bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Inferring Species Trees Using Integrative Models of Species Evolution

Huw A. Ogilvie*[1,2], Timothy G. Vaughan[3,4], Nicholas J. Matzke[1,5], Graham J. Slater[6], Tanja Stadler[3,4], David Welch[2,7], and Alexei J. Drummond[2,7]

[1]Division of Ecology and Evolution, Research School of Biology, Australian National University, Canberra, Australia

[2]Centre for Computational Evolution, University of Auckland, Auckland, New Zealand

[3]Computational Evolution Group, Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland

[4]Swiss Institute of Bioinformatics, Lausanne, Switzerland

[5]School of Biological Sciences, University of Auckland, Auckland, New Zealand

[6]Department of the Geophysical Sciences, University of Chicago, Chicago, Illinois

[7]Department of Computer Science, University of Auckland, Auckland, New Zealand

January 6, 2018

*huw.ogilvie@anu.edu.au

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Abstract

Bayesian methods can be used to accurately estimate species tree topologies, times and other parameters, but only when the models of evolution which are available and utilized sufficiently account for the underlying evolutionary processes. Multispecies coalescent (MSC) models have been shown to accurately account for the evolution of genes within species in the absence of strong gene flow between lineages, and fossilized birth-death (FBD) models have been shown to estimate divergence times from fossil data in good agreement with expert opinion. Until now dating analyses using the MSC have been based on a fixed clock or informally derived node priors instead of the FBD. On the other hand, dating analyses using an FBD process have concatenated all gene sequences and ignored coalescence processes. To address these mirror-image deficiencies in evolutionary models, we have developed an integrative model of evolution which combines both the FBD and MSC models. By applying concatenation and the MSC (without employing the FBD process) to an exemplar data set consisting of molecular sequence data and morphological characters from the dog and fox subfamily Caninae, we show that concatenation causes predictable biases in estimated branch lengths. We then applied concatenation using the FBD process and the combined FBD-MSC model to show that the same biases are still observed when the FBD process is employed. These biases can be avoided by using the FBD-MSC model, which coherently models fossilization and gene evolution, and does not require an *a priori* substitution rate estimate to calibrate the molecular clock. We have implemented the FBD-MSC in a new version of StarBEAST2, a package developed for the BEAST2 phylogenetic software.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

We have vastly more data on biological organisms than at any point in the past; whole genome sequences, ancient DNA, morphological characters and fossil occurrences all contain a fingerprint of past evolutionary processes. With this wealth of data, we should expect coherent estimates of the pattern and timing of evolutionary events. Yet the story told by genomes and molecular clocks is often difficult to reconcile with morphological data and the fossil record (Meyer et al. 2012; O'Leary et al. 2013; Jarvis et al. 2014; dos Reis et al. 2014; Mitchell et al. 2015). These debates are often described as "rocks versus clocks" (Donoghue and Benton 2007) with famous examples including the timing of the origin of placental mammals (O'Leary et al. 2013; dos Reis et al. 2014), birds (Jarvis et al. 2014; Mitchell et al. 2015), flowering plants (Beaulieu et al. 2015), and the Cambrian Explosion (Lee et al. 2013). Most disturbingly, these debates persist even for evolutionarily recent and intensively studied questions like the timing of the human-chimp split, where fossils (Brunet et al. 2002; White et al. 2009; Wood and Harrison 2011; White et al. 2015) give different results than genomic data (Patterson et al. 2006; Langergraber et al. 2012; Meyer et al. 2012; Scally et al. 2012; Scally and Durbin 2012; Callaway 2015; Lipson et al. 2015).

Bayesian inference, the gold-standard in estimating evolutionary history (Huelsenbeck et al. 2001; Ronquist and Huelsenbeck 2003; Nylander et al. 2004; Drummond et al. 2012; Bouckaert et al. 2014; Höhna et al. 2016), provides a theoretical framework that supports the integration of multiple data sources. So called "total-evidence" analyses integrate molecular sequence and morphological character data. Where a fossil record is available, total-evidence data sets can be used with "tip-dating" methods to estimate time-calibrated species trees (Ronquist et al. 2012; Gavryushkina et al. 2014; Zhang et al. 2016; Gavryushkina et al. 2017).

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tip-dating makes an advance over previous methods such as node-dating or a fixed clock by treating fossils as data. Node-dating, where researchers propose parametric prior distributions for the dates of particular nodes based on expert opinion and intuition, can result in misleading node ages (Gavryushkina et al. 2017). An alternative to tip- or node-dating is a fixed molecular clock. Fixing the molecular clock at 1 means that only relative divergence times can be estimated, while using a value from a previous study assumes that the *a priori* rate is accurate for the species and loci in the new study.

Previous implementations of tip-dating have so far made the assumption of a single phylogeny encompassing all molecular loci and morphological characters. This assumption is known as "concatenation" because it is equivalent to concatenating several multiple sequence alignments into a single alignment, and it has been demonstrated to cause biases and overestimated precision when inferring species trees from molecular data (Liu et al. 2015; Ogilvie et al. 2016; Ogilvie et al. 2017). To enable the combined use of molecular, morphological and fossil data with the advantage of tip-dating and without the known problems of concatenation, we propose combining models of genealogical evolution, morphological evolution, and of speciation, extinction and fossilization.

The Fossilized Birth-Death Process

Explicitly including fossils in stochastic models of phylogenies became possible with the birth-death-serial-sampling model (Stadler 2010). This model has three macroevolutionary parameters; the fossil sampling rate ψ, the speciation rate λ and the extinction rate μ. A version of this model, named the fossilized birth-death (FBD) process, allows for sampled ancestors (Gavryushkina et al. 2014; Heath et al. 2014; Zhang et al. 2016); each fossil may be either a direct ancestor of other

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

samples, or a tip branch if no descendants have been sampled. The "skyline" extension to the FBD (Stadler et al. 2013) allows the macroevolutionary parameters to vary through time in an arbitrary and independent fashion.

The Multispecies Coalescent

Modern phylogenetic inference distinguishes between high level phylogenetic relationships across species described by a species tree and relationships between individual alleles described by gene trees. It is now well understood that failure to take this into account can significantly bias results due to the effects of incomplete lineage sorting (ILS) and other processes (Liu et al. 2015; Linkem et al. 2016; Mendes and Hahn 2016; Mendes and Hahn 2018).

*BEAST (Heled and Drummond 2010), StarBEAST2 (Ogilvie et al. 2017), BEST (Liu 2008) and BPP (Yang 2015; Rannala and Yang 2017) are all examples of Bayesian software that explicitly sample the joint posterior distribution over both species and gene trees under the multispecies coalescent (MSC) model, as described by Maddison (1997) and Degnan and Rosenberg (2009). These methods account for the hierarchical nature of the evolutionary process and explicitly model ILS. However none of these implementations allow fossils or other ancestral samples to be placed directly on the species tree, meaning that tip-dating approaches are not possible.

Integrative Models of Species Evolution

Integrative models are desirable because they can integrate over uncertainty rather than assuming fixed parameters, and they can also directly utilize more sources of data than simpler models.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

In this paper we describe an integrative Bayesian phylogenetic model for estimating species trees and divergence times, capable of analyzing multilocus genetic data, fossil occurrence data and morphological data in a coherent probabilistic inference framework. The model reconciles molecular and fossil evidence by explicitly distinguishing two evolutionary processes, with the FBD process describing the distribution over species trees and the MSC model describing the probability distribution of molecular genealogies conditional on the species tree.

The FBD branching model of macroevolution accounts for speciation, extinction and fossilization. The species tree is modeled using the FBD process, with the morphology of all species arising from a stochastic process of evolution that proceeds down the branches of this species tree.

The MSC has become the standard model for describing the relationship between molecular genealogies and species trees. The molecular sequence data (sampled from extant individuals or as ancient DNA) are modeled by multiple independent gene trees, which may differ from each other due to processes such as ILS, but must be consistent with the shared species tree that they have all evolved within (Fig. 1).

The BEAST2 phylogenetic software features "StarBEAST2" — a recent implementation of the MSC — and an implementation of the FBD prior (Gavryushkina et al. 2014). We have updated StarBEAST2 to combine the MSC model with the FBD process, henceforth "FBD-MSC". To demonstrate the utility of the FBD-MSC model, we applied the latest version of StarBEAST2 to an exemplar data set of the dog and fox subfamily Caninae.

Estimates made under the FBD-MSC model are compared with estimates made using FBD with concatenation (henceforth "FBD-concatenation"), the MSC with a fixed molecular clock in-

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

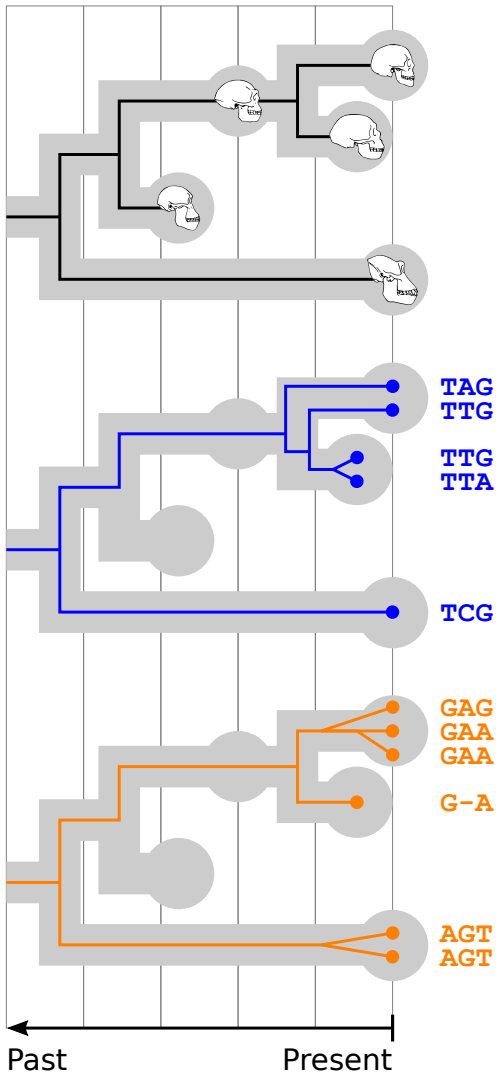

Figure 1: A species tree with a single sampled ancestor and its relationship to morphological data (top) and multilocus sequence alignments (middle and bottom) in a unified model.

stead of an FBD prior, and concatenation with a fixed clock. FBD-MSC results were generally in agreement with fixed clock MSC estimates. Concatenation overestimated tip branch lengths, species divergence times, and the timing of diversification leading to extant Caninae, even when fossil data was incorporated using the FBD model.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Methods

Integrative Model Probability

The integrative model combining the MSC, the FBD process, and morphological evolution can be expressed by combining the component likelihoods. The likelihood of a gene tree is the phylogenetic likelihood (Felsenstein 1981) $\Pr(D_i|G_i)$ where D_i is the multiple sequence alignment (MSA) for the ith gene tree G_i. The MSC probability for that gene tree is $P(G_i|S)$ where S is the species tree. The likelihood contribution to the species tree of a morphological character is the phylogenetic likelihood $\Pr(C_j|S)$ where C_j is the vector of states for the jth character. The prior probability of the species tree under the FBD process is $P(S|\theta)$, where θ is a vector of FBD parameters as described by Gavryushkina et al. (2014). Combining the likelihoods for the integrative model we get the probability of the species tree given the molecular, morphological and fossil data:

$$P(S, G, \theta|D, C) = \frac{1}{Z} \prod_i (\Pr(D_i|G_i) \cdot P(G_i|S)) \cdot \prod_j \Pr(C_j|S) \cdot P(S|\theta), \tag{1}$$

where $Z = 1/\Pr(D, C)$ is the marginal likelihood, an unknown normalizing constant that does not need to be computed when using Markov chain Monte Carlo (MCMC) to sample from the posterior distribution.

Under this model, the MSAs inform the species tree through the gene trees, whereas the morphological characters inform the species tree directly. Ultimately both the MSAs and morphological characters inform the FBD parameters through the species tree (Fig. 2).

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

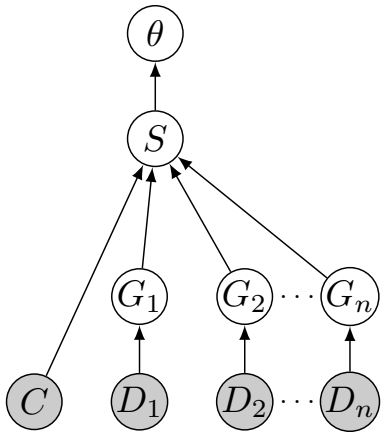

Figure 2: Graphical relationship between key elements of the model. Grey nodes represent data including morphological characters (C) and multiple sequence alignments (D_i). White nodes represent estimated parameters including gene trees (G_i), the species tree (S) and fossilized birth-death parameters (θ).

Sampling and Simulating Trees from the Prior

We tested our implementation of the FBD-MSC model by using it to jointly sample species trees with a single embedded gene tree from the prior, and comparing those distributions with FBD trees and with gene trees produced by direct simulation. Three and four-taxon FBD trees were sampled from the prior using the "SA" package in BEAST2 (Gavryushkina et al. 2014). Following the parameterization in Gavryushkina et al. (2014), these trees were conditioned on an origin time t_{or} of 3, a birth rate λ of 1, a death rate μ of 0.5, a sampling rate ψ of 0.1, a removal probability r of 0 and a present-day sampling probability ρ of 0.1.

The sampled taxa for three-taxon FBD trees were labelled A, B and C, and had fixed ages of 0, 1, and 1.5 respectively. The sampled taxa for the four-taxon FBD trees were labelled A, B, C and D, and had fixed ages of 0, 0, 0.5 and 2 respectively.

MCMC chains to sample FBD trees were run for 100 million steps, and 100,000 trees sampled

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

at a rate of 1 per 1,000 steps. One gene tree was simulated for each sample using custom Java code available as part of the StarBEAST2 package, assuming effective population sizes fixed at 1 for each branch.

When jointly sampling FBD-MSC species and gene trees from the prior using StarBEAST2, identical parameters were used but MCMC chains were run for 500 million steps. Species and gene trees were sampled at a rate of 1 per 5,000 steps for 100,000 species trees and the same number of gene trees.

Compiling Caninae Data

Unphased molecular sequences were retrieved from NCBI GenBank. Sequences from Bardeleben et al. (2005*a*) had accession numbers AY609082–AY609158. Sequences from Bardeleben et al. (2005*b*) had accession numbers AY885308–AY885426. Sequences from Lindblad-Toh et al. (2005) had accession numbers DQ239439–DQ239486 and DQ240289–DQ240817. Outgroup (non-Caninae) and domestic dog sequences were discarded. *Canis aureus* was renamed *Canis anthus* following Koepfli et al. (2015). For each locus, we aligned those sequences to produce an MSA using PRANK (Löytynoja and Goldman 2005). Phased MSAs were generated by duplicating each aligned sequence and randomly phasing heterozygous sites.

Coded morphological data, character names, character state names and tip dates from Slater (2015) were retrieved from Dryad (https://doi.org/10.5061/dryad.9qd51). This data set built on previous monographs (Wang 1994; Wang et al. 1999; Tedford et al. 2009).

Outgroup characters and characters invariable within Caninae were discarded. *Canis aureus*

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

was again renamed *Canis anthus*, and *Cuon javanicus* was renamed *Cuon alpinus*, a synonym used in the molecular sequence data. For species with molecular sequences but no morphological data, all characters were treated as missing data. An extant-only data set was produced by discarding fossil taxon characters, and characters invariable within extant Caninae. BEAST2-compatible NEXUS files were generated containing the coded data and names.

MSC and Concatenation Analyses

The MSC (in practice, StarBEAST2) was configured to estimate a constant population size separately for each branch, with a maximum effective population size of 2, and a $1/X$ prior on the mean population size. Phased sequences were used with StarBEAST2, and unphased sequences with concatenation. For both StarBEAST2 and concatenation, we set the uniform priors $U(0, 2)$ and $U(0, 1)$ on the diversification rate $\lambda - \mu$ and on the turnover parameter $\mu \div \lambda$ respectively.

The mean substitution rate was either fixed at 8×10^{-4} substitutions per million years, or estimated with a lognormal prior which had a mean of 7.5×10^{-4} and a standard deviation of the log rate of 0.6. Substitution rates among loci were allowed to vary with a flat Dirichlet prior. The HKY substitution model was used for molecular data (Hasegawa et al. 1985), and transition/transversion ratios estimated separately for each locus. The Mkv model (Lewis 2001) was used to model the evolution of morphological characters, assuming character state frequencies and transition rates are all equal. A morphological clock was estimated with a $1/X$ prior and an upper bound of 1.

FBD analyses were conditioned on t_{or} which was estimated with a uniform prior $U(0, 1000)$. The sampling proportion $\psi \div (\psi + \mu)$ was also estimated with a uniform prior $U(0, 1)$. The other

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FBD parameters r and ρ were fixed at 0 and 1 respectively.

For each fixed clock analysis, we ran 20 independent MCMC chains of 400 million states each, sampling once every 200,000 states, and discarded the first 10% of samples as burnin. For each fossilized birth-death analysis, we ran 20 independent MCMC chains of 15 billion states each, sampling once every 2 million states, and discarded the first 4% of samples from each chain as burnin. For each type of analysis, the independent chains were concatenated and subsampled for a combined sample of 2,000 states.

Posterior Predictive Simulations

For half (1,000) of the fixed clock StarBEAST2 posterior samples, we resimulated molecular and morphological data. For each locus a gene tree was simulated according to the MSC using DendroPy (Sukumaran and Holder 2010), embedded within the species tree (topology, times and per-branch population sizes) for that sample, with two alleles per extant species. An MSA was simulated for each gene tree using Seq-Gen (Rambaut and Grassly 1997), based on the HKY model with the estimated κ ratio and substitution rate of the locus from the posterior sample, and of the same length as the original locus. Unphased per-species sequences were generated using ambiguity codes for heterozygous sites.

Morphological data was resimulated by simulating a 1,000 character MSA along the posterior sample's species tree with 20 states per character, again using Seq-Gen. Base frequencies and transition rates were all equal, and the substitution rate set to 0.03. Then for each morphological character in the original data set, we sampled without replacement one of the simulated characters

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

with a matching number of observed states.

Each simulation was reanalyzed using concatenation with the same model and priors as for the original data set. However only one chain of 200 million states was run for each simulation, sampling once every 80,000 states, and 20% of samples were discarded as burnin.

Calculating Summary Statistics

Summary statistics were calculated for each estimated distribution of trees using DendroPy. These included the maximum clade credibility (MCC) tree, branch lengths, node heights, branch support and node support. For the purpose of calculating support values and internal node heights, a node is defined as the root of a subtree containing all of, and only, a given set of extant taxa. A branch is defined as the direct connection between parent and child nodes as defined above. Lineages-through-time (LTT) curves for FBD analyses were calculated using a custom script. Summary statistics and LTT plots were visualized using ggplot2 (Wickham 2016) and ggtree (Yu et al. 2017).

Results

FBD-MSC Implementation Correctness

To test the correctness of our FBD-MSC implementation, we first compared distributions of three and four-taxon FBD trees drawn from the prior using BEAST2 without the MSC, to distributions drawn from the prior using the FBD-MSC model in StarBEAST2. The marginal divergence

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

time (Supplementary Fig. S1,S2) and topology (Supplementary Fig. S3,S4) distributions thus generated were found to be identical between implementations. As the BEAST2 implementation of the FBD model has been previously verified (Gavryushkina et al. 2014), this is strong evidence that the new implementation is also correct.

Gene trees were also sampled from the prior under the FBD-MSC model in StarBEAST2, and were compared to a distribution of gene trees simulated evolving within the FBD trees that were drawn from the prior absent StarBEAST2. The distributions of gene tree coalescent times (Supplementary Fig. S5,S6) and topologies (Supplementary Fig. S7,S8) were identical for either method, further supporting the correctness of our implementation.

Compiling an Exemplar Dataset

To demonstrate the effects of estimating species divergence times without accounting for coalescent processes, as when using concatenation, we compiled a data set by combining 19 previously published Caninae nuclear locus sequences from extant Caninae taxa (Table 1) with morphological characters and times from extant and fossil Caninae (Slater 2015). The combined data set included 21 extant taxa with molecular data only, 9 extant taxa with molecular and morphological data, and 31 fossil taxa with tip dates and morphological data. After removing characters with no variation within Caninae, there were 72 morphological characters remaining for FBD analyses. After further removing characters with no variation among the 9 taxa with both molecular and morphological data, there were 55 remaining for fixed clock analyses.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 1: Nineteen nuclear loci used in this study.

Locus name	MSA length*	Publication
APOBS1	702	Lindblad-Toh et al. (2005)
BDNF	489	Lindblad-Toh et al. (2005)
BRCA1S2	741	Lindblad-Toh et al. (2005)
Ch14	921	Lindblad-Toh et al. (2005)
Ch21	601	Lindblad-Toh et al. (2005)
Ch24	730	Lindblad-Toh et al. (2005)
CHRNA1	376	Bardeleben et al. (2005*a*)
CHST12	705	Lindblad-Toh et al. (2005)
CMKOR1	735	Lindblad-Toh et al. (2005)
CYPIA1	619	Bardeleben et al. (2005*a*)
FES	483	Bardeleben et al. (2005*a*)
FGFR3	503	Lindblad-Toh et al. (2005)
GHR	821	Bardeleben et al. (2005*a*)
RAG1	741	Lindblad-Toh et al. (2005)
TMEM20	615	Lindblad-Toh et al. (2005)
TRSP	722	Bardeleben et al. (2005*b*)
VANGL2	546	Lindblad-Toh et al. (2005)
VTN	487	Bardeleben et al. (2005*a*)
VWF	732	Lindblad-Toh et al. (2005)

*the number of sites in the multiple sequence alignment

Calibrating Species Trees Using a Fixed Clock

In the absence of a fossil record for a clade of interest, divergence times can be estimated using a fixed molecular clock. This scales the tree by an *a priori* chosen substitution rate, or a set of substitution rates for a set of genes. Substitution rates have been previously estimated for the nuclear RAG1 gene across multiple tetrapod clades, and for mammals the rate is approximately 1×10^{-3} substitutions per site per million years (Hugall et al. 2007). Exploratory analyses suggested that RAG1 evolves around 25% more quickly than the mean rate for all genes in our study, so we used a substitution rate fixed at 8×10^{-4} for analyses calibrated with a fixed clock.

We compared the posterior distribution of species trees inferred under the MSC and concatenation without any fossil data, including nuclear loci and morphological characters only from extant

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

taxa, and using a birth-death prior for the species tree. The estimated lengths of all tip branches and some internal branches were longer when using concatenation (Fig. 3). A few internal branches were shorter, most of all the 1–2, 5–A and E–J branches.

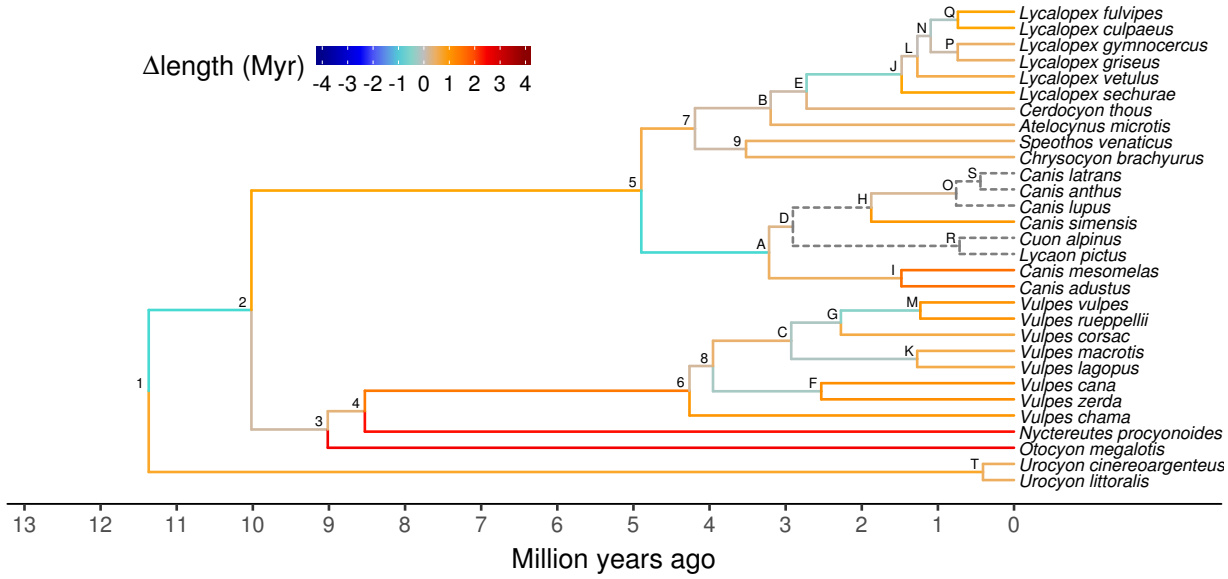

Figure 3: Branch length changes resulting from concatenation using a fixed clock. The color shows how branch length estimates differ when using concentration rather than the multispecies coalescent (MSC). The tree is a maximum clade credibility (MCC) summary tree with mean node ages, generated from the MSC posterior distribution of species trees, inferred using molecular and extant morphological data with a fixed clock. The difference in branch lengths is the mean among concatenation samples including that branch, less the MSC mean. Dashed lines represent branches with less than 0.5% support using concatenation.

To understand whether failing to account for neutral coalescent processes could cause the observed branch length differences, we used posterior predictive simulations to model the expected differences. For 1,000 species tree samples in the fixed clock MSC posterior distribution, we resimulated gene trees according to the MSC. For each simulated gene tree, we simulated an MSA based on that sample's substitution rates and transition/transversion ratios. A set of morphological characters were also simulated along the species tree for each sample. Posterior distributions of species trees using concatenation were then inferred from the simulated data.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

For a given branch, we calculated the distribution of differences in branch length $\Delta(l_b - \bar{l}_b)$ between the true length l of a branch b, and the concatenation estimate $\bar{l}_b$. This calculation was based on the replicates where the species tree used for simulation contained b. $\bar{l}_b$ is the expectation marginalized over all samples containing b. In the case of phylogenetic cherries, only one branch was included, because their lengths are always equal in an ultrametric tree.

All observed differences in branch lengths fell within expectations (Fig. 4). This suggests that the failure to account for neutral coalescent processes, as modeled by the MSC, is responsible for the observed differences.

Calibrating Species Trees Using Fossil Data

Using a fixed molecular clock conditions the estimated divergence times on the accuracy of the *a priori* chosen substitution rate. The rate of molecular evolution is inversely associated with body size in mammals (Bromham 2011) so the substitution rate used for, say, baleen whales would likely be too slow when applied to Muridae. Instead the molecular substitution rate can be inferred jointly with the species tree topology and times by including fossil data and applying an FBD prior to the species tree.

We reran our concatenation and MSC analyses of Caninae after including morphological data with tip dates (fossils), and applied FBD priors to the species trees. The placement of fossil taxa was very uncertain, so to make the FBD results interpretable we pruned the posterior distributions of species trees to include only extant taxa. This also enables direct comparisons of the FBD and fixed clock results. The MCC tree topology inferred by FBD-MSC was identical to fixed clock MSC

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

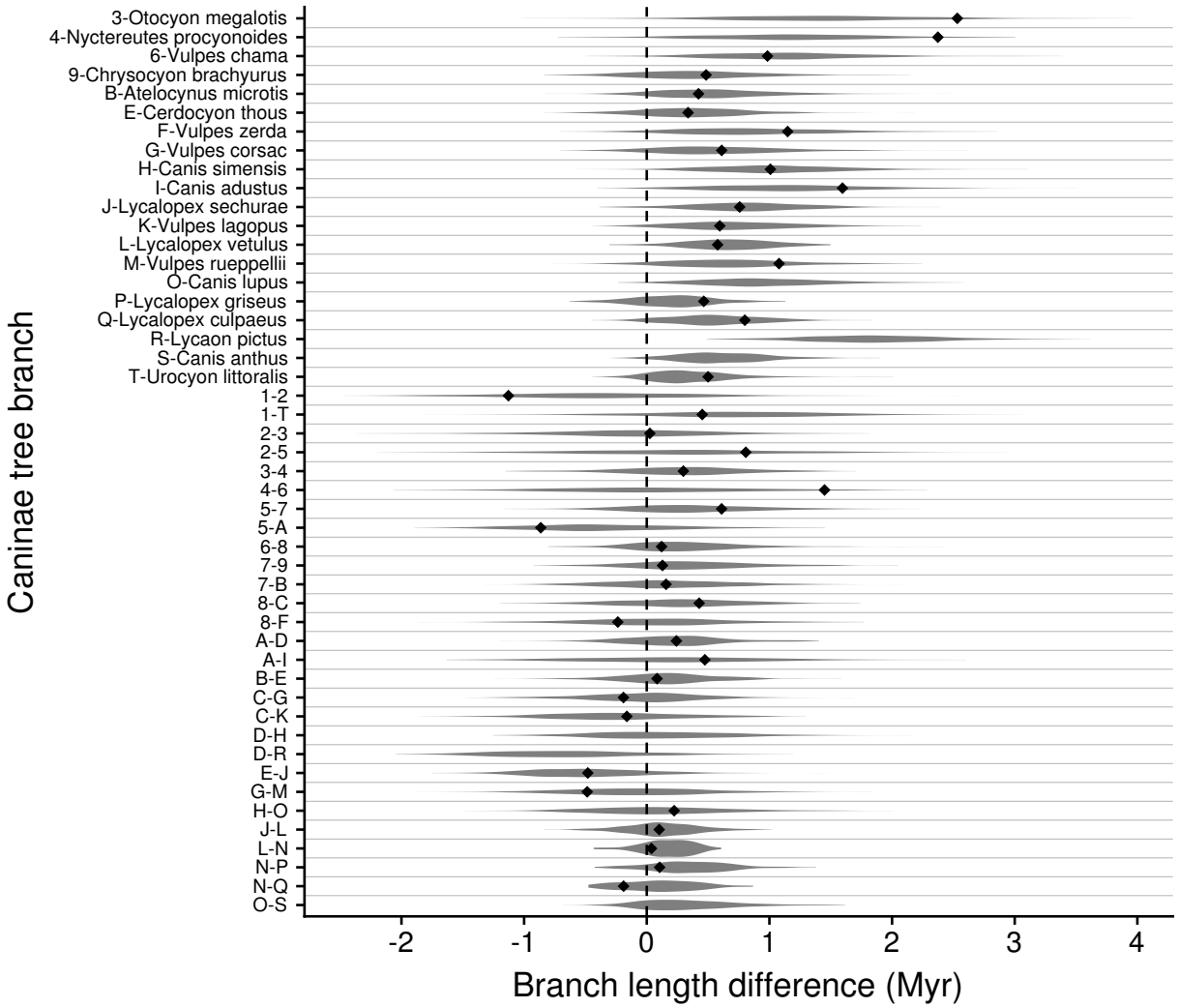

Figure 4: The expected and observed effects of concatenation on branch lengths. Branches correspond to those in Figure 3 and 5. Violin densities represent the distribution of differences between the posterior predictive simulations and corresponding concatenation estimates. Diamonds pinpoint the differences between multispecies coalescent and concatenation posterior mean estimates using molecular and extant morphological data. Diamonds are missing where the support for a branch using concatenation is less than 0.5%. Estimates to the right of zero (dashed) line are longer when using concatenation, estimates to the left are shorter.

(Figs. 3, 5).

The differences in branch lengths observed for FBD-concatenation compared to FBD-MSC were very similar to those seen in the fixed clock scenario (Fig. 6). All branches with longer estimated

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

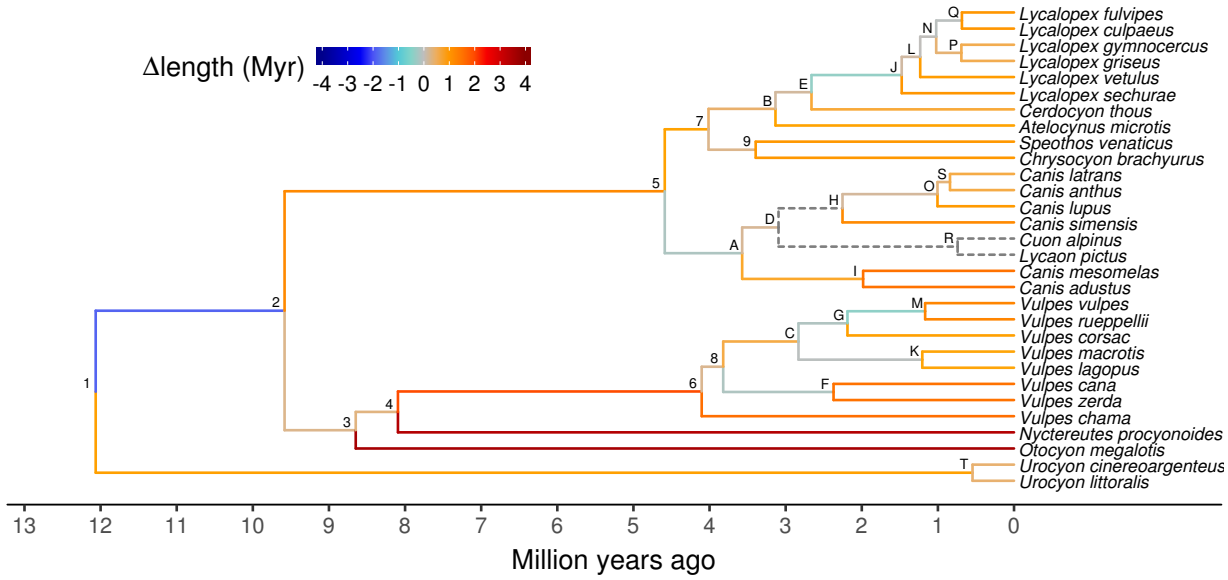

Figure 5: Branch length changes resulting from concatenation using the fossilized birth-death (FBD) model. The color shows how branch length estimates differ when using the concentration rather than the multispecies coalescent (MSC). The tree is a maximum clade credibility (MCC) summary tree with mean node ages, generated from the MSC posterior distribution of species trees, inferred using molecular and morphological data and fossil times. The difference in branch lengths is the mean among concatenation samples including that branch, less the MSC mean. Dashed lines represent branches with less than 0.5% support using concatenation.

lengths using concatenation and a fixed clock compared to MSC and a fixed clock also had longer estimated lengths using FBD-concatenation compared to the corresponding FBD-MSC estimates. The same applied to branches with shorter estimated lengths using concatenation compared to the MSC (Figs. 3, 5).

Similar estimates were made of macroevolutionary parameters using the MSC and concatenation models, as long as the same species tree prior was used (Table 2). When using the FBD prior, the molecular clock rate highest posterior densities (HPDs) included the *a priori* rate of 8×10^{-4} with either the MSC or concatenation. The only non-overlapping HPDs were for the morphological clock rate, which was inferred to be slower when using the FBD compared to a fixed clock. The lower

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

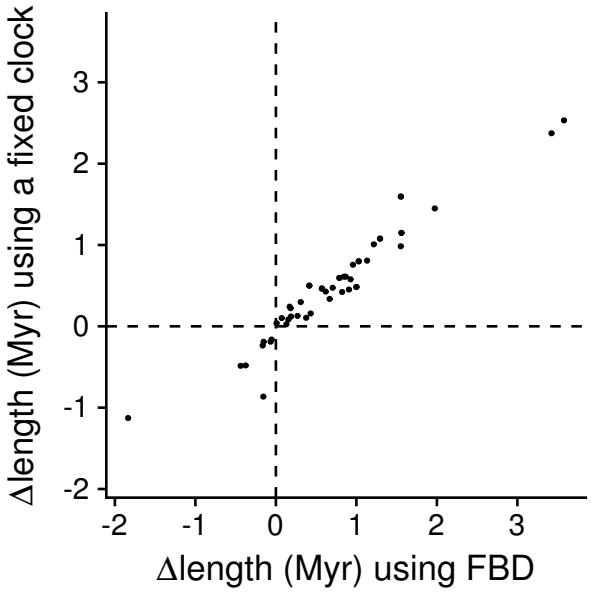

Figure 6: Consistency in how branch lengths change under concatenation rather than the multispecies coalescent (MSC). Differences are when using a fixed clock (y-axis) or when using the fossilized birth-death (FBD) process (x-axis). All branches present in MSC maximum clade credibility trees (Figs. 3, 5) with at least 0.5% support when using concatenation are included. Branch lengths in the top-right quadrant are overestimated by concatenation using a fixed clock or FBD, those in the bottom-left are underestimated by concatenation using either method.

bound for turnover (extinction relative to speciation) was approximately zero when fossil data was not included, but was higher when fossils were explicitly included for FBD analyses.

Clade Ages and Uncertainty

For all clades in the FBD-MSC MCC tree with at least 0.5% support, the divergence time for the root node of that clade according to the FBD-MSC was younger than when estimated using FBD-concatenation (Fig. 7). While the HPD intervals of the two estimates often overlapped substantially, those for the A node (the MRCA of extant sampled *Canis*, *Cuon* and *Lycaon*) and the D node (nested within the A node and excluding *Canis mesomelas* and *C. adustus*) did not, and the FBD-concatenation estimates of those species divergence times were about 2Myr older

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 2: Macroevolutionary parameter estimates.

	MO	FC	FBD
	Multispecies coalescent		
Molecular clock rate $\times 10^{-4}$	8	8	8.24 (6.02–9.78)
Morphological clock rate	NA	0.13 (0.09–0.17)	0.05 (0.04–0.06)
Mean $N_e g$	0.47 (0.37–0.58)	0.50 (0.40–0.62)	0.51 (0.36–0.66)
Diversification rate $\lambda - \mu$	0.23 (0.11–0.35)	0.23 (0.10–0.36)	0.14 (0.02–0.26)
Turnover $\mu \div \lambda$	0.26 (0.00–0.64)	0.29 (0.00–0.66)	0.71 (0.47–0.95)
Sampling proportion $\psi \div (\psi + \mu)$	NA	NA	0.31 (0.14–0.50)
	Concatenation		
Molecular clock rate $\times 10^{-4}$	8	8	7.36 (5.79–8.61)
Morphological clock rate	NA	0.09 (0.07–0.11)	0.05 (0.04–0.06)
Diversification rate $\lambda - \mu$	0.19 (0.10–0.28)	0.19 (0.10–0.29)	0.12 (0.03–0.20)
Turnover $\mu \div \lambda$	0.20 (0.00–0.52)	0.20 (0.00–0.51)	0.63 (0.33–0.91)
Sampling proportion $\psi \div (\psi + \mu)$	NA	NA	0.40 (0.18–0.65)

Parameters were estimated using a fixed clock and molecular only data (MO), a fixed clock with molecular and extant morphological data (FC), or a fossilized birth-death process with molecular, extant morphological and fossil data (FBD).
Values are posterior mean estimates followed in brackets by 95% highest posterior densities.
Clock rates are in units of per-site or per-character per million years. Diversification rate is in per million years.
Mean $N_e g$ refers to the mean of the effective population size N_e distribution, which is scaled by generation time g.

than FBD-MSC.

The Tempo of Caninae Evolution

If species divergence times are always overestimated using concatenation, even when using fossil data and an FBD prior to calibrate the species trees, this is likely to affect macroevolutionary analyses. As an example, we present LTT curves of Caninae diversification estimated using FBD-MSC and FBD-concatenation (Fig. 8).

For both methods the LTT curves are convex, as expected for a birth-death model of evolution with good taxon sampling (Stadler 2008). However the diversification leading to extant Caninae occurs earlier for the FBD-concatenation LTT curve compared to the FBD-MSC curve. The FBD-

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

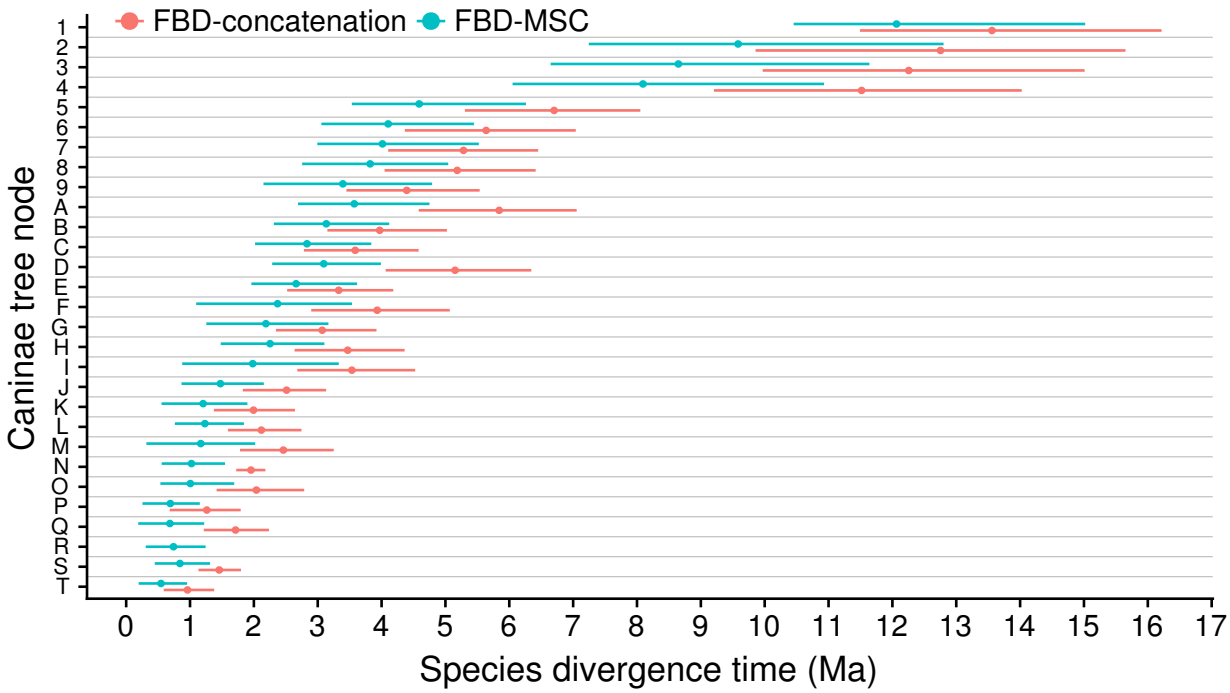

Figure 7: Speciation times estimated by fossilized birth-death with multispecies coalescent (FBD-MSC) and with concatenation (FBD-concatenation) models. Posterior mean FBD-MSC node ages (solid circles) and 95% highest posterior density (HPD) intervals (lines) are estimated from samples where that clade is present. FBD-concatenation ages and intervals are also conditioned on clade presence. Node labels correspond to those in Figure 3 and 5.

concatenation estimate also suggests a diversification slowdown during the last ≈ 2 million years, which is not suggested by the FBD-MSC curve. Diversification slowdown is a predicted spurious effect of concatenation (Ogilvie et al. 2017).

Support for Specific Clades

We considered clade support contradictory between analyses if that clade was highly supported ($> 95\%$) in any analysis and unsupported ($< 5\%$) in any other analysis. Only the R clade met this criterion, which is the clade that unites *Cuon* and *Lycaon* (Table 3).

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

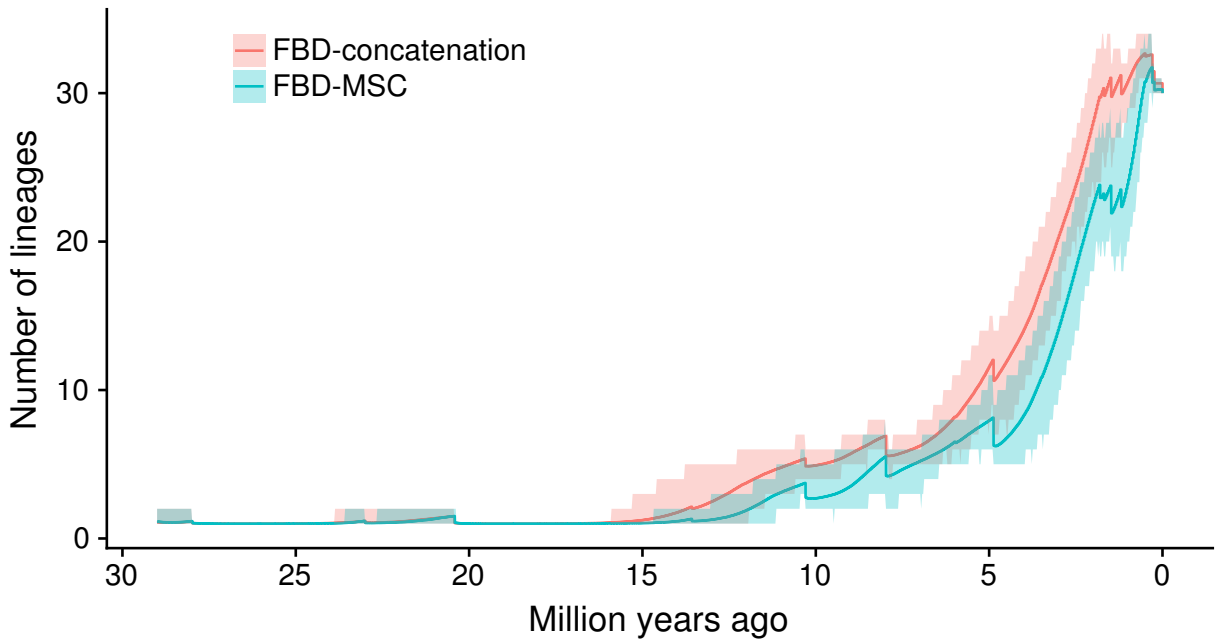

Figure 8: Lineages-through-time (LTT) plot of Caninae diversification. Posterior mean estimates (solid lines) of LTT are calculated for 1,001 evenly spaced time steps spanning 0 to 1, and include extant, fossil and ancestral taxa, and sampled ancestors (which are both fossil and ancestral). 95% highest posterior density (HPD) intervals were also calculated for each step, and are shown as translucent ribbons.

The R clade was highly supported by the MSC regardless of whether a fixed clock or FBD was used to calibrate the species tree. To understand whether this support was driven by coalescent processes alone or by interactions with morphological data, we reran our fixed clock analyses with only molecular data. Without the morphological data there was no support for this clade even when using the MSC, suggesting that unmodeled processes such as selection for convergent morphological evolution might be increasing support for this clade.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table 3: Posterior probabilities of clades.

	MSC			Concatenation		
Clade	MO	FC	FBD	MO	FC	FBD
2	61%	68%	96%	8%	11%	29%
3	77%	82%	94%	69%	77%	92%
4	33%	36%	44%	41%	43%	44%
5	100%	100%	100%	100%	100%	100%
6	100%	100%	100%	100%	100%	100%
7	100%	100%	100%	100%	100%	100%
8	59%	59%	58%	72%	72%	73%
9	100%	100%	100%	100%	100%	100%
A	100%	100%	100%	100%	100%	100%
B	100%	100%	100%	100%	100%	100%
C	100%	100%	100%	100%	100%	100%
D	39%	59%	88%	100%	100%	100%
E	89%	87%	91%	100%	100%	100%
F	100%	100%	100%	100%	100%	100%
G	82%	82%	81%	100%	100%	99%
H	93%	100%	100%	100%	100%	100%
I	100%	100%	100%	100%	100%	100%
J	100%	100%	100%	100%	100%	100%
K	100%	100%	100%	100%	100%	100%
L	50%	47%	51%	100%	100%	100%
M	98%	98%	97%	100%	100%	99%
N	40%	40%	43%	1%	1%	1%
O	100%	100%	100%	100%	100%	100%
P	91%	90%	90%	12%	10%	11%
Q	86%	87%	86%	8%	7%	7%
R	0%	100%	100%	0%	0%	0%
S	16%	73%	46%	0%	0%	1%
T	100%	100%	100%	100%	100%	100%

Clades correspond to node labels used in Figure 3 and 5.
Probabilities were estimated using a fixed clock and molecular only data (MO), a fixed clock with molecular and extant morphological data (FC), or a fossilized birth-death process with molecular, extant morphological and fossil data (FBD).

Discussion

Concatenated Likelihood Methods Are Inaccurate

Several recent studies have demonstrated that methods which use phylogenetic likelihood to estimate species trees from concatenated loci – "concatenated likelihood" for short – are inaccurate

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

under realistic conditions. These studies have been based on simulation and analytical results, and have covered both maximum likelihood (ML) and Bayesian concatenation.

Mendes and Hahn (2016) showed that ML concatenation is systematically biased when estimating the lengths of particular branches on an asymmetric species tree. This is due to substitutions produced by ILS (SPILS), which are artificial substitutions on discordant species tree branches. Mendes and Hahn (2018) went on to show that SPILS is also responsible for the statistical inconsistency of ML concatenation when estimating species tree topologies, even outside of the so-called "anomaly zone" of short branch lengths where the most probable gene tree topology is discordant with the species tree.

Other studies have shown that Bayesian concatenation can be grossly inaccurate when estimating species trees. Bayesian concatenation can overestimate the lengths of tip branches by as much as 350%, and is less accurate than Bayesian MSC using the same number of loci (Ogilvie et al. 2016). Bayesian concatenation is also less accurate at estimating the lengths of internal branches, and reports overly precise credible intervals and support values which can exclude the true values and topologies a majority of the time (Ogilvie et al. 2017).

We have built on previous results by studying the effect of concatenation on an empirical data set of Caninae. Using posterior predictive simulations, we have shown that the observed differences in species tree branch lengths between the MSC and concatenation are expected and caused by a failure to account for coalescent processes. Consistent with previous studies, tip branch lengths were always overestimated, and internal branch lengths were sometimes inaccurate in either direction (Figs. 3, 4).

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

FBD-MSC Results Are More Plausible

Researchers may wonder if the known problems of concatenation are relevant to dated trees inferred using an FBD process. Our study showed that for Caninae, dated species trees inferred using a fixed clock are very similar to dated species trees inferred using an FBD process. We further demonstrated that the differences between MSC and concatenation estimates made under a birth-death process without fossil data are very similar to those made under a FBD process with fossil data (Fig. 6).

Considering coalescent theory and the totality of our results, the FBD-MSC results are more plausible than the FBD-concatenation results. The posterior predictive simulations show that the observed differences in branch lengths between the MSC and concatenation are expected due to a failure to account for coalescent processes.

This has important implications for downstream analyses, as seen in the LTT plots (Fig. 8) where the FBD-concatenation LTT curve suggests a slowdown in Caninae diversification during the past ≈ 2 million years. In contrast, the FBD-MSC LTT curve shows a burst of diversification in the same time frame.

In this study the estimated clock rate of Caninae using the FBD was consistent with the rate inferred by Hugall et al. (2007). Despite this consistency, FBD models are still necessary to account for the correct amount of uncertainty in clock rates, and because the *a priori* clock rate will not always be accurate. If we had studied a different mammalian clade, it would not necessarily have a mean substitution rate consistent with Hugall et al. (2007).

Some unexplored possibilities are that FBD-concatenation would approach FBD-MSC given

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

a morphological matrix covering more taxa and/or when using a relaxed clock. These are hypothetically interesting questions but in practice morphological data sets are usually quite limited in the number of taxa and characters. Concatenation with a relaxed clock is much slower than StarBEAST2 with a strict clock, without any evidence of improved error rates (Ogilvie et al. 2017).

Morphological and Molecular Discordance

We observed that the inclusion or omission of morphological data completely changes the support of the *Lycaon+Cuon* clade from 100% to 0% respectively when using MSC models (Table 3). Support for this clade is ubiquitous in morphological phylogenetic studies of Caninae (Tedford et al. 2009; Prevosti 2010) and probably is due to their specialized dentitions. A previous study of Caninae which combined morphological characters and mitochondrial sequence alignments found that support for this clade came only from the morphological data, and proposed that the responsible characters are likely convergent due to the hypercarnivory of these two species (Zrzavý and Řičánková 2004).

Molecular phylogeneticists should be aware of the potential for morphological model violations when conducting total-evidence studies, and be appropriately cautious when interpreting results. A potential avenue for future research is the development of improved models of morphological evolution, which allow for convergence across many characters at once due to selection. New models could either rule in or out support for *Lycaon+Cuon* by ascribing their similar morphology to convergent evolution. Alternatively, support for this putative clade could be further scrutinized through expanded sampling of fossil representatives of these lineages.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The molecular signal could also be potentially misleading due to unmodeled processes, for example introgression. This could be addressed by integrating the FBD with the multispecies network coalescent, which unlike the MSC does allow for introgression and hybridization (Wen and Nakhleh 2017; Zhang et al. 2017).

Integrative Models Are the Future

The development and implementation of the integrative FBD-MSC model demonstrates how integrative models are made possible within a Bayesian framework. Unlike previous Bayesian implementations of the MSC which are ultrametric and hence limited to contemporary sources of data, using the FBD-MSC we can incorporate morphological and timing information from excavated fossils. The FBD-MSC is a first step, and the future will see further development of integrative models in theory, and the development and use of new implementations in practice.

Funding

This research was funded a Royal Society of New Zealand Marsden award granted to AJD, DW, NJM, TGV and TS (16-UOA-277). HAO was supported by an Australian Laureate Fellowship awarded to Craig Moritz by the Australian Research Council (FL110100104). TS was supported in part by the European Research Council under the Seventh Framework Programme of the European Commission (PhyPD: grant agreement number 335529).

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Acknowledgments

This research was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We thank Craig Moritz for his advice on preparing the manuscript, and the late Colin Groves for his insight into the Caninae fossil record.

References

Bardeleben C., Moore R.L., Wayne R.K. 2005*a*. Isolation and molecular evolution of the selenocysteine tRNA (Cf *TRSP*) and RNase P RNA (Cf *RPPH1*) genes in the dog family, Canidae. Molecular Biology and Evolution 22:347–359.

Bardeleben C., Moore R.L., Wayne R.K. 2005*b*. A molecular phylogeny of the Canidae based on six nuclear loci. Molecular Phylogenetics and Evolution 37:815–831.

Beaulieu J.M., O'Meara B.C., Crane P., Donoghue M.J. 2015. Heterogeneous rates of molecular evolution and diversification could explain the Triassic age estimate for angiosperms. Systematic Biology 64:869–878.

Bouckaert R., Heled J., Kühnert D., Vaughan T., Wu C.H., Xie D., Suchard M.A., Rambaut A., Drummond A.J. 2014. BEAST 2: A software platform for Bayesian evolutionary analysis. PLOS Computational Biology 10:e1003537.

Bromham L. 2011. The genome as a life-history character: why rate of molecular evolution varies

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

between mammal species. Philosophical Transactions of the Royal Society of London B: Biological Sciences 366:2503–2513.

Brunet M., Guy F., Pilbeam D., Mackaye H.T., Likius A., Ahounta D., Beauvilain A., Blondel C., Bocherens H., Boisserie J.R., De Bonis L., Coppens Y., Dejax J., Denys C., Duringer P., Eisenmann V., Fanone G., Fronty P., Geraads D., Lehmann T., Lihoreau F., Louchart A., Mahamat A., Merceron G., Mouchelin G., Otero O., Pelaez Campomanes P., Ponce De Leon M., Rage J.C., Sapanet M., Schuster M., Sudre J., Tassy P., Valentin X., Vignaud P., Viriot L., Zazzo A., Zollikofer C. 2002. A new hominid from the Upper Miocene of Chad, Central Africa. Nature 418:145–151.

Callaway E. 2015. DNA clock proves tough to set. Nature 519:139–140.

Degnan J.H., Rosenberg N.A. 2009. Gene tree discordance, phylogenetic inference and the multispecies coalescent. Trends in Ecology & Evolution 24:332–340.

Donoghue P.C., Benton M.J. 2007. Rocks and clocks: calibrating the tree of life using fossils and molecules. Trends in Ecology & Evolution 22:424–431.

dos Reis M., Donoghue P.C.J., Yang Z. 2014. Neither phylogenomic nor palaeontological data support a Palaeogene origin of placental mammals. Biology Letters 10.

Drummond A.J., Suchard M.A., Xie D., Rambaut A. 2012. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Molecular Biology and Evolution 29:1969–1973.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution 17:368–376.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Gavryushkina A., Heath T.A., Ksepka D.T., Stadler T., Welch D., Drummond A.J. 2017. Bayesian total-evidence dating reveals the recent crown radiation of penguins. Systematic Biology 66:57–73.

Gavryushkina A., Welch D., Stadler T., Drummond A.J. 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. PLOS Computational Biology 10:e1003919.

Hasegawa M., Kishino H., Yano T. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. Journal of Molecular Evolution 22:160–174.

Heath T.A., Huelsenbeck J.P., Stadler T. 2014. The fossilized birth-death process for coherent calibration of divergence-time estimates. Proceedings of the National Academy of Sciences 111:E2957–E2966.

Heled J., Drummond A.J. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27:570–580.

Huelsenbeck J.P., Ronquist F., Nielsen R., Bollback J.P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Hugall A.F., Foster R., Lee M.S.Y., Hedin M. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Systematic Biology 56:543–563.

Höhna S., Landis M.J., Heath T.A., Boussau B., Lartillot N., Moore B.R., Huelsenbeck J.P., Ronquist F. 2016. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. Systematic Biology 65:726–736.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L., Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K., Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X., Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H., Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F., Samaniego J.A., Vargas Velazquez A.M., Alfaro-Núñez A., Campos P.F., Petersen B., Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q., Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K., Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun M., Fjeldsa J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science 346:1320–1331.

Koepfli K.P., Pollinger J., Godinho R., Robinson J., Lea A., Hendricks S., Schweizer R.M., Thalmann O., Silva P., Fan Z. et al. 2015. Genome-wide evidence reveals that African and Eurasian golden jackals are distinct species. Current Biology 25:2158–2165.

Langergraber K.E., Prüfer K., Rowney C., Boesch C., Crockford C., Fawcett K., Inoue E., Inoue-Muruyama M., Mitani J.C., Muller M.N., Robbins M.M., Schubert G., Stoinski T.S., Viola B., Watts D., Wittig R.M., Wrangham R.W., Zuberbühler K., Pääbo S., Vigilant L. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

evolution. Proceedings of the National Academy of Sciences 109:15716–15721.

Lee M.S.Y., Soubrier J., Edgecombe G.D. 2013. Rates of phenotypic and genomic evolution during the Cambrian Explosion. Current Biology 23:1889–1895.

Lewis P.O. 2001. A likelihood approach to estimating phylogeny from discrete morphological character data. Systematic Biology 50:913–925.

Lindblad-Toh K., Wade C.M., Mikkelsen T.S., Karlsson E.K., Jaffe D.B., Kamal M., Clamp M., Chang J.L., Kulbokas E.J., Zody M.C. et al. 2005. Genome sequence, comparative analysis and haplotype structure of the domestic dog. Nature 438:803–819.

Linkem C.W., Minin V.N., Leaché A.D. 2016. Detecting the anomaly zone in species trees and evidence for a misleading signal in higher-level skink phylogeny (Squamata: Scincidae).. Systematic Biology 65:465–477.

Lipson M., Loh P.R., Sankararaman S., Patterson N., Berger B., Reich D. 2015. Calibrating the human mutation rate via ancestral recombination density in diploid genomes. PLOS Genetics 11:e1005550.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. Bioinformatics 24:2542–2543.

Liu L., Xi Z., Wu S., Davis C.C., Edwards S.V. 2015. Estimating phylogenetic trees from genome-scale data. Annals of the New York Academy of Sciences 1360:36–53.

Löytynoja A., Goldman N. 2005. An algorithm for progressive multiple alignment of sequences with insertions. Proceedings of the National Academy of Sciences 102:10557–10562.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Maddison W.P. 1997. Gene trees in species trees. Systematic Biology 46:523–536.

Mendes F.K., Hahn M.W. 2016. Gene tree discordance causes apparent substitution rate variation. Systematic Biology 65:711–721.

Mendes F.K., Hahn M.W. 2018. Why concatenation fails near the anomaly zone. Systematic Biology 67:158–169.

Meyer M., Kircher M., Gansauge M.T., Li H., Racimo F., Mallick S., Schraiber J.G., Jay F., Prüfer K., de Filippo C., Sudmant P.H., Alkan C., Fu Q., Do R., Rohland N., Tandon A., Siebauer M., Green R.E., Bryc K., Briggs A.W., Stenzel U., Dabney J., Shendure J., Kitzman J., Hammer M.F., Shunkov M.V., Derevianko A.P., Patterson N., Andrés A.M., Eichler E.E., Slatkin M., Reich D., Kelso J., Pääbo S. 2012. A high-coverage genome sequence from an archaic Denisovan individual. Science 338:222–226.

Mitchell K.J., Cooper A., Phillips M.J. 2015. Comment on "Whole-genome analyses resolve early branches in the tree of life of modern birds". Science 349:1460.

Nylander J.A., Ronquist F., Huelsenbeck J.P., Nieves-Aldrey J. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology 53:47–67.

Ogilvie H.A., Bouckaert R.R., Drummond A.J. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Molecular Biology and Evolution 34:2101–2114.

Ogilvie H.A., Heled J., Xie D., Drummond A.J. 2016. Computational performance and statistical accuracy of *BEAST and comparisons with other methods. Systematic Biology 65:381–396.

O'Leary M.A., Bloch J.I., Flynn J.J., Gaudin T.J., Giallombardo A., Giannini N.P., Goldberg S.L., Kraatz B.P., Luo Z., Meng J., Ni X., Novacek M.J., Perini F.A., Randall Z.S., Rougier G.W.,

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Sargis E.J., Silcox M.T., Simmons N.B., Spaulding M., Velazco P.M., Weksler M., Wible J.R., Cirranello A.L. 2013. The placental mammal ancestor and the post-K-Pg radiation of placentals. Science 339:662–667.

Patterson N., Richter D.J., Gnerre S., Lander E.S., Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441:1103–1108.

Prevosti F.J. 2010. Phylogeny of the large extinct South American Canids (Mammalia, Carnivora, Canidae) using a "total evidence" approach. Cladistics 26:456–481.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in the Biosciences 13:235–238.

Rannala B., Yang Z. 2017. Efficient Bayesian species tree inference under the multispecies coalescent. Systematic Biology 66:823–842.

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

Ronquist F., Klopfstein S., Vilhelmsen L., Schulmeister S., Murray D.L., Rasnitsyn A.P. 2012. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. Systematic Biology 61:973–999.

Scally A., Durbin R. 2012. Revising the human mutation rate: implications for understanding human evolution. Nature Reviews Genetics 13:745–753.

Scally A., Dutheil J.Y., Hillier L.W., Jordan G.E., Goodhead I., Herrero J., Hobolth A., Lappalainen T., Mailund T., Marques-Bonet T., McCarthy S., Montgomery S.H., Schwalie P.C.,

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tang Y.A., Ward M.C., Xue Y., Yngvadottir B., Alkan C., Andersen L.N., Ayub Q., Ball E.V., Beal K., Bradley B.J., Chen Y., Clee C.M., Fitzgerald S., Graves T.A., Gu Y., Heath P., Heger A., Karakoc E., Kolb-Kokocinski A., Laird G.K., Lunter G., Meader S., Mort M., Mullikin J.C., Munch K., O'Connor T.D., Phillips A.D., Prado-Martinez J., Rogers A.S., Sajjadian S., Schmidt D., Shaw K., Simpson J.T., Stenson P.D., Turner D.J., Vigilant L., Vilella A.J., Whitener W., Zhu B., Cooper D.N., de Jong P., Dermitzakis E.T., Eichler E.E., Flicek P., Goldman N., Mundy N.I., Ning Z., Odom D.T., Ponting C.P., Quail M.A., Ryder O.A., Searle S.M., Warren W.C., Wilson R.K., Schierup M.H., Rogers J., Tyler-Smith C., Durbin R. 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483:169–175.

Slater G.J. 2015. Iterative adaptive radiations of fossil canids show no evidence for diversity-dependent trait evolution. Proceedings of the National Academy of Sciences 112:4897–4902.

Stadler T. 2008. Lineages-through-time plots of neutral models for speciation. Mathematical Biosciences 216:163–171.

Stadler T. 2010. Sampling-through-time in birth-death trees. Journal of Theoretical Biology 267:396–404.

Stadler T., Kühnert D., Bonhoeffer S., Drummond A.J. 2013. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proceedings of the National Academy of Sciences 110:228–233.

Sukumaran J., Holder M.T. 2010. DendroPy: a Python library for phylogenetic computing. Bioinformatics 26:1569–1571.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Tedford R.H., Wang X., Taylor B.E. 2009. Phylogenetic systematics of the North American fossil Caninae (Carnivora: Canidae). Bulletin of the American Museum of Natural History 325.

Wang X. 1994. Phylogenetic systematics of the Hesperocyoninae (Carnivora, Canidae). Bulletin of the American Museum of Natural History 221.

Wang X., Tedford R.H., Taylor B.E. 1999. Phylogenetic systematics of the Borophaginae (Carnivora, Canidae). Bulletin of the American Museum of Natural History 243.

Wen D., Nakhleh L. 2017. Coestimating reticulate phylogenies and gene trees from multilocus sequence data. Systematic Biology. Advance article.

White T.D., Asfaw B., Beyene Y., Haile-Selassie Y., Lovejoy C.O., Suwa G., WoldeGabriel G. 2009. *Ardipithecus ramidus* and the paleobiology of early hominids. Science 326:75–86.

White T.D., Lovejoy C.O., Asfaw B., Carlson J.P., Suwa G. 2015. Neither chimpanzee nor human, *Ardipithecus* reveals the surprising ancestry of both. Proceedings of the National Academy of Sciences 112:4877–4884.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. 2nd ed. New York: Springer-Verlag.

Wood B., Harrison T. 2011. The evolutionary context of the first hominins. Nature 470:347–352.

Yang Z. 2015. The BPP program for species tree estimation and species delimitation. Current Zoology 61:854–865.

Yu G., Smith D.K., Zhu H., Guan Y., Lam T.T. 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution 8:28–36.

bioRxiv preprint doi: https://doi.org/10.1101/242875; this version posted January 7, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Zhang C., Ogilvie H.A., Drummond A.J., Stadler T. 2017. Bayesian inference of species networks from multilocus sequence data. Molecular Biology and Evolution. Advance article.

Zhang C., Stadler T., Klopfstein S., Heath T.A., Ronquist F. 2016. Total-evidence dating under the fossilized birth-death process. Systematic Biology 65:228–249.

Zrzavý J., Řičánková V. 2004. Phylogeny of recent Canidae (Mammalia, Carnivora): relative reliability and utility of morphological and molecular datasets. Zoologica Scripta 33:311–333.