

Sampling the past to understand evolution

PhD in Ecology and Evolutionary Biology

School of Biological Sciences

Jacob D. Gardner

Submitted January 2023

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Chapter 1 is published as:

Gardner, Jacob D., and Chris L. Organ. 2021. Evolutionary Sample Size and Consilience in Phylogenetic Comparative Analysis. *Systematic Biology*, 70(5): 1061–1075.

Author contributions to this publication are as follows:

- Generated hypotheses: **JDG**, CLO
- Collected data and performed analyses: **JDG**
- Writing: **JDG** wrote the initial manuscript; CLO contributed to later drafts.

Chapter 2 is published as:

Gardner, Jacob D., Kevin Surya, and Chris L. Organ. 2019. Early Tetrapodomorph Biogeography: Controlling for Fossil Record Bias in Macroevolutionary Analyses. *Comptes Rendus Palevol*, 18(7): 609–709.

Author contributions to this publication are as follows:

- Generated hypotheses: **JDG**, KS, CLO
- Collected data and performed analyses: **JDG**, KS
- Writing: **JDG** wrote the initial manuscript; CLO, KS contributed to later drafts.

Jacob Gardner

Abstract

Computers have fundamentally advanced all fields of biology over the past 50 years. Within evolutionary biology, for example, researchers can now leverage computational approaches to detect macroevolutionary patterns with unprecedented rigour and objectivity using large phylogenies of species. But evolution is a historical process, and macroevolutionary hypotheses demand direct evidence from the deep past. The work presented in this thesis applies the latest phylogenetic comparative methods to fossil and simulated datasets and highlights how fossils from deep time inform our understanding of evolution. Sampling is central to all science, including comparative analyses, and can shape how we detect patterns and processes. Chapter 1 shows that a large effective (evolutionary) sample size of independent changes is crucial for accurately inferring rates of evolution and differentiating support among evolutionary hypotheses. Chapter 2 shows how disproportionate geopalaeontological sampling affects inferences of dispersal rates and ancestral geographic locations. I offer an innovative approach for assessing geographic sampling biases in the fossil record. The fossil record is also a window into diverse ecosystems and a wide range of global conditions. It thus offers unique and independent data sources for testing evolutionary hypotheses. Chapter 3 describes the utility of fossils for assessing general ecological principles based on extant taxa and modern climates. The study shows that Bergmann's rule does not extend to Mesozoic dinosaurs and mammals and that their poleward dispersal did not drive increases in body size. Chapter 4 describes an approach for directly quantifying change in functional equations along phylogenetic branches and applies it to the locomotor evolution of dinosaurs. The study reveals a close connection between the rate of locomotor evolution and speciation. Together, these works highlight the utility of the fossil record for informing evolutionary models and our understanding of evolution.

Acknowledgements

Palaeontology has been a passion of mine for as long as I can remember. Contributing to the world's knowledge on these extraordinary animals has been a dream come true. All of that and more is thanks to my loving family for nurturing this passion and supporting me every step of the way. Thank you, mom, dad, Savannah, Garet, Ross, and all my cousins, aunts, uncles, and grandparents for your love and support. That kid who was obsessed with *Jurassic Park* is now becoming a dino doctor!

I have made so many wonderful friends along this journey; luckily, for me, there's too many to name. But I would like to give a special thanks to my best friends and closest conspirers in science, Jack, Kevin, and Lauren. I couldn't have done this without you. (And quite literally because you're involved in 3/4s of this thesis!) Thank you to all the friends I made in Montana, China, and the UK for all the fun and adventures. Your friendship makes all this worth it. Extra thanks to my new colleagues at the University of Reading for letting this bushy-haired American talk your ear off about dinosaurs. To Bom and my fellow Meeps (Carolynne, Ciara, and LiLi), you especially have been the best support group I could have ever hoped for. Thank you!

None of this would have been possible without the support and inspiration of my teachers, professors, and mentors. Mr. Johnson, Ms. Toshach, Mr. Shorba, and Ms. Dixon, thank you for fostering my interests early on. Thank you, Jim Schmitt, Dave Lageson, and Dave Varricchio for your mentorship, inspiring me, and helping me become a scientist. And thanks to Qiaomei Fu, Dale Greenwalt, Julia Haggerty, Julie Hawkins, Jack Horner, Ellen Lamm, Michel Laurin, Matt Lavin, Andrew Meade, Scott Taylor, Chris Venditti, and Mark Wilson for all the advice and opportunities.

Last but certainly not least, all my gratitude goes to my super radical advisor and wise sensei, Chris Organ. I can't thank you enough for everything you've done for me, from teaching me all aspects of being a good scientist to being the coolest, funniest, and most sage mentor. I'm eternally grateful for the countless hours you've spent working with me and for all your guidance and support. You're the Bodhi to my Johnny Utah, the Mr. Miyagi to my Daniel-san. Thank you. Now, let's hit the waves!

Contents

Declaration	ii
Abstract	iii
Acknowledgements	iv
Introduction	1
References	7
Chapter 1	
Evolutionary Sample Size and Consilience in Phylogenetic Comparative	
Analysis	15
Abstract	15
Introduction	16
Materials and Methods	25
Results	31
Discussion	37
References	51
Appendix 1	60
Chapter 2	
Early Tetrapodomorph Biogeography: Controlling for Fossil Record Bias in	
Macroevolutionary Analyses	82
Abstract	82
Introduction	83
Materials and Methods	85
Results	96
Discussion	102
Conclusions	104

References	105
Appendix 1	114
Chapter 3	
Latitude Does Not Shape Body Size Evolution in Mammals or Dinosaurs	142
Abstract	142
Introduction	143
Results	145
Discussion	152
Methods	155
References	159
Appendix 1	166
Appendix 2	172
Chapter 4	
Dinosaur Diversity and Ecology was Driven by Limb Functional Evolution	179
Abstract	179
Introduction	180
Results and Discussion	180
References	191
Methods	199
Appendix 1	209
Appendix 2	216
Appendix 3	227
Summary	230
References	236

Introduction

“But just in proportion as this process of extermination has acted on an enormous scale, so must the number of intermediate varieties, which have formerly existed, be truly enormous.”

- Charles Darwin, 1859¹

Sampling the past

How do new species form? How do they change over time? And how do we explain the vast diversity of life around us? These questions are fundamental to our understanding of life on Earth and our place in it. Fossils play an integral role in answering these questions. The origin and extinction of species are most directly observed in the fossil record (Benton, 1995; Foote, 2003; Jablonski, 2004), along with evidence for novel adaptations (Shubin et al., 2009) and climatic conditions unseen in the modern world (Mannion et al., 2014). In his seminal work *On the Origin of Species*, Charles Darwin dedicated two chapters to the geological record and succession of extinct organisms (Darwin, 1859). Yet, Darwin was troubled by the incomplete nature of the fossil record and thought it to be the “most obvious and serious objection” to his theory of evolution (Darwin, 1859). If life evolved gradually from ancestor to descendant, then why does not every geological unit contain fossil intermediates? His answer was that the geological record was imperfect, and scientists using fossil data have since been forced to grapple with this reality. In the past few decades, sophisticated computational and statistical methods have transformed our ability to

¹ Darwin, C. (1859). *On the origin of species by means of natural selection*. London, UK: John Murray 1

analyse fossil data with greater efficiency and at unprecedented scales. However, Darwin's concern about the imperfect nature of the fossil record remains. How do we know whether our inferences are signals of an evolutionary process or merely biases in our data or methodologies?

In the last century, palaeontology has blossomed into a more quantitative field (Polly et al., 2016). Fossils are now treated as data points, providing new insights into the origin, evolution, and extinction of species. Throughout the 20th century, palaeontologists began to treat patterns observed in the fossil record as an accurate reflection of the pace and manner by which species evolved. In *Tempo and Mode in Evolution*, George G. Simpson used fossils to describe how rates (or tempo) of evolution can vary (Simpson, 1944). He proposed his theory of quantum evolution in which gaps in the fossil record are, in part, explained by rapid shifts to new adaptive zones—a set variation in which certain morphologies are adaptive in a species. Niles Eldredge and Stephen J. Gould noticed that species were often static and found that most morphological change occurred between species (Eldredge and Gould, 1972). Their theory of punctuated equilibria supposes that evolution is predominately driven by speciation.

Central to studying evolution is the comparative method (Pagel, 2000). Comparison forms the basis for how we distinguish organisms and their characteristics. For instance, does body size differ between herbivorous and carnivorous species? Such comparisons, however, require we account for shared ancestry. The more time two species share an evolutionary past, the more similar we expect them to be. In other words, comparative biological data are not independent, as assumed in common statistical tests. Joseph Felsenstein's (1985) paper on the non-independence of biological data was ground-breaking for comparative biology

and kickstarted the use of evolutionary trees (or phylogenies) as a statistical comparative framework. Phylogenetic comparative methods have since accelerated, with a wide selection of models for estimating rates of evolution (Eastman et al., 2011; Venditti et al., 2011), rates of speciation and extinction (Kubo and Iwasa, 1995; Beaulieu and O'Meara, 2015; Louca and Pennell, 2020), mode of evolution (Pagel, 1999; Pagel et al., 2006), correlated or dependent evolution (Pagel, 1994; Pagel and Meade, 2006), the states of long-extinct ancestors (Pagel et al., 2004), and even predicting unobserved traits in extinct organisms (Organ et al., 2009, 2007). The ideas on the tempo and mode of evolution brought forward by researchers like Eldredge, Gould, and Simpson can now be statistically tested across large phylogenies. For example, Pagel et al. (2006) proposed a phylogenetic test for punctuated evolution, where the sum of the branch lengths from the root to each tip is correlated with the number of nodes (or lineage-splitting events). Evidence for a correlation suggests that speciation explains the total amount of evolution among those taxa, akin to Eldredge and Gould's punctuated equilibria (Eldredge and Gould, 1972). However, the test by Pagel et al. allows for a gradation in the effect of speciation on evolution.

Each chapter of this thesis capitalises on recent advancements in phylogenetic comparative methods. In Chapter 1, we use Bayesian models for discrete-coded characters (Pagel, 1999; Pagel et al., 2006) to assess the role of sampling on estimated rates of evolution and inferences on correlated evolution (Gardner and Organ, 2021). In Chapter 2, we apply a variable-rates implementation of a biogeographic dispersal model (O'Donovan et al., 2018) and test the effects of disproportionate geographic sampling on estimated ancestral locations and dispersal rates (Gardner et al., 2019). In Chapters 3 & 4, we apply a recently developed variable-rates regression model (Baker et al., 2016; Baker and Venditti, 2019) to study the

geographic distribution of body size and evolution of locomotor lever arms in Mesozoic dinosaurs.

Biases in the fossil record

Questions over the quality of the fossil record have plagued scientists since Darwin's time. Stratigraphic and phylogenetic congruence suggests that the fossil record is quite good for a given geological stage and taxonomic family (Benton et al., 2000). However, several types of biases are widely known. Fossils are more likely to be preserved and recovered in younger geological strata, known as the "Pull of the Recent" bias (Jablonski et al., 2003). Uncertainty in stratigraphic ranges can also influence estimates of diversification and extinction rates (Raup and Boyajian, 1988; Signor and Lipps, 1982). These limitations and biases must be addressed when using fossil data to test hypotheses. Palaeontologists have several approaches to account for possible sampling biases, including subsampling methods (Alroy et al., 2001; Close et al., 2020; Dunne et al., 2018; Jablonski et al., 2003; Lloyd et al., 2012) and incorporating sampling bias proxies as covariates in regression analyses (Benson et al., 2010; Benson and Butler, 2011; Benton et al., 2013; Sakamoto et al., 2016). Formation count often tracks palaeobiodiversity closely through time (Benton et al., 2013). However, when the fossil record is sparse, formation count may be a poor predictor of diversity (Dunhill et al., 2014b, 2014a, 2013, 2012). Moreover, a causal relationship between formation count and palaeobiodiversity is uncertain. Confounding variables, such as sea level, may best explain their association (Benton et al., 2013; Dunhill et al., 2014b). Counter to expectation, palaeobiodiversity could drive the number of formations since fossil taxa are often used to define geologic strata

(Benton, 2015). Formation count is a particularly popular sampling bias proxy in phylogenetic comparative analyses (O'Donovan et al., 2018; Sakamoto et al., 2016; Tennant et al., 2016a, 2016b). However, geographic fossil sampling bias is often not considered in such analyses despite being well-documented (Benson and Upchurch, 2013; Vilhena and Smith, 2013). Disproportionate palaeontological sampling through space and time can influence inferences on ancestral geographic locations, dispersal rates, and the geographic distribution of traits. In Chapters 2 & 3, we apply new geographically informed sampling bias metrics to assess their effect on dispersal and the latitudinal distribution of body size.

Contributions of this thesis

This thesis details the ways in which sampling the past shapes our understanding of evolution. Chapters 1 & 2 explore how sampling biases influence the support for hypotheses on correlated evolution and dispersal rates. And Chapters 3 & 4 demonstrate how fossils can provide unique insights into biogeography, ecology, and the evolution of movement.

Chapter 1 shows how large effective (evolutionary) sample sizes of independent character state changes are crucial for accurately estimating rates of evolution and distinguishing hypotheses of correlated evolution (Gardner and Organ, 2021). There are many morphological traits, especially in fossil taxa, that originate along the same branch of a phylogenetic tree (Maddison and FitzJohn, 2015). Many are surely relics of cladistics, in which shared derived characters are prioritised; however, the lack of independent (convergent/parallel) evolution makes it challenging to statistically test for correlated evolution. We propose new metrics for assessing the

suitability of phylogenetic comparative methods in such cases and offer recommendations for mitigating biased assessments of correlated evolution.

The effects of spatial sampling biases have long been recognized (Benson and Upchurch, 2013; Vilhena and Smith, 2013); yet, studies estimating patterns of biogeographic dispersal fail to account for its effects. Chapter 2 demonstrates how disproportionate fossil sampling across space and time affects estimates on dispersal rates and ancestral geographic locations (Gardner et al., 2019). We propose a new approach for assessing geographic sampling biases and apply it to a case study on fossil tetrapodomorphs.

The latitudinal diversity of species is known to change with climate over deep time (Mannion et al., 2014). Yet, many ecological ‘rules’ are based almost entirely on present-day diversity (Allen, 1877; Bergmann, 1847). Bergmann’s rule states that body size increases with latitude and temperature as an adaptation for retaining heat in endothermic species (Bergmann, 1847; Blackburn et al., 1999; Meiri, 2011). However, tests for Bergmann’s rule have been almost strictly conducted on extant animals and are often selective in terms of which taxa support the rule. Using models that allow variable rates of body size evolution, Chapter 3 reveals that Bergmann’s rule does not extend to Mesozoic dinosaurs and mammals and that their poleward dispersal did not drive increases in body size. Moreover, applying our models to a large dataset of extant mammals, the study finds little evidence for Bergmann’s rule as a driver of body size evolution in mammals.

Chapter 4 describes a new approach for directly quantifying change in functional equations along phylogenetic branches and applies it to the locomotor evolution of dinosaurs. Dinosaurs exhibited an enormous range of sizes and forms of

locomotion (Brusatte et al., 2008; Carrano, 2000), including innovations in flight, quadrupedality, and gigantism, making them an ideal test case for understanding how locomotion evolves. The study applies variable-rates regression models to the parameters of locomotor equations and uncovers the complex ways in which underlying parts of a system can interact to produce enormous changes in function. The study also demonstrates how bouts of functional evolution can coincide with innovations in locomotion and may be driven by speciation.

References

- Allen, J.A., 1877. The influence of physical conditions in the genesis of species. *The Radical Review* 1, 108–140.
- Alroy, J., Marshall, C.R., Bambach, R.K., Bezusko, K., Foote, M., Fürsich, F.T., Hansen, T.A., Holland, S.M., Ivany, L.C., Jablonski, D., Jacobs, D.K., Jones, D.C., Kosnik, M.A., Lidgard, S., Low, S., Miller, A.I., Novack-Gottshall, P.M., Olszewski, T.D., Patzkowsky, M.E., Raup, D.M., Roy, K., Sepkoski, J.J., Sommers, M.G., Wagner, P.J., Webber, A., 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences* 98, 6261–6266.
<https://doi.org/10.1073/pnas.111144698>
- Baker, J., Meade, A., Pagel, M., Venditti, C., 2016. Positive phenotypic selection inferred from phylogenies. *Biological Journal of the Linnean Society* 118, 95–115. <https://doi.org/10.1111/bij.12649>
- Baker, J., Venditti, C., 2019. Rapid change in mammalian eye shape is explained by activity pattern. *Current Biology* 29, 1082-1088.e3.
<https://doi.org/10.1016/j.cub.2019.02.017>

- Beaulieu, J.M., O'Meara, B.C., 2015. Extinction can be estimated from moderately sized molecular phylogenies. *Evolution* 69, 1036–1043.
<https://doi.org/10.1111/evo.12614>
- Benson, R.B.J., Butler, R.J., 2011. Uncovering the diversification history of marine tetrapods: Ecology influences the effect of geological sampling biases. Geological Society, London, Special Publications 358, 191–208.
<https://doi.org/10.1144/SP358.13>
- Benson, R.B.J., Butler, R.J., Lindgren, J., Smith, A.S., 2010. Mesozoic marine tetrapod diversity: mass extinctions and temporal heterogeneity in geological megabiases affecting vertebrates. *Proceedings of the Royal Society B: Biological Sciences* 277, 829–834. <https://doi.org/10.1098/rspb.2009.1845>
- Benson, R.B.J., Upchurch, P., 2013. Diversity trends in the establishment of terrestrial vertebrate ecosystems: Interactions between spatial and temporal sampling biases. *Geology* 41, 43–46. <https://doi.org/10.1130/G33543.1>
- Benton, M.J., 2015. Palaeodiversity and formation counts: Redundancy or bias? *Palaeontology* 58, 1003–1029. <https://doi.org/10.1111/pala.12191>
- Benton, M.J., 1995. Diversification and extinction in the history of life. *Science* 268, 52–58. <https://doi.org/10.1126/science.7701342>
- Benton, M.J., Ruta, M., Dunhill, A.M., Sakamoto, M., 2013. The first half of tetrapod evolution, sampling proxies, and fossil record quality. *Palaeogeography, Palaeoclimatology, Palaeoecology* 372, 18–41.
<https://doi.org/10.1016/j.palaeo.2012.09.005>
- Benton, M.J., Wills, M.A., Hitchin, R., 2000. Quality of the fossil record through time. *Nature* 403, 534–537. <https://doi.org/10.1038/35000558>

- Bergmann, C., 1847. Über die verhältnisse der wärmeökonomie der thiere zu ihrer grösse. Göttinger Studien 3, 595–708.
- Blackburn, T.M., Gaston, K.J., Loder, N., 1999. Geographic gradients in body size: A clarification of Bergmann's rule. Diversity and Distributions 5, 165–174.
<https://doi.org/10.1046/j.1472-4642.1999.00046.x>
- Brusatte, S.L., Benton, M.J., Ruta, M., Lloyd, G.T., 2008. Superiority, competition, and opportunism in the evolutionary radiation of dinosaurs. Science 321, 1485–1488. <https://doi.org/10.1126/science.1161833>
- Carrano, M.T., 2000. Homoplasy and the evolution of dinosaur locomotion. Paleobiology 26, 489–512. [https://doi.org/10.1666/0094-8373\(2000\)026%3C0489:HATEOD%3E2.0.CO;2](https://doi.org/10.1666/0094-8373(2000)026%3C0489:HATEOD%3E2.0.CO;2)
- Close, R.A., Benson, R.B.J., Alroy, J., Carrano, M.T., Cleary, T.J., Dunne, E.M., Mannion, P.D., Uhen, M.D., Butler, R.J., 2020. The apparent exponential radiation of Phanerozoic land vertebrates is an artefact of spatial sampling biases. Proceedings of the Royal Society B: Biological Sciences 287, 20200372. <https://doi.org/10.1098/rspb.2020.0372>
- Darwin, C., 1859. On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life. London: John Murray, 1859.
- Dunhill, A.M., Benton, M.J., Newell, A.J., Twitchett, R.J., 2013. Completeness of the fossil record and the validity of sampling proxies: A case study from the Triassic of England and Wales. Journal of the Geological Society 170, 291–300. <https://doi.org/10.1144/jgs2012-025>
- Dunhill, A.M., Benton, M.J., Twitchett, R.J., Newell, A.J., 2014a. Testing the fossil record: Sampling proxies and scaling in the British Triassic–Jurassic.

- Palaeogeography, Palaeoclimatology, Palaeoecology 404, 1–11.
<https://doi.org/10.1016/j.palaeo.2014.03.026>
- Dunhill, A.M., Benton, M.J., Twitchett, R.J., Newell, A.J., 2012. Completeness of the fossil record and the validity of sampling proxies at outcrop level. *Palaeontology* 55, 1155–1175. <https://doi.org/10.1111/j.1475-4983.2012.01149.x>
- Dunhill, A.M., Hannisdal, B., Benton, M.J., 2014b. Disentangling rock record bias and common-cause from redundancy in the British fossil record. *Nature Communications* 5, 1–9. <https://doi.org/10.1038/ncomms5818>
- Dunne, E.M., Close, R.A., Button, D.J., Brocklehurst, N., Cashmore, D.D., Lloyd, G.T., Butler, R.J., 2018. Diversity change during the rise of tetrapods and the impact of the ‘Carboniferous rainforest collapse.’ *Proceedings of the Royal Society B: Biological Sciences* 285, 20172730.
<https://doi.org/10.1098/rspb.2017.2730>
- Eastman, J.M., Alfaro, M.E., Joyce, P., Hipp, A.L., Harmon, L.J., 2011. A novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* 65, 3578–3589. <https://doi.org/10.1111/j.1558-5646.2011.01401.x>
- Eldredge, N., Gould, S.J., 1972. Punctuated equilibria: An alternative to phyletic gradualism, in: Schopf, T.J.M. (Ed.), *Models in Paleobiology*. Freeman, Cooper and Co., San Francisco, pp. 82–115.
- Felsenstein, J., 1985. Phylogenies and the comparative method. *The American Naturalist* 125, 1–15. <https://doi.org/10.1086/284325>

- Foote, M., 2003. Origination and extinction through the Phanerozoic: A new approach. *The Journal of Geology* 111, 125–148.
<https://doi.org/10.1086/345841>
- Gardner, J.D., Organ, C.L., 2021. Evolutionary sample size and consistency in phylogenetic comparative analysis. *Systematic Biology* 70, 1061–1075.
<https://doi.org/10.1093/sysbio/syab017>
- Gardner, J.D., Surya, K., Organ, C.L., 2019. Early tetrapodomorph biogeography: Controlling for fossil record bias in macroevolutionary analyses. *Comptes Rendus Palevol* 18, 699–709. <https://doi.org/10.1016/j.crpv.2019.10.008>
- Jablonski, D., 2004. Extinction: Past and present. *Nature* 427, 589–589.
<https://doi.org/10.1038/427589a>
- Jablonski, D., Roy, K., Valentine, J.W., Price, R.M., Anderson, P.S., 2003. The impact of the pull of the recent on the history of marine diversity. *Science* 300, 1133–1135. <https://doi.org/10.1126/science.1083246>
- Kubo, T., Iwasa, Y., 1995. Inferring the rates of branching and extinction from molecular phylogenies. *Evolution* 49, 694–704. <https://doi.org/10.1111/j.1558-5646.1995.tb02306.x>
- Lloyd, G.T., Pearson, P.N., Young, J.R., Smith, A.B., 2012. Sampling bias and the fossil record of planktonic foraminifera on land and in the deep sea. *Paleobiology* 38, 569–584. <https://doi.org/10.1666/11041.1>
- Louca, S., Pennell, M.W., 2020. Extant timetrees are consistent with a myriad of diversification histories. *Nature* 580, 502–505. <https://doi.org/10.1038/s41586-020-2176-1>

- Maddison, W.P., FitzJohn, R.G., 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology* 64, 127–136. <https://doi.org/10.1093/sysbio/syu070>
- Mannion, P.D., Upchurch, P., Benson, R.B.J., Goswami, A., 2014. The latitudinal biodiversity gradient through deep time. *Trends in Ecology & Evolution* 29, 42–50. <https://doi.org/10.1016/j.tree.2013.09.012>
- Meiri, S., 2011. Bergmann's Rule – What's in a name? *Global Ecology and Biogeography* 20, 203–207. <https://doi.org/10.1111/j.1466-8238.2010.00577.x>
- O'Donovan, C., Meade, A., Venditti, C., 2018. Dinosaurs reveal the geographical signature of an evolutionary radiation. *Nature Ecology & Evolution* 2, 452. <https://doi.org/10.1038/s41559-017-0454-6>
- Organ, C.L., Janes, D.E., Meade, A., Pagel, M., 2009. Genotypic sex determination enabled adaptive radiations of extinct marine reptiles. *Nature* 461, 389–392. <https://doi.org/10.1038/nature08350>
- Organ, C.L., Shedlock, A.M., Meade, A., Pagel, M., Edwards, S.V., 2007. Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446, 180–184. <https://doi.org/10.1038/nature05621>
- Pagel, M., 2000. Statistical analysis of comparative data. *Trends in Ecology & Evolution* 15, 418. [https://doi.org/10.1016/S0169-5347\(00\)01952-2](https://doi.org/10.1016/S0169-5347(00)01952-2)
- Pagel, M., 1999. Inferring the historical patterns of biological evolution. *Nature* 401, 877–884. <https://doi.org/10.1038/44766>
- Pagel, M., 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences* 255, 37–45. <https://doi.org/10.1098/rspb.1994.0006>

- Pagel, M., Meade, A., 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist* 167, 808–825. <https://doi.org/10.1086/503444>
- Pagel, M., Meade, A., Barker, D., 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology* 53, 673–684. <https://doi.org/10.1080/10635150490522232>
- Pagel, M., Venditti, C., Meade, A., 2006. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* 314, 119–121. <https://doi.org/10.1126/science.1129647>
- Polly, P.D., Stayton, C.T., Dumont, E.R., Pierce, S.E., Rayfield, E.J., Angielczyk, K.D., 2016. Combining geometric morphometrics and finite element analysis with evolutionary modeling: Toward a synthesis. *Journal of Vertebrate Paleontology* 0, e1111225. <https://doi.org/10.1080/02724634.2016.1111225>
- Raup, D.M., Boyajian, G.E., 1988. Patterns of generic extinction in the fossil record. *Paleobiology* 14, 109–125. <https://doi.org/10.1017/S0094837300011866>
- Sakamoto, M., Benton, M.J., Venditti, C., 2016. Dinosaurs in decline tens of millions of years before their final extinction. *Proceedings of the National Academy of Science* 113, 5036–5040. <https://doi.org/10.1073/pnas.1521478113>
- Shubin, N., Tabin, C., Carroll, S., 2009. Deep homology and the origins of evolutionary novelty. *Nature* 457, 818–823. <https://doi.org/10.1038/nature07891>
- Signor, P.W., Lipps, J.H., 1982. Sampling bias, gradual extinction patterns, and catastrophes in the fossil record. *Geological Society of America Special Publication* 190, 291–296. <https://doi.org/10.1130/SPE190-p291>
- Simpson, G.G., 1944. *Tempo and mode in evolution*. Columbia University Press.

- Tennant, J.P., Mannion, P.D., Upchurch, P., 2016a. Environmental drivers of crocodyliform extinction across the Jurassic/Cretaceous transition. *Proceedings of the Royal Society B: Biological Sciences* 283, 20152840. <https://doi.org/10.1098/rspb.2015.2840>
- Tennant, J.P., Mannion, P.D., Upchurch, P., 2016b. Sea level regulated tetrapod diversity dynamics through the Jurassic/Cretaceous interval. *Nature Communications* 7, 12737. <https://doi.org/10.1038/ncomms12737>
- Venditti, C., Meade, A., Pagel, M., 2011. Multiple routes to mammalian diversity. *Nature* 479, 393–396. <https://doi.org/10.1038/nature10516>
- Vilhena, D.A., Smith, A.B., 2013. Spatial bias in the marine fossil record. *PLOS ONE* 8, e74470. <https://doi.org/10.1371/journal.pone.0074470>

Chapter 1

Evolutionary Sample Size and Consilience in Phylogenetic Comparative Analysis

(Published as: Gardner, Jacob D., and Chris L. Organ. 2021. Evolutionary Sample Size and Consilience in Phylogenetic Comparative Analysis. *Systematic Biology* 70(5): 1061—1075.)

Abstract

Phylogenetic comparative methods (PCMs) are commonly used to study evolution and adaptation. However, frequently used PCMs for discrete traits mishandle single evolutionary transitions. They erroneously detect correlated evolution in these situations. For example, hair and mammary glands cannot be said to have evolved in a correlated fashion because each evolved only once in mammals, but a commonly used model (Pagel's Discrete) statistically supports correlated (dependent) evolution. Using simulations, we find that rate parameter estimation, which is central for model selection, is poor in these scenarios due to small effective (evolutionary) sample sizes of independent character state change. Pagel's Discrete model also tends to favor dependent evolution in these scenarios, in part, because it forces evolution through state combinations unobserved in the tip data. This model prohibits simultaneous dual transitions along branches. Models with underlying continuous data distributions (e.g., Threshold and GLMM) are less prone to favor correlated evolution but are still susceptible when evolutionary sample sizes are small. We provide three general recommendations for researchers who encounter these common situations: i) create study designs that evaluate *a priori* hypotheses and maximize evolutionary sample sizes; ii) assess the suitability of evolutionary models—for discrete traits, we introduce

the phylogenetic imbalance ratio; and iii) evaluate evolutionary hypotheses with a consilience of evidence from disparate fields, like biogeography and developmental biology. Consilience plays a central role in hypothesis testing within the historical sciences where experiments are difficult or impossible to conduct, such as many hypotheses about correlated evolution. These recommendations are useful for investigations that employ any type of PCM.

Introduction

Over the past 40 years, biologists have capitalized on computational advances to study evolution with greater efficiency, rigor, and objectivity. Statistical phylogenetics is central to many of these endeavors, from macroecology to cancer genomics (Felsenstein 2003; Keith et al. 2012; Schwartz and Schaffer 2017). In addition to inferring taxonomic relationships, biologists routinely use phylogenies to design studies and analyze comparative datasets, a practice called phylogenetic comparative methods (PCMs) (Harvey and Pagel 1991; Garamszegi 2014; Harmon 2018). Joseph Felsenstein's 1985 paper on the nonindependence of biological data was a watershed moment for PCMs and a clarion call for many fields across biology. Felsenstein (1985) argued that, because of shared ancestry, comparative data and their associated errors are often phylogenetically structured—species with more recent common ancestors tend to have trait values more similar than those in distant relatives (Felsenstein 1985). Comparative data, therefore, violates assumptions of independence common in statistics.

Phylogenetic comparative methods have blossomed since, especially in the past 10 years, coincident with the rise of R as a ubiquitous statistical platform (R Core Team 2019). Researchers now use PCMs to analyze diverse datasets—from

genomes to languages to the fossil record—and model how traits evolved over time, often in association with other traits and abiotic factors (Harvey and Pagel 1991; Garamszegi 2014; Harmon 2018). These methods can help researchers model speciation and extinction rates through time (Nee et al. 1994; Kubo and Iwasa 1995; Morlon 2014; but see Rabosky 2010, and Louca and Pennell 2020 for critiques); reconstruct the ancestral traits of long-extinct common ancestors (Pagel 1999); and predict traits in extinct organisms (i.e., retrodiction [Organ et al. 2007, 2009, 2011]).

Exceptional evolutionary change along single lineages, such as brain size along the human lineage, is inherently interesting to biologists and can be rigorously studied with continuous data using a variety of approaches (McPeck 1995; Revell 2008; Eastman et al. 2011; Organ et al. 2011; Venditti et al. 2011; Uyeda et al. 2018). In general, these approaches compare trait values of target taxa against wider trait distributions. Problems arise, however, when common models for discrete (binary) trait data are applied in these scenarios. For instance, Maddison and FitzJohn (2015) found that Pagel's model for discrete characters (Pagel 1994) supports hypotheses of dependent (correlated) evolution when two binary (absent/present) characters evolve on the same phylogenetic branch and are never replicated (Maddison and FitzJohn 2015). An example would be the coincident evolution of hair and mammary glands in mammals. Maddison and FitzJohn (2015) argue that this cannot be taken as sufficient evidence for correlated evolution because coincident change happened only once; the effective or evolutionary sample size, representing the number of independent character state changes, is only one. Note that the evolutionary sample size may differ markedly from an apparently large sample of taxa. The number of taxa is often an overestimate of the evolutionary sample size, which could lead to overestimated degrees of freedom and inflated Type I error rates (Smith 1994).

Here we extend the work of Maddison and FitzJohn (2015) by showing that all common PCM models of correlated evolution for discrete characters are prone to support correlated evolution in these situations, especially those with underlying continuous model structures. We make three general recommendations for researchers using PCMs who regularly face these issues. First, we suggest that researchers design studies to evaluate *a priori* hypotheses that maximize evolutionary sample sizes (independent originations of trait states). Second, researchers should assess the suitability of their evolutionary model. We develop a metric that gauges the suitability of discrete character models by combining the consistency index and class imbalance ratio (Wang and Yao 2012). Third, we recommend that researchers employ a ‘consilience of inductions’ approach (Whewell 1840), a crucially important concept in the historical sciences, exemplified in Charles Darwin’s *Origin of Species* and other work (Darwin 1860; Ruse 1975; Thagard 1977).

Background: Discrete Character Models & Correlated Evolution

Before we discuss our simulations and recommendations, we provide a brief overview of PCMs for discrete characters (categorical; nominal or ordinal data). The effects that traits can have on the evolution of other traits has long interested biologists. The term ‘correlated evolution’ refers generally to the evolutionary change of one trait influencing that of another. After Felsenstein (1985), there was an explosion of PCMs aimed at testing for correlated evolution with both continuous and discrete traits (Maddison 1990; Pagel 1994, 1999; Huelsenbeck et al. 2003; Felsenstein 2005; Hadfield 2010). We recognize three classes of methods for testing correlated evolution, each with their own assumptions and lines of evidence that they consider. These include methods that test for: i) dependent evolution, in which the rate of change

in one character depends on the state of another (Maddison 1990; Pagel 1994); ii) shared character history, in which there is considerable overlap in the inferred evolutionary history of two character states (Huelsenbeck et al. 2003); and iii) correlated evolution *sensu stricto*, in which the variation of one trait covaries with that of another trait (Felsenstein 1985; Pagel 1999; Felsenstein 2005; Hadfield 2010). This last category also includes methods for continuous trait correlation and mixed models. We also use correlated evolution throughout the manuscript to refer to the association of two traits generally.

We first review the methods for testing dependent evolution. Maddison (1990) first developed a method, called the concentrated-changes test, that assesses whether changes in a character are more likely to be found on phylogenetic branches in association with changes in a second character. The method first infers the ancestral states of both characters separately for each node of the phylogeny, which provides an estimate on the number of gains (state 0 to 1) and losses (state 1 to 0). It then calculates the probability that the inferred changes in one character would, by chance, be as concentrated along branches given the state of a second character. A few notable limitations of this method are that it is not model based, it does not take branch lengths (amount of evolution or time) into account, and the inference of dependent evolution depends on the prior ancestral state inference.

Pagel (1994) developed a method that uses a continuous-time Markov model to characterize the evolution of binary characters (Pagel 1994). This model is a derivation of the Jukes-Cantor model, which assumes uniform transition rates among nucleotides (Jukes and Cantor 1969). However, rather than using transition rates to construct a phylogeny, the Markov model in Pagel's (1994) method estimates the transition rates between the states of two binary characters. The estimated rates

depend on the tree topology, branch lengths, and distribution of states across the tips. Unlike previous methods, Pagel's model treats ancestral states as random variables with some probability for each state of both characters at a given node. This solves the issue of relying on prior ancestral state inferences to test for dependent evolution. Importantly, this model assumes that dual character transitions cannot occur simultaneously; one character must evolve before the other. By comparing differences among rate parameters, we can assess whether the state of one character evolves dependently upon the state of another character. For example, Organ et al. (2009) found that live birth evolved only after genetic sex determination had previously evolved in amniotes.

To test for dependent evolution, two models are fitted: one where the transition rates of each character are independent and another where the transition rates of one character depend on the state of the other character (Fig. 1b). A likelihood ratio test can then be used to discriminate the better fitting model. A Bayesian reversible jump Markov chain Monte Carlo (rjMCMC) implementation of Pagel's (1994) method was also developed to simultaneously account for phylogenetic uncertainty and automatically reduce the number of transition rate parameters (Pagel and Meade 2006).

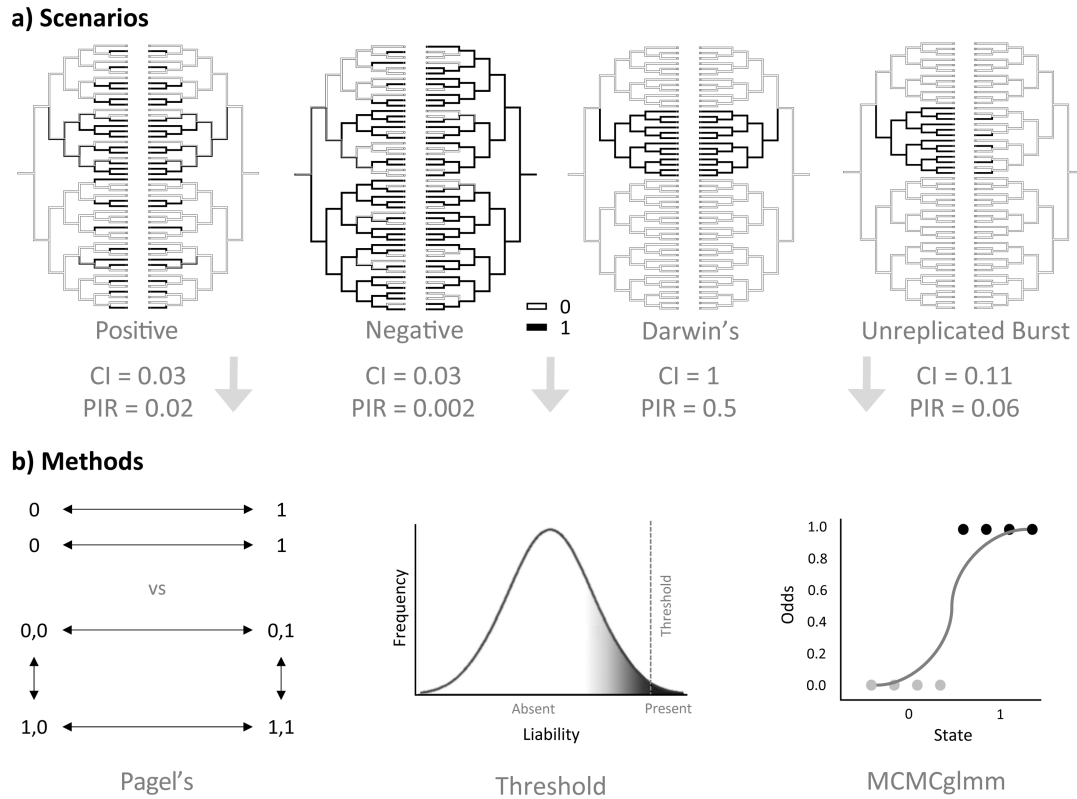


Figure 1. Discrete character scenarios and methods used for testing correlated evolution in this study. a) Character distributions of the four scenarios simulated; character presence is illustrated as black, and inner and outer boxes represent characters one and two. Average consistency index (CI) and phylogenetic imbalance ratio (PIR) are listed under each scenario. b) Figures illustrating the methods used in this study. Pagel's (1994) method showing the difference between the independent (top) and dependent (bottom) models with transition rates between possible character state combinations. The threshold figure illustrates the distribution of an unobserved liability underlying a binary character (absent/present). The MCMCgIimm figure illustrating a logistic regression where the odds of character one being present is a function of the second character's state.

Huelsenbeck et al.'s (2003) test for shared character history adapts a method used to stochastically map mutations on a phylogeny (Nielsen 2002). The method, like Pagel's (1994), uses a continuous-time Markov model to reconstruct the evolution of

multiple binary characters across a tree. The method differs, however, in its test for correlated evolution. Instead of assessing the dependence of character transition rates, Huelsenbeck et al.'s method tests whether the shared evolutionary history (measured in branch lengths or time) between two character states is significantly greater than expected assuming an independent model. Like Pagel's method, this method samples node states according to their estimated probability. A considerable overlap in the inferred character history between two character states suggests that their evolution was correlated. Bianchini and Sánchez-Baracaldo (2020) developed a program, called sMap, that also models discrete character evolution using stochastic mapping, but, instead, it tests for dependent evolution like Pagel's method. Unlike Pagel's method, it allows for simultaneous dual character state transitions and the ability to define conditional probabilities for specific character state transitions.

Discrete character evolution methods have also been developed to model transitions with underlying continuous variables and can allow for correlations between discrete and continuous traits. Felsenstein (2005, 2012) developed a phylogenetic threshold model based on work by Sewall Wright (Wright 1934; Felsenstein 1988, 2005, 2012; Fig. 1b). This model assumes that an unobserved continuous variable, called a liability, determines the state of a binary character across some threshold value (Felsenstein 2012). The liability attempts to characterize underlying mechanisms of discrete character evolution that vary continuously, such as gene expression levels or unobserved environmental factors. Testing for covariation between two discrete characters with the threshold model requires fewer parameters than rate-based Markov models (Felsenstein 2005, 2012). Another advantage of this model is its ability to test for correlated evolution between continuous and discrete

variables by estimating the covariance between a continuous variable and the liability of a discrete character.

Generalized linear mixed models (glmm) are extensions of general linear models, but they relax the assumption that residuals are normally distributed and allow for random effects. Hadfield (2010) developed an R package (MCMCglmm, Fig. 1b) that allows phylogenetic structures to be specified as random effects—essentially, removing the influence of phylogeny from the residual error. The phylogenetic structure is in the form of a variance-covariance matrix, representing the shared branch lengths of taxa. Any type of variance-covariance matrix can be incorporated, including those explaining pedigree relationships or geographic distances. A glmm relaxes the assumption of normality by transforming the response variable, allowing different data types to be modeled, such as binary, nominal, and ordinal data. For binary data, the response implements a logit transformation, amounting to a logistic regression. Like the threshold model, phylogenetic glmm assumes the presence of an unobserved continuously evolving parameter (the phylogenetically structured error), but it uses probabilities of change to dictate character state transitions instead of a threshold value. Other phylogenetic glmm implementations exist, like Ives and Garland's (2010) PGLMM, which differs in how it structures the random effects (Ives and Garland 2010, 2014). PGLMM also uses a frequentist framework, whereas MCMCglmm is Bayesian.

Framing the Problem

Recent work has shown that some PCMs reviewed above support models of correlated evolution in scenarios of single unreplicated evolutionary transitions (Maddison and FitzJohn 2015; Rabosky and Huang 2016; Uyeda et al. 2018). For

example, Maddison and FitzJohn (2015) showed that Maddison's (1990) and Pagel's (1994) methods support dependent evolution in a scenario where two binary characters change states on the same phylogenetic branch (which they dub the "Darwin's scenario"; Fig. 1a). The two characters originate only once in the same clade and are never lost or repeated; they are essentially synapomorphies that define the clade, and their co-occurrence may be coincidental. For example, Maddison and FitzJohn (2015) argue that trait distributions, like the coincident evolution of hair and mammary glands in mammals, lack biological evidence for correlated evolution precisely because each evolved only once. It is possible for two traits to evolve sequentially along a branch, and these may or may not be functionally or developmentally dependent, but we cannot falsify independent evolution because the change occurred on only one branch. The "unreplicated burst scenario", described by Maddison and FitzJohn (2015), occurs when a binary character changes state on a branch where a second character increases its rate of evolution. This results in the presence of one character state across the entire clade with the second character changing states randomly throughout the same clade (Fig. 1a); high rates of character evolution yield random data distributions at the tips. Maddison and FitzJohn (2015) found that methods developed by Maddison (1990) and Pagel (1994) also statistically support dependent evolution in this scenario, despite the associations between the first and second characters being random within one clade. These results suggest that current methods are misled by rare and unreplicated events in evolution. We extend this discussion by assessing whether two additional models of correlated evolution (the phylogenetic threshold and glmm) yield similar results.

Materials and Methods

To assess the performance of correlated evolution models, we simulated 1000 ultrametric trees, each with 100 taxa, under a pure birth process (birth rate = 0.1) using the “geiger” package in R (Pennell et al. 2014; R Core Team 2019). We assigned the values of two characters to a random clade in the phylogeny, which varied in size randomly between 40 and 60 taxa, following the procedure of Maddison and FitzJohn (2015). For Darwin’s scenario, we set both characters to state 1 for all taxa inside this clade and to 0 for all taxa outside of it (Fig. 1a). In the unreplicated burst scenario, we assigned values as above for one character. For the second character, only half the taxa inside the selected clade were randomly assigned a value of 1 (Fig. 1a). The data were not simulated by modeling character evolution along the branches of a tree at a specified rate. We only assigned the character values to the tips, leaving the node states unassigned. Our construction of the unreplicated burst scenario is, therefore, different from Maddison and FitzJohn’s (2015); they allowed trait two to evolve rapidly throughout the entire tree and secondarily changed the trait values to all the taxa outside of the selected clade to state 0. However, note that inferred rates of evolution from random data at the tips (as in the second character of the unreplicated burst scenario) should be high relative to the rest of the tree. To further assess our construction of the two scenarios, we simulated three datasets intermediate between Darwin’s and the unreplicated burst scenarios using the same tree settings noted above. Starting with Darwin’s scenario, we randomly changed 1/8, 1/6, and 1/4 of the taxa in the selected clade from state 1 to 0. This represents a graded transition between our construction of the Darwin’s and unreplicated burst scenarios.

In addition, we simulated positive and negative control scenarios, which were unreported in previous studies (Fig. 1a). In the positive control, we randomly assigned

state 1 to one character throughout the entire tree and mirrored it to create the second character. This scenario produces large evolutionary sample sizes (i.e., many independent character state changes) and we expect it to yield considerable evidence for correlated evolution. In the negative control, we randomly assigned state 1 to both characters independently throughout the tree. In this scenario, any character correspondence is coincidental; we, therefore, expect models of independent evolution to fit this character distribution better.

We replicated Maddison and FitzJohn's (2015) results with Pagel's (1994) model by fitting it to 1000 trees under the four scenarios outlined above. As previously described, this model tests whether the rate of change in one character depends on the state of the second character (Pagel 1994). We fit the models using the *ace* function in the R package "ape" (Paradis et al. 2004) and used the likelihood ratio test to test whether the dependent or independent model better fit the data. Based on Maddison and FitzJohn's (2015) study, we expected Pagel's model to favor dependent evolution in both Darwin's and unreplicated burst scenarios on average.

We also tested for correlated evolution with the same 1000 trees using the rjMCMC implementation of Pagel's (1994) model, the phylogenetic threshold model, and a phylogenetic glmm. We conducted the tests for the rjMCMC model in the program BayesTraits V3 with the 1000 simulated datasets (iterations = 1,010,000; sampling = 1000; burn-in = 10,000; state frequencies were equal). This model allowed us to test if parameter reduction resolved the issues discussed above. The rjMCMC procedure automatically reduces the number of rate parameters to those only supported by the data. Our results could, therefore, differ if model complexity were at fault in the scenarios explored here. The tree and data files for this analysis were created in R using the procedure outlined above. We calculated the proportion of

iterations where a dependent model was chosen over an independent one within each run and across all 1000 simulations. In addition, we calculated the average transition rate parameter estimates and ancestral root state probabilities for each simulation. We failed to find evidence that the MCMC chains had not converged using Tracer 1.7 (Rambaut et al. 2018).

We used a Bayesian implementation of the phylogenetic threshold model with the `threshbayes` function in the “phytools” R package (Revell 2012). Analyses were run for 10,000,000 iterations (burn-in = 2,500,000) with model parameters sampled every 1000 iterations. The mean of the posterior correlation coefficients (r-values) was estimated for each of the 1000 simulated trees ($n = 1000$ avg. r-values). To assess the significance for this distribution of mean r-values, we calculated a pMCMC, which represents two times the proportion of r-values that cross 0. Typically, pMCMCs less than 0.05 are considered good support (Fisher et al. 2013). Using trace plots, we failed to find evidence that the MCMC chains had not converged.

We implemented our phylogenetic glmm model with the “MCMCglmm” package in R (Hadfield 2010). Here, the MCMC procedure samples the posterior distribution of slope parameters for 5,000,000 iterations (burn-in = 1,250,000). We used a logit link function with the residual effect fixed to 1 and a fixed effect prior of $N(0, \sigma^2 \text{ units, } + \pi^{2/3})$, which is flat when using a logit link (Hadfield 2010; Fisher et al. 2013). At the end of each simulation, a pMCMC value was estimated, representing two times the proportion of the posterior slope parameters that cross 0. Model significance was assessed by estimating the 95% interval of the pMCMCs from the 1000 simulations ($n = 1000$ pMCMCs). A 95% interval that does not contain a pMCMC > 0.05 is taken as good evidence for correlated evolution. We assessed convergence in the MCMCglmm

models with trace plots. We also used the “coda” package’s autocorr function to ensure that there was no autocorrelation between iterations (Plummer et al. 2006).

Then, we tested for a correlation between P values and clade size from the analyses using Pagel’s (1994) method (maximum likelihood implementation). We re-ran the analyses fixing the selected clade size to 40 taxa but allowed the tree size to vary between 50 and 1000 taxa. This served as a test for the effect of tree size on the likelihood ratio test statistic, which has been proposed to be a factor for favoring the dependent model in Darwin’s scenario (Uyeda et al. 2018). These correlations were conducted using the `lm` function in the R base package (R Core Team 2019).

Moreover, we calculated the consistency index and class imbalance of all simulated datasets ($n = 1000$) under each of the four scenarios. The consistency index (CI) is a measure of character homoplasy and equals the minimum number of steps divided by the required number of steps taken to explain the data distribution on a given tree (Kluge and Farris 1969; Farris 1989). The metric ranges from 0 to 1. A data distribution with a CI of 1 has no homoplasy—no convergent evolution. We used the CI function in the R package “phangorn” (Schliep 2011) to calculate the average CI of the 1000 simulated datasets for each scenario (Fig. 1a).

Class imbalance is when different classes of data (i.e., character state combinations) are not equally represented in the dataset, which is common in comparative studies. We used a normalized measure of multi-class imbalance, which equals the maximum frequency of a class minus its minimum frequency (the normalized imbalance ratio [NIR]):

$$NIR = \frac{T_{max} - T_{min}}{n}$$

where n is the size of the dataset. The NIR will equal 0 when the data are fully balanced (all character state combinations occur with equal frequency in the dataset) and 1

when fully unbalanced (all the data have the same character states). Ideally, we want lower NIR scores so that transition rate parameters are more accurate and based on actual data. Using a combination of the CI and NIR, we developed a phylogenetic imbalance ratio (PIR):

$$PIR = NIR * CI$$

The PIR ranges from 0 to 1. An example of how these metrics are calculated are provided with a 12-tip phylogeny (Supplementary Fig. S1). A binary character has a minimum number steps of 1. In the example provided, two steps can explain the distribution of data (gray circles in Supplementary Fig. S1). A CI of 0.5 is estimated. Among the four classes of character state combinations ($\{0,0\}$, $\{0,1\}$, $\{1,0\}$, and $\{1,1\}$), the most represented classes have six out of 12 taxa. The least represented classes have zero out of 12 taxa. Therefore, the example dataset has an NIR of 0.5 and a PIR of 0.25.

To understand how different model parameter and tree settings influence the values of these metrics, we ran simulations where we constructed phylogenetic trees under a range of different tree sizes and balance levels (i.e., how disproportionate clades are across the tree), relative clade sizes, and birth rates. We constructed 100 trees using the `simulate_yule` function in the R package “apTreeshape” (Bortolussi et al. 2006). We randomly selected tree sizes to vary between 50 and 1000 taxa. We varied the birth rate between 0.1 and 100. The death rate was fixed to 0, as with the rest of our simulations. The balance level of the trees was simulated with the imbalance index, beta (β), in which a negative β represents an imbalanced tree and a positive β represents a balanced tree. We randomly selected β values to vary between -1.5 and 10, which were two extremes used in a previous simulation study on phylogenetic diversity (Maliot et al. 2018). We fixed the minimum size of unsampled

splits (ϵ) and the clade age-richness index (α) to 0.001 and -1, respectively, as done in the `simulate_yule` example provided in the package documentation. For each of the 100 trees, we simulated datasets under the positive control, negative control, Darwin's, and unreplicated burst scenarios (described previously). We varied the proportion of the selected clade size between 10 and 90% of the total number of taxa in the tree. For the positive and negative control scenarios, the selected clade size refers to the proportion of taxa across the tree that were randomly selected to have a character state of 1. We only tested for the effect of selected clade proportions between 10 and 50% because the slope at which NIR varies with selected clade proportion is expected to inflect. NIR is calculated by taking the relative difference between the most-represented and least-represented character state combinations. As the selected clade proportion increases, the character state combination that is most represented changes. The slope is expected to shift at a selected clade proportion of 50% for Darwin's and the two control scenarios and at about 67% for the unreplicated burst scenario (see Supplementary Figs. S2 and 3 for examples). We ran multiple linear regression models to assess the effects that these parameters had on CI, NIR, and PIR using the `lm` function in the R base package (R Core Team 2019). There was a total of 12 multiple linear regression models (four scenarios and three metrics). Our alpha level for statistical significance was adjusted from 0.05 to 0.001 using a Bonferroni correction (12 models with four P values estimated for each; $0.05/48 = \sim 0.001$). This correction only affected a small number of marginally significant P values. Through these simulations we inferred the range of these summary statistics, allowing us to recommend a cut-off.

To demonstrate how these metrics are used, we calculated the CI, NIR, and PIR for an empirical dataset and tested for correlated evolution using the four methods

described previously. We used a published dataset of 60 primate species with two binary traits, the presence of estrus advertisement and multimale mating (Pagel and Meade 2006). From this dataset, we removed four species (*Macaca cyclopis*, *Mucaca fascicularis*, *Macaca mulatta*, and *Trachypithecus phayrei*) due to the absence of data for at least one of the traits, resulting in a final data set of 56 species. We fit dependent and independent models of character evolution using the R package “ape” (Paradis et al. 2004) and used the likelihood ratio test to compare the fit of both models. We used the rjMCMC implementation of Pagel’s (1994) model in BayesTraits V3 (iterations = 1,000,000; sampling = 1000; burn-in = 250,000; state frequencies were equal). With the “phytools” R package (Revell 2012), we tested for a correlation using the phylogenetic threshold model. The threshold model ran for 1,000,000 iterations (burn-in = 250,000) with model parameters sampled every 1000 iterations. We then tested for a correlation using the phylogenetic glmm model in the R package “MCMCglmm” (Hadfield 2010). The model ran for 1,000,000 iterations (burn-in = 250,000). We used a logit link function with the residual effect fixed to 1 and a fixed effect prior of $N(0, \sigma^2 \text{ units, } + \pi^{2/3})$ (Hadfield 2010; Fisher et al. 2013). For all four analyses, we used a majority-rule consensus tree from the 500 trees stored in Pagel and Meade’s (2006) tree file. We assessed the convergence of the MCMC chains by making trace plots using the R packages and programs specified previously for each method. The code produced for all analyses and results are available in our Supplementary materials on Dryad (<https://doi.org/10.5061/dryad.8931zcrpw>).

Results

The positive and negative controls yield results as expected for all four methods (Table 1; Supplementary Fig. S4). We find support for dependent evolution using Pagel’s

(1994) method under Darwin's and unreplicated burst scenarios, which replicates Maddison and FitzJohn's (2015) results (Darwin's scenario median P value = 0.00207, unreplicated burst scenario median P value = 4.811E-8; Table 1; Supplementary Fig. S4 available on Dryad). The three intermediate scenarios (where 1/8, 1/6, and 1/4 of the states are randomly set back to 0) that transition between Darwin's and unreplicated burst scenarios yield progressively lower P values as the character distribution approaches that of the unreplicated burst scenario (1/8 setback P value = 4.827E-5, 1/6 P value = 1.563E-5, and 1/4 P value = 9.729E-7; Supplementary Fig. S5 available on Dryad). The rjMCMC implementation of Pagel's discrete model also supports dependent evolution under both scenarios. The median of 1000 simulations supported a dependent model for ~100% of its iterations under both scenarios (Table 1; Supplementary Fig. S4). The rjMCMC also fails to fully replicate Maddison and FitzJohn's (2015) results; out of 1000 simulations, the percentage of total iterations favoring the dependent model were slightly lower in Darwin's scenario (99.6%) than in the unreplicated burst scenario (100%). Despite small differences in the unreplicated burst results, our simulations of Pagel's model are consistent with Maddison and FitzJohn's (2015) results.

Table 1. Summary of simulation results.

	Positive	Negative	Darwin's	Unrep. Burst
Pagel's				
<i>P</i> value	<0.001	0.78	0.002	<0.001
rjMCMC				
% dep.	100%	80%	100%	100%
Threshold				
avg. <i>r</i> -value	0.99	-0.002	0.67	0.47
MCMCglmm				
pMCMC	<0.001	0.49	<0.001	0.14

Notes: Listed are the median values obtained from the distributions of results ($n = 1,000$ simulations). The type of value reported is specified in gray underneath the model's name. Gray values represent cases when a correlated model of evolution was not supported. For the reversible jump MCMC (rjMCMC) implementation of Pagel's (1994) method, we evaluated statistical significance as whether the percent number of dependent models supported was greater than 95% of the distribution; for the threshold model, we calculated the pMCMC for the distribution of average *r*-values. We found little evidence for correlated evolution in the unreplicated burst scenario (Unrep. Burst) when using the threshold and MCMCglmm models.

The phylogenetic threshold and glmm methods are less prone to support hypotheses of correlated evolution in the scenarios under study. The threshold model detects correlated evolution in Darwin's scenario but not the unreplicated burst scenario on average (Table 1; Supplementary Fig. S4 available on Dryad). In Darwin's scenario, we estimated a pMCMC of 0.018 from the distribution of average *r*-values (n

= 1000). The minimum average r -value was -0.282 with a maximum and median of 0.999 and 0.666, respectively. In the unreplicated burst scenario, we estimated a pMCMC of 0.094. The minimum average r -value was -0.396 with a maximum and median of 0.926 and 0.470. The MCMCglmm also detects correlated evolution under Darwin's scenario (95% interval pMCMC = 0.00088, 0.00112). The median pMCMC of slope values was 0.000667 with a minimum and maximum of 0.000667 and 0.036 (Table 1; Supplementary Fig. S4 available on Dryad). However, there was little evidence for correlated evolution under the unreplicated burst scenario using this model (95% interval pMCMC = 0.140, 0.147). The median pMCMC was 0.135 with a minimum and maximum of 0.00533 and 0.395, respectively (Table 1; Supplementary Fig. S4 available on Dryad).

We found no evidence for an association between clade size and P values in the analyses using Pagel's (1994) method (Darwin's: P value = 0.696; unreplicated burst: P value = 0.137; Supplementary Figs. S6 and S7 available on Dryad). We do find support for the effect of tree size on the likelihood ratio test statistic (Darwin's: P value = $1.15\text{E-}10$, R^2 = 0.04; unreplicated burst: P value = $2.15\text{E-}12$, R^2 = 0.05; Supplementary Figs. S8—S12). The R^2 values are low, however, suggesting that only 4 and 5% of the variation in the test statistic is explained by the variation in tree size for Darwin's and unreplicated burst scenarios, respectively. In addition, the simulation with the highest P value for Darwin's scenario still favors the dependent model over the independent model (Supplementary Fig. S10 available on Dryad; maximum P value = 0.026). The simulation with the highest P value for the unreplicated burst scenario favors the independent model (highest P value = 0.052), but all other simulations of equal tree size or less produced P value less than 0.01 (Supplementary Fig. S12 available on Dryad). Therefore, even though tree size has a small effect on

the level of support for the dependent model, we find no evidence that it explains why Pagel's method is prone to detecting correlated models of evolution in single transition scenarios.

The Darwin's scenario exhibits high class imbalance with a high CI (median: $\{1,1\} = 47\%$, $\{0,0\} = 53\%$, $\{0,1\} = \{1,0\} = 0\%$; $CI = 1$), which results in a PIR of 0.5. The unreplicated burst scenario, to a lesser degree, also shows high class imbalance but a low CI (median: $\{1,1\} = 24\%$, $\{0,0\} = 52\%$, $\{0,1\} = 24\%$, $\{1,0\} = 0\%$; $CI = 0.11$) and has a PIR = 0.06. Our positive control also shows high class imbalance and a low CI (median: $\{1,1\} = 47\%$, $\{0,0\} = 53\%$, $\{0,1\} = \{1,0\} = 0\%$; $CI = 0.03$), which results in a PIR of 0.02. The negative control is the least imbalanced because each trait state is randomly distributed across the tree with a correspondingly low CI (median: $\{1,1\} = 22\%$, $\{0,0\} = 28\%$, $\{0,1\} = \{1,0\} = 25\%$; $CI = 0.03$) and a PIR = 0.002.

Through our simulations of CI, NIR, and PIR under different tree settings and parameters, we found that there is no evidence for an effect of either birth rate or tree balance on any of our metrics under any of the scenarios, controlling for tree size and the selected clade proportion (P values > 0.25). The influence of tree size was found to influence the CI and PIR values under the unreplicated burst (P values = $4.06E-12$ and $1.15E-10$, respectively), positive control (P values = $7.38E-07$ and $4.40E-06$), and negative control (P values = $9.55E-07$ and $7.77E-05$) scenarios. Under the unreplicated burst scenario, CI and PIR values decrease exponentially with tree size, both leveling off at values between about 0.15 and 0.2 (Supplementary Fig. S13 available on Dryad). The CI and PIR values also decrease exponentially with tree size under the positive and negative control scenarios, but they level off under values of 0.05 and 0.04, respectively (Supplementary Figs. S14 and 15 available on Dryad). Tree size may also help explain the variation observed in NIR under the negative

control scenario (P value = 0.048); however, this effect is non-significant after the Bonferroni correction. The proportion of the selected clade is another major factor in affecting the values of these metrics, explaining all the variation observed in NIR and PIR in Darwin's scenario (P values < $2e-16$). Here, NIR and PIR both decreased with an increase in clade proportion and ranged between 0.5 and 0.9; NIR and PIR increased with the selected clade proportion after 50%, as expected (Supplementary Fig. S16 available on Dryad). Similarly, under the unreplicated burst scenario, the selected clade proportion explains nearly all the variation observed in NIR (P value < $2E-16$), which decreased with an increase in clade proportion until after about 67% (Supplementary Fig. S17 available on Dryad). There was also evidence for an effect of clade proportion on CI (P value = 0.0007) and PIR (P value = $2.54E-05$) under the unreplicated burst scenario, where both CI and PIR decreased with an increase in clade proportion with most values plotting below a value of 0.2 (Supplementary Fig. S17). Under the positive and negative control scenarios, the selected clade proportion explained nearly all the variation observed in NIR, which decreased with an increase in clade proportion (P values < $2E-16$). As in Darwin's scenario, NIR increases with the selected clade proportion after 50% of the taxa are selected (Supplementary Fig. S18 available on Dryad). There may be an effect of selected clade proportion on PIR under both positive and negative control scenarios (P values = 0.0259 and 0.004, respectively); however, these effects were nonsignificant after a Bonferroni correction. The variation observed in these metrics being explained by tree size and selected clade proportion further highlights the necessity of sufficient sample sizes when using PCMs.

We recommend lower PIR values for comparative studies because we seek to both maximize evolutionary sample size (homoplasy, a goal opposite from

phylogenetic inference) and sample enough character state combinations across the tree to facilitate parameter estimation. Following our simulations of CI, NIR, and PIR, we recommend a PIR threshold of 0.1. Our assessment was made based on the positive control scenario because it was simulated to maximize evolutionary sample sizes. Under the positive control, most PIR values fall under 0.1 (Supplementary Fig. S14 available on Dryad) under a variety of parameter and tree settings. Although class imbalance was moderately high in this scenario (average NIR = 0.67), the evolutionary sample size (average CI = 0.03) is suitable for analysis with PCMs.

The empirical dataset of estrus advertisement and multi-male mating in primates had a CI of 0.167, an NIR of 0.5, and a PIR of 0.083, demonstrating a phylogenetic character distribution and level of class imbalance that is conducive for PCMs. We find evidence for correlated evolution between the presence of estrus advertisement and multimale mating across the four methods, consistent with the results of Pagel and Meade (2006). The maximum likelihood implementation of Pagel's (1994) method yields a likelihood ratio test statistic of about 21.98 and a *P* value of about 0.0002. The Bayesian rjMCMC implementation of Pagel's method supported a model of dependent evolution ~100% of the time, closely replicating the results of Pagel and Meade (2006). We find further evidence for correlated evolution using the phylogenetic threshold model (mean *r*-value = 0.61) and the phylogenetic glmm (pMCMC < 0.001).

Discussion

Questions about evolutionary singularities (single unreplicated evolutionary scenarios) emphasize basic problems with modeling data distributions using low evolutionary sample sizes. For instance, in Darwin's and unreplicated burst scenarios, Pagel's

(1994) model favors dependent evolution, but parameter estimation is poor. The average rate parameters (inferred with rjMCMC) under Darwin's and unreplicated burst scenarios have extremely high variances across simulations and are heavily right-skewed (Fig. 2). In Darwin's scenario, rate parameters $q_{2,4}$ and $q_{3,4}$ have medians of 0.579 and 0.589, respectively, but maxima that are two-orders-of-magnitude higher (max avg. $q_{2,4} = 34.36$, max avg. $q_{3,4} = 34.48$). The rate estimates are, in fact, stable within each simulation (i.e., across iterations). Trace plots also confirm that convergence was reached for each simulation. Notably, these rate parameters describe the rate of gain when one character is already present (shifts from states $\{0,1\}$ and $\{1,0\}$, respectively, to state $\{1,1\}$). See Table 2 for a summary of the transition rate parameters. We observe a similar trend for the rate parameters $q_{2,1}$ and $q_{3,1}$, which describe the rate of loss when one character is already present (shifts from states $\{0,1\}$ and $\{1,0\}$, respectively, to $\{0,0\}$). The model cannot consistently estimate rate parameters under these scenarios where only one character state change has occurred; in other words, Darwin's scenario only has an evolutionary sample size of 1, which is insufficient for statistical analysis.

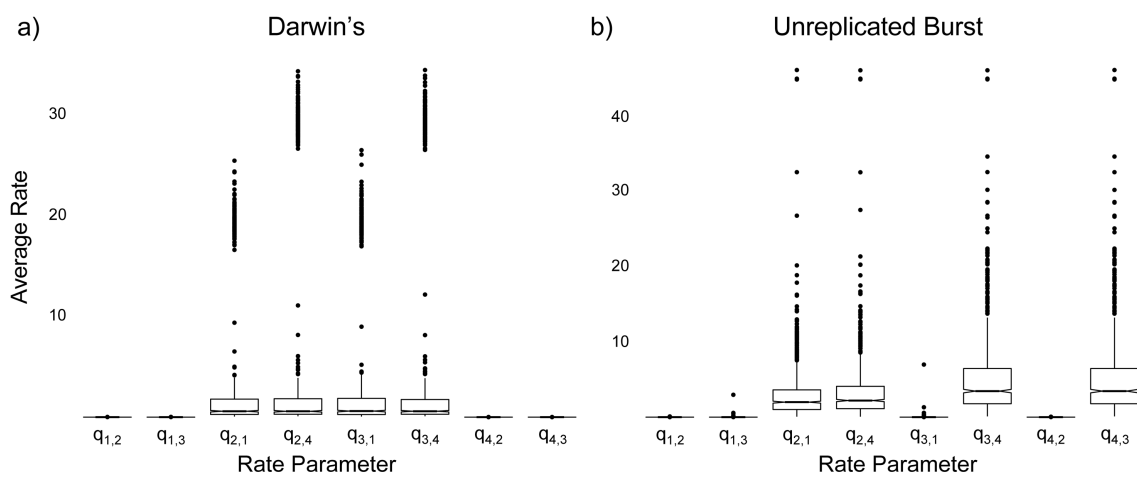


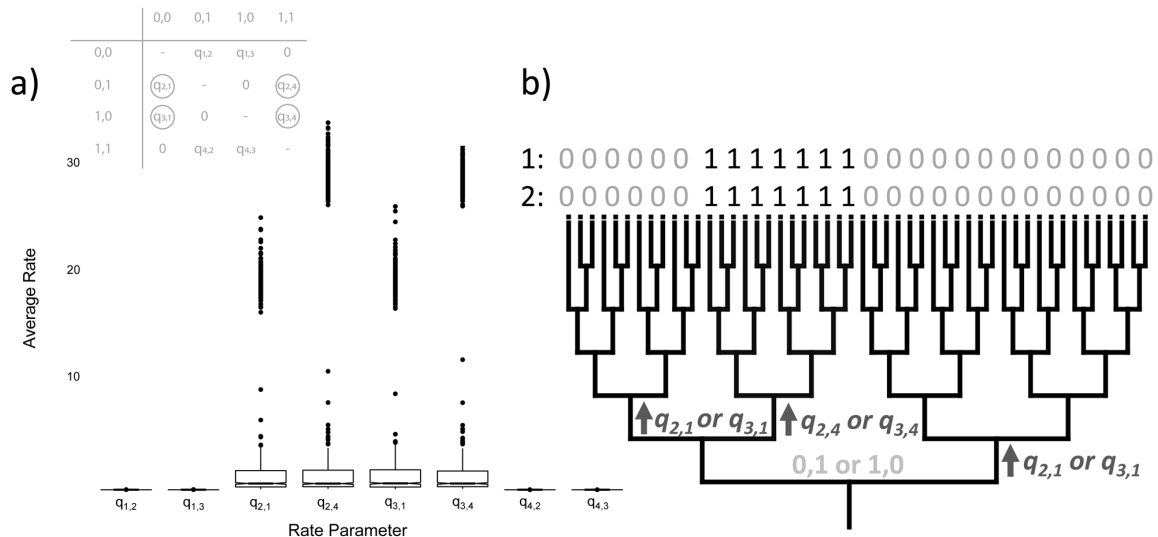
Figure 2. Boxplots of average rate parameter estimates from the 1000 rjMCMC simulations under a) Darwin's scenario and b) the unreplicated burst scenario.

Table 2. Transition rates of character change.

From/To	1	2	3	4
Trait 1,2	0,0	0,1	1,0	1,1
1 0,0	-	$q_{1,2}$	$q_{1,3}$	0
2 0,1	$q_{2,1}$	-	0	$q_{2,4}$
3 1,0	$q_{3,1}$	0	-	$q_{3,4}$
4 1,1	0	$q_{4,2}$	$q_{4,3}$	-

Pagel's (1994) model also prohibits dual character state transitions, which is assumed to have occurred in Darwin's scenario. To explain the tip data, the model must evolve through unobserved character state combinations elsewhere in the tree (Fig. 3). Out of 1000 rjMCMC simulations in Darwin's scenario, we find higher estimated probabilities for a root state of $\{0,1\}$ or $\{1,0\}$ even though they are not observed in the tip data; the average root state probabilities were $P(0,0) = 4.3\%$, $P(0,1) = 47.3\%$, $P(1,0) = 47.2\%$, and $P(1,1) = 0.029\%$ (Supplementary Fig. S19 available on Dryad). This is consistent with the high average estimates for the rate parameters $q_{2,1}$, $q_{2,4}$, $q_{3,1}$, and $q_{3,4}$, which all represent transitions from a state where one character was already present. Rate parameter estimates and root state probabilities both show that Pagel's (1994) model favors dependent evolution in Darwin's scenario because it demands that evolution occurs through unobserved state combinations (states $\{0,1\}$ or $\{1,0\}$). According to Pagel (1994), one can determine the temporal order of these

two traits by testing if their rates of gain or loss are equal between them. However, due to an insufficient evolutionary sample size in Darwin's scenario, we find that the rates of gain and loss between the two characters are indistinguishable (e.g., rate parameter $q_{2,1} = q_{3,1}$ and $q_{2,4} = q_{3,4}$; Fig. 3).



Fixing the root and clade of interest to the states $\{0,0\}$ and $\{1,1\}$ does not resolve the problem because the model reconstructs the unobserved state combination of $\{0,1\}$ or $\{1,0\}$ elsewhere in the tree. Following the analyses of Uyeda et al. (2018), we repeated the simulations setting the character loss rates to 0 while fixing the root and selected clade states (Supplementary Fig. S20 available on Dryad). Pagel's model still favored a dependent model in Darwin's scenario (99.94%

dependence across all iterations; median percent dependence = 100%). Even without loss parameters, the gain of character two from 0 to 1 still depends on the prior gain of character one or vice versa (median avg. $q_{1,2} = 8.8$, median avg. $q_{1,3} = 8.8$, median avg. $q_{2,4} = 9.2$, median avg. $q_{3,4} = 9.1$; Supplementary Fig. S21 available on Dryad). When the method estimates all four gain parameters as nearly equal, the independent model is preferred about half of the time (e.g., run 180: median avg. $q_{1,2} = q_{1,3} = q_{2,4} = q_{3,4} = 8.8$; percent dependence = 49.3%; Supplementary Fig. S22 available on Dryad). This suggests that the data distribution of Darwin's scenario is biased towards inflating the rate of gain parameters, leading to a higher rate of acceptance for a dependent model. No matter how one simulates the two characters, the model is biased towards supporting dependent evolution in Darwin's scenario; Pagel's model fails by design in these instances. These issues do not make Pagel's model obsolete, rather researchers must be aware when model assumptions are violated.

Previous authors have discussed how similar methods perform in these scenarios (Rabosky and Huang 2016; Uyeda et al. 2018), which can be recast as single or coincident evolutionary events. Rabosky and Huang (2016) noted that these models "fail" in these scenarios because they do not account for the number of independent origins. We cannot, in other words, adequately estimate model parameters with small evolutionary sample sizes. Uyeda et al. (2018) found that the support for Darwin's scenario using Pagel's (1994) model was correlated with tree size. They inferred a strong correlation between the expected likelihood difference (based on the branch length to the selected clade and the tree size) and the empirically estimated likelihood difference using Pagel's model. We also find that the likelihood ratio test statistic correlates with tree size, but with low correlation coefficients and the dependent model is almost always preferred regardless. Uyeda et al. (2018) assumed

that transition rates are low because character state changes should be rare in these scenarios. As we demonstrate, the average rates are highly variable across simulations due to low evolutionary sample sizes. Transition rates are poorly estimated even when we use the same simulated tree and data across multiple simulated trials, as in our Darwin's scenario fixed clade state analysis (Supplementary Fig. S21 available on Dryad).

The “failure” of these methods under single evolutionary scenarios has led researchers to reexamine what constitutes as evidence for correlated evolution (Uyeda et al. 2018). Uyeda et al. (2018) call for a reconsideration of how researchers test for correlated evolution. For instance, they argue that the prolonged association of two traits without subsequent loss, as in Darwin's scenario, may be taken as evidence for their correlated evolution. This view of correlated evolution can be modeled by Huelsenbeck et al.'s (2003) method. Indeed, Maddison and FitzJohn (2015) acknowledged that Huelsenbeck et al.'s method would likely support a model of correlated evolution in both Darwin's and unreplicated burst scenarios given the high amount of shared branch lengths between the two characters. However, both Uyeda et al. and Maddison and FitzJohn note that taxonomists often seek out characters with such phylogenetic distributions, increasing the chance that researchers will analyze traits with independent origins. Huelsenbeck et al.'s and Uyeda et al.'s conception of correlated evolution differs from Pagel's (1994) dependent evolution, which tests whether the evolution of one character depends on the existing state of another. Pagel's model can take the lack of character state reversals as evidence for dependent evolution. For example, a model of dependent evolution is supported if the rates of loss from state $\{1,1\}$ to $\{0,1\}$ or $\{1,0\}$ are 0, like in Darwin's scenario ($q_{4,2} = q_{4,3} = 0$; Table 2). However, this is specifically due to the rate of loss

in one character being dependent on the state of the other character. In addition, an independent model of evolution may still be supported when all rates of character loss are set to 0, so long as the other transition rates are equal (Supplementary Fig. S22 available on Dryad). Central to these different interpretations is whether we consider stasis as an important aspect in macroevolution.

The Markov model that underlies Maddison's (1990), Pagel's (1994), and Huelsenbeck et al.'s (2003) models has limitations that may exacerbate the problem of modeling single evolutionary events with discrete characters. First, adjacent branches and neighboring sections of a lineage are treated as independent (Maddison and FitzJohn 2015). Homologous structures evolving within a single clade are treated as independent but may undergo parallel evolution. Another issue with the Markov model is that it assumes the transition rate from one state to another is the same over time (Goldberg and Foo 2019). Goldberg and Foo (2020) developed multiple models, referred to as "memory models", in which the rate of transition depends on the time that two character states are associated. These models are particularly intriguing in the case of Darwin's scenario, where the lack of reverting back to states $\{0,1\}$ or $\{1,0\}$ after evolving the state $\{1,1\}$ may be best represented by a memory model. This is also consistent with the arguments made by Uyeda et al. (2018) in which the prolonged association of the two character states could be taken as evidence for their correlated evolution. However, the current implementations of these memory models are not conducive for testing Darwin's and similar scenarios because they assume that rates of character state gain and loss are equal. The lack of character state losses under Darwin's scenario, for example, would invalidate this assumption. A generalized memory model that allows for rates of gain and loss to vary may make these models more widely applicable. Moreover, these memory models are also susceptible to low

sample sizes, like other models (and statistics in general) discussed here. In Goldberg and Foo's simulations, they found that the shape of the function describing the wait time between character state transitions is inaccurately estimated when the true transition rate is low (i.e., fewer character state changes; see Fig. 6 in Goldberg and Foo 2020).

We also tested two non-Markovian models of discrete character evolution. Like Pagel's model, the threshold and glmm models "fail" in Darwin's scenario despite fundamentally different model structure; an evolutionary sample size of one will plague any statistical model. Our analyses, however, demonstrate that the threshold and glmm models can correctly infer independent evolutionary change in more complex scenarios (e.g., the unreplicated burst scenario).

The tests discussed above are employed to reach conclusions about evolutionary interactions. A "weak conclusion" (Maddison and FitzJohn 2015, pg. 128) is that the traits under study belong to the "same adaptive/functional network". "Stronger" conclusions may be reached, however, with a study design that establishes clear *a priori* hypotheses backed by theoretical expectations and taxon sampling that maximizes independent originations of trait values (see Recommendations below). We agree with Maddison and FitzJohn (2015) that modeling alone for scenarios like Darwin's are insufficient for testing hypotheses about correlated evolution. But how then should evolutionary biologists proceed in testing hypotheses about correlated evolution in these all too common situations?

Recommendations

Our first recommendation (hardly original) is to craft *pre-hoc* hypotheses so that predictions can be assessed with study designs that maximize evolutionary sample

sizes (the number of independent originations of trait states). This also helps from falling prey to data dredging and p-hacking, a problem accentuated by easy access to large data repositories. Biologists can further maximize evolutionary sample sizes by broadening the taxonomic scope of the research question. Maximizing evolutionary sample size is particularly helpful for complex character distributions, such as those described by Maddison and FitzJohn (2015). In these cases, statistical significance is consistent with *pre-hoc* hypotheses of correlated evolution. Our simulations also demonstrate that the phylogenetic threshold model and glmm may be useful in more complex scenarios. Moreover, as our simulations show, we should prioritize meaningful and robust parameter estimates over low P values. In the case of correlated evolution, statistical significance is biologically meaningless if associated with low evolutionary samples sizes and biased transition rate estimates. The effects of low sample sizes and improper parameter estimates applies to all PCMs, not just those that test for correlated evolution.

Sample size can be difficult to define in comparative studies. For instance, it is necessary for comparing molecular and trait evolution models, such as with Bayesian Information Criteria (BIC; Schwarz 1978) or Akaike Information Criteria (Akaike 1974) corrected for sample size (AICc; Hurvich and Tsai 1989). However, what should be used for sample size is often uncertain (Posada and Buckley 2004; Beaulieu et al. 2019). The number of taxa is commonly used for comparing trait evolution models, but BIC and AICc assume that these are independent observations (Bartoszek 2016). Multiple studies have used modified effective sample sizes that better reflect the amount of independent signal in the trait data (Ho and Ané 2014; Bartoszek 2016). Similarly, the number of sites in an alignment may be an overestimate given the nonindependence of sequence data (Posada and Buckley 2004). Through

simulations, Beaulieu et al. (2019) found that the number of sites multiplied by the number of taxa worked best for comparing codon selection models. More simulation work is needed to determine the proper use of sample size in phylogenetic models.

Our second recommendation is to use the consistency index and class imbalance to examine the suitability of the data for modeling. The consistency index (CI) is the minimum number of character state changes divided by the number of state changes mapped onto the tree. It is a simple nonmodel based measure of character homoplasy. Researchers can explore data distributions in the context of phylogeny using reported CIs from programs such as Mesquite and the R packages “phangorn” and “phytools” (Schliep 2011; Revell 2012; Maddison and Maddison 2014). For PCMs, we aim to maximize evolutionary sample sizes (maximize homoplasy, the opposite goal for phylogenetic inference), which yield low consistency index scores. Darwin’s scenario, for example, suffers from a high consistency index—such characters would be good for inferring a phylogeny, but are problematic for studying with PCMs (Fig. 1a; Table 3).

Table 3. Metric combinations for different character distribution scenarios.

	Low CI	High CI
Low		
Imbalance	Negative	–
High	<i>Positive</i>	
Imbalance	Unrep. Burst	<i>Darwin’s</i>

Notes: A combination of consistency index (CI), class imbalance, and correlation (black italics = correlated evolution) can distinguish the four scenarios. Evidence for correlated evolution in the unreplicated burst scenario (Unrep. Burst) depends on the method used.

Class imbalance (when classes of data are unrepresented) is common in comparative studies. It is most cited as a concern for learning algorithms and predictive modeling, in which estimated probabilities are biased against the unrepresented class (Japkowicz and Stephen 2002; Oommen et al. 2011; Wang and Yao 2012; Wallace and Dahabreh 2014; Kaur et al. 2019). Notably, for logistic regression, having a sample that is representative of the true population is more important than having a perfectly balanced dataset (Oommen et al. 2011). We contend that data representation also applies to PCMs for discrete characters; these methods rely on accurate probability and rate estimates. To assess the adequacy of PCM models for discrete characters, we developed a phylogenetic class imbalance metric (which ranges from 0 to 1). Darwin's scenario exhibits a high PIR (= 0.5), whereas the unreplicated burst scenario and positive and negative controls have lower PIRs (= 0.06, 0.02, and 0.002, respectively). We recommend low PIR values ($\text{PIR} < 0.1$) to maximize evolutionary sample sizes and minimize class imbalance. This metric distinguishes Darwin's scenario as being problematic ($\text{PIR} = 0.5$) but not the negative control ($\text{PIR} = 0.002$) because the PIR is a quality control metric and not a test for correlated evolution. The unreplicated burst scenario's suitability for PCMs depends on tree size and the proportion of the selected unreplicated burst clade. A PIR can be low enough if the number of independent originations in the unreplicated burst clade is large enough. The PIR is meant to aid researchers during the early phases of investigating whether their discrete data are suitable for phylogenetic modeling. It can be used in all types of PCM studies for discrete characters, not just those focusing on correlation. Note, however, that different combinations of CI and NIR can yield the same PIR value. It is important to also analyze individual CI and NIR values to

determine the exact cause of a low PIR, either due to little homoplasy or high data imbalance.

Our third, and most general, recommendation is to apply the principle of consilience early in the study design. William Whewell developed the concept of consilience in his masterwork *The Philosophy of the Inductive Sciences*, in which he argued that a theory's strength lies in its ability to coherently connect facts from multiple unrelated fields—even without direct observation of the underlying cause (Whewell 1840). Larry Laudan (1981) later summarized Whewell's approach:

...the real strength of such an hypothesis is usually that *it shows that events previously thought to be of different kinds are, as a matter of fact, the 'same' kind of event* (Laudan 1981, pp. 166; emphasis not our own).

The “consilience of inductions” approach made an immediate and lasting impact on the historical sciences, such as biology and astronomy, where direct experimentation is difficult or impossible. Charles Darwin, for example, closely followed Whewell's advice to structure his arguments for natural selection by drawing on evidence from many disparate fields (Ruse 1975; Thagard 1977). The argument for consilience was apparent in the second edition of Darwin's seminal book, *On the Origin of Species*:

I cannot believe that a false theory would explain, as it seems to me that the theory of natural selection does explain, the several large classes of facts above specified (Darwin 1860, pp. 480-481).

In this passage, Darwin argues that his theory of natural selection could adequately and coherently explain the evidence he described from the independent fields of biogeography, comparative anatomy, domestication, and embryology. The theory of natural selection has since grown to ubiquitous acceptance because of the convergence of evidence from unconnected fields that independently support it, including modern developmental biology, genetics, and genomics.

Evolutionary biologists, especially those using PCMs, can and should apply Whewell's consilience of inductions to evaluate primary and alternative hypotheses when possible (Fig. 4). Mexican tetra fish (*Astyanax mexicanus*), a model organism for studying convergent and parallel evolution, provides an excellent example of how consilience can be applied (Yamamoto and Jeffery 2000; Jeffery 2001; Wilkens and Strecker 2003; Protas et al. 2006, 2007; Gross et al. 2009; McGaugh et al. 2014). Over 20 populations of Mexican tetra are known to inhabit restricted karst regions, such as caves, where eye loss and reduced pigmentation evolve independently (Wilkens and Strecker 2003). Researchers studying Mexican tetra find empirical evidence for evolutionary associations using genetics and developmental biology. Crosses among separate cave fish populations result in offspring with more well-developed eyes and pigments, strongly suggesting that these populations independently reduced their pigmentation and ceased eye development through different mutational mechanisms (Wilkens and Strecker 2003). Quantitative trait loci mapping paired with expression analyses show that different mutations in the same genes and changes in their expression levels are associated with similar adaptations among cave-dwelling tetra fish populations (Protas et al. 2006, 2007; Gross et al. 2009; McGaugh et al. 2014). These results are further bolstered by developmental studies. Cave fish develop diminished eyes early in development, which later

regresses completely through elevated apoptosis of cells in the lens (Jeffery 2001). Transplanting a surface-dwelling fish's lens into a cave-dwelling fish embryo counteracts this regression (Yamamoto and Jeffery 2000; Jeffery 2001).

We realize that the level of consilience exhibited in Mexican tetra fish research may be impractical for many study systems. However, research programs may be designed in such a way that models can be applied to hundreds of taxa. For example, it is impractical to study the biomechanics of >100 bird species, but it is feasible to model a proxy for many species and pair that with a biomechanical analysis of representative species from each group of interest. A comparative analysis of >100 bird species found that claw curvature is associated with grasping behavior (Cobb and Sellers 2020), which aligns with a structural analysis of claws from select bird species specializing in different prey sizes (Tsang et al. 2019). It can also be expensive to study macroevolution with many different types of analyses. Some studies achieve this by comparing representative species from each major clade. A study on hummingbirds used a combination of behavioral experiments, comparative genetics, and gene expression analysis to demonstrate the independent evolution of sweet taste reception within archosaurs (Baldwin et al. 2014). These examples show how consilience can be leveraged in studies that involve relatively large samples of taxa or broad taxonomic focuses.

Consilience simply helps to more rigorously test hypotheses in evolutionary biology than assessment from a single isolated field. Consilience should be the central philosophy of all evolutionary biologists, including (and maybe especially) those who specialize in PCMs.

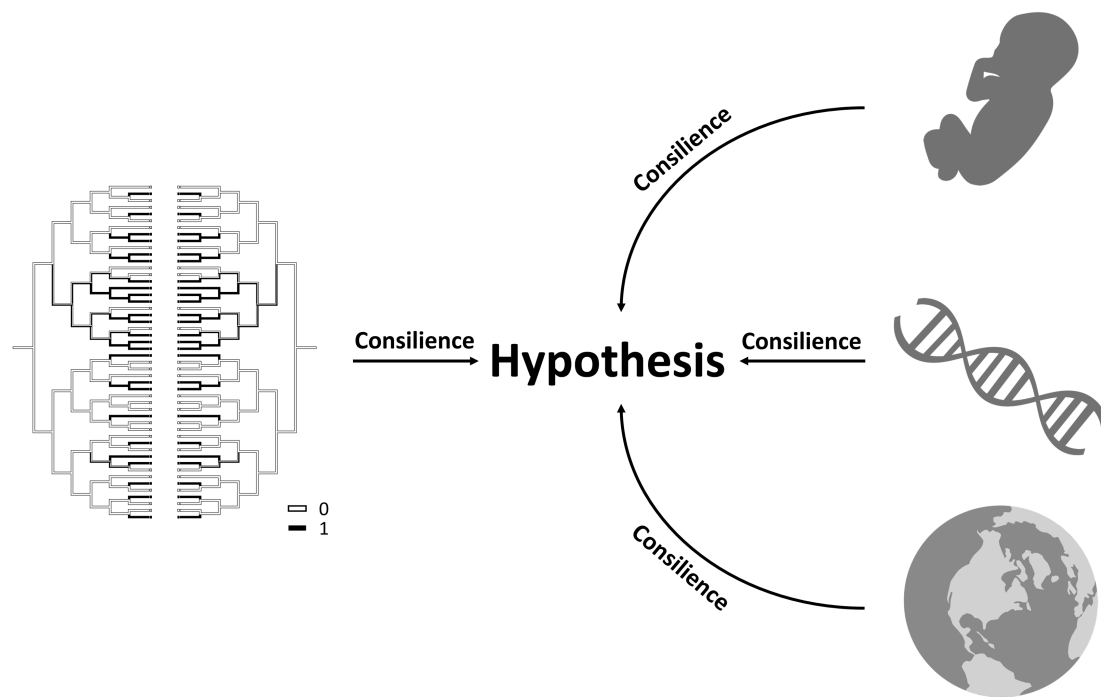


Figure 4. Roadmap for a consilience of inductions approach in comparative biology, combining phylogenetic comparative methods (left), developmental biology (top right), genetics (middle right), and biogeography (bottom right), among other fields. Infant, DNA, and globe illustrations are open source on pixabay.com.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19, 716–723.
- Baldwin, M.W., Toda, Y., Nakagita, T., O’Connell, M.J., Klasing, K.C., Misaka, T., Edwards, S.V., and Liberles, S.D. (2014). Evolution of sweet taste perception in hummingbirds by transformation of the ancestral umami receptor. *Science* 345, 929–933.
- Bartoszek, K. (2016). Phylogenetic effective sample size. *J. Theor. Biol.* 407, 371–386.

- Beaulieu, J.M., O'Meara, B.C., Zaretzki, R., Landerer, C., Chai, J., and Gilchrist, M.A. (2019). Population genetics based phylogenetics under stabilizing selection for an optimal amino acid sequence: a nested modeling approach. *Mol. Biol. Evol.* 36, 834–851.
- Bianchini, G., and Sánchez-Baracaldo, P. (2020). sMap: Evolution of independent, dependent and conditioned discrete characters in a Bayesian framework. *Methods Ecol. Evol.* 00, 1–8.
- Bortolussi, N., Durand, E., Blum, M., and François, O. (2006). apTreeshape: statistical analysis of phylogenetic tree shape. *Bioinformatics* 22, 363–364.
- Cobb, S.E., and Sellers, W.I. (2020). Inferring lifestyle for Aves and Theropoda: A model based on curvatures of extant avian ungual bones. *PLOS ONE* 15, e0211173.
- Darwin, C.R. (1860). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life* (London: Murray).
- Eastman, J.M., Alfaro, M.E., Joyce, P., Hipp, A.L., and Harmon, L.J. (2011). A novel comparative method for identifying shifts in the rate of character evolution on Trees. *Evolution* 65, 3578–3589.
- Farris, J.S. (1989). The retention index and the rescaled consistency index. *Cladistics* 5, 417–419.
- Felsenstein, J. (1985). Phylogenies and the comparative method. *Am. Nat.* 125, 1–15.
- Felsenstein, J. (1988). Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.* 19, 445–471.
- Felsenstein, J. (2003). *Inferring phylogenies* (Sunderland, Mass: Sinauer Associates is an imprint of Oxford University Press).

- Felsenstein, J. (2005). Using the quantitative genetic threshold model for inferences between and within species. *Philos. Trans. R. Soc. B Biol. Sci.* 360, 1427–1434.
- Felsenstein, J. (2012). A comparative method for both discrete and continuous characters using the threshold model. *Am. Nat.* 179, 145–156.
- Fisher, R.M., Cornwallis, C.K., and West, S.A. (2013). Group formation, relatedness, and the evolution of multicellularity. *Curr. Biol.* CB 23, 1120–1125.
- Garamszegi, L.Z. (2014). *Modern Phylogenetic comparative methods and their application in evolutionary biology: concepts and practice* (New York: Springer).
- Goldberg, E.E., and Foo, J. (2019). Memory in trait macroevolution. *Am. Nat.* 195, 300–314.
- Gross, J.B., Borowsky, R., and Tabin, C.J. (2009). A novel role for *Mc1r* in the parallel evolution of depigmentation in independent populations of the cavefish *Astyanax mexicanus*. *PLOS Genet.* 5, e1000326.
- Hadfield, J.D. (2010). MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* 33, 1–22.
- Harmon, L.J. (2018). *Phylogenetic comparative methods: learning from trees* (CreateSpace Independent Publishing Platform).
- Harvey, P.H., and Pagel, M.D. (1991). *The comparative method in evolutionary biology* (Oxford ; New York: Oxford University Press).
- Ho, L.S.T., and Ané, C. (2014). Intrinsic inference difficulties for trait evolution with Ornstein-Uhlenbeck models. *Methods Ecol. Evol.* 5, 1133–1146.
- Huelsenbeck, J.P., Nielsen, R., and Bollback, J.P. (2003). Stochastic mapping of morphological characters. *Syst. Biol.* 52, 131–158.
- Hurvich, C.M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* 76, 297–307.

- Ives, A.R., and Garland, T. (2014). Phylogenetic regression for binary dependent variables. In *Modern phylogenetic comparative methods and their application in evolutionary biology: concepts and practice*, L.Z. Garamszegi, ed. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 231–261.
- Ives, A.R., and Garland, T., Jr. (2010). Phylogenetic logistic regression for binary Dependent Variables. *Syst. Biol.* 59, 9–26.
- Japkowicz, N., and Stephen, S. (2002). The class imbalance problem: a systematic study. *Intell. Data Anal.* 6, 429–449.
- Jeffery, W.R. (2001). Cavefish as a model system in E=evolutionary developmental biology. *Dev. Biol.* 231, 1–12.
- Jukes, T.H., and Cantor, C.R. (1969). CHAPTER 24 - Evolution of protein molecules. In *mammalian protein metabolism*, H.N. Munro, ed. (Academic Press), pp. 21–132.
- Kaur, H., Pannu, H.S., and Malhi, A.K. (2019). A systematic review on imbalanced data challenges in machine learning: applications and solutions. *ACM Comput. Surv.* 52, 79:1-36.
- Keith, S.A., Webb, T.J., Böhning-Gaese, K., Connolly, S.R., Dulvy, N.K., Eigenbrod, F., Jones, K.E., Price, T., Redding, D.W., Owens, I.P.F., et al. (2012). What is macroecology? *Biol. Lett.* 8, 904–906.
- Kluge, A.G., and Farris, J.S. (1969). Quantitative phyletics and the evolution of Anurans. *Syst. Zool.* 18, 1–32.
- Kubo, T., and Iwasa, Y. (1995). Inferring the rates of branching and extinction from molecular phylogenies. *Evolution* 49, 694–704.

- Laudan, L. (1981). William Whewell on the consilience of inductions. In *Science and hypothesis: historical essays on scientific methodology*, L. Laudan, ed. (Dordrecht: Springer Netherlands), pp. 163–180.
- Louca, S., and Pennell, M.W. (2020). Extant timetrees are consistent with a myriad of diversification histories. *Nature* 580, 502–505.
- Maddison, W.P., and Maddison, D.R. (2014). Mesquite: a modular system for evolutionary analysis. Version 3.01 <http://mesquiteproject.org>.
- Maddison, W.P. (1990). A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution* 44, 539–557.
- Maddison, W.P., and FitzJohn, R.G. (2015). The unsolved challenge to phylogenetic correlation tests for categorical characters. *Syst. Biol.* 64, 127–136.
- Maliet, O., Gascuel, F., and Lambert, A. (2018). Ranked tree shapes, nonrandom extinctions, and the loss of phylogenetic diversity. *Syst. Biol.* 67, 1025–1040.
- McGaugh, S.E., Gross, J.B., Aken, B., Blin, M., Borowsky, R., Chalopin, D., Hinaux, H., Jeffery, W.R., Keene, A., Ma, L., et al. (2014). The cavefish genome reveals candidate genes for eye loss. *Nat. Commun.* 5.
- McPeck, M.A. (1995). Testing hypotheses about evolutionary change on single branches of a phylogeny using evolutionary contrasts. *Am. Nat.* 145, 686–703.
- Morlon, H. (2014). Phylogenetic approaches for studying diversification. *Ecol. Lett.* 17, 508–525.
- Nee, S., Holmes, E.C., May, R.M., Harvey, P.H., Lawton, J.H., and May, R.M. (1994). Extinction rates can be estimated from molecular phylogenies. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 344, 77–82.
- Nielsen, R. (2002). Mapping mutations on phylogenies. *Syst. Biol.* 51, 729–739.

- Oommen, T., Baise, L.G., and Vogel, R.M. (2011). Sampling bias and class imbalance in maximum-likelihood logistic regression. *Math. Geosci.* 43, 99–120.
- Organ, C., Nunn, C.L., Machanda, Z., and Wrangham, R.W. (2011). Phylogenetic rate shifts in feeding time during the evolution of *Homo*. *Proc. Natl. Acad. Sci. U. S. A.* 108, 14555–14559.
- Organ, C.L., Shedlock, A.M., Meade, A., Pagel, M., and Edwards, S.V. (2007). Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446, 180–184.
- Organ, C.L., Janes, D.E., Meade, A., and Pagel, M. (2009). Genotypic sex determination enabled adaptive radiations of extinct marine reptiles. *Nature* 461, 389–392.
- Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. B* 255, 37–45.
- Pagel, M. (1999). Inferring the historical patterns of biological evolution. *Nature* 401, 877–884.
- Pagel, M., and Meade, A. (2006). Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *Am. Nat.* 167, 808–825.
- Paradis, E., Claude, J., and Strimmer, K. (2004). APE: analyses of phylogenetics and evolution in {R} language. *Bioinformatics* 20, 289–290.
- Pennell, M.W., Eastman, J.M., Slater, G.J., Brown, J.W., Uyeda, J.C., FitzJohn, R.G., Alfaro, M.E., and Harmon, L.J. (2014). geiger v2.0: an expanded suite of methods for fitting macroevolutionary models to phylogenetic trees. *Bioinforma. Oxf. Engl.* 30, 2216–2218.

- Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: convergence diagnosis and output analysis for MCMC. *R News* 6, 7–11.
- Posada, D., and Buckley, T.R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53, 793–808.
- Protas, M., Conrad, M., Gross, J.B., Tabin, C., and Borowsky, R. (2007). Regressive evolution in the Mexican cave tetra, *Astyanax mexicanus*. *Curr. Biol. CB* 17, 452–454.
- Protas, M.E., Hersey, C., Kochanek, D., Zhou, Y., Wilkens, H., Jeffery, W.R., Zon, L.I., Borowsky, R., and Tabin, C.J. (2006). Genetic analysis of cavefish reveals molecular convergence in the evolution of albinism. *Nat. Genet.* 38, 107–111.
- R Core Team (2019). R: a language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).
- Rabosky, D.L. (2010). Extinction rates should not be estimated from molecular phylogenies. *Evolution* 64, 1816–1824.
- Rabosky, D.L., and Huang, H. (2016). A robust semi-parametric test for detecting trait-dependent diversification. *Syst. Biol.* 65, 181–193.
- Rambaut, A., Drummond, A.J., Xie, D., Baele, G., and Suchard, M.A. (2018). Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904.
- Revell, L.J. (2008). On the analysis of evolutionary change along single branches in a phylogeny. *Am. Nat.* 172, 140–147.
- Revell, L.J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223.

- Ruse, M. (1975). Darwin's debt to philosophy: an examination of the influence of the philosophical Ideas of John F.W. Herschel and William Whewell on the development of Charles Darwin's theory of evolution. *Stud. Hist. Philos. Sci. Part A* 6, 159–181.
- Schliep, K.P. (2011). phangorn: phylogenetic analysis in R | *Bioinformatics* | Oxford Academic. *Bioinformatics* 27, 592–593.
- Schwartz, R., and Schäffer, A.A. (2017). The evolution of tumour phylogenetics: principles and practice. *Nat. Rev. Genet.* 18, 213–229.
- Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.
- Smith, R.J. (1994). Degrees of freedom in interspecific allometry: An adjustment for the effects of phylogenetic constraint. *Am. J. Phys. Anthropol.* 93, 95–107.
- Thagard, P.R. (1977). Darwin and Whewell. *Stud. Hist. Philos. Sci. Part A* 8, 353–356.
- Tsang, L.R., Wilson, L.A.B., Ledogar, J., Wroe, S., Attard, M., and Sansalone, G. (2019). Raptor talon shape and biomechanical performance are controlled by relative prey size but not by allometry. *Sci. Rep.* 9, 7076.
- Uyeda, J.C., Zenil-Ferguson, R., and Pennell, M.W. (2018). Rethinking phylogenetic comparative methods. *Syst. Biol.* 67, 1091–1109.
- Venditti, C., Meade, A., and Pagel, M. (2011). Multiple routes to mammalian diversity. *Nature* 479, 393–396.
- Wallace, B.C., and Dahabreh, I.J. (2014). Improving class probability estimates for imbalanced data. *Knowl. Inf. Syst.* 41, 33–52.
- Wang, S., and Yao, X. (2012). Multiclass imbalance problems: analysis and potential Solutions. *IEEE Trans. Syst. Man Cybern. Part B Cybern.* 42, 1119–1130.
- Whewell, W. (1840). *The Philosophy of the inductive sciences, founded upon their history* (London: J. W. Parker).

- Wilkens, H., and Strecker, U. (2003). Convergent evolution of the cavefish *Astyanax* (Characidae, Teleostei): genetic evidence from reduced eye-size and pigmentation. *Biol. J. Linn. Soc.* 80, 545–554.
- Wright, S. (1934). An analysis of variability in number of digits in an inbred strain of guinea pigs. *Genetics* 19, 506–536.
- Yamamoto, Y., and Jeffery, W.R. (2000). Central role for the lens in cave fish eye degeneration. *Science* 289, 631–633.

Appendix 1

Supplementary figures

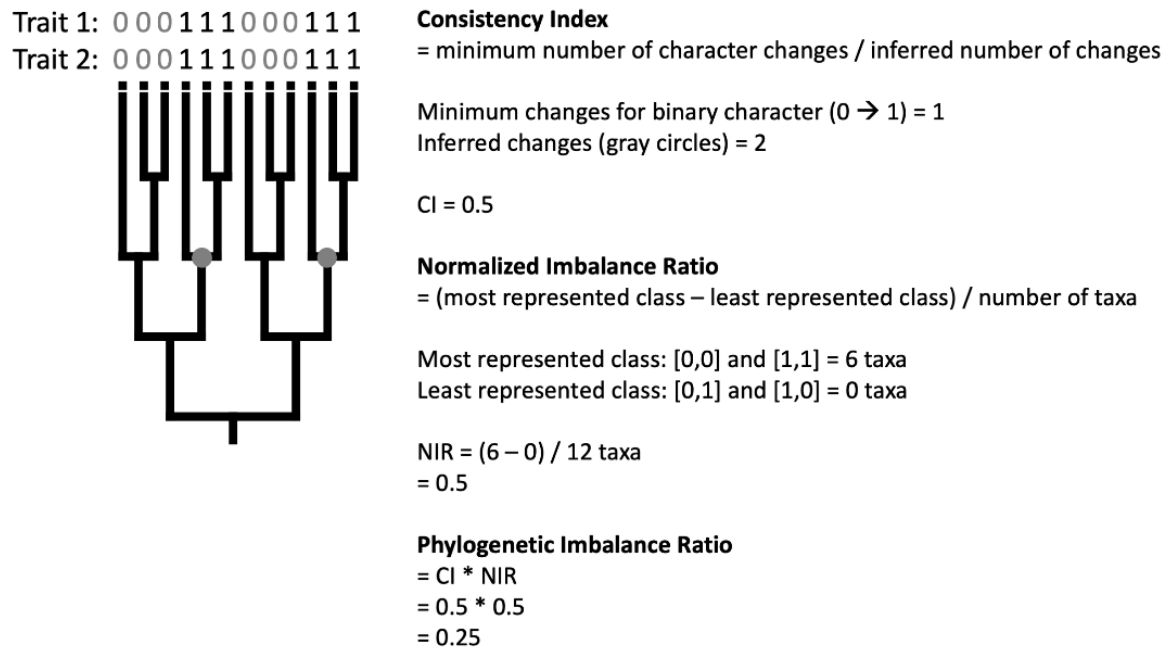
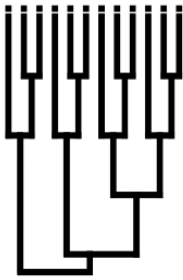


Figure S1: Example of how the consistency index (CI), normalized imbalance ratio (NIR), and phylogenetic imbalance ratio (PIR) are calculated. A 12-tip phylogeny with two trait distributions listed above. Gray 0s represent absence of traits and black 1s represent presence of traits. Dark gray circles represent the inferred character state changes.

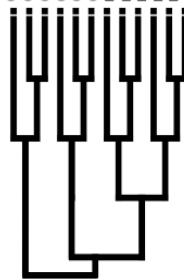
Trait 1: 0 0 0 0 0 0 0 0 0 1 1 1
 Trait 2: 0 0 0 0 0 0 0 0 0 1 1 1



Clade proportion = 0.25

NIR = most – least/total
 = 9 – 0/12
 = 0.75

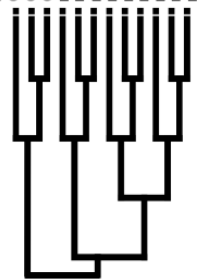
Trait 1: 0 0 0 0 0 0 1 1 1 1 1 1
 Trait 2: 0 0 0 0 0 0 1 1 1 1 1 1



Clade proportion = 0.50

NIR = most – least/total
 = 6 – 0/12
 = 0.50

Trait 1: 0 0 0 1 1 1 1 1 1 1 1 1
 Trait 2: 0 0 0 1 1 1 1 1 1 1 1 1

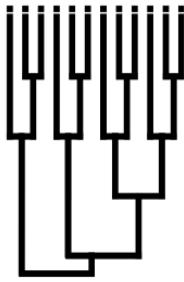


Clade proportion = 0.75

NIR = most – least/total
 = 9 – 0/12
 = 0.75

Figure S2: Example of how the effect of clade proportion on NIR inflects under Darwin's scenario. With a 12-tip phylogeny, the most-represented character state combination {0,0} changes to being {1,1} after the selected clade proportion increases beyond 50%. This results in the slope inflection observed in Supplementary Fig. 16.

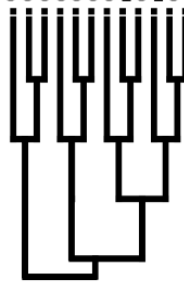
Trait 1: 000000000111
 Trait 2: 000000000101



Clade proportion = 0.25

NIR = most – least/total
 = 9 – 0/12
 = 0.75

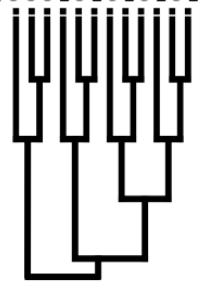
Trait 1: 000000111111
 Trait 2: 00000010101



Clade proportion = 0.50

NIR = most – least/total
 = 6 – 0/12
 = 0.50

Trait 1: 000111111111
 Trait 2: 0001010101



Clade proportion = 0.75

NIR = most – least/total
 = 5 – 0/12
 = 0.50

Figure S3: Example of how the effect of clade proportion on NIR inflects under the unreplicated burst scenario. With a 12-tip phylogeny, the most-represented character state combination {0,0} changes to being {1,1} after the selected clade proportion increases beyond about 67%. This results in the slope inflection observed in Supplementary Fig. 17. A clade proportion between 50-75% because another monophyletic clade cannot be selected with this 12-tip example.

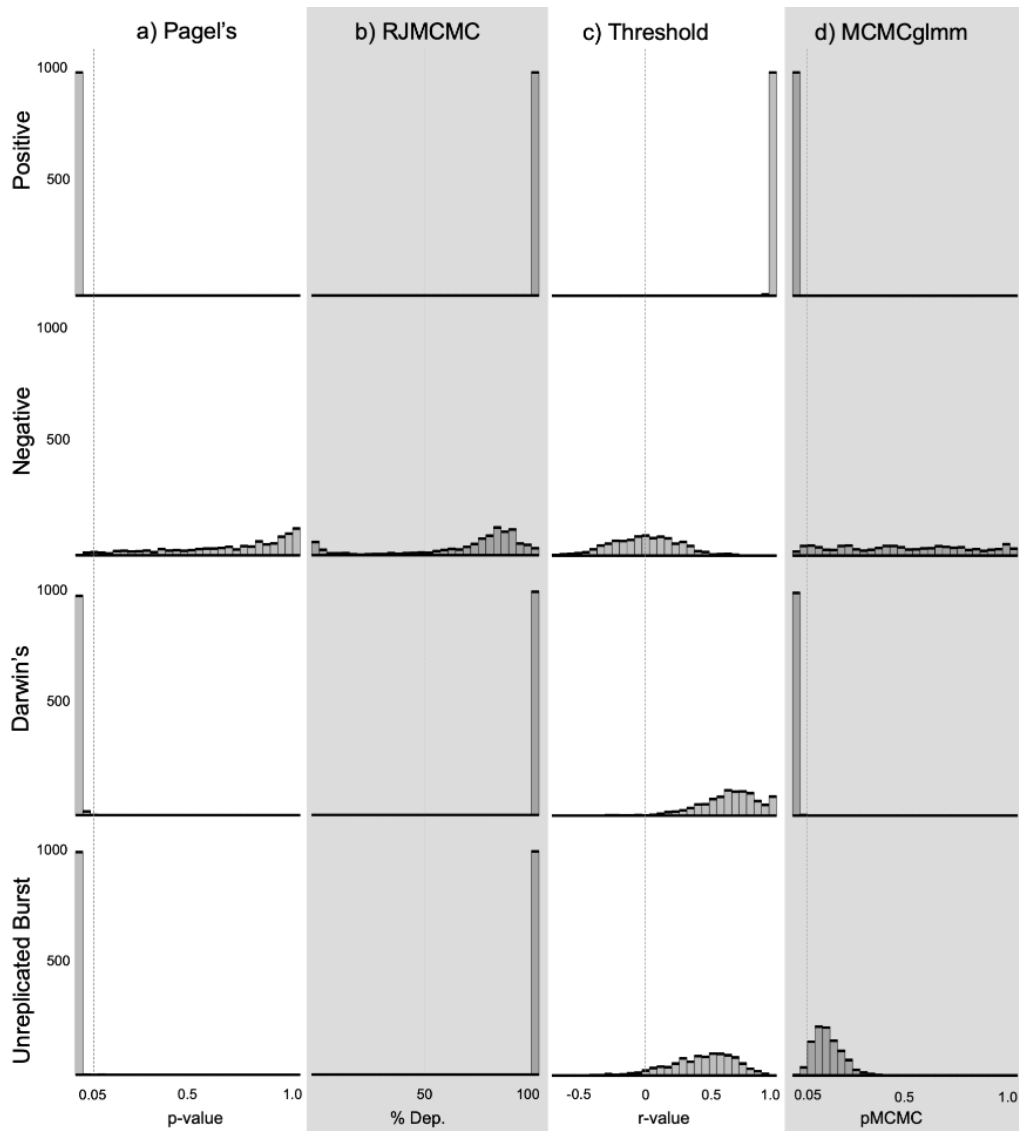


Figure S4: Histograms illustrating the distribution of results from simulations (n = 1,000 simulations) with rows representing the simulated scenario and columns representing the method used: a) P values from Pagel's discrete method (vertical dashed line: P value = 0.05); b) mean percent dependence from the RJMCMC implementation of Pagel's discrete method (vertical dashed line: % dependence = 50); c) mean correlation coefficient values (r-value) from the threshold model (vertical dashed line: r-value = 0); d) mean pMCMC from the MCMCglmm analyses (vertical dashed line: pMCMC = 0.05).

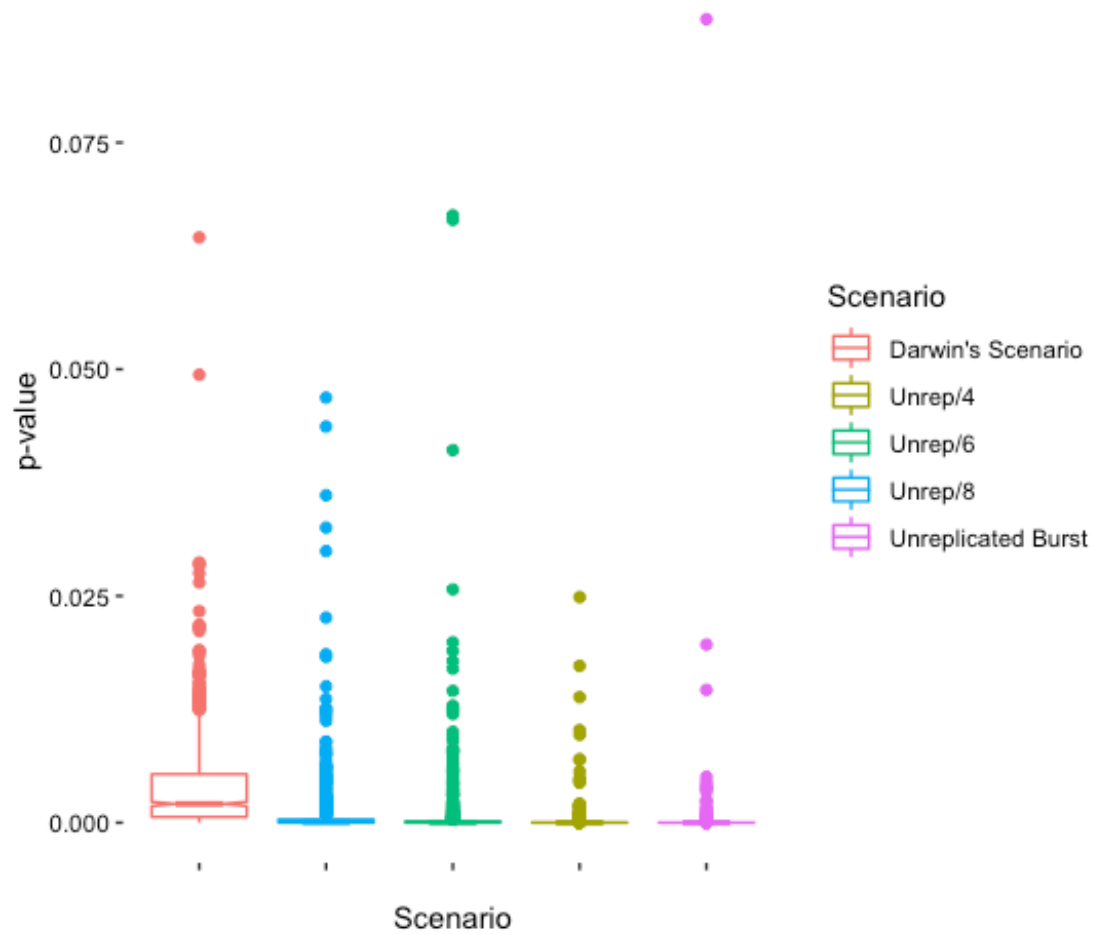
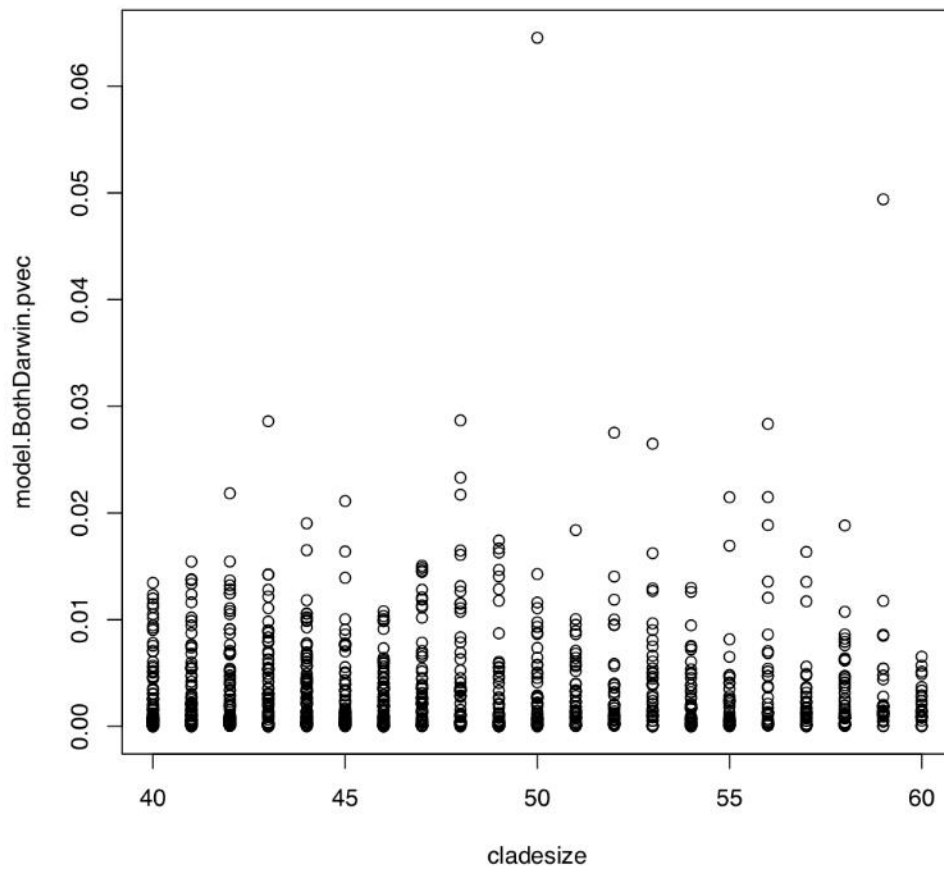


Figure S5: Side-by-side boxplots of P values estimated from the 1,000 Pagel's method under Darwin's, unreplicated burst, and intermediate (Unrep/4, Unrep/6, Unrep/8) scenarios. The intermediate scenarios progress between Darwin's and the unreplicated burst scenario from left to right, starting with the scenario where 1/8 of the character states are changed back to 0 (Unrep/8).



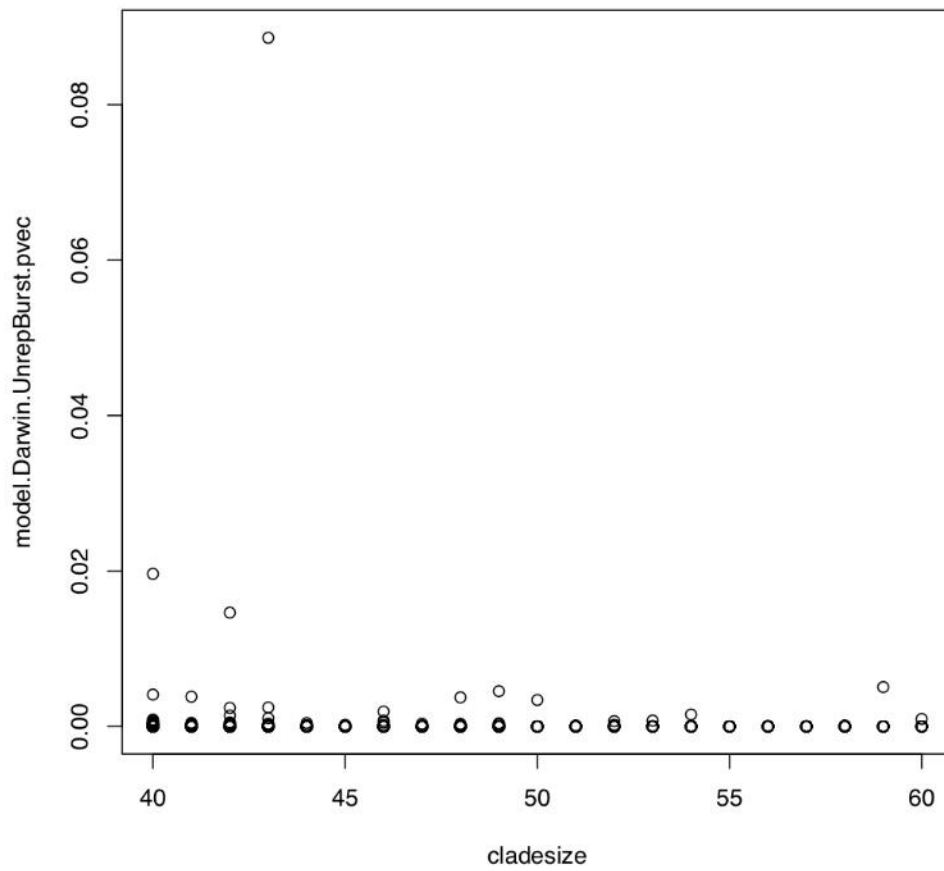


Figure S7: Scatter-plot comparing the P values from the Pagel's discrete model unreplicated burst scenario analysis with the size of the selected clade.

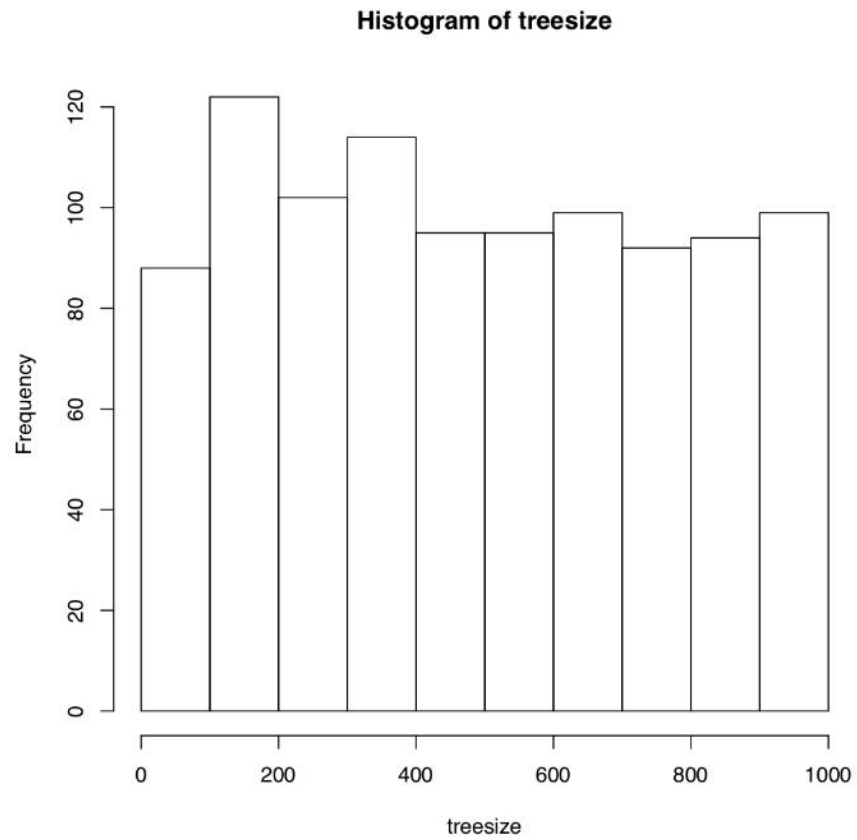


Figure S8: Histogram of tree size from the Pagel's discrete model analysis where tree size was allowed to vary between 50 and 1,000 taxa.

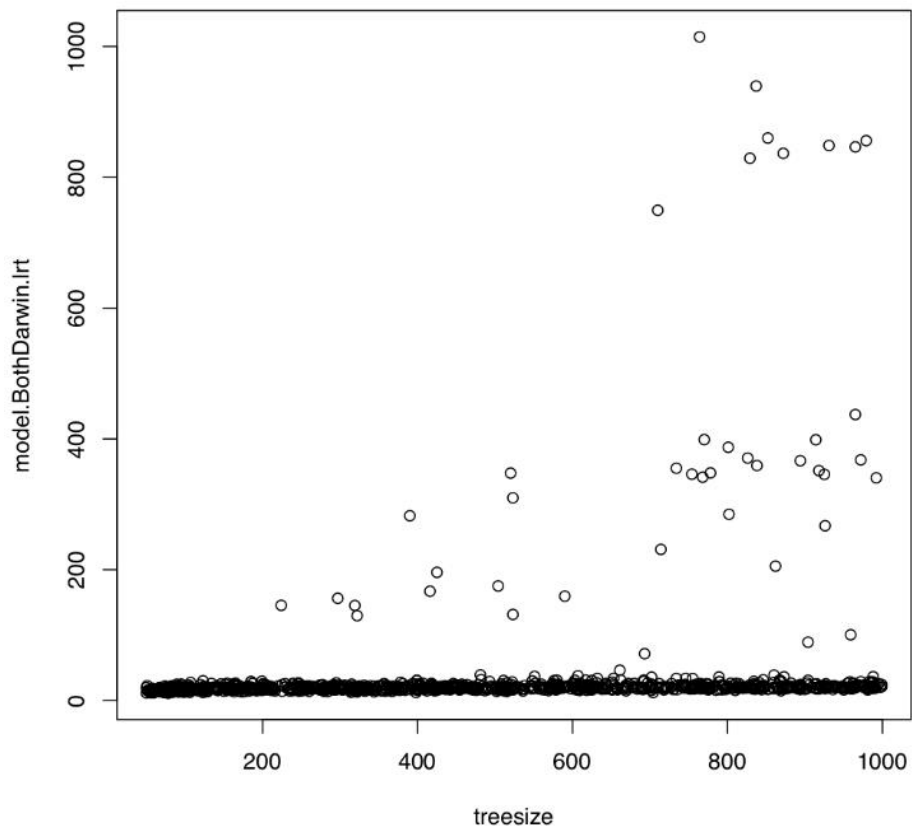


Figure S9: Scatter-plot comparing the likelihood ratio test statistics from the Pagel's discrete model Darwin's scenario analysis with the size of the simulated tree.

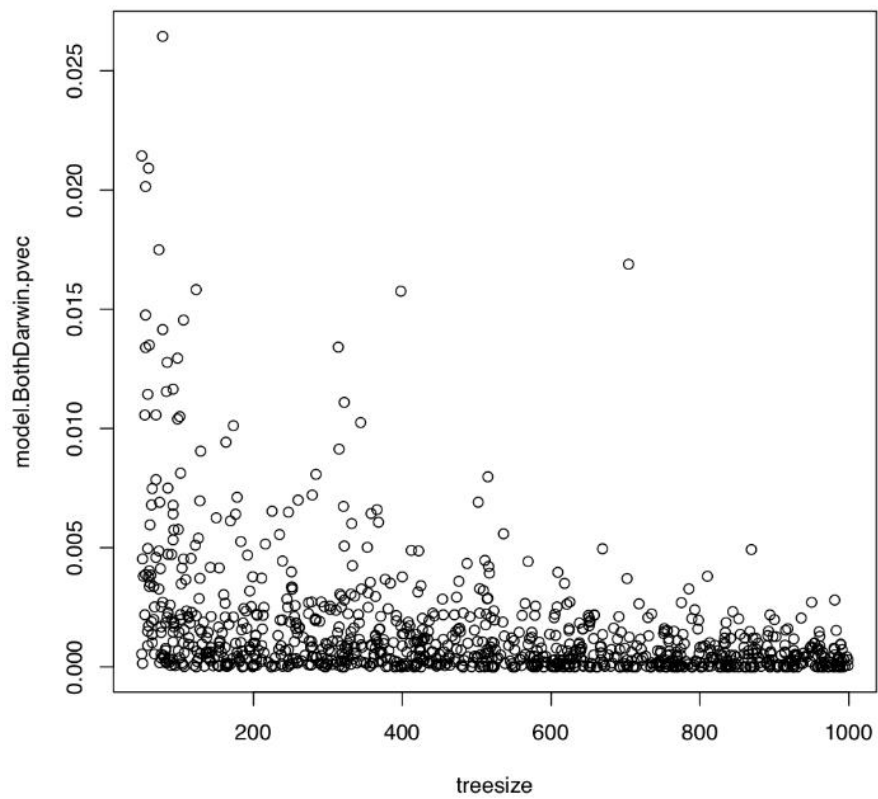


Figure S10: Scatter-plot comparing the P values from the Pagel's discrete model Darwin's scenario analysis with the size of the simulated tree.

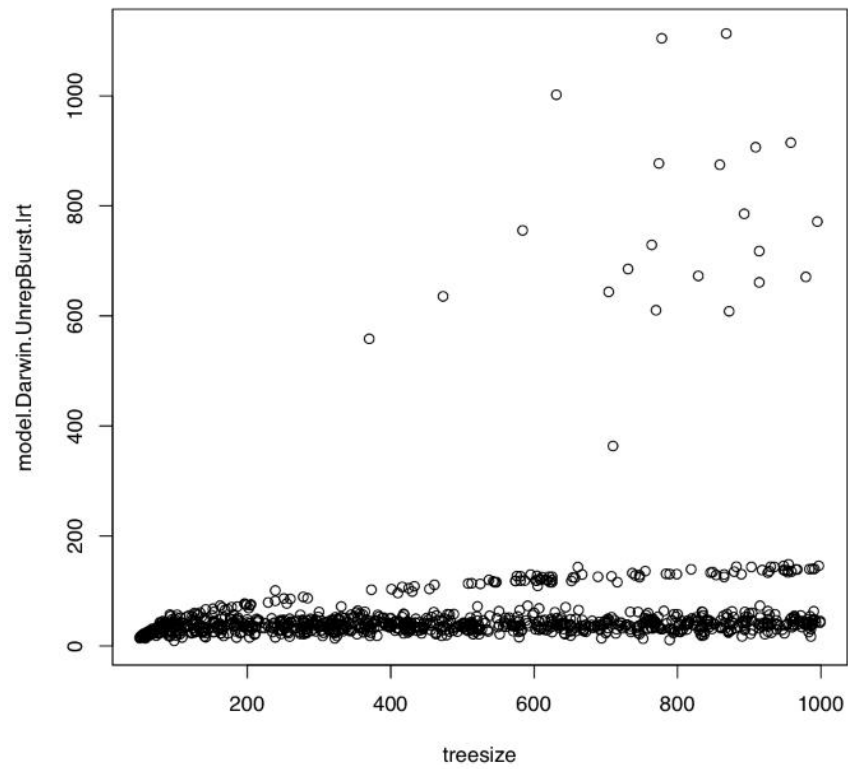


Figure S11: Scatter-plot comparing the likelihood ratio test statistics from the Pagel's discrete model unreplicated burst scenario analysis with the size of the simulated tree.

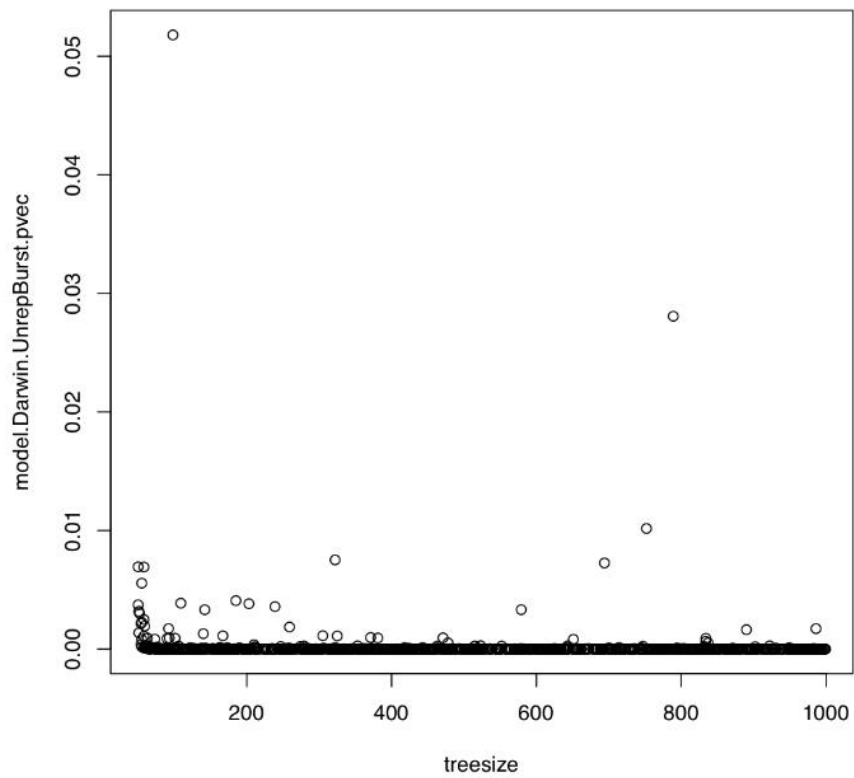


Figure S12: Scatter-plot comparing the P values from the Pagel's discrete model unreplicated burst scenario analysis with the size of the simulated tree.

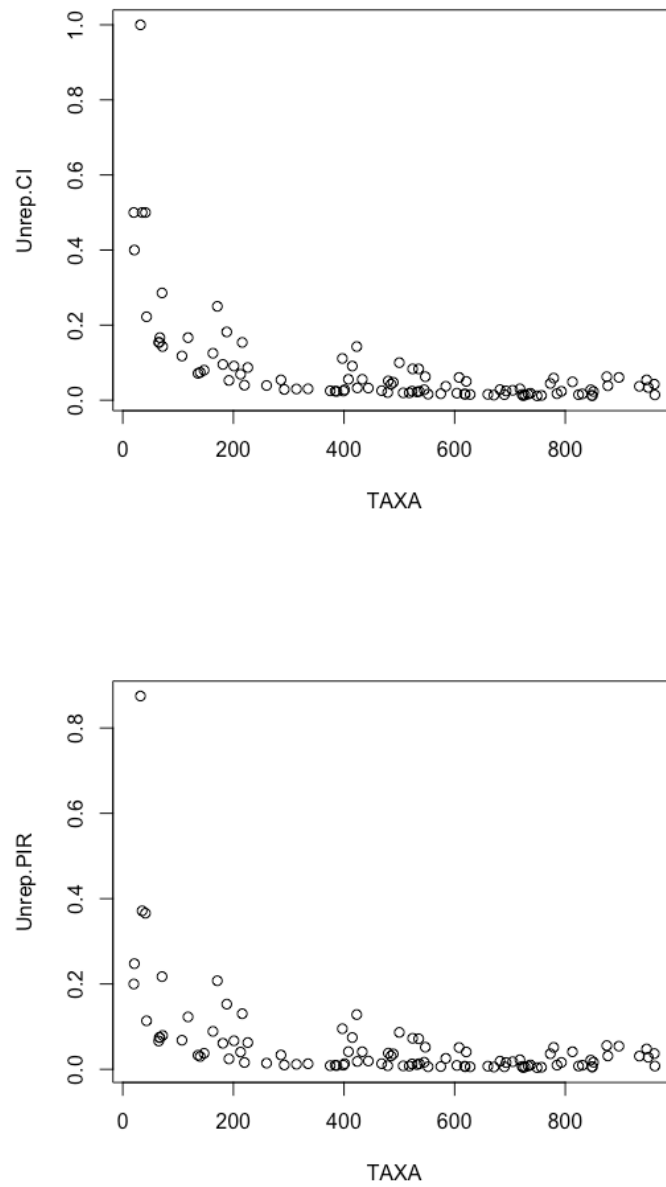


Figure S13: Variation in consistency index (CI) and phylogenetic imbalance ratio (PIR) explained by the number of taxa (tree size) under the unreplicated burst scenario. CI and PIR values exponentially decrease with an increase in tree size and level off between 0.15 and 0.2.

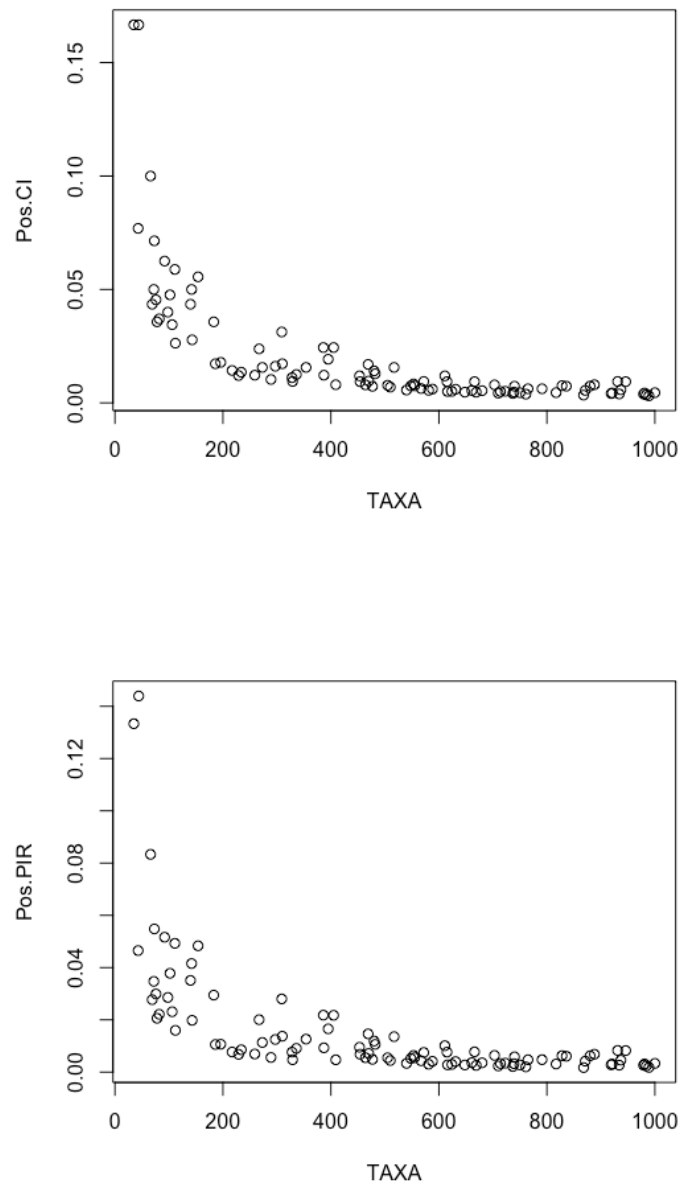


Figure S14: Variation in consistency index (CI) and phylogenetic imbalance ratio (PIR) explained by the number of taxa (tree size) under the positive control scenario. CI and PIR values exponentially decrease with an increase in tree size and level off under values of 0.05 and 0.04, respectively.

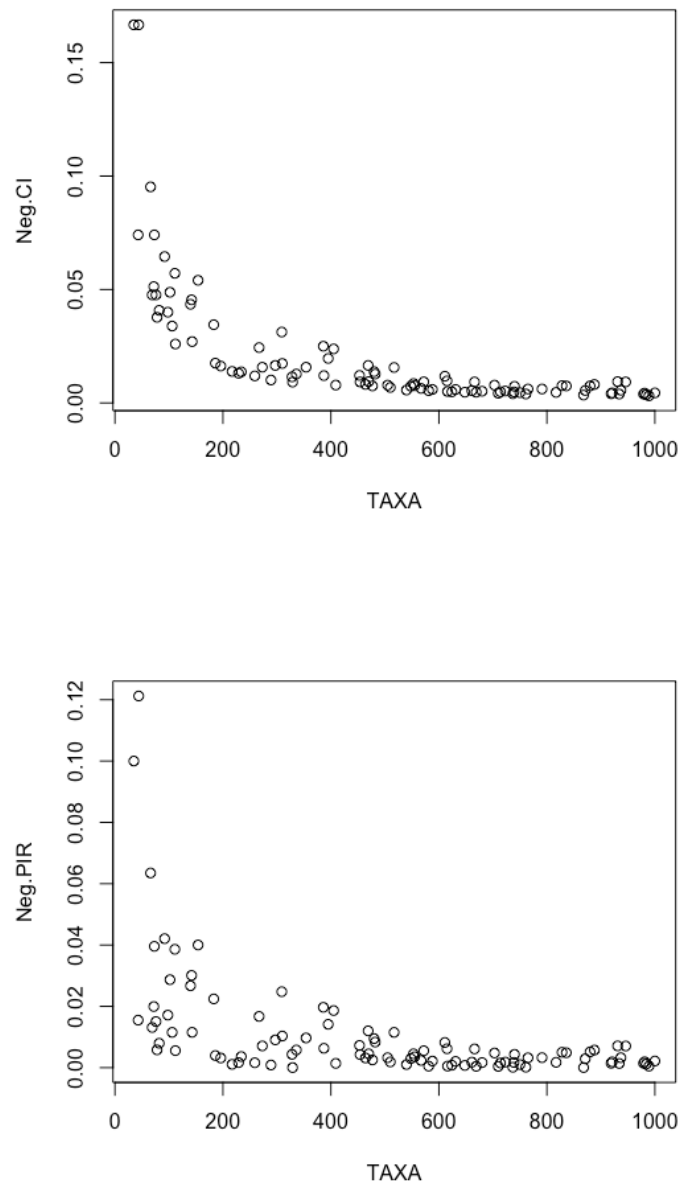


Figure S15: Variation in consistency index (CI) and phylogenetic imbalance ratio (PIR) explained by the number of taxa (tree size) under the negative control scenario. CI and PIR values exponentially decrease with an increase in tree size and level off under values of 0.05 and 0.04, respectively.

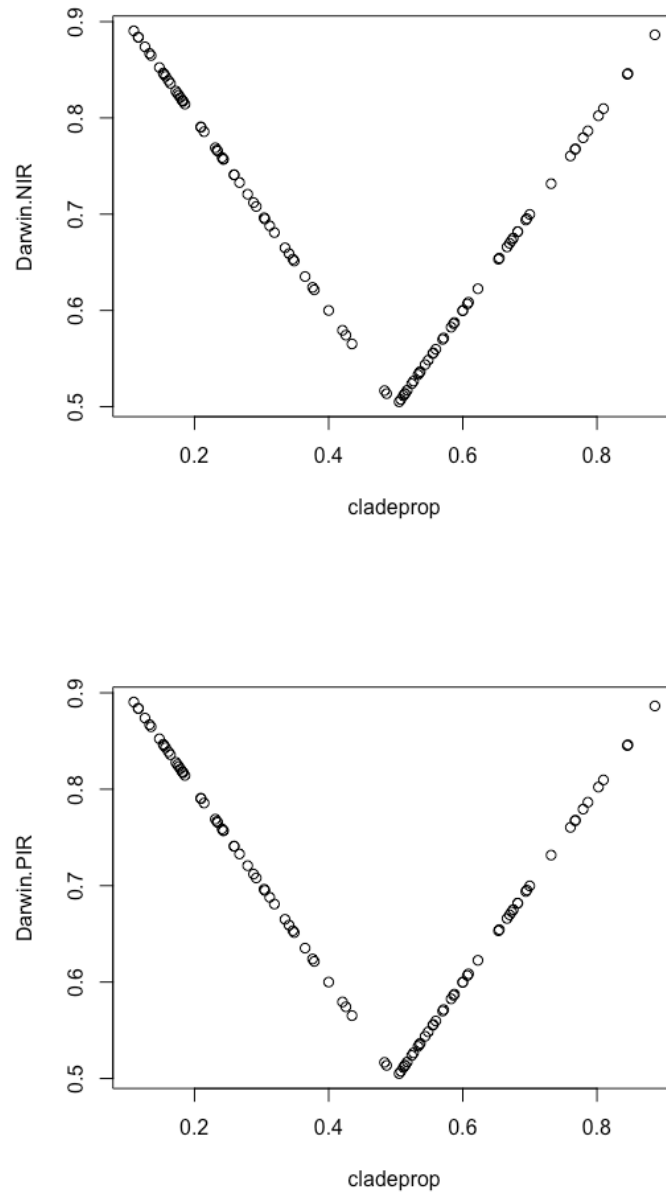


Figure S16: Variation in normalized imbalance ratio (NIR) and phylogenetic imbalance ratio (PIR) explained by the proportion of the selected clade size under Darwin's scenario. NIR and PIR values decrease with an increase in clade proportion until 50% of the taxa are selected.

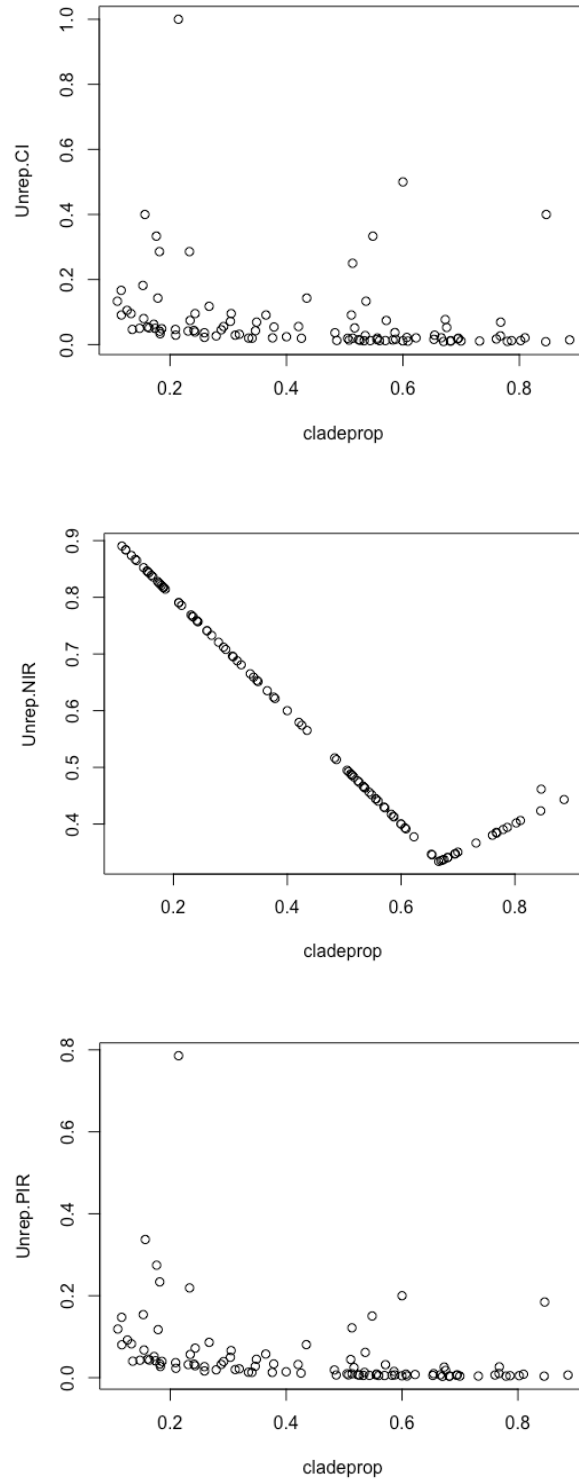


Figure S17: Variation in consistency index (CI), normalized imbalance ratio (NIR), and phylogenetic imbalance ratio (PIR) explained by the proportion of the selected clade size under the unreplicated burst scenario. CI and PIR decrease with an increase in clade proportion, with most values plotting below a value of 0.2. NIR decreases with an increase in clade proportion until about 67% of the taxa are selected.

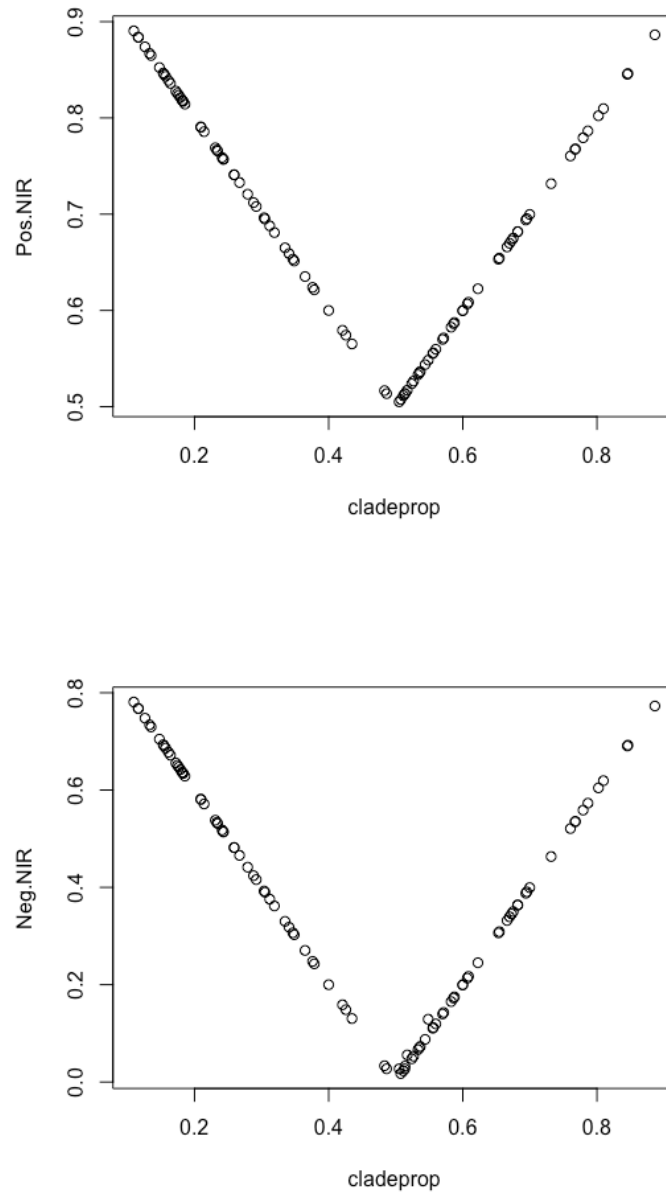


Figure S18: Variation in normalized imbalance ratio (NIR) explained by the proportion of the selected clade size under the positive and negative control scenarios. NIR decreases with an increase in clade proportion until 50% of the taxa are selected.

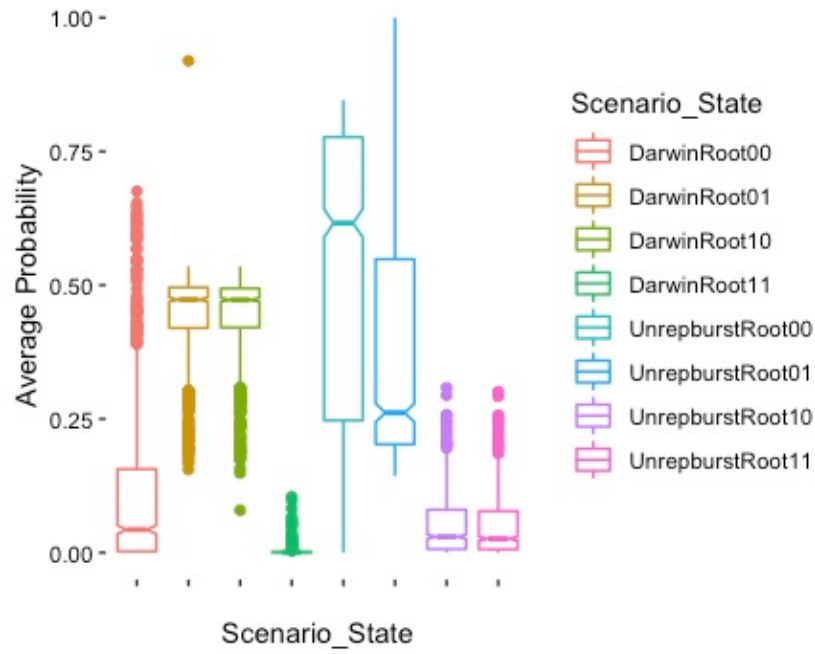


Figure S19: Side-by-side boxplots of root state probability estimates from the 1,000 RJMCMC experimental runs. Prefix indicates the scenario and the numbered suffix represents the root state probability estimate.

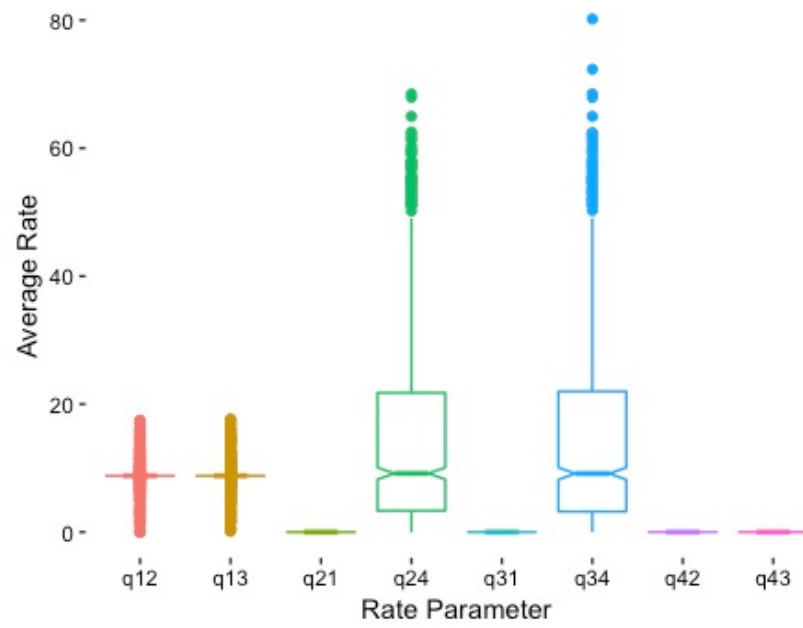


Figure S21: Side-by-side boxplots of average rate parameter estimates from the 1,000 RJMCMC Darwin's scenario runs when node states are fixed and rate of loss parameters are set to 0.

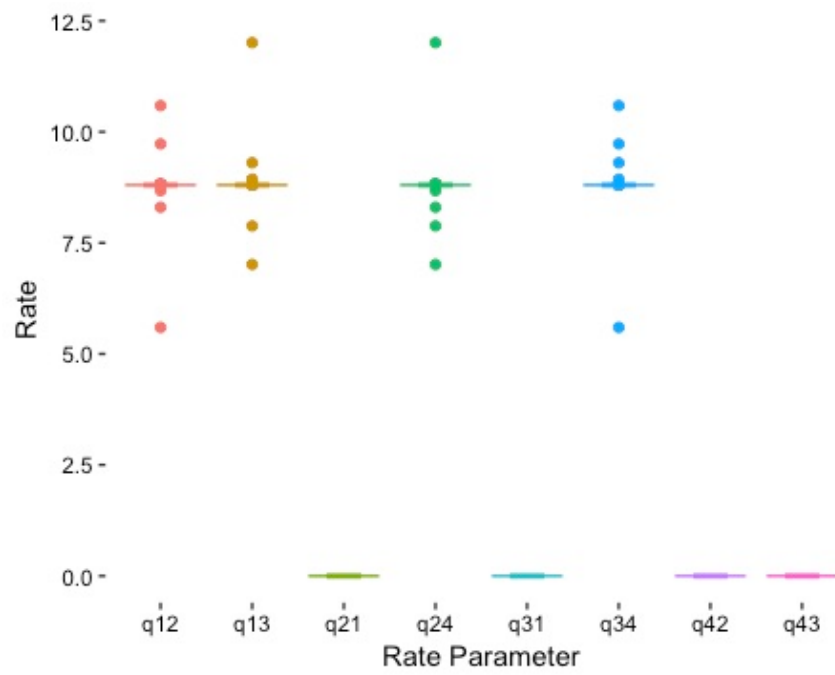


Figure S22: Side-by-side boxplots of rate parameter estimates from the RJMCMC Darwin's scenario run #180 (node states fixed and rate of loss parameters are set to 0).

Chapter 2

Early Tetrapodomorph Biogeography: Controlling for Fossil Record Bias in Macroevolutionary Analyses

(Published as: Gardner, Jacob D., Kevin Surya, and Chris L. Organ. Early Tetrapodomorph Biogeography: Controlling for Fossil Record Bias in Macroevolutionary Analyses. *Comptes Rendus Palevol*, 18(7): 609—709.)

Rendus Palevol, 18(7): 609—709.)

Abstract

The fossil record provides direct empirical data for understanding macroevolutionary patterns and processes. Inherent biases in the fossil record are well known to confound analyses of this data. Sampling bias proxies have been used as covariates in regression models to test for such biases. Proxies, such as formation count, are associated with paleobiodiversity, but are insufficient for explaining species dispersal owing to a lack of geographic context. Here, we develop a sampling bias proxy that incorporates geographic information and test it with a case study on early tetrapodomorph biogeography. We use recently-developed Bayesian phylogeographic models and a new supertree of early tetrapodomorphs to estimate dispersal rates and ancestral habitat locations. We find strong evidence that geographic sampling bias explains supposed radiations in dispersal rate (potential adaptive radiations). Our study highlights the necessity of accounting for geographic sampling bias in macroevolutionary and phylogenetic analyses and provides an approach to test for its effect.

1. Introduction

Our understanding of macroevolutionary patterns and processes are fundamentally based on fossils. The most direct evidence for taxonomic origination and extinction rates come from the rock record, as do evidence for novelty and climate change unseen in data sets gleaned from extant sources. There are no perfect data sets in science; there are inherent limitations and biases in the rock record that must be addressed when we form and test paleobiological hypotheses. For instance, observed stratigraphic ranges of fossils can mislead inferences about diversification and extinction rates (Raup and Boyajian, 1988; Signor and Lipps, 1982). Observed species diversity is also known to increase with time due to the preferential preservation and recovery of fossils in younger geological strata—referred to as “the Pull of the Recent” (Jablonski et al., 2003). Large and long-surviving clades with high rates of early diversification tend to result in an illusory rate slow-down as diversification rates revert back to a mean value—referred to as “the Push of the Past” (Budd and Mann, 2018). Paleobiologists test and account for these biases when analyzing diversification and extinction at local and global scales (Alroy et al., 2001; Benson and Butler, 2011; Benson and Upchurch, 2013; Benson et al., 2010, 2013; Foote, 2003; Jablonski et al., 2003; Koch, 1978; Lloyd, 2012; Sakamoto et al., 2016a, 2016b). These bias-detection and correction techniques include fossil occurrence subsampling (Alroy et al., 2001; Dunne et al., 2018; Close et al., 2019; Jablonski et al., 2003; Lloyd, 2012); correcting origination, extinction, and sampling rates using evolutionary predictive models (Foote, 2003); the use of residuals from diversity-sampling models (Benson et al., 2010; Benson and Upchurch, 2013; Sakamoto et al., 2016b); and the incorporation of sampling bias proxies as covariates in regression models (Benson et al., 2010; Benson and Butler, 2011; Benton et al., 2013; Sakamoto et al., 2016a).

Benton et al. (2013), studying sampling bias proxies, demonstrated that diversity through time closely tracks formation count (Benton et al., 2013).

However, case studies in England and Wales suggest that proxies for terrestrial sedimentary rock volume (such as formation count) do not accurately explain paleobiodiversity, particularly if the fossil record is patchy (Dunhill et al., 2013, 2014a, 2014b). Marine outcrop area and paleoecological-associated facies changes are, however, associated with shifts in paleobiodiversity (Dunhill et al., 2013, 2014b). Moreover, Benton et al. (2013) argue that the direction of causality between paleobiodiversity and formation count is unclear; there may be a common cause to explain their covariation, such as sea level (Benton et al., 2013). Nonetheless, formation count is a widely-used sampling bias proxy in phylogenetic analyses of macroevolution (O'Donovan et al., 2018; Sakamoto et al., 2016a, 2016b; Tennant et al., 2016a, 2016b). The advent of computational modeling approaches, particularly phylogenetic comparative methods, has made it easier to include proxies, like formation count, into models. Additional sampling bias proxies used in these studies include occurrence count, valid taxon count, and specimen completeness and preservation scores. Absent from these proxies is geographic context, which could confound many types of macroevolutionary analyses.

Despite advancements made in understanding the origin and evolution of early tetrapodomorphs, biogeographical studies are hindered by the incompleteness of the early tetrapodomorph fossil record. For example, “Romer’s Gap” represents a lack of tetrapodomorph fossils from the end-Devonian to mid-Mississippian, a period crucial for understanding early tetrapodomorph diversification. Recent collection efforts recovered tetrapodomorph specimens from “Romer’s Gap”, suggesting that a collection and preservation bias explains this gap (Clack et al., 2017; Marshall et al.,

2019). In addition, a trackway site in Poland demonstrates the existence of digit-bearing tetrapodomorphs 10 million years before the earliest elpistostegalian body fossil, showcasing the limitation of body fossils to reveal evolutionary history (Niedzwiedzki et al., 2010). A recent study by Long et al. (2018) leveraged phylogenetic reconstruction of early tetrapodomorphs to frame hypotheses about the origin of major clades, as well as their dispersal patterns, including the hypothesis that stem tetrapodomorphs dispersed from eastern Gondwana to Euramerica. However, this study did not use phylogenetic comparative methods to estimate ancestral geographic locations or to model dispersal patterns.

Here, we present a phylogeographic analysis of early tetrapodomorphs. Our goals are: (1) to construct a phylogenetic supertree of early tetrapodomorphs that synthesizes previous phylogenetic reconstructions; (2) to estimate the paleogeographic locations of major early tetrapodomorph clades using recently-developed phylogeographic models that account for the curvature of the Earth; and (3) to test for the influence of geographic sampling bias on dispersal rates. Our results indicate that geographic sampling bias substantially confounds analyses of dispersal and paleogeography. We conclude with a discussion about the necessity of controlling for fossil record biases in macroevolutionary analyses.

2. Materials and Methods

2.1. Nomenclature

Tetrapoda has been informally defined historically to include all terrestrial vertebrates with limbs and digits (Laurin, 1998). Gauthier et al. (1989) first articulated a phylogenetic definition of Tetrapoda as the clade including the last common ancestor of amniotes and lissamphibians. This definition excludes stem-tetrapodomorphs, like

Acanthostega and *Ichthyostega*. Stegocephalia was coined by E.D. Cope in 1868 (Cope, 1868), but was more recently used to describe fossil taxa more closely related to tetrapods than other sarcopterygians. A recent cladistic redefinition of Stegocephalia includes all vertebrates more closely related to temnospondyls than *Panderichthys* (Laurin, 1998). Here, we use the definitions of Laurin (1998) for a monophyletic Stegocephalia and of Gauthier et al. (1989) for Tetrapoda, which refers specifically to the crown group. We use Tetrapodomorpha to refer to all taxa closer to the tetrapod crown-group than the lungfish crown-group (Ahlberg, 1998). We additionally use Elpistostegalia (= Panderichthyida) to refer to the common ancestor of all stegocephalians and Panderichthys as well as Eotetrapodiformes to refer to the common ancestor of all tristichopterids, elpistostegalians, and tetrapods (Coates and Friedman, 2010).

2.2. Supertree

We inferred a supertree of 69 early tetrapodomorph taxa from five edited, published morphological data matrices, focusing on tetrapodomorphs whose previously inferred phylogenetic position bracket the water–land transition (Clack et al., 2017; Friedman et al., 2007; Pardo et al., 2017; Swartz, 2012; Zhu et al., 2017). Since downstream analyses might be sensitive to unequal sample sizes between taxa pre- and post-water–land transition, we did not include several crownward stem-tetrapodomorphs from the original matrices (Supplementary Material). For each matrix, we generated a posterior distribution of phylogenetic trees using MrBayes 3.2.6 (Ronquist et al., 2012b). In each case, we ran two Markov chain Monte Carlo (MCMC) replicates for 20,000,000 generations with 25% burn-in, each with four chains and a sampling frequency of 1000. We used one partition, except for Clack et al.’s (2017) matrix, which

was explicitly divided into cranial and postcranial characters. To time-calibrate the trees, we constrained the root ages and employed a tip dating approach (Ronquist et al., 2012a). Tip dates (last occurrence) were acquired from the Paleobiology Database (PBDB; <https://www.paleobiodb.org/>) and the literature (Supplementary Table 2). Root calibrations (minimum and soft maximum age estimates) were collected from the PBDB and Benton et al. (2015). We also used the fossilized birth-death model as the branch length prior (Didier and Laurin, 2018; Didier et al., 2012, 2017; Gavryushkina et al., 2014; Heath et al., 2014; Stadler, 2010; Zhang et al., 2016). All pairs of MCMC replicates converged as demonstrated by low average standard deviation of split frequencies (< 0.005 ; Lakner et al., 2008; Supplementary Table 3).

Next, we used the five maximum clade credibility trees (source trees; Appendix 1 Figs. S1–10) to compute a distance supermatrix using SDM 2.1 (Criscuolo et al., 2006). We then inferred an unweighted neighbor-joining tree (UNJ by Gascuel, 1997) from the distance supermatrix using PhyD* 1.1 (Criscuolo and Gascuel, 2008). The UNJ* algorithm is preferable for matrices based on morphological characters. Unlike most supertree methods, the SDM-PhyD* combination produces a supertree with branch lengths. We rooted the supertree using phytools 0.6.60 (Revell, 2012) by adding an arbitrary branch length of 0.00001 to break the trichotomy at the basal-most node in R 3.5.2 (R Core Team, 2018), designating the dipnomorph *Glyptolepis* as the outgroup. We qualitatively compared the supertree topology with the published source trees and Marjanovic' and Laurin's (2019). We also calculated normalized Robinson-Foulds (nRF) distances (Robinson and Foulds, 1981) using phangorn 2.4.0 (Schliep, 2011) in R to assess the congruency of topologies. In each comparison, polytomies in the supertree or the source tree were resolved in all possible ways using phytools. We

then calculated all nRF distances and took an average (Supplementary Table 4). The supplementary materials include a more detailed description of this approach.

2.3. Phylogeography

We obtained paleocoordinate data (paleolatitude and paleolongitude) for 63 early tetrapodomorphs from the PBDB using the GPlates software setting (<https://www.gws.gplates.org/>). By default, GPlates estimates paleocoordinates from the midpoint of each taxon's age range. Among the 63 taxa sampled, 16 did not have direct paleocoordinate data in the PBDB. For these taxa, we searched for the geological formations and geographic regions within the time range from which they are known and averaged the paleolocations across each valid taxonomic occurrence in the PBDB. If the paleolocation of the formation was not listed in the PBDB, we used published geographic locations of the formations. This level of precision is adequate for world-wide phylogeographic analyses, such as conducted here. Present-day coordinates for these geographic locations were obtained from Google Earth and matched with PBDB entries that date within each taxon's age range (Supplementary Table 5). Four additional taxa, *Kenichthys*, *Koilops*, *Ossirarus*, and *Tungsenia*, had occurrences in the PBDB but the GPlates software could not estimate their paleocoordinates. For *Koilops* and *Ossirarus*, we used all tetrapodomorph occurrences from the Ballagan Formation of Scotland, UK—a formation in which these two taxa are found (Clack et al., 2017). For *Kenichthys* and *Tungsenia*, we calculated paleocoordinate data from the GPlates website directly using the present-day coordinates from the PBDB (<https://www.gws.gplates.org/#recon-p>). This approach did not work for the 16 previously mentioned taxa (Supplementary Table 5). We, therefore, obtained paleocoordinate data from nearby entries in the PBDB that date

within each taxon's age range. We excluded the following taxa from our analyses due to the lack of data and comparable entries in the PBDB: *Jarvikina*, *Koharalepis*, *Spodichthys*, and *Tinirau*. We excluded the outgroup taxon, *Glyptolepis*, in our analysis to focus on the dispersal trends within early Tetrapodomorpha. We also excluded *Eusthenodon* and *Strepsodus* because their high estimated dispersal rates—being reported from multiple continents—masked other rate variation throughout the phylogeny and inhibited our downstream analyses from converging on a stable likelihood. We do, however, discuss their geographic implications in Section 4.

A model that incorporates phylogeny is crucial for paleobiogeographic reconstruction because it accounts for both species relationships and the amount of evolutionary divergence (branch lengths). Using continuous paleocoordinate data, rather than discretely-coded regions, allows dispersal trends to be estimated at finer resolutions. Discretely-coded geographic regions also limit ancestral states to the same regions inhabited by descendant species. However, standard phylogenetic comparative methods for continuous data assume a flat Earth because they do not account for spherically structured coordinates (i.e., the proximity of -179° and 179° longitudes). Recently-developed phylogenetic comparative methods for modeling continuous paleocoordinate data, implemented as the 'geo' model in the program BayesTraits V3, overcome this hurdle by "evolving" continuous coordinate data on the surface of a globe (O'Donovan et al., 2018). The model is implemented with a Bayesian reversible jump MCMC algorithm to estimate rates of geographic dispersal and ancestral paleolocations simultaneously. To account for the spheroid shape of the globe, the 'geo' model converts latitude and longitude data into three-dimensional coordinates while prohibiting moves that penetrate the inside of the globe. Ancestral

states, which are converted back to standard latitude and longitude, are estimated for each node of the phylogeny. The method includes a variable rates model to estimate variation in dispersal rate (Venditti et al., 2011). The 'geo' model makes no assumptions about the location of geographic barriers or coastlines, but a study on dinosaur biogeography found 99.2% of mean ancestral state reconstructions to be located within the bounds of landmasses specific to the time at which they occurred (O'Donovan et al., 2018). We ran three replicate independent analyses using the Bayesian phylogenetic 'geo' model for 100 million iterations each with a 25% burnin and sampling every 1000 iterations. We estimated log marginal likelihoods using the Stepping Stone algorithm with 250 stones sampling every 1000 iterations (Xie et al., 2011). We used Bayes factors (BF) to test whether a variable rates model explained the data better than a uniform rate model. Bayes factors greater than two are considered good evidence in support of the model with the greater log marginal likelihood. We compared estimated rate scalars and ancestral states among the three independent variable rates analyses to check for consistency in our results. Rates of dispersal were estimated for each branch by dividing the average rate scalars by the original branch lengths (scaled by time). We assessed the MCMC convergence of all analyses using Tracer 1.7 (Rambaut et al., 2018).

To test for the effect of sampling bias on dispersal rates, we developed a sampling bias proxy that incorporates geographic context: a regional-level formation count. Formation counts are meant to capture multiple biases: uneven global rock exposure, uneven fossil collection and database efforts, and global variation in sediment deposition in environments conducive to preservation. Stage-level (stage-specific) formation count represents the mean number of formations, or distinct rock units, globally known to produce relevant fossils along each terminal branch of a

phylogeny. Following the protocol of Sakamoto et al. (2016a) and O'Donovan et al. (2018), stage-level formation counts are calculated by taking the average number of formations known from each geological stage (age) across the globe that encompass the time period between the taxon's tip date and its preceding node. These average stage-level formation counts are weighted by the proportion that each terminal branch length covers each geological stage. For example, if a terminal branch covers two geological stages (e.g., Frasnian and Famennian) at 30% and 70%, respectively, then the formation counts from each geological stage are weighted by those proportions and then divided by the number of geological stages covered:

$$\text{Stage-Level Formation Count} = \frac{\text{Frasnian Count} \times 0.3 + \text{Famennian Count} \times 0.7}{2}$$

Stage-level formation count is not informed by geography; it is a global metric. It is therefore an inadequate proxy if bias has a strong geographic component (e.g., if the majority of formations recorded are from a specific region or if few formations are exposed within a region). The number of fossil-bearing geological formations, accounting for geographic distribution, is expected to be an important confounding bias in the fossil record. We developed a proxy that includes geographic sampling bias. Our approach breaks down stage-level formation count by geographic region. To account for the arrangement of the continents during the Devonian, Carboniferous, and Permian, we recognized five major regions: northern Euramerica (including northeastern Eurasia and central Asia), southern Euramerica (North America, Greenland, and western Europe), western Gondwana (South America and Africa), eastern Gondwana (Antarctica, Australia, and southern Asia), and East Asia (e.g., China). These regions generally resemble traditional bioregionalizations of the

Devonian period, but note that regions based on biotic similarities of fossil assemblages are known to change through time (Dowding and Ebach, 2019). Future studies could modify this approach to capture temporal changes in biotic connectivity. For each branch in the phylogeny, we used the average ancestral state and taxon paleolocation estimates to determine if the branch crossed multiple geographic regions. The number of formations within this time window are totaled for every region covered by the branch and then divided by the number of regions covered. For example, if ancestral state estimates at node 1 and 2 are located in eastern Gondwana and southern Euramerica, respectively, then the number of formations recorded in eastern Gondwana, southern Euramerica, and the regions in between (i.e., western Gondwana or northern Euramerica + East Asia) are counted for that geological stage; this total is then divided by the number of geographic regions covered by the entire branch (three for the western Gondwana route and four for the northern Euramerica + East Asia route). If the dispersal path between two consecutive ancestral states does not cross any of the five regions, then the number of formations in the inhabited region is counted alone. Fig. 1 illustrates an example of how this proxy is measured. This results in the average number of formations present along the dispersal path (at geographic region scale) for each branch in the phylogeny. As with stage-level formation counts, the regional-level formation counts are weighted by the proportion that the branch length covers each geological stage. We hypothesize that dispersal rate will inversely correlate with regional-level formation count because we expect that the lack of formations in intermediate regions will lead to inflated dispersal rates. The 'geo' model will increase the dispersal rate along a branch to account for the geographic variation observed when there is a lack of intermediate geographic fossil occurrences. This hypothesis can be falsified if high dispersal rates are associated

with larger average numbers of formations along dispersal paths. Benton et al. (2013) provide a global sample of tetrapod-bearing rock formations known for each geological stage from the Middle Devonian through the Triassic. We supplemented these lists with stratigraphic units known to produce sarcopterygian fossils entered in the PBDB (collected on December 10th, 2018).

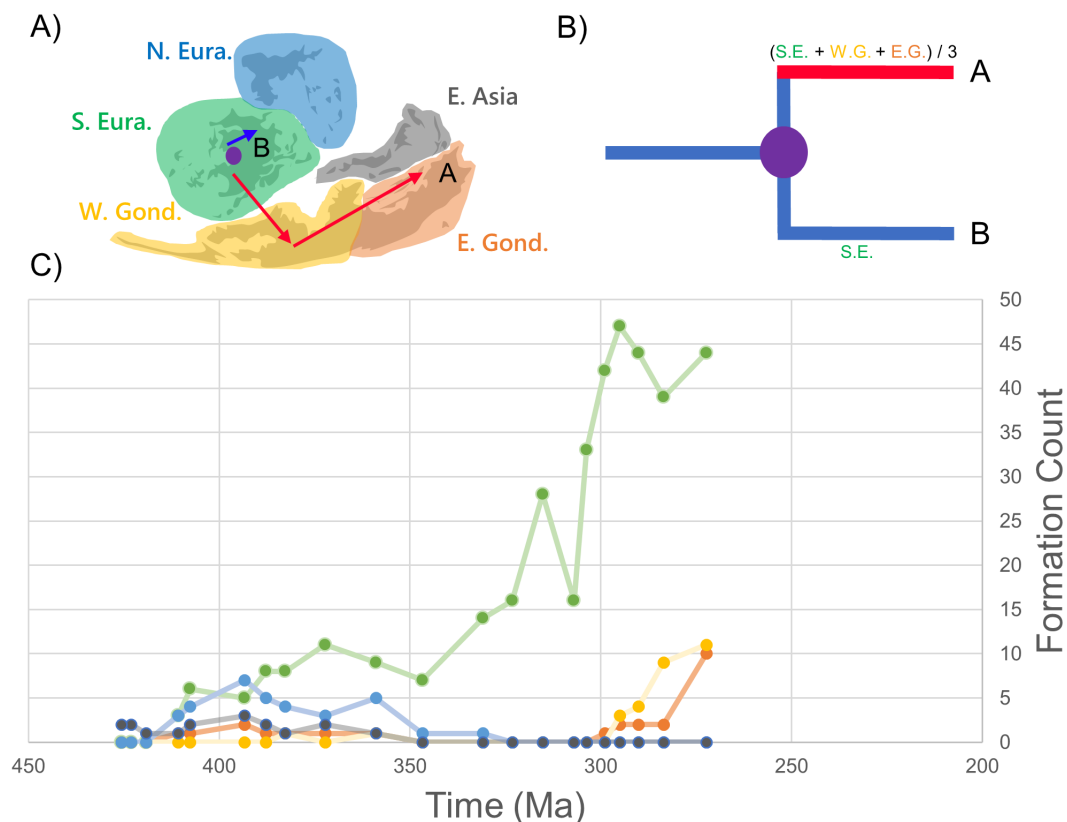


Fig. 1. Example of how the regional-level formation count proxy is calculated. A. Five major geographic regions are highlighted by color in the Devonian map. Red arrows represent a branch-specific dispersal path to species A, beginning in southern Euramerica and ending in eastern Gondwana. The blue arrow represents the dispersal path to species B. The Devonian map is modified and reproduced with permission from © 2016 Colorado Plateau Geosystems Inc. B. The phylogeny of species A and B scaled by time, with equal branch lengths to both species, and colored to represent the rate of dispersal (red is fast, blue is slow). For every

branch of the tree, the number of formations is counted for every region and for each geological stage covered by the dispersal pathway. It is then weighted by the number of geological stages and geographic regions covered. Under the western Gondwana route scenario, the branch to species A covers three geographic regions, while the branch to species B only covers one. Assuming both branches cover only one geological stage, the high dispersal rate for species A can be explained by the lack of recorded geological formations in western Gondwana. C. A line plot of the formation counts through time, colored by geographic region according to the Devonian map above, shows temporal and geographic variability.

Period	Epoch	Age	Min.	Northern Euramerica	Southern Euramerica	Western Gondwana	Eastern Gondwana	Total
			Time (Ma)					
Permian	Cisuralian	Kungurian	272.95	10	44	11	0	65
Permian	Cisuralian	Artinskian	283.5	2	39	9	0	50
Permian	Cisuralian	Sakmarian	290.1	2	44	4	0	50
Permian	Cisuralian	Asselian	295	2	47	3	0	52
Pennsylvanian	Late	Gzhelian	298.9	1	42	0	0	43
Pennsylvanian	Late	Kasimovian	303.7	0	33	0	0	33
Pennsylvanian	Middle	Moscovian	307	0	16	0	0	16
Pennsylvanian	Early	Bashkirian	315.2	0	28	0	0	28
Mississippian	Late	Serpukhovian	323.2	0	16	0	0	16
Mississippian	Middle	Visean	330.9	0	14	0	1	15
Mississippian	Early	Tournaisian	346.7	0	7	0	1	8
Devonian	Late	Famennian	358.9	1	9	1	6	17
Devonian	Late	Frasnian	372.2	1	11	0	5	17
Devonian	Middle	Givetian	382.7	1	8	1	5	15
Devonian	Middle	Eifelian	387.7	1	8	0	7	16
Devonian	Early	Emsian	393.3	2	5	0	8	15
Devonian	Early	Pragian	407.6	1	6	0	6	13

Devonian	Early	Lochkovian	410.8	1	3	0	4	8
Silurian	Pridoli	Pridoli	419.2	0	0	0	1	1
Silurian	Ludlow	Ludfordian	423	0	0	0	2	2
Silurian	Ludlow	Gorstian	425.6	0	0	0	2	2

Table 1: Regional- and stage-level (Total) formation counts.

To test for the effect of regional-level formation count bias on dispersal rate, we conducted a non-parametric two sample, upper-tailed Mann-Whitney U-test using the base package ‘stats’ in R (R Core Team, 2018). This approach ranks all branches of the phylogeny by their regional level formation count and tests if the branches with lower dispersal rates rank higher on average than branches with higher rates. We define “high” vs “low” dispersal rates based on whether or not they are two standard deviations greater than the average rate across the tree. Due to the vast difference in sample size between the two groups (“high rates”: $n = 9$, “low rates”: $n = 111$), we bootstrapped the regional-level formation counts from each group with 100,000 replicates. From this bootstrap analysis, we obtained a 95% confidence interval for the summed ranks of the branches with low dispersal rates ($n = 100,000$ U-statistic values). The expected U-statistic is 499.5 given the null hypothesis that only 50% of the regional-level formation counts along branches with low rates rank higher than the formation counts with high rates (half of all possible combinations = $\frac{9 \times 111}{2}$). A 95% confidence interval of bootstrapped U-statistics that does not include the null expected U-statistic is considered good evidence for higher mean dispersal rates along branches with lower regional-level formation counts. The full dataset and code for the phylogeographic analyses can be requested by email to the corresponding author.

Estimated ancestral states do not identify specific dispersal routes, so we conducted sensitivity analyses to test if the dispersal route chosen for counting formations influenced our results. We conceived of three scenarios for dispersal routes between eastern Gondwana and southern Euramerica or vice versa: (1) a dispersal route through western Gondwana; (2) a route through northern Euramerica and East Asia; and (3) a direct route between eastern Gondwana and southern Euramerica. For the first scenario, we averaged the number of formations found in eastern and western Gondwana and southern Euramerica for a given time period. The second scenario is similar to the first but included formation counts from northern Euramerica and East Asia in place of western Gondwana. The third scenario only averaged formation counts from eastern Gondwana and southern Euramerica.

3. Results

3.1. Supertree

Topological differences resulted among our supertree, the published source trees, and Marjanovic' and Laurin's (2019) tree (Fig. 2). In our tree, a polyphyletic "Megalichthyiformes" is the basal-most tetrapodomorph group instead of Rhizodontida (Swartz, 2012; Zhu et al., 2017). Canowindrids and rhizodontids formed an unexpected sister clade to Eotetrapodiformes. Clack et al.'s (2017) five Tournaisian tetrapod taxa cluster together. Colosteidae is rootward of *Crassigyrinus*. *Caerorhachis* is next to Baphetidae. Baphetidae moved crownward compared to previous topologies (likely because of a small character sample size [Marjanovic' and Laurin, 2019]). Two crownward nodes are unresolved (polytomous). We retained *Tungsenia* and *Kenichthys* as the oldest and second oldest tetrapodomorphs. Tristichopteridae, Elpistostegalia, Stegocephalia, Aïstopoda, Whatcheeriidae, Colosteidae,

Anthracosauria, Dendrerpetidae, and Baphetidae remain monophyletic. Aïstopoda (*Lethiscus* and *Coloraderpeton*) fell rootward to Tetrapoda as reported in Pardo et al. (2017). The average nRF distances quantify differences in topology (Supplementary Table 4). On average, there are 39.7% different or missing bipartitions in the source trees compared to the supertree.

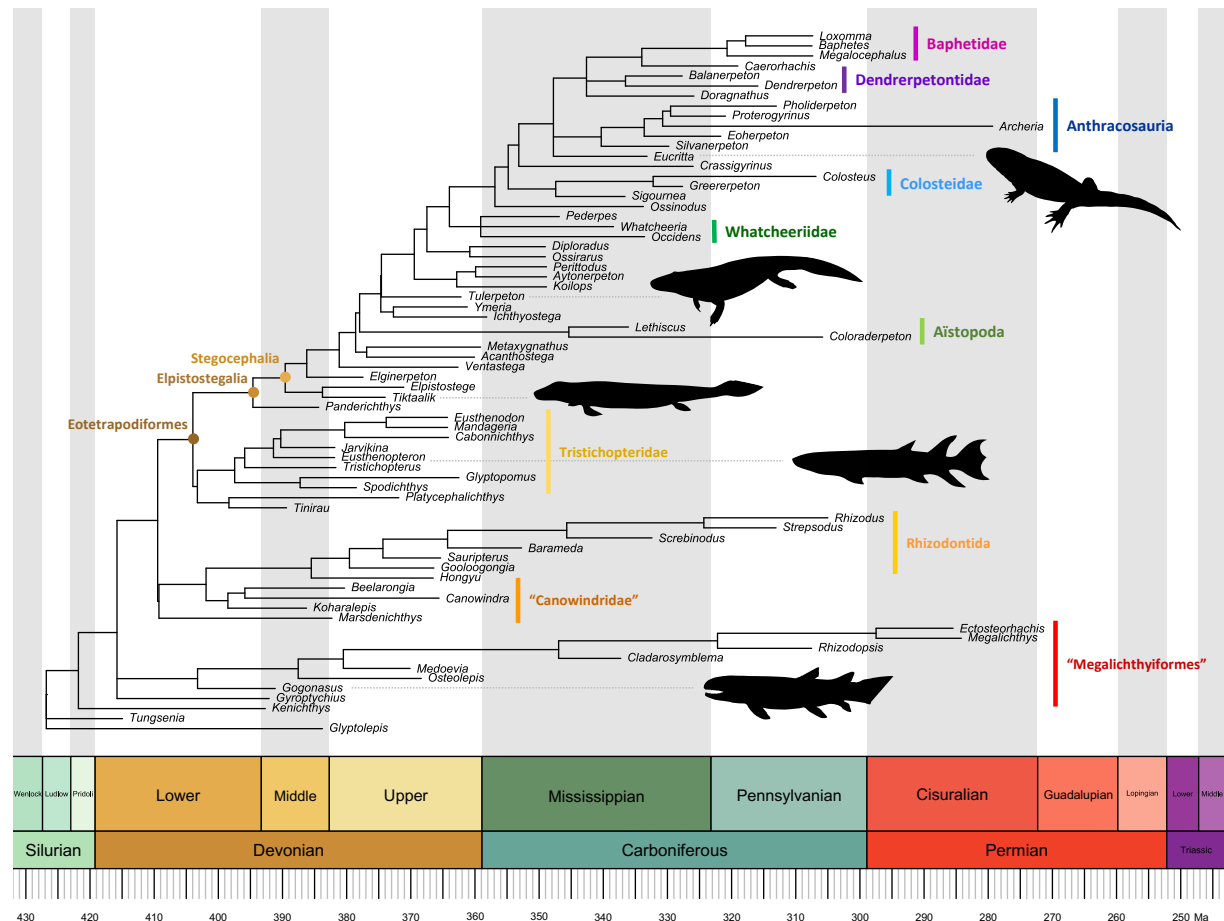


Fig. 2. The time-scaled tetrapodomorph supertree. Taxonomic groups in quotes are not monophyletic. Here, *Glyptolepis*, a dipnomorph, is the outgroup. We downloaded the silhouettes from phylopic.org: *Eucritta* and *Greererpeton* by Dmitry Bogdanov (vectorized by T. Michael Keesey), *Eusthenopteron* by Steve Coombs (vectorized by T. Michael Keesey), and *Gogonasus* and *Tiktaalik* by Nobu Tamura (CC BY-SA 3.0).

3.2. *Phylogeography*

We found overwhelming support for a variable rates model of geographic dispersal in early tetrapodomorphs (BF = 632.3; Fig. 3). The estimated rates across the three replicate runs are consistent (out of 122 branches, only three had a median rate scalar with an absolute value difference among the three runs greater than 3). All rate shifts that were two standard deviations greater than the average dispersal rate were reconstructed dispersal events moving from East Asia to southern Euramerica, from eastern Gondwana to southern Euramerica, or southern Euramerica to eastern Gondwana. The fastest estimated dispersal rate occurs along the branch leading to Eotetrapodiformes, moving from eastern Gondwana to southern Euramerica ($14.34^\circ \times$ the average rate). As Long et al. (2018) suggest, we find evidence for an East Asian origin for Tetrapodomorpha but with moderate uncertainty (average estimate \pm standard deviation of posterior distribution; longitude avg = $81.5^\circ \pm 10.1^\circ$, latitude avg = $-6.4^\circ \pm 8.5^\circ$). We also reconstruct an origin for “Megalichthyiformes” that borderlines East Asia and eastern Gondwana (longitude avg = $107.2^\circ \pm 14.1^\circ$, latitude avg = $-22.6^\circ \pm 8.7^\circ$), along with an eastern Gondwana origin for the clade uniting “Canowindridae” and Rhizodontida (longitude avg = $137.1^\circ \pm 8.2^\circ$, latitude avg = $-32.0^\circ \pm 4.7^\circ$). We recover a southern Euramerican origin for Eotetrapodiformes, consistent with previous studies (longitude avg = $-12.5^\circ \pm 7.0^\circ$, latitude avg = $-19.4^\circ \pm 6.4^\circ$). A southern Euramerican origin was also found for Tristichopteridae (longitude avg = $-12.7^\circ \pm 6.9^\circ$, latitude avg = $-19.7^\circ \pm 6.3^\circ$) and Elpistostegalia (longitude avg = $-12.3^\circ \pm 5.5^\circ$, latitude avg = $-13.5^\circ \pm 5.3^\circ$). As expected in a phylogenetic comparative analysis, uncertainty in estimated node states increases toward the root. However, despite the level of uncertainty within a single run, only three nodes have mean

ancestral state values that are greater than an absolute value of 5° among the three replicate runs.

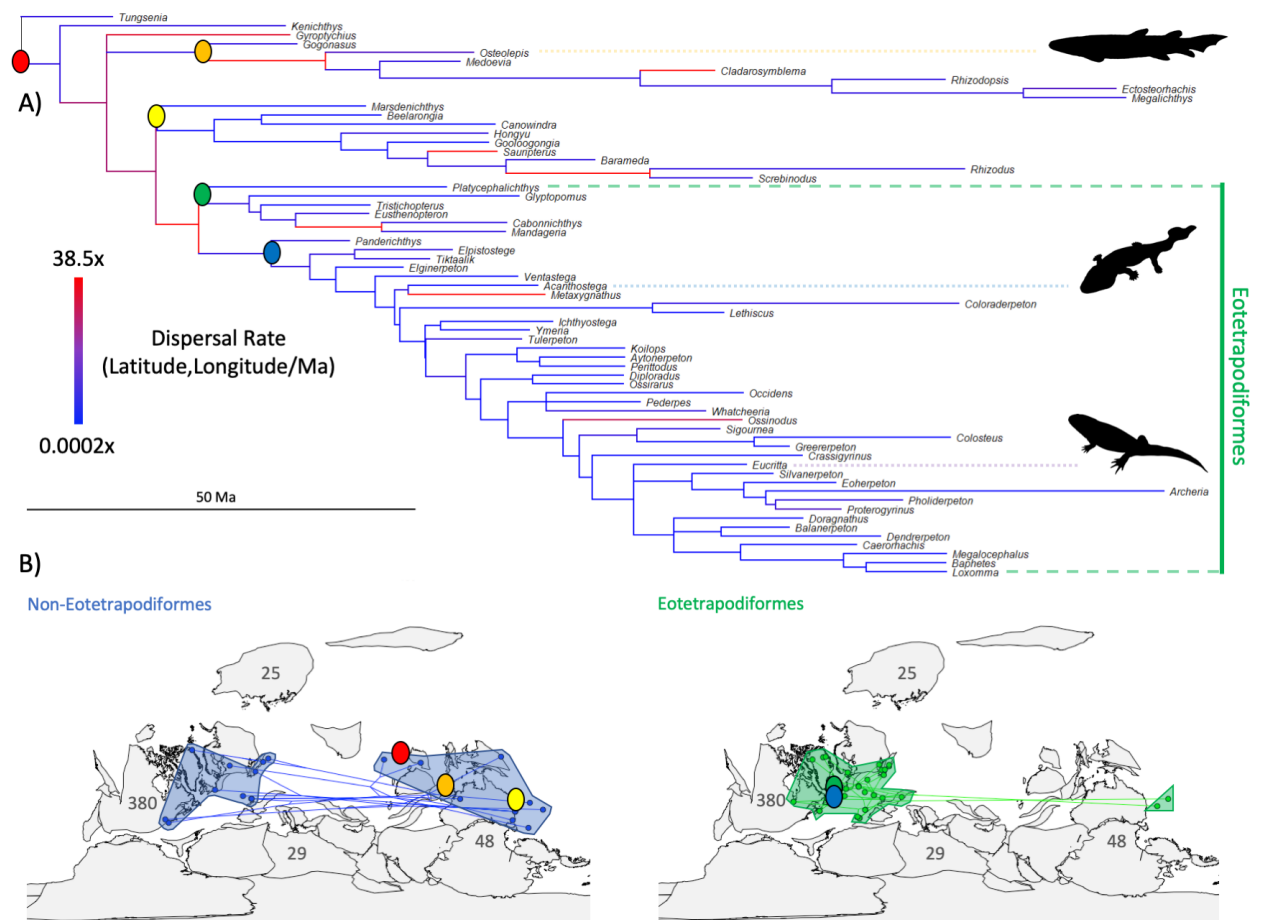


Fig. 3. A. Trimmed tetrapodomorph phylogeny with mapped rates of dispersal. Cooler (bluish) colors represent slower rates and warmer (reddish) colors represent faster rates. B. Non-eotetrapodiform (left in blue) and eotetrapodiform (right in green) trees and taxon paleolocations plotted on a map of the Middle Devonian. Transparent polygons illustrate broad geographic regions of sampled taxa in southern Euramerica, eastern Gondwana, and East Asia. Numbers show the total number of geological formations recorded from each major geographic region (eastern Gondwana and East Asia combined). Colored circles show average paleolocations of major clades estimated by the 'geo' model and indicated in the tree above. Red circle: Tetrapodomorpha, orange: "Megalichthyiformes", yellow: "Canowindridae" + Rhizodontidae, green: Tristichopteridae, and blue: Elpistostegalia. Phylogeny with mapped dispersal rates was produced in BayesTrees (<http://www.evolution.rdg.ac.uk/>)

BayesTrees.html). Middle Devonian tree and paleolocation plots were made using the 'phylo-to-map' function in the R package, phytools (Revell, 2012). Middle Devonian map was sourced from the R package, paleoMap (Rothkugel and Varela, 2015). Tetrapodomorph silhouettes were sourced from phylopic.org: *Eucritta* by Dmitry Bogdanov (vectorized by T. Michael Keeseey), *Osteolepis* by Nobu Tamura, and *Acanthostega* by Mateus Zica (CC BY-SA 3.0).

We find good evidence that geographic sampling bias influences dispersal rate estimates, regardless of the route used (95% CI: western Gondwana route $U = [800,928]$; northern Euramerica + East Asia route $U = [832,946]$; direct route $U = [729,889]$; no scenario includes the null $U = 499.5$; Fig. 4 and Supplementary Figs. 12 and 13). A U -statistic considerably higher than 499.5 suggests that branches with high dispersal rates have lower regional-level formation counts, on average, than branches with low rates. One can also interpret the null U -statistic of 499.5 as a 50% probability that a random branch with a low dispersal rate will rank higher in its regional-level formation count than a random branch with a high dispersal rate. With bootstrapping, we are 95% confident that the probability of a random branch with a low dispersal rate having a higher regional-level formation count than a random branch with a high rate is 72.97–88.99% for the more conservative 'direct route' scenario. Under the more liberal 'northern Euramerica + East Asia route' scenario, the probabilities are 83.28–94.69%. In sum, branches with high dispersal rates (two standard deviations greater than average) have a smaller number of recorded formations, on average, along their reconstructed dispersal path. Our results cannot be explained by a fossil record that is more complete through time (Pull of the Recent). A regression model relating regional-level formation count to the minimum age of each branch shows only a weak relationship (slope = -0.044 , $r^2 = 0.1$, $P < 0.001$). However, the total global (stage-level) formation count (which does not account for geographic variation) does show

potential bias from Pull of the Recent (slope = -0.3 , $r^2 = 0.71$, $P < 0.0001$). If dispersal rates are biased by the increase in number of formations globally, we would also expect to see elevated dispersal rates decrease toward the tips, but a regression model relating stretched branch lengths with time is not supported (slope = -0.025 , $r^2 = 0.006$, $P = 0.41$).

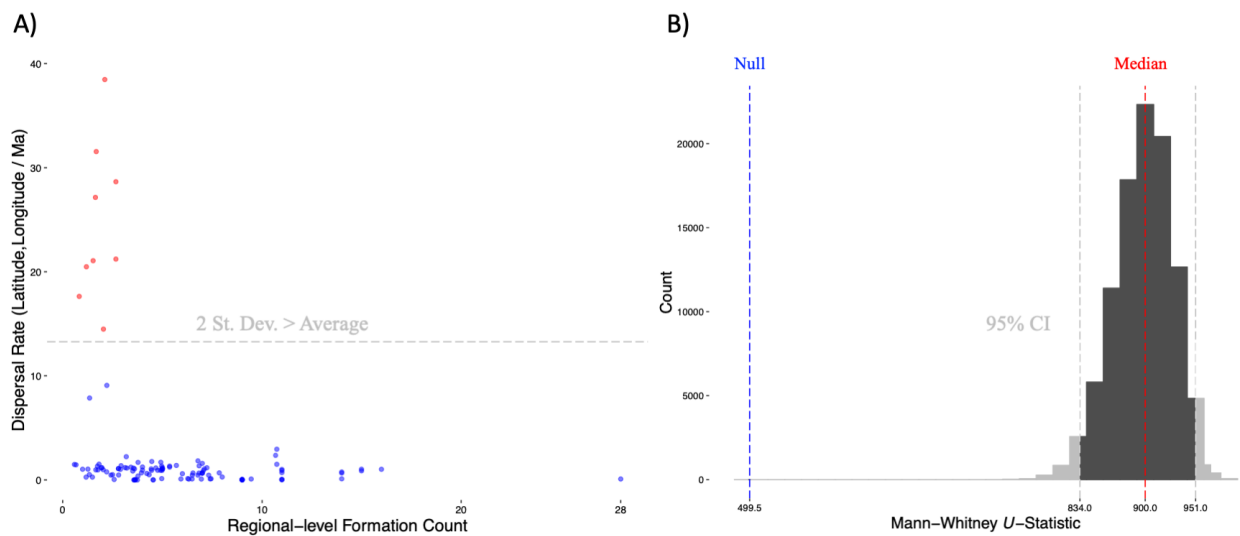


Fig. 4. A. Scatter-plot of the average dispersal rates over the regional-level formation counts for each branch of the phylogeny, using the northern Euramerica + East Asia route scenario. Points colored by the dispersal rate being above (red) or below (blue) two standard deviations greater than the average rate across the tree. B. Histogram of the bootstrapped U-statistics. Values outside of the 95% confidence interval are grayed out. The median and null expected U-statistics are indicated by the red and blue dotted lines, respectively. The null expected U-statistic is based on the null hypothesis that 50% of the branches with low dispersal rates will have greater regional-level formation counts than branches with higher rates. Rejecting the null hypothesis suggests that estimated dispersal rates are biased and correlate with regional-level formation count.

Our results cannot be explained by a fossil record that is more complete through time (Pull of the Recent). A regression model relating regional-level formation count to the maximum age of each branch shows only a weak relationship (slope = -0.045, $r^2 = 0.1$, $P < 0.001$). However, total global (stage-level) formation count (which does not account for geographic variation) does show potential bias from Pull of the Recent (slope = -0.31, $r^2 = 0.72$, $P < 0.0001$). We would also expect to see elevated dispersal rates decrease with time if this artifact is present in our data, but a regression model relating stretched branch lengths with time is not supported (slope = -0.025, $r^2 = 0.006$, $P = 0.41$).

4. Discussion

We expected to infer high dispersal rates for closely related taxa that are distributed across the globe. Our results, unadjusted for geographic bias in the fossil record, confirm this notion. However, we also find a compelling statistical association between high dispersal rates and a low number of formations along dispersal paths—a patchy fossil record is driving inferences of high dispersal rates. Although we did not test for a correlation between dispersal rate and previously used proxies, such as valid taxon count and stage-level formation count, these proxies do not offer clear predictions for explaining dispersal rate variation. High dispersal rate variation is inferred when closely related taxa are geographically separate. For example, valid taxon count cannot explain geographic rate variation because spatial information is lacking in this bias proxy and because sister taxa are likely to have similar counts (these data are phylogenetically structured). Stage level formation counts will also not explain dispersal rate variation, particularly if high rate variation exists within the same geological stage. Assuming geological formations are evenly exposed and sampled

worldwide, low stage level formation counts should yield geographically variable fossil species and, therefore, drive high dispersal rate variation. However, formations are not evenly exposed or recorded in geological/paleontological databases, including the PBDB. Our formation count table demonstrates this bias (Table 1). Without geographic context, stage-level formation count cannot distinguish between global and local regions. For example, the geological stages that have the highest recorded number of formations are restricted to southern Euramerica where the majority of eotetrapodiform taxa have been discovered. The association between high formation counts in specific regions and high paleobiodiversity in those regions is likely not a coincidence and has a clear impact on how we interpret dispersal history. The earliest tetrapodomorphs are known from China and Australia at geological stages where relatively few formations are recorded outside of East Asia and eastern Gondwana. The basal-most ancestral state estimates reconstruct paleolocations in East Asia (unsurprisingly). This inference (hypothesis) is predicated on the lack of geological formations recorded outside of East Asia during this time period. In addition, the majority of more crownward taxa and their reconstructed ancestral states are located in North America and Europe at geological stages in which relatively fewer formations are known elsewhere. This bias may heavily influence any conclusions made on the location and habitat of the tetrapodomorph water–land transition. Recently discovered taxa could help mitigate this problem by increasing the power of taxon sampling (Heath et al., 2014), such as *Tutusius* and *Umzantsia* from South Africa (Gess and Ahlberg, 2018). However, the current lack of cladistic coding for these taxa excludes them from phylogeny-based analyses. The taxonomic resolution of globally-occurring species, like *Eusthenodon* and *Spodichthys*, also impacts current models of species dispersal history because of their relatively uniform distribution (Long et al., 2018). *Eusthenodon*

and *Spodichthys* represent possible cases where taxonomic resolution is too coarse for phylogeographic analyses. Including these species inhibited our MCMC algorithms from reaching convergence. Widely distributed cosmopolitan species that lack intermediate geographic occurrences increase the uncertainty of parameter estimates within phylogeographic models, as is the case here for these two species.

Phylogenetic studies on macroevolution also often fail to incorporate data from the fossil record itself, such as trace fossil occurrences. Non-anatomical data often contribute to our understanding of taxonomic originations, including chiropteran (or digit-possessing) tetrapodomorphs for which trace fossil evidence exists about 10 million years before the first elpistostegalian body fossils (Niedzwiedzki et al., 2010). The inclusion of additional data from trace fossils could radically alter our current models of species dispersal history. Finally, it is important to note that the sampling bias proxies are also constrained by database curation biases. Phylogenetic studies on macroevolutionary trends now regularly leverage public databases, such as the PBDB, which allows larger and broader studies. It is unclear how patchy entries, on taxonomic occurrences and geological formations, for example, interact with other biases inherent in the fossil record. Caution is therefore warranted when these databases are mined, as is the case here.

5. Conclusions

Phylogenetic studies on macroevolution have not previously incorporated geographic context, which could influence a wide variety of analyses. We demonstrate here that phylogeographic methods are influenced by geographic sampling variability. We develop a simple sampling bias proxy that incorporates geographic information and show that it explains variation in estimated dispersal rates. The majority of elevated

dispersal rates are associated with large-scale movements between major landmasses that have very few, if any, relevant geological formations in between. Our analysis is also unlikely to be influenced by “Pull of the Recent”-like effects. Although not the first supertree for early tetrapodomorphs (Ruta et al., 2003), this study presents the first (to our knowledge) with branch lengths, making it useable for phylogenetic comparative analyses. The new supertree comprises many of the major clades previously inferred, but also recovers new ones that will be subject to scrutiny in future studies (discussed further in the Supplementary Material). This supertree should be useful to researchers who aim to use phylogenetic comparative methods to test hypotheses on the evolution of early tetrapodomorphs. In sum, our study estimates ancestral geographical reconstructions consistent with previously hypothesized dispersal patterns in early tetrapodomorphs. We also find that rates of dispersal are strongly influenced by geographic sampling bias. We suggest that researchers incorporate this proxy in phylogeny-based macroevolutionary studies that could be influenced by spatial distribution of the fossil record.

References

- Ahlberg, P.E., 2018. Follow the footprints and mind the gaps: A new look at the origin of tetrapods. *Earth Env. Sci. T. R. So.* 1–23.
<https://doi.org/10.1017/S1755691018000695>
- Alroy, J., Marshall, C.R., Bambach, R.K., Bezusko, K., Foote, M., Fürsich, F.T., Hansen, T.A., Holland, S.M., Ivany, L.C., Jablonski, D., Jacobs, D.K., Jones, D.C., Kosnik, M.A., Lidgard, S., Low, S., Miller, A.I., NovackGottshall, P.M., Olszewski, T.D., Patzkowsky, M.E., Raup, D.M., Roy, K., Sepkoski, J.J., Sommers, M.G., Wagner, P.J., Webber, A., 2001. Effects of sampling

- standardization on estimates of Phanerozoic marine diversification. *Proc. Natl. Acad. Sci. USA* 98, 6261–6266, <http://dx.doi.org/10.1073/pnas.111144698>.
- Benson, R.B.J., Butler, R.J., 2011. Uncovering the diversification history of marine tetrapods: ecology influences the effect of geological sampling biases. *Geol. Soc. Lond. Spec. Publ.* 358, 191–208, <http://dx.doi.org/10.1144/SP358.13>.
- Benson, R.B.J., Upchurch, P., 2013. Diversity trends in the establishment of terrestrial vertebrate ecosystems: Interactions between spatial and temporal sampling biases. *Geology* 41, 43–46, <http://dx.doi.org/10.1130/G33543.1>.
- Benson, R.B.J., Butler, R.J., Lindgren, J., Smith, A.S., 2010. Mesozoic marine tetrapod diversity: mass extinctions and temporal heterogeneity in geological megabiases affecting vertebrates. *Proc. R. Soc. B* 277, 829–834, <http://dx.doi.org/10.1098/rspb.2009.1845>.
- Benton, M.J., Donoghue, P.C.J., Asher, R.J., Friedman, M., Near, T.J., Vinther, J., 2015. Constraints on the timescale of animal evolutionary history. *Palaeontol. Electron.* 18, 1–106. <https://doi.org/10.26879/424>
- Benton, M.J., Ruta, M., Dunhill, A.M., Sakamoto, M., 2013. The first half of tetrapod evolution, sampling proxies, and fossil record quality. *Palaeogeogr. Palaeoclimatol. Palaeoecol., Vertebrate palaeobiodiversity patterns and the impact of sampling bias* 372, 18–41. <https://doi.org/10.1016/j.palaeo.2012.09.005>
- Budd, G.E., Mann, R.P., 2018. History is written by the victors: the effect of the push of the past on the fossil record. *Evolution* 72, 2276–2291, <http://dx.doi.org/10.1111/evo.13593>.
- Clack, J.A., Bennett, C.E., Carpenter, D.K., Davies, S.J., Fraser, N.C., Kearsey, T.I., Marshall, J.E.A., Millward, D., Otoo, B.K.A., Reeves, E.J., Ross, A.J., Ruta, M., Smithson, K.Z., Smithson, T.R., Walsh, S.A., 2017. Phylogenetic and

- environmental context of a Tournaisian tetrapod fauna. *Nat. Ecol. Evol.* 1, 0002.
<https://doi.org/10.1038/s41559-016-0002>
- Close, R.A., Benson, R.B.J., Alroy, J., Behrensmeyer, A.K., Benito, J., Carrano, M.T., Cleary, T.J., Dunne, E.M., Mannion, P.D., Uhen, M.D., Butler, R.J., 2019. Diversity dynamics of Phanerozoic terrestrial tetrapods at the local-community scale. *Nat. Ecol. and Evol.* 3, 590–597, <http://dx.doi.org/10.1038/s41559-019-0811-8>.
- Coates, M.I., Clack, J.A., 1991. Fish-like gills and breathing in the earliest known tetrapod. *Nature* 352, 234. <https://doi.org/10.1038/352234a0>
- Cope, E.D., 1868. Synopsis of the Extinct Batrachia of North America. *Proc. Acad. Nat. Sci. Philadelphia* 20, 208–221.
- Criscuolo, A., Gascuel, O., 2008. Fast NJ-like algorithms to deal with incomplete distance matrices. *BMC Bioinform.* 9, 166, <http://dx.doi.org/10.1186/1471-2105-9-166>.
- Criscuolo, A., Berry, V., Douzery, E.J.P., Gascuel, O., 2006. SDM: A fast distance-based approach for (super)tree building in phylogenomics. *Syst. Biol.* 55, 740–755. <https://doi.org/10.1080/10635150600969872>
- Didier, G., Fau, M., Laurin, M., 2017. Likelihood of tree topologies with fossils and diversification rate estimation. *Syst. Biol.* 66, 964–987, <http://dx.doi.org/10.1093/sysbio/syx045>.
- Didier, G., Laurin, M., 2018. Exact distribution of divergence times from fossil ages and topologies. *bioRxiv*, 490003, <http://dx.doi.org/10.1101/490003>.
- Didier, G., Royer-Carenzi, M., Laurin, M., 2012. The reconstructed evolutionary process with the fossil record. *J. Theor. Biol.* 315, 26–37, <http://dx.doi.org/10.1016/j.jtbi.2012.08.046>.

- Dunne, E.M., Close, R.A., Button, D.J., Brocklehurst, N., Cashmore, D.D., Lloyd, G.T., Butler, R.J., 2018. Diversity change during the rise of tetrapods and the impact of the Carboniferous rainforest collapse. *Proc. R. Soc. B* 285, <http://dx.doi.org/10.1098/rspb.2017.2730>, 20172730.
- Dunhill, A.M., Benton, M.J., Newell, A.J., Twitchett, R.J., 2013. Completeness of the fossil record and the validity of sampling proxies: a case study from the Triassic of England and Wales. *J. Geol. Soc.* 170, 291–300, <http://dx.doi.org/10.1144/jgs2012-025>.
- Dunhill, A.M., Benton, M.J., Twitchett, R.J., Newell, A.J., 2014a. Testing the fossil record: Sampling proxies and scaling in the British Triassic–Jurassic. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* 404, 1–11, <http://dx.doi.org/10.1016/j.palaeo.2014.03.026>.
- Dunhill, A.M., Hannisdal, B., Benton, M.J., 2014b. Disentangling rock record bias and common-cause from redundancy in the British fossil record. *Nat. Commun.* 5, 4818, <http://dx.doi.org/10.1038/ncomms5818>.
- Dowding, E.M., Ebach, M.C., 2019. Evaluating Devonian bioregionalization: quantifying biogeographic areas. *Paleobiology*, <http://dx.doi.org/10.1017/pab.2019.30>.
- Foote, M., 2003. Origination and extinction through the Phanerozoic: a new approach. *J. Geol.* 111, 125–148, <http://dx.doi.org/10.1086/345841>.
- Friedman, M., Coates, M.I., Anderson, P., 2007. First discovery of a primitive coelacanth fin fills a major gap in the evolution of lobed fins and limbs. *Evol. Dev.* 9, 329–337. <https://doi.org/10.1111/j.1525-142X.2007.00169.x>

- Gascuel, O., 1997. Concerning the NJ algorithm and its unweighted version, UNJ, in: Roberts, F., Rzhetsky, A. (Eds.), *Mathematical hierarchies and biology*. American Mathematical Soc., Providence, RI, pp. 149–170.
- Gauthier, J., Cannatella, D., de Queiroz, K., Kluge, A.G., Rowe, T., 1989. Tetrapod phylogeny, in: Fernholm, B., Bremer, K., Jörnvall, H. (Eds.), *The hierarchy of life*. Elsevier Science Publishers B. V. (Biomedical Division), Amsterdam, Netherlands.
- Gavryushkina, A., Welch, D., Stadler, T., Drummond, A.J., 2014. Bayesian inference of sampled ancestor trees for epidemiology and fossil calibration. *PLoS Comput. Biol.* 10, e1003919. <https://doi.org/10.1371/journal.pcbi.1003919>
- Gess, R., Ahlberg, P.E., 2018. A tetrapod fauna from within the Devonian Antarctic Circle. *Science* 360, 1120–1124, <http://dx.doi.org/10.1126/science.aag1645>.
- Heath, T.A., Huelsenbeck, J.P., Stadler, T., 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2957–E2966. <https://doi.org/10.1073/pnas.1319091111>
- Jablonski, D., Roy, K., Valentine, J.W., Price, R.M., Anderson, P.S., 2003. The impact of the pull of the recent on the history of marine diversity. *Science* 300, 1133–1135, <http://dx.doi.org/10.1126/science.1083246>.
- Koch, C.F., 1978. Bias in the Published Fossil Record. *Paleobiology* 4, 367–372.
- Lakner, C., van der Mark, P., Huelsenbeck, J.P., Larget, B., Ronquist, F., 2008. Efficiency of Markov chain Monte Carlo tree proposals in Bayesian phylogenetics. *Syst. Biol.* 57, 86–103. <https://doi.org/10.1080/10635150801886156>
- Laurin, M., 1998. The importance of global parsimony and historical bias in understanding tetrapod evolution. Part II. Vertebral centrum, costal ventilation,

and paedomorphosis. *Ann. Sci. Nat. Zoo.* 19, 99–114.

[https://doi.org/10.1016/S0003-4339\(98\)80004-X](https://doi.org/10.1016/S0003-4339(98)80004-X)

Lloyd, G.T., 2012. A refined modelling approach to assess the influence of sampling on palaeobiodiversity curves: new support for declining Cretaceous dinosaur richness. *Biol. Lett.* 8, 123–126, <http://dx.doi.org/10.1098/rsbl.2011.0210>.

Long, J.A., Gordon, M.S., 2004. The greatest step in vertebrate history: A paleobiological review of the fish-tetrapod transition. *Physiol. Biochem. Zool.* 77, 700–719. <https://doi.org/10.1086/425183>

Marjanović, D., Laurin, M., 2019. Phylogeny of Paleozoic limbed vertebrates reassessed through revision and expansion of the largest published relevant data matrix. *PeerJ* 6, e5565. <https://doi.org/10.7717/peerj.5565>

Marshall, J.E.A., Reeves, E.J., Bennett, C.E., Davies, S.J., Kearsey, T.I., Millward, D., Smithson, T.R., Browne, M.A.E., 2019. Reinterpreting the age of the uppermost “Old Red Sandstone” and Early Carboniferous in Scotland. *Earth Environ. Sci. Trans. R. Soc.* 109, 265–278, <http://dx.doi.org/10.1017/S1755691018000968>.

Niedźwiedzki, G., Szrek, P., Narkiewicz, K., Narkiewicz, M., Ahlberg, P.E., 2010. Tetrapod trackways from the early Middle Devonian period of Poland. *Nature* 463, 43–48. <https://doi.org/10.1038/nature08623>

O'Donovan, C., Meade, A., Venditti, C., 2018. Dinosaurs reveal the geographical signature of an evolutionary radiation. *Nat. Ecol. Evol.* 2, 452–458. <https://doi.org/10.1038/s41559-017-0454-6>

Pardo, J.D., Szostakiwskyj, M., Ahlberg, P.E., Anderson, J.S., 2017. Hidden morphological diversity among early tetrapods. *Nature* 546, 642–645. <https://doi.org/10.1038/nature22966>

- R Core Team., 2018. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. <https://www.R-project.org/> (accessed 20 December 2018)
- Rambaut, A., Drummond, A.J., Xie, D., Baele, G., Suchard, M.A., 2018. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* 67, 901–904. <https://doi.org/10.1093/sysbio/syy032>
- Raup, D.M., Boyajian, G.E., 1988. Patterns of generic extinction in the fossil record. *Paleobiology* 14, 109–125.
- Revell, L.J., 2012. phytools: An R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* 3, 217–223. <https://doi.org/10.1111/j.2041-210X.2011.00169.x>
- Robinson, D.F., Foulds, L.R., 1981. Comparison of phylogenetic trees. *Math. Biosci.* 53, 131–147. [https://doi.org/10.1016/0025-5564\(81\)90043-2](https://doi.org/10.1016/0025-5564(81)90043-2)
- Ronquist, F., Klopstein, S., Vilhelmsen, L., Schulmeister, S., Murray, D.L., Rasnitsyn, A.P., 2012a. A total-evidence approach to dating with fossils, applied to the early radiation of the Hymenoptera. *Syst. Biol.* 61, 973–999. <https://doi.org/10.1093/sysbio/sys058>
- Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., Huelsenbeck, J.P., 2012b. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* 61, 539–542. <https://doi.org/10.1093/sysbio/sys029>
- Rothkugel, S., Varela, S., 2015. paleoMap: An R-package for getting and using paleontological maps. R package version 0.0.0.9001. <https://github.com/NonaR/paleoMap>.

- Ruta, M., Jeffery, J.E., Coates, M.I., 2003. A supertree of early tetrapods. *Proc. R. Soc. B* 270, 2507–2516, <http://dx.doi.org/10.1098/rspb.2003.2524>
- Sakamoto, M., Benton, M.J., Venditti, C., 2016. Dinosaurs in decline tens of millions of years before their final extinction. *Proc. Natl. Acad. Sci. U.S.A.* 113, 5036–5040. <https://doi.org/10.1073/pnas.1521478113>
- Sakamoto, M., Venditti, C., Benton, M.J., 2016b. ‘Residual diversity estimates’ do not correct for sampling bias in palaeodiversity data. *Methods Ecol. Evol.* 8, 453–459, <http://dx.doi.org/10.1111/2041-210X.2666>.
- Schliep, K.P., 2011. phangorn: Phylogenetic analysis in R. *Bioinformatics* 27, 592–593. <https://doi.org/10.1093/bioinformatics/btq706>
- Signor, P.W., Lipps, J.H., 1982. Sampling bias, gradual extinction patterns, and catastrophes in the fossil record. *Geol. Soc. Am. Spec. Publ.* 190, 291–296.
- Stadler, T., 2010. Sampling-through-time in birth–death trees. *J. Theor. Biol.* 267, 396–404. <https://doi.org/10.1016/j.jtbi.2010.09.010>
- Swartz, B., 2012. A marine stem-tetrapod from the Devonian of Western North America. *PLoS ONE* 7, e33683. <https://doi.org/10.1371/journal.pone.0033683>
- Tennant, J.P., Mannion, P.D., Upchurch, P., 2016a. Environmental drivers of crocodyliform extinction across the Jurassic/Cretaceous transition. *Proc. R. Soc. B* 283, 20152840, <http://dx.doi.org/10.1098/rspb.2015.2840>.
- Tennant, J.P., Mannion, P.D., Upchurch, P., 2016b. Sea level regulated tetrapod diversity dynamics through the Jurassic/Cretaceous interval. *Nat. Commun.* 7, 12737, <http://dx.doi.org/10.1038/ncomms12737>.
- Venditti, C., Meade, A., Pagel, M., 2011. Multiple routes to mammalian diversity. *Nature* 479, 393–396. <https://doi.org/10.1038/nature10516>

- Xie, W., Lewis, P.O., Fan, Y., Kuo, L., Chen, M.-H., 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* 60, 150–160. <https://doi.org/10.1093/sysbio/syq085>
- Zhang, C., Stadler, T., Klopfstein, S., Heath, T.A., Ronquist, F., 2016. Total-evidence dating under the fossilized birth–death process. *Syst. Biol.* 65, 228–249. <https://doi.org/10.1093/sysbio/syv080>
- Zhu, M., Ahlberg, P.E., Zhao, W.-J., Jia, L.-T., 2017. A Devonian tetrapod-like fish reveals substantial parallelism in stem tetrapod evolution. *Nat. Ecol. Evol.* 1, 1470–1476. <https://doi.org/10.1038/s41559-017-0293-5>

Appendix 1

Supplementary Material

1. Supertree

1.1. Data

1.1.1. Morphological data matrices

To maximize stem-tetrapodomorph taxon sample size, we collected five published data matrices containing unordered, multistate morphological characters (Friedman et al., 2007; Swartz, 2012; Clack et al., 2017; Pardo et al., 2017; Zhu et al., 2017) (see Table 1). Since downstream analyses might be sensitive to unequal sample sizes between taxa pre- and post-water-land transition, we didn't include several crownward stem-tetrapods from the original matrices. All taxa more crownward than *Baphetes* and *Eucritta* in Pardo et al. (2017), except *Balanerpeton* and *Dendrerpeton*, were disregarded. *Asaphestera*, *Casineria*, *Discosauriscus*, *Edops*, *Eryops*, *Gephyrostegus*, *Hylopleuron*, *Microbrachis*, *Paleothyris*, *Seymouria*, and *Westlothiana* were removed from Clack et al.'s (2017) data matrix. Future studies should include more taxa on either end of the water-land transition. Lastly, we chose the supertree approach because it's infeasible to construct a morphological supermatrix. One needs to recode characters and assess redundancies.

Source	# of	# of	Outgroup
	taxa	characters	
Friedman et al. (2007)	13	216	<i>Glyptolepis</i>
Swartz et al. (2012)	41	204	<i>Glyptolepis</i>
Clack et al. (2017)	33	213	<i>Eusthenopteron</i>
Pardo et al. (2017)	18	370	<i>Eusthenopteron</i>
Zhu et al. (2017)	33	169	<i>Glyptolepis</i>

Table 1. Data matrix summary. Number of taxa refers to the number of stem-tetrapodomorphs included in our analyses. For three data matrices, we chose *Glyptolepis* as the outgroup in subsequent Bayesian phylogenetic inferences because it seemed to be the skeletally most complete dipnomorph.

We made several corrections to the data matrices. We changed *Ymeria*'s state in character 343 of Pardo et al.'s (2017) matrix from 3 to a question mark (?) because that character only has two states. In Clack et al.'s (2017) matrix, we changed the states of *Silvanerpeton* in character 18, *Proterogyrinus* in character 31, and *Pholiderpeton* in character 66 from 2, 3, and 2, respectively, to question marks (?). These character states exceed the maximum number of states. Additionally, we substituted parentheses in Swartz's (2012) and Pardo et al.'s (2017) matrices with curly brackets for consistency.

Here, we must explicitly acknowledge that we didn't guarantee that every specimen used to score character states is skeletally mature. We didn't order characters (see Rineau et al., 2015, 2018). Further, we didn't check correlations

between characters (see Guillerme and Brazeau [2018] for discussions regarding this issue). These caveats could have biased phylogenetic inference.

1.1.2. Tip dates

We collected tip dates (genus-level minimum ages) from the Paleobiology Database (PBDB; <https://paleobiodb.org/>). If minimum age data were unavailable from the PBDB, we checked the localities where researchers found the genus, chose the youngest one and collected the upper boundary of the locality's age based on <http://fossilworks.org> (see Table 2).

Taxon	Tip date (Ma)	Geochronologic unit	Reference
<i>Acanthostega</i>	358.90	Late Famennian	PBDB
<i>Archeria</i>	279.30	Middle Kungurian	PBDB
<i>Aytonerpeton</i>	346.70	Late Tournaisian	PBDB
<i>Balanerpeton</i>	326.40	Middle Serpukhovian	PBDB
<i>Baphetes</i>	307.00	Late Moscovian	PBDB
<i>Barameda</i>	346.70	Late Tournaisian	PBDB
<i>Beelarongia</i>	376.10	Frasnian	Long, (1987)
<i>Cabonnichthys</i>	360.70	Famennian	Ahlberg and Johanson, (1997)
<i>Caerorhachis</i>	318.10	Middle Bashkirian	PBDB
<i>Canowindra</i>	360.70	Late Devonian	Thomson, (1973)
<i>Cladarosymblema</i>	326.40	Viséan	Fox et al., (1995)
<i>Coloraderpeton</i>	307.00	Late Moscovian	PBDB
<i>Colosteus</i>	306.95	Early Kasimovian	PBDB

<i>Crassigyrinus</i>	323.20	Late Serpukhovian	PBDB
<i>Dendrerpeton</i>	314.60	Early Moscovian	PBDB
<i>Diploradus</i>	346.70	Late Tournaisian	PBDB
<i>Doragnathus</i>	323.20	Late Serpukhovian	PBDB
<i>Ectosteorhachis</i>	272.30	Early Roadian	PBDB
<i>Elginerpeton</i>	376.10	Middle Frasnian	PBDB
<i>Elpistostege</i>	372.20	Late Frasnian	PBDB
<i>Eoherpeton</i>	318.10	Middle Bashkirian	PBDB
<i>Eucritta</i>	330.90	Late Viséan	PBDB
<i>Eusthenodon</i>	360.70	Late Famennian	Clement, (2002)
<i>Eusthenopteron</i>	372.20	Late Frasnian	PBDB
<i>Glyptolepis</i>	382.40	Early Frasnian	PBDB
<i>Glyptopomus</i>	360.70	Late Famennian	Lebedev and Lukševičs, (2017)
<i>Gogonasus</i>	382.40	Early Frasnian	Long et al., (2006)
<i>Gooloogongia</i>	360.70	Famennian	Johanson and Ahlberg, (1998)
<i>Greererpeton</i>	323.20	Late Serpukhovian	PBDB
<i>Gyroptychius</i>	383.70	Middle Devonian	Newman et al., (2015)
<i>Hongyu</i>	360.70	Famennian	Zhu et al., (2017)
<i>Ichthyostega</i>	358.90	Late Famennian	PBDB
<i>Jarvikina</i>	379.50	Middle Frasnian	Lebedev et al., (2010)
<i>Kenichthys</i>	382.70	Late Givetian	PBDB
<i>Koharalepis</i>	382.40	Early Frasnian	Young et al., (1992)
<i>Koilops</i>	346.70	Late Tournaisian	PBDB
<i>Lethiscus</i>	336.00	Middle Viséan	PBDB
<i>Loxomma</i>	306.95	Early Kasimovian	PBDB

<i>Mandageria</i>	360.70	Famennian	Johanson and Ahlberg (1997)
<i>Marsdenichthys</i>	376.10	Frasnian	Holland et al., (2010)
<i>Medoevia</i>	360.70	Late Devonian	Lebedev, (1995)
<i>Megalichthys</i>	272.30	Early Roadian	PBDB
<i>Megalocephalus</i>	306.95	Early Kasimovian	PBDB
<i>Metaxygnathus</i>	358.90	Late Famennian	PBDB
<i>Occidens</i>	330.90	Late Viséan	PBDB
<i>Ossinodus</i>	330.90	Late Viséan	PBDB
<i>Ossirarus</i>	346.70	Late Tournaisian	PBDB
<i>Osteolepis</i>	358.90	Late Famennian	PBDB
<i>Panderichthys</i>	382.40	Early Frasnian	PBDB
<i>Pederpes</i>	345.30	Early Viséan	PBDB
<i>Perittodus</i>	346.70	Late Tournaisian	PBDB
<i>Pholiderpeton</i>	311.45	Middle Moscovian	PBDB
<i>Platycephalichthys</i>	360.70	Late Famennian	Lebedev et al., (2010)
<i>Proterogyrinus</i>	318.10	Middle Bashkirian	PBDB
<i>Rhizodopsis</i>	298.90	Late Gzhelian	PBDB
<i>Rhizodus</i>	298.90	Late Gzhelian	PBDB
<i>Sauripterus</i>	358.90	Late Famennian	PBDB
<i>Screbinodus</i>	326.40	Viséan	Andrews, (1985)
<i>Sigournea</i>	336.00	Middle Viséan	PBDB
<i>Silvanerpeton</i>	326.40	Middle Serpukhovian	PBDB
<i>Spodichthys</i>	376.10	Frasnian	Snitting, (2008)
<i>Strepsodus</i>	307.00	Late Moscovian	PBDB
<i>Tiktaalik</i>	372.20	Late Frasnian	PBDB

<i>Tinirau</i>	383.70	Late Givetian	Swartz, (2012)
<i>Tristichopterus</i>	379.50	Early Frasnian	Bishop, (2013)
<i>Tulerpeton</i>	360.70	Late Famennian	PBDB
<i>Tungsenia</i>	407.60	Late Pragian	PBDB
<i>Ventastega</i>	358.90	Late Famennian	PBDB
<i>Whatcheeria</i>	336.00	Middle Viséan	PBDB
<i>Ymeria</i>	358.90	Late Famennian	PBDB

Table 2. Tip dates.

1.1.3. Root calibrations

We collected clade minimum and soft maximum ages from the PBDB and Benton et al. (2015) to calibrate tree roots. For Friedman et al. (2007), Swartz (2012), and Zhu et al. (2017), the least inclusive clade with age estimates is Rhipidistia (minimum age = 408.0 Ma; soft maximum age = 427.9 Ma). For Clack et al. (2017) and Pardo et al. (2017), we used the maximum ages of *Eusthenopteron*, *Panderichthys*, and *Spodichthys* (one of the basalmost taxa in Eotetrapodiformes) as the minimum age of the least inclusive clade (383.7 Ma). And the mean age of the clade is represented by the minimum age of Tetrapodomorpha (407.6 Ma).

1.2. Analyses

1.2.1. Bayesian phylogenetic inference

For each matrix, we generated a posterior distribution of phylogenetic trees using MrBayes 3.2.6 (Ronquist et al., 2012b). We wanted to have trees with branch length information and to standardize the inference process as much as possible.

Here, we included details absent from the manuscript. Aside from using outgroups designated in Table 1, we also constrained the ingroup. For Clack et al.'s (2017) matrix, we allowed the gamma shape parameter, state frequency, and rate to vary across partitions. Next, we conditioned on coding only variable characters (Lewis Mk model [Lewis, 2001] corrects for ascertainment bias). Therefore, 150, 8, 6, 103, and 6 constant characters in Friedman et al.'s (2007), Swartz's (2012), Clack et al.'s (2017), Pardo et al.'s (2017), and Zhu et al.'s (2017) matrices, respectively, were ignored. Moreover, we used gamma-shaped rate variation across sites (four categories; Yang, 1994). Harrison and Larsson (2015) found that the four rate category discrete approximation is sufficient to approximate a gamma rate distribution. We used an exponentially-distributed prior for the gamma shape parameter. An exponentially-distributed prior for the gamma shape parameter results in higher marginal likelihoods than a uniformly-distributed prior (Harrison and Larsson, 2015). To allow variable evolutionary rates over time, we used the Independent Gamma Rate (IGR) model (Lepage et al., 2007). As a prior for the morphological clock rate, we used a truncated normal distribution. Further, we used offset exponential priors for root and tree ages and fixed tip dates (see Ronquist et al., 2012a). Although we employed the fossilized birth-death model (FBD; Stadler, 2010; Didier et al., 2012, 2015; Heath et al., 2014; Gavryushkina et al., 2014; Zhang et al., 2016; Didier and Laurin 2018) as a branch length prior, we didn't allow for sampled ancestors.

In each inference, we ran two Markov chain Monte Carlo (MCMC) replicates for 20,000,000 generations, each with four chains, a sampling frequency of 1,000, and a diagnostics frequency of 5,000. MrBayes employs Metropolis-coupled version of the MCMC (Metropolis et al., 1953; Hastings, 1970; Geyer, 1991). We discarded the first 25% samples as burn-in. We also used BEAGLE 2.1 (Ayres et al., 2012) to decrease

computational time. Unless specified above, we used the default settings. Finally, we chose to output maximum clade credibility trees (Fig. 1-10).

We diagnosed MCMC convergence between runs using the average standard deviation (SD) of split frequencies (Lakner et al., 2008). The values in all five inferences were less than 0.005 (see Table 3). We also assessed convergence using minimum effective sample size (ESS) and potential scale reduction factor (PSRF by Gelman and Rubin, 1992) values. All these metrics showed that within each inference, runs converged.

Inference	Average SD of split frequencies
Friedman et al. (2007)	0.002578
Swartz et al. (2012)	0.004706
Clack et al. (2017)	0.003893
Pardo et al. (2017)	0.003184
Zhu et al. (2017)	0.004766

Table 3. The average standard deviation of split frequencies values between runs in all inferences were less than 0.005.

1.2.2. Distance supermatrix

First, we converted the maximum clade credibility trees (source trees) to Newick files using FigTree 1.4.3 (Rambaut, 2017). Then, we combined all the Newick

trees into a single PHYLIP file. We inputted this file to SDM 2.1 (Criscuolo et al., 2006) and computed a distance supermatrix. Trees were weighted using their sizes.

1.2.3. *Unweighted neighbor-joining*

We inferred the supertree from the distance supermatrix using a modified unweighted neighbor-joining (Gascuel, 1997) algorithm (UNJ*) implemented in PhyD* 1.1 (Criscuolo and Gascuel, 2008). We allowed polytomies and only positive branch lengths. Furthermore, we chose to output confidence values at branches (Guénoche and Garreta, 2000), which were suited for incomplete distance matrices. Most values are above 50. However, we didn't understand why our supertree contains branches with zero confidence values (Fig. 11), especially when nearby nodes had high posterior probabilities in the source trees. Unless specified above, we used the default settings.

1.2.4. *Rooting and plotting*

We read the supertree into R 3.5.2 (R Core Team, 2018) using APE 5.2 (Paradis and Schliep, 2019), rooted and saved it using phytools 0.6.60 (Revell, 2012), converted it to a Newick file using FigTree, and converted it again to a .trees file using BayesTreesConverter 1.3 (<http://www.evolution.rdg.ac.uk/BayesTrees.html>). BayesTraits 3.0.1 (Pagel 1999; <http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.1/BayesTraitsV3.0.1.html>) can't process a tree with a polytomous root node. So, we added an arbitrary branch length of 0.00001 to break the trichotomy.

Lastly, we plotted the complete supertree using strap 1.4 (Bell and Lloyd, 2015), Cairo 1.5.9 (Urbanek and Horner, 2015), and APE in R. Since *Archeria* has the longest

path length, we scaled the tree using *Archeria*'s tip date (279.3 Ma) despite *Megalichthys* and *Ectosteorhachis*' tip dates (272.3 Ma).

1.2.5. Tree comparisons

We compared the supertree with the published source tree and Marjanović and Laurin's (2019) Paleozoic limbed vertebrate tree regarding topology. Due to small stem-tetrapod sample size, we ignored Friedman et al.'s (2007) topology. Additionally, we prioritized published Bayesian over maximum parsimony trees whenever possible.

Lastly, we compared the supertree topology with the published source tree topologies using normalized Robinson-Foulds (nRF) distances (Robinson and Foulds, 1981) implemented in phangorn 2.4.0 (Schliep, 2011) in R. We first wrote Newick files for the source tree topologies. For Clack et al.'s (2017) Bayesian tree, we designated *Eusthenopteron* as the outgroup. Afterward, we pruned the supertree to match the tips in individual source trees using APE in R. To match Swartz's (2012) tree, we collapsed several clades (Rhizodontidae, Megalichthyidae, Whatcheeridae, Colosteidae, Baphetidae, total-group Lissamphibia, and Embolomeri). There appears to be no actual taxon in the clade "other stem-group amniotes" in Swartz's (2012) tree figure. In each comparison, polytomies in the supertree or the source tree were resolved in all possible ways using phytools. Then, we calculated all possible nRF distances and took an average (see Table 4).

Inference	Average nRF distance (%)
Friedman et al. (2007)	25.0
Swartz et al. (2012)	27.1
Clack et al. (2017)	67.7
Pardo et al. (2017)	33.3
Zhu et al. (2017)	45.6
Average	39.7

Table 4. The average of average normalized Robinson-Foulds (nRF) distances is 39.7%. Thus, there are, on average, 39.7% different or missing bipartitions in the source trees compared to the supertree.

2. Phylogeography

2.1. Data

2.1.1. Paleocoordinate locations

We obtained paleocoordinate data (paleolatitude and paleolongitude) for 65 early tetrapodomorphs from the PBDB using the GPlates software setting (<https://gws.gplates.org/>). For 16 taxa that did not have direct paleocoordinate data in the PBDB, we searched for the geologic formations and geographic regions, while encapsulating the time range, from which they were discovered and averaged all valid tetrapodomorph occurrences from those formations and regions. If not the paleolocation of the formation entry given by the PBDB, we used the closest geographic location from where a publication stated the formation is located. Although not the precise location, a more-general geographic location (e.g. township, county,

or country) should suffice for the global scale that we're conducting analyses. Below is a table of the locations we used for each of the 16 taxa.

Taxon	Paleolocation source	Reference	Notes
<i>Acanthostega</i>	PBDB	-	
<i>Archeria</i>	PBDB	-	
<i>Aytonerpeton</i>	PBDB	-	
<i>Balanerpeton</i>	PBDB	-	
<i>Baphetes</i>	PBDB	-	
<i>Barameda</i>	PBDB	-	
<i>Beelarongia</i>	PBDB - Avon River Group	-	
<i>Cabonnichthys</i>	PBDB - New South Wales	Long et al. (2018)	
<i>Caerorhachis</i>	PBDB	-	
<i>Canowindra</i>	PBDB - New South Wales	Long et al. (2018)	
<i>Cladarosymblema</i>	PBDB - Queensland	Long et al. (2018)	
<i>Coloraderpeton</i>	PBDB	-	
<i>Colosteus</i>	PBDB	-	
<i>Crassigyrinus</i>	PBDB	-	
<i>Dendrerpeton</i>	PBDB	-	
<i>Diploradus</i>	PBDB	-	
<i>Doragnathus</i>	PBDB	-	
<i>Ectosteorhachis</i>	PBDB	-	
<i>Elginerpeton</i>	PBDB	-	
<i>Elpistostege</i>	PBDB	-	
<i>Eoherpeton</i>	PBDB	-	

<i>Eucritta</i>	PBDB	-	
<i>Eusthenodon</i>	PBDB - Celsius Bjerg Group, Tula Region, Evieux Formation, New South Wales, Witpoort Formation	Clement et al. (2009), Long et al. (2018)	Not included: outlier rates
<i>Eusthenopteron</i>	PBDB	-	
<i>Glyptolepis</i>	-	-	Not included: outgroup
<i>Glyptopomus</i>	PBDB - Latvia	Lebedev and Lukševičs (2017)	
<i>Gogonasus</i>	PBDB	-	
<i>Gooloogongia</i>	PBDB - New South Wales	Long et al. (2018)	
<i>Greererpeton</i>	PBDB	-	
<i>Gyroptychius</i>	PBDB - Estonia, Scotland	Newman et al. (2015)	
<i>Hongyu</i>	PBDB - Zhongning	-	
<i>Ichthyostega</i>	PBDB	-	
<i>Jarvikina</i>	Russia	Young et al. (2013)	Not included: specific region in Russia unknown
<i>Kenichthys</i>	portal.gplates	-	Used present-day coordinates from PBDB
<i>Koharalepis</i>	PBDB - Mount Crean, Antarctica	Long et al. (2018)	Not included: no Antarctica entries in PBDB

<i>Koilops</i>	PBDB - Ballagan Formation	-	
<i>Lethiscus</i>	PBDB	-	
<i>Loxomma</i>	PBDB	-	
<i>Mandageria</i>	PBDB - New South Wales	Long et al. (2018)	
<i>Marsdenichthys</i>	PBDB - Mount Howitt, Victoria	Long et al. (2018)	
<i>Medoevia</i>	PBDB - Latvia	Lebedev (1995)	From Belarus, used Latvia occurrences
<i>Megalichthys</i>	PBDB	-	
<i>Megalocephalus</i>	PBDB	-	
<i>Metaxygnathus</i>	PBDB	-	
<i>Occidens</i>	PBDB	-	
<i>Ossinodus</i>	PBDB	-	
<i>Ossirarus</i>	PBDB - Ballagan Formation	-	
<i>Osteolepis</i>	PBDB	-	
<i>Panderichthys</i>	PBDB	-	
<i>Pederpes</i>	PBDB	-	
<i>Perittodus</i>	PBDB	-	
<i>Pholiderpeton</i>	PBDB	-	
<i>Platycephalichthys</i>	PBDB - Latvia	Boisvert et al. (2008)	
<i>Proterogyrinus</i>	PBDB	-	
<i>Rhizodopsis</i>	PBDB	-	
<i>Rhizodus</i>	PBDB	-	
<i>Sauripterus</i>	PBDB	-	
<i>Screbinodus</i>	PBDB - Scotland	Andrews (1985)	

<i>Sigournea</i>	PBDB	-	
<i>Silvanerpeton</i>	PBDB	-	
<i>Spodichthys</i>	East Greenland	Snitting (2008)	Not included: PBDB entries for Greenland are outside of age range
<i>Strepsodus</i>	PBDB (North American <i>Strepsodus</i> entries), PBDB - Queensland	Parker et al. (2005)	Not included: outlier rates
<i>Tiktaalik</i>	PBDB	-	
<i>Tinirau</i>	Eureka County, Nevada	Swartz (2012)	Not included: PBDB entries for Eureka County, NV, are outside of age range
<i>Tristichopterus</i>	PBDB - Scotland	-	
<i>Tulerpeton</i>	PBDB	-	
<i>Tungsenia</i>	portal.gplates	-	Used present-day coordinates from PBDB
<i>Ventastega</i>	PBDB	-	
<i>Whatcheeria</i>	PBDB	-	
<i>Ymeria</i>	PBDB	-	

Table 5. Taxon Paleolocations.

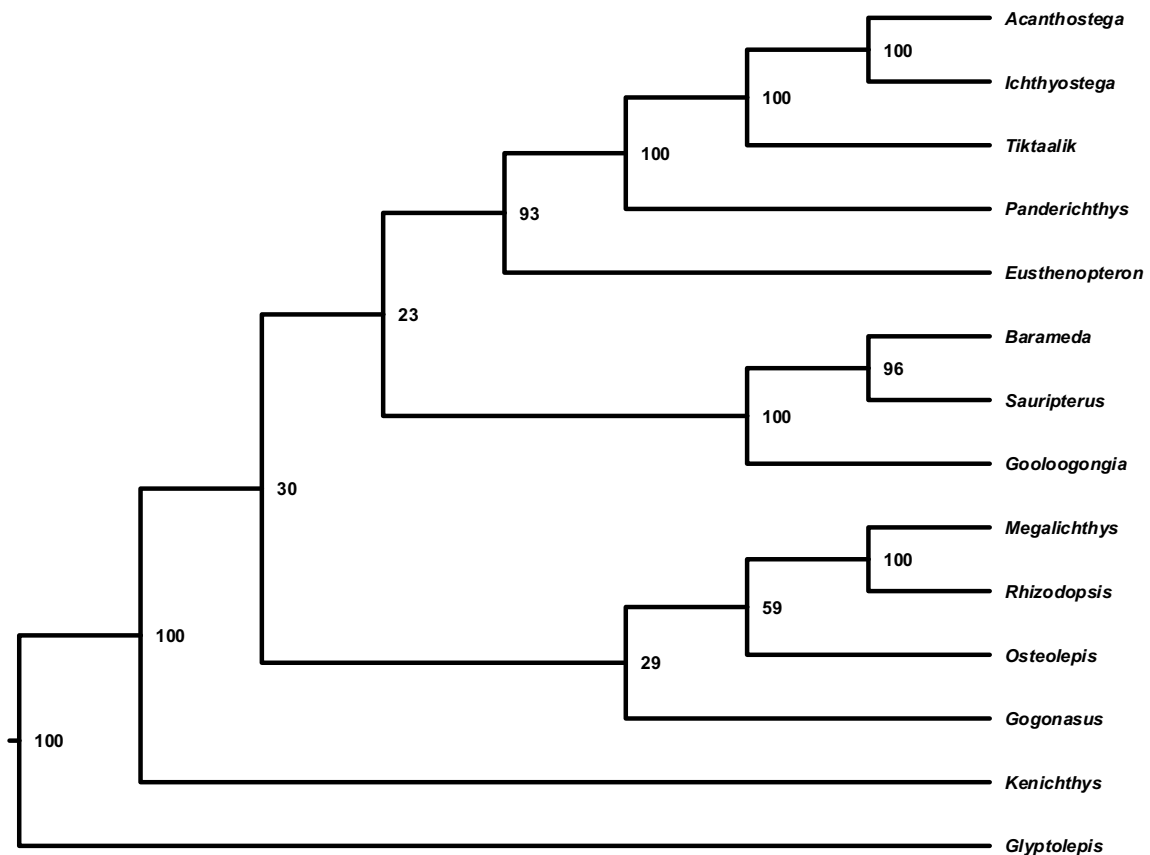


Fig. S1. The phylogenetic tree inferred using Friedman et al.'s (2007) matrix. Node values represent percent node posterior probabilities. Note that the hypothesized relationship between Megalichthyiformes, Rhizodontida, and Eotetrapodiformes have low support (30 and 23). We used FigTree to produce this figure.

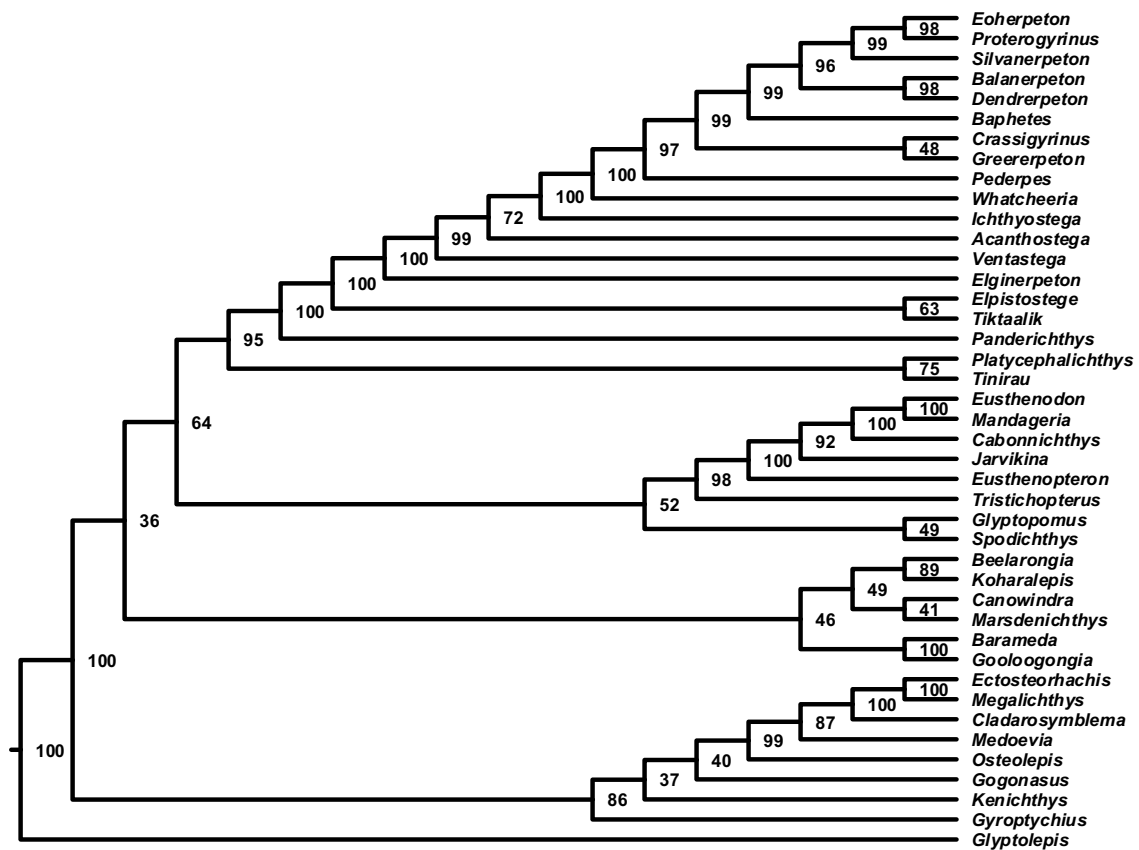


Fig. S2. The phylogenetic tree inferred using Swartz's (2012) matrix. Node values represent posterior probabilities (%). The sister taxon relationship between a group of canowindrids and rhizodontids and Eotetrapodiformes has low support (36).

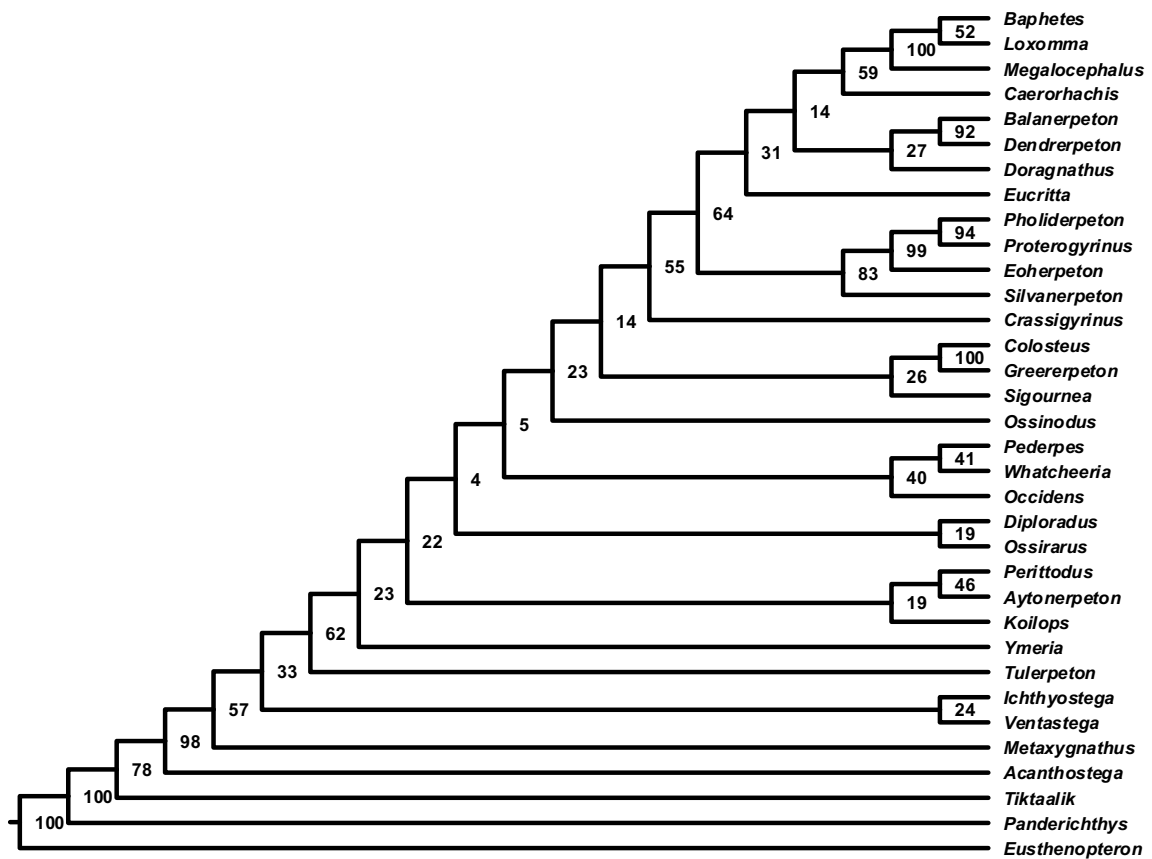


Fig. S3. The phylogenetic tree inferred using Clack et al.'s (2017) matrix. Node values represent posterior probabilities (%). Note the low support for multiple backbone nodes.

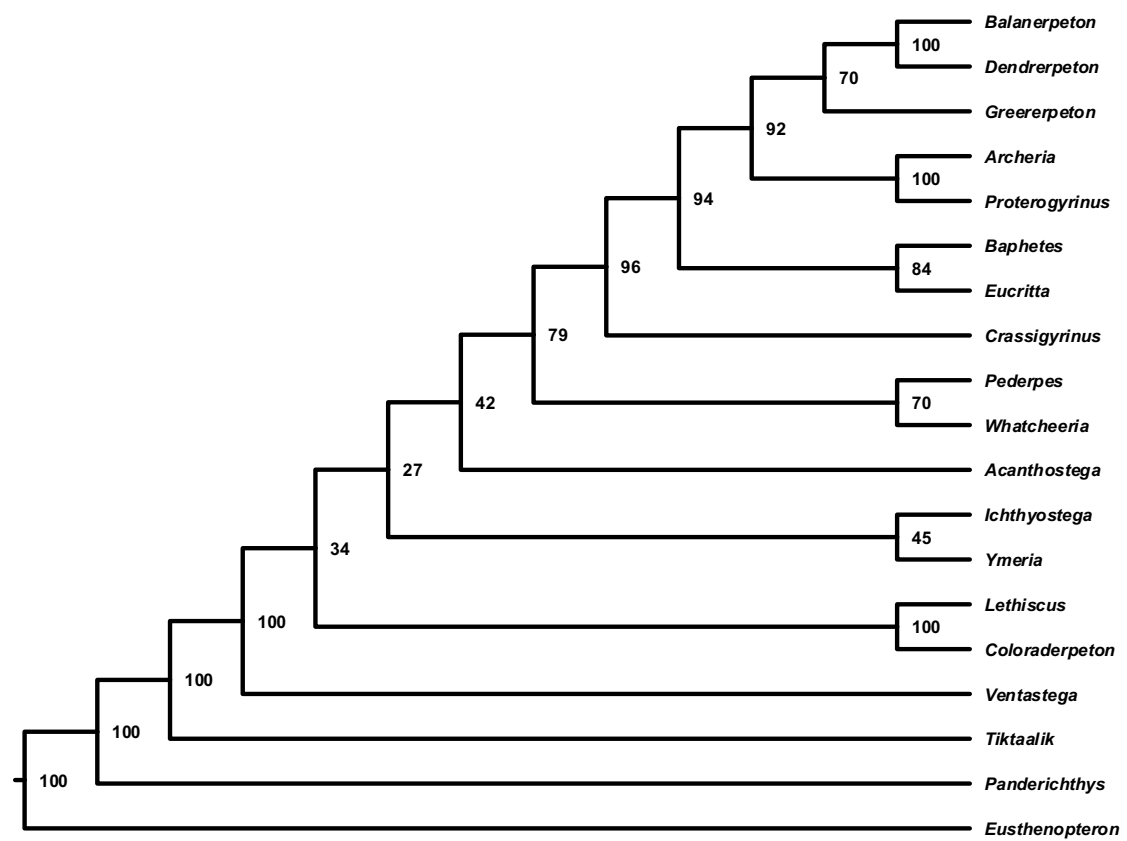


Fig. S4. The phylogenetic tree inferred using Pardo et al.'s (2017) matrix. Node values represent posterior probabilities (%). Note the low support for some backbone nodes (34, 27, and 42).

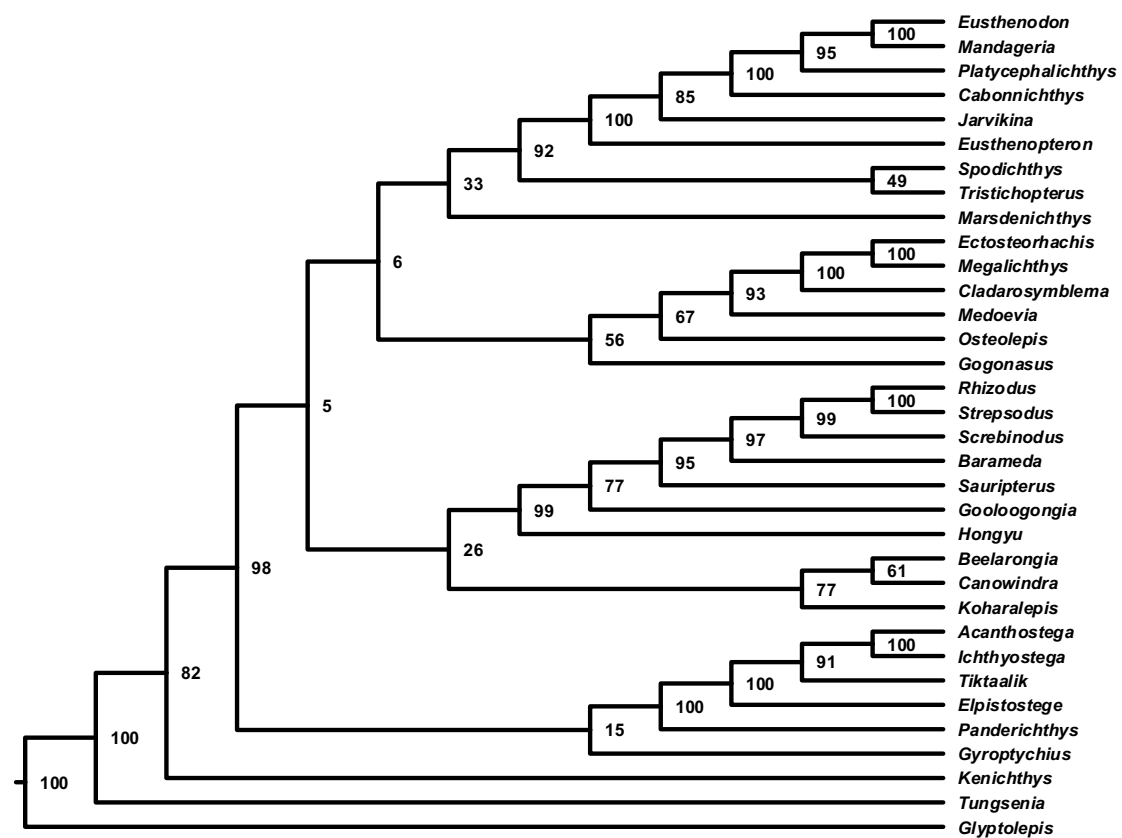


Fig. S5. The phylogenetic tree inferred using Zhu et al.'s (2017) matrix. Node values represent posterior probabilities (%). Note that the hypothesized relationship between Canowindridae, Rhizodontida, Megalichthyiformes, and Tristichopteridae have low support (5, 6, and 26). This unconventional topology shows an early divergence of Elpistostegalia from the rest of stem-tetrapods.

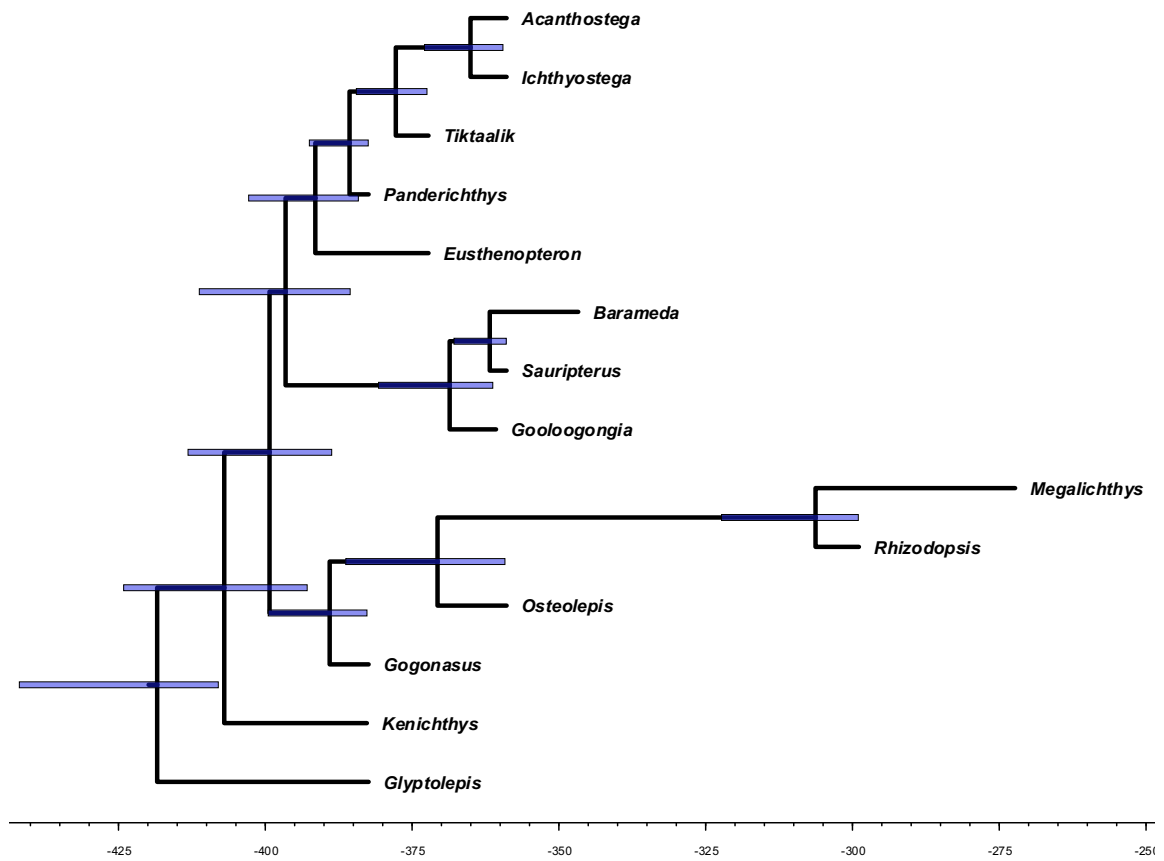


Fig. S6. The time-scaled phylogenetic tree inferred using Friedman et al.'s (2007) matrix. Node bars represent 95% highest posterior density (HPD) of node age estimates.

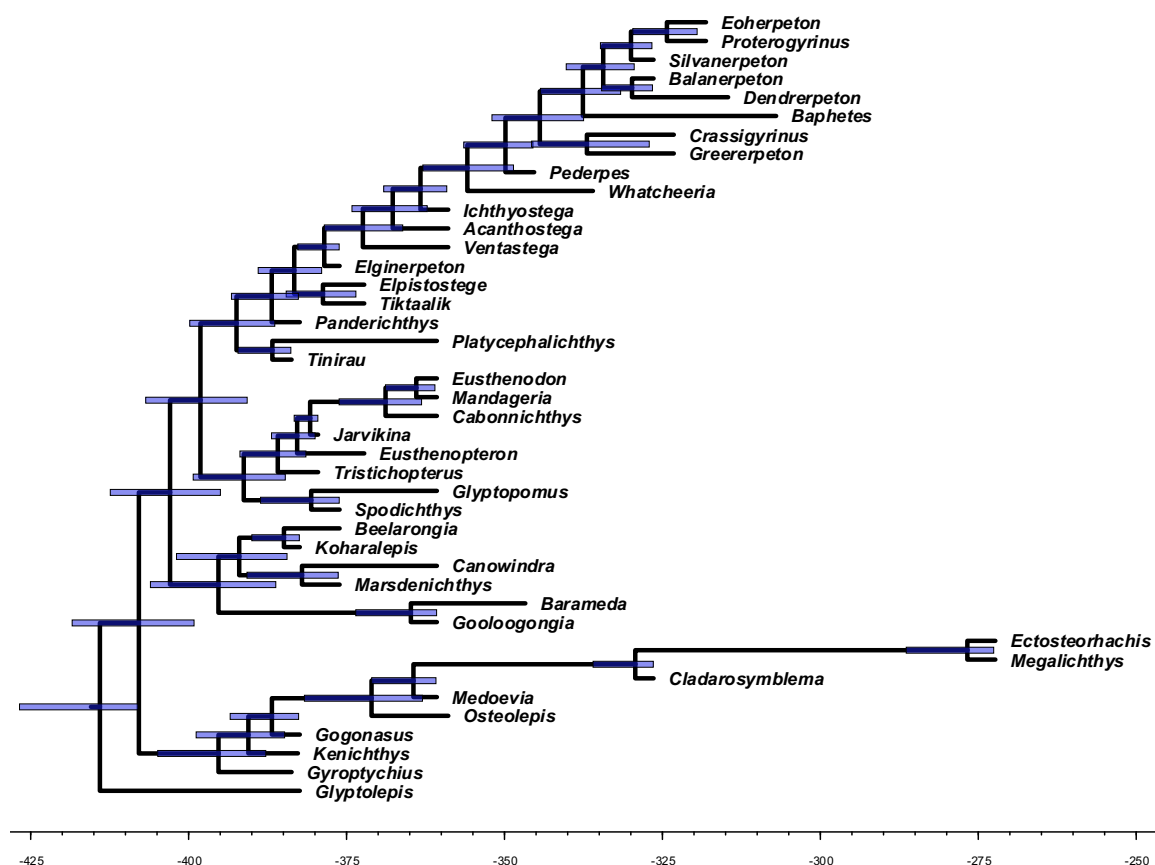


Fig. S7. The time-scaled phylogenetic tree inferred using Swartz's (2012) matrix. Node bars represent 95% highest posterior density (HPD) of node age estimates.

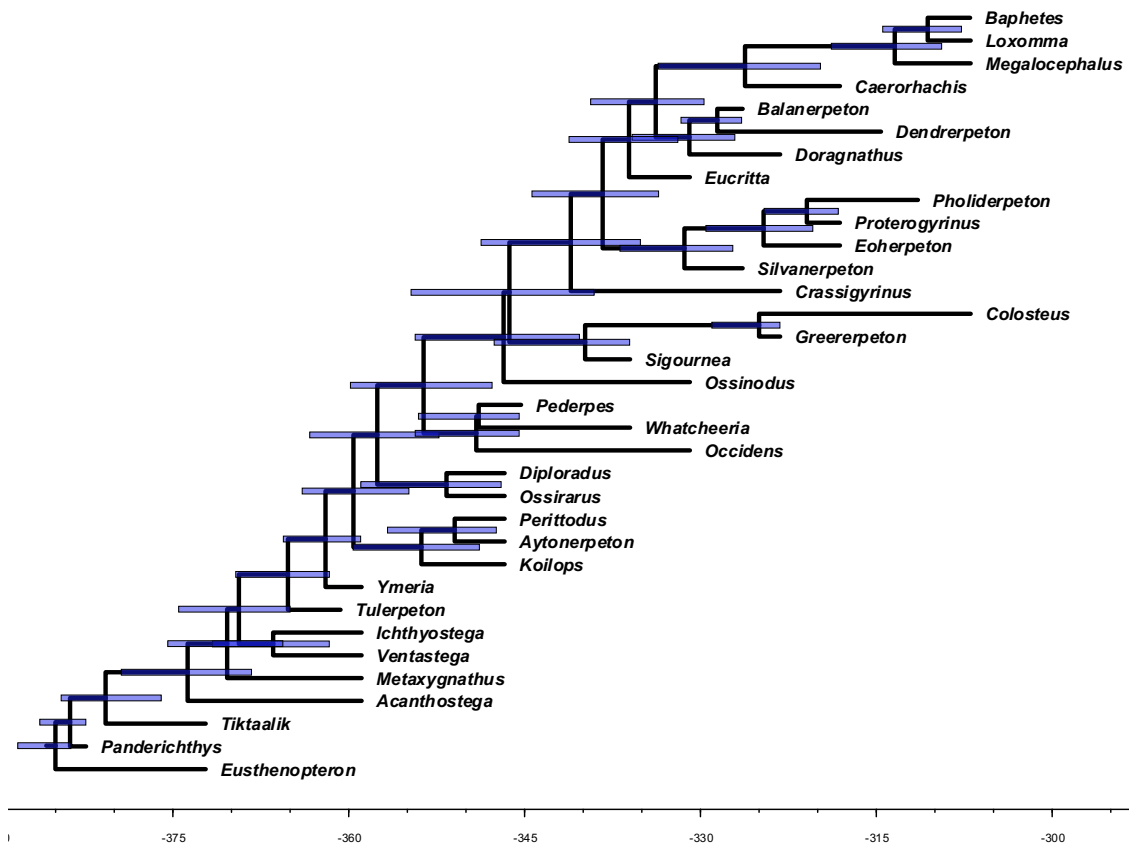


Fig. S8. The time-scaled phylogenetic tree inferred using Clack et al.'s (2017) matrix. Node bars represent 95% highest posterior density (HPD) of node age estimates.

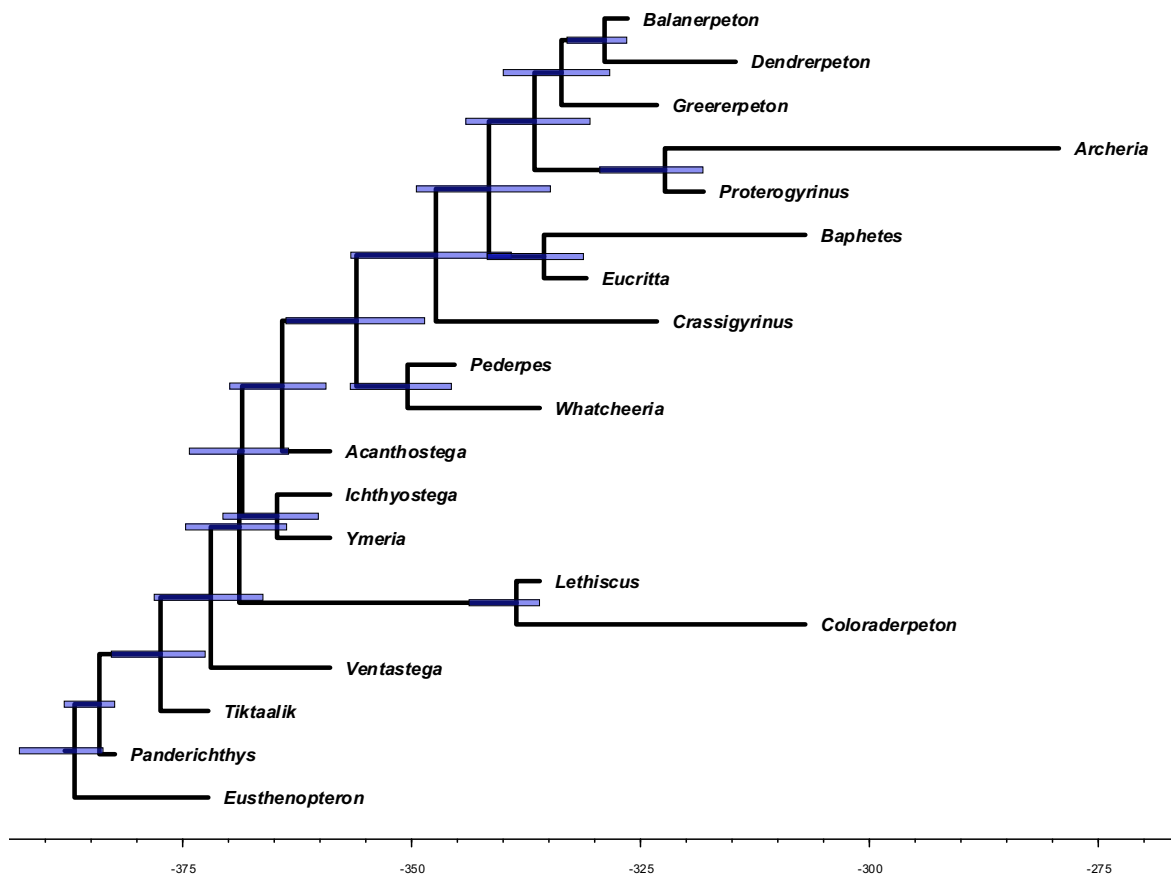


Fig. S9. The time-scaled phylogenetic tree inferred using Pardo et al.'s (2017) matrix. Node bars represent 95% highest posterior density (HPD) of node age estimates.

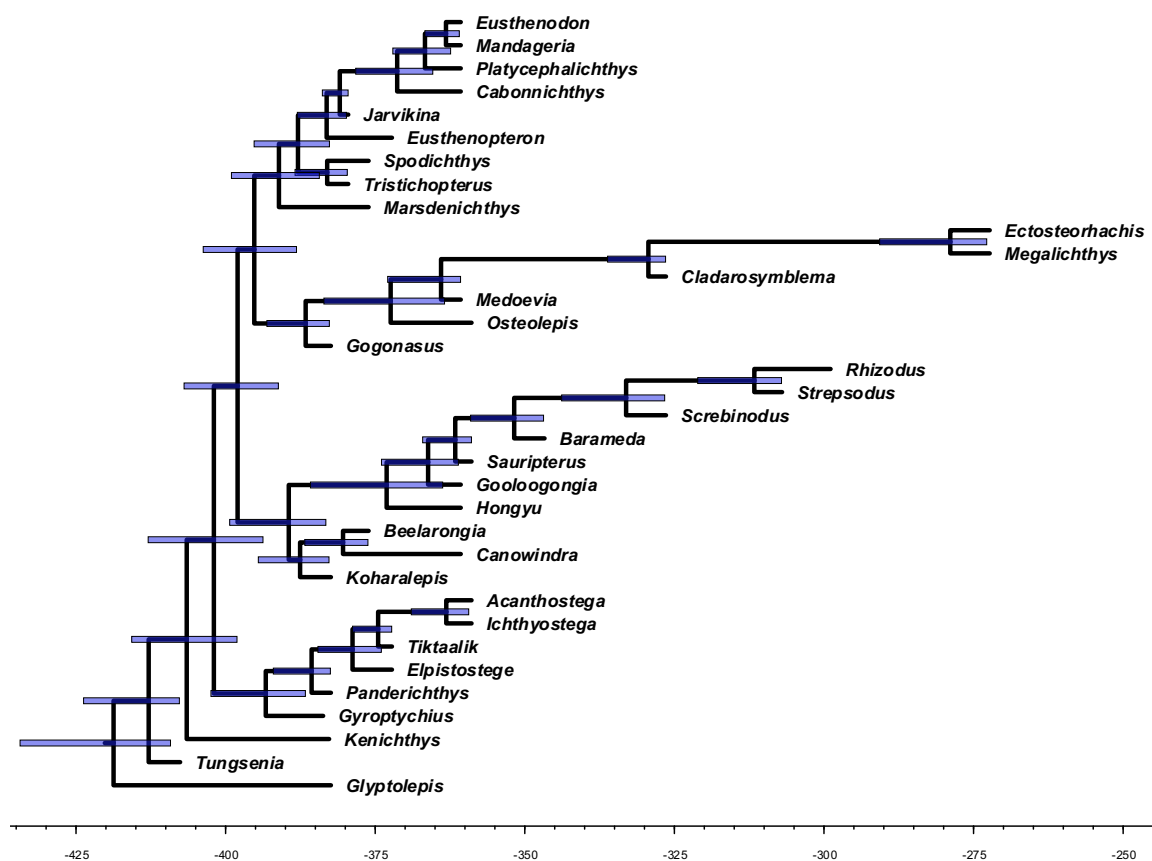


Fig. S10. The time-scaled phylogenetic tree inferred using Zhu et al.'s (2017) matrix. Node bars represent 95% highest posterior density (HPD) of node age estimates.

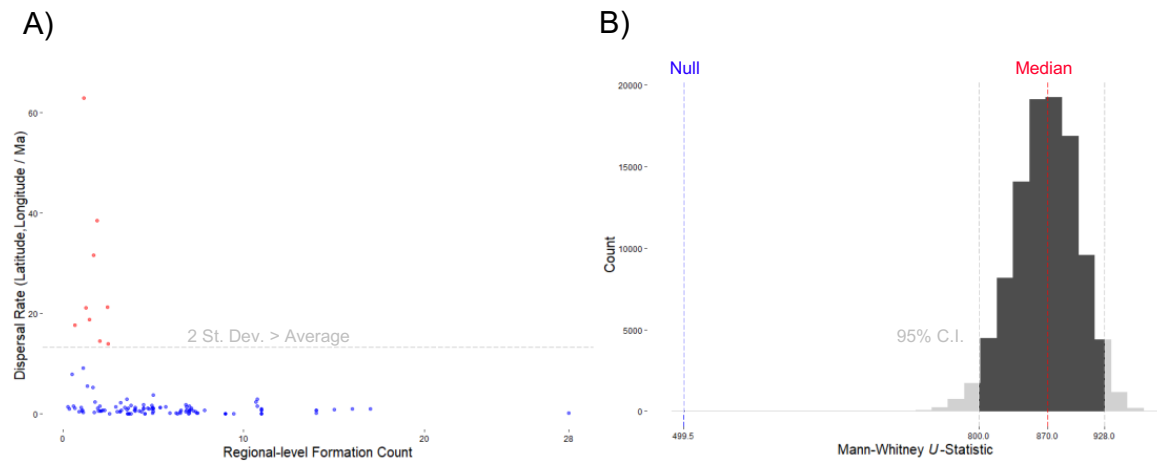


Fig. S12. A) Scatter-plot of the average dispersal rates over the regional-level formation counts for each branch of the phylogeny, using the Western Gondwana route scenario. Points colored by the dispersal rate being above or below two standard deviations greater than the average rate across the tree. **B)** Histogram of the bootstrapped *U*-statistics with values outside of the 95% confidence interval grayed out. The median and null expected *U*-statistics are indicated by the red and blue dotted lines, respectively. The null expected *U*-statistic is based on the null hypothesis that 50% of the regional-level formation counts with low dispersal rates will rank higher than formation counts with higher rates.

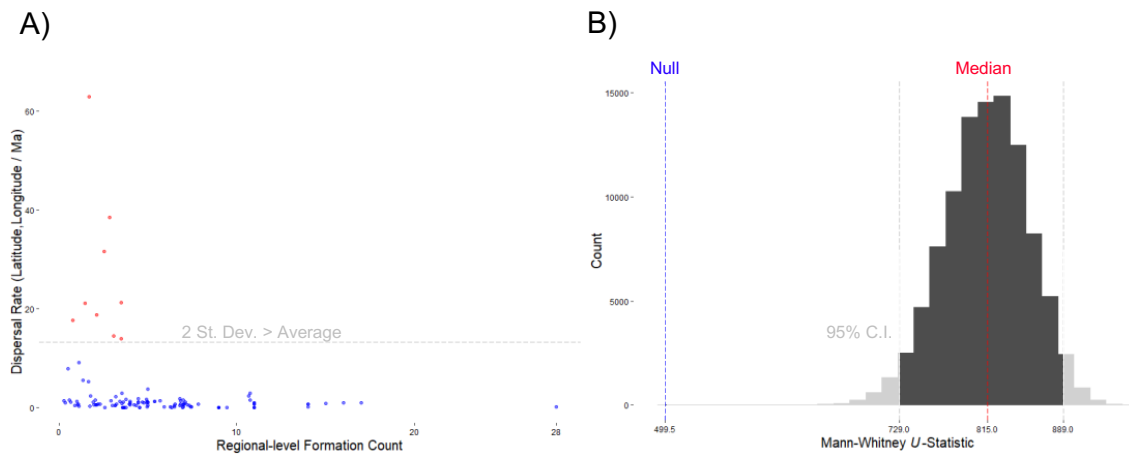


Fig. S13. A) Scatter-plot of the average dispersal rates over the regional-level formation counts for each branch of the phylogeny, using the direct route scenario. Points colored by the dispersal rate being above or below two standard deviations greater than the average rate across the tree. **B)** Histogram of the bootstrapped U -statistics with values outside of the 95% confidence interval grayed out. The median and null expected U -statistics are indicated by the red and blue dotted lines, respectively. The null expected U -statistic is based on the null hypothesis that 50% of the regional-level formation counts with low dispersal rates will rank higher than formation counts with higher rates.

Chapter 3

Latitude Does Not Shape Body Size Evolution in Mammals or Dinosaurs

Abstract

Global climate patterns help fundamentally shape the distribution of species and ecosystem structure. For example, warm-blooded animals like mammals inhabiting high latitudes are thought to be larger (and more extinction prone) than relatives at lower latitudes to better conserve body heat, a pattern known as Bergmann's rule. The modern world, however, lacks the comparative data to evaluate such ecological rules rigorously. Here, we create the first null model of Bergmann's rule using taxa that radiated under largely temperate Mesozoic climate conditions. We use newly described fossils from the Prince Creek Formation of northern Alaska along with data from Mesozoic mammaliaforms and dinosaurs to model the coevolution of body size and paleolatitude, accounting for fossil record biases and nuanced evolutionary rate variation. As predicted, we find no relationship between paleolatitude and body size in these two warm-blooded groups. When our model is applied to a large sample of extant mammals, we also find that body size evolution is independent of poleward dispersal (no evidence for Bergmann's rule). Our results suggest that latitude will not influence ecological interactions with body size, including extinction risk in mammals, as global warming advances. Our study provides a general approach for studying ecological rules and highlights the fossil record's power for studying longstanding and general principles in ecology.

Biologists have long sought general rules that describe broad ecological patterns and the processes that generate them¹. Global temperature variation, for example, plays a central role in shaping the genetic and biogeographical distribution of species^{2,3}. Bergmann's rule is the most well-known ecological rule associated with this variation. In its original framing, Bergmann's rule states that endothermic animals from cooler climates (higher latitudes) tend to be larger than close relatives from warmer, more equatorial climates^{4–6}. Initially proposed for mammals, the rule has also been applied to birds^{7,8} and (arguably inappropriately, given the hypothesised mechanism of endothermic heat retention) to ectotherms like amphibians⁹, reptiles¹⁰, fish¹¹, and invertebrates^{12–14}.

A strength of ecological rules is that their hypotheses yield clear predictions⁶ that can be tested with phylogenetically-informed statistical models¹⁵. Bergmann's rule predicts that endothermic lineages will tend to increase in body mass as they disperse to higher latitudes (colder climates). Research testing this prediction has, however, been hampered by three problems. First, it is common to find examples of taxa that fit Bergmann's rule by post hoc subsampling larger datasets (cherry-picking)^{5,7}. This is a serious problem because any sufficiently large dataset can be subdivided into groups, each of which may show a trend. Bergmann's rule is a "rule" precisely because it is hypothesised to apply across mammals, not to a few specific subgroups. If the "rule" only applies during post hoc analyses to particular groups, it is simply not an ecological rule (though associations in smaller groups may be interesting for other ecological or adaptive reasons). Second, ecological rules demand a model that allows the rate of evolution to vary across lineages rather than assuming a homogeneous process. This is especially important for rules like Bergmann's because it is exactly this variation regarding biogeographical dispersal and body size that are of interest¹⁶. Only recently

have variable-rate phylogenetic models become available, and they have yet to be applied to Bergmann's rule. Third, and perhaps most importantly, ecological rules often lack null models because they are hypothesised to operate broadly (e.g., across Mammalia) where controls are difficult and limited by current climate data. The fossil record provides repeated "natural experiments" in different climatic conditions across geological time that can be used to test general ecological rules. With rare exceptions^{17,18}, however, research on Bergmann's rule has focused on extant animal diversity and present-day climatic patterns. Mesozoic dinosaurs and mammals are ideal for establishing a null expectation of Bergmann's rule because they were both endothermic^{19,20} and inhabited more broadly temperate climates than today^{21,22}. Moreover, non-avian dinosaurs dispersed globally by the Late Triassic and persisted for over 180 million years²³, during which they evolved body sizes across orders of magnitude from several kilograms to over 50 tonnes. Living alongside non-avian dinosaurs, Mesozoic mammals represent a second, phylogenetically distinct clade of endotherms that underwent an independent geographic radiation. These two groups provide null models that predict a lack of association between latitude and body size against which extant mammals can be assessed.

Here, we develop null models for Bergmann's rule using data for 444 Mesozoic mammaliaforms and dinosaurs. We model the coevolution of body size and paleolatitude, accounting for fossil record biases²⁴ and nuanced evolutionary rate variation²⁵. We supplement this analysis with high-palaeolatitude dinosaur fossils from the Prince Creek Formation of Northern Alaska. After establishing the null expectation in Mesozoic dinosaurs and mammals, we apply our approach to 2,566 extant mammals where Bergmann's rule has important implications for how ecosystems are structured along latitudinal gradients²⁶. Large-bodied mammals, for example, tend to

have smaller population sizes and are more vulnerable to extinction^{25,26}. Factors driving both latitudinal dispersal and body size evolution are therefore potentially crucial for navigating the current climatically-induced biodiversity crisis²⁷ and understanding why some groups evolve large disparities in body size.

RESULTS

Establishing a null model in the Mesozoic Era

To establish a null model for Bergmann's rule, we regressed the femur lengths (a body size proxy²⁷, log₁₀ millimeters) of 382 Mesozoic dinosaurs and four dinosauromorphs onto paleolatitude (absolute values – distance from the equator). We then compared the likelihood fit of several models that distinguished the effect of paleolatitude between the northern and southern hemispheres, across the three geologic periods of the Mesozoic, and among the major dinosaur clades (Extended Data Table 1; Extended Data Figure 1). Our model selection procedure favoured a simple model without differences in effect between hemispheres, geologic periods, and clades (BF = 27.4 to 116.7). A variable-rates version of this model was favoured over the simple uniform-rate model (BF = 119.94). Under this model, we found no support for a relationship between body size and paleolatitude among dinosaurs and closely related taxa ($p_{\text{MCMC}} = 0.41$, Figure 1a, Extended Data Table 3). Even so, the estimated change in body size with paleolatitude was not biologically meaningful ($\beta = 0.0002$ (1.0 mm/degree), $R^2 = -0.031$). Bergmann's rule predicts that ancestral increases in body size are explained by positive shifts in absolute latitude along phylogenetic lineages. To visualise this, we plotted the branch-specific changes in body size over the inferred absolute latitudinal change along the same branches (Figure 1b).

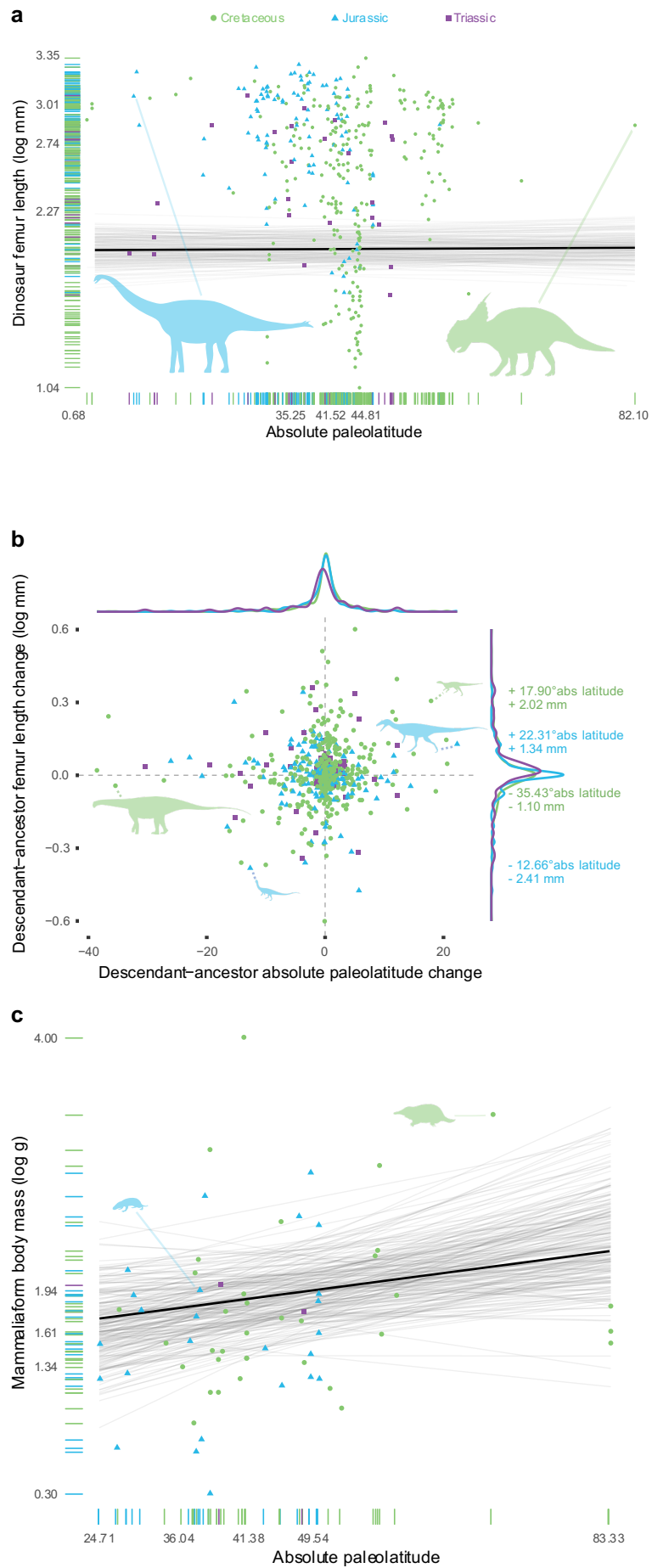


Figure 1. Mesozoic dinosaur and mammal body size does not covary with paleolatitude.

a, Femur length (\log_{10} mm) as a function of absolute paleolatitude in 382 extinct dinosaurs and dinosauriforms. Silhouettes in blue, *Spinophorosaurus nigerensis* by Remes K. et al. (CC BY-SA 3.0 license), and green, *Pachyrhinosaurus* by Andrew A. Farke (CC BY 3.0 license), were taken from phylopic.org. **b**, Estimated branch-specific changes in femur length (\log_{10} mm) as a function of estimated branch-specific changes in absolute paleolatitude. Overlapping density plots indicate the distribution of estimated branch-specific changes for each geological period, distinguished by colour. Values listed on the right-hand side indicate the estimated amount and direction of absolute paleolatitude and unlogged femur length changes along the terminal branch lengths to each species highlighted as silhouettes. Silhouettes in panel b (from left to right): *Nigersaurus* by Jagged Fang Designs (Public Domain), *Anchisaurus* by Tasman Dixon (Public Domain Dedication 1.0 license), *Cryolophosaurus ellioti* by Scott Hartman (CC BY-SA 3.0), and *Parksosaurus warreni* by Caleb M. Brown (CC BY-SA 3.0 license). **c**, Body mass (\log_{10} g) as a function of absolute paleolatitude in Mesozoic mammaliaforms. Silhouettes in blue, *Morganucodon watsoni* by Michael B. H. (CC BY-SA 3.0), and green, *Steropodon galmani* by Nobu Tamura (vectorised by T. Michael Keesey; CC BY 3.0), were obtained from phylopic.org. Outside of colour changes, no alterations were made to the silhouettes. Posterior (gray) and average (black) regression lines were derived from the simple phylogenetic generalised least squares regression models. Axes labels represent the minimum, 25% quantile, median, 75% quantile, and maximum values.

Throughout our initial model selection, we used the centroid position of each taxon but found consistent results when incorporating a distribution of body sizes and paleolatitudes to account for our uncertainty in body size estimates and geographical range ($\beta = 0.0001$ (1.0 mm/degree), $p_{\text{MCMC}} = 0.44$, $R^2 = 5.1\text{E-}5$). We further replicated

our results using a smaller dataset of imputed body masses ($n = 289$, $\beta = 0.003$ (1.01 mm/degree), $p_{\text{MCMC}} = 0.27$, $R^2 = 0.001$) and after accounting for the minimum age of taxa ($\beta_2 = 0.0003$ (1.0 mm/degree), $p_{\text{MCMC}} = 0.38$, total $R^2 = 0.02$).

We repeated these analyses using body mass data (\log_{10} grams) for 62 Mesozoic mammaliaforms. Again, we found no support for an association between body mass and absolute paleolatitude (Figure 1c; Extended Data Tables 2 and 3). Our model selection procedure preferred the simplest model with no differences in effect between hemispheres ($\text{BF} = 32.87\text{--}43.7$). Under the variable-rates model, which was favoured over the simplest uniform-rate model ($\text{BF} = 5.1$), the change in body mass with paleolatitude was negligible ($\beta = 0.009$ (1.02 g/degree), $p_{\text{MCMC}} = 0.12$, $R^2 = -0.0087$). We additionally show that branch-specific changes in body mass do not correlate with branch-specific changes in paleolatitude (Extended Data Figure 2). Together, these results satisfy the null expectation that Bergmann's rule did not drive evolution and interspecific variation in body size during more temperate global temperatures.

Sampling bias is a pervasive challenge for comparative analyses of fossil data²⁴. To test whether these biases influenced our regression results, we developed a geographic- and time-specific sampling metric and included it as a covariate in our regression analyses (Figure 2, see Methods). For dinosaurs, our model selection procedure best supported a model that excluded a geographic sampling bias effect ($\text{BF} = 24.4$). When we assessed the results of that model, we also found no evidence of sampling bias ($\beta_2 = -0.0003$ (1.0 mm/degree), $p_{\text{MCMC}} = 0.062$, $R^2 = 0.0049$). The number of vertebrate fossil-bearing formations in each geographic region and geologic period does not explain the variation observed in dinosaur body size. Our model selection procedure for Mesozoic mammaliaforms also supported a model excluding

a geographic sampling bias effect ($BF = 25.0$). We also found no evidence of sampling bias when our geographic sampling metric was included in the model ($\beta_2 = -0.0003$ (1.0 g/formation), $p_{MCMC} = 0.336$, $R^2 = 0.027$). Moreover, we would expect the preferential preservation of larger-bodied taxa in the fossil record²⁸ to bias our results in favour of larger body sizes at the high-latitudes, rather than against it. We thus find insufficient evidence that sampling biases explain body size variation across latitude (Figure 2).

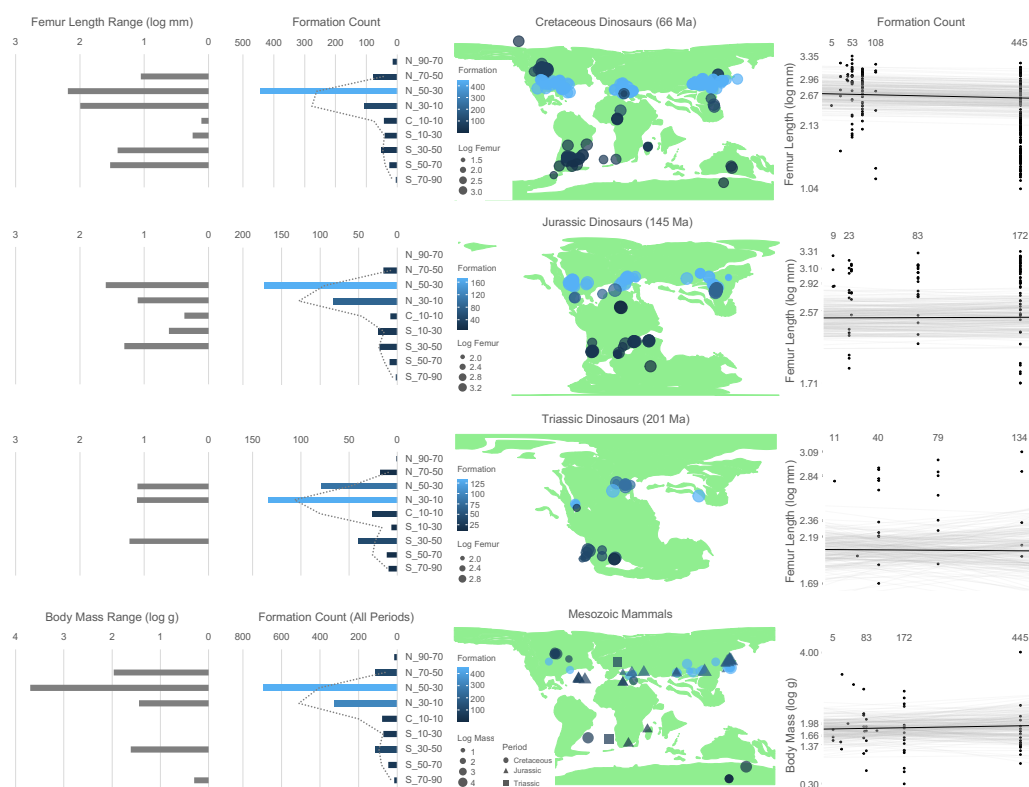


Figure 2. Gaps in the fossil record do not influence the latitudinal distribution of body size. Top three rows show the geographic distribution of Mesozoic dinosaurs and dinosauriforms in the Cretaceous, Jurassic, and Triassic. The bottom row shows the geographic distribution of Mesozoic mammaliaformes. Left two columns show the distribution of body size ranges (maximum - minimum) and formation counts for each of the nine latitudinal regions. Colours of formation count bars match colours in the map. Paleogeographic maps show the locations of fossil taxa and were obtained from the R package *chronosphere*⁵⁷. Scatter plots relate logged body sizes of taxa as a function of formation count. Posterior (grey)

and average (black) regression lines were derived from phylogenetic generalised least squares regressions. The y-axes represent the minimum, 25% quantile, median, 75% quantile, and maximum values.

Bergmann's rule does not operate across extant mammals

We next analysed 2,566 extant mammal species (Figure 3), the largest dataset analysed for Bergmann's rule. Like the results with the extinct species, a variable-rates model is strongly favoured over a uniform-rate model (BF = 639). We find modest support for a relationship between body mass (\log_{10} grams) and absolute mid-range latitude ($p_{\text{MCMC}} = 0.021$) but with coefficients that aren't biologically meaningful ($\beta = 0.00096$ (1.002 g/degree), $R^2 = 0.0019$; Figure 3a, Extended Data Table 3). To account for geographic range, we also created a model that randomly sampled the absolute minimum and maximum latitudes for each species with similar results. A model allowing the effect of absolute latitude on body mass to differ across the 14 most-represented mammalian orders and super-orders was supported over a single effect of absolute latitude for all mammals (BF = 154.34) and a model that assumed a single effect but allowed average body mass to vary (BF = 242.75). Under the separate effects model, we found that absolute mid-range latitude cannot explain body mass variation among species in at least nine of the 14 tested mammalian orders and super-orders ($p_{\text{MCMC}} > 0.05$). Of the other five mammalian orders, four showed an inconsequential increase in body size with mid-range absolute latitude ($\beta_{\text{Artiodactyla}} = 0.005$ (1.01 g/degree), $\beta_{\text{Chiroptera}} = 0.008$ (1.02 g/degree), $\beta_{\text{Didelphimorpha}} = 0.02$ (1.05 g/degree), and $\beta_{\text{Soricomorpha}} = 0.006$ (1.01 g/degree), $p_{\text{MCMC}} < 0.01$). Lagomorpha (pikas, rabbits, and hares) showed a slight decrease in body size with mid-range absolute latitude ($\beta_{\text{Lagomorpha}} = -0.009$ (-0.97 g/degree), $p_{\text{MCMC}} < 0.001$).

Ursidae (bears) provides a good example of what Bergmann's rule would look like if it were found to operate across Mammalia (Figure 3c). We confirmed an interspecific relationship between body mass and absolute mid-range latitude among ursids ($\beta = 0.0097$ (1.02 g/degree), $R^2 = 0.754$, $p\text{-value} = 0.0027$). The greatest positive co-directional changes occur in the common ancestor of *Ursus arctos* (brown bear) and *U. maritimus* (polar bear) and along the terminal branch to *U. maritimus*. We also see the greatest negative co-directional change along the branch to the Southeast Asian *Helarctos malyanus* (sun bear). These results demonstrate the efficacy of our approach and the ability to provide explanations of Bergmann's rule by post hoc selection of lineages.

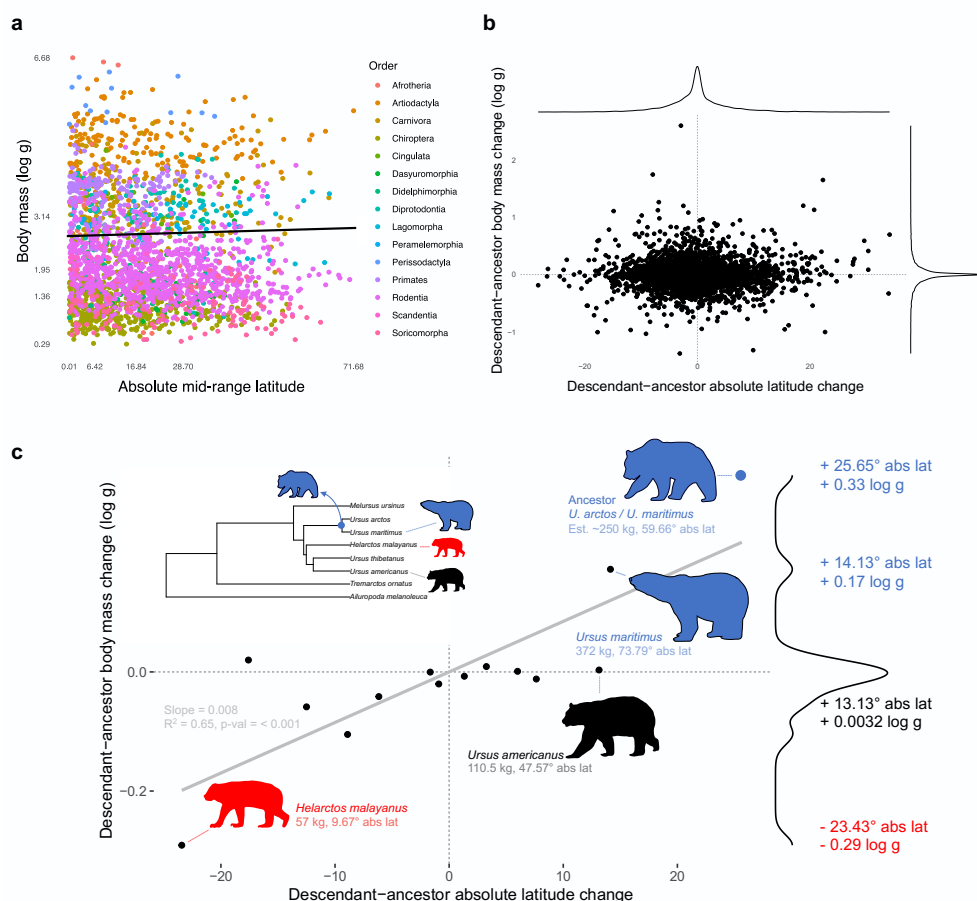


Figure 3. Bergmann's rule does not explain body size evolution among extant mammals. a, Body mass (\log_{10} g) as a function of absolute mid-range latitude among 2,566 extant mammals. The black best-fit line is an average of the posterior distribution of regression

model estimates. Axes labels represent the minimum, 25% quantile, median, 75% quantile, and maximum values. **b**, Estimated changes in body mass (\log_{10} g) as a function of changes in absolute mid-range latitude along branches of the mammal phylogeny. **c**, Estimated branch-specific changes in body mass (\log_{10} g) and absolute mid-range latitude among ursids.

DISCUSSION

The fossil record is a rich but underutilized data source for testing ecological principles²⁹ because it provides a wealth of unique climatic and biodiversity data unavailable in the modern world. In the Mesozoic, for example, high seasonality and arid conditions in the Late Triassic gave way ultimately to more stable, humid conditions in the Late Cretaceous²². During the Early to Late Triassic in the Southern Hemisphere and Late Triassic to Early Jurassic in the Northern Hemisphere, mid to high-latitude seasonal temperature contrasts (STC) reached 40°C. This contrast dropped to 20–30°C in the Late Cretaceous at mid to high latitudes in both hemispheres²². It was under this Mesozoic climate regime that both mammals and dinosaurs diversified and dispersed globally, providing a context to assess general ecological rules. Our results show that extinct groups can provide an important null model for evaluating such rules.

Although the Cretaceous was largely more temperate compared to the Present, the Late Cretaceous paleoArctic exhibited extended periods of winter darkness and freezing winter temperatures. While there are multiple well-known high-latitude dinosaur fossil assemblages, the Late Cretaceous Prince Creek Formation of Northern Alaska is exceptional because it is the highest-latitude dinosaur-bearing unit currently known (~82–85°N paleolatitude) and the only one exhibiting evidence for freezing temperatures and occasional snowfall. The Prince Creek Formation records strong seasonality with a cold month mean annual temperature of $-2.0 \pm 3.9^{\circ}\text{C}$ ³⁰. Despite this

challenging climatic regime, there is compelling evidence that dinosaurs endured these colder periods and were year-round residents of the Arctic^{30–34}. Overwintering in these conditions presumably would have required thermophysiological and life history adaptations specific to polar conditions. However, there is no evidence that the endemic, non-migratory adult dinosaurs found within the Prince Creek Formation ecosystem were larger, on average, than related species found in lower latitude formations³⁰.

An unnamed troodontid from the Prince Creek Formation, with teeth approximately twice the length and width of *Troodon* teeth from Montana and southern Alberta, is the sole example of a relatively large taxon³⁵. In contrast, all other dromaeosaurid teeth are comparable in size to those from more southerly formations. Other dinosaurs from the Prince Creek Formation, representing nine families, including tyrannosaurid and ceratopsid dinosaurs, are comparable in size to their relatives from more southern Late Cretaceous North American localities³⁰. For example, while originally described as a dwarf taxon, newly collected fossils of *Nanuqsaurus hoglundi*, the only known tyrannosaurid from the Prince Creek Formation, exhibit adult body sizes within the range of contemporaneous southern taxa like *Albertosaurus sarcophagus*³⁰. Additionally, *Pachyrhinosaurus perotorum* is known from multiple partial adult skulls^{36,37}, all smaller than adult skulls of *Pachyrhinosaurus canadensis* from southern Alberta. An unnamed leptoceratopsid has teeth comparable in size to *Leptoceratops gracilis* from the Scollard Formation of Alberta³⁰ and an unnamed thescelosaurid has teeth comparable in size to *Parksosaurus warreni* from the Horseshoe Canyon Formation of Alberta³⁰. Together, these fossils demonstrate that almost all dinosaurs inhabiting high-latitude climates had body sizes similar to relatives at lower latitudes.

For extant mammals, Bergmann's rule is theorised to operate at many taxonomic levels, from intraspecific relationships³⁸ to monophyletic groups⁵. It has been cited as an explanation for the evolution of large body size in mammals such as the short-faced bear *Arctodus simus*³⁹, the giant bison *Bison priscus*⁴⁰, and the giant ground sloth *Megalonyx jeffersonii*⁴¹. But most studies supporting Bergmann's rule generally evaluate the trend among assemblages of species^{7,42} or the strength and direction of intraspecific support across a few species^{43,44}. Among extant mammal species, as well as among species within 14 major mammalian orders and super-orders, we do not find a widespread positive association between body mass and absolute latitude. Our species-level analyses are consistent with those from Gohli and Voje⁴⁵, which found that neither latitude nor temperature were major drivers of body size evolution at the family level in a smaller dataset of extant mammals.

Bergmann's rule provides specific evolutionary predictions. As lineages of homeothermic species peripatrically diversify and disperse, the rule predicts selection of larger body size in lineages moving towards the poles by factors associated with latitude, such as mean annual temperature, seasonality, or precipitation⁶. We tested this evolutionary prediction of Bergmann's rule using an approach that assessed whether body size and latitude evolved at varying speeds along the branches of an evolutionary tree. This approach is, arguably⁶, a direct test for the evolutionary signature of Bergmann's rule, and we find that, across lineages of dinosaurs and mammals, ancestral increases in body size are unexplained by poleward shifts in latitude. These findings do not preclude climate in general from driving body size evolution. For example, Chiarenza et al.⁴⁶ found that the global distribution of sauropods was restricted to lower latitudes (warmer climates), suggesting they may have been poikilothermic. However, we found that latitude did not have a different

effect on sauropodomorph body size evolution compared with other groups of dinosaurs (BF=20.49, Extended Data Table 1). Our results highlight the limitations of attributing biogeographical patterns observed in select taxa to a general evolutionary process and the value of fossil data for evaluating long-held general principles in ecology.

METHODS

Data

Bergmann's rule, in its original text, operates among closely related taxa^{4–6} and should have a phylogenetic structure. We used phylogenetic regression models to test for Bergmann's rule in Mesozoic dinosaurs and mammaliaformes but focused primarily on the former given their larger sample size and range in body sizes. To test for a relationship between body size and latitude, we collected femur length (\log_{10} millimeters) and paleogeographic data for 378 dinosaur species and four other dinosauromorph archosaurs from Benson and colleagues²⁷ and O'Donovan and colleagues²³. Femur length was used as a proxy for body size via the conventions set in Benson and colleagues²⁷. We supplemented the femur length data with a smaller dataset of imputed body masses ($n = 289$) from Benson and colleagues²⁷. Paleolatitude was used as a proxy for average environmental temperature under the standard assumption that climate becomes cooler as absolute latitude increases. O'Donovan and colleagues²³ obtained the paleogeographic data from the Paleobiology Database (PBDB), which converts the present-day latitudes and longitudes of fossil sites into paleolatitude and paleolongitude values using GPlates software (<https://www.gplates.org/>). For any taxa with multiple paleolatitude data points, an average was taken.

The disproportionate sampling of fossils in different geographic regions has been shown to influence comparative analyses of diversification and geographic dispersal^{24,47–50}. It is conceivable that the known variation in body size is correlated with the number of fossil-bearing rock formations in a particular region and point in time. To test for such an effect on our regression results, we followed the approach of Gardner and colleagues²⁴ and collected the number of unique vertebrate fossil-bearing rock formations across multiple geographic zones. Rather than the broad geographic regions used by Gardner and colleagues²⁴, we collected formation counts across nine 20-degree latitudinal zones (Figure 2; Appendix 2, section 3). We further subdivided these geographic-specific formation counts into the three Mesozoic geologic periods, the Triassic, Jurassic, and Cretaceous. Based on their average age and paleolatitude, we assigned each taxon a geographic- and time-specific formation count as an additional independent variable in our regression analyses.

Interspecific regression analyses

We conducted Bayesian phylogenetic generalised least squares regressions using \log_{10} -transformed femur length as the dependent variable. Given our interest in testing for the effect of temperature gradients, using latitude as a proxy, on body size, both northerly and southerly, we used the absolute value of paleolatitude as our primary independent variable. We also ran additional models of increasing complexity that included dummy-coded indicator variables for hemisphere location (northern or southern hemisphere), geologic period (Triassic, Jurassic, or Cretaceous), and the clade (Theropoda, Sauropodomorpha, and Ornithischia/non-dinosaur Dinosauromorpha) as well as their interactions with absolute paleolatitude. These indicator variables let us test for a difference in the effect of paleolatitude on body size

across space, time, and taxonomic groups. We also tested if absolute paleolatitude explains body size after accounting for the tip ages of species in the phylogeny.

We used BayesTraits V4 to conduct our interspecific regression analyses (<http://www.evolution.reading.ac.uk/BayesTraitsV4.0.0/BayesTraitsV4.0.0.html>). All analyses ran for 1,000,000 iterations with a 250,000-iteration burn-in and sampling frequency of 1,000. We estimated log marginal likelihoods using the Stepping Stone algorithm⁵⁰ with 100 stones sampled every 1,000 iterations. In addition, we allowed the model to sample a distribution of values for taxa with multiple femur lengths and paleolatitude records using the 'DistData' command. We also estimated phylogenetic signal in the relationship between body size and paleolatitude using Pagel's lambda. A lambda of 1 indicates high phylogenetic signal. Then, we compared the fit of eight models of varying complexity by calculating BayesFactors (BF) from their estimated log marginal likelihoods, where a BF > 5 is considered good evidence for the model with the higher marginal likelihood. We selected the model with the highest log marginal likelihood and assessed the statistical significance of each regression coefficient by calculating the proportion of slope (β) parameter estimates that crossed a value of 0 (p_{MCMC}). A low p_{MCMC} means that a considerable proportion of the slope estimates deviates from a flat line. We ensured that our independent variables did not carry similar information (i.e., multicollinearity) by calculating variance inflation factors (VIFs) with the package car in R (Table S1, Supplementary Materials)⁵¹. After model selection, we also assessed the assumptions of equal variance and normality while accounting for phylogenetic non-independence (Figures S1–3, Supplementary Materials)⁵².

We repeated the regression analyses with 62 Mesozoic mammaliaformes. We obtained body mass data (\log_{10} grams) from Slater and colleagues⁵³, paleolatitude

data from the PBDB, and a phylogeny of extinct mammals from Huttenlocker and colleagues⁵⁴. Like our dinosaur analysis, we compared four regression models of varying complexity to test for differences in the effect of paleolatitude on body mass between hemispheres and geologic periods as well as the effect of geographic sampling bias on our regression results. Due to our small sample of Triassic taxa ($n = 2$), we did not test for differences in the effect across geologic periods. We assessed statistical support using the proportion of parameter estimates that cross a value of 0 (p_{MCMC}). As with our dinosaur models, we ensured that our independent variables were not multicollinear and assessed the regression model assumption of equal variance and normality.

We also tested for a relationship between body mass (\log_{10} grams) and absolute mid-range latitude using 2,566 extant mammals from the PanTHERIA database⁵⁵ with additional latitudinal data for *Ursus maritimus* (polar bear) from the southern Beaufort Sea (averaged across three decades⁵⁶). We used a Bayesian independent contrasts regression model in BayesTraits V4. We further tested for a difference in the effect of absolute latitude across the 14 most-represented mammalian orders in the dataset ($n \geq 10$ species) by applying 13 dummy-coded indicators, using Rodentia as our baseline group (all indicator variables = 0). We removed species from the least-represented orders, including the Dermoptera ($n = 1$), Erinaceomorpha ($n = 8$), Microbiotheria ($n = 1$), Monotremata ($n = 3$), Notoryctemorphia ($n = 1$), Paucituberculata ($n = 3$), Pholidota ($n = 2$), and Pilosa ($n = 9$). We followed the protocol of Baker et al.¹⁵ and combined the following orders into the monophyletic super-order Afrotheria: Afrosoricida ($n = 39$), Hyracoidea ($n = 4$), Macroscelidea ($n = 11$), Proboscidea ($n = 3$), and Tubulidentata ($n = 1$). We compared the separate effects model to one where absolute latitude had the same effect on body

mass across all mammals and one where the average body mass differed among the orders but absolute latitude had the same effect. Our regression analyses ran for 125 million iterations, sampling every 10,000 iterations, and discarding the first 25 million as burn-in. To account for variation in geographic range, we used the 'DistData' command to randomly sample the absolute minimum and maximum latitudes throughout the analyses. For model comparisons, we estimated the log marginal likelihoods of each model using a stepping stones algorithm, using 100 stones and sampling for 10,000 iterations. We assessed statistical significance using the proportion of parameter estimates that crossed a value of 0 (p_{MCMC}). We made sure our independent variables were not multicollinear and assessed the regression model assumption of equal variance and normality.

References

1. Lawton, J. H. Are there general laws in ecology? *Oikos* **84**, 177–192 (1999).
2. Theodoridis, S. *et al.* Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nature Communications* **11**, 2557 (2020).
3. Ackerly, D. D. *et al.* The geography of climate change: implications for conservation biogeography. *Diversity and Distributions* **16**, 476–487 (2010).
4. Bergmann, C. Über die Verhältnisse der Wärmeökonomie der Thiere zu ihrer Grösse. *Göttinger Studien* **3**, 595–708 (1847).
5. Blackburn, T. M., Gaston, K. J. & Loder, N. Geographic gradients in body size: a clarification of Bergmann's rule. *Diversity and Distributions* **5**, 165–174 (1999).
6. Meiri, S. Bergmann's Rule – what's in a name? *Global Ecology and Biogeography* **20**, 203–207 (2011).

7. Blackburn, T. M. & Gaston, K. J. Spatial patterns in the body sizes of bird species in the New World. *Oikos* **77**, 436–446 (1996).
8. Salewski, V. & Watt, C. Bergmann's rule: a biophysiological rule examined in birds. *Oikos* **126**, (2017).
9. Adams, D. C. & Church, J. O. Amphibians do not follow Bergmann's rule. *Evolution* **62**, 413–420 (2008).
10. Ashton, K. G. & Feldman, C. R. Bergmann's rule in nonavian reptiles: turtles follow it, lizards and snakes reverse it. *Evolution* **57**, 1151–1163 (2003).
11. Belk, M. C. & Houston, D. D. Bergmann's rule in ectotherms: A test using freshwater fishes. *The American Naturalist* **160**, 803–808 (2002).
12. Berke, S. K., Jablonski, D., Krug, A. Z., Roy, K. & Tomasovych, A. Beyond Bergmann's rule: size–latitude relationships in marine Bivalvia world-wide. *Global Ecology and Biogeography* **22**, 173–183 (2013).
13. Brehm, G. & Fiedler, K. Bergmann's rule does not apply to geometrid moths along an elevational gradient in an Andean montane rain forest. *Global Ecology and Biogeography* **13**, 7–14 (2004).
14. Shelomi, M. Where are we now? Bergmann's rule sensu lato in insects. *The American Naturalist* **180**, 511–519 (2012).
15. Baker, J., Meade, A., Pagel, M. & Venditti, C. Adaptive evolution toward larger size in mammals. *Proceedings of the National Academy of Science* **112**, 5093–5098 (2015).
16. Blackburn, T. M. & Gaston, K. J. Spatial Patterns in the Body Sizes of Bird Species in the New World. *Oikos* **77**, 436–446 (1996).

17. Kulik, Z. T. & Sidor, C. A. A test of Bergmann's rule in the Early Triassic: latitude, body size, and sampling in *Lystrosaurus*. *Paleobiology* 1–15 (2022)
doi:10.1017/pab.2022.25.
18. Prothero, D. R. *et al.* Size and shape stasis in late Pleistocene mammals and birds from Rancho La Brea during the Last Glacial–Interglacial cycle. *Quaternary Science Reviews* **56**, 1–10 (2012).
19. Erickson, G. M. On dinosaur growth. *Annual Review of Earth and Planetary Sciences* **42**, 675–697 (2014).
20. Erickson, G. M. On Dinosaur Growth. *Annual Review of Earth and Planetary Sciences* **42**, 675–697 (2014).
21. Wiemann, J. *et al.* Fossil biomolecules reveal an avian metabolism in the ancestral dinosaur. *Nature* 1–5 (2022) doi:10.1038/s41586-022-04770-6.
22. Landwehrs, J., Feulner, G., Petri, S., Sames, B. & Wagreich, M. Investigating Mesozoic climate trends and sensitivities with a large ensemble of climate model simulations. *Paleoceanography and Paleoclimatology* **36**, e2020PA004134 (2021).
23. O'Donovan, C., Meade, A. & Venditti, C. Dinosaurs reveal the geographical signature of an evolutionary radiation. *Nature Ecology & Evolution* **2**, 452–458 (2018).
24. Gardner, J. D., Surya, K. & Organ, C. L. Early tetrapodomorph biogeography: Controlling for fossil record bias in macroevolutionary analyses. *Comptes Rendus Palevol* **18**, 699–709 (2019).
25. Pimm, S. L., Jones, H. L. & Diamond, J. On the risk of extinction. *The American Naturalist* **132**, 757–785 (1988).

26. Cardillo, M. *et al.* Multiple causes of high extinction risk in large mammal species. *Science* **309**, 1239–1241 (2005).
27. Benson, R. B. J., Hunt, G., Carrano, M. T. & Campione, N. Cope's rule and the adaptive landscape of dinosaur body size evolution. *Palaeontology* **61**, 13–48 (2018).
28. Brown, C. M., Campione, N. E., Mantilla, G. P. W. & Evans, D. C. Size-driven preservational and macroecological biases in the latest Maastrichtian terrestrial vertebrate assemblages of North America. *Paleobiology* 1–29 (2021) doi:10.1017/pab.2021.35.
29. Mannion, P. D., Upchurch, P., Benson, R. B. J. & Goswami, A. The latitudinal biodiversity gradient through deep time. *Trends in Ecology & Evolution* **29**, 42–50 (2014).
30. Druckenmiller, P. S., Erickson, G. M., Brinkman, D., Brown, C. M. & Eberle, J. J. Nesting at extreme polar latitudes by non-avian dinosaurs. *Current Biology* S0960982221007399 (2021) doi:10.1016/j.cub.2021.05.041.
31. Bell, P. R. & Snively, E. Polar dinosaurs on parade: a review of dinosaur migration. *Alcheringa: An Australasian Journal of Palaeontology* **32**, 271–284 (2008).
32. Chinsamy, A., Thomas, D. B., Tumarkin-Deratzian, A. R. & Fiorillo, A. R. Hadrosaurs were perennial polar residents. *The Anatomical Record* **295**, 610–614 (2012).
33. Chiarenza, A. A. *et al.* The first juvenile dromaeosaurid (Dinosauria: Theropoda) from Arctic Alaska. *PLOS ONE* **15**, e0235078 (2020).
34. Herman, A. B., Spicer, R. A. & Spicer, T. E. V. Environmental constraints on terrestrial vertebrate behaviour and reproduction in the high Arctic of the Late

- Cretaceous. *Palaeogeography, Palaeoclimatology, Palaeoecology* **441**, 317–338 (2016).
35. Fiorillo, A. R. On the occurrence of exceptionally large teeth of *Troodon* (Dinosauria: Saurischia) from the Late Cretaceous of Northern Alaska. *paleo* **23**, 322–328 (2008).
36. Fiorillo, A. R. & Tykoski, R. S. A New Maastrichtian Species of the Centrosaurine Ceratopsid *Pachyrhinosaurus* from the North Slope of Alaska. *Acta Palaeontologica Polonica* **57**, 561–573 (2012).
37. Tykoski, R. S., Fiorillo, A. R. & Chiba, K. New data and diagnosis for the Arctic ceratopsid dinosaur *Pachyrhinosaurus perotorum*. *Journal of Systematic Palaeontology* **17**, 1397–1416 (2019).
38. Rensch, B. Some problems of geographical variation and species-formation. *Proceedings of the Linnean Society of London* **150**, 275–285 (1938).
39. Gillette, D. D. & Madsen, D. B. The short-faced bear *Arctodus simus* from the late Quaternary in the Wasatch Mountains of central Utah. *Journal of Vertebrate Paleontology* **12**, 107–112 (1992).
40. Raymond, K. R. & Prothero, D. R. Did climate changes affect size in Late Pleistocene bison? *New Mexico Museum of Natural History and Science, Bulletin* **53**, 5 (2011).
41. Hill, C., Wilson, M. & McDonald, H. Fossil ground sloths, *Megalonyx* and *Paramylodon* (Mammalia: Xenarthra), from the Doeden Local Fauna (Illinoian/Sangamonian?) eastern Montana. *Current Research in the Pleistocene* **22**, 83–85 (2005).
42. Blackburn, T. M. & Hawkins, B. A. Bergmann's rule and the mammal fauna of northern North America. *Ecography* **27**, 715–724 (2004).

43. Ashton, K. G., Tracy, M. C. & Queiroz, A. de. Is Bergmann's rule valid for mammals? *The American Naturalist* **156**, 390–415 (2000).
44. Meiri, S. & Dayan, T. On the validity of Bergmann's rule. *Journal of Biogeography* **30**, 331–351 (2003).
45. Gohli, J. & Voje, K. L. An interspecific assessment of Bergmann's rule in 22 mammalian families. *BMC Evolutionary Biology* **16**, 222 (2016).
46. Chiarenza, A. A., Mannion, P. D., Farnsworth, A., Carrano, M. T. & Varela, S. Climatic constraints on the biogeographic history of Mesozoic dinosaurs. *Current Biology* **32**, 570-585.e3 (2022).
47. Benson, R. B. J. & Upchurch, P. Diversity trends in the establishment of terrestrial vertebrate ecosystems: Interactions between spatial and temporal sampling biases. *Geology* **41**, 43–46 (2013).
48. Close, R. A. *et al.* The apparent exponential radiation of Phanerozoic land vertebrates is an artefact of spatial sampling biases. *Proceedings of the Royal Society B: Biological Sciences* **287**, 20200372 (2020).
49. Jones, L. A., Dean, C. D., Mannion, P. D., Farnsworth, A. & Allison, P. A. Spatial sampling heterogeneity limits the detectability of deep time latitudinal biodiversity gradients. *Proceedings of the Royal Society B: Biological Sciences* **288**, 20202762 (2021).
50. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology* **60**, 150–160 (2011).
51. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (Sage, 2019).
52. Freckleton, R. P. The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology* **22**, 1367–1375 (2009).

53. Slater, G. J. Phylogenetic evidence for a shift in the mode of mammalian body size evolution at the Cretaceous-Palaeogene boundary. *Methods in Ecology and Evolution* **4**, 734–744 (2013).
54. Huttenlocker, A. K., Grossnickle, D. M., Kirkland, J. I., Schultz, J. A. & Luo, Z.-X. Late-surviving stem mammal links the lowermost Cretaceous of North America and Gondwana. *Nature* **558**, 108–112 (2018).
55. Jones, K. *et al.* PanTHERIA: a species-level database of life history, ecology, and geography of extant and recently extinct mammals. (2016)
doi:10.6084/m9.figshare.c.3301274.v1.
56. Durner, G. M., Douglas, D. C. & Atwood, T. C. Are polar bear habitat resource selection functions developed from 1985–1995 data still useful? *Ecology and Evolution* **9**, 8625–8638 (2019).
57. Kocsis, Á. T. & Raja, N. B. chronosphere: Earth system history variables. (2020)
doi:10.5281/zenodo.3530703.

Appendix 1

Extended Data and Figures

Extended Data Table 1. Model selection results for the Mesozoic dinosaur analyses.

Each row includes the regression model description (uniform rate), the number of estimated parameters (No. Par.), the estimated log marginal likelihood (LH), and Bayes factor (BF) compared to Model 1. The number of parameters also include the y-intercept and Pagel's lambda, the phylogenetic signal parameter.

Model	No. Par.	LH	BF
1. Femur ~ AbsLat	3	14.314	-
2. Femur ~ AbsLat + Formations	4	2.096	24.437
3. Femur ~ AbsLat + Formations + Hemisphere	5	0.602	27.426
4. Femur ~ AbsLat + Formations + Hemisphere + AbsLat:Hemisphere	6	-9.249	47.127
5. Femur ~ AbsLat + Formations + Hemisphere + AbsLat:Hemisphere + Jurassic + Cretaceous	8	-17.474	63.576
6. Femur ~ AbsLat + Formations + Hemisphere + AbsLat:Hemisphere + Jurassic + Cretaceous + AbsLat:Jurassic + AbsLat:Cretaceous	10	-42.052	112.733
7. Femur ~ AbsLat + Formations + Hemisphere + AbsLat:Hemisphere + Jurassic + Cretaceous + AbsLat:Jurassic + AbsLat:Cretaceous + Sauropodomorpha + Theropoda	13	-44.991	118.611

8. Femur ~ AbsLat + Formations + Hemisphere + AbsLat:Hemisphere + Jurassic + Cretaceous + AbsLat:Jurassic + AbsLat:Cretaceous + Sauropodomorpha + Theropoda + AbsLat:Sauropodomorpha + AbsLat:Theropoda	14	-44.032	116.692
9. Femur ~ AbsLat + Hemisphere	4	7.176	14.276
10. Femur ~ AbsLat + Hemisphere + AbsLat:Hemisphere	5	-1.521	31.670
11. Femur ~ AbsLat + Jurassic + Cretaceous	5	3.292	22.044
12. Femur ~ AbsLat + Jurassic + Cretaceous + AbsLat:Jurassic + AbsLat:Cretaceous	7	-15.149	58.927
13. Femur ~ AbsLat + Sauropodomorpha + Theropoda	5	4.070	20.489
14. Femur ~ AbsLat + Sauropodomorpha + Theropoda + AbsLat:Sauropodomorpha + AbsLat:Theropoda	7	-16.666	61.960
15. Femur ~ Tip age + AbsLat	4	9.815	8.998

Extended Data Table 2. Model selection results for the Mesozoic mammal analyses.

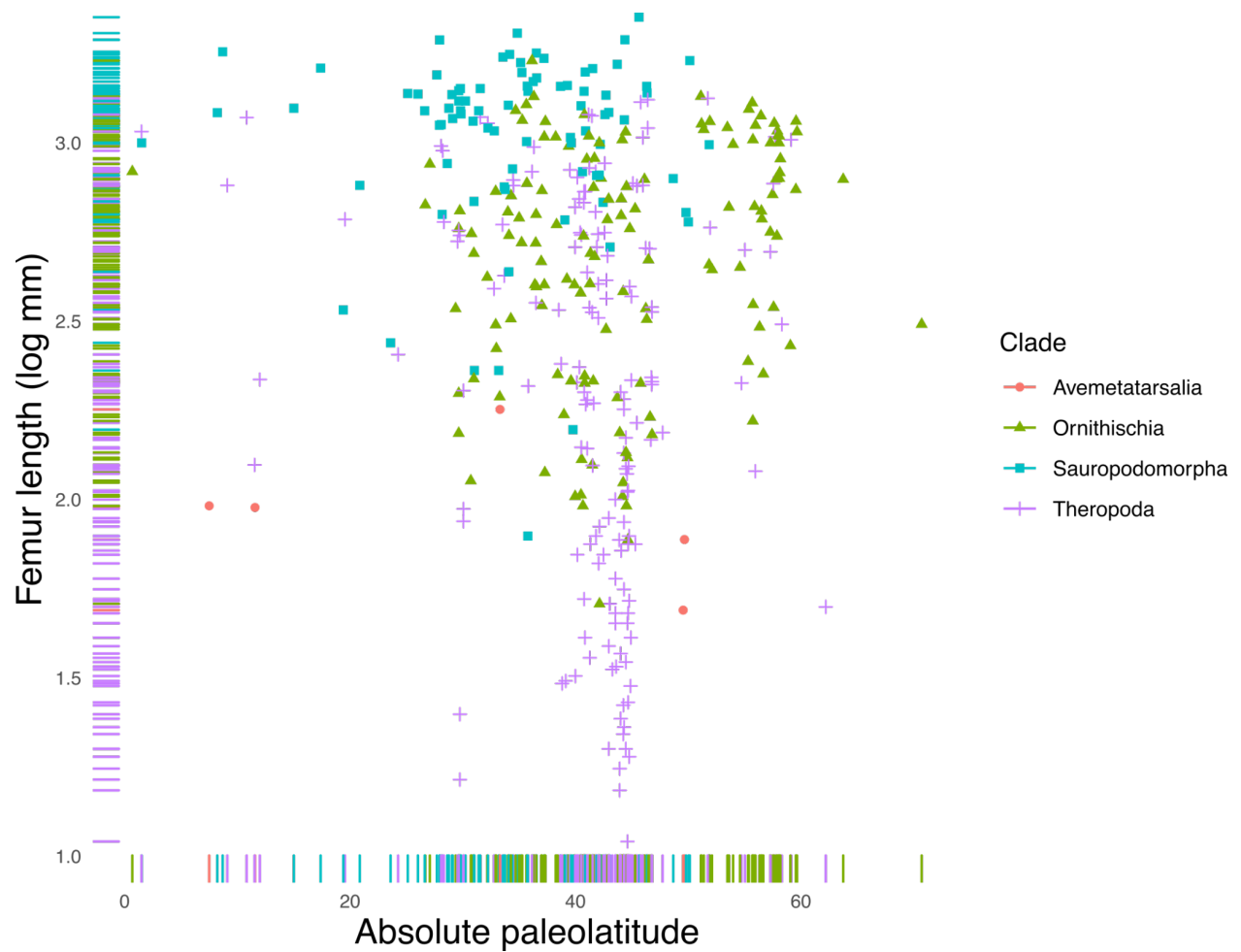
Each row includes the regression model description (uniform rate), the number of estimated parameters (No. Par.), the estimated log marginal likelihood (LH), and Bayes factor (BF) compared to Model 1. The number of parameters also include the y-intercept and Pagel's lambda, the phylogenetic signal parameter.

Model	No. Par.	LH	BF
1. Body Mass ~ AbsLat	3	-78.331	-
2. Body Mass ~ AbsLat + Formations	4	-90.825	24.988
3. Body Mass ~ AbsLat + Formations + Hemisphere	5	-94.767	32.873
4. Body Mass ~ AbsLat + Formations + Hemisphere + AbsLat:Hemisphere	6	-100.200	43.738

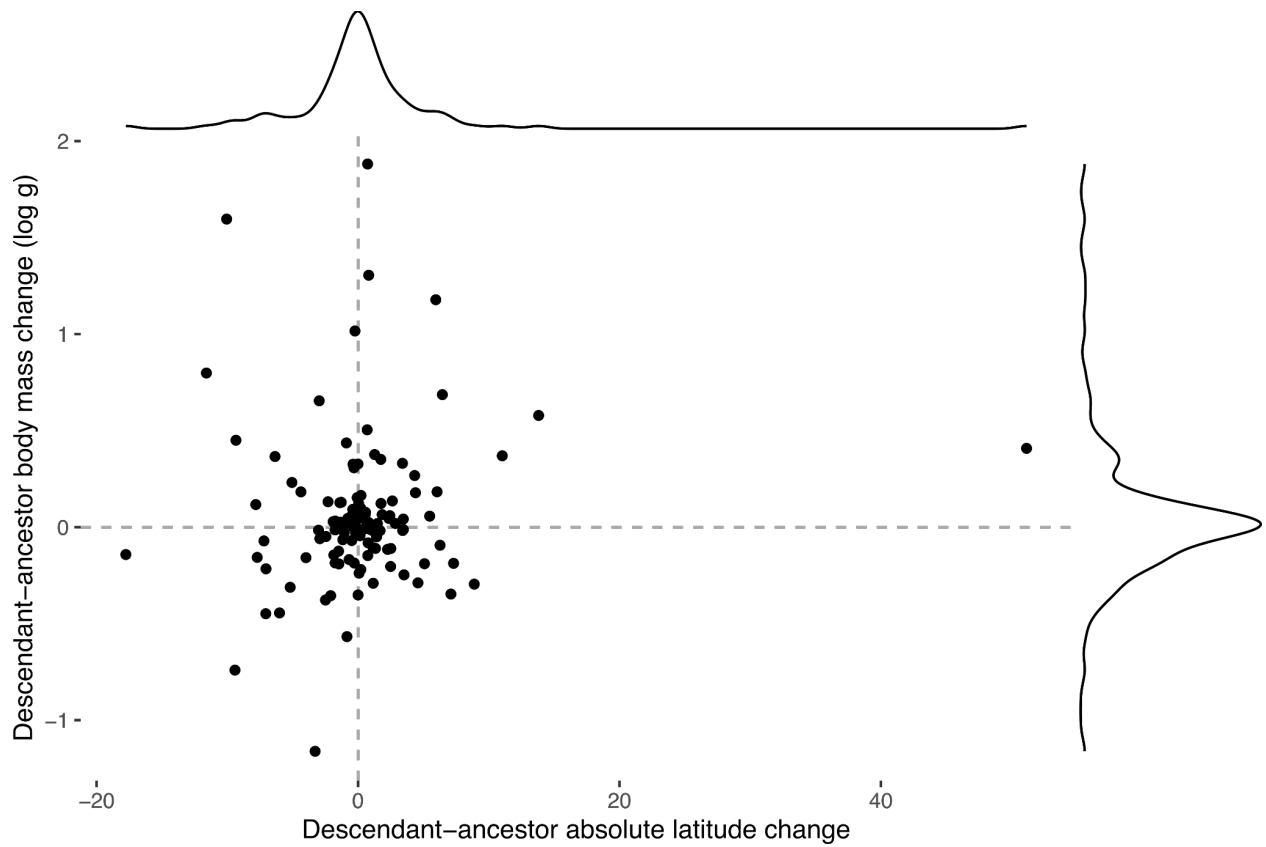
Extended Data Table 3. Results of the final variable-rates regression models for Mesozoic dinosaurs and mammals, and extant mammals. Each row includes the sample size of each analysis (N), the average estimated slope for the effect of paleolatitude (β_1) and associated Bayesian credible interval (95% CI), the proportion of slope estimates that crossed a value of 0 (p_{MCMC}), and the average estimated correlation coefficient (R^2) and associated 95% CI.

Group	N	B_1 (95% CI)	p_{MCMC}	R^2 (95% CI)
Dinosaurs	382	0.0002 (-0.001, 0.0017)	0.41	-0.031 (-0.0467, -0.0185)
Mammaliaforms	62	0.009 (-0.007, 0.0261)	0.12	-0.0087 (-0.082, 0.04)
Extant mammals	2566	0.001 (3.86E-5, 0.0019)	0.021	0.0019 (-0.0004, 0.0052)
Extant mammals (sample lat min & max)	2566	0.001 (-1.3E-5, 0.0024)	0.026	0.007 (-0.0003, 0.025)

Extended Data Figure 1. Body size does not covary with paleolatitude across major clades of Mesozoic dinosaurs and dinosauromorphs. Femur length (\log_{10} mm) as a function of absolute paleolatitude in 382 Mesozoic dinosaurs and other dinosauromorphs with colours and symbols showing the distribution of different clades.



Extended Data Figure 2. Ancestral shifts in paleolatitude do not explain ancestral body size change in Mesozoic mammals. Estimated branch-specific changes in body mass (\log_{10} g) as a function of estimated branch-specific changes in absolute paleolatitude.



Appendix 2

Supplementary Information

1. Regression model assumptions

We ensured that our independent variables did not carry redundant information (i.e., multicollinearity) by calculating variance inflation factors (VIFs) with the package *car* in R¹. Two or more variables are collinear if they share a $VIF > 5.0$. After removing the interaction variables, we found that multicollinearity was absent in our full models. We did not estimate VIFs for our final models because they only included one independent variable (absolute paleolatitude).

Table S1. Variance inflation factors for all independent variables in the full regression models, excluding interactions, for the Mesozoic dinosaurs and mammaliaformes.

Model	AbsLat	Formations	Hemisphere	Jurassic	Cretaceous	Sauropod	Theropod
<i>Mesozoic dinosaur full model, no interactions</i>	1.34	2.08	1.55	3.77	4.62	1.65	1.30
<i>Mesozoic mammal full model, no interactions</i>	1.26	1.30	1.47	-	-	-	-

Table S2. Variance inflation factors for all independent variables in the full separate-effects regression model, excluding interactions, for the extant mammals.

Model	AbsLat	Afrotheria	Artiodactyla	Carnivora	Chiroptera	Cingulata	Dasyromorphia	Didelphimorphia
<i>Extant mammal full model, no interactions</i>	1.34	1.04	1.11	1.11	1.32	1.01	1.04	1.03

-	Diprotodonta	Lagomorpha	Peramorphia	Perissodactyla	Primates	Scandentia	Soricomorpha	-
-	1.06	1.05	1.01	1.01	1.19	1.02	1.09	

After model selection, we also checked the assumptions of equal variance and normality while accounting for phylogenetic non-independence². We minimized the violations in regression model assumptions by log-transforming femur length (Mesozoic dinosaurs and other dinosauromorphs) and body mass (Mesozoic and extant mammals). However, there were still violations of equal variance (Figures S1–2). The final model with Mesozoic mammaliaformes also deviates from normality (Figure S2). We further found convergence in the Markov-chain Monte Carlo chains of our final models.

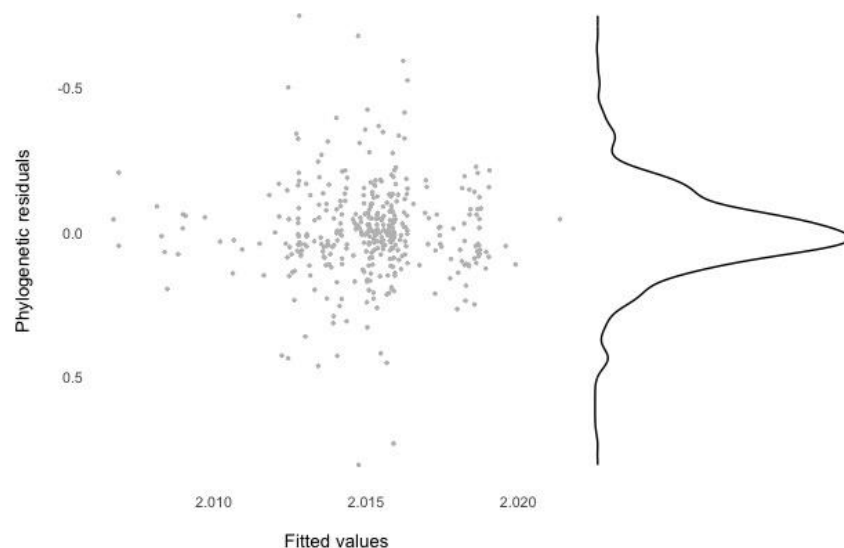


Figure S1. Assessment of regression model assumptions for the final model with Mesozoic dinosaurs and other dinosauromorphs. After log-transforming femur length, we still see minor violations in equal variance assumption.

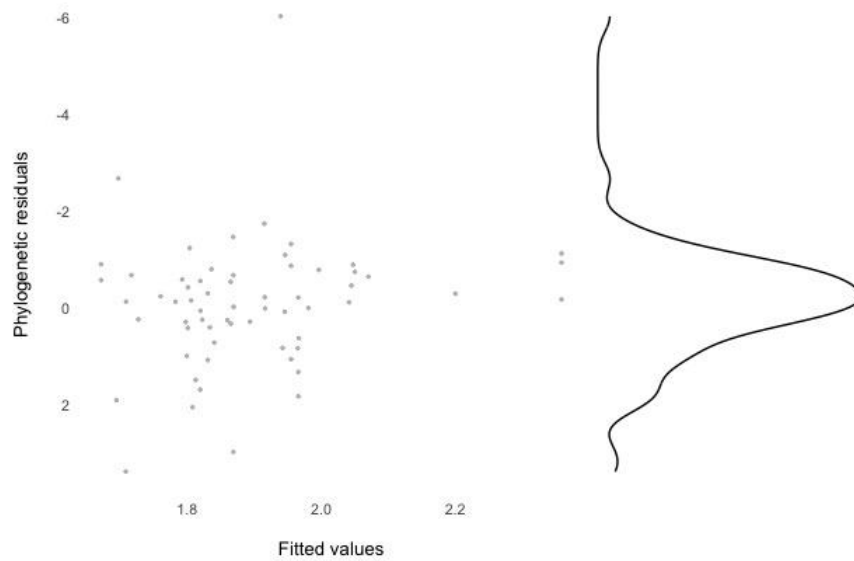


Figure S2. Assessment of regression model assumptions for the final model with Mesozoic mammaliaformes. After log-transforming body mass, we still see small violations in the equal variance and normality assumptions.

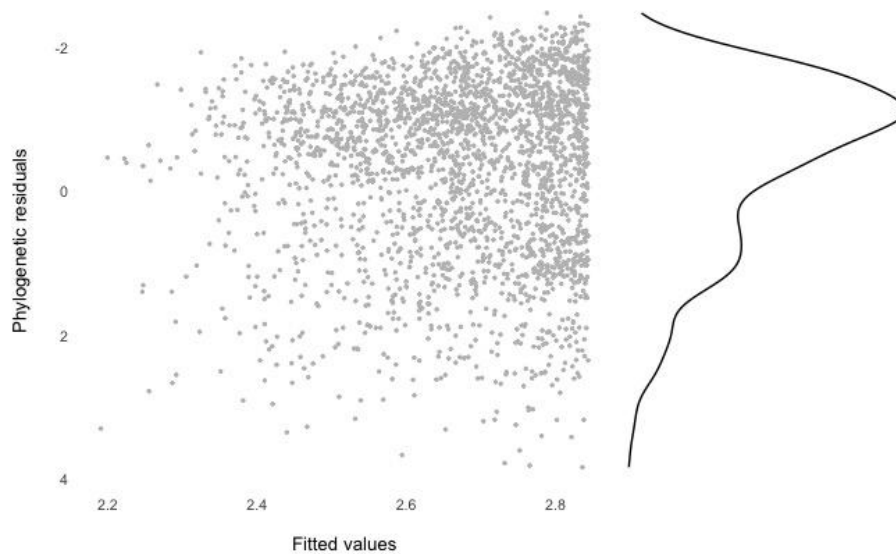


Figure S3. Assessment of regression model assumptions for the single slope model with extant mammals. After log-transforming body mass, we still see a violation in the normality assumption.

2. Full regression results

Here, we provide more complete results from the models referenced in the main text, including the final regression models (Table S3) and the models including our geographic sampling bias metric (Table S4), and the dinosaur models using imputed body masses and distributions of femur length and paleolatitude data (Table S5).

Table S3. Full results for the final models with Mesozoic dinosaurs and mammaliaformes. The extant mammal analysis was conducted using phylogenetic independent contrasts, in which the sum of squared error (SSE), total sum of squares (SST), and the standard errors for the alpha (y-intercept) parameter were not estimated, and Pagel's lambda (phylogenetic signal) was assumed to be 1.0.

Model	Alpha (95% CI)	Beta 1 (95% CI)	p _{MCMC}	Var (95% CI)	R ² (95% CI)	SSE (95% CI)	SST (95% CI)	s.e. Alpha (95% CI)	s.e. Beta-1 (95% CI)	Lambda (95% CI)
<i>Mesozoic dinosaur, Log femur length ~ absolute paleolatitude</i>	2.01 (1.81, 2.17)	0.0002 (-0.002, 0.0025)	0.42	0.0019 (0.0017, 0.0021)	5.0E-5 (3.3E-5, 7.0E-5)	0.71 (0.63, 0.79)	0.71 (0.63, 0.79)	0.093 (0.089, 0.096)	0.00116 (0.00115, 0.00118)	0.96 (0.94, 0.98)
<i>Mesozoic mammal, Log body mass ~ absolute paleolatitude</i>	1.38 (0.56, 2.18)	0.012 (-0.004, 0.027)	0.075	0.45 (0.31, 0.67)	0.035 (0.030, 0.037)	27.02 (19.10, 39.92)	27.98 (19.82, 41.16)	0.40 (0.33, 0.48)	0.0080 (0.0072, 0.0088)	0.51 (0.10, 0.93)
<i>Extant mammal, Log body mass ~ absolute mid-range latitude</i>	2.844 (2.842, 2.845)	-0.0091 (-0.0092, - 0.0090)	<0.001	0.0132 (0.0132, 0.0132)	0.9213 (0.9212, 0.9213)	-	-	-	5.291E-5 (5.29E-5, 5.294E-5)	1.0
<i>Extant ursids, Log body mass ~ absolute mid-range latitude</i>	4.82 (4.21, 5.27)	0.0097 (0.0043, 0.015)	0.0027	0.0017 (0.0008, 0.0032)	0.754 (0.687, 0.832)	0.0092 (0.0065, 0.013)	0.04 (0.021, 0.078)	0.116 (0.08, 0.16)	0.0023 (0.0016, 0.0033)	0.57 (0.06, 0.99)

Table S4. Full results for the models, including the geographic-specific formation count as a covariate.

Model	Alpha (95% CI)	Beta 1 (95% CI)	p _{MCMC} 1	Beta 2 (95% CI)	p _{MCMC} 2	Var (95% CI)	R ² (95% CI)	SSE (95% CI)	SST (95% CI)	s.e. Alpha (95% CI)	s.e. Beta- 1 (95% CI)	s.e. Beta-2 (95% CI)	Lambda a (95% CI)
<i>Mesozoic dinosaur, Log femur length ~ absolute Paleolatitude + formation count</i>	2.03 (1.83, 2.21)	-0.0003 (-0.0029, 0.0020)	0.39	-0.0001 (-0.0003, 3.13E-5)	0.062	0.0018 (0.0016, 0.0021)	0.0049 (1.08E-7, 0.0073)	0.701 (0.617, 0.781)	0.704 (0.623, 0.787)	0.094 (0.090, 0.098)	0.00120 (0.00119, 0.00121)	7.95E-05 (7.8E-5, 8.1E-5)	0.96 (0.94, 0.98)
<i>Mesozoic mammal, Log body mass ~ absolute paleolatitude + formation count</i>	1.35 (0.43, 2.13)	0.012 (-0.0037, 0.028)	0.072	0.0003 (-0.001, 0.002)	0.34	0.47 (0.32, 0.68)	0.027 (0.0005, 0.04)	27.63 (19.12, 40.65)	28.39 (19.82, 41.84)	0.42 (0.36, 0.49)	0.0081 (0.0074, 0.0089)	6.16E-4 (5.71E-4, 6.67E-4)	0.52 (0.089, 0.91)

Table S5. Full results for the dinosaur models, including body mass as the dependent variable, when incorporating a femur length estimate for the Prince Creek Formation's *Pachyrhinosaurus perotorum* as well as a distribution of femur lengths and paleolatitudes for 195 species (PCF Pachy-DistData), and when accounting for time (species tip ages).

Model	Alpha (95% CI)	Beta 1 (95% CI)	p _{MCMC} 1	Beta 2 (95% CI)	p _{MCMC} 2	Var (95% CI)	R ² (95% CI)	SSE (95% CI)	SST (95% CI)	s.e. Alpha (95% CI)	s.e. Beta-1 (95% CI)	s.e. Beta-2 (95% CI)	Lambda (95% CI)
<i>Mesozoic dinosaur, Log mass ~ abs paleolatitude</i>	2.03 (1.83, 2.21)	-0.0003 (- 0.0029, 0.0020)	0.39	-	-	0.0018 (0.0016, 0.0021)	0.0049 (1.08E- 7, 0.0073)	0.701 (0.617, 0.781)	0.704 (0.623, 0.787)	0.094 (0.090, 0.098)	0.00120 (0.00119, 0.00121)	-	0.96 (0.94, 0.98)
<i>Dinosaur, PCF Pachy-DistData Log femur length ~ abs paleolatitude</i>	1.35 (0.43, 2.13)	0.012 (- 0.0037, 0.028)	0.072	-	-	0.47 (0.32, 0.68)	0.027 (0.0005, 0.04)	27.63 (19.12, 40.65)	28.39 (19.82, 41.84)	0.42 (0.36, 0.49)	0.0081 (0.0074, 0.0089)	-	0.52 (0.089, 0.91)
<i>Mesozoic dinosaur, Log mass ~ time + abs paleolatitude</i>	1.35 (0.43, 2.13)	0.012 (- 0.0037, 0.028)	0.072	0.0003 (-0.001, 0.002)	0.34	0.47 (0.32, 0.68)	0.027 (0.0005, 0.04)	27.63 (19.12, 40.65)	28.39 (19.82, 41.84)	0.42 (0.36, 0.49)	0.0081 (0.0074, 0.0089)	6.16E-4 (5.71E-4, 6.67E-4)	0.52 (0.089, 0.91)

3. Latitudinal formation count data

Zone	Formation count			
	Cretaceous	Jurassic	Triassic	All
N_90-70	15	0	1	16
N_70-50	77	18	18	113
N_50-30	445	172	79	696
N_30-10	108	83	134	325
C_10-10	43	9	26	78
S_10-30	41	25	6	72
S_30-50	53	23	40	116
S_50-70	26	10	11	47
S_70-90	5	2	9	16

4. References cited

1. Fox, J. & Weisberg, S. *An R Companion to Applied Regression*. (Sage, 2019).
2. Freckleton, R. P. The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology* **22**, 1367–1375 (2009).

Chapter 4

Dinosaur Diversity and Ecology was Driven by Limb Functional Evolution

Abstract

Animals are defined by their ability to move through environments. They move to find food¹, new habitats², and mates³. As such, locomotor adaptations are central to species' success across ecosystems and through evolution. Comparative and biomechanical studies have clarified many shared and unique features of functional systems^{4,5}. But how selection creates and shapes them from ancestor to descendant has remained a mystery. Here, we develop an approach that allows parameters of functional equations to evolve at varying rates along branches of an evolutionary tree and use it to investigate the evolution of limb retraction in dinosaurs. Many individual parameters of these lever systems evolve gradually; however, the levers themselves evolve at varying rates, revealing emergent evolutionary patterns. Evolutionary rates of limb retraction are correlated with lineage divergence, suggesting that locomotor evolution is associated with speciation through dispersal or niche partitioning. Accelerated evolutionary rates of hindlimb retraction were also associated with the modular uncoupling of the hindlimb and tail that preceded multiple origins of flight in maniraptoran dinosaurs^{6,7}. Our results help explain how dinosaurs diversified ecologically and dispersed globally and provide a rigorous framework to study how components of a functional system interact to produce adaptive change.

Locomotion is crucial for species success and survival¹. Novel adaptations in vertebrate evolution often involved transitions to new forms of movement, including terrestriality⁵ and flight⁶. Comparative and biomechanical studies have provided insights, from major biomechanical innovations^{4,8–11} to ecomorphometrics and niche partitioning in both extant^{12–14} and extinct¹⁵ species. In such studies, comparing ecomorphological traits is a common approach for assessing functional diversity and evolution. These simple but useful metrics reflect broad relationships with climate¹², diet¹⁶, diversification¹⁷, and extinction¹³. However, how functional systems evolve from ancestor to descendant remains poorly understood^{18,19}, and there has been no study on the rate at which these systems evolve along lineages. Such a gap hinders our ability to adequately test for drivers of functional diversity and evolution, including biogeographic dispersal, climate, and environmental change. Recently developed methods have enabled researchers to study how the rate of evolution varies through time and across the branches of a phylogenetic tree^{20–24}. These models have enhanced our understanding of archosaur evolution through the study of body size^{25–27}, limb proportions²⁸, and morphological characters²⁹. Yet, such characters and traits are tangential to studying functional change. Here, we advance the study of functional evolution by applying newly developed variable-rate phylogenetic models to functional (lever arm) parameters describing limb retraction in dinosaurs, allowing us for the first time to quantify the evolution of a functional system along ancestor-descendant lineages.

Results and Discussion

We measured the lever gear ratios of the primary forelimb and hindlimb retractor muscles in 106 dinosaur species (see Methods for lever arm descriptions; Extended

Data Fig. 1). The gear ratio is a simple but widely used metric for characterizing general locomotor function^{30,31}. It equals the ratio of two parameters: the lever_{out}, the distance to the loading force (in this case, postural limb length), and lever_{in}, the muscle moment arm or distance to the muscle's line of action. Animals with high gear ratios (e.g., the cheetah, *Acinonyx jubatus*) have relatively longer limbs (longer lever_{out}), yielding a greater range of motion. Lower gear ratios are seen in animals with larger muscle attachments (longer lever_{in}; e.g., the African elephant, *Loxodonta africana*), which increase torque in limb movement. Most archosaurs, including dinosaurs, retracted their hindlimbs with a large muscle (the *m. caudofemoralis longus*) that connects the hindlimb with the tail. Birds and many non-avian maniraptoran dinosaurs have reduced their tail-driven hindlimb retractors, emphasizing more knee-driven movements^{6,32,33}. To uncover the macroevolutionary effects of this transition in maniraptorans, we calculated the gear ratios of the *m. iliotibialis lateralis*, a muscle that retracts the hindlimb and extend the knee (see Methods for further description).

We explored gear ratio variation among dinosaurs through Bayesian phylogenetic generalized least squares (PGLS) regression analyses³⁴ on the gear ratio components (\log_{10} lever_{out} and lever_{in}) and estimated Bayes factors (BF) to compare the likelihood fit of each model (BF > 2.0 is good evidence for the model with the higher marginal likelihood). The gear ratio components are strongly correlated in the forelimbs and hindlimbs ($R^2 = 0.87$ and 0.81 , $p_{\text{MCMC}} < 0.0001$; Fig. 1) and show moderate to high phylogenetic signal (mean $\lambda^{\wedge} = 0.93$ and 0.73 ; BF = 31.05 and 24.44). Body size is associated with the lever_{out} and lever_{in} through limb length. To remove its confounding effects, we tested for gear ratio differences among dinosaur clades using phylogenetic analysis of covariance (ANCOVA). Our final forelimb model shows pennaraptoran dinosaurs (oviraptorosaurs and paravians) with lower-than-

average gear ratios (BF = 48.31). Pennaraptorans retracted their shoulders with greater torque, on average, given their body size (Fig. 1A, Extended Data Fig. 2), governed by their more distally inserting shoulder muscles³⁵. The hindlimb gear ratios do not scale differently among dinosaur clades (BF = 37.38), indicating that body size is the dominant factor in explaining variation in hindlimb retraction (Fig. 1B, Extended Data Fig. 3).

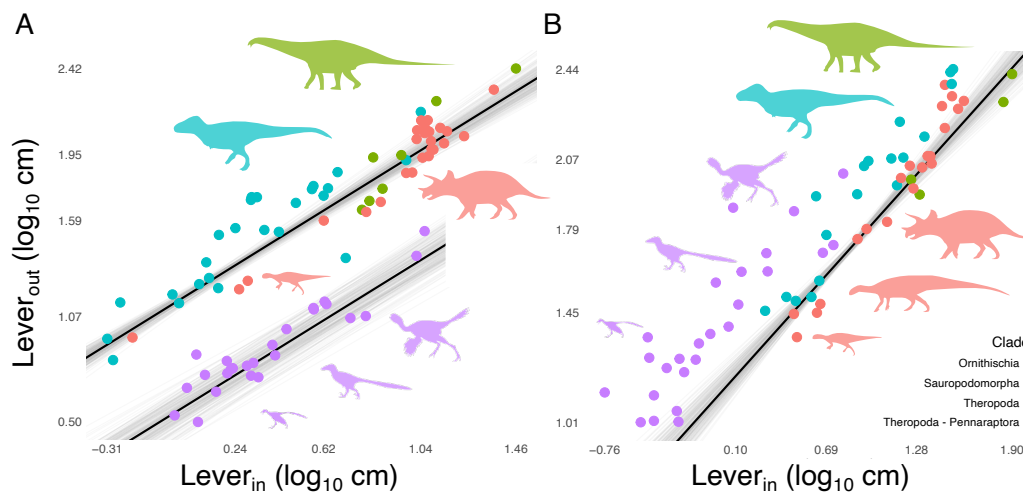


Fig. 1. Gear ratio variation in dinosaurs. Scatter plots of \log_{10} lever_{out} vs. \log_{10} lever_{in} for the A) forelimb and B) hindlimb retractor muscles. Silhouette and data point colours correspond to clades of dinosaurs. Grey regression lines are a random sample of 100 estimates from the variable-rates regression analysis. The mean estimated regression line is shown in black. The following silhouettes were collected from phylopic.org: *Apatosaurus louisae* (Scott Hartman, CC BY-NC-SA 3.0), *Archaeopteryx lithographica* (Scott Hartman, CCO 1.0), *Deinonychus antirrhopus* and *Tyrannosaurus rex* (Emily Willoughby, CC BY-SA 3.0), *Luoyanggia liudianensis* (Brad McFeeters, CCO 1.0), *Psittacosaurus mongoliensis* (Skye McDavid, CC BY 3.0), and *Tenontosaurus tilletti* and *Triceratops prorsus* (Matt Dempsey, CC BY 3.0).

Next, we estimated rates of evolution in the forelimb and hindlimb gear ratio components using a variable-rates phylogenetic model²¹. The model allows the

lever_{out} and lever_{in} parameters to evolve at varying rates along phylogenetic branches^{18,36,37}. It uses a Bayesian reversible jump Markov-chain Monte Carlo algorithm to propose sets of rate scaling parameters (varying $0 < r < \infty$) that either compress ($r < 1.0$) or stretch ($r > 1.0$) phylogenetic branches to reflect the rate of evolution. There are two types of rate scalars: one that scales individual branches, reflecting a mean shift in trait evolution, and a second that scales the branches of an entire clade, reflecting a change in trait variance¹⁸. We first modelled the gear ratio parameters individually to assess the underlying evolutionary processes driving variation in limb morphology. We found no evidence of rate variation in two of the four components comprising the forelimb and hindlimb gear ratios, including the forelimb lever_{out} (BF < 1.77) and hindlimb lever_{out} (BF = 1.13), both representing postural limb length. The hindlimb lever_{in}, representing the muscle moment arm of the hindlimb retractor, shows marginal support for evolutionary rate variation (BF = 2.77), stretching three branches towards higher average gear ratios in the ancestors of alvarezsaurids, paravians, and *Microraptor* (median $r = 1.23x$, $1.17x$, and $1.06x$ background rate, respectively; posterior probability < 70%). However, most branches remained unscaled (median $r = 1$ in 129/132 branches), showing mostly clock-like change in the evolution of the hindlimb retractor muscle.

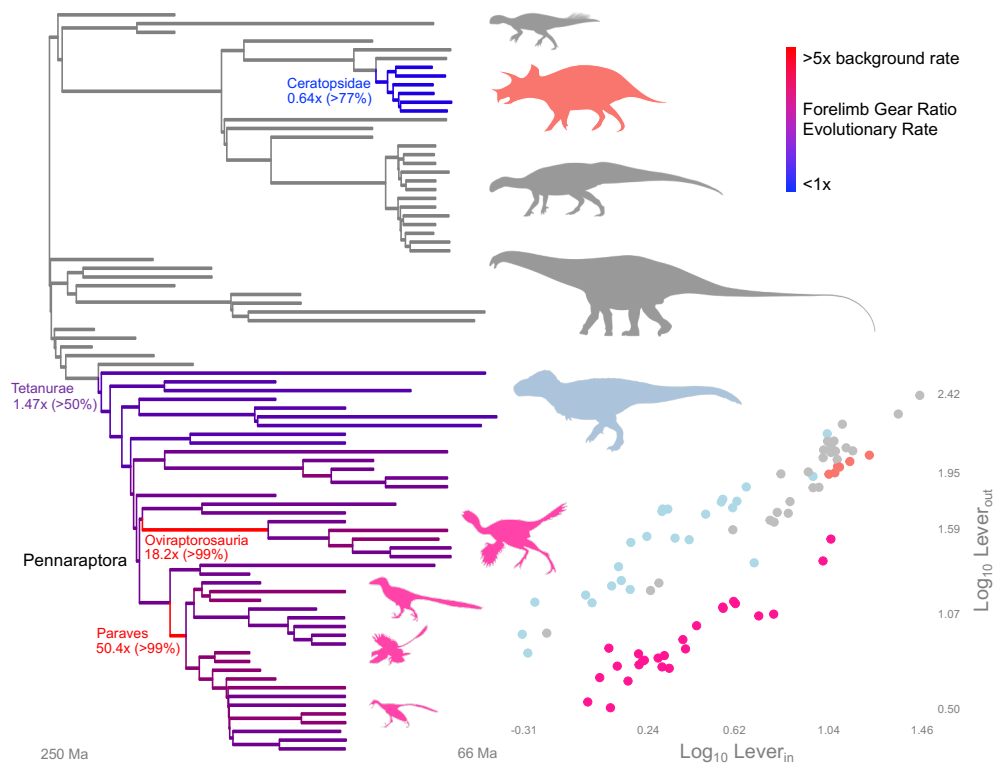


Fig. 2. Evolution of forelimb function. Rates of gear ratio evolution for the primary retractor muscles of the forelimbs mapped onto a dinosaur time-scaled phylogeny. Warmer colours indicate increased rates of evolution relative to the background rate of the tree. Cooler colours indicate reduced rates of evolution. Numbers indicate the median estimated rate shift among the coloured branches, along with the associated percent posterior probability. Silhouette colours correspond to data points in the scatter plot showing the correlation between the forelimb lever_{out} and lever_{in}. The coral-coloured silhouette represents ceratopsid species, blue represents non-maniraptoran theropods, and pink represents pennaraptoran theropods. The *Microaptor zhaoianus* (Emily Willoughby, CC BY-SA 3.0) silhouette obtained from phylopic.org, along with those in Fig. 1.

Yet, these analyses on the gear ratio components do not describe the evolution of limb retraction as a functional system. To do so, we apply a newly developed

variable-rates regression model^{18,38}, which proposes shifts in evolutionary rate by stretching and compressing branches based on trait variation within a regression framework. This approach allows us to both measure gear ratio evolution along phylogenetic branches and assess how its components interact to produce such change. While most gear ratio components evolved gradually, we find that the forelimb and hindlimb gear ratios themselves evolved at varying rates (FL: BF = 29.1; HL: BF = 3.0; Figs. 2–3). This reveals an emergent process by which the underlying morphological components interact to produce substantial changes in the gear ratios; the evolution of a complex functional system is more than the sum of its parts³⁹. Moreover, when comparing the overall evolutionary rates (background σ^2), we find that the lever_{out} and lever_{in} (standardized with each other and the gear ratio) evolve faster on average than the gear ratios themselves (mean difference in background σ^2 = 1.60 and 1.97 for forelimb lever_{out} and lever_{in}, and 1.58 and 2.33 for hindlimb lever_{out} and lever_{in}; ANOVA: p-value < 0.0001; Extended Data Fig. 4). Given the components' relationship with body size, this suggests that body size and associated morphology evolves faster than the functional systems driving limb retraction.

Shifts in the rate of evolution can follow bouts of adaptive change¹⁸. The ancestors of oviraptorosaurs and paravians exhibited large increases in the evolutionary rate towards lower average forelimb gear ratios (median r = 18.2x and 50.4x background rate over periods of 57.9 and 7.33 million years, respectively, $pp > 99\%$; Fig. 2), consistent with our ANCOVA results. Increased mechanical leverage in shoulder retraction would have been beneficial in the evolution of flight among paravians, particularly during wing upstroke⁴⁰, as well as for prey apprehension among theropods generally⁴¹. These rate shifts also precede several physiological and neurological changes seen in pennaraptorans, including symmetrical vaned wing and

tail feathers, an increase in basal metabolic rate, and expansion of the cerebrum⁴². We also see a modest ~50% increase in the evolutionary rate of the forelimb gear ratios among tetanuran theropods (median $r = 1.47\times$ background rate, $pp > 50\%$). This suggests that theropods, as bipedal animals, had fewer constraints in their forelimbs to explore a wider ecomorphospace as they diversified (e.g., predation and flight). Although, the support for rate variation vanishes when we include a dummy-coded indicator variable for pennaraptorans into the variable-rates regression model ($BF = 0.98$), which suggests that they are the predominant source of rate variation.

The high rates of forelimb gear ratio evolution among pennaraptorans coincide with the accelerated evolutionary rates in the hindlimb gear ratios among maniraptorans (Fig. 3). We see a stepwise increase in the rate of hindlimb retractor evolution along the ancestral line to birds – first, an estimated 85% increase in the evolutionary rates among ornithomimosaur and their shared ancestral branches with maniraptorans (median $r = 1.85\times$ background rate, $pp > 70\%$) and then a more than 200% rate increase among maniraptorans (median $r = 3.3\times$ background rate, $pp > 90\%$). We also detect additional large rate increases among oviraptorids (median $r = 5.5\times$ background rate, $pp > 95\%$) and two paravians, *Microraptor gui* and *Archaeopteryx lithographica* (median $r = 5.5\times$ background rate, $pp > 95\%$). The accelerated evolutionary rates of the hindlimb retractor occurred as maniraptorans explored an expanded ecomorphospace (e.g., gliding/flight). This began with the release of the hindlimb musculature from the tail, creating two new locomotor modules^{6,33}. The modularity of the hindlimb and tail helped repurpose the tail musculature for flight⁶ and explains the independent evolution of gliding or flight potential in several paravian species⁷.

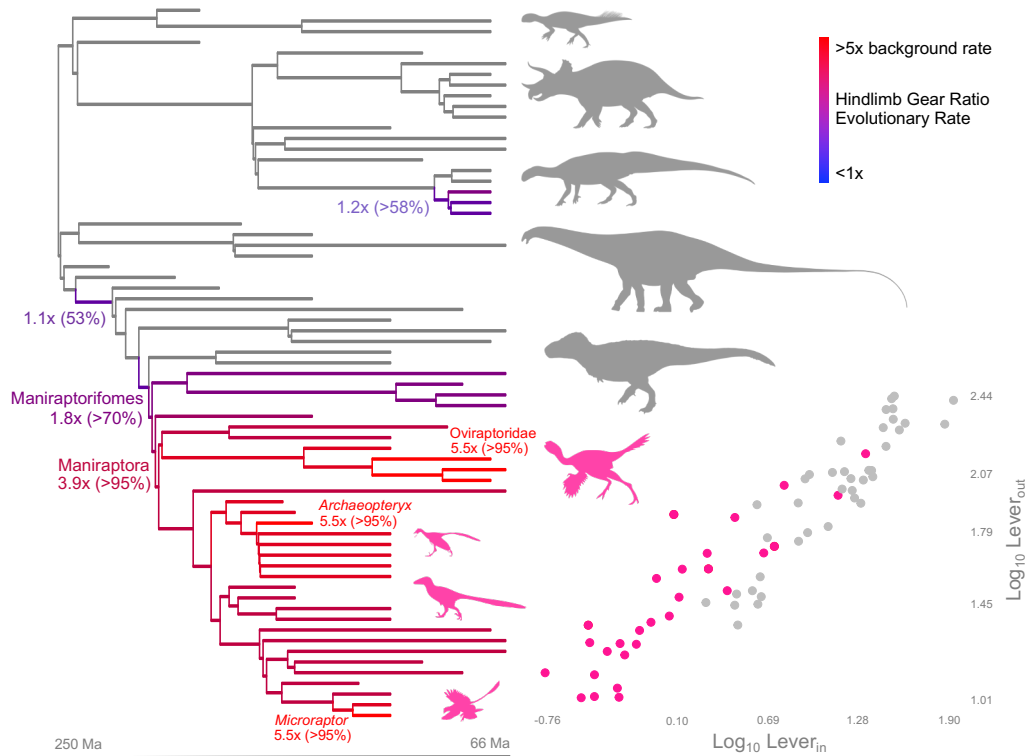


Fig. 3. Evolution of hindlimb function. Rates of gear ratio evolution for the primary retractor muscles of the hindlimbs mapped onto a dinosaur phylogeny. Warmer colours indicate increased rates of evolution relative to the background rate of the tree. We did not detect reductions in the rate of evolution. Numbers indicate the median estimated rate shift among the coloured branches, along with the associated percent posterior probability. Pink silhouettes represent maniraptoran species and correspond to data points in the scatter plot showing the correlation between the hindlimb lever_{out} and lever_{in}.

The horned ceratopsid dinosaurs exhibited a 64% reduction in the rate of forelimb retractor evolution (median $r = 0.64x$ background rate, $pp > 77\%$). This is consistent with increased constraints on forelimb locomotor evolution due to a transition to quadrupedalism and enlarging the head with more elaborate frills and horns⁴³. The forelimb lever_{in} (muscle moment arm of the shoulder) is the only gear

ratio component that shows substantial rate variation across the tree (BF = 2.64; Extended Data Fig. 5). In particular, we see a strong reduction in the evolutionary rate among iguanodontian ornithopods (median 0.12x background rate, pp > 70%). The evolution of the shoulder retractor slowed dramatically in concert with increased body sizes and facultative quadrupedalism in derived ornithopods, consistent with an overall conservation of postcranial anatomy⁴⁴. Interestingly, the sauropodomorph dinosaurs, boasting the largest body sizes of any terrestrial species, showed no evidence for changes in the evolutionary rate of their forelimb and hindlimb gear ratios. Our sauropodomorph dataset, although limited in sample size, spans their record from the Early Jurassic bipedal or facultatively quadrupedal ‘prosauropods’ (e.g., *Massospondylus*) to the Late Cretaceous gigantic titanosaurs (e.g., *Alamosaurus*). This suggests that the functional mechanics driving limb retraction evolved in step with gigantism and quadrupedalism in sauropodomorphs. This differs from the reduced rates of forelimb gear ratio and shoulder retractor evolution seen in ceratopsids and derived ornithopods and highlights multiple paths toward adapting to large-bodied herbivorous ecologies.

We then used the results from our variable-rates regression analyses as data to test for predictors of the rate of limb retractor evolution. For each species, we summed the median-estimated rates along each branch from the root to each tip (path rate) as a measure of the total amount of functional gear ratio evolution. Through a phylogenetic t-test, we found that quadrupedal species overall did not evolve slower on average (pMCMC = 0.18 and 0.39 for forelimbs and hindlimbs). In both limbs, rates of evolution were highly correlated with the number of nodes (or lineage divergence events) along each phylogenetic path (pMCMC < 0.0001). After accounting for variation in the amount of time to each species from the root (time root-to-tip path

length), we found that over 70% and 92% of the rate variation in the forelimb and hindlimb gear ratios, respectively, are attributable to speciation ($R^2 = 0.70$ and 0.92 , $pMCMC < 0.0001$). This suggests that locomotion plays a role in generating species through biogeographic dispersal or niche partitioning. For example, decoupling of the hindlimb and tail locomotor modules likely influenced the foot-driven prey apprehension ecology characteristic of maniraptoran dinosaurs during their diversification.

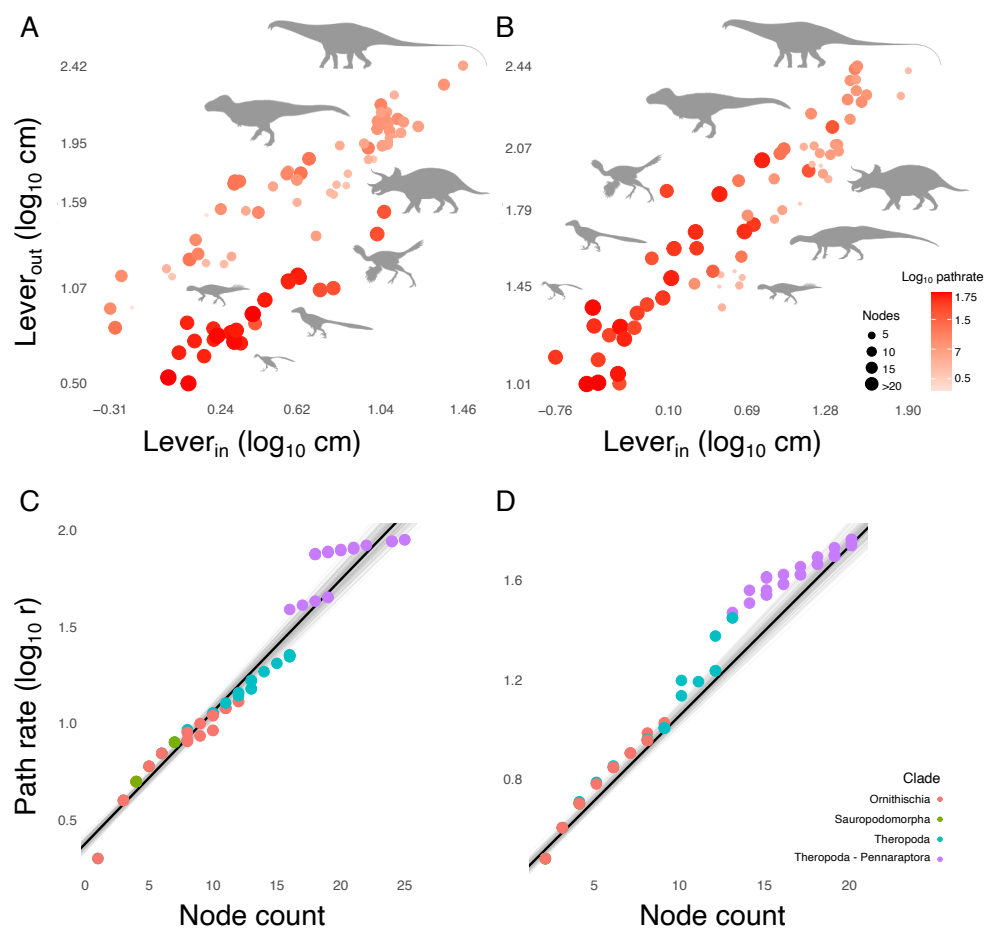


Fig. 4. Speciation drives rate of gear ratio evolution. Scatter plots relating the average rate of A) forelimb and B) hindlimb gear ratio evolution (root-to-tip sum of the median estimated rates) to net speciation (node counts along path lengths). Darker red and larger points represent species with higher average rates and higher net speciation, respectively. Scatter plots relating the root-to-tip sum of the median estimated evolutionary rates (r) of the C) forelimb and D) hindlimb gear ratio to net speciation. Data point colours correspond to

dinosaur clades. Grey regression lines are a random sample of 100 estimates from the Bayesian PGLS regression analysis with Pagel's λ estimated. The mean estimated regression lines are shown in black.

Last, we find evidence that evolutionary rates in the forelimb and hindlimb gear ratios were coupled. We reran the variable-rates regression model using a subset of taxa with gear ratio data for both limbs (N = 48 species; BF = 19.28 and 3.36, favouring variable-rates model for forelimbs and hindlimbs) and found that the median estimated branch-wise rates were correlated between the forelimbs and hindlimbs (ordinary least squares regression, adjusted $R^2 = 0.47$, p-value < 0.0001); the rate of hindlimb gear ratio evolution explains about 47% of the rate of forelimb gear ratio evolution. This increases substantially to 88% after removing the ancestral branches to oviraptorids and paravians (adjusted $R^2 = 0.88$, p-value < 0.0001; Extended Data Fig. 6), which saw large shifts in forelimb gear ratio evolution (Fig. 2). However, the reduced dataset only recovered evolutionary rate variation among maniraptorans (median $r = 2.09x$ and $2.52x$ for forelimb and hindlimb). All other branches remained unscaled (median $r = 1.0$), limiting our scope of inference to Maniraptora. Evolutionary coupling in other clades is ambiguous. Regardless, these results demonstrate that functional changes in forelimb and hindlimb retraction were coupled throughout the evolution of a major dinosaur group. As obligate bipeds, this suggests there were developmental constraints on locomotor limb evolution.

Our study highlights the complex ways functional systems evolve. Past comparative ecomorphological and biomechanical studies have revealed many insights into the diversification and evolution of extant^{1,2,12–14,45,46} and extinct^{4,5,8–10,15,32,47–49} species. However, through our system-centred approach paired with variable-rates phylogenetic modelling, we shed new light on the evolution of functional

systems. For example, past studies argued for gradual locomotor evolution and accumulation of characteristics along the avian-stem lineage^{4,29,32}. Evidence for gradual change in the individual gear ratio components ($\text{lever}_{\text{out}}$ and lever_{in}) are consistent with these findings. Yet, the gear ratios themselves show variable rates of evolution, revealing a complex process in which bouts of functional evolution emerge from gradual change in underlying morphology. Our variable-rates analyses detected instances of both increased and decreased rates of limb retractor evolution, coinciding with transitions to novel modes of locomotion, such as flight in paravians and large-bodied quadrupedalism in ceratopsids. Last, we find that the rates at which forelimb and hindlimb retraction evolve were coupled in maniraptoran dinosaurs and highly associated with lineage divergence, suggesting potential developmental constraints on limb retraction and highlighting the role of locomotion in speciation.

References

1. Wilson, A. M. *et al.* Biomechanics of predator–prey arms race in lion, zebra, cheetah and impala. *Nature* **554**, 183–188 (2018).
2. Hein, A. M., Hou, C. & Gillooly, J. F. Energetic and biomechanical constraints on animal migration distance. *Ecol. Lett.* **15**, 104–110 (2012).
3. Greenwood, P. J. Mating systems, philopatry and dispersal in birds and mammals. *Anim. Behav.* **28**, 1140–1162 (1980).
4. Allen, V., Bates, K. T., Li, Z. & Hutchinson, J. R. Linking the evolution of body shape and locomotor biomechanics in bird-line archosaurs. *Nature* **497**, 104–107 (2013).

5. Dickson, B. V., Clack, J. A., Smithson, T. R. & Pierce, S. E. Functional adaptive landscapes predict terrestrial capacity at the origin of limbs. *Nature* **589**, 242–245 (2021).
6. Gatesy, S. M. & Dial, K. P. Locomotor modules and the evolution of avian flight. *Evolution* **50**, 331–340 (1996).
7. Pei, R. *et al.* Potential for powered flight neared by most close avialan relatives, but few crossed its thresholds. *Curr. Biol.* **30**, 4033–4046.e8 (2020).
8. Sullivan, C., Hone, D. W. E., Xu, X. & Zhang, F. The asymmetry of the carpal joint and the evolution of wing folding in maniraptoran theropod dinosaurs. *Proc. R. Soc. B Biol. Sci.* rspb20092281 (2010) doi:10.1098/rspb.2009.2281.
9. Pierce, S. E. *et al.* Vertebral architecture in the earliest stem tetrapods. *Nature* **494**, 226–229 (2013).
10. Neenan, J. M., Ruta, M., Clack, J. A. & Rayfield, E. J. Feeding biomechanics in *Acanthostega* and across the fish–tetrapod transition. *Proc. R. Soc. Lond. B Biol. Sci.* **281**, 20132689 (2014).
11. Heers, A. M. & Dial, K. P. Wings versus legs in the avian bauplan: development and evolution of alternative locomotor strategies. *Evolution* **69**, 305–320 (2015).
12. Polly, P. D. Tiptoeing through the trophics: geographic variation in carnivoran locomotor ecomorphology in relation to environment. in *Carnivoran Evolution: New Views on Phylogeny, Form, and Function*. (eds. Goswami, A. & Friscia, A.) 374–410 (Cambridge University Press, 2010).
13. Polly, P. D. & Sarwar, S. Extinction, extirpation, and exotics: effects on the correlation between traits and environment at the continental level. *Ann. Zool. Fenn.* **51**, 209–226 (2014).

14. Anderson, P. S. L., Claverie, T. & Patek, S. N. Levers and linkages: mechanical trade-offs in a power-amplified system. *Evolution* **68**, 1919–1933 (2014).
15. Button, D. J., Rayfield, E. J. & Barrett, P. M. Cranial biomechanics underpins high sauropod diversity in resource-poor environments. *Proc. R. Soc. Lond. B Biol. Sci.* **281**, 20142114 (2014).
16. Morales-García, N. M., Gill, P. G., Janis, C. M. & Rayfield, E. J. Jaw shape and mechanical advantage are indicative of diet in Mesozoic mammals. *Commun. Biol.* **4**, 1–14 (2021).
17. Grossnickle, D. M. & Newham, E. Therian mammals experience an ecomorphological radiation during the Late Cretaceous and selective extinction at the K–Pg boundary. *Proc R Soc B* **283**, 20160256 (2016).
18. Baker, J., Meade, A., Pagel, M. & Venditti, C. Positive phenotypic selection inferred from phylogenies. *Biol. J. Linn. Soc.* **118**, 95–115 (2016).
19. Holzman, R. *et al.* Biomechanical trade-offs bias rates of evolution in the feeding apparatus of fishes. *Proc. R. Soc. B Biol. Sci.* rspb20111838 (2011)
doi:10.1098/rspb.2011.1838.
20. Eastman, J. M., Alfaro, M. E., Joyce, P., Hipp, A. L. & Harmon, L. J. A Novel comparative method for identifying shifts in the rate of character evolution on trees. *Evolution* **65**, 3578–3589 (2011).
21. Venditti, C., Meade, A. & Pagel, M. Multiple routes to mammalian diversity. *Nature* **479**, 393–396 (2011).
22. Thomas, G. H. & Freckleton, R. P. MOTMOT: models of trait macroevolution on trees. *Methods Ecol. Evol.* **3**, 145–151 (2012).
23. Revell, L. J. On the analysis of evolutionary change along single branches in a phylogeny. *Am. Nat.* **172**, 140–147 (2008).

24. Rabosky, D. L. *et al.* Rates of speciation and morphological evolution are correlated across the largest vertebrate radiation. *Nat. Commun.* **4**, (2013).
25. Puttick, M. N., Thomas, G. H. & Benton, M. J. High rates of evolution preceded the origin of birds. *Evolution* **68**, 1497–1510 (2014).
26. Lee, M. S. Y., Cau, A., Naish, D. & Dyke, G. J. Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds. *Science* **345**, 562–566 (2014).
27. Benson, R. B. J. *et al.* Rates of dinosaur body mass evolution indicate 170 million years of sustained ecological innovation on the avian stem lineage. *PLoS Biol* **12**, e1001853 (2014).
28. Benson, R. B. J. & Choiniere, J. N. Rates of dinosaur limb evolution provide evidence for exceptional radiation in Mesozoic birds. *Proc. R. Soc. B Biol. Sci.* **280**, 20131780 (2013).
29. Brusatte, S. L., Lloyd, G. T., Wang, S. C. & Norell, M. A. Gradual assembly of avian body plan culminated in rapid rates of evolution across the dinosaur-bird transition. *Curr. Biol.* **24**, 2386–2392 (2014).
30. Gregersen, C. S. & Carrier, D. R. Gear ratios at the limb joints of jumping dogs. *J. Biomech.* **37**, 1011–1018 (2004).
31. Short, R. A. & Lawing, A. M. Geography of artiodactyl locomotor morphology as an environmental predictor. *Divers. Distrib.* **27**, 1818–1831 (2021).
32. Allen, V. R., Kilbourne, B. M. & Hutchinson, J. R. The evolution of pelvic limb muscle moment arms in bird-line archosaurs. *Sci. Adv.* **7**, eabe2778 (2021).
33. Rhodes, M. M., Henderson, D. M. & Currie, P. J. Maniraptoran pelvic musculature highlights evolutionary patterns in theropod locomotion on the line to birds. *PeerJ* **9**, e10855 (2021).

34. 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 Berlin Heidelberg, 2014). doi:10.1007/978-3-662-43550-2_5.
35. Jasinowski, S. C., Russell, A. P. & Currie, P. J. An integrative phylogenetic and extrapolatory approach to the reconstruction of dromaeosaur (Theropoda: Eumaniraptora) shoulder musculature. *Zool. J. Linn. Soc.* **146**, 301–344 (2006).
36. Holzman, R. *et al.* Biomechanical trade-offs bias rates of evolution in the feeding apparatus of fishes. *Proc. R. Soc. B Biol. Sci.* rspb20111838 (2011)
doi:10.1098/rspb.2011.1838.
37. Obbard, D. J., Jiggins, F. M., Halligan, D. L. & Little, T. J. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr. Biol.* **16**, 580–585 (2006).
38. Baker, J. & Venditti, C. Rapid change in mammalian eye shape is explained by activity pattern. *Curr. Biol.* **29**, 1082–1088.e3 (2019).
39. Kane, E. A. & Higham, T. E. Complex systems are more than the sum of their parts: using integration to understand performance, biomechanics, and diversity. *Integr. Comp. Biol.* **55**, 146–165 (2015).
40. Dial, K. P. Avian forelimb muscles and nonsteady flight: can birds fly without using the muscles in their wings? *The Auk* **109**, 874–885 (1992).
41. Burch, S. H. Complete forelimb myology of the basal theropod dinosaur *Tawa hallae* based on a novel robust muscle reconstruction method. *J. Anat.* **225**, 271–297 (2014).
42. Xu, X. *et al.* An integrative approach to understanding bird origins. *Science* **346**, 1253293 (2014).

43. Dodson, P., Forster, C. & Sampson, S. D. Ceratopsidae. in *The Dinosauria* (eds. Weishampel, D. B., Dodson, P. & Osmólska, H.) 494–513 (University of California Press, 2004).
44. Horner, J. R., Weishampel, D. B. & Forster, C. A. Hadrosauridae. in *The Dinosauria* (eds. Weishampel, D. B., Dodson, P. & Osmólska, H.) 438–463 (University of California Press, 2004).
45. Alexander, R. M. *Animal Mechanics*. (Blackwell Scientific Publication, 1983).
46. Hutchinson, J. R. Biomechanical modeling and sensitivity analysis of bipedal running ability. I. Extant taxa. *J. Morphol.* **262**, 421–440 (2004).
47. Holtz, T. R. The arctometatarsalian pes, an unusual structure of the metatarsus of Cretaceous Theropoda (Dinosauria: Saurischia). *J. Vertebr. Paleontol.* **14**, 480–519 (1995).
48. Dececchi, T. A., Mloszewski, A. M., Jr, T. R. H., Habib, M. B. & Larsson, H. C. E. The fast and the frugal: divergent locomotory strategies drive limb lengthening in theropod dinosaurs. *PLOS ONE* **15**, e0223698 (2020).
49. Hutchinson, J. R. Biomechanical modeling and sensitivity analysis of bipedal running ability. II. Extinct taxa. *J. Morphol.* **262**, 441–461 (2004).
50. Sakamoto, M., Benton, M. J. & Venditti, C. Strong support for a heterogeneous speciation decline model in Dinosauria: a response to claims made by Bonsor et al. (2020). *R. Soc. Open Sci.* **8**, 202143 (2021).
51. Lloyd, G. T., Bapst, D. W., Friedman, M. & Davis, K. E. Probabilistic divergence time estimation without branch lengths: dating the origins of dinosaurs, avian flight and crown birds. *Biol. Lett.* **12**, 20160609 (2016).

52. Wang, M., O'Connor, J. K., Xu, X. & Zhou, Z. A new Jurassic scansoriopterygid and the loss of membranous wings in theropod dinosaurs. *Nature* **569**, 256–259 (2019).
53. Lü, J. & Brusatte, S. L. A large, short-armed, winged dromaeosaurid (Dinosauria: Theropoda) from the Early Cretaceous of China and its implications for feather evolution. *Sci. Rep.* **5**, 11775 (2015).
54. Gatesy, S. M. Caudofemoral musculature and the evolution of theropod locomotion. *Paleobiology* **16**, 170–186 (1990).
55. Romer, A. S. Crocodilian pelvic muscles and their avian and reptilian homologues. *Bull. Am. Mus. Nat. Hist.* **48**, 533–552 (1923).
56. Romer, A. S. The pelvic musculature of saurischian dinosaurs. *Bull. Am. Mus. Nat. Hist.* **48**, 605–617 (1923).
57. Romer, A. S. The pelvic musculature of ornithischian dinosaurs. *Acta Zool.* **8**, 225–275 (1927).
58. Gatesy, S. M., Bäker, M. & Hutchinson, J. R. Constraint-based exclusion of limb poses for reconstructing theropod dinosaur locomotion. *J. Vertebr. Paleontol.* **29**, 535–544 (2009).
59. Hutchinson, J. R. The evolution of pelvic osteology and soft tissues on the line to extant birds (Neornithines). *Zool. J. Linnean Soc.* **131**, 123–168 (2001).
60. Hutchinson, J. R. The evolution of femoral osteology and soft tissues on the line to extant birds (Neornithines). *Zool. J. Linnean Soc.* **131**, 169–197 (2001).
61. Maidment, S. C. R. & Barrett, P. M. The locomotor musculature of basal ornithischian dinosaurs. *J. Vertebr. Paleontol.* **31**, 1265–1291 (2011).
62. Otero, A. Forelimb musculature and osteological correlates in Sauropodomorpha (Dinosauria, Saurischia). *PLOS ONE* **13**, e0198988 (2018).

63. Smith, D. K. Forelimb musculature and function in the therizinosaur *Nothronychus* (Maniraptora, Theropoda). *J. Anat.* **239**, 307–335 (2021).
64. Burch, S. H. Myology of the forelimb of *Majungasaurus crenatissimus* (Theropoda, Abelisauridae) and the morphological consequences of extreme limb reduction. *J. Anat.* **231**, 515–531 (2017).
65. Senter, P. & Robins, J. H. Resting orientations of dinosaur scapulae and forelimbs: a numerical analysis, with implications for reconstructions and museum mounts. *PLoS ONE* **10**, e0144036 (2015).
66. 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).
67. 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).
68. Rambaut, A., Drummond, A. J., Xie, D., Baele, G. & Suchard, M. A. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. *Syst. Biol.* **67**, 901–904 (2018).
69. Xie, W., Lewis, P. O., Fan, Y., Kuo, L. & Chen, M.-H. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Syst. Biol.* **60**, 150–160 (2011).
70. Carrano, M. T. Locomotion in non-avian dinosaurs: integrating data from hindlimb kinematics, in vivo strains, and bone morphology. *Paleobiology* **24**, 450–469 (1998).
71. Venditti, C., Meade, A. & Pagel, M. Multiple routes to mammalian diversity. *Nature* **479**, 393–396 (2011).

72. Green, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732 (1995).
73. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**, 119–121 (2006).
74. Revell, L. J. phytools: an R package for phylogenetic comparative biology (and other things). *Methods Ecol. Evol.* **3**, 217–223 (2012).

Methods

No statistical methods were used to predetermine sample size. Additional details on data collection and statistical analyses can be found in Appendix 2 – Supplementary Materials. An abbreviated data table can be found in Appendix 3.

Phylogenies. We acquired a set of 100 time-scaled meta-trees consisting of 961 taxa from Sakamoto et al.⁵⁰, originally sourced from a recent study by Lloyd et al.⁵¹. These trees were trimmed down to the taxa in our dataset (forelimb, N = 85; hindlimb, N = 67). The following taxa in our dataset were not originally included in the meta-trees but were added based on the age and phylogenetic position of close relatives: *Amblopteryx longibrachium*⁵² (replaced *Epidendrosaurus ningchengensis*), *Anzu wyliei* (replaced *Elmisaurus rarus*), *Camarasaurus lentus* (replaced *Camarasaurus grandis*), *Galeamopus pabsti* (replaced *Barosaurus lentus*), *Meraxes gigas* (replaced *Carcharodontosaurus saharicus*), *Nemegtomykus citus* (replaced *Shuvuuia deserti*), *Oksoko avarsan* (replaced *Banji long*), *Serikornis sungei* (replaced *Eosinopteryx brevipenna*), *Stegouros elengassen* (replaced *Antarctopelta oliveroi*), *Xingxiulong chengi* (replaced *Jingshanosaurus xinwaensis*), *Zhenyuanlong suni*⁵³ (replaced *Tianyuraptor ostromi*), and *Zhongjianosaurus yangi* (replaced *Graciliraptor*

lujiatunensis). To account for our uncertainty in the node ages and branch lengths, we sampled across the entire set of 100 meta-trees in the analyses described below.

Limb lever arm data. To measure locomotor limb function, we estimated the lever gear ratios of the primary humeral and femoral retractor muscles. The gear ratio is equal to the distance between the point of axial movement (fulcrum, e.g., shoulder joint) and the loading point (e.g., manus) divided by the perpendicular distance from the fulcrum to the line of muscle action. The distance in the numerator is called the lever_{out}, or load arm, which in our case is the effective postural limb length. The denominator is called the lever_{in}, or muscle moment arm. Gear ratios are a simple but useful metric for characterizing basic locomotor function^{30,31}. Although we do not expect gear ratios for individual muscles to fully characterize locomotor performance, the simplicity of the metric enables us to quantify the evolution of a single functional system across numerous species. Future studies can extend our analytical approach to jointly sample several muscle groups, paired with biomechanical and physical models, to characterize locomotor function in more detail.

The primary femoral retractor of archosaurs is the *m. caudofemoralis longus* (*CFL*). This muscle originates along the ventral aspect of the caudal transverse processes and the lateral surfaces of the caudal centra and extends ventrally to insert onto the 4th trochanter of the femur^{54–57}. We took the following skeletal measurements to calculate the muscle moment arm for the *CFL*: 1) the length from the femoral head to the 4th trochanter on the femur and 2) the length from the acetabulum to the last preserved transverse process-bearing caudal vertebra. If specimens were fully articulated, we took the straight-line distance from the acetabulum to the last preserved transverse process. If specimens were disarticulated, we measured the

distance from the acetabulum to the posterior-most extent of the ilium, and then added the lengths of each individual vertebral centrum to the last caudal vertebra bearing a transverse process. For species where a 4th trochanter measurement was not available or inaccessible, we predicted the size of the *CFL* for species based on a its correlation with femur length, accounting for phylogenetic relationships (see details in ‘Retrodicting muscle attachments’ section below). Non-avian dinosaurs were digitigrade, making the end of the metatarsus the loading point. Therefore, we took the following skeletal measurements to calculating the lever_{out}: 1) the length of femur from the femoral head to its distal articular condyles; 2) the length of the tibia from proximal to distal articular ends; 3) the length of astragalus/calcaneum, when separate from the tibia, but incorporated into tibia length when fused to the tibia in theropods; and 4) the length of metatarsal III when preserved, or replaced by the next longest preserved metatarsal when metatarsal III was not preserved (Extended Data Fig. 1). Each of these elements were measured individually regardless of articulation. We assumed a mid-stance pose for our joint angles to calculate the moment arms from the limb bone measurements. For the mid-stance pose of all dinosaur species (except sauropodomorphs), we used the following joint angles based on the ‘extreme pose 1’ for *Tyrannosaurus rex*⁵⁸, which was the most flexed pose modelled by Gatesy et al.—hip angle: 50° from horizontal; knee angle: 108°; ankle angle: 132°. This flexed pose also falls within the range of mid-stance poses for the Ostrich (*Struthio*)⁵⁸. For derived sauropodomorph species, we assumed a more graviportal pose—hip angle: 85° from horizontal; knee angle: 175°; ankle angle: 175°.

Birds and their close non-avian maniraptoran relatives underwent an important biomechanical shift that functionally decoupled the hindlimb and tail⁵⁴. This change in musculature also accompanied a shift toward knee-driven locomotion, in which limb

displacement is mostly accomplished by movement about the knee^{32,54}. To model the macroevolutionary consequences of this functional transition, we measured the lever_{in} of the *m. iliotibialis lateralis pars postacetabularis* (*ITL*), which serves as both a hip and knee extensor in birds^{59,60}. The *ITL* originates on the dorsal and posterior margins of the ilium and inserts on the proximolateral surface of the femur and/or the tibial cnemial crest. In extinct birds and maniraptoriforms without obvious *CFL* scars on the femur, we assumed that locomotion was dominated by knee extension and used the *ITL* lever_{in}. Please refer to the abbreviated data table in Appendix 3 or the full data table for notes on which species were assigned a *CFL* vs. *ITL* lever_{in}.

The *m. deltoideus scapularis* (*DS*) is one of the primary humeral retractor muscles in archosaurs^{35,61,41}. In birds and crocodilians, the *DS* originates along the posterior lateral face of the scapula and inserts on the dorsolateral aspect of the deltopectoral crest of the humerus. We measured the following elements to calculate the lever_{in} of the forelimb: 1) the entire length of the scapula (and then take three-quarters this length as the midpoint of the origin); and 2) the length from the head of the humerus to the end of the deltopectoral crest (Extended Data Fig. 1). The insertion of the *DS* is known to extend distally along the line to birds from a more proximal position, in a depression closer to the humeral head, as seen in extant archosaurs, to the entire margin of the deltopectoral crest. We, therefore, approximated the *DS* insertion site using a percentage of the deltopectoral crest length, measured from the head of the humerus. ImageJ measurements of humeri found that the *DS* insertion site from the head of the humerus was about 40% of the deltopectoral crest length in ornithischians⁶¹, 40% in sauropodomorphs⁶², and 30% in non-paravian theropods^{41,63} (except *Majungasaurus*, which showed about 50%⁶⁴). Muscle reconstructions of dromaeosaurs, troodontids, and oviraptorids showed a broader *DS* insertion on the

dorsal surface of the deltopectoral crest, as in modern birds (see Figure 11C in ³⁵). Therefore, for paravian dinosaurs, we approximated the *DS* insertion site by taking the entire length of the deltopectoral crest from the head of the humerus.

For the *DS* lever_{out}, we measured the following: 1) the length of humerus from the head to its distal articular condyles; 2) the length of the radius from proximal to distal articular ends; and 3) the length of metacarpal III when preserved or replaced by the next longest preserved metacarpal when metacarpal III was not preserved. As with the hindlimb analyses, we calculated the gear ratios assuming a mid-stance pose. We used the following recommended joint angles for the average theropod without a semilunate carpal—shoulder angle: 54°; elbow angle: 106°; wrist angle: 158°⁶⁵. For ornithischians and basal sauropodomorphs, we used a shoulder angle of 88° and the same elbow and wrist angles as theropods. For graviportal species, we used 88° (shoulder), 175° (elbow), and 175° (wrist). A semilunate carpal, in which the manus is oriented posteriorly, is a distinctive feature of pennaraptoran theropods. We, therefore, used the following angles for pennaraptoran species—shoulder angle: 54° from horizontal; elbow angle: 46°; wrist angle: 99°⁶⁵.

We collected the data detailed above directly from museum specimens and cast mounts based on real fossil material and supplemented these data with measurements from the literature and photographs whenever scalebars were provided. Please refer to our full data table for extensive notes on specimen measurements.

Retrodicting muscle attachments. To calculate the *CFL* and *DS* lever_{in}, we respectively require the lengths from the femoral head to 4th trochanter and humeral head to the deltopectoral crest. However, we lack these data from many species in

our dataset either due to the absence of measurements in the literature or a genuine osteological absence. To estimate the missing measurements, we conducted a retrodiction (imputation) analysis^{66,67} using a phylogenetic generalized least-squares (PGLS) regression model in BayesTraits V4 (<http://www.evolution.reading.ac.uk/BayesTraitsV4.0.0/BayesTraitsV4.0.0.html>). The retrodiction protocol involves an initial regression analysis that uses a Markov-chain Monte Carlo (MCMC) algorithm to estimate model parameters while ignoring the species with missing data. That model is saved and then used to estimate the missing values of those species based on the phylogeny and statistical relationship between two traits. We further tested if there was phylogenetic signal by comparing a model that estimates Pagel's λ with one that assumes negligible phylogenetic signal ($\lambda = 0.000001$). All regression analyses ran for 12,500,000 iterations, sampling every 1,000 iterations and discarding the first 25,000,000 as burn-in. We assessed the convergence of the MCMC chains using trace plots of the estimated model parameters in the program Tracer v1.7.2⁶⁸. For model comparisons, we estimated the log marginal likelihoods with a stepping-stones algorithm⁶⁹, using 100 stones and sampling every 10,000 iterations, and calculated Bayes factors (BF). A BF > 2 indicates positive support for the model with the higher log marginal likelihood.

We further regressed the length of the deltopectoral crest on humerus length (\log_{10} centimetres) using a dataset of 75 species. The initial regression analysis found compelling support for including Pagel's λ (phylogenetic signal) in the model (mean $\hat{\lambda} = 0.92$, BF = 25.35). Humerus length strongly explains deltopectoral crest length in dinosaurs (mean $R^2 = 0.92$, pMCMC < 0.0001) (Extended Data Fig. 7A). We used this model to retrodict the deltopectoral lengths of 12 dinosaur species. To predict the distance to the 4th trochanter, we regressed 4th trochanter distance on femur length

(log₁₀ centimetres) using an expanded dataset of 117 species from the literature⁷⁰. The initial analysis strongly supports the inclusion of λ in the model (mean $\hat{\lambda} = 0.88$, BF = 72.39) and a correlation between 4th trochanter distance and femur length (mean $R^2 = 0.96$, pMCMC < 0.0001) (Extended Data Fig. 7B). We used this model to retrodict the 4th trochanter distances of 17 dinosaurs.

To account for the variation in our retrodicted estimates, we sampled the distribution of gear ratio values using a random sample of 100 estimates of the retrodicted deltopectoral crest lengths and 4th trochanter distances throughout the analyses detailed below. We log₁₀-transformed the data prior to analysis and assessed regression model assumptions of normality and equal variance using the phylogenetically corrected residuals.

Phylogenetic multiple regression. To characterize the evolution of the forelimb and hindlimb gear ratio, we first conducted PGLS regressions on its two components, lever_{out} and lever_{in}, in BayesTraits V4. We tested if the gear ratio varied among dinosaur clades by including ‘dummy-coded’ indicator variables (binary: 0 or 1) as additional explanatory variables. We first tested if gear ratios varied among the three major dinosaur clades (Ornithischia, Sauropodomorpha, Theropoda) using Ornithischia as a baseline (all indicator variables = 0). We also included interaction variables to test if the gear ratio components scaled more steeply or gently among those clades, amounting to a phylogenetic ANCOVA. We further tested if the gear ratio components scaled differently in each clade individually, in more specific theropod subclades, Maniraptora and Avialae, and between obligate bipeds and facultative and obligate quadrupeds. In addition, we tested if there was phylogenetic signal in the variation of the gear ratio components by comparing a model that estimates Pagel’s λ

with one that assumes negligible phylogenetic signal ($\lambda = 0.000001$). We used Bayes factors to compare the log marginal likelihoods of the competing models. Lastly, we reran the best-fitting model using a distribution of gear ratio estimates based on a random sample of 100 estimates of the retrodicted deltopectoral crest lengths and 4th trochanter distances. We log₁₀-transformed the data prior to analysis and assessed regression model assumptions. We used the same MCMC settings as the retrodiction analyses and assessed the convergence of the MCMC chains using trace plots in the program Tracer⁶⁸.

Modelling functional evolution. We modelled the evolution of the two gear ratio components using a recently developed phylogenetic independent contrasts regression model that allows for variable rates of evolution^{18,71}. This approach uses a Bayesian reversible jump MCMC algorithm⁷² to propose shifts in the rate of evolution across a phylogeny under a regression model framework¹⁸. The rates of evolution are inferred based on phylogenetic shifts in the residual variance. The model can propose two types of rate scalars: 1) branch-specific scalars, through which the rate of evolution shifts along a single branch on the tree, representing a mean or intercept shift; 2) clade-wide scalars, through which the rate of evolution shifts across an entire clade, representing a shift in trait variance. The method produces a posterior sample of trees in which branch lengths are scaled to represent the amount of trait evolution. Rate shifts relative to the background rate can be identified in reference to the original time-calibrated tree without prior specification as to their location or magnitude within the phylogeny. The reversible jump MCMC, additionally, reduces the number of model parameters to those only supported by the data. Additionally, we estimated the rates of evolution in the two components of the gear ratio ($\text{lever}_{\text{out}}$ and lever_{in}) individually to

identify how the underlying parameters of the gear ratio evolve at different rates throughout the tree. We ran an additional set of variable-rates analyses on the gear ratio components where they were standardized to each other and between the forelimbs and hindlimbs. This ensured that all parameters were on the same scale, allowing us to compare their overall background rates (background σ^2 ; Extended Data Fig. 4). As with the multiple regression analyses, we \log_{10} -transformed the data prior to analysis and conducted the analyses while sampling the entire set of 100 meta-trees and distribution of gear ratio estimates resulting from the retrodiction analyses.

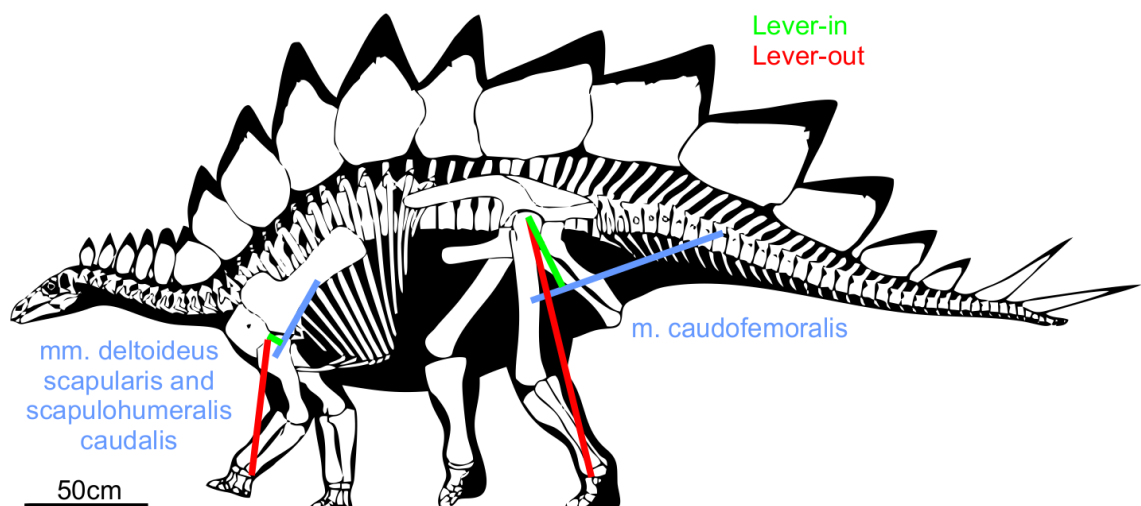
To further assess the effect of speciation on gear ratio evolution, we also ran regression analyses on the root-to-tip sum of the median estimated rate scalars from the variable-rates regression analyses. We used dummy-coded indicator variables to test if the total rates of gear ratio evolution differed between gaits (0 = facultative and obligate quadruped, 1 = obligate biped), amounting to a phylogenetic t-test. We also tested if the net number of speciation events (node count) along the root-to-tip path length explains the variation in the total rates gear ratio evolution. This approach derives from the path length and node count test, originally developed by⁷³. We extracted phylogenetic path lengths and node counts using the R package *fallpaddy*, a package for testing punctuated trait evolution (<https://github.com/suryakevin/fallpaddy>) and obtained median estimated rate scalars using the variable-rates analysis post-processor (<http://www.evolution.reading.ac.uk/VarRatesWebPP/>). A statistically significant effect of node count on the total rate of evolution suggests that speciation explains variation in the rate of evolution. We used the same regression settings in BayesTraits V4 as the Bayesian PGLS regression analyses described previously. Regression diagnostics indicate minimal modelling violations. The node-density artifact, an

underestimation of branch lengths in parts of the tree with fewer taxa, is absent ($\Delta < 1$). Lastly, we tested for an evolutionary coupling between the rates of the forelimb and hindlimb gear ratios by regressing the median estimated branch-wise rates of evolution from the variable-rates regression analyses. In this analysis, to ensure we were comparing the same branches on the tree, we reran the variable-rates regression analyses on a smaller dataset in which species had both forelimb and hindlimb gear ratio data ($N = 47$).

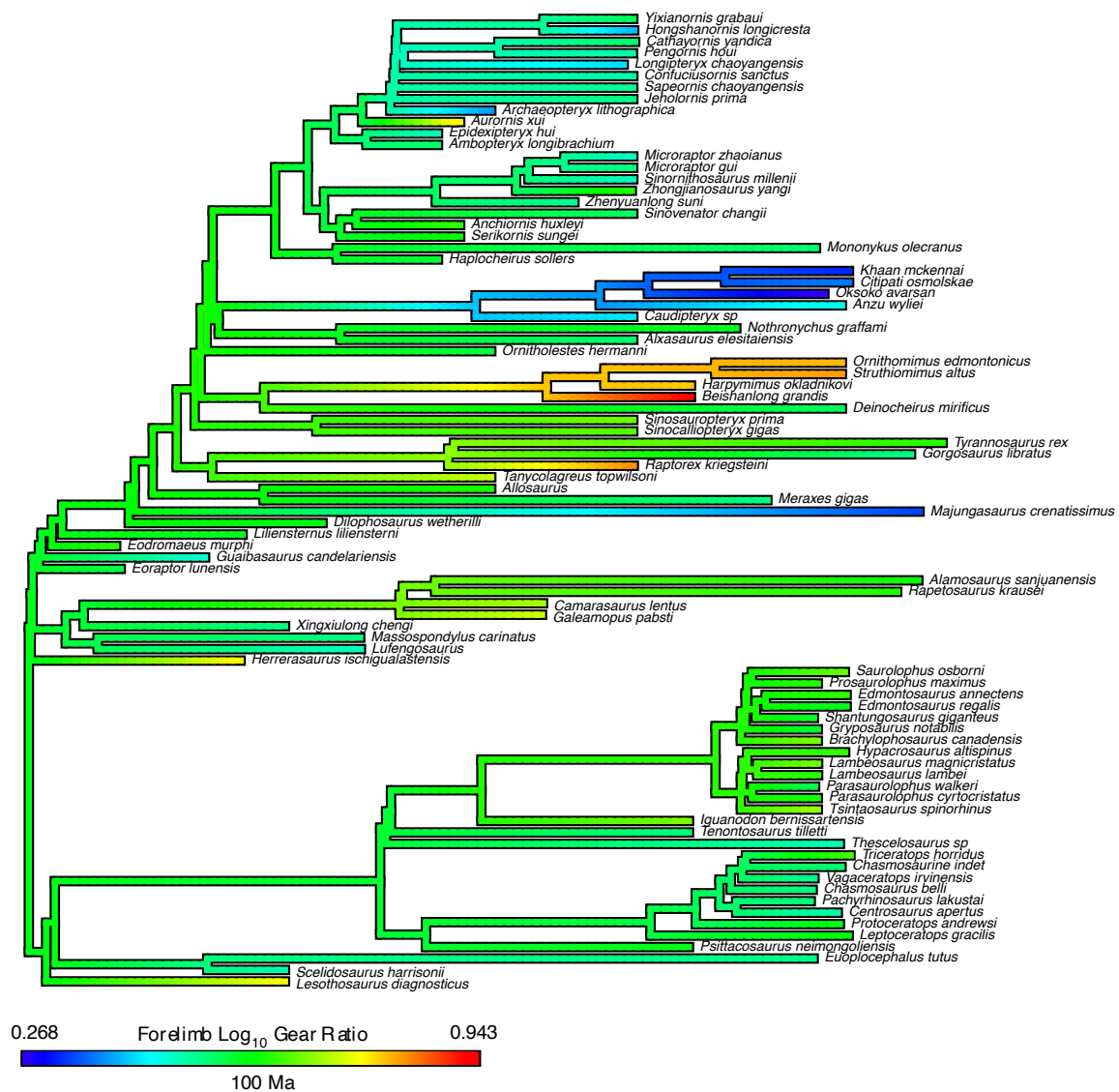
Appendix 1

Extended Data and Figures

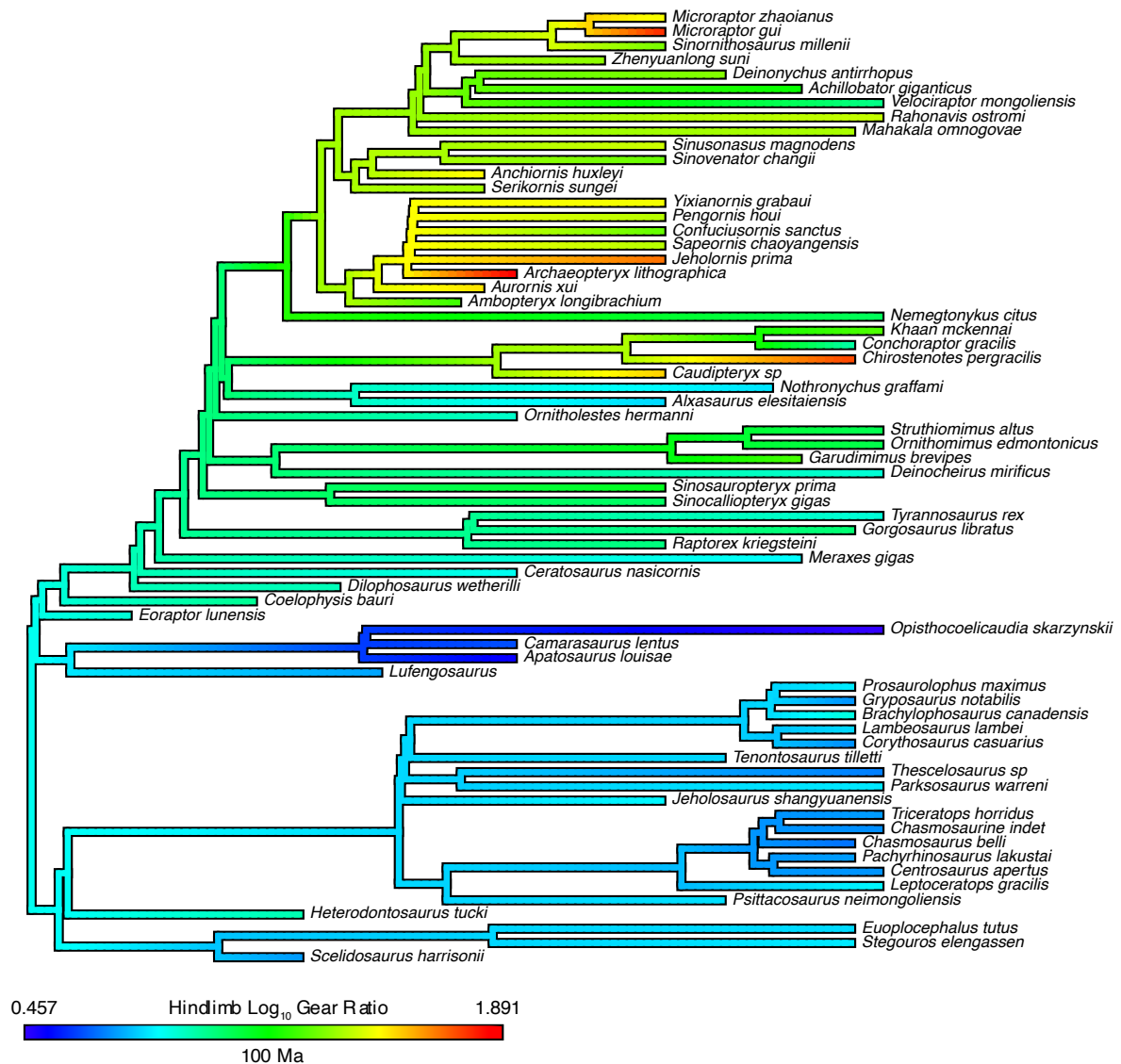
Extended Data Fig. 1. Mechanical gear ratio measurements. Diagrams of gear ratio measurements for the *m. deltoideus scapularis* and *m. caudofemoralis longus*. Red lines show the lever_{out}, green shows the lever_{in}, and blue shows the lines of action for each muscle. The skeletal of *Stegosaurus stenops* is copy written (2013) and reproduced with permission by Scott Hartman. Descriptions of the gear ratio measurements are provided in the Methods and Appendix 2.



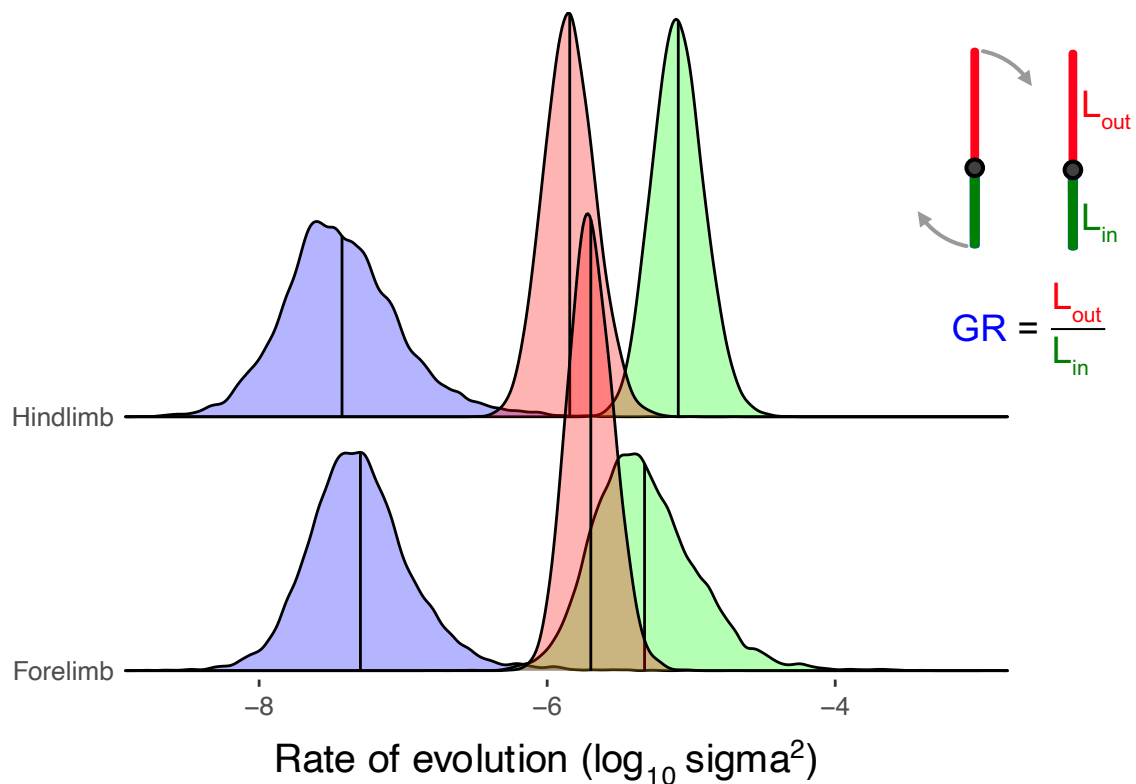
Extended Data Fig. 2. Ancestral states of the forelimb gear ratio. A) Phylogenetic trace map of ancestral states for the \log_{10} -scaled forelimb gear ratio, estimated using maximum likelihood through the fastanc function in the R package *phytools*⁷⁴. Colours were mapped onto a time-scaled phylogeny using the contMap function in *phytools*. Warmer colours indicate higher gear ratios and cooler colours indicate lower gear ratios. The length of the scale represents 100 million years.



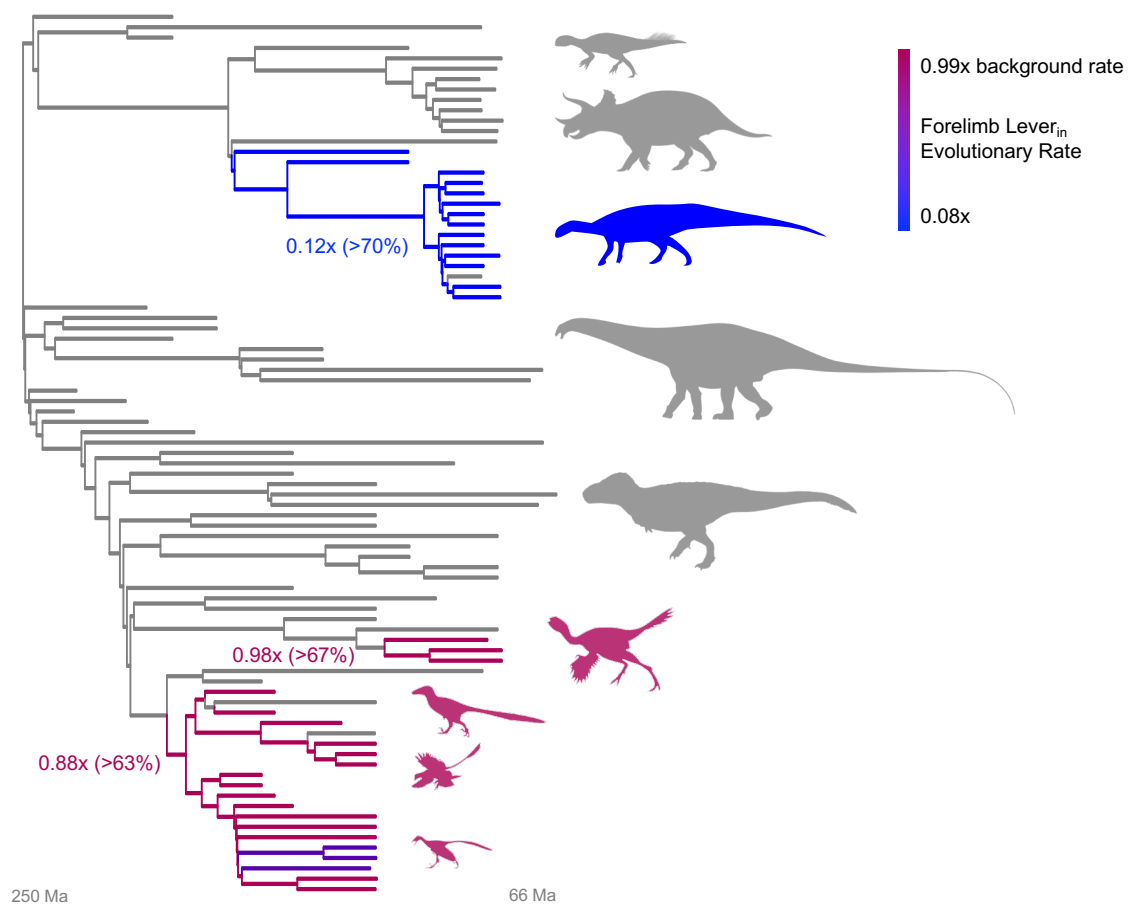
Extended Data Fig. 3. Ancestral states of the hindlimb gear ratio. A) Phylogenetic trace map of ancestral states for the \log_{10} -scaled hindlimb gear ratio, estimated using maximum likelihood through the `fastanc` function in the R package *phytools*⁷⁴. Colours were mapped onto a time-scaled phylogeny using the `contMap` function in *phytools*. Warmer colours indicate higher gear ratios and cooler colours indicate lower gear ratios. The length of the scale represents 100 million years.



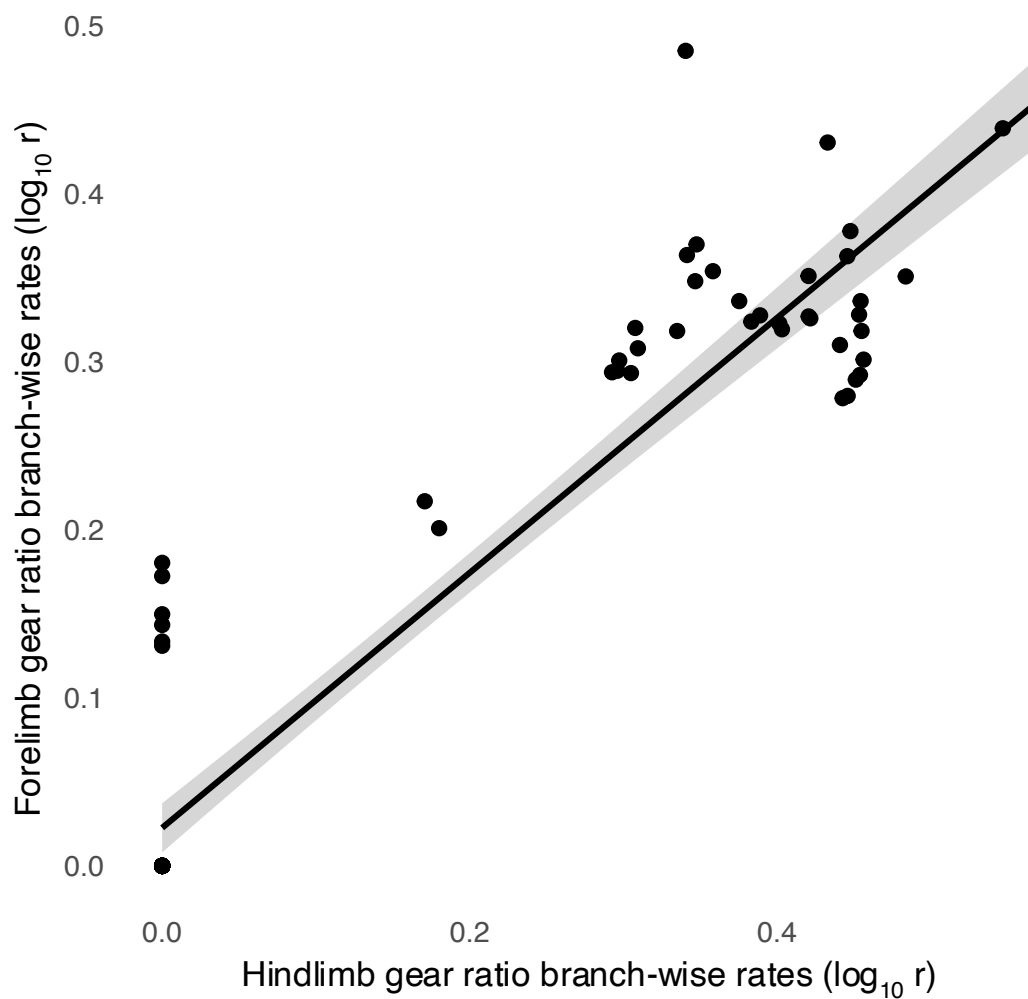
Extended Data Fig. 4. Evolutionary rates of functional parameters. Overlapping density plots comparing the posterior distributions of estimated \log_{10} -scaled background rates (σ^2) between forelimb and hindlimb gear ratios (GR) and functional parameters, lever_{out} (L_{out}) and lever_{in} (L_{in}), accounting for shifts in rate when variable-rates model was supported (BF > 2.0). Hindlimb and forelimb Lever_{out}, Lever_{in}, and gear ratios were standardized with each other to ensure the data varied on the same scale. Two-sample t-tests (forelimb vs hindlimb), ANOVA (GR vs L_{in} vs L_{out}), and Tukey-Kramer post-hoc tests all yield p-values < 0.0001) after accounting for multiple hypothesis testing.



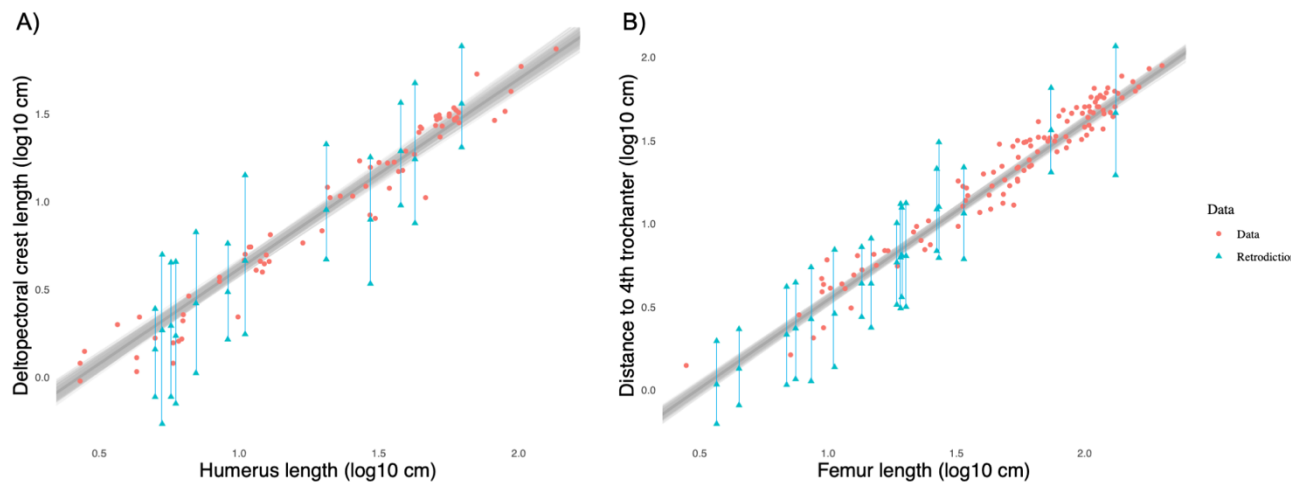
Extended Data Fig. 5. Evolution of forelimb lever_{in}. Rates of lever_{in} evolution for the primary forelimb retractor muscles mapped onto a time-scaled phylogeny. The rate of forelimb lever_{in} change only reduces throughout dinosaur evolution. Cooler colours indicate more reduced rates of evolution. Gray branches are unscaled ($r = 1$). Silhouette colours correspond to lineages with reduced rates. The blue silhouette represents ornithomimid species, and magenta represents pennaraptoran theropods. All silhouettes were obtained from phylopic.org and referenced in Figs. 1-2.



Extended Data Fig. 6. Branch-wise rate correlation. Scatter plot showing the correlation of the median branch-wise rates of evolution ($\log_{10} r$) between the forelimb and hindlimb gear ratios after removing the highly scaled branches to oviraptorosaurs and paravians. Median branch-wise rates were obtained from the variable-rates analysis on the reduced dataset of species with both forelimb and hindlimb gear ratio data ($n = 47$).



Extended Data Fig. 7. Retrodiction of muscle insertion sites. Scatter plots showing the retrodicted A) \log_{10} deltopectoral crest lengths and B) \log_{10} 4th trochanter distances. Original data represented by red circles and retrodicted values by blue triangles, with fading bars representing the 95% credible intervals of retrodictions. Grey regression lines are a random sample of 100 estimates from the Bayesian PGLS analysis with Pagel's λ . The mean estimated regression line is shown in black.



This equals the sum of the humerus, ulna, and metacarpus lengths when assuming a straight limb (all 180 degrees).

$$\text{Lever}_{\text{out}} = \text{SQRT}((\text{Shoulder to wrist}^2) + (\text{MC}^2) - (2 * \text{Shoulder to wrist} * \text{MC} * \text{COS}(\text{RADIANS}(\text{New Angle}))))$$

Forelimb lever_{in} (*m. deltoideus scapularis*)

LEVER-IN (DS)		4.99995							
Straight arm		54.4625932	90	90					
			7						
		8.602296205							
				35.537407					
		4.068640415							
		*Bisect DLP length and assume right angle							
		*Treat scapula length as hypotenuse and calculate length to bisected point (Lever-in)							
	DPC length	Half scap length	Shoulder Angle	DS	Angle B	Lever-in	Angle A	Angle Check	
Test	7	4.99995	90	8.60229621	54.462593	4.06864041	35.537407	180	
T. rex	15	97.725	90	98.8694878	8.7263356	14.8263638	81.273664	180	

LEVER-IN (DS)		4.99995							
Postured arm		81.11350511	54						
			7						
		5.731923219							
				44.88649489					
		4.939932125	4.939932125	4.93993212					
		*Bisect DLP length and assume right angle							
		*Treat scapula length as hypotenuse and calculate length to bisected point (Lever-in)							
	DPC length	Half scap length	Shoulder Angle	DS	Angle B	Lever-in	Angle A	Angle Check	
Test	7	4.99995	54	5.73192322	81.113505	4.93993212	44.886495	180	
T. rex	15	97.725	54	89.7325817	7.7723825	13.2161336	118.22762	180	

The *m. deltoideus scapularis* originates about 1/2 to 2/3 along the scapular blade and inserts along the dorsal aspect of the deltopectoral crest (DPC). Use distance to origin (Half scap length), shoulder angle, and DPC length to calculate the *m. deltoideus* line of action (DS).

$$\text{DS} = \text{SQRT}((\text{DPC}^2) + (\text{Half scap length}^2) - (2 * \text{DPC} * \text{Half scap length} * \text{COS}(\text{RADIANS}(\text{Shoulder Angle}))))$$

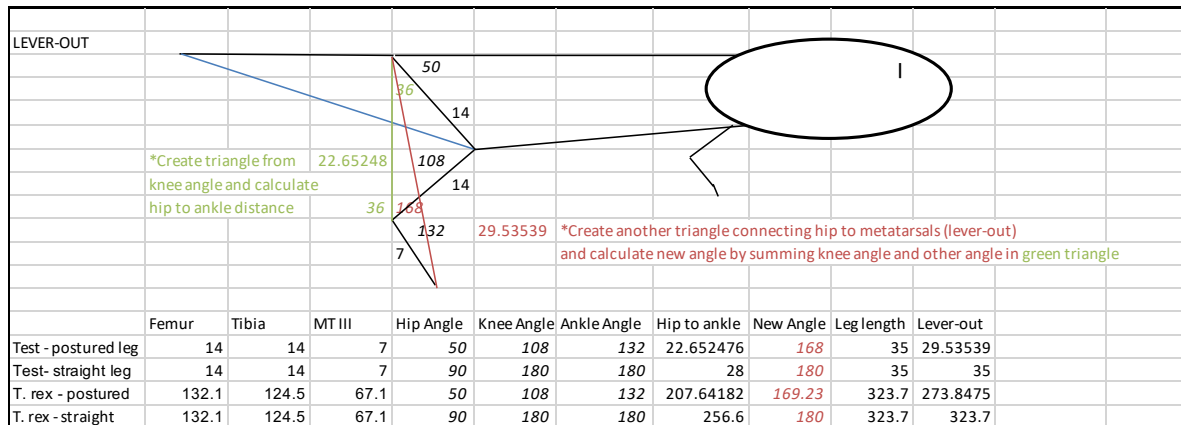
Calculate the new Angle B using DS, DPC, and shoulder angle.

$$\text{Angle B} = \text{DEGREES}(\text{ASIN}(\text{DPC} * (\text{SIN}(\text{RADIANS}(\text{Shoulder Angle}))) / \text{DS}))$$

Green line bisects DS, assumes a right triangle, and calculates length of green line (Lever_{in}) using Angle B and Half scap length.

$$\text{Lever}_{\text{in}} = \text{Half scap length} * \text{SIN}(\text{RADIANS}(\text{New Angle}))$$

Hindlimb lever_{out}



The hindlimb lever_{out} is calculated from the femur, tibia/fibula, and metatarsus lengths. The first step is to calculate the distance between the hip and ankle by using the law of cosines with the femur length, tibia length, and knee joint angle. This distance equals the sum of the femur and tibia lengths when the knee joint is 180 degrees.

$$\text{Hip to ankle} = \text{SQRT}((\text{Tibia}^2) + (\text{Femur}^2) - (2 * \text{Tibia} * \text{Femur} * \text{COS}(\text{RADIANS}(\text{Ankle Angle}))))$$

Create a triangle by connecting the hip to the metatarsus and get the red italicized angle by calculating one of the small green angles (through law of sines) and adding it the wrist angle.

$$\text{New Angle} = \text{Ankle Angle} + \text{DEGREES}(\text{ASIN}(\text{Femur} * (\text{SIN}(\text{RADIANS}(\text{Knee Angle}))) / \text{Hip to ankle distance}))$$

Then, you use the law of cosines with the new angle, hip to ankle distance, and metatarsus length (MT III) to calculate the red line or postured hindlimb length (lever_{out}). This equals the sum of the femur, tibia, and metatarsus lengths when assuming a straight limb (all 180 degrees).

$$\text{Lever}_{\text{out}} = \text{SQRT}((\text{Hip to ankle}^2) + (\text{MTIII}^2) - (2 * \text{Hip to ankle} * \text{MTIII} * \text{COS}(\text{RADIANS}(\text{New Angle}))))$$

Hindlimb lever_{in} (*m. caudofemoralis longus*)

LEVER-IN (CFL)		7							
Straight leg		35.5377	90	90					
		8.602325		5					
				54.4623					
				4.068667					
				*Bisect CFL length and assume right angle					
				*Treat tail length as hypotenuse and calculate length to bisected point (Lever-in)					
	4th troch	Tail	Hip Angle	CFL	Angle B	Lever-in	Angle A	Angle Check	
Test	5	7	90	8.602325	35.5377	4.0686674	54.462322	180	
T. rex	46.5	392.737	90	395.4802	6.75238	46.177457	83.247624	180	

LEVER-IN (CLF)		7							
Postured leg		20.556	130	50					
		10.90849		5					
				29.444					
				2.457862					
				*Bisect CFL length and assume right angle					
				*Treat tail length as hypotenuse and calculate length to bisected point (Lever-in)					
	4th troch	Tail	Hip Angle	CFL	Angle B	Lever-in	Angle A	Angle Check	
Test	5	7	130	10.90849	20.556	2.4578615	29.443977	180	
T. rex	46.5	392.737	130	424.1251	4.81779	32.984867	45.182215	180	

The *m. caudofemoralis longus* originates on the transverse processes of the tail caudal vertebrae (Tail) and inserts on the 4th trochanter of the femur (4th troch). Use distance to origin (Tail), hip angle, and 4th troch length to calculate the *m. caudofemoralis* line of action (CFL).

$$\text{CFL} = \text{SQRT}((4\text{thtroch}^2) + (\text{Tail}^2) - (2 * 4\text{thtroch} * \text{Tail} * \text{COS}(\text{RADIANS}(\text{Hip Angle}))))$$

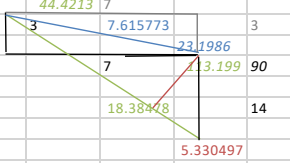
Calculate the new Angle B using CFL, 4th troch, and hip angle.

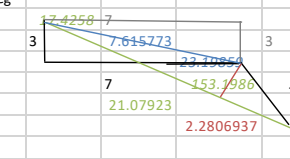
$$\text{Angle B} = \text{DEGREES}(\text{ASIN}(4\text{th troch} * (\text{SIN}(\text{RADIANS}(\text{Hip Angle}))) / \text{CFL}))$$

Green line bisects DS, assumes a right triangle, and calculates length of green line (Lever_{in}) using Angle B and Half scap length.

$$\text{Lever}_{in} = \text{Tail} * \text{SIN}(\text{RADIANS}(\text{Angle B}))$$

Hindlimb lever_{in} (*m. iliotibialis lateralis*)

LEVER-IN (ITL)					Steps:					
Straight leg					*Assume a rectangular-shaped ilium and estimate distance from acetabulum to anteroposterior ilium (hypotenuse) *Take the inverse tangent of ilium height and post-acetabular ilium length to calculate alpha angle *Create new triangle using femur length and newly estimated hypotenuse *Calculate big angle between femur and anteroposterior ilium (new hypotenuse) = 180 - hip angle + alpha angle *Calculate ITL length using Law of Cosines with big angle and sides (hypotenuse and femur) of new triangle *Calculate small angle between hypotenuse and ITL length *Bisect ITL length and uses small angle and hypotenuse to calculate distance to bisected point (lever-in)					
					Acetabulum	Alpha Angle	Big Angle	ITL Length	Small Angle	Lever-in
Test	14	3	7	90	7.6157731	23.1985905	113.199	18.38478	44.42127	5.330497
Archaeopteryx	5.14	0.923	1.063	90	1.407799	40.967693	130.968	6.15548	39.08795	0.887635

LEVER-IN (ITL)					Steps:					
Postured leg					*Assume a rectangular-shaped ilium and estimate distance from acetabulum to anteroposterior ilium (hypotenuse) *Take the inverse tangent of ilium height and post-acetabular ilium length to calculate alpha angle *Create new triangle using femur length and newly estimated hypotenuse *Calculate big angle between femur and anteroposterior ilium (new hypotenuse) = 180 - hip angle + alpha angle *Calculate ITL length using Law of Cosines with big angle and sides (hypotenuse and femur) of new triangle *Calculate small angle between hypotenuse and ITL length *Bisect ITL length and uses small angle and hypotenuse to calculate distance to bisected point (lever-in)					
					Acetabulum	Alpha Angle	Big Angle	ITL Length	Small Angle	Lever-in
Test	14	3	7	50	7.6157731	23.198591	153.199	21.07923	17.42576	2.280694
Archaeopteryx	5.14	0.923	1.063	50	1.407799	40.967693	170.968	6.534081	7.093934	0.173858

Assume a rectangular-shaped ilium and estimate distance from acetabulum to anteroposterior ilium (hypotenuse)

$$\text{Acetabulum} = \text{SQRT}(\text{Ilium height}^2 + \text{Post-acetabular length}^2)$$

Take the inverse tangent of ilium height and post-acetabular ilium length to calculate alpha angle

$$\text{Alpha Angle} = \text{DEGREES}(\text{ATAN}(\text{Ilium height} / \text{Post-acetabular length}))$$

Create new triangle using femur length and newly estimated hypotenuse

Calculate big angle between femur and anteroposterior ilium (new hypotenuse)

$$\text{Big Angle} = 180 - \text{Hip Angle} + \text{Alpha Angle}$$

Calculate ITL length using Law of Cosines with big angle and sides (hypotenuse and femur) of new triangle

$$\text{ITL Length} = \text{SQRT}((\text{Femur}^2) + (\text{Acetabulum}^2) - (2 * \text{Femur} * \text{Acetabulum} * \text{COS}(\text{RADIANS}(\text{Big Angle}))))$$

Calculate small angle between hypotenuse and ITL length

$$\text{Small Angle} = \text{DEGREES}(\text{ASIN}(\text{Femur} * (\text{SIN}(\text{RADIANS}(\text{Big Angle}))) / \text{ITL}))$$

Bisect ITL length and uses small angle and hypotenuse to calculate distance to bisected point (lever-in)

$$\text{Lever}_{\text{in}} = \text{Acetabulum} * \text{SIN}(\text{RADIANS}(\text{Small Angle}))$$

2. Regression multicollinearity assumptions

We ensured that our independent variables did not carry redundant information (i.e., multicollinearity) by calculating variance inflation factors (VIFs) using the package car in R¹. Two or more variables are collinear if they share a VIF > 5.0. We found that multicollinearity was absent in our full PGLS models after removing the interaction variables, as well as between time and node count in our path rate regression analyses.

Table 1. Variance inflation factors for all independent variables in the full regression models.

Model	Lever _{in}	Clade
<i>Forelimb, lever_{out}</i> (response)	1.867104	1.867104
<i>Hindlimb, lever_{out}</i> (response)	3.354968	3.354968

Model	Time	Node count
<i>Forelimb, path rate</i> (response)	1.025295	1.025295
<i>Hindlimb, path rate</i> (response)	1.023972	1.023972

3. Detailed model results

Here, we provide more complete results from the models referenced in the main text. In each table, the green row represents the best supported model.

Forelimb results

Table 2. Model selection for variable-rates analysis of the forelimb lever_{out}.

Univariate, Log ₁₀ Lever _{out}		
Model	LogMarginalLh	Bayes factor
Single rate	-27.393205	-
Variable rates	-26.506676	1.773058

Table 3. Model selection for variable-rates analysis of the forelimb lever_{in}.

Univariate, Log ₁₀ Lever _{in}		
Model	LogMarginalLh	Bayes factor
Variable rates	-34.678577	-
Single rate	-35.999252	2.64135

Table 4. Model selection for variable-rates analysis of the forelimb gear ratio.

Univariate, Log ₁₀ Gear ratio		
Model	LogMarginalLh	Bayes factor
Variable rates	25.821865	-
Single rate	14.606918	22.429894

Table 5. Model selection for the multiple linear regression of lever arm components. A variable-rates (VR) analysis with Pennaraptora included as a dummy-coded indicator variable did not improve over the uniform-rate model. The simple VR model fit the data better than the simple uniform-rate model (Bayes factor = 29.68).

Independent Contrasts Regression, Log ₁₀ Lever _{out} ~ Log ₁₀ Lever _{in}		
Model	LogMarginalLh	Bayes factor
Pennaraptor, VR	63.038319	-
Pennaraptora	62.550068	0.976502
Pennaraptor interact	62.244518	0.6111
Simple, VR	53.238828	29.682764
Simple	38.397446	48.305244
Theropod	32.208333	60.68347
Theropod interact	52.452036	20.196064
Sauropod	32.00496	61.090216
Sauropod interact	27.231758	70.63662
Ornithischia	32.161134	60.777868
Ornithischia interact	27.117119	70.865898
Clades	51.155357	22.789422
Clades int	46.455407	32.189322

Table 6. Model selection for presence of phylogenetic signal in the relationship between forelimb lever_{out} and lever_{in}, accounting for an intercept difference among pennaraptorans and incorporating a distribution of lever_{in} values calculated from a random set of predicted deltopectoral crest measurements.

PGLS Regression, Log₁₀ Lever_{out} ~ Log₁₀ Lever_{in} + Pennaraptora, dist data		
Model	LogMarginalLh	Bayes factor
Lambda	51.519226	-
Lambda 0	35.992149	31.054154

Table 7. Average parameter estimates of the posterior distribution (post-burnin) from the final forelimb model.

PGLS Regression, Log₁₀ Lever_{out} ~ Log₁₀ Lever_{in} + Pennaraptora, lambda est, dist data	
Parameter	Avg estimate
Alpha	1.173588705
Beta 1	0.774860533
Beta 2	-
Alpha.Pennaraptor	0.613186882
Alpha	0.560401823
Var	0.000154197
R ²	0.866851778
SSE	0.012788279
SST	0.095852611
s.e. Alpha	0.028845144
s.e. Beta-1	0.03627741
s.e. Beta-2	0.059838658
Lambda	0.926750878
pMCMC - Beta 1	0
pMCMC - Beta 2	0

Hindlimb results

Table 8. Model selection for variable-rates analysis of the hindlimb lever_{out}.

Univariate, Log₁₀ Lever_{out}		
Model	LogMarginalLh	Bayes factor
Single rate	-27.163277	-
Variable rates	-27.727019	-1.127484

Table 9. Model selection for variable-rates analysis of the hindlimb lever_{in}.

Univariate, Log₁₀ Lever_{in}		
Model	LogMarginalLh	Bayes factor
Variable rates	-50.213488	-
Single rate	-51.598473	2.76997

Table 10. Model selection for variable-rates analysis of the hindlimb gear ratio.

Univariate, Log ₁₀ Gear ratio		
Model	LogMarginalLh	Bayes factor
Variable rates	-2.685686	-
Single rate	-10.577472	15.783572

Table 11. Model selection for the multiple linear regression of lever arm components. A variable-rates (VR) analysis of the simple model fit the data better than the simple uniform-rate model (Bayes factor = 8.43).

Independent Contrasts Regression, Log ₁₀ Lever _{out} ~ Log ₁₀ Lever _{in}		
Model	LogMarginalLh	Bayes factor
Simple, VR	36.742658	-
Simple	32.526145	8.433026
Avialae	26.017145	13.018
Avialae interact	21.523485	22.00532
Maniraptor	26.699189	11.653912
Maniraptor interact	23.977563	17.097164
Theropod	26.131019	12.790252
Theropod interact	30.816318	3.419654
Sauropod	25.858193	13.335904
Sauropod interact	20.847784	23.356722
Clades	13.836998	37.378294
Clades interact	10.325665	44.40096

Table 12. Model selection for presence of phylogenetic signal in the relationship between forelimb lever_{out} and lever_{in}, incorporating a distribution of lever_{in} values calculated from a random set of predicted 4th trochanter measurements.

PGLS Regression, Log ₁₀ Lever _{out} ~ Log ₁₀ Lever _{in} , dist data		
Model	LogMarginalLh	Bayes factor
Lambda	25.354978	-
Lambda 0	13.134964	24.440028

Table 13. Average parameter estimates of the posterior distribution (post-burnin) from the final hindlimb model.

PGLS Regression, Log ₁₀ Lever _{out} ~ Log ₁₀ Lever _{in} , lambda est, dist data	
Parameter	Avg estimate
Alpha	1.215123925
Beta 1	0.616616102
Var	0.000185475
R ²	0.808794905
SSE	0.012044912
SST	0.062822307
s.e. Alpha	0.045883709
s.e. Beta-1	0.037166028
Lambda	0.732617183
pMCMC - Beta 1	0

Path rate results

Table 14. Results of the forelimb path rate regression analyses using the root-to-tip sum of the median estimated rates obtained from the variable-rates regression analyses.

Forelimb, Log ₁₀ median root-to-tip path rate (response) ~ time path length (Predictor 1)							
Predictor 2	Alpha	Beta 1	pMCMC 1	Beta 2	pMCMC 2	R ²	Lambda
Node count	0.4515	0.05921	0.0906	0.05921	0	0.7082	0.9804
Gait	0.6146	0.0014	0.0081	-0.0297	0.3557	0.0676	0.9891

Table 15. Results of the hindlimb path rate regression analyses using the root-to-tip sum of the median estimated rates obtained from the variable-rates regression analyses.

Hindlimb, Log ₁₀ median root-to-tip path rate (response) ~ time path length (Predictor 1)							
Predictor 2	Alpha	Beta 1	pMCMC 1	Beta 2	pMCMC 2	R ²	Lambda
Node count	0.3678	0.00019	0.1682	0.0672	0	0.9214	0.9620
Gait	0.4195	0.0024	0	0.0418	0.29	0.2119	0.9833

Reduced dataset rates analyses

Table 16. Model selection for variable-rates regression analysis of the forelimb leverout on leverin.

Regression, forelimb, trimmed data, Log10 Lever-out ~ Log10 Lever-in		
Model	LogMarginalLh	BayesFactor
Variable rates	20.759301	-
Single rate	11.118107	19.282388

Table 17. Model selection for variable-rates regression analysis of the hindlimb leverout on leverin.

Regression, hindlimb, trimmed data, Log10 Lever-out ~ Log10 Lever-in		
Model	LogMarginalLh	BayesFactor
Variable rates	24.718495	-
Single rate	23.037632	3.361726

Results for the ordinary least-squares regression of the median estimated forelimb branch-wise rates of gear ratio evolution on the median estimated hindlimb branch-wise rates with long pennaraptoran branches.

```
#
# Estimate Std. Error t value Pr(>|t|)
# (Intercept) 0.0125613 0.0333224 0.377 0.707
# Original BL 0.0004320 0.0005904 0.732 0.466
# LogHL_median 0.8959234 0.0988221 9.066 2.3e-14 ***
# ---
# Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# Residual standard error: 0.1871 on 91 degrees of freedom
# Multiple R-squared: 0.4767, Adjusted R-squared: 0.4652
# F-statistic: 41.44 on 2 and 91 DF, p-value: 1.601e-13
```

Results for the ordinary least-squares regression of the median estimated forelimb branch-wise rates of gear ratio evolution on the median estimated hindlimb branch-wise rates without long pennaraptoran branches.

#	Estimate	Std. Error	t value	Pr(> t)
# (Intercept)	0.0162547	0.0098893	1.644	0.104
# Original BL	0.0001709	0.0001758	0.972	0.334
# LogHL_median	0.7638230	0.0296088	25.797	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.05546 on 89 degrees of freedom
Multiple R-squared: 0.8841, Adjusted R-squared: 0.8815
F-statistic: 339.3 on 2 and 89 DF, p-value: < 2.2e-16

5. References cited

1. Fox, J., and Weisberg, S. (2019). An R Companion to Applied Regression (Sage).
2. Freckleton, R.P. (2009). The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology* 22, 1367–1375. 10.1111/j.1420-9101.2009.01757.x.

Appendix 3

Data table

Hindlimb (HL) and forelimb (FL) lever_{out} (Lout), lever_{in} (Lin), and gear ratios, along with the clade and specimen information, for each studied species. ITL indicates whether the hindlimb gear ratio was based on the *m. caudofemoralis longus* measurement (ITL = 0) or the *m. iliotibialis lateralis* (ITL = 1).

Species	Clade	Subclade	Specimen ID	ITL	HL_Lout	HL_Lin	HL_Gear_Ratio	FL_Lout	FL_Lin	FL_Gear_Ratio
Brachylophosaurus_canadensis	Omithischia	Omithopoda	MOR 794	0	236.510137	29.1595681	8.11089298	138.61734	11.8758656	11.6721884
Centrosaurus_apertus	Omithischia	Marginocephalia	TMP 1981.003.0001	0	109.718066	20.6983399	5.30081478	88.2986591	11.9885172	7.36526945
Chasmosaurine_indet	Omithischia	Marginocephalia	CMN 8547	0	121.79542	23.44204	5.19559815	86.6191665	11.2760025	7.68172643
Chasmosaurus_belli	Omithischia	Marginocephalia	TMP 1982.052.0002	0	113.775088	23.8758645	4.76527616	103.732529	13.9799478	7.42009417
Corythosaurus_casuarius	Omithischia	Omithopoda	ROM 845	0	203.731818	38.7947223	5.25153439	-	-	-
Edmontosaurus_annectens	Omithischia	Omithopoda	YPM 2182	-	-	-	-	121.649286	11.645808	10.4457574
Edmontosaurus_regalis	Omithischia	Omithopoda	NMC 2288	-	-	-	-	125.956618	13.4614514	9.35683782
Euoplocephalus_tutus	Omithischia	Thyreophoran	TMP 1989.007.0009	0	99.4741349	15.0552865	6.6072562	71.8801157	10.1800545	7.0608773
Heterodontosaurus_tucki	Omithischia	Heterodontosauridae	TMP 1984.172.0001	0	28.0772371	3.00054968	9.35736454	-	-	-
Hypacrosaurus_altispinus	Omithischia	Omithopoda	NMC 8501	-	-	-	-	117.950019	11.517619	10.2408336
Iguanodon_bemissartensis	Omithischia	Omithopoda	IRSNM 1534	-	-	-	-	135.855951	11.4701686	11.8442855
Jeholosaurus_shangyuanensis	Omithischia	Thescelosaurinae	IVPP V12542	0	22.6218787	3.14125589	7.20153961	-	-	-
Gryposaurus_notabilis	Omithischia	Omithopoda	TMP 1980.051.0002	0	189.245084	35.5646008	5.32116431	120.647923	14.405708	8.37500827
Lambeosaurus_lambeii	Omithischia	Omithopoda	TMP 1982.038.0001 (HL)	0	212.786099	32.4510429	6.55714207	121.440424	11.8063717	10.2860072
Lambeosaurus_magnicristatus	Omithischia	Omithopoda	TMP 66.4.1	-	-	-	-	122.239878	10.6808146	11.4448085
Leptoceratops_gracilis	Omithischia	Marginocephalia	CMN 8888	0	56.3444452	7.82667366	7.19902831	39.3288341	4.20881199	9.34440269
Lesothosaurus_diagnosticus	Omithischia	Neomithischia	BMNH RUB17	-	-	-	-	9.06592539	0.62378273	14.5337872
Pachyrhinosaurus_lakustai	Omithischia	Marginocephalia	TMP 1993.003.0001 (HL)	0	121.97088	22.4791827	5.42594814	96.6934927	12.3296376	7.84236292
Parasaurolophus_cyrtocristatus	Omithischia	Omithopoda	FMNH P27393	-	-	-	-	120.055201	12.0564322	9.95777186
Parasaurolophus_walkerii	Omithischia	Omithopoda	ROM 768	-	-	-	-	106.628882	12.2784219	8.68424968
Parksosaurus_warreni	Omithischia	Thescelosaurinae	ROM 804	0	61.8985917	8.96794605	6.90220384	-	-	-
Prosaurolophus_maximus	Omithischia	Omithopoda	TMP 1984.000.0009 (HL)	0	194.201867	27.9149641	6.95690907	109.382333	10.6312663	10.2887398
Protoceratops_andrewsi	Omithischia	Marginocephalia	ZPAL MgD-II/3	-	-	-	-	18.4716015	1.97108179	9.37130136
Psittacosaurus_neimongoliensis	Omithischia	Marginocephalia	IVPP 12-0888-2	0	28.3894103	4.24730888	6.68409365	16.6047347	1.80825278	9.18275083
Saurolophus_osborni	Omithischia	Omithopoda	AMNH 5220	-	-	-	-	130.619385	11.4925932	11.3655275
Scelidosaurus_harrisonii	Omithischia	Thyreophoran	BRSMG Ce12785	0	66.1337302	12.2058163	5.41821445	49.7096288	7.43050752	6.68993721
Shantungosaurus_giganteus	Omithischia	Omithopoda	IVPP no number (from HL)	-	-	-	-	203.454876	22.9726125	8.85641004
Stegouros_elengassen	Omithischia	Thyreophoran	CPAP-3165	0	30.8154448	4.46963937	6.89439176	-	-	-
Tenontosaurus_tilleti	Omithischia	Omithopoda	OU 11	0	111.462894	17.1891923	6.48447537	71.4632637	9.58132055	7.45860274
Thescelosaurus_sp	Omithischia	Thescelosaurinae	MOR 979	0	90.2103541	18.1497308	4.97034116	43.8323176	6.43002493	6.81681924
Triceratops_horridus	Omithischia	Marginocephalia	TMP 1982.006.0001	0	158.345764	29.1127107	5.43905944	113.476246	17.0760234	6.6453555
Tsintaosaurus_spinorhinus	Omithischia	Omithopoda	PMNH V728	-	-	-	-	138.276486	11.1209514	12.433872
Vagaceratops_irvinensis	Omithischia	Marginocephalia	CMN 41357	-	-	-	-	95.8254392	12.5685956	7.62419626

Apatosaurus_louisae	Sauropodomorpha	Sauropoda	CM 3018	0	261.477226	80.1969218	3.26043968	-	-	-
Alamosaurus_sanjuanensis	Sauropodomorpha	Sauropoda	USNM 15560	-	-	-	-	265.275392	28.6456274	9.26058937
Camarasaurus_lentus	Sauropodomorpha	Sauropoda	CM 11338	0	85.2395486	19.946493	4.27341031	86.9240358	6.89077384	12.6145536
Galeamopus_pabsti	Sauropodomorpha	Sauropoda	SMA 0011	-	-	-	-	176.146085	12.946252	13.6059521
Lufengosaurus	Sauropodomorpha	Massospondylidae	IVPP V15	0	97.9511539	17.524251	5.58946309	45.0493767	6.16371347	7.30880449
Massospondylus_carinatus	Sauropodomorpha	Massospondylidae	BP/1/4934	-	-	-	-	50.3578825	6.63724202	7.58716984
Opisthocoelicaudia_skarzynskii	Sauropodomorpha	Sauropoda	ZPAL MgD-1/48*	0	201.668768	70.3618969	2.86616445	-	-	-
Rapetosaurus_krausei	Sauropodomorpha	Sauropoda	FMNH PR 2209	-	-	-	-	89.4383515	9.12602205	9.80036548
Xingxiulong_chengi	Sauropodomorpha	Sauropodiformes	LFGT-D0003	-	-	-	-	58.6835399	7.56737594	7.754807
Eoraptor_lunensis	Theropoda	Theropoda	PVSJ 512	0	32.8726686	3.91055567	8.40613748	13.9467258	0.99253887	14.0515664
Guaibasaurus_candelariensis	Theropoda	Theropoda	MCN-PV 2355, MCN-PV	-	-	-	-	16.8500668	1.46743967	11.48263
Herrerasaurus_ischigualastensis	Theropoda	Herrerasauridae	PVSJ 373	-	-	-	-	32.8564379	1.47783745	22.232782
Achillobator_giganticus	Theropoda	Maniraptoriformes	MNUFR 15	1	103.615978	6.30594334	16.4314794	-	-	-
Allosaurus	Theropoda	Allosauroidea	MOR 693	-	-	-	-	60.6999246	3.79665621	15.9877327
Alxasaurus_elesitaiensis	Theropoda	Maniraptoriformes	IVPP 88402	0	92.9548964	14.1693963	6.56025806	59.0181942	4.38963351	13.4449024
Ambopteryx_longibrachium	Theropoda	Maniraptoriformes	IVPP V24192	1	10.3003763	0.52767875	19.520165	5.78313335	1.60730124	3.59803951
Anchiomis_huxleyi	Theropoda	Maniraptoriformes	IVPP V14378	1	13.1617776	0.36488259	36.0712682	4.82006584	1.0739074	4.48834402
Anzu_wyliei	Theropoda	Maniraptoriformes	CM 78000	-	-	-	-	34.4646886	11.5123064	2.99372578
Beishanlong_grandis	Theropoda	Maniraptoriformes	FRDC-GS GJ (06) 01-18	-	-	-	-	83.9708706	9.5811619	8.76416363
Caudipteryx_sp	Theropoda	Maniraptoriformes	IVPP V12430	1	37.5853103	0.92645331	40.5690281	7.24506028	2.58998127	2.7973408
Ceratosaurus_nasicornis	Theropoda	Ceratosauria	USNM 4735	0	120.249385	15.5963354	7.71010507	-	-	-
Chirostenotes_pergracilis	Theropoda	Maniraptoriformes	TMP 79.20.1	1	75.4363665	1.20221059	62.748047	-	-	-
Citipati_osmolskae	Theropoda	Maniraptoriformes	IGM 100/1004	-	-	-	-	25.2617244	10.6182457	2.37908645
Coelophysis_bauri	Theropoda	Coelophysoidea	AMNH 7229	0	31.7015195	3.08679645	10.2700389	-	-	-
Conchoraptor_gracilis	Theropoda	Maniraptoriformes	TMP 2013.010.0004	0	53.2947986	5.44117	9.79473139	-	-	-
Deinococheirus_mirificus	Theropoda	Maniraptoriformes	MPC-D 100/18 (forelimb),	0	266.200451	31.665912	8.40653037	153.864849	11.0844118	13.881192
Deinonychus_antirrhopus	Theropoda	Maniraptoriformes	MCZ 4371	1	73.0124253	3.00689816	24.2816422	-	-	-
Dilophosaurus_wetherilli	Theropoda	Neotheropoda	UCMP 37302	0	119.212387	13.0148881	9.15969361	49.103139	3.18518827	15.4160869
Eodromaues_murphi	Theropoda	Theropoda	PVSJ 562	-	-	-	-	15.5637006	0.92954452	16.7433622
Epidexipteryx_hui	Theropoda	Maniraptoriformes	IVPP V15471	-	-	-	-	4.59168142	1.43601977	3.19750571
Garudimimus_brevipes	Theropoda	Maniraptoriformes	GIN 100/13	1	83.8883639	4.19213604	20.0108878	-	-	-
Gorgosaurus_libratus	Theropoda	Tyrannosauroidea	TMP 1991.036.0500, TM	0	167.773987	14.7858723	11.3469117	34.1587112	2.68806723	12.707536
Haplocheirus_sollers	Theropoda	Maniraptoriformes	IVPP V15988	-	-	-	-	17.7167718	1.2135949	14.5985879
Harpyimimus_okladnikovi	Theropoda	Maniraptoriformes	IGM 100/29	-	-	-	-	52.7579363	2.15994388	24.4256051
Khaan_mckennai	Theropoda	Maniraptoriformes	IGM 100/1127	1	41.7062545	2.02104496	20.6359854	11.5755021	5.48999747	2.1084713
Liliensternus_lilienstermi	Theropoda	Coelophysoidea	MB.R.2175	-	-	-	-	34.9542401	2.325655	15.0298475

Mahakala_omnogovae	Theropoda	Maniraptoriformes	IGM 100/1033	1	23.3115264	0.85287447	27.3328929	-	-	-
Majungasaurus_crenatissimus	Theropoda	Ceratosauria	FMNH PR 2836	-	-	-	-	24.5937985	5.2390982	4.69428087
Meraxes_gigas	Theropoda	Allosauroidae	MMCh-PV 65	0	238.505173	32.312678	7.38116393	53.7649231	4.20084315	12.7986028
Microraptor_gui	Theropoda	Maniraptoriformes	IVPP V13352	1	22.5877962	0.3323553	67.9627989	10.0729654	2.90165289	3.47145776
Microraptor_zhaoianus	Theropoda	Maniraptoriformes	CAGS 20-8-001	1	18.4932641	0.53612404	34.4943757	6.58941621	2.08422307	3.16156956
Mononykus_olecranus	Theropoda	Maniraptoriformes	GI N107/6	-	-	-	-	6.82985573	0.51373997	13.2943827
Nemegtomykus_citus	Theropoda	Maniraptoriformes	MPC D-100/203	0	32.8862655	2.68055081	12.2684731	-	-	-
Nothronychus_graffami	Theropoda	Maniraptoriformes	UMNH VP 16420	0	146.159488	21.4283836	6.82083591	72.2527246	4.82985761	14.9595972
Oksoko_avarsan	Theropoda	Maniraptoriformes	MPC-D 100/33	-	-	-	-	11.8700987	6.39906734	1.85497324
Omitholestes_hermanni	Theropoda	Coelurosauria	TMP 2006.003.0002	0	38.2536767	4.39953768	8.69493102	19.1344256	1.33958357	14.283861
Omithomimus_edmontonicus	Theropoda	Maniraptoriformes	TMP 1995.110.0001	0	111.198545	8.63261575	12.881211	50.9763746	2.03402489	25.0618244
Rahonavis_ostromi	Theropoda	Maniraptoriformes	UA 8656	1	21.3319911	0.71841552	29.6931099	-	-	-
Raptorex_kriegsteini	Theropoda	Tyrannosauroidae	LH PV18	0	85.5063541	7.78265533	10.9867841	14.0711912	0.55223279	25.4805428
Sinocalliopteryx_gigas	Theropoda	Compsognathidae	JMP-V-05-8-01	0	58.5491348	4.89754986	11.9547808	23.2877723	1.30240587	17.8805799
Sinomithosaurus_milleni	Theropoda	Maniraptoriformes	IVPP V12811	1	30.6368582	1.30023235	23.5626027	13.7858759	4.3212882	3.19022367
Sinosauropteryx_prima	Theropoda	Compsognathidae	TMP 2006.013.0001	0	28.900903	1.94589676	14.8522284	8.92116514	0.48511471	18.3898056
Sinovenator_changii	Theropoda	Maniraptoriformes	IVPP V20378	1	24.948197	1.12279017	22.2198213	6.16647076	1.69750391	3.63266954
Sinuronasus_magnodens	Theropoda	Maniraptoriformes	IVPP V11527	1	41.572226	1.36416999	30.4743737	-	-	-
Struthiomimus_altus	Theropoda	Maniraptoriformes	TMP 1985.008.0003	0	116.537126	9.2143837	12.6473055	52.7962283	2.04870036	25.7705955
Tanycolagrus_topwilsoni	Theropoda	Coelurosauria	TPII 2000-09-29	-	-	-	-	35.7825511	1.74202296	20.5408034
Tyrannosaurus_rex	Theropoda	Tyrannosauroidae	FMNH PR2081	0	273.847534	32.9848668	8.30221736	58.3806048	3.73911388	15.6134867
Velociraptor_mongoliensis	Theropoda	Maniraptoriformes	IGM 100/986	0	49.5893359	4.64053934	10.6861148	-	-	-
Zhenyuanlong_suni	Theropoda	Maniraptoriformes	JPM-0008	1	49.4168563	1.98383698	24.9097365	12.8911943	3.80928286	3.38415255
Zhongjianosaurus_yangi	Theropoda	Maniraptoriformes	IVPP V22775	-	-	-	-	5.67323986	1.28565446	4.41272522
Archaeopteryx_lithographica	Theropoda	Maniraptoriformes	MOR Cast	1	13.5232552	0.17385813	77.7832786	5.50279254	2.19177163	2.51065962
Auromis_xui	Theropoda	Maniraptoriformes	YFGP-T5198	1	16.9730949	0.44208447	38.3933303	7.31956786	1.17972668	6.20446072
Cathayomis_yandica	Theropoda	Maniraptoriformes	IVPP V9769	-	-	-	-	3.41039535	0.94989503	3.59028655
Confuciusornis_sanctus	Theropoda	Maniraptoriformes	IVPP V11374	1	11.3784127	0.51474695	22.1048666	6.36203511	1.95532617	3.25369507
Hongshanornis_longicresta	Theropoda	Maniraptoriformes	IVPP V14533	1	-	-	-	3.14165406	1.19880967	2.62064458
Jeholomis_prima	Theropoda	Maniraptoriformes	IVPP V13274	1	18.6742255	0.34005043	54.9160471	13.1266567	3.79675519	3.45733554
Longipteryx_chaoyangensis	Theropoda	Maniraptoriformes	IVPP V12325	1	-	-	-	5.61110562	2.03578912	2.75623127
Pengornis_houi	Theropoda	Maniraptoriformes	IVPP V15336	1	10.3815303	0.3646438	28.4703327	8.26914204	2.52073614	3.2804473
Sapeornis_chaoyangensis	Theropoda	Maniraptoriformes	IVPP V13275	1	16.3309066	0.57498921	28.4021097	14.2254228	4.24344637	3.3523277
Serikornis_sungei	Theropoda	Maniraptoriformes	PMOL-AB00200	1	18.3465167	0.68441602	26.8060891	6.75594325	1.59933049	4.22423213
Yixianornis_grabau	Theropoda	Maniraptoriformes	IVPP V12631	1	10.2467699	0.29897137	34.2734148	6.09848895	1.66665528	3.65911836

Summary

"The evolutionary trees that adorn our textbooks have data only at the tips and nodes of their branches; the rest is inference, however reasonable, not the evidence of fossils."

- Stephen Jay Gould, 1977¹

Comparative analyses and modelling have made major contributions to the study of evolution in the past four decades. We can now estimate rates of evolution, speciation, and extinction with unprecedented efficiency and rigour. This thesis utilises the latest advancements in inferring varying rates of evolution across space and ancestor-descendant lineages. These new techniques uncovered the geographic dispersal history of early tetrapodomorphs; however, geographic sampling bias inflates dispersal rates across regions with a sparse terrestrial rock record (Chapter 2; Gardner et al., 2019). For the first time, an established ecogeographic rule (Bergmann's rule) was assessed while accounting for variable rates of body size evolution in dinosaurs and mammals (Chapter 3). The results suggest that dispersals to high latitudes (and relatively colder climates) did not drive body size evolution in dinosaurs and mammals, likely making Bergmann's rule an invalid generalisation for these groups. The variable-rates regression model (Baker et al., 2016; Baker and Venditti, 2019) was also applied to the functional lever arm equations that describe dinosaur limb retraction (Chapter 4). The study shows that, while individual lever parts evolve gradually, the lever arm systems themselves evolve at varying rates. This

¹ Stephen Jay Gould, "Evolution's erratic pace," *Natural History*, Vol. 86, No. 5, pp.12-16, 230 May 1977

finding reveals an emergent evolutionary process by which gradual changes in individual functional parts can give way to bursts of adaptive change at the system level. However, inferences are not direct observations of the past, and it is crucial that we pair comparative methods with fossils to bolster our inferences. For example, Baker et al. use Mesozoic mammal fossils to demonstrate that ancestral body size estimates from variable-rates models are more congruent with the fossil record than those assuming a single uniform rate (Baker et al., 2015).

Another example where the fossil record may provide more clarity is the Darwin's scenario outlined in Chapter 1 (Gardner and Organ, 2021). This hypothetical (but common) scenario is when two characters originate on the same branch of a phylogeny (Maddison and FitzJohn, 2015). Due to cladistics and the focus on shared derived characters, we know many morphological characters that are unique to clades of interest. Naturally, researchers may be interested in what adaptive functions those characters might have (Maddison and FitzJohn, 2015). However, character and trait distributions alone are not indicative of adaptation, as they may have arisen from developmental effects or historical contingencies (Gould and Lewontin, 1979). Fossils can shed light on character evolution. They provide invaluable information on ancestral states (Baker et al., 2015), character transition rates and correlated evolution (Organ et al., 2009), and divergence time estimates (Heath et al., 2014; Stadler et al., 2018). A commonly used example of Darwin's scenario is the co-origination of fur and middle ear bones in mammals (Maddison and FitzJohn, 2015). Interestingly, the mammalian fossil record suggests that middle ear bones evolved multiple times independently (Han et al., 2017; Meng et al., 2011). Exceptionally preserved fossils with soft-tissue remains will continue to shed light on the evolution of fur and

associated integumentary features (Pickrell, 2019). Such fossils reveal a more complex evolutionary history than extant character distributions might suggest.

Chapter 1 also discussed how we ought to study evolution, particularly when using comparative methods (Gardner and Organ, 2021). Phylogenetic comparative analyses are not unique in that they require careful study designs that maximise effective sample sizes. Studies continue to investigate character distributions like Darwin's scenario to establish a model-based solution for avoiding the extreme pseudoreplication of Darwin's and like scenarios (Boyko and Beaulieu, 2022; Uyeda et al., 2018). These include the use of graphical models to clarify the direction of causality (Uyeda et al., 2018) and hidden Markov models to detect tree-wide variability in character transition rates (Boyko and Beaulieu, 2022). However, no matter how complex, new statistical models will never adequately correlate a character with an effective (evolutionary) sample size of one independent character state change—as is the case with Darwin's scenario (Boyko and Beaulieu, 2022; Gardner and Organ, 2021). Chapter 1 concluded with three recommendations for researchers: 1) Researchers should design studies that prioritise *a priori* hypotheses and maximising evolutionary sample sizes. Correlation is not necessarily causal, but it can be consistent with *a priori* hypotheses that have clear and testable predictions. 2) Future studies should also assess the suitability of statistical models. For discrete traits, Chapter 1 introduced the phylogenetic imbalance ratio, which already has been used by researchers (Cosme, 2022) and implemented into R packages (Minter, 2021). 3) Researchers should seek a consilience of evidence from disparate fields, like biogeography and developmental biology, to evaluate evolutionary hypotheses. Consilience has long been considered an indicator of a strong hypothesis or theory

(Whewell, 1840), and it was a major influence on Charles Darwin as he built evidence for his theory of evolution by natural selection (Ruse, 1979, 1975; Thagard, 1977).

Fossils are integral to our understanding of evolution. Although the fossil record is patchy and contains well-known biases (Jablonski et al., 2003; Raup and Boyajian, 1988; Signor and Lipps, 1982), it also records species, ecosystems, and climates that are vastly different from the modern world (Benton, 1995; Mannion et al., 2014). As such, it provides a rich and independent data source to test hypotheses on the biogeography, ecology, and evolution of species. There is no such thing as a perfect data set, but the inherent limitations and biases in the rock record can be addressed with a variety of approaches (Alroy et al., 2001; Benson and Upchurch, 2013; Benton et al., 2013). The more complex our models become, however, the more pressing it is to understand different types of biases. For example, formation count is commonly used as a proxy for sampling bias in comparative analyses (O'Donovan et al., 2018; Sakamoto et al., 2016; Tennant et al., 2016a, 2016b); however, particularly in studies on biogeography, it is crucial to account for the geographic variation in fossil sampling rate (Benson and Upchurch, 2013; Close et al., 2020; Gardner et al., 2019; Jones et al., 2021). Inferences on biogeographic dispersal and ancestral locations might be influenced by disproportionate fossil sampling of different regions around the globe, including countries and communities that historically have had fewer resources to collect or publish new fossils (Raja et al., 2022). In Chapter 2, we established a metric and approach to assess geographic sampling biases and apply it to the dispersal history of early tetrapodomorphs (Gardner et al., 2019). The study found that estimated dispersal rates are exceptionally high in regions with undersampled terrestrial rock records. Estimated ancestral locations also coincided with regions that

are well-sampled for that stage in geologic time, including the ancestral location of the first terrestrial stem-tetrapod.

Sampling the past is also necessary for establishing general rules of ecology. Researchers have long sought to establish ecological rules to organise and explain the global distribution of species. For example, Bergmann's rule posits that warm-blooded animals from higher latitudes (and colder climates) tend to be larger than those from lower latitudes. The rule would have implications for the global distribution of biodiversity and structure of ecosystems, as well as our ability to manage biodiversity due to recent climate change. However, ecological rules in general suffer from three problems: they are often tested in a post hoc fashion through the subsampling of larger datasets, they lack models allowing the evolutionary rate to shift from ancestor to descendant, and they lack null models for proper hypothesis testing. In Chapter 3, these problems were addressed by using a new phylogenetic approach to assess Bergmann's rule, in which the rate of poleward dispersals vary across the tree along with the rate of body size evolution. The study also builds the first null models for Bergmann's rule by leveraging the Mesozoic dinosaur and mammal fossil record, when the climate was more temperate than the Present, including recent fossil discoveries from the Cretaceous Prince Creek Formation of Northern Alaska. This framework was then used to analyse the largest biogeographical dataset of extant mammals yet compiled, for which there is no support for Bergmann's rule as an ecological generalisation. The study sets a new standard for studying ecological rules in extant taxa. It also showcases the fossil record's value for testing broad ecological hypotheses.

The evolution of function is crucial for understanding how changes in climate and environment influence ecosystem structures through time. Changes in the

environment can drive functional adaptation. The study of morphological adaptations have long been typological (Padian and Horner, 2002). Yet, quantifying how functional systems evolve from ancestors to descendants has remained difficult to study due to an absence of methods that model such evolution. Chapter 4 provides a new approach for modelling the evolution of functional systems. The study applies recently developed variable rates models to evolve the parameters of generalised functional equations across phylogenetic lineages of dinosaurs. The analyses reveal that innovations in locomotor function gave way to shifts in the rate of evolution. Multiple dinosaur lineages that evolved large-bodied quadrupedality independently showed major reductions in the rate of forelimb locomotor evolution. This demonstrates that locomotor form can strongly dictate functional adaptation. There was also accelerated evolution in the hindlimb retractor mechanics of birds and bird-like (maniraptoran) dinosaurs. This was driven by a shift to a style of locomotion where the primary hindlimb retractor musculature was reduced and detached from the tail, indicating that locomotor shifts can accelerate the evolution of functional systems. The study provides a framework for studying functional evolution along lineages and reveals how locomotor innovations dictate evolutionary rates.

Altogether, these four studies highlight the utility of the fossil record for informing evolutionary models and our understanding of evolution. When paired with phylogenetic comparative models, fossils can provide invaluable insights into sampling biases (through time, space, and across phylogenies), general ecological trends, and the evolution of function systems and innovative locomotor adaptations.

References

- Alroy, J., Marshall, C.R., Bambach, R.K., Bezusko, K., Foote, M., Fürsich, F.T., Hansen, T.A., Holland, S.M., Ivany, L.C., Jablonski, D., Jacobs, D.K., Jones, D.C., Kosnik, M.A., Lidgard, S., Low, S., Miller, A.I., Novack-Gottshall, P.M., Olszewski, T.D., Patzkowsky, M.E., Raup, D.M., Roy, K., Sepkoski, J.J., Sommers, M.G., Wagner, P.J., Webber, A., 2001. Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Science* 98, 6261–6266. <https://doi.org/10.1073/pnas.111144698>
- Baker, J., Meade, A., Pagel, M., Venditti, C., 2016. Positive phenotypic selection inferred from phylogenies. *Biological Journal of the Linnean Society* 118, 95–115. <https://doi.org/10.1111/bij.12649>
- Baker, J., Meade, A., Pagel, M., Venditti, C., 2015. Adaptive evolution toward larger size in mammals. *Proceedings of the National Academy of Science* 112, 5093–5098. <https://doi.org/10.1073/pnas.1419823112>
- Baker, J., Venditti, C., 2019. Rapid change in mammalian eye shape Is explained by activity pattern. *Current Biology* 29, 1082-1088.e3. <https://doi.org/10.1016/j.cub.2019.02.017>
- Benson, R.B.J., Upchurch, P., 2013. Diversity trends in the establishment of terrestrial vertebrate ecosystems: Interactions between spatial and temporal sampling biases. *Geology* 41, 43–46. <https://doi.org/10.1130/G33543.1>
- Benton, M.J., 1995. Diversification and extinction in the history of life. *Science* 268, 52–58. <https://doi.org/10.1126/science.7701342>
- Benton, M.J., Ruta, M., Dunhill, A.M., Sakamoto, M., 2013. The first half of tetrapod evolution, sampling proxies, and fossil record quality. *Palaeogeography*,

- Palaeoclimatology, Palaeoecology 372, 18–41.
<https://doi.org/10.1016/j.palaeo.2012.09.005>
- Boyko, J.D., Beaulieu, J.M., 2022. Reducing the biases in false correlations between discrete characters. *Systematic Biology* syac066.
<https://doi.org/10.1093/sysbio/syac066>
- Close, R.A., Benson, R.B.J., Alroy, J., Carrano, M.T., Cleary, T.J., Dunne, E.M., Mannion, P.D., Uhen, M.D., Butler, R.J., 2020. The apparent exponential radiation of Phanerozoic land vertebrates is an artefact of spatial sampling biases. *Proceedings of the Royal Society B: Biological Sciences* 287, 20200372. <https://doi.org/10.1098/rspb.2020.0372>
- Cosme, M., 2022. Mycorrhizas shape the evolution of plant adaptation to drought. *bioRxiv*. <https://doi.org/10.1101/2022.05.13.491064>
- Gardner, J.D., Organ, C.L., 2021. Evolutionary sample size and concision in phylogenetic comparative analysis. *Systematic Biology* 70, 1061–1075.
<https://doi.org/10.1093/sysbio/syab017>
- Gardner, J.D., Surya, K., Organ, C.L., 2019. Early tetrapodomorph biogeography: Controlling for fossil record bias in macroevolutionary analyses. *Comptes Rendus Palevol* 18, 699–709. <https://doi.org/10.1016/j.crpv.2019.10.008>
- Gould, S.J., Lewontin, R.C., 1979. The spandrels of San Marco and the Panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society of London, Series B* 205, 581–598.
<https://doi.org/10.1098/rspb.1979.0086>
- Han, G., Mao, F., Bi, S., Wang, Y., Meng, J., 2017. A Jurassic gliding euharamiyidan mammal with an ear of five auditory bones. *Nature* 551, 451–456.
<https://doi.org/10.1038/nature24483>

- Heath, T.A., Huelsenbeck, J.P., Stadler, T., 2014. The fossilized birth–death process for coherent calibration of divergence-time estimates. *Proceedings of the National Academy of Science* 111, E2957–E2966.
<https://doi.org/10.1073/pnas.1319091111>
- Jablonski, D., Roy, K., Valentine, J.W., Price, R.M., Anderson, P.S., 2003. The impact of the pull of the recent on the history of marine diversity. *Science* 300, 1133–1135. <https://doi.org/10.1126/science.1083246>
- Jones, L.A., Dean, C.D., Mannion, P.D., Farnsworth, A., Allison, P.A., 2021. Spatial sampling heterogeneity limits the detectability of deep time latitudinal biodiversity gradients. *Proceedings of the Royal Society B: Biological Sciences* 288, 20202762. <https://doi.org/10.1098/rspb.2020.2762>
- Maddison, W.P., FitzJohn, R.G., 2015. The unsolved challenge to phylogenetic correlation tests for categorical characters. *Systematic Biology* 64, 127–136.
<https://doi.org/10.1093/sysbio/syu070>
- Mannion, P.D., Upchurch, P., Benson, R.B.J., Goswami, A., 2014. The latitudinal biodiversity gradient through deep time. *Trends in Ecology & Evolution* 29, 42–50. <https://doi.org/10.1016/j.tree.2013.09.012>
- Meng, J., Wang, Y., Li, C., 2011. Transitional mammalian middle ear from a new Cretaceous Jehol eutriconodont. *Nature* 472, 181–185.
<https://doi.org/10.1038/nature09921>
- Minter, K.A. and A., 2021. windex: analysing convergent evolution using the Wheatsheaf Index. <https://doi.org/10.4137/EBO.S20968>
- O'Donovan, C., Meade, A., Venditti, C., 2018. Dinosaurs reveal the geographical signature of an evolutionary radiation. *Nature Ecology & Evolution* 2, 452.
<https://doi.org/10.1038/s41559-017-0454-6>

- Organ, C.L., Janes, D.E., Meade, A., Pagel, M., 2009. Genotypic sex determination enabled adaptive radiations of extinct marine reptiles. *Nature* 461, 389–392. <https://doi.org/10.1038/nature08350>
- Padian, K., Horner, J.R., 2002. Typology versus transformation in the origin of birds. *Trends in Ecology & Evolution* 17, 120–124. [https://doi.org/10.1016/S0169-5347\(01\)02409-0](https://doi.org/10.1016/S0169-5347(01)02409-0)
- Pickrell, J., 2019. How the earliest mammals thrived alongside dinosaurs. *Nature* 574, 468–472. <https://doi.org/10.1038/d41586-019-03170-7>
- Raja, N.B., Dunne, E.M., Matiwane, A., Khan, T.M., Nätscher, P.S., Ghilardi, A.M., Chattopadhyay, D., 2022. Colonial history and global economics distort our understanding of deep-time biodiversity. *Nature Ecology & Evolution* 6, 145–154. <https://doi.org/10.1038/s41559-021-01608-8>
- Raup, D.M., Boyajian, G.E., 1988. Patterns of generic extinction in the fossil record. *Paleobiology* 14, 109–125. <https://doi.org/10.1017/S0094837300011866>
- Ruse, M., 1979. Falsifiability, consilience, and systematics. *Systematic Biology* 28, 530–536. <https://doi.org/10.2307/sysbio/28.4.530>
- Ruse, M., 1975. Darwin's debt to philosophy: An examination of the influence of the philosophical ideas of John F.W. Herschel and William Whewell on the development of Charles Darwin's theory of evolution. *Studies in History and Philosophy of Science Part A* 6, 159–181. [https://doi.org/10.1016/0039-3681\(75\)90019-9](https://doi.org/10.1016/0039-3681(75)90019-9)
- Sakamoto, M., Benton, M.J., Venditti, C., 2016. Dinosaurs in decline tens of millions of years before their final extinction. *Proceedings of the National Academy of Science* 113, 5036–5040. <https://doi.org/10.1073/pnas.1521478113>

- Signor, P.W., Lipps, J.H., 1982. Sampling bias, gradual extinction patterns, and catastrophes in the fossil record. *Geological Society of America Special Publication* 190, 291–296. <https://doi.org/10.1130/SPE190-p291>
- Stadler, T., Gavryushkina, A., Warnock, R.C.M., Drummond, A.J., Heath, T.A., 2018. The fossilized birth-death model for the analysis of stratigraphic range data under different speciation modes. *Journal of Theoretical Biology* 447, 41–55. <https://doi.org/10.1016/j.jtbi.2018.03.005>
- Tennant, J.P., Mannion, P.D., Upchurch, P., 2016a. Environmental drivers of crocodyliform extinction across the Jurassic/Cretaceous transition. *Proceedings of the Royal Society B: Biological Sciences* 283, 20152840. <https://doi.org/10.1098/rspb.2015.2840>
- Tennant, J.P., Mannion, P.D., Upchurch, P., 2016b. Sea level regulated tetrapod diversity dynamics through the Jurassic/Cretaceous interval. *Nature Communications* 7, 12737. <https://doi.org/10.1038/ncomms12737>
- Thagard, P.R., 1977. Darwin and Whewell. *Studies in History and Philosophy of Science Part A* 8, 353–356. [https://doi.org/10.1016/0039-3681\(77\)90026-7](https://doi.org/10.1016/0039-3681(77)90026-7)
- Uyeda, J.C., Zenil-Ferguson, R., Pennell, M.W., 2018. Rethinking phylogenetic comparative methods. *Systematic Biology* 67, 1091–1109. <https://doi.org/10.1093/sysbio/syy031>
- Whewell, W., 1840. *The Philosophy of the Inductive Sciences, Founded Upon their History*. J. W. Parker, London.