen coronavirus HKU21	*Gallinula chloropus*	HKU21-8295	JQ065049	2007
Magpie robin coronavirus HKU18	*Copsychus saularis*	HKU18-chu3	JQ065046	2007
Munia coronavirus HKU13	*Lonshura striata*	HKU13-3514	FJ376622	2007
Night-heron coronavirus HKU19	*Nycticorax nycticorax*	HKU19-6918	JQ065047	2007
Sparrow coronavirus HKU17	*Passer montanus*	HKU17-6124	JQ065045	2007
Thrush coronavirus HKU12	*Turdus hortulorum*	HKU12-600	FJ376621	2007
White-eye coronavirus HKU16	*Zosterops* sp.	HKU16-6847	JQ065044	2007
Wigeon coronavirus HKU20	*Anas penelope*	HKU20-9243	JQ065048	2008

[a] Viruses are classified as a single species (*Miniopterus bat coronavirus 1*).
[b] Viruses are classified as a single species (*Alphacoronavirus 1*).
[c] Viruses are classified as a single species (*Severe acute respiratory syndrome-related coronavirus*).
[d] Viruses are classified as a single species (*Betacoronavirus 1*).
[e] Viruses are classified as a single species (*Avian coronavirus*).
[f] Host species cannot be determined.

combinant sequences and/or sequence misalignments, we specifically selected partial viral sequences from five relatively conserved genomic regions (Fig. 1 and Table 1): Nsp15-16 (1,320 nucleotides [nt]), the matrix protein (640 nt), papain-like protease 2 (PLP2; 620 nt), the RdRp (1,860 nt), and the Y domain (400 nt) (18). These nucleotide sequences were aligned at the amino acid level by MUSCLE (19) as described previously (20).

Phylogenetic analyses. Each of the five target sequences was screened for recombination with a genetic algorithm tool, GARD (21). Maximum-likelihood phylogenies were constructed for each nonrecombinant region

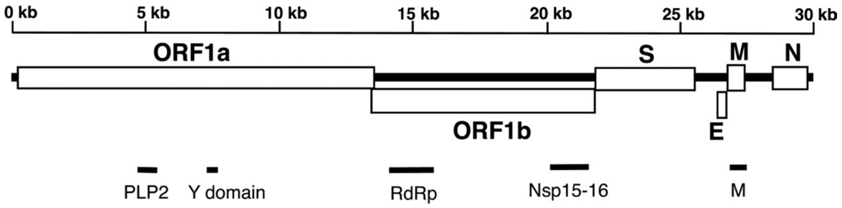

FIG 1 Schematic diagram of the SARS coronavirus genome. Open reading frame 1a (ORF1a) and ORF1b, encoding the nonstructural polyproteins, and those encoding the S, E, M, and N structural proteins are indicated. The approximate locations of the viral sequences used in this study are shown.

with PHYML 3.0 (22), implemented in Seaview 4.0 (23), with a subtree-pruning and regrafting search algorithm and the general time-reversible substitution model with a four-bin gamma rate distribution (GTR + Γ_4).

With these topologies, branch lengths were re-estimated in HyPhy (24) under GTR + Γ_4 and a branch site random effects likelihood (BS-REL) model (25). The BS-REL model was implemented to account for the effects of variable selection pressure across codon sites and phylogenetic lineages by inferring three unconstrained ω classes (dN/dS: nonsynonymous substitution rate/synonymous substitution rate) for each branch and estimating the proportion of sites in the alignment that evolved under each ω class.

To estimate the variance in branch length estimates produced by the BS-REL model, we used a modified Latin hypercube sampling (LHS) importance resampling scheme ($M = 500$ samples), described in detail previously (26). LHS is a standard technique that can be efficiently parallelized, an important consideration here. It is used to assess the variability of high-dimensional distributions by discretizing the volume of parameter space around the maximum-likelihood estimate into N bins (10,000 in our case) of approximately equal probability along each coordinate and sampling the bin for each parameter in a way that no two parameters share the same bin index (this technique maximizes space coverage). Samples are reweighted according to their likelihood, and resamples ($M = 500$) are drawn from this distribution. A modified version of LHS (26), which we use here, has also been shown to compare favorably to other techniques for the assessment of parameter uncertainty.

RESULTS

Recombination detection. Phylogenies with interpretable branch lengths can be inferred only when analyzing nonrecombinant regions. Therefore, we screened each of the five highly conserved coronavirus target regions with GARD (21). Within the Nsp15-16 region, two recombination breakpoints were detected, and phylogenetic incongruity was confirmed by a Kishino-Hasegawa test ($P < 0.01$) (27, 28). The first nonrecombinant section encompassed Nsp15 and 80 nt of Nsp16, which we refer to as Nsp15$^+$; the second section fell entirely within Nsp16. Because of its short length (300 nt), the third section was not included in later analyses. Within the RdRp region, a single recombination breakpoint was detected with GARD, but it is unlikely that this breakpoint represents a true recombination event since the two topologies were not significantly different according to the Kishino-Hasegawa test. This putative breakpoint detected in the RdRp region was likely due to elevated rate variation across the sequence (21). No recombination breakpoints were detected in the other three loci.

The maximum-likelihood phylogenies inferred from the six nonrecombinant loci were almost all significantly different from each other (Shimodaira-Hasegawa test, $P < 0.05$) (29). This finding suggests that the different regions of the coronavirus genome have distinct evolutionary histories and should not be treated as a single region in a phylogenetic analysis. The lone exception was the comparison of the topologies from RdRp and Nsp15$^+$ ($P = 0.131$). Nevertheless, we chose be conservative and analyze each of the six loci independently.

Branch length expansion. We employed two models of nucleotide sequence evolution (GTR + Γ_4 and BS-REL) to re-estimate the branch lengths of the inferred maximum-likelihood phylogenies. GTR + Γ_4 is a standard nucleotide substitution model that has been commonly in many evolutionary virology studies. In contrast, the BS-REL model is a codon model that explicitly accounts for variation in selection pressures across sites and lineages (25). Both evolutionary models produced comparable length estimates for shorter branches (≤0.05 substitution per site in Fig. 2). However, compared to results inferred from BS-REL, long branches in the coronavirus phylogenies were underestimated by GTR + Γ_4 (Fig. 2). Notably, there were a substantial number of long branches in which the number of substitutions per site approached saturation in the BS-REL model, ranging from 8 branches in Nsp16 to 16 branches in the Y domain (Table 2). The lengths of these saturated branches cannot be reliably estimated, although a meaningful lower bound on their lengths can be obtained (e.g., with LHS; see below).

Among the six nonrecombinant regions, there was substantial variation in the total tree length expansion between GTR + Γ_4 and BS-REL, ranging over 3 orders of magnitude (Table 2). Differing selective pressures along lineages among loci appear to account for this variation, as we observed a correlation between the strength of purifying selection along a branch (i.e., lower mean ω) and the relative increase in branch length under BS-REL compared to GTR + Γ_4 (Spearman's nonparametric correlation, $P \leq 0.0001$; branches in which either model inferred a length of zero substitutions per site were excluded from this analysis). Thus, lineages that bear the mark of stronger purifying selection generally experienced a greater increase in length under the BS-REL model. Moreover, the number of branches approaching saturation was correlated with a greater expansion in the total length under BS-REL (Spearman's nonparametric correlation, $P < 0.05$). Interestingly, simpler, site-based measurements of selection did not correlate with total tree length expansion (e.g., mean dN/dS [$P = 0.87$] and number of sites of pervasive purifying selection [$P = 0.46$]), though analyzing only six loci limits the statistical power of these tests. Nonetheless, the relationship between ω and branch length expansion supports the hypothesis that increased purifying selection is responsible for the underestimation of the lengths of long branches in RNA virus phylogenies.

The variance in branch length expansion from GTR + Γ_4 to BS-REL, as approximated with LHS, differed among coronavirus loci (Table 2). For the Nsp15$^+$, Nsp16, matrix protein, and PLP2 loci, the upper and lower 95% confidence limits differed by orders

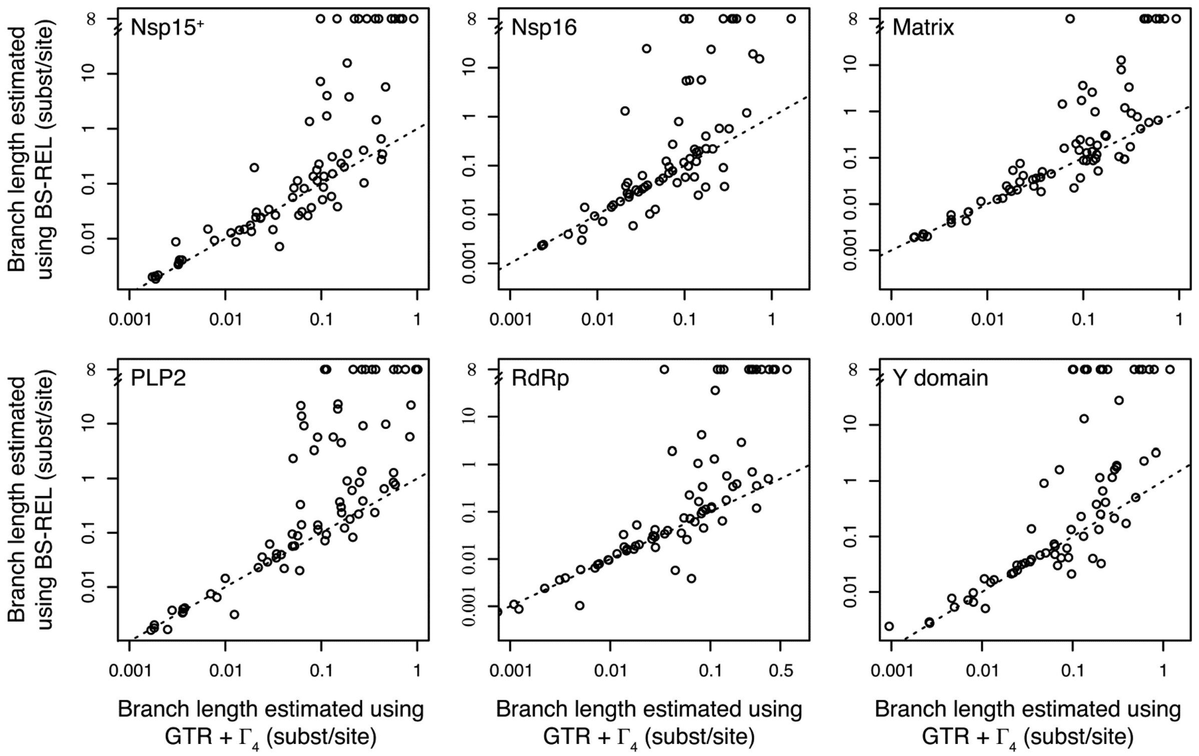

FIG 2 Branch length expansion under the BS-REL model compared to that under the GTR + Γ_4 model for six coronavirus sequences: Nsp15$^+$, Nsp16, matrix protein, PLP2, RdRp, and the Y domain. Each point represents a branch in the coronavirus phylogeny. For ease of visualization, BS-REL branch lengths experiencing extreme saturation (>50 substitutions per site) are depicted with infinite length.

of magnitude, suggesting highly imprecise estimates of branch length expansion in these loci. This lack of precision is not surprising, given the extent to which these loci have experienced saturation along their phylogeny. Once a site becomes saturated with repeated substitutions, it is very difficult for an evolutionary model to determine the exact number of substitutions that have taken place. Therefore, the maximum-likelihood estimate will be imprecise. Nevertheless, these results clearly suggested that there is substantial underestimation of the evolutionary history of coronaviruses.

To examine the effects of saturation and determine the limit of reliable divergence inference with BS-REL, we simulated sequence alignments in HyPhy (24) under the maximum-likelihood parameter values from the BS-REL model fitted to the RdRp locus. HyPhy scripts and input files needed to perform the simulations can be downloaded at https://github.com/veg/pubs/tree/master/SARS. The tree inferred by using this locus experienced a dramatic expansion in a previous BS-REL analysis. For example, a single internal branch at this locus accounted for >70% of the total inferred tree length, but it also experienced incredibly strong purifying selection (dN/dS = 0) at 97.5% of the sites and very strong diversifying selection (dN/dS > 100) at the remaining 2.5% of the sites. In the simulations (n = 100), we kept the relative lengths of branches fixed and scaled the entire tree length from 1 to 10,000 expected substitutions/site. We then fitted the GTR + Γ_4 and BS-REL models to the replicates to evaluate the saturation

TABLE 2 Tree lengths generated under GTR+Γ_4 and BS-REL evolutionary models

Viral sequence	Mean dN/dS[a]	No. of branches approaching saturation under BS-REL[b]	Proportion of sites under pervasive purifying selection[c]	Tree length (no. of substitutions/site) GTR+Γ_4	Tree length (no. of substitutions/site) BS-REL	Expansion under BS-REL	95% CI expansion under BS-REL[d] (lower-upper)
Nsp15$^+$	0.17	12	0.77	11.3	10,530	935	659–100,725
Nsp16	0.12	8	0.85	10.6	6,488	610	439–157,128
Matrix	0.19	9	0.67	14.8	38,287	2,581	1,868–162,126
PLP2	0.26	13	0.70	15.3	67,831	4,432	3,261–166,920
RdRp	0.11	14	0.87	8.0	230,399	28,860	18,699–48,256
Y domain	0.21	16	0.85	13.3	362,318	27,253	16,504–27,434

[a] Inferred by a single-likelihood ancestral counting method (50).
[b] Out of a total of 83 (40 internal) branches in the coronavirus phylogeny.
[c] Inferred by a fixed-effects likelihood method (50).
[d] CIs were estimated by an LHS importance resampling scheme (M = 500 samples).

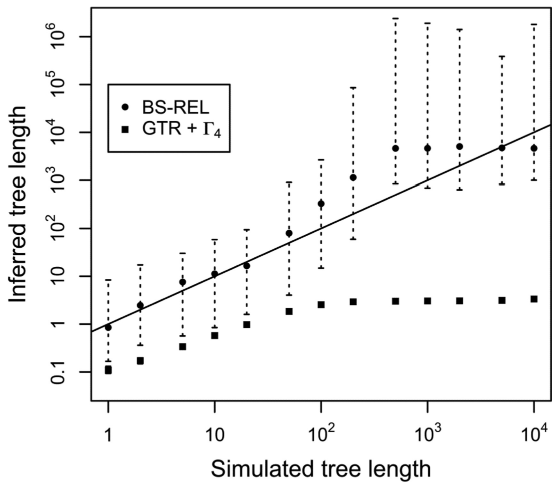

FIG 3 Substitution saturation curves obtained with the GTR + Γ_4 and BS-REL models on the basis of RdRp simulations. The solid line represents the simulated values, symbols are plotted at median values (100 replicates for each tree length value), and dotted whiskers show the 2.5-to-97.5% range.

behavior of the models. There were striking differences between the models (Fig. 3). GTR + Γ_4 consistently, and progressively more badly, underestimated the simulated tree length, starting at the lowest simulated values; this behavior is due to a long internal branch predominantly subject to strong purifying selection, which poses a challenge for the nucleotide-based model (30). Note that this type of saturation does not lead to infinite branch lengths because a majority of the sites are maintained by strong purifying selection. In contrast, the BS-REL model did not saturate until the total length of the tree was on the order of 1,000 substitutions per site. Interestingly, BS-REL estimates are biased slightly upward, largely because of the saturation of nonsynonymous substitutions at positively selected sites and corresponding infinite values of dN/dS.

tMRCA extrapolation. The current best estimate of the tMRCA of all four coronavirus genera of 10,141 years ago was obtained by Woo et al. (10) by using the RdRp locus; their tMRCA estimate is the only one that includes all four coronavirus genera. Notably, their tMRCA estimates for alpha- and betacoronaviruses are similar to the dates inferred by a similar methodology from a different data set (31). Our results suggest that, despite the comprehensive nature of these studies, there are intrinsic shortcomings in the standard evolutionary models used to estimate branch lengths and the tMRCA of ancient viral lineages like coronaviruses. Nevertheless, Woo et al. calibrated the coronavirus molecular clock with serially sampled sequences, which rely predominantly on the relationship between the branches near the tips of the phylogeny to estimate substitution rates. Notably, it is possible to infer consistent RNA virus substitution rates from these types of serially sampled sequences, even if the tMRCA of older lineages deep in the tree cannot be reliably estimated because of saturation along internal branches (e.g., simian immunodeficiency virus/human immunodeficiency virus [32, 33] and avian influenza virus [17, 34]).

When variation in selection pressure was taken into account using BS-REL, we observed a dramatic expansion of the total length of the RdRp phylogeny: 28,860.4-fold (95% confidence interval [CI], 18,695.7-fold to 48,256.0-fold) (Table 2). If one were to extrapolate the evolutionary ages estimated by Woo et al. to the BS-REL phylogeny, the adjusted tMRCA of coronaviruses would be 293 (95% CI, 190 to 489) million years ago, 4 orders of magnitude greater than that previously inferred. The robustness of this tMRCA estimate is highly dependent on both the rate of evolution estimated by Woo et al. (10), which is inferred primarily from short branches near the tips of the phylogeny, and the accuracy of the new branch lengths, which appear extremely variable. Nevertheless, even with the conservative lower bounds of the tMRCA of coronaviruses and branch length expansion (974 BCE [10] and 18,695.7-fold expansion [Table 2]), the adjusted coronavirus tMRCA of 55.8 million years ago would still be 3 orders of magnitude greater than the current estimate.

DISCUSSION

Our results indicate that the evolutionary history of coronaviruses likely extends much further back in time than previous estimates have suggested. Across the coronavirus genome, there is evidence that standard nucleotide models underestimate the amount of evolution that has occurred by orders of magnitude; strong purifying selection had masked the evidence of thousands or millions of years of evolution in the coronavirus phylogeny. Like many other DNA and RNA viruses—including herpesviruses (35, 36), lentiviruses (37), bornaviruses (38), filoviruses (38, 39), and foamy viruses (40, 41)—coronaviruses appear to be an ancient viral lineage. Furthermore, our results demonstrate that purifying selection masking an ancient evolutionary history is not a phenomenon constrained to negative-sense RNA viruses (17) but can be seen in positive-sense RNA viruses like coronaviruses as well.

Interestingly, our extrapolated estimate of the tMRCA of coronaviruses infecting mammalian (alphacoronaviruses and betacoronaviruses) and avian (gammacoronaviruses and deltacoronaviruses) species of 190 to 489 (mean of 293) million years ago corresponds to the inferred tMRCA of these host species based on fossil and molecular evidence of around 325 million years ago (42, 43). It is tempting to speculate that this correspondence between dates is evidence of coevolution and codivergence between bat and avian species and the coronavirus genera. However, given the uncertainty associated with branch length estimation under strong selection (17) and the extreme saturation leading to a dramatic increase in the inferred tree length, our analyses may not provide an accurate estimation of the tMRCA of coronaviruses. Nevertheless, our results strongly suggest that the 10,000-year-ago tMRCA of coronaviruses is underestimated by orders of magnitude. These results leave open the possibility that coronaviruses have been infecting bats and/or birds since the origin of these clades tens of millions of years ago or possibly since their divergence from each other in the carboniferous period, over 300 million years ago. This extrapolation, rather than be considered a reliable estimate of the coronavirus tMRCA, should be viewed as a biologically plausible hypothesis based on realistic parameters (e.g., patterns of substitution rates and selection profiles). We can no longer reject an ancient coevolutionary relationship based on the molecular clock.

The degree to which host switching and codivergence have shaped coronavirus diversity remains unresolved. The observation of recent viral cross-species transmission events among mammalian (9, 11) and avian (8) species is clear evidence of recent host switching. Conversely, the separation of mammalian and avian coronaviruses into distinct genera is suggestive of codivergence: mammalian coronaviruses (i.e., alphacoronaviruses and betacoronaviruses) are generally inferred to be reciprocally monophyletic (8, 10, 11). Therefore, a formal analysis of host switching and codivergence in coronaviruses would be useful for disentangling which sections of the coronavirus phylogeny represent codivergence and which represent host-switching events.

For the shorter branches (≤0.05 substitution per site) in the coronavirus phylogenetic trees, we found general agreement in the inferred branch lengths between evolutionary models (GTR + Γ_4 and BS-REL); there was no evidence of underestimation of the lengths of short branches. Therefore, for closely related viral lineages involved in recent zoonotic transmission events (e.g., SARS coronavirus in humans, bats, and civet cats), previous dating estimates (44) are consistent with our findings. Furthermore, this observation suggests that substitution rate estimates inferred from these recent outbreaks are robust to the biasing effects of purifying selection with respect to branch length estimation (though coalescent effects may still have an impact on these recent evolutionary rate estimates [45–47]). However, it remains unclear how the lower-than-average mutation rate of coronaviruses (10^{-5} to 10^{-6} mutation per site per replication) (15) translates into a typical short-term substitution rate (10^{-3} substitution per site per year) (44). Further investigation on this topic is needed.

BS-REL is an attractive tool for future studies of ancient virus evolution. The use of evolutionary models that allow for variable selection pressure across all branches in a phylogeny when estimating branch lengths is an advance over previous implementations by Wertheim and Kosakovsky Pond (17, 25), which necessitate *a priori* identification of long internal branches bearing the mark of strong purifying selection (30). Unlike the phylogenetic trees in this previous study, which were characterized by closely related isolates separated by long internal branches, the coronavirus phylogenies are complicated, with a mixture of long and short branches interspersed throughout the tree. The BS-REL framework represents a flexible and powerful approach to the modeling of natural selection in the evolution of viruses (25), and it does not necessitate designating which branches experienced stronger or weaker selection pressures.

Moreover, like our previous approach to the analysis of Ebola and avian influenza viruses, BS-REL found evidence of branches experiencing saturation. Because of the way in which BS-REL parameterizes variable selection, the saturated branches were estimated to be longer in coronaviruses than in Ebola and avian influenza viruses.

The BS-REL model almost certainly overfits data on short branches with simple patterns of natural selection, which likely affects the accuracy of branch length estimates across the tree. In the case of coronaviruses, this overfitting is not a serious problem because the expansion of branch length estimates due to saturation is extreme and unlikely to produce precise estimates. A more parsimonious implementation of BS-REL would be useful for addressing issues in viral evolution in which more precise branch length estimates are needed. Appropriate modeling of variation in the strength of natural selection will be integral for obtaining more accurate tMRCAs of viral lineages. Furthermore, it is possible that employing more realistic evolutionary models, for example, in maximum-likelihood or Bayesian tree searches, could improve the quality of viral phylogenetic inference.

In summary, our results indicate that coronaviruses have an evolutionary history much longer than those suggested by phylogenetic trees inferred by using standard nucleotide evolution models. This finding allows for a coevolutionary relationship between coronaviruses and their natural hosts. It is possible that such a long-term relationship has allowed some animal species (e.g., bats) to evolve strategies to coexist with coronaviruses and vice versa (48, 49). Further investigation of this topic might help us to better understand virus-host coevolution, the origin of coronaviruses, and other related viral families in the order *Nidovirales*.

ACKNOWLEDGMENTS

We thank Vijaykrishna Dhanasekaran (Duke-NUS Graduate Medical School Singapore) and two anonymous reviewers for their helpful comments.

This work was supported in part by a National Institutes of Health training fellowship (AI43638) to J.O.W. This project was supported by the National Institutes of Health (NIAID contract HHSN266200700005C, grants AI47745, GM093939, AI74621, and DA034978), the UCSD Center for AIDS Research (AI36214), UC Laboratory Fees Research Program grant 12-LR-236617, and EMPERIE (grant EU FP7 223498).

REFERENCES

1. **de Groot RJ, Baker SC, Baric R, Enjuanes L, Gorbalenya AE, Holmes KV, Perlman S, Poon L, Rottier PJM, Talbot PJ, Woo PCY, Ziebuhr J.** 2011. Family *Coronaviridae*, p 806–828. *In* King AMQ, Adams MJ, Carstens EB, Lefkowitz EJ (ed), Virus taxonomy: classification and nomenclature of viruses: ninth report of the International Committee on Taxonomy of Viruses. Academic Press, Ltd., London, United Kingdom.
2. **Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, Nicholls J, Yee WK, Yan WW, Cheung MT, Cheng VC, Chan KH, Tsang DN, Yung RW, Ng TK, Yuen KY.** 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. Lancet **361:**1319–1325.
3. **Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LL, Guan Y, Rozanov M, Spaan WJ, Gorbalenya AE.** 2003. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J. Mol. Biol. **331:**991–1004.
4. **Poon LL, Chu DK, Chan KH, Wong OK, Ellis TM, Leung YH, Lau SK, Woo PC, Suen KY, Yuen KY, Guan Y, Peiris JS.** 2005. Identification of a novel coronavirus in bats. J. Virol. **79:**2001–2009.
5. **Wevers BA, van der Hoek L.** 2009. Recently discovered human coronaviruses. Clin. Lab Med. **29:**715–724.
6. **Woo PC, Lau SK, Huang Y, Yuen KY.** 2009. Coronavirus diversity, phylogeny and interspecies jumping. Exp. Biol. Med. (Maywood) **234:**1117–1127.
7. **Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA.** 2012. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N. Engl. J. Med. **367:**1814–1820.
8. **Chu DK, Leung CY, Gilbert M, Joyner PH, Ng EM, Tse TM, Guan Y, Peiris JS, Poon LL.** 2011. Avian coronavirus in wild aquatic birds. J. Virol. **85:**12815–12820.
9. **Vijaykrishna D, Smith GJ, Zhang JX, Peiris JS, Chen H, Guan Y.** 2007. Evolutionary insights into the ecology of coronaviruses. J. Virol. **81:**4012–4020.
10. **Woo PC, Lau SK, Lam CS, Lau CC, Tsang AK, Lau JH, Bai R, Teng JL, Tsang CC, Wang M, Zheng BJ, Chan KH, Yuen KY.** 2012. Discovery of seven novel mammalian and avian coronaviruses in the genus deltacoronavirus supports bat coronaviruses as the gene source of alphacoronavirus and betacoronavirus and avian coronaviruses as the gene source of gammacoronavirus and deltacoronavirus. J. Virol. **86:**3995–4008.
11. **Lau SK, Li KS, Tsang AK, Shek CT, Wang M, Choi GK, Guo R, Wong BH, Poon RW, Lam CS, Wang SY, Fan RY, Chan KH, Zheng BJ, Woo PC, Yuen KY.** 2012. Recent transmission of a novel alphacoronavirus, bat coronavirus HKU10, from Leschenault's rousettes to Pomona leaf-nosed bats: first evidence of interspecies transmission of coronavirus between bats of different suborders. J. Virol. **86:**11906–11918.
12. **Gorbalenya AE.** 2008. Genomics and evolution of the *Nidovirales*, p 15–28. *In* Perlman S, Gallagher T, Snijder EJ (ed), Nidoviruses. ASM Press, Washington, DC.
13. **Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ.** 2006. Nidovirales: evolving the largest RNA virus genome. Virus Res. **117:**17–37.

14. **Minskaia E, Hertzig T, Gorbalenya AE, Campanacci V, Cambillau C, Canard B, Ziebuhr J.** 2006. Discovery of an RNA virus 3′→5′ exoribonuclease that is critically involved in coronavirus RNA synthesis. Proc. Natl. Acad. Sci. U. S. A. **103**:5108–5113.
15. **Eckerle LD, Becker MM, Halpin RA, Li K, Venter E, Lu X, Scherbakova S, Graham RL, Baric RS, Stockwell TB, Spiro DJ, Denison MR.** 2010. Infidelity of SARS-CoV Nsp14-exonuclease mutant virus replication is revealed by complete genome sequencing. PLoS Pathog. **6**:e1000896. doi:10.1371/journal.ppat.1000896.
16. **Duffy S, Shackelton LA, Holmes EC.** 2008. Rates of evolutionary change in viruses: patterns and determinants. Nat. Rev. Genet. **9**:267–276.
17. **Wertheim JO, Kosakovsky Pond SL.** 2011. Purifying selection can obscure the ancient age of viral lineages. Mol. Biol. Evol. **28**:3355–3365.
18. **Ziebuhr J, Thiel V, Gorbalenya AE.** 2001. The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond. J. Biol. Chem. **276**:33220–33232.
19. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**:1792–1797.
20. **Chu DK, Peiris JS, Chen H, Guan Y, Poon LL.** 2008. Genomic characterizations of bat coronaviruses (1A, 1B and HKU8) and evidence for co-infections in *Miniopterus* bats. J. Gen. Virol. **89**:1282–1287.
21. **Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD.** 2006. GARD: a genetic algorithm for recombination detection. Bioinformatics **22**:3096–3098.
22. **Guindon S, Delsuc F, Dufayard JF, Gascuel O.** 2009. Estimating maximum likelihood phylogenies with PhyML. Methods Mol. Biol. **537**:113–137.
23. **Gouy M, Guindon S, Gascuel O.** 2010. SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol. Biol. Evol. **27**:221–224.
24. **Kosakovsky Pond SL, Frost SD, Muse SV.** 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics **21**:676–679.
25. **Kosakovsky Pond SL, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K.** 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol. Biol. Evol. **28**:3033–3043.
26. **Kosakovsky Pond SL, Scheffler K, Gravenor MB, Poon AF, Frost SD.** 2010. Evolutionary fingerprinting of genes. Mol. Biol. Evol. **27**:520–536.
27. **Hasegawa M, Kishino H.** 1989. Confidence limits on the maximum-likelihood estimate of the hominoid tree from mitochondrial-DNA sequences. Evolution **43**:672–677.
28. **Kishino H, Hasegawa M.** 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J. Mol. Evol. **29**:170–179.
29. **Shimodaira H, Hasegawa J.** 1999. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol. Biol. Evol. **16**:1114–1116. (Letter.)
30. **Kosakovsky Pond S, Muse SV.** 2005. Site-to-site variation of synonymous substitution rates. Mol. Biol. Evol. **22**:2375–2385.
31. **Pfefferle S, Oppong S, Drexler JF, Gloza-Rausch F, Ipsen A, Seebens A, Muller MA, Annan A, Vallo P, Adu-Sarkodie Y, Kruppa TF, Drosten C.** 2009. Distant relatives of severe acute respiratory syndrome coronavirus and close relatives of human coronavirus 229E in bats, Ghana. Emerg. Infect. Dis. **15**:1377–1384.
32. **Wertheim JO, Worobey M.** 2009. Dating the age of the SIV lineages that gave rise to HIV-1 and HIV-2. PLoS Comput. Biol. **5**:e1000377. doi:10.1371/journal.pcbi.1000377.
33. **Worobey M, Telfer P, Souquiere S, Hunter M, Coleman CA, Metzger MJ, Reed P, Makuwa M, Hearn G, Honarvar S, Roques P, Apetrei C, Kazanji M, Marx PA.** 2010. Island biogeography reveals the deep history of SIV. Science **329**:1487.
34. **Chen R, Holmes EC.** 2010. Hitchhiking and the population genetic structure of avian influenza virus. J. Mol. Evol. **70**:98–105.
35. **McGeoch DJ, Cook S.** 1994. Molecular phylogeny of the alphaherpesvirinae subfamily and a proposed evolutionary timescale. J. Mol. Biol. **238**:9–22.
36. **McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EA.** 1995. Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. J. Mol. Biol. **247**:443–458.
37. **Gifford RJ, Katzourakis A, Tristem M, Pybus OG, Winters M, Shafer RW.** 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.
38. **Belyi VA, Levine AJ, Skalka AM.** 2010. Unexpected inheritance: multiple integrations of ancient Bornavirus and Ebolavirus/Marburgvirus sequences in vertebrate genomes. PLoS Pathog. **6**:e1001030. doi:10.1371/journal.ppat.1001030.
39. **Taylor DJ, Leach RW, Bruenn J.** 2010. Filoviruses are ancient and integrated into mammalian genomes. BMC Evol. Biol. **10**:193. doi:10.1186/1471-2148-10-193.
40. **Han GZ, Worobey M.** 2012. An endogenous foamy-like viral element in the coelacanth genome. PLoS Pathog. **8**:e1002790. doi:10.1371/journal.ppat.1002790.
41. **Katzourakis A, Gifford RJ, Tristem M, Gilbert MT, Pybus OG.** 2009. Macroevolution of complex retroviruses. Science **325**:1512. doi:10.1126/science.1174149.
42. **Blair JE, Hedges SB.** 2005. Molecular phylogeny and divergence times of deuterostome animals. Mol. Biol. Evol. **22**:2275–2284.
43. **Shedlock AM, Edwards SV.** 2009. Amniotes (Amniota), p 375–379. *In* Hedges SB, Kumar S (ed), The timetree of life. Oxford University Press, Oxford, United Kingdom.
44. **Hon CC, Lam TY, Shi ZL, Drummond AJ, Yip CW, Zeng F, Lam PY, Leung FC.** 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.
45. **Nicolaisen LE, Desai MM.** 2012. Distortions in genealogies due to purifying selection. Mol. Biol. Evol. **29**:3589–3600.
46. **O'Fallon BD.** 2010. A method to correct for the effects of purifying selection on genealogical inference. Mol. Biol. Evol. **27**:2406–2416.
47. **Ho SY, Lanfear R, Bromham L, Phillips MJ, Soubrier J, Rodrigo AG, Cooper A.** 2011. Time-dependent rates of molecular evolution. Mol. Ecol. **20**:3087–3101.
48. **Wang LF, Walker PJ, Poon LL.** 2011. Mass extinctions, biodiversity and mitochondrial function: are bats 'special' as reservoirs for emerging viruses? Curr. Opin. Virol. **1**:649–657.
49. **Zhang G, Cowled C, Shi Z, Huang Z, Bishop-Lilly KA, Fang X, Wynne JW, Xiong Z, Baker ML, Zhao W, Tachedjian M, Zhu Y, Zhou P, Jiang X, Ng J, Yang L, Wu L, Xiao J, Feng Y, Chen Y, Sun X, Zhang Y, Marsh GA, Crameri G, Broder CC, Frey KG, Wang LF, Wang J.** 2013. Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. Science **339**:456–460.
50. **Kosakovsky Pond SL, Frost SD.** 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. **22**:1208–1222.