A BAYESIAN ANALYSIS OF THE GENETIC EVOLUTION IN LIVING FOSSILS

by

Kevin Surya

A thesis submitted in partial fulfillment
of the requirements for the degree

of

Master of Science

in

Statistics

MONTANA STATE UNIVERSITY
Bozeman, Montana

December 2023

DEDICATION

This thesis is for all the friends I have made in my journey. Memories of our mini arcs are reminders that somewhere in the vast darkness, there will always be an iris to marvel at.

No more sappy lines allowed. Yes, I will treat you to lunch or dinner somewhere sometime after I finish this thesis (probably not all of you, but you get the point).

ACKNOWLEDGEMENTS

I thank my Masters-en-Route committee members, Chris Organ, John Borkowski, and Andrew Hoegh, for their never-ending patience and constructive comments. I received financial support from the Department of Mathematical Sciences and the Molecular Biosciences Fellowship Program.

TABLE OF CONTENTS

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

LIST OF FIGURES CONTINUED

Figure Page

GLOSSARY

Alignment (or multiple sequence alignment): an alignment of DNA or amino acids sequences from multiple species. It is a matrix with species as rows and DNA or amino acids sites (locations) as columns. “Letters” within a column are homologous or comparable across species. Phylogenetic trees are most often inferred from an alignment.

Amino acids (typically abbreviated as AA): “letters” (A, R, N, …) that, when stringed together, form proteins. An amino acids letter maps to particular combinations of DNA letter triplets in a coding sequence (e.g., AGC corresponds with the amino acid letter S). However, the map between all possible DNA triplets to amino acids is not one-to-one, as an amino acids letter can correspond with up to six triplets (e.g., TCT, TCC, TCA, TCG, AGT, and AGC map to S). Note this definition oversimplifies what happens during transcription (DNA to RNA) and translation (RNA to proteins), where RNA is an intermediate stage.

Amino acids sequence (or protein sequence): a sequence of amino acids “letters.” Different sequences create different proteins, the building blocks of life.

Clade: a set of species that includes a common ancestor and all its descendants. For example, a set of birds and bats is not a clade because it excludes non-bird reptiles and non-bat mammals that share a common ancestor with birds and bats. A clade can be a single species. Also, we often use the term ‘clade’, ‘lineage’, and ‘taxonomic group’ interchangeably.

Coding sequence (or protein-coding sequence): a sequence of DNA “letters” (A, C, T, and G) that directly “make” proteins (i.e., transcribed and translated into proteins), the building blocks of life. In this study, the coding sequences are those in the nuclear genome (inside the nucleus), not those in the mitogenome (inside the mitochondria).

Gene family tree: a phylogenetic tree where the tips or terminal edges are not species but genes. The tree represents the evolutionary history of gene samples, where present-day genes descended from ancestral genes through speciation and gene duplication events. An example of a gene family tree is a tree of the globin superfamily, which includes genes encoding myoglobin, hemoglobin, etc.

Genome: a complete set of an organism’s genetic material. We often use the adjective ‘genomic’ not to indicate the totality of DNA sequences but to emphasize that our data cover more than a few genes.

Node calibration: a constraint on the age estimate of a node in a timetree (e.g., age of the tree root ~ Uniform(66, 70) means that the age estimate must fall between 66 and 70 million years ago). Information for node calibrations typically come from fossil specimens (e.g., the estimated age of the oldest known fossil mammal). But, they can come from other sources as well, such as the timing of plate tectonics events.

GLOSSARY CONTINUED

Paralogy: a situation where the evolutionary history of a sample of genes involves a gene duplication event(s). Consequently, the lineage-branching pattern of a particular gene sampled from multiple species may not match the branching pattern of the species (e.g., some human genes are more closely related with genes in gorillas than those in chimps). This situation can be avoided by only comparing genes of the same version (e.g., α-globins) and not mixing different versions (e.g., α- and β-globins) in the same sample. A gene that has no duplicates is a single-copy gene.

Phylogenetic tree (or phylogeny): akin to a family tree that records relatedness structure and information about evolutionary history.

Site: a specific location in a DNA or amino acid sequence. In the context of an alignment, a site is a column.

Speciation: a lineage-splitting event. A small subpopulation may split from a larger ancestral population (e.g., dispersal into a new environment), creating a new species. Or an ancestral species may branch into two descendant species through isolation by environmental barriers (e.g., the emergence of a mountain range).

Supermatrix: a concatenation (column bind) of multiple sequence alignments, where each alignment, for most cases, represents a gene.

Substitution: a replacement of a DNA or amino acid letter with a different letter (e.g., An A in a DNA sequence is replaced with a T).

Taxon (plural: taxa): a taxonomic unit, which can be a species or a genus, for example. This term helps biologists refer to a set of biological entities of different taxonomic ranks (e.g., a set that includes genus, species, and subspecies) and paleontologists to refer to extinct entitities without assigning a taxonomic rank.

Vertebrates: animals with a backbone, which include fishes, amphibians, reptiles, birds, and mammals.

ABSTRACT

Most animal species appear radically different from their ancestors that lived hundreds of millions of years ago, yet a few are remarkably similar to their ancient ancestors. Charles Darwin called these ancient-looking creatures living fossils. Genomes of several living fossils, such as the African coelacanth (*Latimeria chalumnae*) and reptile tuatara (*Sphenodon punctatus*), were sequenced recently, revealing the low average evolutionary rates of their protein-coding sequences. Such slow genomic and anatomical evolution (i.e., stasis) are poorly understood, especially in the era of big data. For example, how can species evolve slowly despite continual environmental change? Here, we used Bayesian models to (1) quantify how much slower is the genomic (coding sequence) evolution in living fossils than in other species and (2) explain why from a macroevolutionary standpoint. Among vertebrates, we found that, on average, the coding sequence has changed less in living fossil species but only by negligible amounts (-0.02 ± 0.03 DNA substitutions/site; -0.01 ± 0.02 amino acids substitutions/site). However, this result is likely due to how researchers classify living fossils. When we focus on the widely recognized living fossils (coelacanth and tuatara), we found that their genomic sequences have evolved less than predicted based on the other species (Bayesian p-values < 0.5). A caveat is that this prediction is sensitive to species sampling. The predictive distributions capture their slow evolution slightly better (increased Bayesian p-values) when the model includes node count, the number of net speciation or branching events along a lineage. The lack of net speciation events in living fossil lineages partly corresponds with their slow genomic evolution, consistent with the theory of punctuated evolution, which posits a coupling between speciation and evolutionary change. Population- and ecological-level processes also dictate rates of evolution, so including them in future models would be crucial. Elucidating speciation as a driver of the pace of living fossils' genomic evolution sheds light on a central problem in evolutionary theory and is essential for conserving these species. An evolutionary history marked by minimal change and a lack of diversity could be costly in today's world of unprecedented climate change.

CHAPTER ONE

QUANTIFYING THE SLOW GENETIC EVOLUTION IN LIVING FOSSILS

Introduction

The sequencing of living fossil genomes has progressed rapidly over the past decade[1–8]. So far, the motivation has been to unravel the evolutionary origins of their more diverse sister clades (next-of-kin taxonomic groups). For example, the comparison between the African coelacanth (*L. chalumnae*) genome and its sister clade (comprising amphibians, birds, mammals, and reptiles) helps identify genomic changes associated with the animal's water-to-land transition circa 400 million years ago[1].

These studies above show that the DNA or amino acids substitution rate averaged across all the protein-coding regions is low for all sequenced living fossils. Yet, these individual studies did not investigate, in detail, the mechanism that led to molecular stasis. My contribution is to study these genomes comparatively under one umbrella. It is also unclear how the low evolutionary rates are distributed across gene families and other genomic aspects such as the non-coding sequences, karyotype (chromosome set), and genome size[9]; hence the need to analyze the genomes part by part. In this chapter, we quantify, using Bayesian "*t*-tests" and posterior predictive distributions, the extent to which vertebrate living fossils evolve more slowly in one genomic partition: the coding sequences—the series of "letters" that make proteins, the building blocks of life.

Living Fossil Species

Here, we classify a vertebrate species as a living fossil a priori based on four linchpin publications[10–13]. In *On the Origin of Species* (1859), Darwin[10] coined the term "living fossils" to capture the anatomical and environmental conservativeness of the platypus (*Ornithorhynchus anatinus*) and South American lungfish (*Lepidosiren paradoxa*). He wrote:

> "... All fresh-water basins, taken together, make a small area compared with that of the sea or of the land; and, consequently, the competition between fresh-water productions will have been less severe than elsewhere; new forms will have been more slowly formed, and old forms more slowly exterminated. And it is in fresh water that we find seven genera of Ganoid fishes, remnants of a once preponderant order: and in fresh water we find some of the most anomalous forms now known in the world, as the *Ornithorhynchus* and *Lepidosiren*, which, like fossils, connect to a certain extent orders now widely separated in the natural scale. These anomalous forms may almost be called living fossils; they have endured to the present day, from having inhabited a confined area, and from having thus been exposed to less severe competition."

Since then, paleontologists have recognized more living fossils. In his 1944 classic, *The Tempo and Mode in Evolution*, George Gaylord Simpson[11] dedicated an entire chapter to cases of exceedingly low rates of evolution, which he called bradytely. These cases range from the marsupial opossums (since the Late Cretaceous, 100-66 million years ago) to the mollusk *Lingula* (now known to have changed little since the Cambrian[14], more than 500 million years ago).

Later, in his 1979 book *Macroevolution, Pattern and Process*, Steven Stanley included a table of 17 clades with relatively continuous fossil records that have changed little for at least the past 15 million years. Stanley also mentioned or discussed the, by that time, flagship living fossils such as the coelacanth, tuatara, and lungfish[15–17].

In 1984, Niles Eldredge and Stanley[13] published the book *Living Fossils*, where each chapter, written by distinct, invited paleontologists, weighs the evidence for and against a taxon's

living fossil label. The book has 18 chapters covering vertebrate taxa. Most consist of detailed morphological arguments, except for Joel Cracraft's bird chapter. Cracraft acknowledged the challenge of identifying living fossils among birds because of the poor fossil record and taxonomic uncertainties. So, he cast a wide net, postulating as living fossils, bird taxa with at least an identifiable fossil specimen dating far back in the rock record (i.e., assumed stratigraphic or fossil record longevity).

We compiled a list of 205 vertebrate living fossil species from these four sources (see supplemental files). We exclude species without a published whole-genome sequence, transcriptome, or, at the very least, mitogenome. For bird taxa, I only select those with whole-genome data to keep the list concise. This list is not exhaustive; the goal is a reasonably representative list.

Researchers have constructed metrics to capture what the public qualitatively regard as living fossils. In 2011, Lionel Cavin and Anne Kemp[18] developed the metric Evolutionary Distinctiveness (ED) based on the observation that living fossil species tend to have persisted for a long time[11]. Living fossil species should have high ED values. However, the ED value is sensitive to sampling bias because it depends on the species in the dataset and the sample's taxonomic scale. We do not use the ED metric to classify species as living fossils a priori since there is no objective cutoff. At the very least, Cavin and Kemp[18] demonstrate that time is a characteristic of living fossils.

Bennett et al.[19] developed the Evolutionary Performance Index (EPI) and a proxy for EPI (pEPI). Their motivation was to clarify the core features of the living fossil label, which researchers have interpreted differently since Darwin (for example, see Simpson[11], Casane and

Laurenti[20], and Grandcolas et al.[21]). The EPI is based on success (number of extant species), time, and change (ecological and morphological) since the origin of a lineage. The pEPI is a less general metric that does not include the 'change' component should data be unavailable. Still, the EPI and pEPI are sensitive to sampling bias and taxonomic scale. For instance, sharks have high pEPI scores (not living-fossil-like) because they "outperform" their sisters, the chimaeras (holocephalians). But, together, these cartilaginous fishes (chrondrichthyans) have a low pEPI score.

An a priori specification minimizes confirmation bias. One drawback is that we may miss species whose genomes have evolved slowly but are not historically known as living fossils. Therefore, we explore a second approach where we classify the top 10% of the slowest-evolving species as living fossils (see below). This approach is akin to unsupervised learning. Simpson[11] postulated three distributions of rates of evolution: bradytely (low rate), horotely (average rate), and tachytely (high rate). He noted that it might be difficult to distinguish whether a lineage falls into the bradytely distribution or the low end of the horotely distribution. We expect the arbitrary 10% cutoff is sufficiently liberal to capture the so-called low end of the horotely distribution.

Phylogenetic Comparative Methods

Before proceeding further, we need to emphasize the non-independence between species. In 1985, Joseph Felsenstein[22] wrote a seminal paper demonstrating how non-independence due to shared ancestry could mislead comparative studies. A simple example is that the average human body mass is more similar to that of its closest living relative, the chimpanzee, than a more distant relative, such as a mouse. For any collection of species, there is a relatedness structure resembling a family tree called a phylogenetic tree. A phylogenetic tree is impractical to

build analytically due to the massive search space. Inferring a tree of present-day species requires molecular sequence data: DNA or amino acids sequences. Therefore, due to shared ancestry, the degrees of freedom are always less than if the species were independent. Felsenstein accounts for the non-independence using independent contrasts[22]. In 1989, Alan Grafen[23] developed a generalized least squares (GLS) method, later dubbed the phylogenetic generalized least squares (PGLS; see Symonds and Blomberg[24] for an introductory walkthrough). Grafen[23] converted the phylogenetic tree into a variance-covariance matrix. Here, we use PGLS. A generalized linear mixed model (GLMM) approach exists[25]. But, we will not use it in this study.

A trait of interest, for example, average body mass, might not have evolved following the species' evolutionary history down to the tee. In other words, the variance-covariance matrix derived from a tree may not be the best to represent the correlation among species. In 2002, Robert Freckleton, Paul Harvey, and Mark Pagel[26] introduced a parameter, λ, that scales the off-diagonals of the matrix, essentially "stretching" the tree's internal branches to fit the trait data better. A $\lambda = 0$ means that the trait evolved independently of the species' evolutionary history (off-diagonals go to zero), and a $\lambda = 1$ indicates no scaling is necessary as the tree already explains trait data well (off-diagonals multiplied by one). In this report, since all variables come from the trees themselves, we always fix $\lambda = 1$.

Phylogenetic trees, however, are not only for accounting non-independence. Integrating trees with various statistical methods can reveal the patterns and processes of deep-time evolution[27–32]. The set of tools centered around phylogenies is called the phylogenetic comparative methods (PCMs). Evolutionary virologists use PCMs-like methods, but they call it phylodynamics[33]. This work falls under PCMs.

Data

Phylogenetic Trees

The data we use are from previously inferred phylogenetic trees. Ideally, we would account for uncertainties in the trees. But, for simplicity, we treat the trees we use here as fixed observations.

Irisarri et al. (2017) We downloaded two vertebrate phylogenetic trees, where the branch lengths represent the amount of molecular evolution[34,35]. Irisarri et al.'s[34] tree comprises 100 jawed vertebrate species (Figures 1-2). The concatenated multiple sequence alignment used to build this tree consists of 1,964,439 nuclear amino acids (AA) sites, spanning 4,593 genes. This alignment is free from deep paralogs. Therefore, this tree's branch lengths reflect the evolution in most protein-coding regions comparable across jawed vertebrates. The tree is a majority rule consensus from 100 Bayesian inference analyses of 100 gene bootstrap replicates (alignments with ~50,000 AA sites each), inferred using PhyloBayes[36] with the CAT+G substitution rate model.

Once we downloaded the tree file, I converted it to the NEXUS format using BayesTreesConverter v1.3 (https://www.evolution.reading.ac.uk/BayesTrees.html). All tip labels were renamed to complete species names. Some names are different between the tree file and the manuscript figure. The reason is unclear, but these discrepancies are unimportant in the scale of jawed vertebrates. *Basiliscus vittatus*, *Sceloporus occidentalis*, *Sternotherus carinatus*, and *Hymenochirus boettgeri* were named *Basiliscus plumifrons*, *Sceloporus undulatus*, *Sternotherus odoratus*, and *Hymenochirus curtipes* in Figure 1 of Irisarri et al.[34]. We use the species names in their tree file.

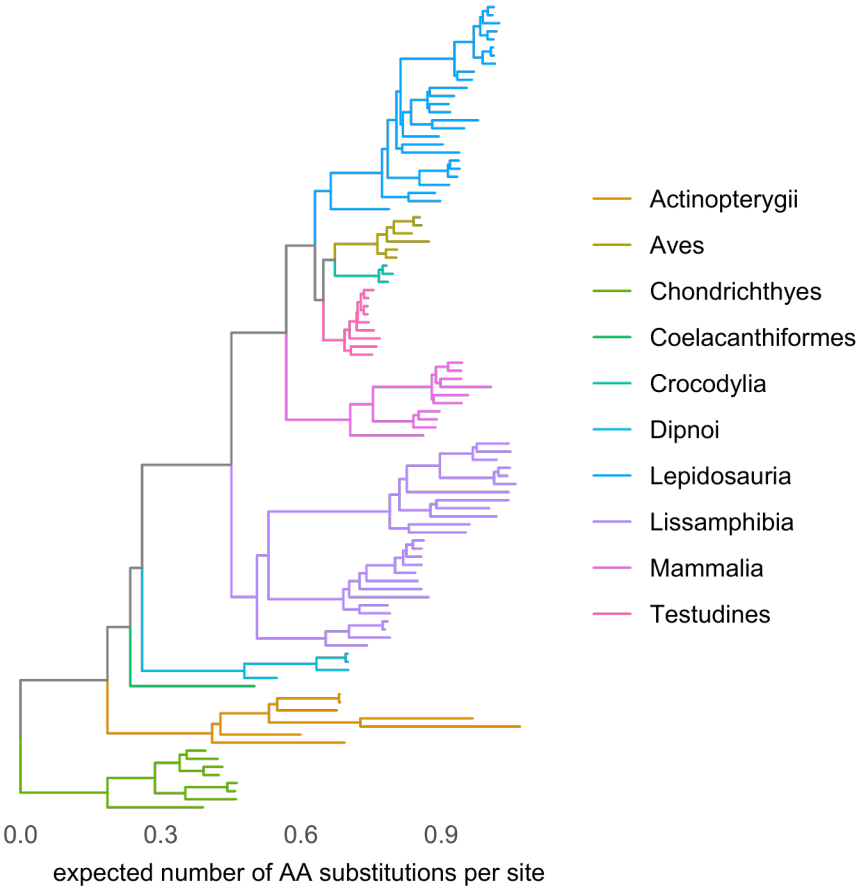

Figure 1. Irisarri et al.'s[34] molecular tree comprises 100 jawed vertebrate species. Colors represent clades. The branch length unit is the expected number of amino acids (AA) substitutions per site. Note the uneven taxonomic sampling. Birds and ray-finned fishes (Aves and Actinopterygii) are vastly undersampled relative to reptiles and amphibians (Lepidosauria and Lissamphibia). The ratios of sampled to known living species across clades, per the International Union for Conservation of Nature[37], are as follows. Chondrichthyes: 8/1,234 (0.65%); Actinopterygii: 7/23,995 (0.03%); Coelacanthiformes: 1/2 (50.00%); Dipnoi: 4/6 (66.67%); Lissamphibia: 26/7,486 (0.35%); Mammalia: 10/5,973 (0.18%); Testudines: 9/260 (3.46%); Crocodylia: 3/28 (10.71%); Aves: 6/11,188 (0.05%); Lepidosauria: 26/9,954 (0.26%).

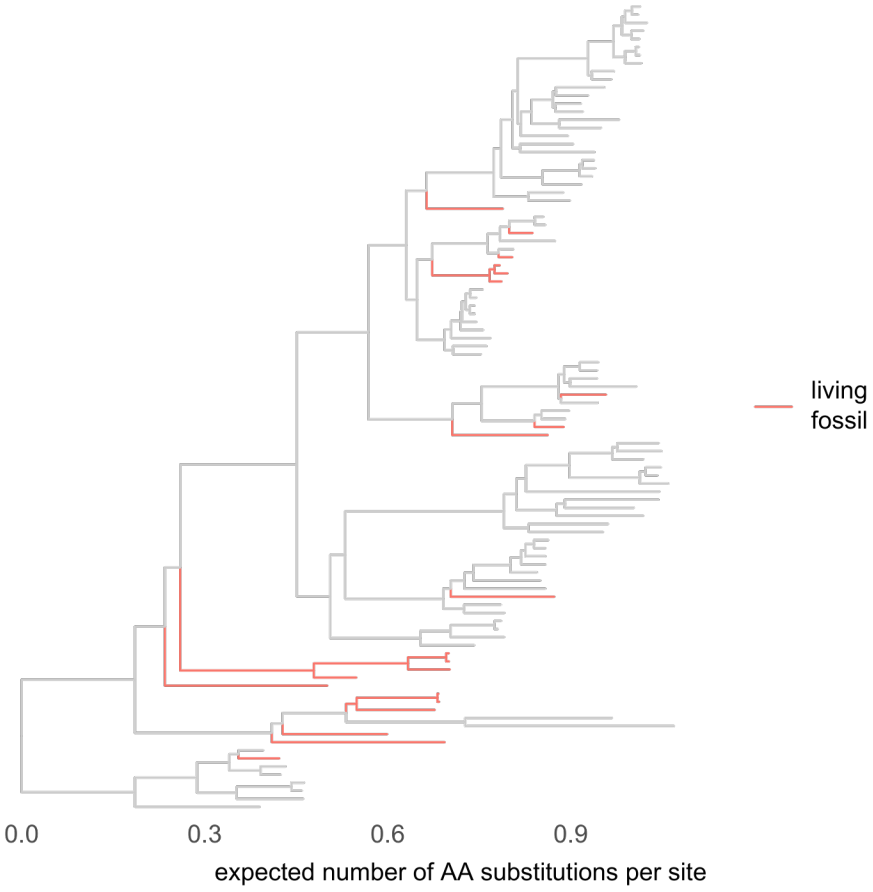

Figure 2. Of the 100 vertebrate species in Irisarri et al.'s[34] molecular tree, 21 are recognized living fossils, per the four publications[10–13] (red branches).

Irisarri et al.[34] inferred a time-calibrated tree, where the branch length unit is million of years (Figures 3-4). Its topology (relatedness structure) is the same as the molecular tree above. This timetree is a Bayesian majority-rule consensus, where the posterior trees were estimated using 16 node calibrations and an autocorrelated lognormal clock model. Whereas a molecular tree records the amount of evolution, a timetree contains the timing of speciations or lineage-splitting events. We expect living fossil taxa or lineages to have long time-calibrated branches since they have persisted for so long yet have not speciated as much as their contemporaries[11]. Time-related variables may be a covariate we need.

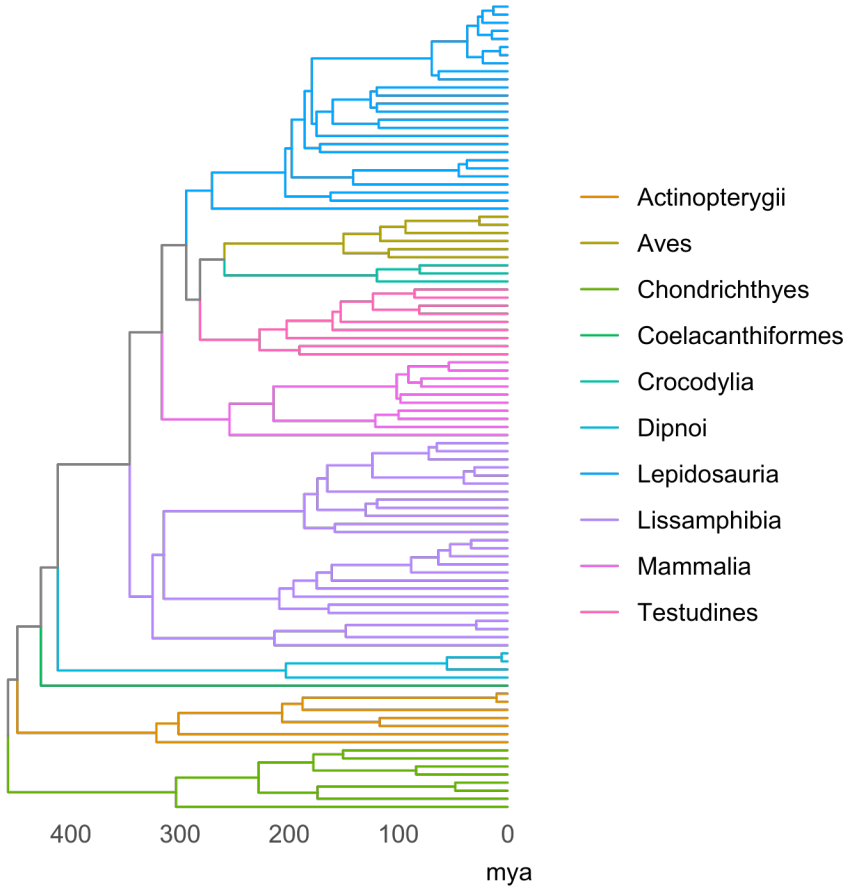

Figure 3. Irisarri et al.'s[34] timetree contains the same 100 jawed vertebrate species and relatedness structure as the tree in Figure 1. But, the branch length unit is millions of years. The x-axis scale is in the unit millions of years ago (mya).

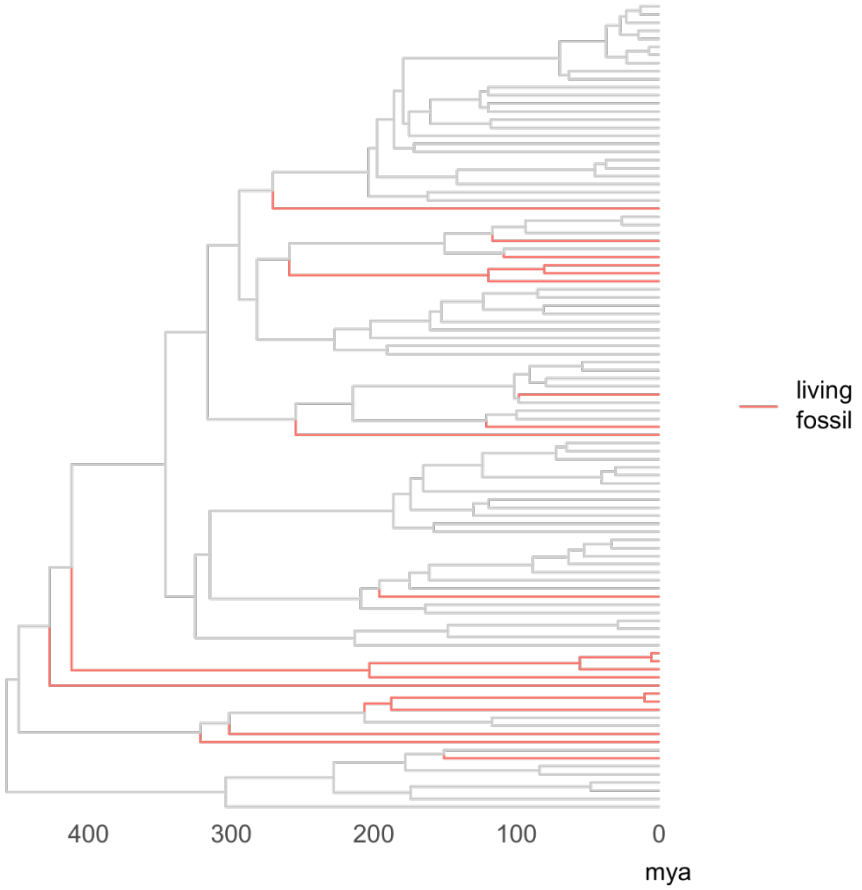

Figure 4. Of the 100 jawed vertebrate species in Irisarri et al.'s (2017) timetree, 21 are recognized living fossils, per the four publications[10–13] (red branches).

Morel et al. (2022) We downloaded a second molecular tree from Morel et al.[35], which comprises 188 vertebrate species (Figures 5-6). They inferred their molecular tree using a different approach. Previously, Irisarri et al.'s[34] general approach was to concatenate 4,593 sequence alignments (one per gene) into one supermatrix and then infer a tree using this matrix.

Morel et al.[35] developed the program SpeciesRax and used it to infer a vertebrate species tree with 31,612 gene family trees (GFTs) as inputs. The GFTs were acquired from Ensembl Compara[38]. The total number of gene copies included is 3,725,332. SpeciesRax accounts for gene duplication, loss, and transfer. They built the GFTs using the ParGenes[39] pipeline, where

each of the 31,162 GFT alignments went through a RAxML-NG[40] maximum likelihood tree search under the GTR+G4 model with one random starting tree. Then, they used their novel method, MiniNJ[35], to generate a starting species tree. MiniNJ accounts for paralogy. This starting tree is then modified using a hill-climbing algorithm to maximize the reconciliation likelihood[35], which is the probability of observing a set of GFTs given a a rooted species tree and parameters on gene duplication, loss, and transfer. SpeciesRax estimates each species' tree branch length independently, averaging over the branch lengths of the reconciled GFTs.

The sole chondrichthyan species in this tree, the Australian ghostshark (*Callorhinchus milii*), is placed as the sister taxon of Sarcopterygii. This topology conflicts with the traditional arrangement of Chondrichthyes diverging before the split of Actinopterygii and Sarcopterygii[34,41]. In their paper, Morel et al.[35] mentioned this issue and four other topological disagreements between their vertebrate species tree and the NCBI taxonomy[42]. This erroneous placement may be due to having only one chondrichthyan species in the tree.

The Irisarri et al.'s[34] and Morel et al.'s[35] trees differ in the relatedness structure, branch length units (amino acids vs. DNA substitution rates), and the species sampled. The DNA substitution rates have higher magnitudes than those for amino acids because there are three times as many sites for DNA and Morel et al.[35] used more genes to build the tree.

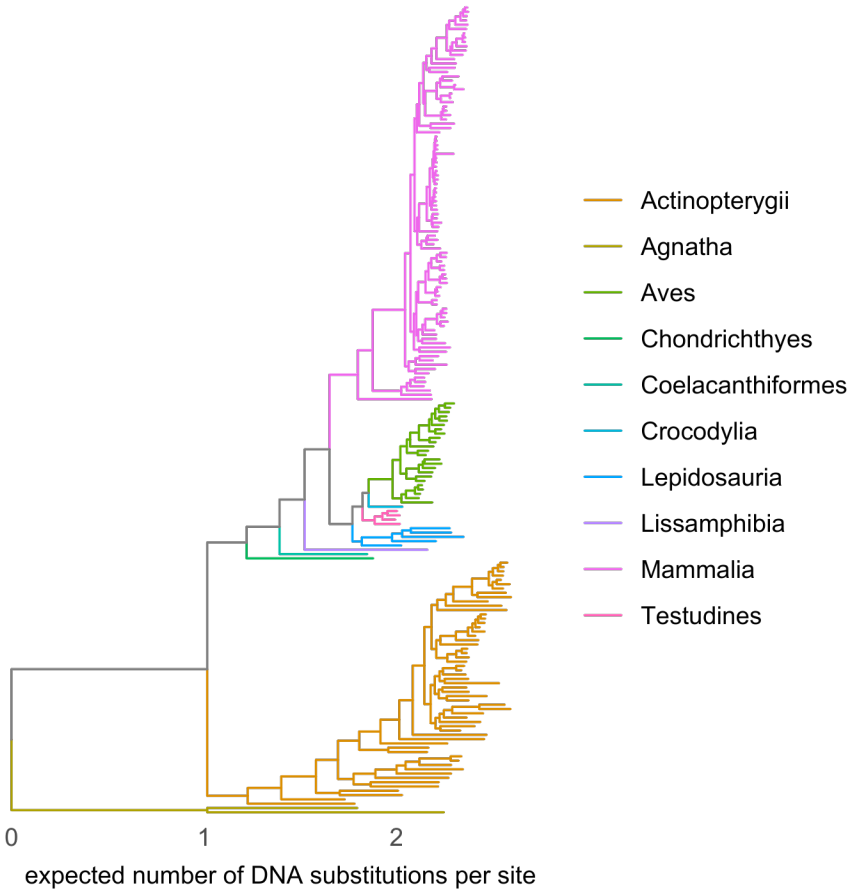

Figure 5. Morel et al.'s[35] molecular tree comprises 188 vertebrate species. Colors represent clades. The branch length unit is the expected number of DNA substitutions per site. Chondrichthyes, Lissamphibia, and Lepidosauria are underrepresented in this tree. The ratios of sampled to known living species across clades, per the International Union for Conservation of Nature[37], are as follows. Agnatha: 2/114 (1.75%); Chondrichthyes: 1/1,234 (0.08%); Actinopterygii: 57/23,995 (0.24%); Coelacanthiformes: 1/2 (50.00%); Lissamphibia: 1/7,486 (0.01%); Lepidosauria: 5/9,954 (0.05%); Testudines: 4/260 (1.54%); Crocodylia: 1/28 (3.57%); Aves: 24/11,188 (0.21%); Mammalia: 92/5,973 (1.54%).

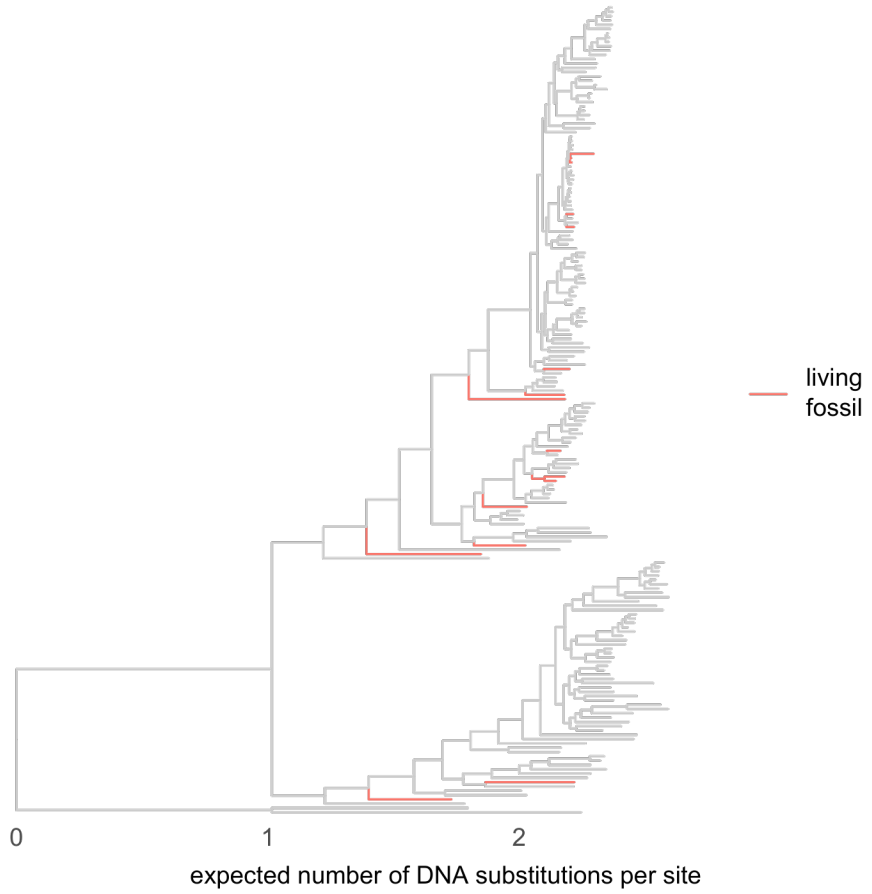

Figure 6. Of the 188 vertebrate species in Morel et al.'s[35] molecular tree, 16 (~9%) are recognized living fossils, per the four publications[10–13] (red branches).

Unlike Irisarri et al.[34], Morel et al.[35] did not infer a timetree. We had to make one. Since Morel et al.[35] used gene family trees as inputs instead of a supermatrix, we could not use popular tree dating programs like BEAST2[43], MrBayes[44], and treePL[45]. So, we dated Morel et al.'s[35] tree using the penalized likelihood algorithm in the R package ape v5.7.1[46,47]. Paradis[47] generalized Sanderson's[46] original penalized likelihood method, allowing the evolutionary rates to vary across the tree branches in an uncorrelated manner (i.e., uncorrelated relaxed clock). We fixed the rate-smoothing parameter to one, the default value. Ideally, as Sanderson[46] did, this parameter is optimized through a leave-one-out-cross-validation. However, Paradis's[47] method

does not have this functionality. To rescale molecular branch lengths into the absolute time unit, we needed fossil calibrations at key internal nodes in the tree. An internal node is the most recent common ancestor (MRCA) of its descendant nodes. Using previously vetted calibration points, we constrained 16 node ages. Irisarri et al.[34] list 30 calibration points, and we selected 15 applicable ones. The root calibration (Vertebrata) is from Benton et al.[48]. Whenever Irisarri et al.[34] only provide a minimum age for a node, we added the maximum age from Benton et al.[48].

Table 1. The 16 node calibrations, taken from Irisarri et al.[34], place lower and upper bounds on the age estimates of the 16 internal nodes in Morel et al.'s[35] tree. The unit of the calibrations is millions of years ago (mya).

Node	Minimum age (mya)	Maximum age (mya)
Vertebrata	457.5	636.1
Actinopterygii vs. Sarcopterygii	421.8	462.5
Lissamphibia vs. Amniota	330.4	350.1
Diapsida vs. Synapsida	288.0	338.0
Lepidosauria vs. Archosauria	252.0	257.0
Crocodylia vs. Aves	243.0	251.0
Gallus vs. *Taeniopygia*	66.0	86.5
Sphenodon vs. Squamata	238.0	252.7
Iguania	165.0	230.0
Theria vs. Monotremata	162.5	191.4
Placenta vs. Marsupialia	124.6	138.4
Boreoeutheria vs. Xenartha	95.3	113.0
Canis vs. *Felis*	42.8	63.8
Homo vs. *Mus*	61.5	100.5
Ameridelphia vs. Australidelphia	61.5	71.2
Neopterygii	345.0	392.0

Below (Figures 7-8) are plots of the resulting timetree.

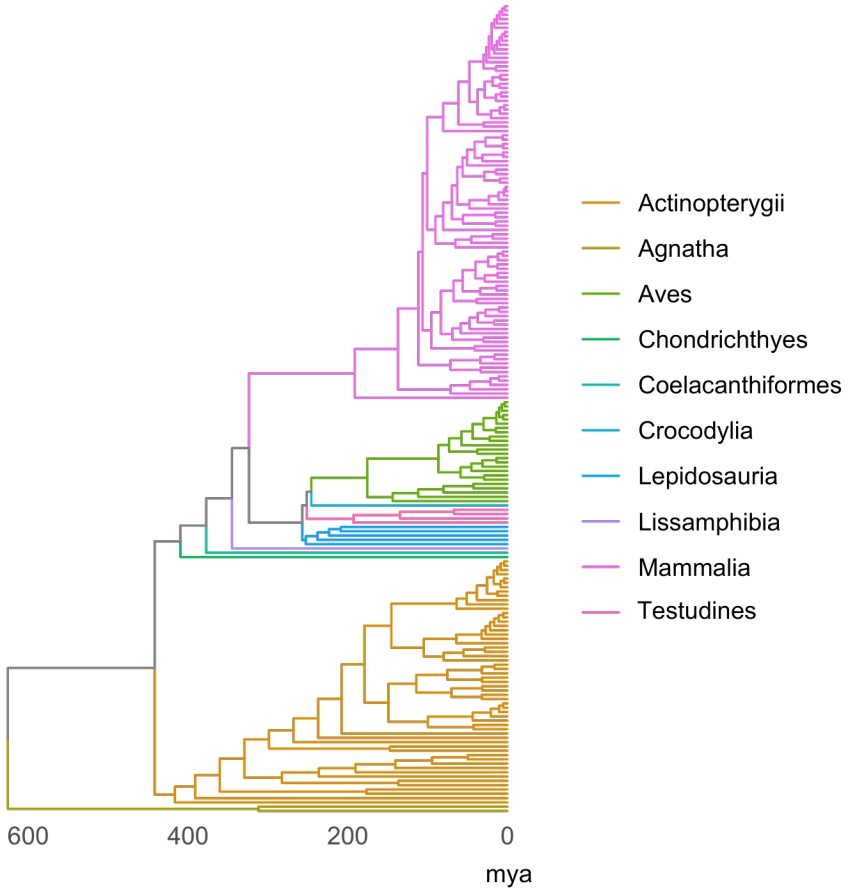

Figure 7. The timetree inferred from Morel et al.'s[35] tree contains the same 188 vertebrate species and relatedness structure as in Figure 5. The branch length unit is millions of years. The x-axis scale is in the unit millions of years ago (mya).

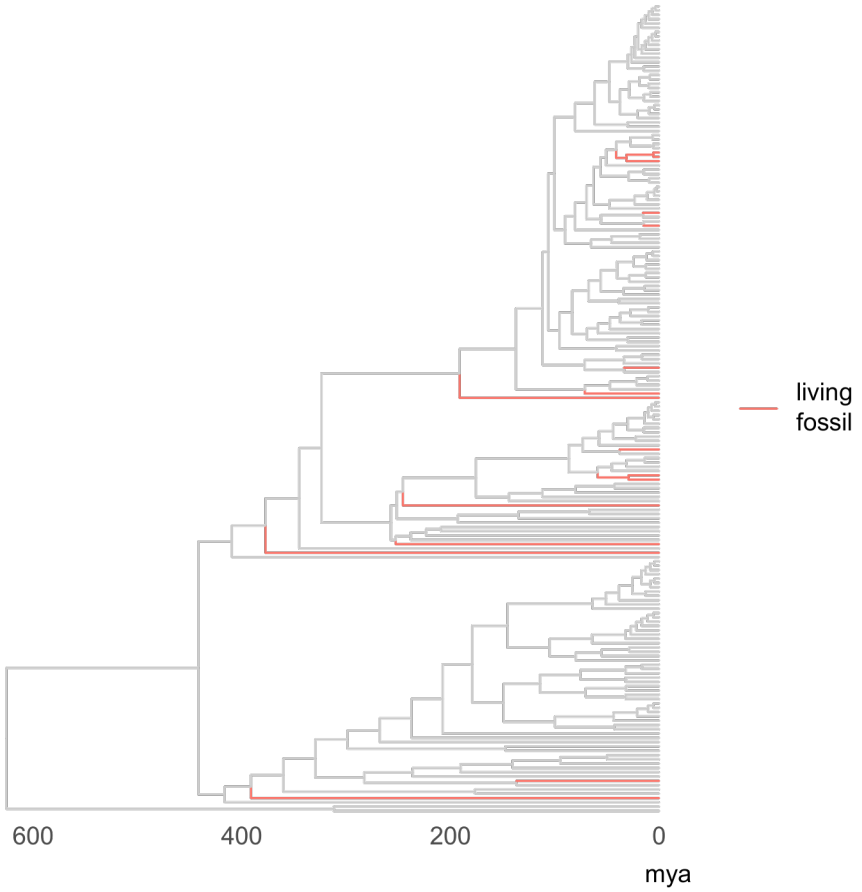

Figure 8. Of the 188 vertebrate species in Morel et al.'s[35] timetree, 16 (~9%) are recognized living fossils, per the four publications[10–13] (red branches).

Variables

We extracted variables of interest from the trees. Below, we walk through each variable.

Phylogenetic Path Length The phylogenetic path length is the root-to-tip distance, measured by summing branch lengths from the tree root to a particular tip[49,50]. It reflects the net evolution from a common ancestor to a descendant species. For example, the African coelacanth's path length (*Latimeria chalumnae*) in Irisarri et al.'s[34] tree represents how much the coelacanth's amino acids sequence has deviated from the most recent common ancestor (MRCA)

of all jawed vertebrates more than 400 mya (Figure 9). Living fossils should have shorter path lengths than other species in the tree. In a timetree comprising present-day species, all tips have the same time path length (Figures 3 and 7). The path length is only meaningful for the molecular trees. Because all species have survived the same duration since the jawed vertebrate MRCA, the path length is a proxy for the net rate of molecular evolution.

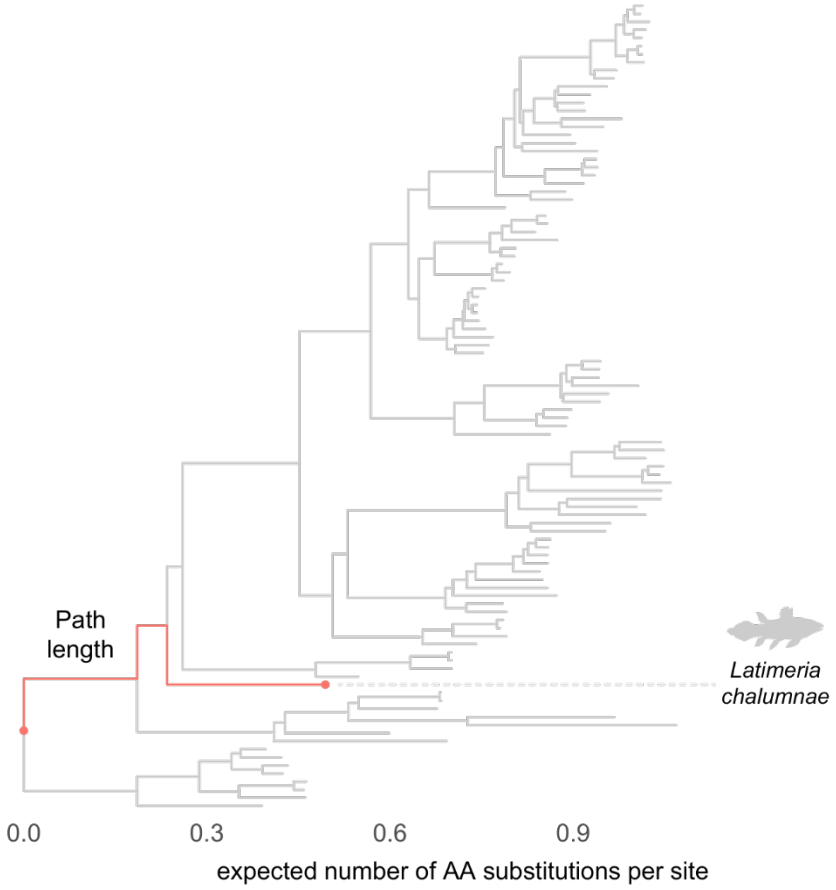

Figure 9. For example, the African coelacanth's (*Latimeria chalumnae*) phylogenetic path length is in red. Here, the path length is the sum of horizontal branch lengths; the vertical ones do not contain any information. We expect living fossil species to have short path lengths (i.e., low net coding sequence evolution). PhyloPic silhouette: *Latimeria* by Maija Karala (CC BY-NC-SA 3.0).

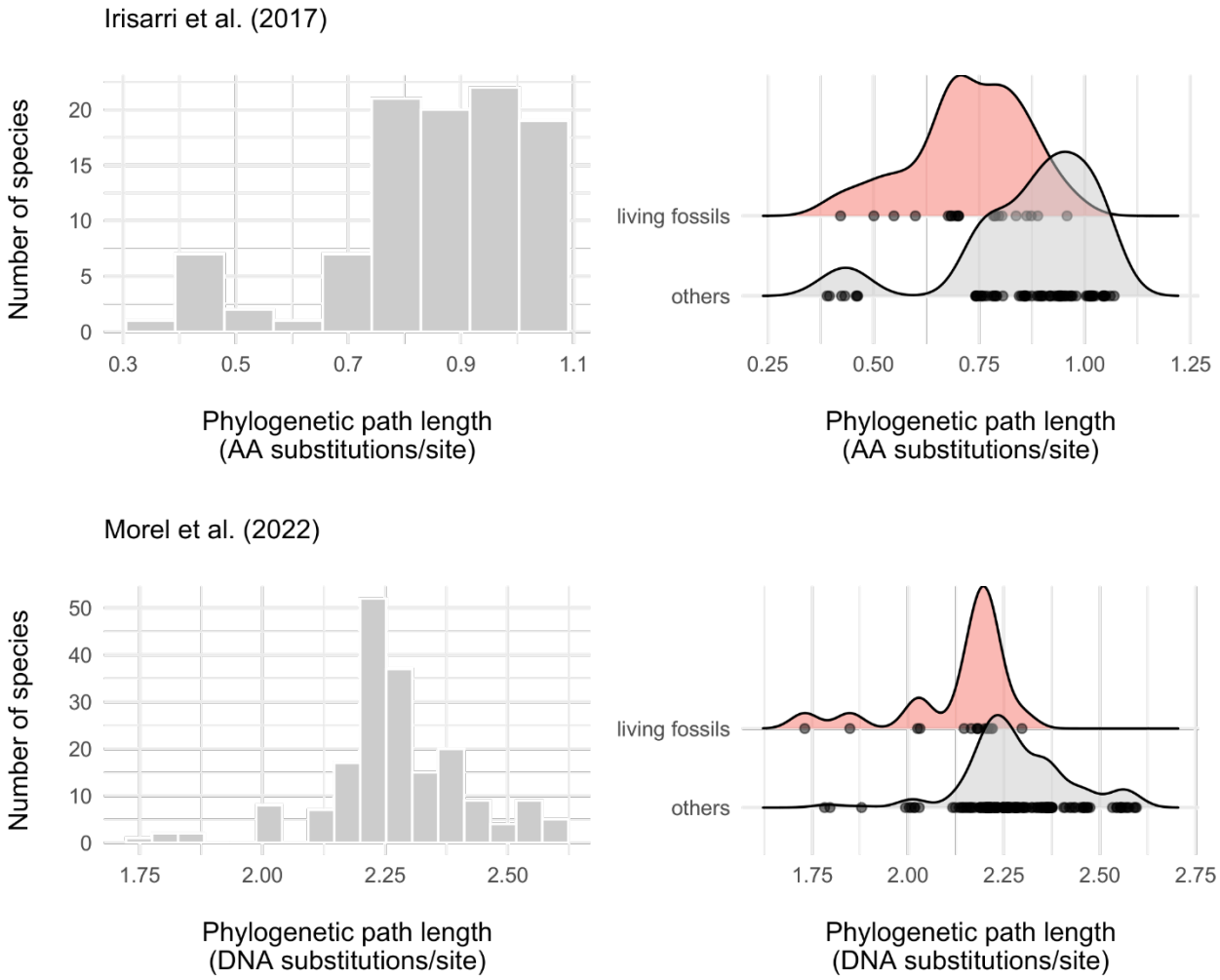

Figure 10. The path length distribution is not normal. On the left are histograms of path lengths of species in Irisarri et al.'s[34] and Morel et al.'s[35] molecular trees. We used Freedman and Diaconis's[51] rule for specifying the bin width. On the right are the corresponding densities separated by whether the species is a living fossil. Each dot is a species. In the 'others' (non-living-fossil) group, the data points at the left tail of Irisarri et al.'s[34] gray curve are sharks and rays not classified as living fossils per the four sources mentioned in the text above (Chondrichthyes; see Figure 1).

Based on Figure 10, "testing" if the coding sequence evolution is slower in living fossils may not be as simple as running a phylogenetic "t-test" (we performed it, regardless.) The reason is that the path length alone is scale-dependent. For example, the gray short-tailed opossum

(*Monodelphis domestica*) may be a living fossil among mammals. But it is probably not among the broader vertebrates. Its path length is short relative to other mammal species but is quite average at the vertebrate scale. Simpson[11] dealt with the same issue. He wrote, "Discussion of low and high evolutionary rates requires designation of the standard of reference." Conversely, some species have short path lengths, but paleontologists do not typically regard them as living fossils. These species drag down the average for the 'others' groups in Figure 10.

We could also be naive and define living fossil species as those whose path lengths lie in the bottom 10%. The 10% is an arbitrary cutoff, as mentioned previously. This definition catches only living fossils at the vertebrate scale.

Table 2. The bottom 10% of path lengths from Irisarri et al.'s[34] tree.

Species	Clade	Path length (AA subs/site)
Callorhinchus milii	Chondrichthyes	0.390
Carcharodon carcharias	Chondrichthyes	0.396
Scyliorhinus canicula	Chondrichthyes	0.422
Ginglymostoma cirratum	Chondrichthyes	0.425
Chiloscyllium griseum	Chondrichthyes	0.433
Raja clavata	Chondrichthyes	0.459
Neotrygon kuhlii	Chondrichthyes	0.461
Leucoraja erinacea	Chondrichthyes	0.463
Latimeria chalumnae	Coelacanthiformes	0.501
Neoceratodus forsteri	Dipnoi	0.549

And here's the list for Morel et al.'s[35] tree.

Table 3. The bottom 10% of path lengths from Morel et al.'s[35] tree.

Species	Clade	Path length (DNA subs/site)
Lepisosteus oculatus	Actinopterygii	1.73
Erpetoichthys calabaricus	Actinopterygii	1.78
Petromyzon marinus	Agnatha	1.80
Latimeria chalumnae	Coelacanthiformes	1.85

Table 3 Continued.

Species	Clade	Path length (DNA subs/site)
Callorhinchus milii	Chondrichthyes	1.88
Chrysemys picta bellii	Testudines	1.99
Gopherus agassizii	Testudines	2.00
Scleropages formosus	Actinopterygii	2.01
Chelonoidis abingdonii	Testudines	2.02
Pelodiscus sinensis	Testudines	2.02
Sphenodon punctatus	Lepidosauria	2.02
Paramormyrops kingsleyae	Actinopterygii	2.03
Crocodylus porosus	Crocodylia	2.03
Dromaius novaehollandiae	Aves	2.12
Apteryx rowi	Aves	2.12
Apteryx owenii	Aves	2.13
Apteryx haastii	Aves	2.14
Notamacropus eugenii	Mammalia	2.14

The two lists (Tables 2 and 3) differ due to different species in the trees and molecular data used (AA vs. DNA). If Morel et al.'s[35] tree includes more sharks and rays (chrondrichthyans), those species will dominate the list, as with Irisarri et al.'s[34] tree. Flagship vertebrate living fossils such as the African coelacanth (*Latimeria chalumnae*) are not necessarily the species with the slowest evolving protein-coding sequences. And the tuatara, which might have been the slowest-evolving reptile, is only in the list on Table 2.

While we are here, here are the top 10% fastest-evolving species in Irisarri et al.'s[34] tree.

Table 4. The top 10% of path lengths from Irisarri et al.'s[34] tree.

Species	Clade	Path length (AA subs/site)
Takifugu rubripes	Actinopterygii	1.07
Rana chensinensis	Lissamphibia	1.06
Espadarana prosoblepon	Lissamphibia	1.05
Pelophylax lessonae	Lissamphibia	1.05
Megophrys nasuta	Lissamphibia	1.04
Atelopus zeteki	Lissamphibia	1.04
Hymenochirus boettgeri	Lissamphibia	1.04

Table 4 Continued.

Species	Clade	Path length (AA subs/site)
Pelophylax nigromaculatus	Lissamphibia	1.04
Thamnophis elegans	Lepidosauria	1.02
Micrurus fulvius	Lepidosauria	1.02

And here's the list for Morel et al.'s[35] tree.

Table 5. The top 10% of path lengths from Morel et al.'s[35] tree.

Species	Clade	Path length (DNA subs/site)
Cyprinodon variegatus	Actinopterygii	2.60
Tetraodon nigroviridis	Actinopterygii	2.59
Xiphophorus couchianus	Actinopterygii	2.59
Poecilia latipinna	Actinopterygii	2.58
Gouania willdenowi	Actinopterygii	2.57
Fundulus heteroclitus	Actinopterygii	2.57
Takifugu rubripes	Actinopterygii	2.56
Poecilia formosa	Actinopterygii	2.56
Poecilia mexicana	Actinopterygii	2.56
Gambusia affinis	Actinopterygii	2.55
Poecilia reticulata	Actinopterygii	2.55
Oryzias melastigma	Actinopterygii	2.55
Xiphophorus maculatus	Actinopterygii	2.54
Cynoglossus semilaevis	Actinopterygii	2.53
Kryptolebias marmoratus	Actinopterygii	2.47
Betta splendens	Actinopterygii	2.47
Hippocampus comes	Actinopterygii	2.47
Maylandia zebra	Actinopterygii	2.46

These lists are not necessarily biased by clades with the most species in the respective trees. Irisarri et al.'s[34] tree includes 26 reptiles (lepidosaurs) and 26 amphibians, yet the species with the longest path is the Japanese pufferfish (*Takifugu rubripes*). Morel et al.'s[35] has 57 ray-finned fishes (actinopterygians) and 92 mammals. None of the mammals are on the top 10. Ray-finned fishes seem to be the fastest-evolving vertebrate group, which may not be too surprising

given that teleost fishes (a subset of Actinopterygii) underwent an additional third round of whole-genome duplication event compared to other vertebrates[52].

Node Count The node count is the number of internal nodes along the root-to-tip path[49,50]. It represents the number of net-speciation events since a common ancestor. Net-speciations are lineage-branching events, excluding those erased by extinction. If the tree were to include extinct species, the node count of the living species would increase. Given ~99% of all species that have ever lived are extinct[53], the node count derived from a complete tree of life, which is unattainable, is much different from the node count extracted from any inferred phylogenetic tree. Hence the interpretation of the node count as a proxy of cumulative net-speciation.

The node count may predict stasis since past studies have documented the link between disparity and diversity[10,49,50,54–58]. We will discuss the mechanistic hypotheses later below. For now, it is sufficient to note the low diversity of clades to which living fossils belong[13,55,56]. For example, the clade Actinistia (coelacanths) was more diverse in the past. However, this diversity has always been minuscule compared to their contemporaries, the Actinopterygii (ray-finned fishes), Lissamphibia, and Amniota (birds, reptiles, and mammals). This phenomenon is likely not due to high extinction rates in Actinistia but a combination of low speciation and low extinction rates[59]. Figure 11 below shows the African coelacanth's (*Latimeria chalumnae*) node count in Irisarri et al.'s[34] tree. We expect living fossils to have low node counts.

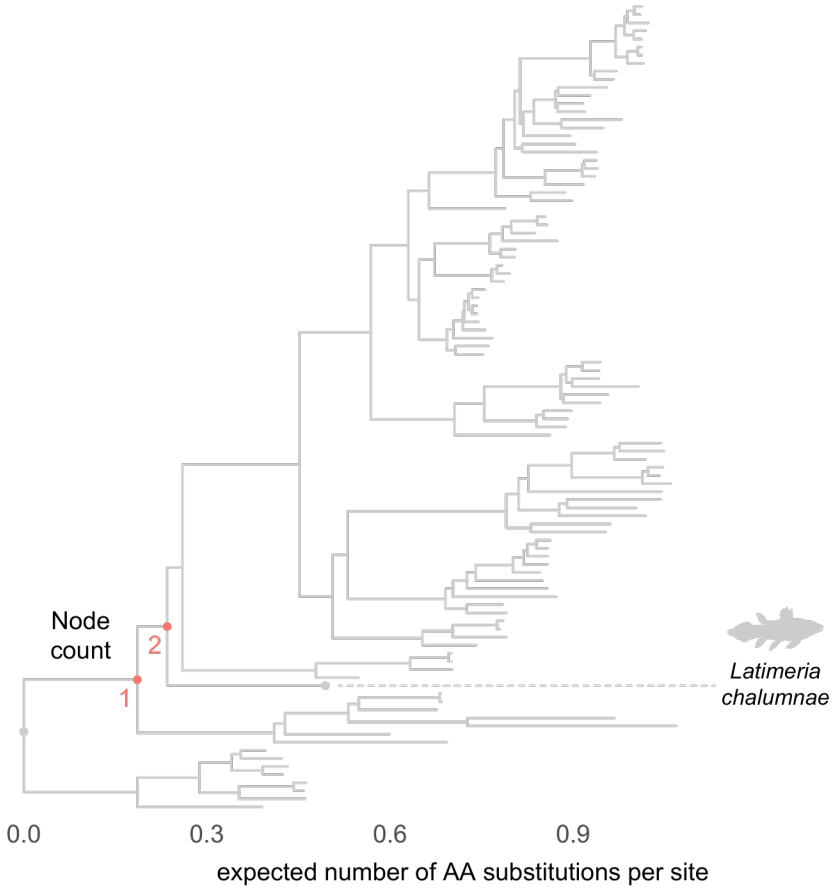

Figure 11. For example, the African coelacanth's (*Latimeria chalumnae*) node count is 2. We expect living fossil species to have low node counts (i.e., low cumulative net-speciation events).

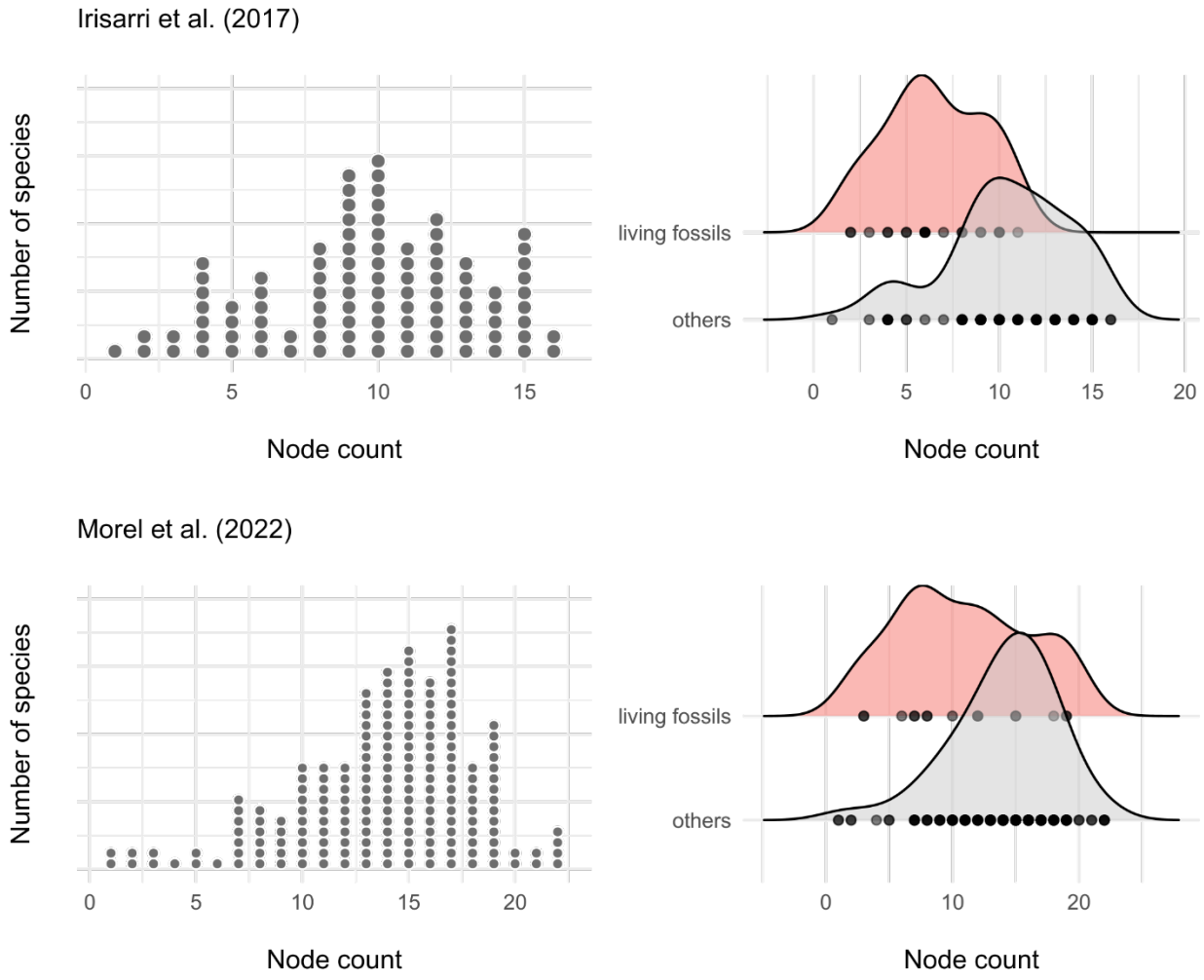

Figure 12. The node count distribution is not always symmetric. On the left are dot plots of node counts of species in Irisarri et al.'s[34] and Morel et al.'s[35] molecular trees. On the right are the corresponding densities separated by whether the species is a living fossil. Each dot is a species.

Figure 12 above shows that, on average, living fossils exhibit lower node counts than other vertebrate species.

Terminal Time Branch Length A terminal time branch length represents the time since a lineage's origin. Living fossil lineages have persisted for extended periods[11]. So, they should have long terminal branches. The African coelacanth (*Latimeria chalumnae*) line last split from

its common ancestor with lungfishes and tetrapods (four-legged animals, from amphibians to mammals) ~400 million years ago[34] (Figure 13). By contrast, for example, humans have only split from the chimpanzees 9.0-6.9 million years ago[60]. Therefore, time-related variables seem crucial in capturing the essence of living fossils[18,19], at least, in our case, as covariates for explaining path length.

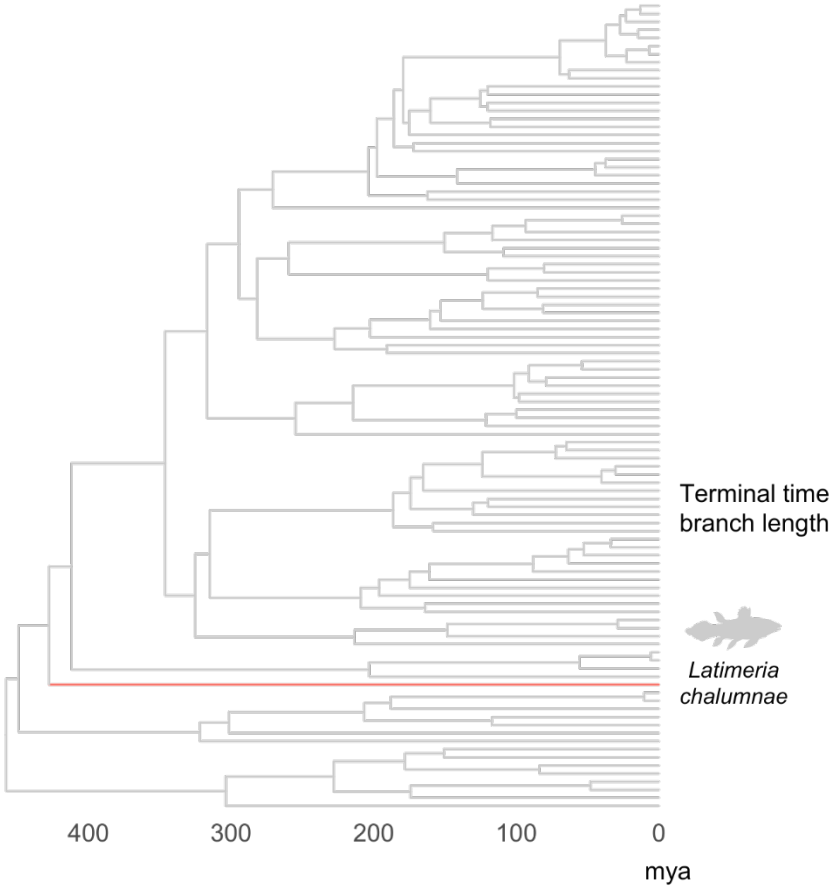

Figure 13. For example, the African coelacanth's (*Latimeria chalumnae*) terminal branch is long (red).

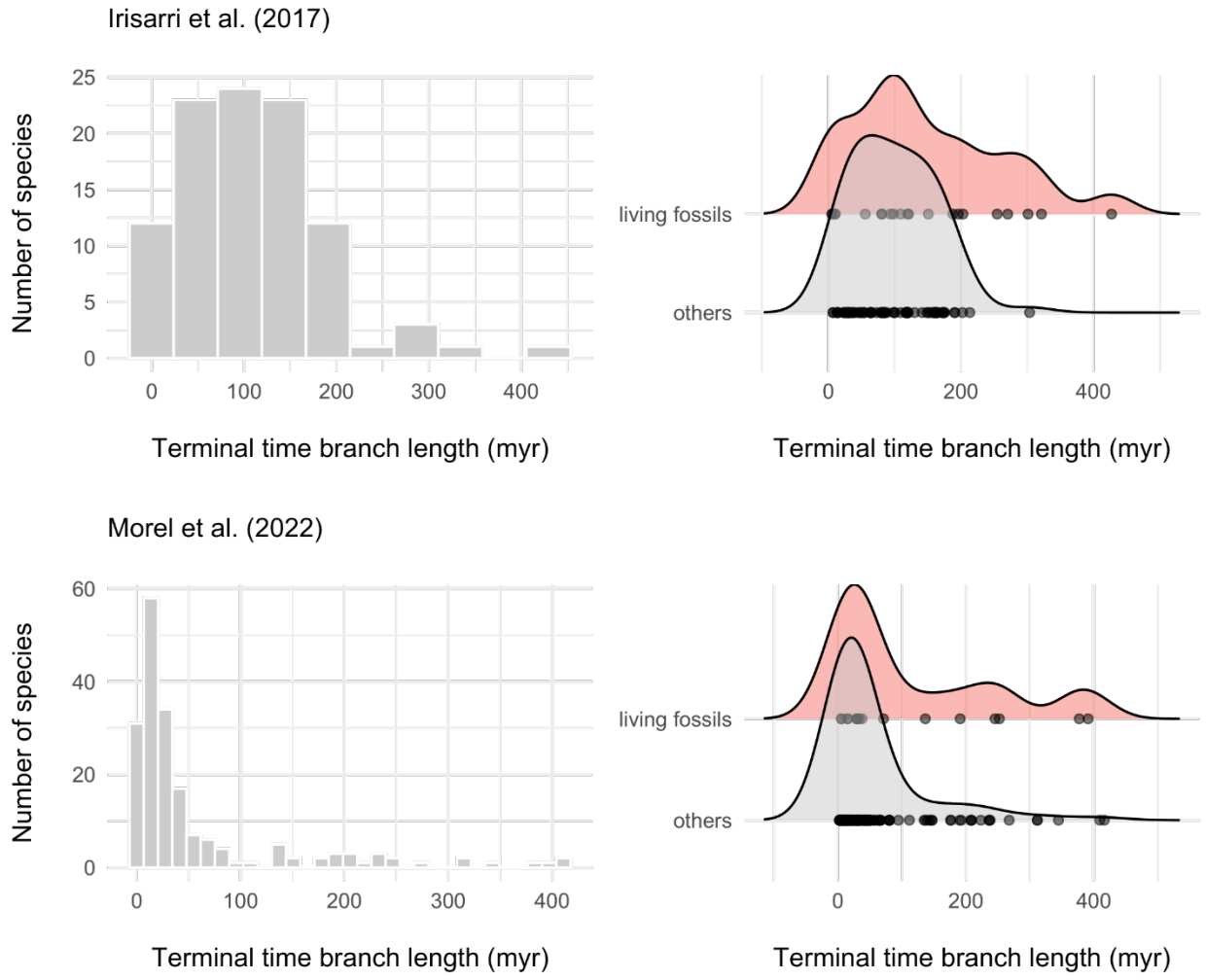

Figure 14. The terminal time branch length distributions are right-skewed, but living fossil species do not necessarily have long terminal branches, on average. Some living fossils have short terminals, and some species typically not regarded as living fossils have long terminals, such as the agnathans (jawless fishes). Branch lengths are in millions of years (myr).

Contrary to prior expectations, living fossil species do not necessarily tend to have long terminal branches. And some species not typically regarded as living fossils have persisted for long periods. Should we look at the following list of the top six species with the shortest terminal time branch in Irisarri et al.'s[34] tree, we recognize a familiar issue:

Table 6. The top six species with the shortest terminal time branch in Irisarri et al.'s[34] tree.

Species	Clade	Living fossil	Terminal time branch length (millions of years)
Protopterus aethiopicus	Dipnoi	Yes	5.4
Protopterus annectens	Dipnoi	Yes	5.4
Crotalus adamanteus	Lepidosauria	No	6.9
Sistrurus miliarius	Lepidosauria	No	6.9
Lepisosteus oculatus	Actinopterygii	Yes	10.2
Lepisosteus platyrhincus	Actinopterygii	Yes	10.2

Two lungfish (*Protopterus*) and two gar species (*Lepisosteus*), widely regarded as living fossils, have short terminal time branch lengths. This time variable is susceptible to species sampling. The terminal time branch length of a living fossil depends on whether its closely related living fossil sister is also in the tree, which is the case here. The African coelacanth's (*Latimeria chalumnae*) terminal would decrease from ~400 million years to 30-40 million years had its sister, the Indonesian coelacanth (*Latimeria menadoensis*), been included in the tree[61]. Perhaps, extant living fossil lines are generally more recent than traditionally thought[62–64]. Either way, the terminal time branch length is likely not a good covariate.

Simpson[11] proposed the analogy that living fossils are species that had been passed in the evolutionary race. For his example, he noted how some pelecypod lineages (now known as bivalves) have existed since ~450 million years ago. What is impressive is that such lines slowed down in the race a long time ago and has not picked up its pace. Therefore, a variable of interest may not be the terminal time branch length but the time since a lineage slowed down. However, there is no clear, objective way to measure this variable, nor would it be applicable across all vertebrate species.

Node-Density Artifact

In 2006, Chris Venditti and colleagues studied a phenomenon called the node-density artifact and proposed a way to test for it. Branch lengths in tree regions with fewer species are more likely to be underestimated[65]. The culprit is the under-detection of hidden DNA/AA substitutions. For example, the letter 'A' in a species' DNA sequence might have undergone a transition to the letter 'G' in its history before transitioning back to 'A' (A→G→A). Having close relatives for comparisons helps in uncovering such substitutions. We care about this artifact because living fossil branch lengths may be artificially slightly shorter.

Venditti et al.[65] noticed that the symptom of the artifact is a convex curvilinear relationship between node count (n) and path length (x). The equation is $n = \beta x^{\delta}$ where β is the rate of change between node count and path length, and δ is the degree of convexity. $\delta > 1$ when the node-density artifact is present.

Irisarri et al.'s[34] and Morel et al.'s[35] trees seem to suffer from the node-density artifact ($\hat{\delta}_{Irisarri} = 1.3 \pm 0.6$, $\hat{\delta}_{Morel} = 2.7 \pm 0.8$). The 95% confidence interval (95% CI) for Irisarri et al.'s[34] tree includes the null value of one, whereas the 95% CI for Morel et al.'s[35] does not. See codes and model output in the supplemental files. However, we have reasons to doubt that the curvilinear relationships are due to the artifact.

First, Venditti et al.[65] ran simulations and found that inferring branch lengths using an overly simplistic model of DNA sequence evolution leads to the node-density artifact. Irisarri et al.[34] and Morel et al. used state-of-the-art phylogenetic inference pipelines. So, it is doubtful that the node-density artifact in our case is due to model insufficiency.

Second, tree shape or imbalance[66,67] is a prime candidate for explaining the node-density artifact. Few species with low path length and node count at the lower left corner of a scatter plot (i.e., living fossils) would likely cause curvilinearity. The presence of living fossil branches in a tree means that the tree is imbalanced, with more species on one side of the tree than the other. Venditti et al.[65] did not find evidence for this idea. But, Venditti et al.'s[65] simulated trees were relatively balanced. Using the Colless' index (I_c, ranging from 0 [perfectly balanced] to 1 [most imbalanced][66]), the mean $\hat{I}_c$ of their trees was 0.12 ± 0.03. It is well known, however, that many empirical trees are imbalanced[59,67–69]. Unsurprisingly, Irisarri et al.'s[34] and Morel et al.'s[35] trees are highly imbalanced ($\hat{I}_{c,Irisarri} = 0.62$, $\hat{I}_{c,Morel} = 0.72$). Venditti et al.[65] argued that balanced trees are susceptible to node-density artifacts. Still, we suspect that tree shape would have been a better predictor in their study had they simulated more imbalanced trees.

Inferring a new tree with more species would not reduce the node-density artifact since it would only increase tree imbalance.

Lastly, as we will see in the analysis sections, the curved fit line for Morel et al.'s[35] node count and path length regression (the inverse of regressions fitted here) is due to heterogeneity across the tree rather than the node-density artifact. Evolutionary patterns and processes vary in large-scale phylogenetic trees[58,70,71]. One fitted line as a test for the node-density artifact may be too simple.

Phylogenetic "*t*-test"

If we were to designate the top 10% of the slowest evolving vertebrates as genomic living fossils, then we can skip this section. However, there is value in investigating to what extent vertebrate species with conservative anatomy have slowly evolving protein-coding sequences.

Do species classified as living fossils by the four paleontological sources[10–13] merit the living fossil label from a genomic standpoint? We hypothesize that the phylogenetic path lengths of living fossils for thousands of nuclear genes are shorter than the vertebrate average.

Methods

Based on Figure 10, this method is likely too simplistic. It fails to account that some species, such as the macaques (*Macaca*), are probably living fossils relative to other primates but certainly not to most vertebrates. Conversely, there are species that paleontologists do not traditionally classify as living fossils (sharks and rays, i.e., chrondrichthyans) despite their slowly evolving genomes.

Regardless, below is the model notation:

$$y_i = \beta_0 + \beta_1 x_{i,LF=1} + \epsilon_i$$

- y_i: the phylogenetic path length of the i^{th} species in the phylogenetic tree
- β_0: the predicted average path length of non-living-fossil species
- β_1: the predicted difference in average path length between non-living-fossil and living fossil species
- $x_{i,LF=1}$: an indicator variable of whether the i^{th} species is a living fossil (1: yes; 0: no)
- ϵ_i: the error term for the i^{th} species

A phylogenetic "*t*-test" uses a variance-covariance matrix derived from a phylogenetic tree to account for the non-independence between species. It is a generalized least squares (GLS) approach, as briefly mentioned above. See Freckleton[72] and Symonds and Blomberg[24] for detailed phylogenetic GLS (PGLS) walkthroughs. The errors are not independently and

identically distributed. Instead, they follow a multivariate normal distribution with the variance-covariance matrix $\boldsymbol{V}$ (see below). This matrix uses branch length information from a phylogenetic tree to represent similarities between species. In a molecular tree, each species has a different phylogenetic path length. Those with longer paths have higher variances.

$$\epsilon_i \sim N(\mathbf{0}, \sigma^2 \boldsymbol{V})$$

Here, we ran the "*t*-test" on the Irisarri et al.'s[34] and Morel et al.'s[35] datasets, each using the Bayesian regression modeling R package rstanarm v2.26.1[73] and the binary program BayesTraits v4.0.1 (https://www.evolution.reading.ac.uk/BayesTraitsV4.0.1/BayesTraitsV4.0.1.html). rstanarm does not have a specific GLS function or a way for users to supply a variance-covariance matrix (as far as we know). However, we figured out a workaround by using linear algebra to transform the variables before feeding them into stan_glm (steps shown below). I used BayesTraits, a popular PCMs (phylogenetic comparative methods) software program developed by colleagues at the University of Reading, for benchmarking.

We need to de-correlate our data before model-fitting with rstanarm. First, let $\boldsymbol{V}$ be the $n \times n$ variance-covariance matrix, where n is the number of species in a phylogenetic tree. Each cell in the matrix represents the shared evolutionary history for a pair of species (covariance), whereas each cell on the matrix's diagonals is the path length (variance). See Symonds and Blomberg[24] on how to convert a phylogenetic tree into a variance-covariance matrix. Then, we modify and decompose $\boldsymbol{V}$ using Cholesky whitening[74], which, as far as we know, requires fewer matrix calculations than other whitening methods, such as the singular value decomposition

(SVD). Using Cholesky decomposition, we get a matrix $\boldsymbol{D}$ such that $\boldsymbol{DD^T = V^{-1}}$. We will use this matrix to normalize the data against their evolutionary history.

Next, we multiplied the $\boldsymbol{y}$ vector of path lengths and the $\boldsymbol{X}$ design matrix (the first column is for the intercept, and the second for the living fossil indicator [1: yes; 0: no]) with the $\boldsymbol{D}$ matrix. Note that the resulting intercept column was no longer a vector of 1s. The normalized values are no longer interpretable.

We then fed the normalized data into stan_glm. I removed the default intercept given by the function (path_length ~ intercept + living_fossil - 1). This function's default priors are weakly informative, more appropriate than flat priors[75]. The model-fitting was fast, thanks to the Hamiltonian Monte Carlo algorithm[76]. All other Markov chain Monte Carlo (MCMC[77]) settings were left unchanged. See codes in the supplemental files.

Results and Discussion

For Irisarri et al.'s[34] dataset, the predicted average phylogenetic path length of non-living-fossil vertebrate species is between 0.44 and 0.63 AA substitutions per site with a 95% probability. The estimated difference in the average for living fossil species is, with a 95% probability, between -0.04 and 0.01 AA substitutions per site. Practically, no difference between the two groups. For Morel et al.'s[35] dataset, there is a 95% probability that the estimated average path length of non-living fossils is 1.95 ± 0.14 DNA substitutions per site, and the predicted difference for living fossils is only -0.01 ± 0.02. Similar results.

BayesTraits produced almost identical estimates. The MCMC simulations ran for 11,000,000 iterations with 1,000,000 burn-ins and a sampling period of 1,000. I used the default flat priors (uninformative, but they get "swamped" by the data anyway).

This result is unsurprising. A "*t*-test" may be too simple after all. It fails to detect living fossils at multiple taxonomic levels. Another perspective is that there are many vertebrate species whose anatomy paleontologists have not recognized as evolving slowly, yet their protein-coding sequences are. And many living fossil species do not possess the most conservative genomes at the vertebrate scale. At the extreme, this result shows the subjectivity in what species paleontologists claimed as living fossils.

Phylogenetic Prediction

A more sophisticated method is the phylogenetic prediction[29,78]. The idea is to assess how much a living fossil species evolved more slowly than predicted from the phylogenetic tree alone (its relationship with other species). Incorporating the tree reduces prediction uncertainty[78,79].

Methods

The analysis begins with fitting a model, $y_i = \beta_0 + \epsilon_i$ (see variable and parameter descriptions in the previous section), without the living fossil species of interest. The resulting model is the basis for predicting the distribution of the living fossil's path length. We expect a living fossil's observed path length to fall at the left tail or outside the posterior predictive distribution[75]. If so, the interpretation would be that the living fossil is an outlier in its exceptionally sluggish protein-coding sequence evolution.

Should we predict all living fossils one by one? Before that, we chose six species in Irisarri et al.'s[34] and Morel et al.'s[35] trees for demonstrations. These species are *Latimeria chalumnae* (African coelacanth), *Sphenodon punctatus* (tuatara), *Monodelphis domestica* (gray, short-tailed opossum), *Lepisosteus oculatus* (spotted gar), *Homo sapiens* (human), and *Takifugu*

rubripes (Japanese pufferfish). The first four are living fossils. The method should identify *Latimeria* as an outlier in both trees. It may be for *Sphenodon*, but not *Monodelphis*; it is probably not an outlier in the grand scheme of vertebrates. *Lepisosteus oculatus* may be an outlier in Morel et al.'s[35] dataset but not Irisarri et al.'s[34] because this dataset includes its fellow living fossil sister, *Lepisosteus platyrhincus*. *Homo* is universally regarded as not a living fossil, so it serves as a "negative control." Its path length should fall near the center of the posterior predictive distribution. Lastly, *Takifugu* has the longest path in Irisarri et al.'s[34] tree, meaning that its observed path length should fall on the right tail of the predictive distribution. However, in Morel et al.'s[35] tree, *Takifugu*'s path is not the longest since the tree includes many ray-finned fishes. So, its path length may fall within the prediction interval for that tree. Altogether, the six species should demonstrate the power and drawback of the phylogenetic prediction.

To measure how much an observed path length falls under the posterior predictive distribution, we used posterior predictive *p*-values[80], sometimes called Bayesian *p*-values. Here, we refer to them as p_{MCMC}. For practicality, we used a left-tailed p_{MCMC}. The p_{MCMC} should ideally be ~0 for living fossils, ~0.5 for species that fall near the center of the predictive distribution (e.g., *Homo*), and ~1 for those that evolved faster than predicted (e.g., *Takifugu*).

To streamline the phylogenetic predictions, we wrote a function, phy_out (see the supplemental files). It relies on the rstanarm functions stan_glm and posterior_predict. But, the normalized path lengths are uninterpretable. Given our workaround with rstanarm's functions, there is no obvious "back-transforming" method. Therefore, we repeated the posterior predictions using BayesTraits.

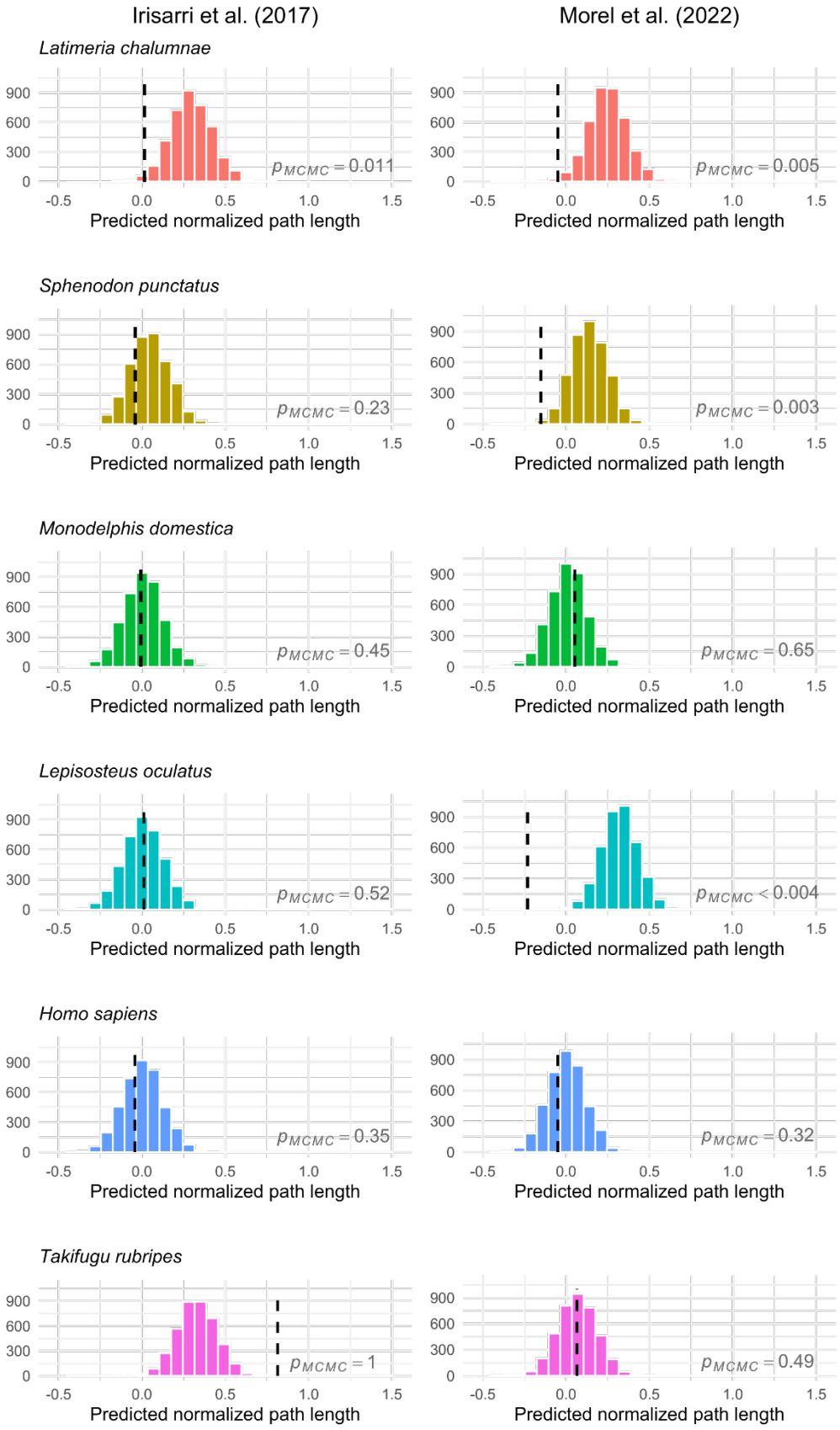
Irisarri et al. (2017)
Morel et al. (2022)
Latimeria chalumnae
$p_{MCMC} = 0.011$
$p_{MCMC} = 0.005$
Sphenodon punctatus
$p_{MCMC} = 0.23$
$p_{MCMC} = 0.003$
Monodelphis domestica
$p_{MCMC} = 0.45$
$p_{MCMC} = 0.65$
Lepisosteus oculatus
$p_{MCMC} = 0.52$
$p_{MCMC} < 0.004$
Homo sapiens
$p_{MCMC} = 0.35$
$p_{MCMC} = 0.32$
Takifugu rubripes
$p_{MCMC} = 1$
$p_{MCMC} = 0.49$
900
600
300
0
-0.5
0.0
0.5
1.0
1.5
Predicted normalized path length

Figure 15 (previous page). Posterior predictive distributions of the normalized path lengths of six species based on Irisarri et al.'s[34] and Morel et al.'s[35] trees generally support prior expectations. Histogram colors represent species. See the main text for detailed interpretations. Dashed black lines indicate the observed path lengths. The normalized path length is unitless. Its absolute value or scale is meaningless; focus on the comparisons between the observed and predicted.

Results and Discussion

The results in Figure 15 are as we expected. The African coelacanth (*Latimeria chalumnae*) is a phylogenetic outlier in both trees ($p_{MCMC} < 0.05$). The tuatara's (*Sphenodon punctatus*) path length is not exceptionally short in Irisarri et al.'s[34] tree but is in Morel et al.'s[35] tree. Irisarri et al.'s[34] tree contains several sharks and rays, all of which have evolved more slowly than the tuatara. Morel et al.'s[35] tree only has one. This comparison shows how the phylogenetic prediction is susceptible to species selection in the tree. The gray, short-tailed opossum's (*Monodelphis domestica*) protein-coding sequence has evolved as expected. It is not a genomic living fossil at the vertebrate scale. The spotted gar's (*Lepisosteus oculatus*) case further exemplifies the susceptibility to species sampling. Irisarri et al.'s[34] tree includes its sister, *Lepisosteus platyrhincus*, which has also evolved slowly; hence the high p_{MCMC}. As a control, the human's (*Homo sapiens*) path length is as predicted. And the Japanese pufferfish's (*Takifugu rubripes*) path length falls on the right tail as anticipated. It falls within the prediction interval in Morel et al.'s[35] tree because this tree includes more ray-finned fishes that evolved even faster.

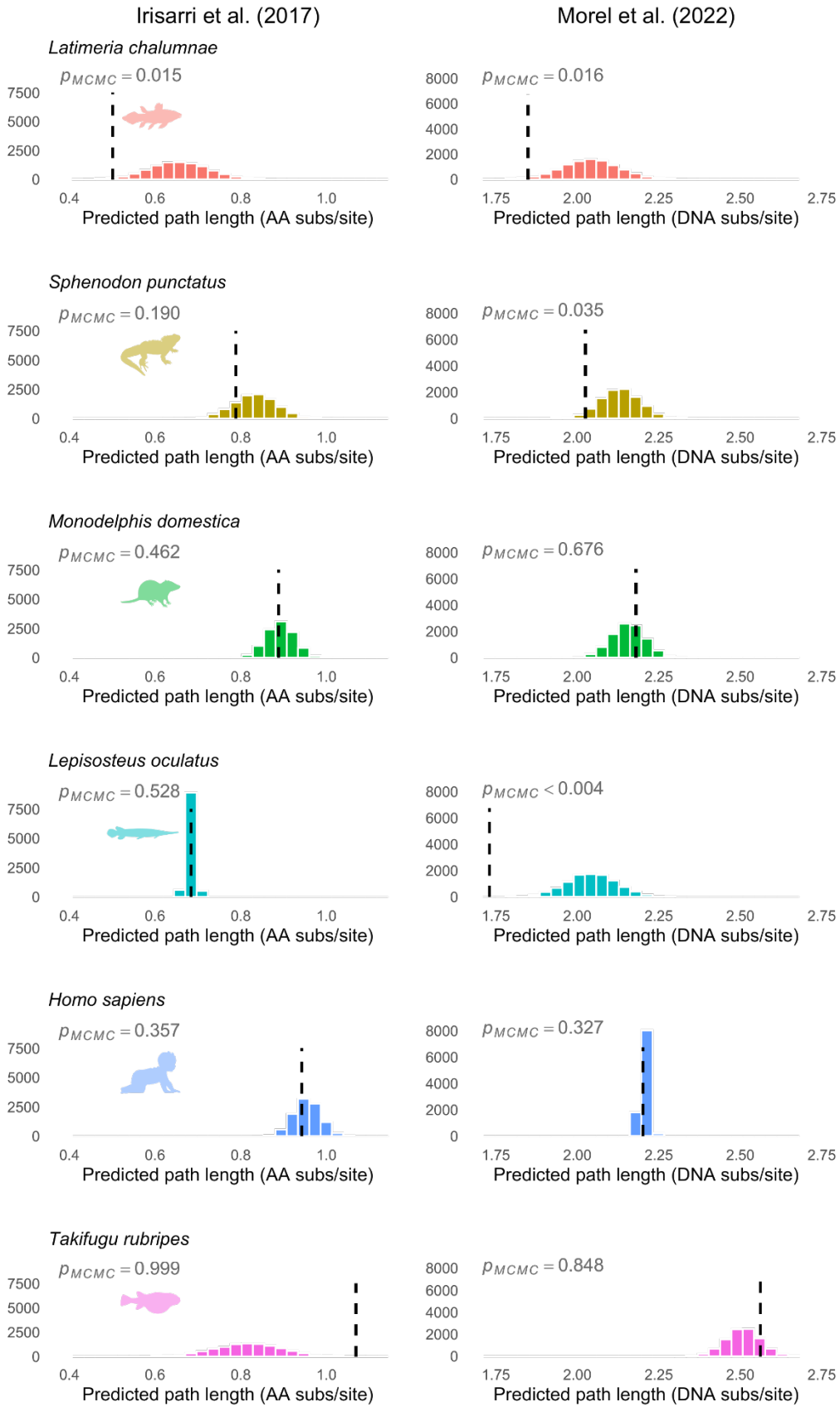
Irisarri et al. (2017)
Morel et al. (2022)
Latimeria chalumnae
$p_{MCMC} = 0.015$
$p_{MCMC} = 0.016$
Sphenodon punctatus
$p_{MCMC} = 0.190$
$p_{MCMC} = 0.035$
Monodelphis domestica
$p_{MCMC} = 0.462$
$p_{MCMC} = 0.676$
Lepisosteus oculatus
$p_{MCMC} = 0.528$
$p_{MCMC} < 0.004$
Homo sapiens
$p_{MCMC} = 0.357$
$p_{MCMC} = 0.327$
Takifugu rubripes
$p_{MCMC} = 0.999$
$p_{MCMC} = 0.848$
Predicted path length (AA subs/site)
Predicted path length (DNA subs/site)

Figure 16 (previous page). Posterior predictions of path lengths of six species based on Irisarri et al.'s[34] and Morel et al.'s[35] using BayesTraits produced similar results. Histogram colors represent species. Dashed black lines indicate the observed path lengths. PhyloPic silhouettes: *Latimeria* by Maija Karala (CC BY-NC-SA 3.0); *Sphenodon* by Steven Traver (CC0 1.0); *Monodelphis* by Milena Cavalcanti, Patricia Pilatti, and Diego Astúa (CC BY 4.0); *Lepisosteus* by Milton Tan (CC BY-NC-SA 3.0); *Homo* by Andrew Farke (CC BY 3.0); *Takifugu* by seung9park (CC0 1.0).

The BayesTraits posterior predicted values are relatively similar. However, the units are now interpretable. For example, with 95% probability, the African coelacanth's (*Latimeria chalumnae*) protein-coding sequence has diverged 0.52-0.79 amino acids substitutions per site and 1.87-2.21 DNA substitutions per site less than predicted based on its evolutionary relationships with other vertebrates.

In summary, our Bayesian phylogenetic prediction demonstrated the ability to quantify how much slower are the genomic sequence evolution for the two widely recognized living fossils (coelacanth and tuatara). However, a caveat is that the predicted values are sensitive to the species sampled in the phylogenetic tree. This downside stresses that the "living fossil" concept will always be relative, and that defining or classifying living fossil species using absolute criteria, as researchers have continued to attempt, is fruitless. Instead, we can focus on species with very slow genomic evolution relative to their contemporaries and shed light on what might have driven such stasis.

CHAPTER TWO

EXPLAINING THE SLOW GENETIC EVOLUTION IN LIVING FOSSILS

Introduction

In *On the Origin of Species*, Darwin[10] envisioned that variation emerges gradually, where evolutionary change accumulates incrementally over long timespans. Since then, we know the rate of evolution is not constant; it can speed up or slow down[11]. But how? In 1954, the biologist Ernst Mayr[81] noted that the branching out of a subpopulation into a new environment, splitting (or speciating) from its parent population, may be associated with rapid evolution. In biology, splitting can happen in two manners, allopatric or peripatric, but we will not discuss them here. The increased evolutionary rate during speciation could be due to genetic drift related to small population size (mutations not necessarily beneficial may get fixed simply due to random sampling) or adaptive as a subpopulation enters a new environment. In 1972, paleontologists Niles Eldredge and Stephen Jay Gould[54] noticed that new species can appear abruptly in the fossil record. Borrowing Mayr's idea, they hypothesized that a lineage's history comprises long periods of stasis punctuated by rapid evolution around speciation events. Their hypothesis is called punctuated equilibrium.

Based on punctuated equilibrium, long-term stasis in some lineages may be attributable to such lines having speciated less often than their diverse counterparts[54–56]. One can view speciation as a mechanism to "lock" mutations arising in local settings, thereby "breaking the chains" of regressing toward the mean[82]. Net divergence should correspond with the number of net speciations.

In 2006, Mark Pagel and colleagues[50] devised a method to detect punctuated evolution, which estimates how much molecular variation is attributable to net-speciation events (0%-100%). The idea is to regress the phylogenetic path length (measured in substitutions per site) on node count. Using this phylogenetic regression method, they inferred that net speciation explains ~16% of the total molecular evolution in extant plants, fungi, and animals. Punctuated molecular evolution is, therefore, widespread across the tree of life. This path length ~ node count regression forms the bedrock of our analyses below.

As a side note, Uyeda et al.[83] argued that phenotypic evolution tends to speed up after one million years. They found that body size divergences across mammals, reptiles, and birds are generally higher at timescales beyond one million years. Curiously, this one-million-year duration coincides with the estimated species lifespan of mammals[84], ranging from 1-2 myr, and the estimated species lifespan generally of 10^3-10^7 years[85]. This coincidence further hints that evolution tends to be rapid during speciation.

Stephen Jay Gould's quote in his 2002 parting tome[56] sums up the idea above best:

> "But punctuated equilibrium suggests another, remarkably simple, explanation once you begin to think in this alternative mode—an insight that ranks in the exhilarating, yet frustrating, category of obvious 'scales falling from eyes' propositions, once one grasps the new phrasing of a basic question. If evolutionary rate correlates primarily with frequency of speciation—the cardinal prediction of punctuated equilibrium—then living fossils may simply represent those groups at the left tail of the distribution for numbers of speciation events through time. In other words, living fossils may be groups that have persisted through geological time at consistently and unvaryingly low species diversity."

Therefore, the question of interest is whether speciation explains the slow evolution of living fossil protein-coding sequences. We hypothesize that short phylogenetic path lengths are due to a lack of net-speciation events (i.e., low node count).

Variables

Path length and node count are the outcome and predictor variables of interest. We previously mentioned that the terminal time branch length could be a covariate to include. However, it generally carries similar information with node count (Figures 17-18). Long terminal time branches are reflections of low net speciation rate. Surprisingly, the variance inflation factors (VIFs) of node count and terminal time branch length in the Irisarri et al.'s[34] and Morel et al.'s[35] cases are 2.3 and 1.9, which are not too extreme (see the supplemental files). So, there's insufficient reason to exclude them from subsequent model buildings. Therefore, we have five variables (path length, node count, terminal time branch length, clade, and living fossil [yes/no]).

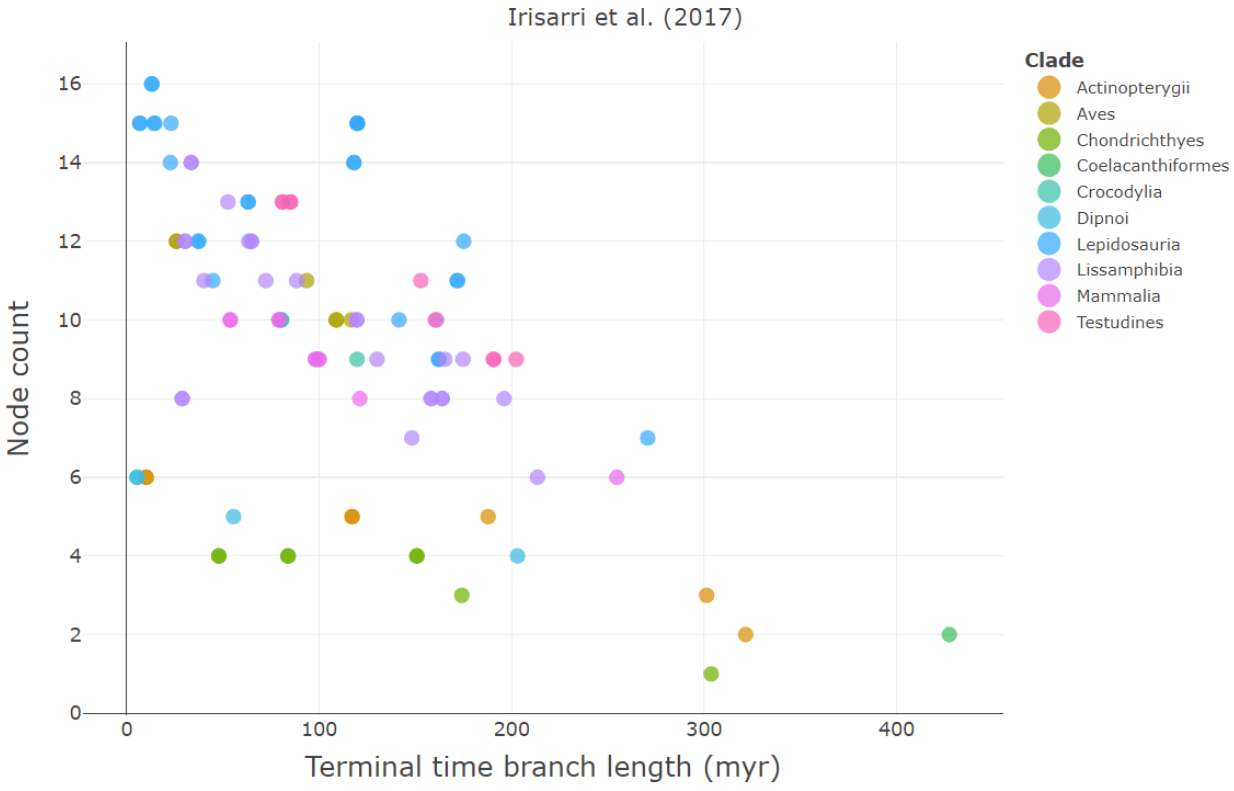

Figure 17. Terminal time branch lengths seem to covary with node count in Irisarri et al.'s[34] tree. See the interactive version in the supplemental files.

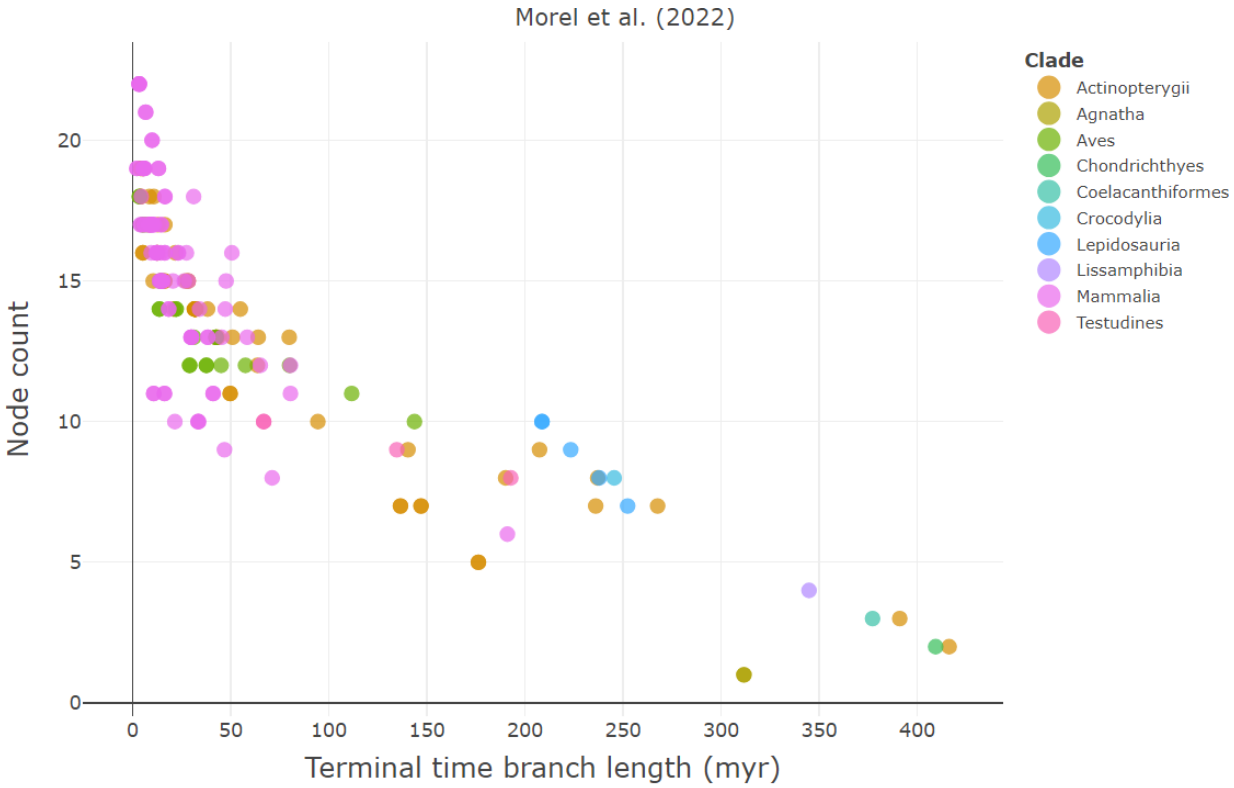

Figure 18. Terminal time branch lengths seem to covary non-linearly with node count in Morel et al.'s[35] tree. See the interactive version in the supplemental files.

Models

Irisarri et al. (2017)

Below are two scatter plots to help guide model selection for Irisarri et al.'s[34] dataset.

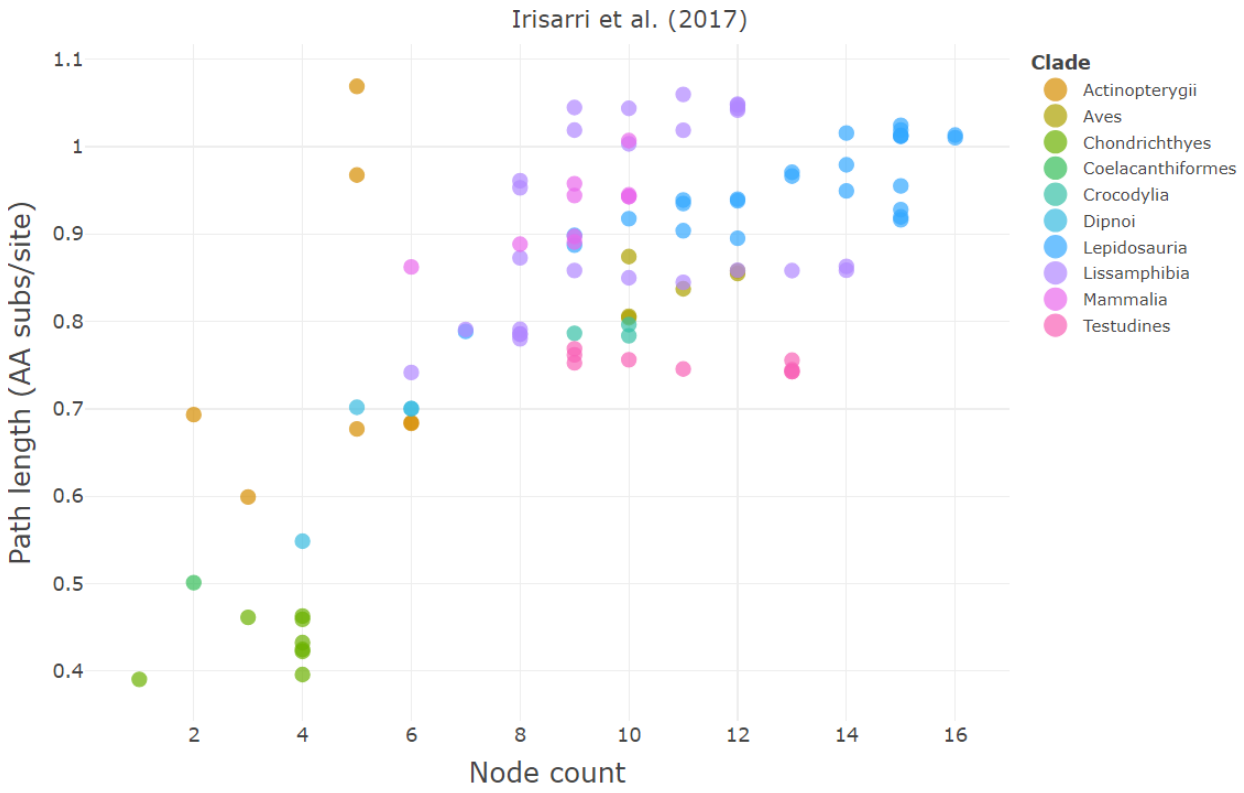

Figure 19. Node count seems to correlate with the phylogenetic path length in Irisarri et al.'s[34] tree, as expected from the hypothesis of punctuated evolution. The relationship differs by clade. Lissamphibians (frogs, salamanders, etc.) may have a flatter fit line than the rest, while testudines (turtles) exhibit an inexplicable negative relationship. See the interactive version in the supplemental files.

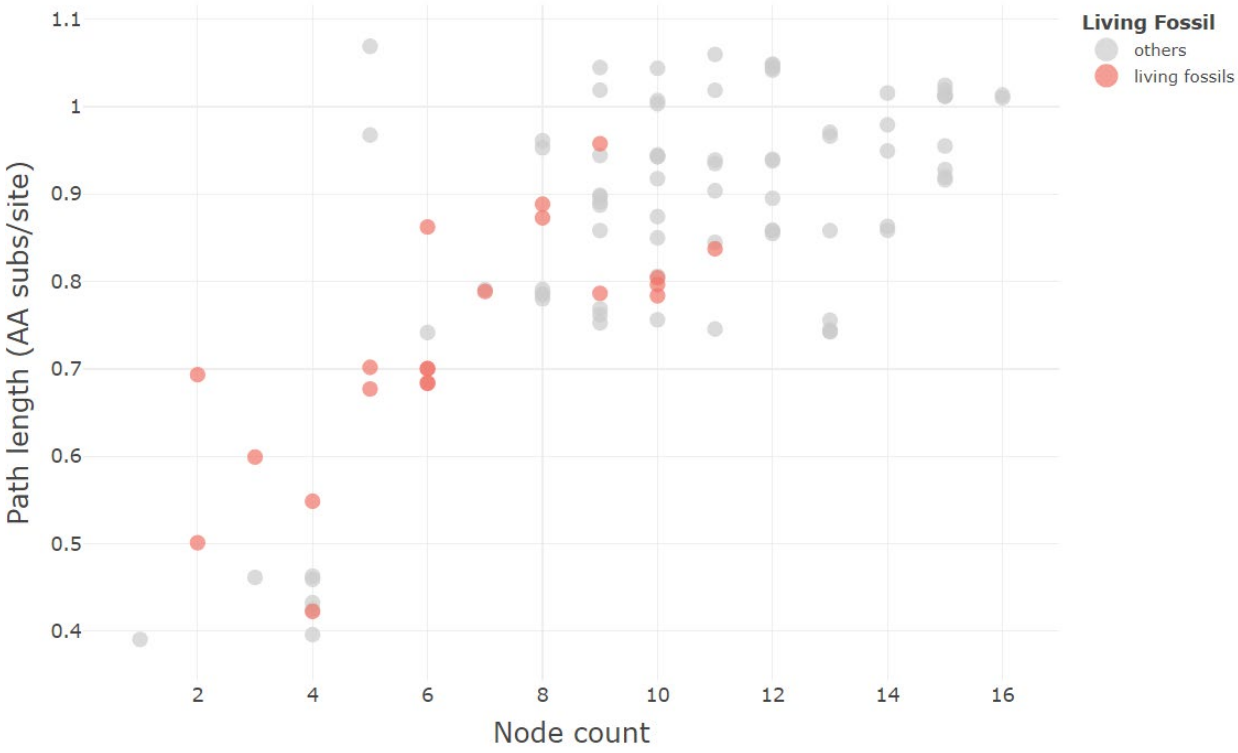

Figure 20. Node count seems to correlate with the phylogenetic path length in Irisarri et al.'s[34] tree, as expected from the hypothesis of punctuated evolution. Incorporating the indicator variable 'living fossil' could help explain path length. See the interactive version in the supplemental files.

Based on these scatter plots (Figures 19-20), below are the models I considered,

- Model 1: $y_i = \beta_0 + \epsilon_i$
- Model 2: $y_i = \beta_0 + \beta_1 x_{i,node} + \epsilon_i$
- Model 3: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_2 x_{i,liss=1} + \beta_3 x_{i,node} x_{i,liss=1} + \epsilon_i$
- Model 4: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_4 x_{i,test=1} + \beta_5 x_{i,node} x_{i,test=1} + \epsilon_i$
- Model 5: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_2 x_{i,liss=1} + \beta_3 x_{i,node} x_{i,liss=1} + \beta_4 x_{i,test=1} + \beta_5 x_{i,node} x_{i,test=1} + \epsilon_i$
- Model 6: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_6 x_{i,LF=1} + \epsilon_i$

- Model 7: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_6 x_{i,LF=1} + \beta_7 x_{i,node} x_{i,LF=1} + \epsilon_i$
- Model 8: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_8 x_{i,time} + \epsilon_i$

where:

- y_i: the phylogenetic path length of the i^{th} species in the phylogenetic tree
- β_0: the predicted average path length when all other variables (if present) are zero
- β_1: the predicted difference in average path length between two species that differ by one node count, holding all other variables constant (if applicable)
- β_2: the predicted difference in average path length between amphibians and other species, holding all other variables constant (if applicable)
- β_3, β_5, β_7: the predicted difference in slope for the node count comparing groups (interaction term)
- β_4: the predicted difference in average path length between turtles and other species, holding all other variables constant (if applicable)
- β_6: the predicted difference in average path length between non-living-fossil and living fossil species, holding all other variables constant (if applicable)
- β_8: the predicted difference in average path length between two species that differ by one million years in the terminal time branch length, holding all other variables constant (if applicable)
- $x_{i,node}$: the node count of the i^{th} species
- $x_{i,liss=1}$: an indicator variable of whether the i^{th} species is a lissamphibian (1: yes; 0: no)
- $x_{i,test=1}$: an indicator variable of whether the i^{th} species is a turtle (1: yes; 0: no)
- $x_{i,LF=1}$: an indicator variable of whether the i^{th} species is a living fossil (1: yes; 0: no)
- $x_{i,time}$: the terminal time branch length of the i^{th} species
- ϵ_i: the error term for the i^{th} species

We then fitted each phylogenetic model. I checked for MCMC convergence of each model using Gelman and Rubin's[86] potential scale reduction statistic $\hat{R}$. This statistic is "the ratio of the average variance of the draws within each chain to the variance of the pooled draws across chains"[86]. So, a convergence is indicated by $\hat{R} \approx 1$, which is the case for all parameters we estimated here. Moreover, we generated a posterior distribution of Bayesian R^2 values for each model and took the median[87]. One should not compare Bayesian R^2 values across models since the denominator in the Bayesian R^2 calculation changes with every model, unlike with the frequentist R^2. Instead, tracking the increase in model fit should be done through the decline in σ^2.

We conducted posterior predictive checks (Figure 21) to diagnose how well the distributions of normalized path lengths simulated from each model mimic that of the observed.

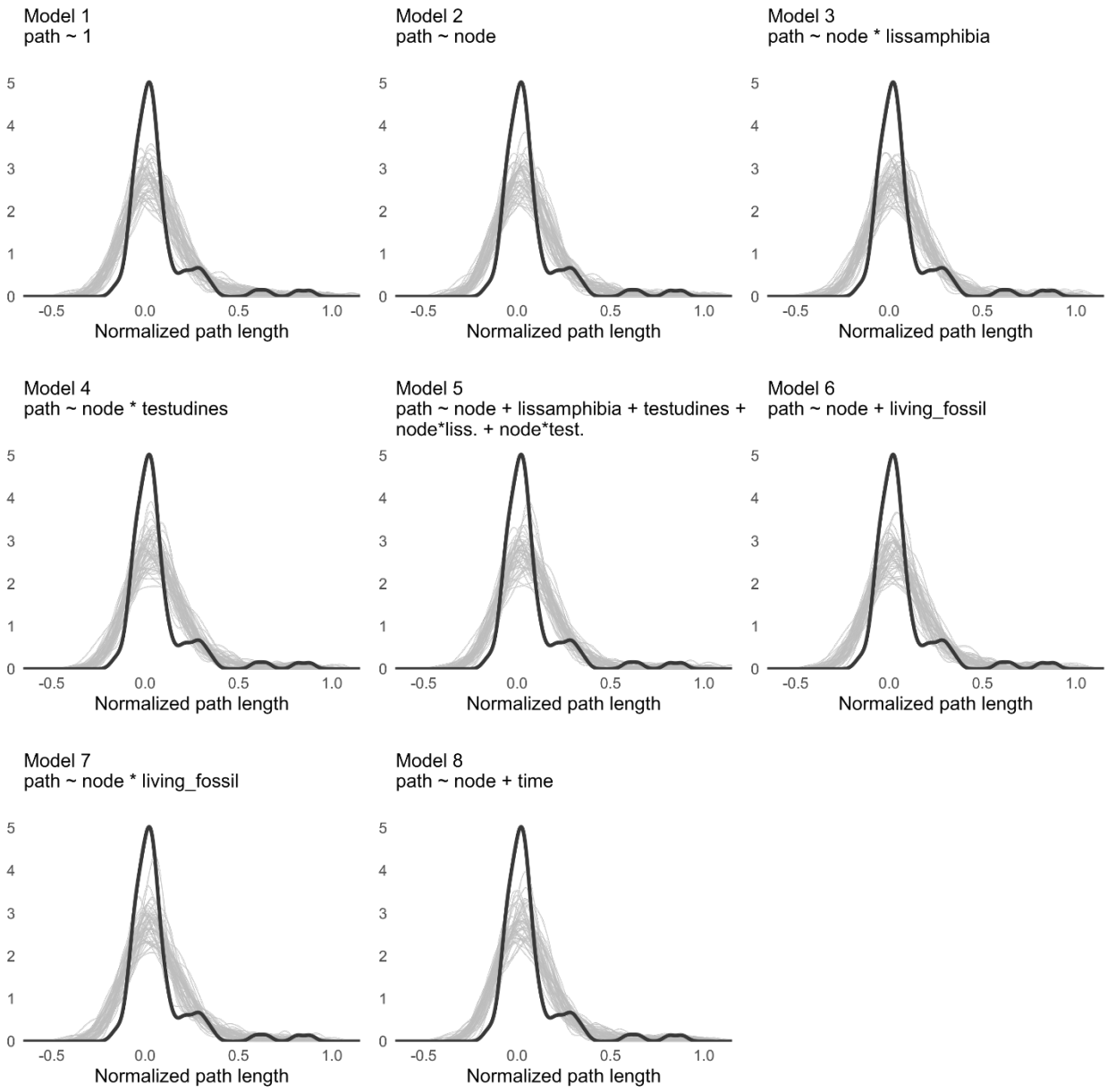

Figure 21. The distributions of posterior predicted normalized path lengths across the eight models of Irisarri et al.'s[34] dataset roughly resemble the observed distribution, though not as peaky. The black line is the observed, and the gray lines are the simulated distributions. Note that the normalized path length distribution is not the same as the distribution at the original scale (amino acids substitutions per site).

We then compared the predictive abilities of the eight models using a Leave-One-Out (LOO) cross-validation[88,89]. For models 4, 5, and 8, we had to run the LOO using a Pareto shape

parameter *k* threshold of 0.7[88,89]. A couple of observations have likely had a high influence on the posterior distributions in these models. Best-fitting models (among the set) have the highest expected log predictive density (ELPD[90]). ELPD incorporates uncertainty in the predictions. The rule-of-thumb for interpreting differences in ELPD scores is similar to differences in AIC or BIC scores[91].

Though, we did not compare all eight models simultaneously.

Table 7. Comparisons of models 1, 2, 6, and 7 (Irisarri et al.'s[34] dataset).

Model	ELPD difference	SE of the difference	95% credible interval
Model 2	0.0		
Model 7	-0.1	2.3	(-4.7, 4.5)
Model 6	-0.2	0.8	(-1.7, 1.3)
Model 1	-4.1	3.1	(-10.3, 2.0)

The first set (models 1, 2, 6, and 7; Table 7) is a series of nested models involving node count and the indicator variable 'living fossil' as predictors of path length. Based on the ELPD score, the best-fitting model (among the set) has node count as the sole predictor (ELPD difference = 0). Models 6 and 7, which have the 'living fossil' variable forming same-slope and separate-slopes models (respectively), trail closely behind (mean ELPD difference < 2). Models 2, 6, and 7 have similar predictive capabilities, which indicate that the living fossil status designation does not help much in explaining path length. This result echoes the outcome of the phylogenetic "*t*-test." The intercept-only model 1 is the worst-fitting among the set (mean ELPD difference > 2).

So, what does this model selection tell us biologically? Model 2 is the classic punctuated evolution model[49,50], where node count predicts path length. Vertebrate species that paleontologists classified as living fossils based on anatomy have slowly evolving protein-coding

sequences (low path length). However, this genomic stasis is as expected from the punctuational model (low node count). If we were to venture into the causal manner of writing, the node count effect is not much different among the so-called living fossil species (no evidence to support a separate-slopes model).

Table 8. Comparisons of models 1, 2, 3, 4, 5, and 8 (Irisarri et al.'s[34] dataset).

Model	ELPD difference	SE of the difference	95% credible interval
Model 4	0.0		
Model 5	-1.6	3.1	(-7.7, 4.5)
Model 2	-4.9	3.4	(-11.7, 1.8)
Model 8	-4.9	4.0	(-12.9, 3.0)
Model 3	-6.3	5.1	(-16.5, 3.8)
Model 1	-9.1	4.8	(-18.7, 0.6)

Now, we expand on model 2 by comparing it with its more complex counterparts (models 3, 4, 5, and 8). Model 4 is the best-fitting one among the set, which allows a separate slope for Testudines (turtles). Model 5 allows Lissamphibia (amphibians) to have a different path length-node count relationship, but this extra complexity seems unwarranted (mean ELPD difference = -1.85). Based on the 95% uncertainty intervals of the ELPD differences, models 2, 3, 4, 5, and 8 may not be that indistinguishable in their predictive capability. However, we stick with selecting model 4 since it has the highest ELPD score.

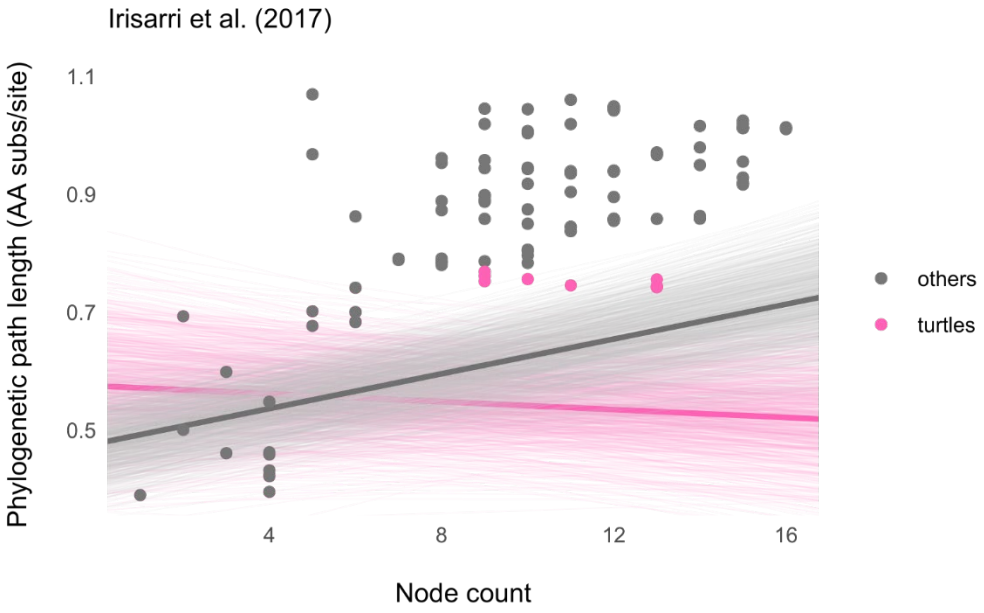

Figure 22. Vertebrate species with slowly evolving protein-coding sequences have, on average, speciated less than their faster-evolving counterparts. The regression fit lines are off from the data points because we accounted for species non-independence (i.e., normalized the data points against their evolutionary history). Plotting the normalized data points is impossible because the intercept column in the design matrix is no longer a vector of 1s.

There is some evidence that node count predicts path length. Among two non-turtle vertebrate species that differ by one net-speciation event, with a 95% probability, the one that has speciated more is expected to have a longer phylogenetic path by, on average, 0.008-0.020 amino acids (AA) substitutions per site. Turtles (testudines) exhibit some evidence for a weaker relationship between path length and node count, if not a negative relationship, which is unheard of. If we pick two turtle species that differ by one node count, the species that have speciated more is predicted, with a 95% probability, to have a phylogenetic path between, on average, 0.014 AA subs/site shorter to 0.007 AA subs/site longer. The expected standard deviation is 0.108, meaning that the observed path length will be within 2 × 0.108 AA subs/site from the linear predictor approximately 95% of the time; relatively high unexplained variability. And the

residuals vs. predicted plot did not show evidence for missed curvature (see the supplemental files).

To put numbers in perspective, we used an equation in Pagel et al.[50] to estimate the punctuational contribution to the total molecular divergence (see the supplemental files). For non-turtles, there is a 95% probability that 13.5%-34.7% of the amino acids sequence divergence is attributable to punctuated evolution, while only 0%-11.5% for turtles (hardbound at 0%).

Morel et al. (2022)

Below are two scatter plots to help guide model selection for Morel et al.'s[35] dataset.

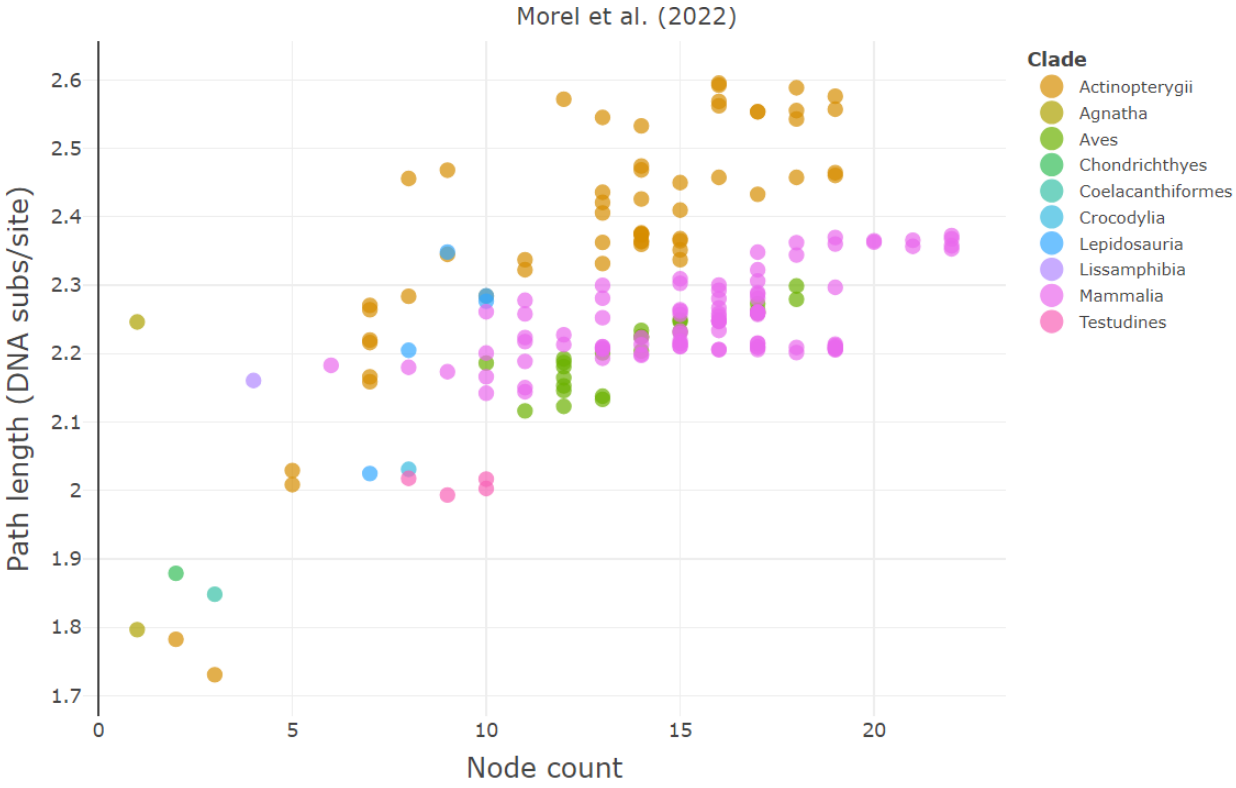

Figure 23. Node count seems to correlate with the phylogenetic path length in Morel et al.'s[35] tree, as expected from the hypothesis of punctuated evolution. The relationship differs by clade. Mammals may have a flatter fit line than the rest. See the interactive version in the supplemental files.

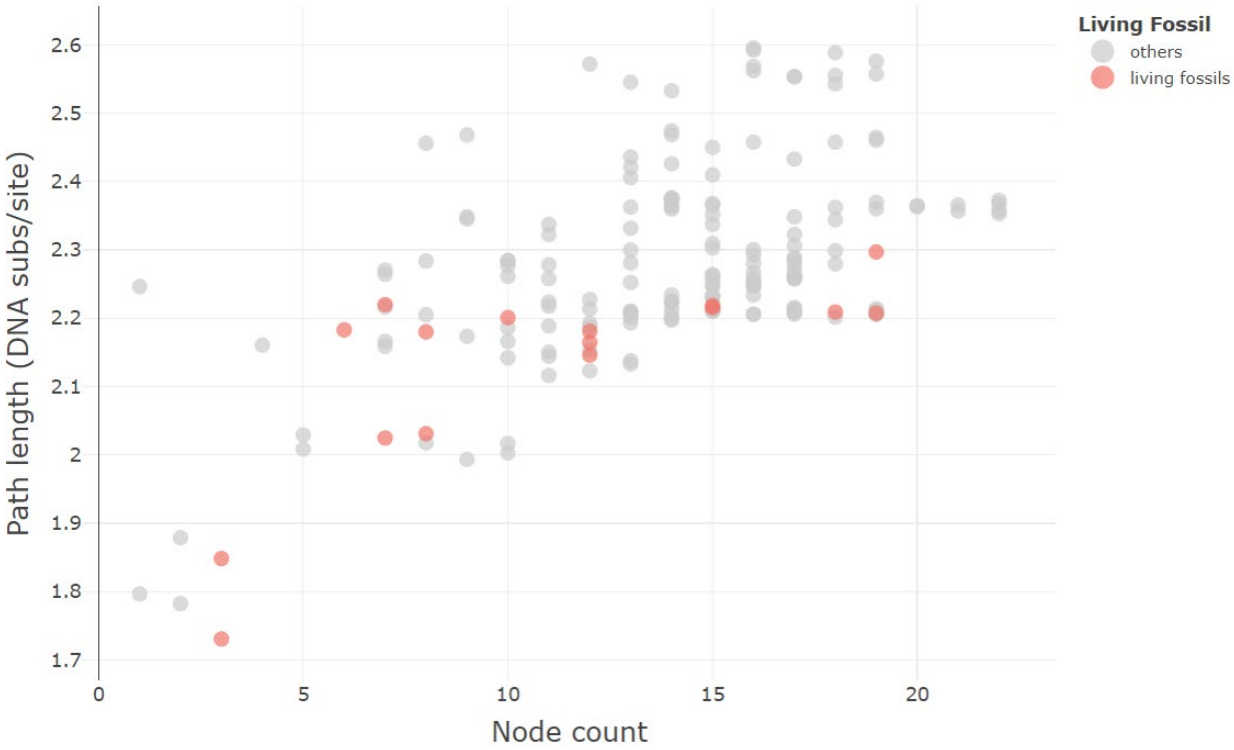

Figure 24. Node count seems to correlate with the phylogenetic path length in Morel et al.'s[35] tree, as expected from the hypothesis of punctuated evolution. Incorporating the indicator variable 'living fossil' could help explain path length. See the interactive version in the supplemental files.

Based on these scatter plots (Figures 23-24), below are the models I considered,

- Model 1: $y_i = \beta_0 + \epsilon_i$
- Model 2: $y_i = \beta_0 + \beta_1 x_{i,node} + \epsilon_i$
- Model 3: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_2 x_{i,mamm=1} + \beta_3 x_{i,node} x_{i,mamm=1} + \epsilon_i$
- Model 4: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_4 x_{i,LF=1} + \epsilon_i$
- Model 5: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_4 x_{i,LF=1} + \beta_5 x_{i,node} x_{i,LF=1} + \epsilon_i$
- Model 6: $y_i = \beta_0 + \beta_1 x_{i,node} + \beta_6 x_{i,time} + \epsilon_i$

where:

- y_i: the phylogenetic path length of the i^{th} species in the phylogenetic tree
- β_0: the predicted average path length when all other variables (if present) are zero
- β_1: the predicted difference in average path length between two species that differ by one node count, holding all other variables constant (if applicable)
- β_2: the predicted difference in average path length between mammals and other species, holding all other variables constant (if applicable)
- β_3, β_5: the predicted difference in slope for the node count comparing groups (interaction term)
- β_4: the predicted difference in average path length between non-living-fossil and living fossil species, holding all other variables constant (if applicable)
- β_6: the predicted difference in average path length between two species that differ by one million years in the terminal time branch length, holding all other variables constant (if applicable)
- $x_{i,node}$: the node count of the i^{th} species
- $x_{i,mamm=1}$: an indicator variable of whether the i^{th} species is a mammal (1: yes; 0: no)
- $x_{i,LF=1}$: an indicator variable of whether the i^{th} species is a living fossil (1: yes; 0: no)
- $x_{i,time}$: the terminal time branch length of the i^{th} species
- ϵ_i: the error term for the i^{th} species

Below are the plots for posterior predictive checks (Figure 25).

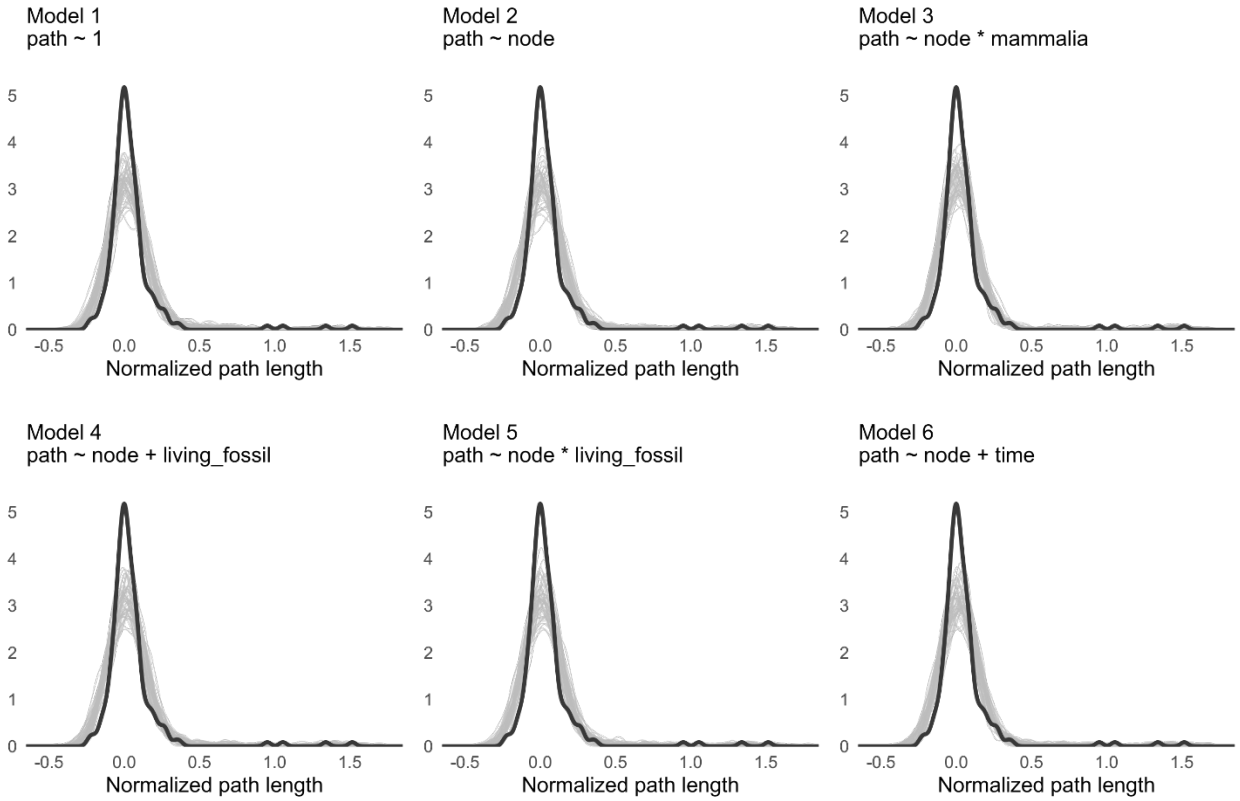

Figure 25. The distributions of posterior predicted normalized path lengths across the six models of Morel et al.'s[35] dataset roughly resemble the observed distribution, though not as peaky. The black line is the observed, and the gray lines are the simulated distributions. Note that the normalized path length distribution is not the same as the distribution at the original scale (DNA substitutions per site).

Table 9. Comparisons of models 1, 2, 4, and 5 (Morel et al.'s[35] dataset).

Model	ELPD difference	SE of the difference	95% credible interval
Model 2	0.0		
Model 4	-0.8	0.8	(-2.4, 0.7)
Model 5	-1.1	3.1	(-7.2, 5.1)
Model 1	-8.2	4.0	(-16.2, -0.1)

As before, this first set (models 1, 2, 4, and 5; Table 9) is a series of nested models involving node count and the indicator variable 'living fossil' as predictors of path length. Based on the ELPD score, the best-fitting model among the set has node count as the sole predictor

(ELPD difference = 0). Models 4 and 5 trail closely behind (mean ELPD difference < 2). Again, vertebrate species labeled as living fossils have path lengths expected from the punctuated evolution model.

Table 10. Comparisons of models 1, 2, 3, and 6 (Morel et al.'s[35] dataset).

Model	ELPD difference	SE of the difference	95% credible interval
Model 3	0.0		
Model 2	-7.9	5.5	(-19.0, 3.1)
Model 6	-8.3	4.2	(-16.7, 0.1)
Model 1	-16.1	8.2	(-32.6, 0.4)

Expanding on model 2, model 3 is the better fitting, allowing a separate slope and intercept for mammals (ELPD difference = 0).

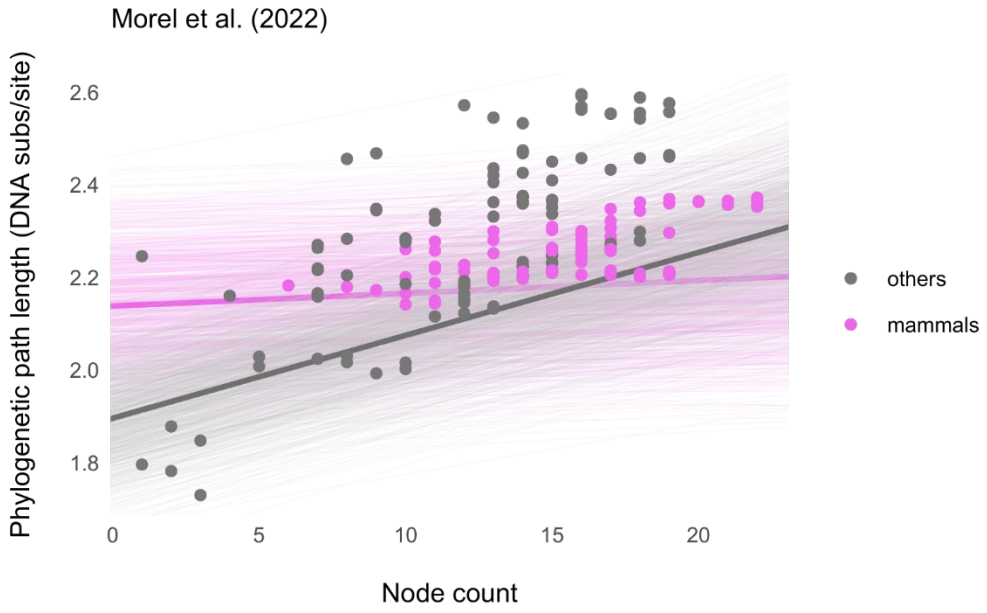

Figure 26. Vertebrate species with slowly evolving protein-coding sequences have, on average, speciated less than their faster-evolving counterparts. The regression fit lines are off from the data points because we accounted for species non-independence (i.e., normalized the data points against their evolutionary history). Plotting the normalized data points is impossible because the intercept column in the design matrix is no longer a vector of 1s.

The evidence, yet again, suggests that node count correlates with path length. Among two non-mammal vertebrate species that differ by one net-speciation event, with a 95% probability, the one that has speciated more is predicted to have a longer phylogenetic path by, on average, 0.012-0.024 DNA substitutions per site. The relationship between path length and node count seems weaker in mammals. This result could be due to the relatively slower rates of evolution in mammals than ray-finned fishes, both of which dominate Morel et al.'s[35] tree. If we pick two mammal species that differ by one node count, the species that have speciated more is expected, with a 95% probability, to have a phylogenetic path between, on average, 0.001 DNA subs/site shorter to 0.007 DNA subs/site longer. For non-mammals, there is a 95% probability that 14.2%-29.5% of the protein-coding sequence divergence is attributable to punctuated evolution, while only 0%-8.21% for mammals (hardbound at 0%). Morel et al.'s[35] tree only has four turtle species, so the weaker punctuational "effect" in this clade, as we inferred earlier, may not be apparent enough to warrant separate parameters. And there might not have been enough mammal species in Irisarri et al.'s[34] tree since we did not find visual evidence for a weaker "effect." The expected standard deviation is 0.103, meaning that the "average" distance between an observed path length from its expected value is 0.103 DNA subs/site; high unexplained variability. The residuals vs. predicted plot did not show evidence for missed curvature (see supplemental files).

Updated Phylogenetic Predictions

We can also assess the node count contribution to explaining path length by updating the posterior predictions. I expect species recognized as living fossils are not as much of outliers under the punctuated evolution model (after) as under the gradual evolution model (before). Below are the p_{MCMC} comparisons for Irisarri et al.'s[34] and Morel et al.'s[35] datasets.

Table 11. Posterior predictive distributions tend to capture the observed phylogenetic path lengths of the six species below slightly better (increased p_{MCMC}) when the model includes node count as a predictor.

Species	p_{MCMC} (mean-only) Irisarri	p_{MCMC} (+node count) Irisarri	p_{MCMC} (mean-only) Morel	p_{MCMC} (+node count) Morel
Latimeria chalumnae	0.01	0.04	< 0.01	0.03
Sphenodon punctatus	0.24	0.31	< 0.01	0.14
Monodelphis domestica	0.45	0.59	0.65	0.67
Lepisosteus oculatus	0.52	0.50	< 0.01	< 0.01
Homo sapiens	0.35	0.21	0.32	0.33
Takifugu rubripes	1.00	1.00	0.49	0.47

Overall, the p_{MCMC} values increased slightly from using models without node count (gradual) to those with node count (punctuational) (Table 11). However, we must not be too optimistic about the predictive capability of the punctuational model. The table above shows the inexplicable p_{MCMC} decrease for *Homo sapiens* (0.35 to 0.21), and node count does not seem to help much in predicting the path length of *Takifugu rubripes* in Irisarri et al.'s[34] dataset and *Lepisosteus oculatus* in Morel et al.'s[35] dataset. Therefore, the punctuational model does not fully account for the slowest (and fastest) evolving genomes, as was indicated by interpreting the parameter estimates alone.

Rabosky[57] and Pennell et al.[92] have criticized Pagel et al.'s[50] punctuated evolution model because the inverse interpretation is possible. What if node count is the response variable? It could be that species that have diverged more have had high evolvability and, thereby, have speciated more[93]. This idea arises from the hypothesis that natural selection acts at or above the species level[12,94], not just at the individual level. Lineages with a deep cauldron of genetic variation perhaps had more opportunities to branch out, occupying varied ecological niches.

However, the genetics of flagship living fossils suggests that the inverse interpretation is unlikely. Gould[56] recounted how Selander et al.[95] demonstrated, using the then-novel molecular technique of gel electrophoresis, that the genetic variability of horseshoe crabs (*Limulus*) is not any lower than their arthropod relatives. The coelacanth and tuatara genomes contain active transposable elements, selfish DNA sequences that can copy themselves and jump around the genome[5,96]. These elements can readily generate genetic variation. Some researchers have thought that slowly evolving organisms might have genetic or developmental constraints—e.g., changes in a gene would affect others in a delicate network[97,98]. Limitations need not be internal and can arise from the biotic environment through coevolution[99]. However, as numerous case studies have demonstrated, the above constraints are easily overcome[82].

Phylogenetic Principal Component Analysis (pPCA)

An exploratory, non-parametric approach to assess the correlation between the genomic sequence evolution and its potential covariates (node count and terminal time branch length) is the phylogenetic principal component analysis (PCA)[100]. The phylogenetic aspect means we account for the non-independence between species[100]. Using the phyl_pca function in the R package phytools v1.9.16[101], we extracted the principal components from the three variables (path length, node count, and terminal time). To prevent one variable dominating, we utilized the correlation matrix instead of the variance-covariance matrix (equivalent to extracting principal components from standardized variables).

Table 12. Eigenvectors used to calculate the principal components (Irisarri et al.'s[34] dataset).

	PC1	PC2	PC3
Path length	-0.61	0.79	0.02
Node count	-0.89	-0.29	0.35
Terminal time branch length	0.90	0.25	0.36

Table 13. Eigenvectors used to calculate the principal components (Morel et al.'s[35] dataset).

	PC1	PC2	PC3
Path length	0.58	0.81	0.00
Node count	0.88	-0.27	0.40
Terminal time branch length	-0.88	0.27	0.40

Tables 12-13 list the eigenvectors used to calculate the principal component scores (PC). Every PC is a linear combination of the original variables, and is uncorrelated to the other PCs. PC1 in Irisarri et al.'s (2017) dataset accounts for 66% of the total variability in the original variables, while PC2 captures 26%. Since PC1 and PC2 account for 92% of the variability, we could ignore PC3. The proportion of variance explained is similar for Morel et al.'s (2022) dataset (PC1: 62%; PC2: 27%). The eigenvectors for PC2 and PC3 are unintuitive. But, the signs of PC1 eigenvectors show that species with high PC1 values have lower net genomic evolution, lower cumulative net speciation events, and longer terminal time branches. This linear combination is consistent with the theory of punctuated evolution. So, we expect living fossil species to fall on the right side of the scale. The PC1 interpretation for Morel et al.'s (2022) dataset is similar, but the signs are reversed.

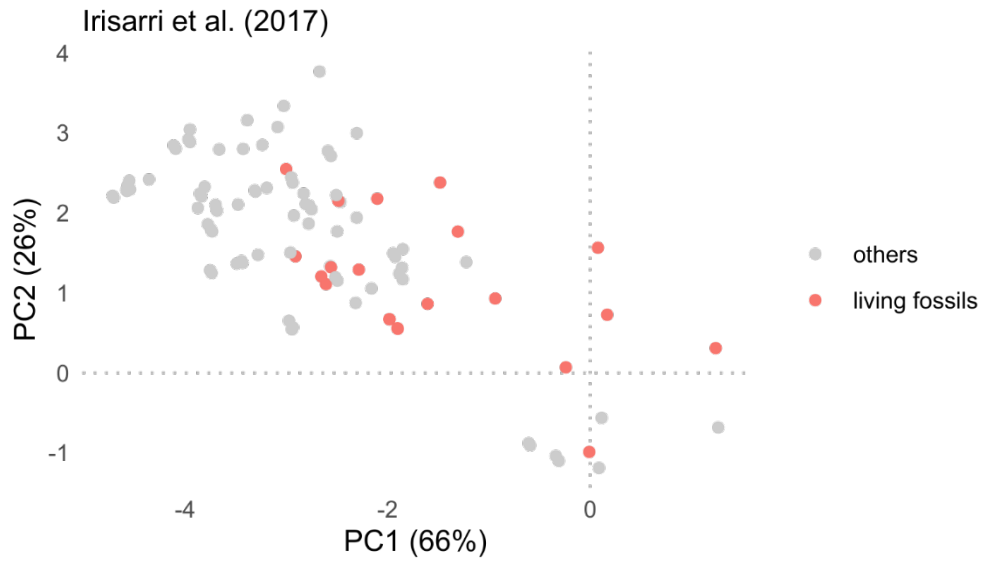

Figure 27. Species classified as living fossils a priori (red data points) do not form a distinct cluster on the principal component space (Irisarri et al.'s[34] dataset).

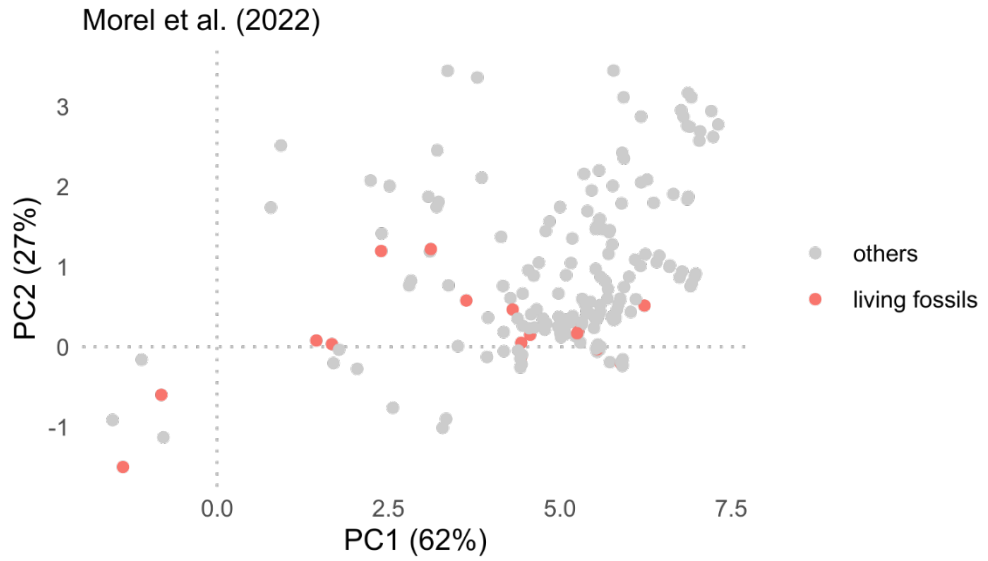

Figure 28. Species classified as living fossils a priori (red data points) do not form a distinct cluster on the principal component space (Morel et al.'s[35] dataset).

Figures 27-28 show the scatter plots of PC2 vs. PC1. No apparent clusters exist for species labeled as living fossils a priori due to the previously discussed subjectivity around

classifying living fossils. Still, PCA manages to compress three variables, with "living-fossilness" plotting to the left side and right side of Figure 27 and 28, respectively. Overall, the phylogenetic PCA is consistent with previous results.

Conclusion

Living fossils are fascinating creatures, racers who, sometime deep in the past, slowed down, have since not picked up the pace, yet have managed to survive to the present (analogy inspired by Simpson[11]). Let us place this analogy under Richard Dawkins' gene-centric view of life[102,103], where bodies are just the vehicles of the replicators, genes. We know from fossils that the cars of these sluggish racers have not changed much over geologic time. But what about the racers themselves? Here we analyzed living fossil genomes under one framework. Our method is flawed: sensitive to sampling strategies and dependent on the taxonomic scale of the dataset, which, in this study, is vertebrate. But with cases that fit the method, such as the African coelacanth (*Latimeria chalumnae*), we found their protein-coding sequences, at the DNA and amino acids levels, to have evolved much more slowly than expected from other species. Living fossils exemplify the coupling between the phenotype (vehicle) and its direct genetic source (replicator) over geologic time. If my method of phylogenetic prediction is not so flawed, it could have helped identify stasis in entities with little-to-no fossil records, such as microbes and viruses.

Next, why did these racers slow down in the first place? Here, we looked at the large-scale picture. As the established model of punctuated evolution predicts, net rates of the protein-coding sequence evolution follow cumulative net-speciation events. Slow genomic change in vertebrates is attributable to low net speciations, events typically associated with a subset of a

population branching off into a different environment. And living fossils, such as the coelacanth, fit this model to a certain extent. These sluggish racers are certainly not incapable of generating innovations (genetic variations) yet have not undergone as many innovation-associated events (net speciations) as their faster compatriots. The fossil record is filled with relatively few extinct relatives of living fossils[104,105], suggesting that lineage branchings did occur, but these relatives did not further diversify and survive. Genomic stasis and the scarcity of branching events hint that living fossil species may not be as adaptable as their contemporaries to changing environments, a serious concern in today's unprecedented climate change[106,107].

REFERENCES CITED

1. Amemiya, C. T. *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496**, 311–316 (2013).

2. Green, R. E. *et al.* Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**, 1254449 (2014).

3. Braasch, I. *et al.* The spotted gar genome illuminates vertebrate evolution and facilitates human-teleost comparisons. *Nat. Genet.* **48**, 427–437 (2016).

4. Du, K. *et al.* The sterlet sturgeon genome sequence and the mechanisms of segmental rediploidization. *Nat. Ecol. Evol.* **4**, 841–852 (2020).

5. Gemmell, N. J. *et al.* The tuatara genome reveals ancient features of amniote evolution. *Nature* **584**, 403–409 (2020).

6. Meyer, A. *et al.* Giant lungfish genome elucidates the conquest of land by vertebrates. *Nature* **590**, 284–289 (2021).

7. Thompson, A. W. *et al.* The bowfin genome illuminates the developmental evolution of ray-finned fishes. *Nat. Genet.* **53**, 1373–1384 (2021).

8. Zhou, Y. *et al.* Platypus and echidna genomes reveal mammalian biology and evolution. *Nature* **592**, 756–762 (2021).

9. Organ, C. L., Moreno, R. G. & Edwards, S. V. Three tiers of genome evolution in reptiles. *Integr. Comp. Biol.* **48**, 494–504 (2008).

10. Darwin, C. *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. (John Murray, 1859).

11. Simpson, G. G. *Tempo and Mode in Evolution*. (Columbia University Press, 1944).

12. Stanley, S. M. *Macroevolution, Pattern and Process*. (W. H. Freeman, 1979).

13. Eldredge, N. & Stanley, S. M. *Living Fossils*. (Springer, 1984). doi:10.1007/978-1-4613-8271-3.

14. Forchielli, A., Steiner, M., Hu, S., Lüter, C. & Keupp, H. Taphonomy of the earliest Cambrian linguliform brachiopods. *Acta Palaeontol. Pol.* **59**, 185–207 (2012).

15. Smith, J. L. B. A living fish of Mesozoic type. *Nature* **143**, 455–456 (1939).

16. Westoll, T. S. On the evolution of the Dipnoi. in *Genetics, Paleontology, and Evolution* (eds. Jepsen, G. L., Simpson, G. G. & Mayr, E.) 121–184 (Princeton University Press, 1949).

17. Bogert, C. M. The tuatara: Why is it a lone survivor? *Sci. Mon.* **76**, 163–170 (1953).

18. Cavin, L. & Kemp, A. The impact of fossils on the Evolutionary Distinctiveness and conservation status of the Australian lungfish. *Biol. Conserv.* **144**, 3140–3142 (2011).

19. Bennett, D. J., Sutton, M. D. & Turvey, S. T. Quantifying the living fossil concept. *Palaeontol. Electron.* **21**, 1–25 (2018).

20. Casane, D. & Laurenti, P. Why coelacanths are not 'living fossils'. *BioEssays* **35**, 332–338 (2013).

21. Grandcolas, P., Nattier, R. & Trewick, S. Relict species: A relict concept? *Trends Ecol. Evol.* **29**, 655–663 (2014).

22. Felsenstein, J. Phylogenies and the comparative method. *Am. Nat.* **125**, 1–15 (1985).

23. Grafen, A. The phylogenetic regression. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **326**, 119–157 (1989).

24. Symonds, M. R. E. & Blomberg, S. P. A Primer on Phylogenetic Generalised Least Squares. in *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice* (ed. Garamszegi, L. Z.) 105–130 (Springer, 2014). doi:10.1007/978-3-662-43550-2_5.

25. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).

26. Freckleton, R. P., Harvey, P. H. & Pagel, M. Phylogenetic analysis and comparative data: A test and review of evidence. *Am. Nat.* **160**, 712–726 (2002).

27. Harvey, P. H. & Pagel, M. D. *The Comparative Method in Evolutionary Biology*. (Oxford University Press, 1991).

28. Pagel, M. Inferring the historical patterns of biological evolution. *Nature* **401**, 877–884 (1999).

29. Organ, C. L., Shedlock, A. M., Meade, A., Pagel, M. & Edwards, S. V. Origin of avian genome size and structure in non-avian dinosaurs. *Nature* **446**, 180–184 (2007).

30. Organ, C. L., Janes, D. E., Meade, A. & Pagel, M. Genotypic sex determination enabled adaptive radiations of extinct marine reptiles. *Nature* **461**, 389–392 (2009).

31. Organ, C., Nunn, C. L., Machanda, Z. & Wrangham, R. W. Phylogenetic rate shifts in feeding time during the evolution of *Homo*. *Proc. Natl. Acad. Sci.* **108**, 14555–14559 (2011).

32. Cornwell, W. & Nakagawa, S. Phylogenetic comparative methods. *Curr. Biol.* **27**, R333–R336 (2017).

33. Volz, E. M., Koelle, K. & Bedford, T. Viral phylodynamics. *PLOS Comput. Biol.* **9**, e1002947 (2013).

34. Irisarri, I. *et al.* Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat. Ecol. Evol.* **1**, 1370–1378 (2017).

35. Morel, B. *et al.* SpeciesRax: A tool for maximum likelihood species tree inference from gene family trees under duplication, transfer, and loss. *Mol. Biol. Evol.* **39**, msab365 (2022).

36. Lartillot, N. & Poujol, R. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. *Mol. Biol. Evol.* **28**, 729–744 (2011).

37. IUCN. The IUCN Red List of Threatened Species. Version 2022-2. *IUCN Red List of Threatened Species* https://www.iucnredlist.org/en (2023).

38. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

39. Morel, B., Kozlov, A. M. & Stamatakis, A. ParGenes: A tool for massively parallel model selection and phylogenetic tree inference on thousands of genes. *Bioinformatics* **35**, 1771–1773 (2019).

40. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

41. Venkatesh, B. *et al.* Elephant shark genome provides unique insights into gnathostome evolution. *Nature* **505**, 174–179 (2014).

42. Federhen, S. The NCBI Taxonomy database. *Nucleic Acids Res.* **40**, D136–D143 (2012).

43. Bouckaert, R. *et al.* BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLOS Comput. Biol.* **15**, e1006650 (2019).

44. Ronquist, F. *et al.* MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542 (2012).

45. Smith, S. A. & O'Meara, B. C. treePL: Divergence time estimation using penalized likelihood for large phylogenies. *Bioinformatics* **28**, 2689–2690 (2012).

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

47. Paradis, E. Molecular dating of phylogenies by likelihood methods: A comparison of models and a new information criterion. *Mol. Phylogenet. Evol.* **67**, 436–444 (2013).

48. Benton, M. J. *et al.* Constraints on the timescale of animal evolutionary history. *Palaeontol. Electron.* **18**, 1–106 (2015).

49. Webster, A. J., Payne, R. J. H. & Pagel, M. Molecular phylogenies link rates of evolution and speciation. *Science* **301**, 478–478 (2003).

50. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**, 119–121 (2006).

51. Freedman, D. & Diaconis, P. On the histogram as a density estimator: L_2 theory. *Z. Für Wahrscheinlichkeitstheorie Verwandte Geb.* **57**, 453–476 (1981).

52. Schilling, T. *et al.* Comparative genomics provides evidence for an ancient genome duplication event in fish. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **356**, 1661–1679 (2001).

53. Raup, D. *Extinction: Bad Genes or Bad Luck*. (W. W. Norton & Company, 1992).

54. Eldredge, N. & Gould, S. J. Punctuated equilibria: An alternative to phyletic gradualism. in *Models in Paleobiology* (ed. Schopf, T. J. M.) 82–115 (Freeman, Cooper, 1972).

55. Stanley, S. M. A theory of evolution above the species level. *Proc. Natl. Acad. Sci.* **72**, 646–650 (1975).

56. Gould, S. J. *The Structure of Evolutionary Theory*. (Belknap Press, 2002).

57. Rabosky, D. L. Positive correlation between diversification rates and phenotypic evolvability can mimic punctuated equilibrium on molecular phylogenies. *Evolution* **66**, 2622–2627 (2012).

58. Rabosky, D. L. *et al.* Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* **4**, 1958 (2013).

59. Bennett, D. J., Sutton, M. D. & Turvey, S. T. Evolutionarily distinct “living fossils” require both lower speciation and lower extinction rates. *Paleobiology* **43**, 34–48 (2016).

60. Kuderna, L. F. K. *et al.* A global catalog of whole-genome diversity from 233 primate species. *Science* **380**, 906–913 (2023).

61. Inoue, J. G., Miya, M., Venkatesh, B. & Nishida, M. The mitochondrial genome of Indonesian coelacanth Latimeria menadoensis (Sarcopterygii: Coelacanthiformes) and divergence time estimation between the two coelacanths. *Gene* **349**, 227–235 (2005).

62. Nagalingum, N. S. *et al.* Recent synchronous radiation of a living fossil. *Science* **334**, 796–799 (2011).

63. Brownstein, C. D., Yang, L., Friedman, M. & Near, T. J. Phylogenomics of the ancient and species-depauperate gars tracks 150 million years of continental fragmentation in the Northern Hemisphere. *Syst. Biol.* **72**, 213–227 (2023).

64. Brownstein, C. D., Harrington, R. C. & Near, T. J. The biogeography of extant lungfishes traces the breakup of Gondwana. *J. Biogeogr.* **50**, 1191–1198 (2023).

65. Venditti, C., Meade, A. & Pagel, M. Detecting the node-density artifact in phylogeny reconstruction. *Syst. Biol.* **55**, 637–643 (2006).

66. Colless, D. H. Review of phylogenetics: The theory and practice of phylogenetic systematics. *Syst. Zool.* **31**, 100–104 (1982).

67. Mooers, A. Ø. & Heard, S. B. Inferring evolutionary process from phylogenetic tree shape. *Q. Rev. Biol.* **72**, 31–54 (1997).

68. Blum, M. G. B. & François, O. Which random processes describe the tree of life? A large-scale study of phylogenetic tree imbalance. *Syst. Biol.* **55**, 685–691 (2006).

69. Alfaro, M. E. *et al.* Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proc. Natl. Acad. Sci.* **106**, 13410–13414 (2009).

70. Venditti, C., Meade, A. & Pagel, M. Multiple routes to mammalian diversity. *Nature* **479**, 393–396 (2011).

71. Landis, M. J. & Schraiber, J. G. Pulsed evolution shaped modern vertebrate body sizes. *Proc. Natl. Acad. Sci.* **114**, 13224–13229 (2017).

72. Freckleton, R. P. Fast likelihood calculations for comparative analyses. *Methods Ecol. Evol.* **3**, 940–947 (2012).

73. Goodrich, B., Gabry, J. & Brilleman, S. rstanarm: Bayesian applied regression modeling via Stan. (2023).

74. Kessy, A., Lewin, A. & Strimmer, K. Optimal whitening and decorrelation. *Am. Stat.* **72**, 309–314 (2018).

75. Gelman, A., Hill, J. & Vehtari, A. *Regression and Other Stories*. (Cambridge University Press, 2020). doi:10.1017/9781139161879.

76. Carpenter, B. *et al.* Stan: A probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).

77. Gelfand, A. E. & Smith, A. F. M. Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**, 398–409 (1990).

78. Garland, Jr., T. & Ives, A. R. Using the past to predict the present: Confidence intervals for regression equations in phylogenetic comparative methods. *Am. Nat.* **155**, 346–364 (2000).

79. Nunn, C. L. & Zhu, L. Phylogenetic Prediction to Identify "Evolutionary Singularities". in *Modern Phylogenetic Comparative Methods and Their Application in Evolutionary Biology: Concepts and Practice* (ed. Garamszegi, L. Z.) 481–514 (Springer, 2014). doi:10.1007/978-3-662-43550-2_21.

80. Meng, X.-L. Posterior predictive *p*-values. *Ann. Stat.* **22**, 1142–1160 (1994).

81. Mayr, E. Change of Genetic Environment and Evolution. in *Evolution as a Process* (eds. Huxley, J., Hardy, A. C. & Ford, E. B.) 157–180 (Allan & Unwin, 1954).

82. Eldredge, N. *et al.* The dynamics of evolutionary stasis. *Paleobiology* **31**, 133–145 (2005).

83. Uyeda, J. C., Hansen, T. F., Arnold, S. J. & Pienaar, J. The million-year wait for macroevolutionary bursts. *Proc. Natl. Acad. Sci.* **108**, 15908–15913 (2011).

84. May, R. M., Lawton, J. H. & Stork, N. E. Assessing Extinction Rates. in *Extinction Rates* (eds. May, R. M. & Lawton, J. H.) 1–24 (Oxford University Press, 1995).

85. Raup, D. Extinction: Bad genes or bad luck. *Acta Geol. Hisp.* **16**, 25–33 (1981).

86. Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992).

87. Gelman, A., Goodrich, B., Gabry, J. & Vehtari, A. R-squared for Bayesian regression models. *Am. Stat.* **73**, 307–309 (2019).

88. Vehtari, A., Gelman, A. & Gabry, J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Stat. Comput.* **27**, 1413–1432 (2017).

89. Vehtari, A., Simpson, D., Gelman, A., Yao, Y. & Gabry, J. Pareto smoothed importance sampling. Preprint at https://doi.org/10.48550/arXiv.1507.02646 (2022).

90. Gelman, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. *Stat. Comput.* **24**, 997–1016 (2014).

91. Burnham, K. P. & Anderson, D. R. Multimodel inference: Understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33**, 261–304 (2004).

92. Pennell, M. W., Harmon, L. J. & Uyeda, J. C. Is there room for punctuated equilibrium in macroevolution? *Trends Ecol. Evol.* **29**, 23–32 (2014).

93. Schopf, T. J. M. Rates of evolution and the notion of "living fossils". *Annu. Rev. Earth Planet. Sci.* **12**, 245–292 (1984).

94. Gould, S. J. Darwinism and the expansion of evolutionary theory. *Science* **216**, 380–387 (1982).

95. Selander, R. K., Yang, S. Y., Lewontin, R. C. & Johnson, W. E. Genetic variation in the horseshoe crab (*Limulus polyphemus*), a phylogenetic "relic". *Evolution* **24**, 402–414 (1970).

96. Naville, M., Chalopin, D., Casane, D., Laurenti, P. & Volff, J.-N. The coelacanth: Can a "living fossil" have active transposable elements in its genome? *Mob. Genet. Elem.* **5**, 55–59 (2015).

97. Hansen, T. F. & Houle, D. Evolvability, stabilizing selection, and the problem of stasis. in *Phenotypic Integration* (eds. Pigliucci, M. & Preston, K. A.) 130–150 (Oxford University Press, 2004).

98. Blows, M. W. & Hoffmann, A. A. A reassessment of genetic limits to evolutionary change. *Ecology* **86**, 1371–1384 (2005).

99. Mayr, E. *What Evolution Is*. (Basic Books, 2001).

100. Revell, L. J. Size-correction and principal components for interspecific comparative studies. *Evolution* **63**, 3258–3268 (2009).

101. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

102. Dawkins, R. *The Selfish Gene*. (Oxford University Press, 1976).

103. Dawkins, R. *The Extended Phenotype: The Gene as the Unit of Selection*. (Oxford University Press, 1983).

104. Friedman, M. & Coates, M. I. A newly recognized fossil coelacanth highlights the early morphological diversification of the clade. *Proc. R. Soc. B Biol. Sci.* **273**, 245–250 (2005).

105. Cavin, L., Piuz, A., Ferrante, C. & Guinot, G. Giant Mesozoic coelacanths (Osteichthyes, Actinistia) reveal high body size disparity decoupled from taxic diversity. *Sci. Rep.* **11**, 11812 (2021).

106. Cahill, A. E. *et al.* How does climate change cause extinction? *Proc. R. Soc. B Biol. Sci.* **280**, 20121890 (2013).

107. Ceballos, G. *et al.* Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Sci. Adv.* **1**, e1400253 (2015).