Phylotranscriptomic consolidation of the jawed vertebrate timetree

Iker Irisarri, Denis Baurain, Henner Brinkmann, Frédéric Delsuc, Jean-Yves Sire, Alexander Kupfer, Jörn Petersen, Michael Jarek, Axel Meyer, Miguel Vences, et al.

Phylotranscriptomic consolidation of the jawed vertebrate timetree

Iker Irisarri[1*], Denis Baurain[2], Henner Brinkmann[3], Frédéric Delsuc[4], Jean-Yves Sire[5], Alexander Kupfer[6], Jörn Petersen[3], Michael Jarek[7], Axel Meyer[1], Miguel Vences[8], Hervé Philippe[9,10,*]

[1] Chair in Zoology and Evolutionary Biology, Department of Biology, University of Konstanz, Universitätsstr. 10, 78464 Konstanz, Germany

[2] InBioS – Eukaryotic Phylogenomics, Department of Life Sciences and PhytoSYSTEMS, University of Liège, Chemin de la Vallée 4, 4000 Liège, Belgium

[3] Leibniz-Institut DSMZ – German Collection of Microorganisms and Cell Cultures, Inhoffenstraße 7 B, 38124 Braunschweig, Germany

[4] Institut des Sciences de l'Evolution, UMR 5554, CNRS, IRD, EPHE, Université de Montpellier, Place Eugène Bataillon, 34095 Montpellier, France.

[5] Institut de Biologie Paris-Seine, UMR7138, Sorbonne Universities, Quai Saint Bernard, 75005 Paris, France

[6] Department of Zoology, Stuttgart State Museum of Natural History, Rosenstein 1, 70191 Stuttgart, Germany

[7] Department of Genome Analytics, Helmholtz Centre for Infection Research, Inhoffenstr. 7, 38124 Braunschweig, Germany

[8] Zoological Institute, Braunschweig University of Technology, Mendelssohnstr. 4, 38106 Braunschweig, Germany

[9] Centre for Biodiversity Theory and Modelling, UMR CNRS 5321, Station of Theoretical and Experimental Ecology, 2 route du CNRS, 09200 Moulis, France

[10] Departement de Biochimie, Université de Montréal, Succursale Centre-Ville, Montréal, Quebec H3C3J7, Canada

* Corresponding authors: irisarri.iker@gmail.com, herve.philippe@sete.cnrs.fr

Key Words: cross-validation, jackknifing, Gnathostomata, molecular dating, phylogeny, RNA-Seq, substitution rates, transcriptome

Abstract

Phylogenomics is extremely powerful but introduces new challenges as no agreement exists on "standards" for data selection, curation and tree inference. We use jawed vertebrates (Gnathostomata) as model to address these issues. Despite considerable efforts in resolving their evolutionary history and macroevolution, few studies have included a full phylogenetic diversity of gnathostomes and some relationships remain controversial. We tested a novel bioinformatic pipeline to assemble large and accurate phylogenomic datasets from RNA sequencing and find this phylotranscriptomic approach successful and highly cost-effective. Increased sequencing effort up to ca. 10Gbp allows recovering more genes, but shallower sequencing (1.5Gbp) is sufficient to obtain thousands of full-length orthologous transcripts. We reconstruct a robust and strongly supported timetree of jawed vertebrates using 7,189 nuclear genes from 100 taxa, including 23 new transcriptomes from previously unsampled key species. Gene jackknifing supports the robustness of our tree and allows calculating genome-wide divergence times by overcoming gene sampling bias. Mitochondrial genomes prove insufficient to resolve the deepest relationships due to limited signal and among-lineage rate heterogeneity. Our analyses emphasize the importance of large curated nuclear datasets to increase the accuracy of phylogenomics and provide a reference framework for the evolutionary history of jawed vertebrates.

Introduction

Understanding the evolutionary relationships among organisms is a prerequisite for any biological study aiming at explaining key processes such as adaptive radiations or evolutionary convergences. Evolutionary relatedness is generally represented with phylogenetic trees, which need to be robust and accurate if one aims at obtaining credible macroevolutionary inferences. In the last decade, genome-scale datasets (phylogenomics) have revolutionized molecular phylogenetics thanks to their ability to yield precise estimates of phylogeny and more precise divergence times by reducing sampling error, one of the major hurdles in the pre-genomics era. Several methodologies have been used to generate raw data for phylogenomics using high-throughput sequencing. But besides the obvious advantages, phylogenetic inference based on genomic data poses numerous challenges. For the assembly of genome-scale datasets, these include the removal of contaminants (from symbionts, pathogens or food items in the original sample or introduced during processing steps such as human DNA or sample cross-contaminations), misalignments due to erroneous sequence stretches (often produced by sequencing and annotation errors in low-coverage genome assemblies), the effective detection and removal of paralogs, and the presence of large amounts of missing data, often aggravated by the difficulty of identifying orthology in only partially assembled transcripts.. Paralogy in particular can have dramatically detrimental effects on phylogenomic analyses[1], but the robustness of tree topology to the inclusion of paralogs is generally not evaluated.

Because phylogenomics relies on hundreds or thousands of genes and taxa, manual data curation has become unfeasible and automatic solutions need to be devised. Phylogenomic analyses have generally relied on pooling evidence from multiple genes by concatenation or used “summary” coalescent-based gene-tree/species tree methods. The size of genomic datasets also makes phylogenomics more sensitive to model misspecification (systematic error), which often translates into long-branch attraction problems[2]. Systematic error may be

reduced with complex mixture models, but their application to large-scale phylogenomic matrices can sometimes become computationally intractable. In addition, phylogenomic alignments are known to inflate non-parametric bootstrap support values and Bayesian clade posterior probabilities, a precision not always accompanied by increased accuracy, thus rendering the interpretation of these support metrics difficult.

The above challenges regarding the quality of the data and the robustness of analytical approaches need to be carefully taken into account in order to produce reliable estimates of both phylogeny and divergence times. Jawed vertebrates (Gnathostomata) represent a good system to benchmark these challenges because of the availability of genomic data for many species but the remarkable absence of several species with key phylogenetic positions, and the relatively good knowledge of their phylogeny except for some nodes that were controversial. In addition, jawed vertebrates are among the best-studied organisms and include astonishing examples of convergent evolution (e.g., flight, echolocation, limb loss) and prominent instances of classic paraphyletic taxa such as "fishes" or "reptiles". Biologists have long been interested in understanding the evolutionary relationships among jawed vertebrates, first using morphological characters and later with sequence data. Molecular phylogenies have greatly contributed towards shaping the jawed vertebrate tree, in many instances corroborating classical morphology-based classifications, but sometimes establishing novel hypotheses such as the close relationship of turtles with crocodiles and birds. Studies relying on mitochondrial genomes (mitogenomes) have resolved several controversial issues[3], but also recovered some unorthodox relationships[4]. Earlier molecular studies based on multiple genes obtained by classical Sanger-sequencing approaches have generally been limited by the number of genes or taxa, and were generally restricted to particular lineages such as ray-finned fishes[5], amphibians[6], squamate reptiles[7], mammals[8] or birds[9]. With the rise of genome-scale molecular datasets, it became possible to use ever larger datasets in an attempt at solving the relationships in the Tree of Life, and many nodes of the jawed vertebrate tree have been

103 confirmed by phylogenomic analyses based on datasets obtained by second-generation
104 sequencing and typically focusing on particular gnathostome clades[10-17]. Despite this growing
105 consensus, some phylogenomic studies have also challenged important relationships, such as
106 the monophyly and internal relationships of amphibians[18] or the position of turtles[19],
107 demonstrating that crucial aspects of the jawed vertebrate tree still require careful attention.
108 Further evolutionary relationships also remain controversial because of incongruence among
109 molecular phylogenies or with morphological evidence, such as the close relationship of
110 iguanian lizards with snakes[20-22] or the relationships among tongueless frogs[23,24].
111 Convincingly resolving difficult nodes requires more than just a large number of genes, and
112 instead a focus is needed on carefully avoiding and removing contaminations and errors in the
113 data, and avoiding model misspecifications[25].
114 Since their origin in the Ordovician (~470 Mya), jawed vertebrates have diversified into
115 lineages with markedly different morphologies and life histories, including hyperdiverse
116 radiations such as spiny-rayed fishes, birds, modern frogs (Neobatrachia) and placental
117 mammals. As an appealing hypothesis, the main diversification bursts of these hyperdiverse
118 radiations have been proposed to coincide with the Cretaceous-Paleogene boundary[5,15], but
119 due to uncertainties in timetree reconstruction and methodological disputes on molecular
120 dating, this hypothesis remains contentious, especially for mammals[26-28].
121 Here, we use a phylotranscriptomic approach to reconstruct the backbone of the jawed
122 vertebrate tree based on a dataset of unprecedented size composed of 7,189 genes for 100
123 species representing all main gnathostome lineages (a total of 3,791,500 aligned amino acid
124 positions). The dataset includes 23 newly generated transcriptomes from previously
125 underrepresented clades occupying key phylogenetic positions, particularly early-branching
126 ray- and lobe-finned fishes, lungfishes, amphibians and squamate reptiles. We devised a novel
127 bioinformatic pipeline to assemble the largest and most informative dataset ever analysed for
128 vertebrates (Supplementary Fig. 1) while focusing on the comprehensive removal of

129 contaminants and paralogs. This dataset is subjected to thorough phylogenetic and molecular
130 dating analyses. We present a strongly supported phylogenetic hypothesis, which is fossil-
131 calibrated to yield robust divergence time estimations, thus providing a reference framework
132 for the evolutionary history of jawed vertebrates.
133
134 **Results and discussion**
135 Phylotranscriptomic pipeline to assemble clean datasets
136 We developed a new bioinformatic pipeline (Supplementary Fig. 2) to assemble an
137 informative and "clean" genome-scale dataset of jawed vertebrates using genome and
138 transcriptome sequence data. For this study we collected RNA-Seq data for 23 previously
139 unsampled gnathostome species representing key lineages. Sequencing effort for the new
140 transcriptomes varied considerably among species (total sequenced base pairs ranged from 1.5
141 to 26 Gbp; Fig. 1 and Supplementary Table 1) and it correlated positively with (i) the average
142 length of reconstructed transcripts ($r=0.78$; $p=8.207x10^{-6}$), (ii) transcriptome completeness,
143 measured as the proportion of recovered core vertebrate genes[29] ($r=0.78$; $p=6.173x10^{-6}$) and
144 (iii) the total number of amino acids in final phylogenomic datasets ($r=0.82$ $p=0.00066$)
145 (Supplementary Table 2). Despite considerable differences in sequencing effort, all
146 transcriptomes were relatively complete (58.8 to 100% of the 233 core vertebrate genes were
147 recovered; Fig. 1) and thousands of genes readily usable for phylogenomics were
148 reconstructed (2,274 to 13,642 high-coverage genes per species, measured as human peptides
149 at ≥70% length coverage; Fig. 1). Hence, deeper sequencing increased the completeness of
150 transcriptome assemblies and the number of genes and amino acid positions in final
151 alignments. Nevertheless, this tendency stabilized at approximately 10Gb of total data (e.g.,
152 50 Mio. 2x100bp reads), after which a higher sequencing effort did not significantly increase
153 the above performance metrics ($r<0.5$ and $p>0.05$ for correlations for transcriptomes with
154 >10Gbp of total data; Supplementary Table 2). Interestingly, genes missing in final

155 phylogenomic matrices were essentially not different in species with shallow or deeply
156 sequenced transcriptomes (assessed by GO enrichment tests with FDR<0.05 against the
157 annotated set of 7,189 genes, run in Blast2GO, which suggests that sequencing effort does not
158 significantly bias the types of genes present in final alignments.
159 The new bioinformatic pipeline established herein warranted the high quality of alignments
160 by addressing key issues in data integrity[25], including several steps to minimize possible
161 contaminations, resolution of paralogy, masking of misalignments, and minimizing missing
162 data. During decontamination steps, BLAST similarity searches were used to identify
163 potentially contaminant sequences from non-vertebrates and human sequences (in this latter
164 case requiring high-identity at the nucleotide level). To remove any remaining contamination,
165 we devised a sensitive protocol that identifies extremely long branches estimated on a fixed
166 reference tree to flag possibly erroneous sequences, which are then removed. Per-sequence
167 missing data was minimized by merging conspecific sequences (typically overlapping
168 partially reconstructed transcripts) with SCaFoS[30], and unreliably aligned regions discarded.
169 A new HMM-based tool was used to mask erroneous sequence stretches typically produced
170 by frame shifts in ORFs or incorrect structural annotation. We implemented an innovative
171 paralog-splitting pipeline that specifically targets distant paralogs (those particularly
172 problematic for resolving the backbone of the tree) and further assessed the effect in the tree
173 stability of including various levels of deep paralogy in the datasets. In order to do that, genes
174 were classified into three sets that contained zero (NoDP), one (1DP) and two (2DP) deep
175 paralogs (i.e., duplication events predating the origin of major jawed vertebrate lineages),
176 which were then concatenated into three datasets that were separately analysed: NoDP (4,593
177 genes, 1,964,439 amino acids, 32% missing data), 1DP (1,162 genes, 668,132 amino acids,
178 36% missing data), and 2DP (1,434 genes, 1,158,929 amino acids, 39% missing data).
179
180 Backbone phylogeny of jawed vertebrates

181 The phylogeny was estimated based on concatenated alignments by (i) maximum
182 likelihood (ML) under the site-homogeneous LG+F+Γ and GTR+Γ models and 100
183 bootstrap replicates in RAxML and (ii) Bayesian inference (BI) under the more realistic site-
184 heterogeneous CAT+Γ model in PhyloBayes. For computational tractability of large datasets
185 under complex and computationally expensive models and to further assess the effect of gene
186 sampling, BI analyses were performed on 100 gene jackknife replicates (~50,000 amino acids
187 and ~180 genes per replicate), which were summarized in a final majority-rule consensus tree.
188 Gene jackknifing measures the repeatability of the phylogenetic relationships across genes,
189 which are randomly sampled without replacement from the total set of genes[31]. We employed
190 gene jackknife proportions (GJP) as a stringent test for the robustness of the obtained
191 relationships because they were estimated under the more realistic CAT model and based on
192 virtually independent gene sets, each containing ~2.5% of the total alignment, as compared to
193 the ~66% of the total alignment used in non-parametric bootstrapping. In addition, we carried
194 out coalescent-based species tree analyses with ASTRAL-II with 100 replicates of multi-locus
195 bootstrapping on the three nuclear datasets separately. All phylogenetic analyses of the
196 paralog-free dataset (NoDP), including BI (Fig. 2a) and ML on the concatenated super-matrix
197 and species tree analyses (Supplementary Figs. 3-5), reconstructed fully resolved and almost
198 identical trees that were highly supported: 88% and 95% of the nodes in Fig. 2 received
199 respectively full (100%) or high (>75%) GJP. All major uncontroversial vertebrate clades
200 were recovered with full support: cartilaginous fishes (Chondrichthyes) were the sister group
201 of bony fishes (Osteichthyes), including ray-finned (Actinopterygii) and lobe-finned
202 (Sarcopterygii) fishes; within sarcopterygians, tetrapods (Tetrapoda) were monophyletic and
203 encompassed amphibians (Lissamphibia), mammals (Mammalia), turtles (Testudines), birds
204 (Aves), crocodiles (Crocodylia), lepidosaurian reptiles (Lepidosauria) and snakes (Serpentes).
205 Even using relatively small alignments of ~5,000 amino acids (Fig. 2b, Supplementary Table
206 3), all the above nodes were recovered with strong support. In fact, these uncontroversial

nodes were also recovered by a large proportion of single-gene trees (58-96% of the genes; Supplementary Table 4) though with varying levels of support.

In contrast, some of the relationships that remained hotly discussed during the past decades were not unambiguously recovered by single genes nor by relatively small-sized gene jackknife replicates (Fig. 2b and Supplementary Tables 3, 4). Thanks to the use of a larger dataset, our analyses however effectively resolved these controversial relationships with maximum support (Fig. 2a). (i) Lungfishes (Dipnoi) were the sister group of tetrapods, in agreement with the latest phylogenomic results[12,32], and topology tests rejected the alternative hypothesis where coelacanth and tetrapods are sister taxa[33] (Supplementary Table 5). (ii) Amphibians (Lissamphibia) were monophyletic and salamanders (Caudata) were the sister group of frogs (Anura) to the exclusion of caecilians (Gymnophiona) (Batrachia[34,35]). Both the paraphyly of amphibians and the alternative sister group of caecilians and salamanders (Procera[18]) were rejected by topological tests. (iii) Turtles were the sister group of crocodiles and birds (Archosauria), in agreement with the majority of previous phylogenomic studies[10,11] and the latest morphological evidence[36]. Topology tests rejected the traditional view of turtles as primarily anapsids (early-branching within "reptiles") as well as possible sister group relationships with either lepidosaurians or crocodiles[18]. (iv) The earliest offshoot within salamanders was *Andrias* (Cryptobranchidae) plus *Hynobius* (Hynobiidae)[34] and the alternative position of *Siren* (Sirenidae) as the earliest-branching salamander clade[37] was statistically rejected. (v) Lastly, our BI tree supports a close relationship between snakes and iguanian and anguimorph lizards (*Elgaria*) (Toxicofera[7]).

Only 4 out of 98 nodes in our phylogeny received relatively low support (<75% GJP; Fig. 2) and we consider these nodes in need of further confirmation. Besides relationships within crown-group iguanians and turtles, this applies to the sister group relationship between anguimorph (*Elgaria*) and iguanian lizards which was sensitive to the use of alternative models (GTR+Γ and LG+Γ+F in ML; Supplementary Figs. 3, 4) or the inclusion of deep

paralogy (BI on the 1DP and 2DP datasets; Supplementary Figs. 6, 7), which recovered anguimorphs as the sister group of snakes (rejected however by topology tests; Supplementary Table 5). In agreement with Fig. 2a, coalescent-based analyses reconstructed an anguimorph + iguanian clade, which was robust to the inclusion of deep paralogy (Supplementary Figs. 5, 8, 9). In addition, only moderate support (75% GJP) was recovered for the controversial position[14,17] of armadillo (Xenarthra) plus elephant (Afrotheria) sister to the remaining placental mammals (Atlantogenata[13,16]), in agreement with coalescent analyses (Supplementary Fig. 5), and the two alternative resolutions were rejected by topology tests (Supplementary Table 5). These problematic nodes correspond to fast radiations whose resolution requires extended taxon sampling in addition to accounting for incomplete lineage sorting. Our study minimized the possibility of model misspecification by using also complex evolutionary models and assessing the stability of tree topology to the effect of gene sampling and deep paralogy. For definitively resolving the above nodes, we argue for a careful exploration using suitable methodology.

Robustness to gene sampling: size of gene jackknife replicates and gene trees

The use of gene jackknifing (100 replicates of ~50,000 amino acids each) allowed recovering an almost fully supported tree and resolving a number of controversial relationships. To explore the stability of the nodes in our tree and assess the amount of data required to recover them, we further analysed four sets of 100 gene jackknife replicates of increasing total length (~2,500, ~5,000, ~10,000 and ~25,000 amino acids) under ML. Relatively short replicates (~2,500 amino acids) recovered 33% and 76% of the nodes with full and high GJP, respectively (Fig. 2b). Increasing alignment length to 25,000 amino acids led to an increase of 47% of fully supported nodes (Fig. 2b; Supplementary Table 3). The relationships among the earliest-branching salamander lineages (*Andrias, Hynobius, Siren*) were particularly unstable and required long replicates (~50,000 amino acids) to be recovered

with strong support. Gene length positively correlated with the proportion of final-tree bipartitions, more strongly for deep (>150 myr; r=0.21, $p<2.2 \times 10^{-16}$) than for recent (<150 myr; r=0.13, $p<2.2 \times 10^{-16}$) relationships, suggesting that longer genes correctly resolve more ancient nodes.

Mitogenomes and limits to phylogenetic resolution

To assess the phylogenetic resolution power of mitogenomes, we assembled a mitogenomic dataset matching the species in our nuclear datasets. Mitogenomic trees inferred by both ML and BI (Supplementary Figs. 10-17) correctly recovered some major clades with strong support, but failed to recover well-established relationships such as the monophyly of ray-finned and lobe-finned fishes or the sister-group position of platypus to all other mammals[4], even after excluding the fastest evolving taxa and using complex mixture models to minimize long-branch attraction artefacts (Supplementary Figs. 14-15). Besides stochastic error due to limited alignment length, these incongruences most likely originate from long-branch attraction (despite using sophisticated models such as CAT-GTR), suggesting that mitogenomes are inadequate for resolving ancient divergences (>400 Mya) using currently available models of sequence evolution. The correlation between nuclear and mitochondrial rates is low (r=0.35; $p<2.49 \cdot 10^{-5}$; Supplementary Fig. 18) but still higher than expected from random datasets ($r=0.13 \pm 0.08$; $p> 0.05$ averaged for 100 replicates). Hence, commonly assumed determinants of substitution rates, such as demography (population size changes, bottlenecks) or life history traits (body size, metabolic rate, generation time, genome size have to some extent influenced both genomes similarly, but other factors must be invoked to explain the observed rate disparity between the two genomes at many branches (Supplementary Fig. 18). These might include clade-specific variation in mitochondrial effective population sizes, genome-specific mutation rate, or acceleration of mitochondrial genes due to selection shifts in respiratory function.

No general relationship among evolutionary rates, species diversity and genome size

Comparing 44 main clades of jawed vertebrates of ages >150 myr confirmed enormous differences in species diversity, from 1 to 31,826 species (Supplementary Table 6). Species diversity was not overall correlated with substitution rate ($r=0.18$, $p=0.25$) nor were higher rates significantly associated with higher species diversity in a sister group approach (Sign test, $p=0.13$). Our dataset includes the entire range of genome sizes in vertebrates (from 0.4 pg in pufferfish to 109 pg in lungfishes). Yet we found no association of genome size with evolutionary rate or species diversity ($r=-0.28$, $p=0.061$ and $r=-0.13$, $p=0.44$, respectively). Previous studies have also suggested that genome size might be associated with indels in coding regions[38], but we detected no significant correlation, neither within conserved ($r=0.1983$, $p=0.0722$) nor within variable coding regions ($r=0.0533$, $p=0.6325$) as defined by BMGE. These correlation analyses were confirmed by a Bayesian joint modelling of the above traits with parameters of the evolutionary process at the sequence level (see Supplementary Table 7).

Divergence times of major lineages of jawed vertebrates

Genome-scale datasets have been shown to produce more precise and accurate divergence time estimates[39], but this ultimately depends on the use of realistic evolutionary and clock models that appropriately account for among-lineage heterogeneities[40] and of multiple calibration intervals whose uncertainty and internal consistency is accounted for[41,42]. We applied an auto-correlated lognormal relaxed clock model and best-fitting sequence evolution model (CAT-GTR) to estimate genome-wide divergence times, averaged over 100 gene jackknife replicates. We used a conservative approach to setting calibrations, starting from multiple well-established calibrations with solid paleontological evidence and used conservative intervals to account for dating and phylogenetic uncertainty[43] (Supplementary

311 Table 8). On top of that, the internal congruence among these calibrations was verified
312 through extensive cross-validation procedures in order to remove any poorly performing
313 calibration, either examining the performance of single calibrations[41] or removing one
314 calibration at a time to check the congruence between estimated ages and priors[42]. The
315 performance of each calibration scheme (named C16 and C30) derived from the above cross-
316 validation strategies was assessed in independent dating analyses with a test dataset (a subset
317 of the 14,352 most complete amino acid positions from the NoDP dataset that was
318 computationally tractable with PhyloBayes). Both schemes produced largely congruent
319 divergence times, but C16 yielded more reasonable dates within turtles, frogs, neoavian birds,
320 modern frogs (overestimated in C30 if compared with previous data; www.timetree.org) or
321 iguanian squamates and snakes (underestimated in C30) (Supplementary Table 9).
322 To estimate genome-wide divergence times, we calculated averaged divergence times (and
323 conservative 95% credibility intervals; CrI) across 100 timetrees based of jackknife sampling
324 of ~15,000 positions and the more stringently cross-validated C16 calibration scheme (Fig. 3).
325 The genome-averaged timetree places the divergence of cartilaginous, ray-finned and lobe-
326 finned fishes in the Ordovician, between 458 (CrI: 465–438) to 449 (462–431) Mya. The first
327 split within lobe-finned fishes occurred in the Silurian *ca*. 427 (444–413) Mya and lungfishes
328 separated from tetrapods in the Early Devonian *ca*. 412 (419–408) Mya. The split between
329 amphibians and amniotes occurred in the Early Carboniferous *ca*. 346 (351–333) Mya and the
330 three amphibian orders separated during the Carboniferous from 325 (338–307) to 315 (332–
331 293) Mya, as did synapsids (mammals) and diapsids (turtles, archosaurs and lepidosaurs) *ca*.
332 317 (330–299) Mya. The origins of the main sauropsid groups, i.e., turtles, crocodiles, birds,
333 squamates and tuatara (*Sphenodon*), took place in the Permian from 294 (313–273) to 259
334 (288–226) Mya. The crown diversification of extant frogs, salamanders and caecilians
335 occurred in the late Triassic to Early Jurassic between 213 (270–151) to 186 (231–153) Mya,
336 almost simultaneously with the crown splits within squamates *ca*. 204 (228–183) Mya,

337 cryptodiran turtles *ca*. 202 (243–159) Mya, pleurodiran turtles *ca*. 191 (248–116) Mya, and
338 therian mammals *ca*. 214 (257–169) Mya.
339 Estimated divergences are generally in line with previous time-calibrated phylogenies
340 using different dating methodologies, molecular data and calibrations, particularly for the
341 deepest splits in the backbone[44,45] as well as divergences within amphibians[6], squamates[46],
342 snakes[47] and placental mammals[8]. Estimated ages for crown-groups of cartilaginous and ray-
343 finned fishes are younger compared to previous analyses[48,49], which is likely caused by the
344 removal of incongruent calibrations in the C16 scheme (the C30 scheme produced estimates
345 more similar to previous studies for these groups; see Supplementary Table 9). The younger
346 age of cartilaginous fishes, however, is consistent with recent paleontological analyses[50].
347 Compared to previous time-calibrated phylogenies, we obtain older divergences for turtles[51]
348 and birds[15] but our estimates are in line with the ages of recently discovered fossils of stem
349 turtles[36] and an ornithuromorph bird that pushes back the origin of the group to at least 130.7
350 Mya[52]. The Cretaceous-Paleogene boundary (67 Mya) in our tree is not associated with a
351 notable concentration of divergences, but our dataset does not capture the crown
352 diversification of several species-rich taxa that might have occurred in this period, such as
353 spiny-ray fishes, modern birds (Neoaves), boreoeutherian mammals, ranoid frogs, gekkonid
354 geckos, or skinks. We support a diversification of placental mammals prior to the Cretaceous-
355 Paleogene boundary *ca*. 102 (139–73) Mya, in agreement with most previous molecular and
356 macroevolutionary studies[8,39].
357
358 Reliability of phylogenomic analyses
359 Inferring phylogenies can be difficult, particularly in the presence of ancient or closely
360 spaced speciation events, and the use of genome-scale datasets poses additional challenges
361 related to poor data quality and more importantly systematic error[25]. In principle, the jawed
362 vertebrate phylogeny is a solvable problem, being devoid of excessively old divergences and

mostly long internal branches (Fig. 2a). It thus represents a good benchmark to test the abovementioned challenges. Yet, poor data quality[18] (Supplementary Fig. 1) can lead to incorrect results (e.g., non-monophyly of amphibians, misplacement of turtles). We adopt a phylotranscriptomic approach to assemble an alignment of >7,000 genes for 100 species with rigorous quality controls. The quality and resolving power of our NoDP dataset are higher than those of previous studies, including Fong et al.[18] and the most comprehensive dataset analysed to date[12], with 70% vs. 17% and 61% mean congruence respectively for the two datasets, measured as the proportion of final-tree bipartitions recovered by single genes. The higher congruence of NoDP persisted after correcting for gene length (Supplementary Fig. 1). We thus confirm that RNA-Seq is a cost-effective method to anchor phylogenomic analyses, which can result in robust fossil-dated trees, provided that careful data curation and appropriate analytical methods are used. Moreover, we show that gene jackknifing allows stringently testing phylogenetic relationships and overcoming the limitations and possible biases of small datasets that aim to represent the entire genome, and that phylogenomics is resilient to limited levels of deep paralogy, as long as a large number of genes (>1,000) are used and internal branches are relatively long. In such cases, realistic models allow recovering correct phylogenetic hypotheses, even in the presence of extreme among-lineage evolutionary rate variation. In contrast, resolving closely spaced radiations, which were not targeted in this work, could require a detailed study with specific gene and taxon sampling[14] and testing the robustness to model misspecification[32]. Overall, our results highlight the importance of data quality in phylogenomics, as well as the application of realistic evolutionary and clock models, and the validation of calibrations in timetree estimation, both *a priori* (based on paleontological data) and *a posteriori* (cross-validation).

Materials and Methods

388 An extended description of our bioinformatic pipeline and detailed Materials and Methods are
389 available as Supplementary Information.
390
391 Assembly of phylogenomic datasets
392 New RNA-Seq data was generated for 23 gnathostome species using Illumina MiSeq
393 (2x250 bp) and HiSeq2000 (2x50 bp, 2x100 bp) technologies. Available RNA-Seq data were
394 downloaded from NCBI SRA. Transcriptomes were assembled *de novo* with Trinity or
395 MIRA. Species names and accession numbers are available in Supplementary Table S10.
396 Nuclear datasets were assembled using a new pipeline summarized in Supplementary Fig.
397 2. Briefly, proteomes of 21 vertebrate genomes (ENSEMBL) were grouped into ortholog
398 clusters and those not containing data for all major jawed vertebrate lineages were discarded.
399 The resulting 11,656 protein clusters were aligned and positions of unreliable homology
400 removed. To identify and resolve paralogy issues, we implemented a paralog-splitting
401 pipeline based on gene trees. The obtained 9,852 ortholog clusters were complemented with
402 new genomes and transcriptomes using the software Forty-Two
403 (https://bitbucket.org/dbaurain/42/). Several decontamination steps were carried out. Any
404 sequence contamination from non-vertebrates and human was detected by BLAST and
405 eliminated. We searched for cross-contamination that can arise during library preparation
406 using gene trees, and removed contaminants based on expression data. After eliminating
407 overlapping redundant sequences that were too divergent, we filtered out incomplete or short
408 sequences and alignments, leading to 7,687 genes. The paralogy splitting procedure was
409 repeated to resolve any paralogy caused by the addition of new species, and gene alignments
410 were classified into three datasets that contained zero (NoDP), one (1DP) and two or more
411 (2DP) deep paralogs. Sequence stretches with unusually low similarity (usually due to frame
412 shifts) were masked with HMMCleaner (R. Poujol) and alignments were trimmed. For each
413 gene, we used SCaFoS[30] to merge conspecific sequences and resolve putative remaining

414 paralogy. A third decontamination step used extremely long branches estimated on a fixed
415 reference tree as proxy for contamination.
416 Mitochondrial datasets were assembled from mitogenomes available at NCBI with a taxon
417 sampling mirroring the nuclear datasets plus a few additional species to reduce long-branch
418 attraction artefacts expected in mitogenomic trees (Supplementary Table 11). The resulting
419 alignments consisted of 106 species (2,773 amino acid positions) and 95 species (2,866 amino
420 acid positions) after removing the fastest evolving species.
421
422 Phylogenetic inference
423 Concatenated nuclear gene sets (NoDP, 1DP and 2DP) were analysed separately using ML
424 with RAxML v.8[53] under LG+F+Γ and GTR+Γ models and BI with PhyloBayes MPI v1.5[54]
425 under the better fitting CAT+Γ model (selected after 10-fold cross-validation). The Bayesian
426 consensus tree was calculated from 100 post-burnin tree collections, each from gene jackknife
427 replicates of ~50,000 amino acid positions. Convergence was verified with the diagnostic
428 tools of PhyloBayes. Branch support was computed from 100 bootstrap pseudo-replicates in
429 ML, and from gene jackknife proportions (GJP) in BI. To assess the robustness to gene
430 sampling, we analysed by ML gene jackknife replicates of *ca*. 2,500, 5,000, 10,000 and
431 25,000 aligned positions under the LG+Γ model. Coalescent analyses were run in ASTRAL-
432 II v.4.10.12 using ML gene trees as input (estimated under best-fit models in RAxML) and
433 node stability was assessed as local posterior support and 100 replicates of multi-locus
434 bootstrapping.
435 The mitochondrial datasets were analysed by ML under MTREV+Γ and GTR+Γ models,
436 and by BI under CAT+Γ and CAT-GTR+Γ models.
437
438 Molecular dating

Divergence times were estimated in PhyloBayes v.4.1 using best-fit CAT-GTR+Γ and auto-correlated lognormal clock models (selected after 10-fold cross-validation), a birth-death prior on divergence times and 30 calibration points with uniform priors and soft bounds (see Supplementary Table 8). After cross-validation procedures (see SI Materials and Methods), we applied the C16 and C30 calibration sets to compute timetrees based on a subset of 14,352 amino acid positions from NoDP (two independent chains). To estimate genome-wide divergence times, we estimated 100 timetrees from 100 gene jackknife replicates of ~15,000 amino acids from the NoDP dataset, using the most stringent C16 calibration scheme. Divergence times were averaged and conservative 95% credibility intervals (CrI) calculated as the absolute maximum and minimum values of 95% confidence intervals across 100 timetrees.

Nuclear and mitochondrial rates

Substitution rates were measured as branch lengths optimized under CAT+Γ and a reference tree (Fig. 2a) in PhyloBayes, independently for the nuclear (NoDP) and mitochondrial datasets, both pruned to a common subset of 78 species. Correlation between mitochondrial and nuclear rates was assessed by Pearson's correlation among all pairs of internal and terminal branches. We simulated 100 random alignments characterized by the amino acid proportions of either mitochondrial or nuclear datasets, then branch lengths were optimized on a reference tree and rates correlated as above.

Association of life history traits and molecular features

We estimated Pearson's correlation after correcting for phylogenetic non-independence among the following life history traits and molecular features: (i) genome size (retrieved from www.genomesize.com) versus number of gaps in either conserved or variable gene regions

(defined by BMGE on untrimmed gene alignments), (ii) genome size versus nuclear substitution rate, and (iii) substitution rate versus species diversity (tabulated from the literature), for 44 lineages divided by an ad-hoc cut-off of >150 myr defined to capture sister groups characterized by obvious differences in species diversity. Nuclear substitution rates and species diversity were also compared in a sister group approach, assessing by non-parametric Sign test whether higher substitution rates (tested by relative-rate tests) were associated with higher species diversity. We further used a Bayesian joint modelling to study the correlation between substitution rates, genome size and the number of gaps in conserved and variable gene regions.

Acknowledgements

We thank the following people for tissue samples: T. Ziegler (*Andrias*), M. Hasselmann (*Neoceratodus*) and O. Guillaume (*Calotriton*, *Proteus*) and V. Michael for technical assistance. N. Galtier kindly provided access to the *Salamandra* transcriptome. We acknowledge the use of computational resources from the University of Konstanz (HPC2), the Institute of Physics of Cantabria IFCA-CSIC (Altamira), the Fédération Wallonie-Bruxelles (Tier-1; funded by Walloon Region, grant #1117545), the Consortium des Équipements de Calcul Intensif (CÉCI; funded by the F.R.S.-FNRS, grant #2.5020.11) and the Réseau Québecois de Calcul de Haute Performance for computational resources. II was supported by the Alexander von Humboldt Foundation (1150725) and the European Molecular Biology Organization (EMBO ALTF 440-2013). Further support came from the University of Konstanz (II), the Deutsche Forschungsgemeinschaft (DFG) (AM), the European Research Council (ERC Advanced Grant "Genadapt" No. 273900 to AM), and the Agence Nationale de la Recherche (TULIP Laboratory of Excellence ANR-10-LABX-41 to HP and ANR-12-BSV7-020 project "Jaws" to FD and JYS). This is contribution ISEM 2017-106 of the Institut des Sciences de l'Evolution de Montpellier.

Ethics statement and data availability

Animal experiments conformed to the European Parliament and council of 22/09/2010 (Directive 2010/63/EU) and the French Rural Code (Articles R214-87 to R214-137, decree No. 2013-118 of 01/02/2013). Experiments performed in France were authorized by the certificate No. 75-600. New RNA-Seq data are available at the SRA (Supplementary Table 10) and phylogenetic datasets, trees and custom scripts in Dryad <to be provided>.

Author contributions

MV, HP and II designed research. II, FD, JYS, AK, MJ, AM and MV contributed new data. II, DB, HB, MV and HP performed analyses. II, MV and HP drafted manuscript and all authors read and approved the final manuscript.

Competing financial interests

The authors declare no competing financial interests.

References

1 Brown, J. M. & Thomson, R. C. Bayes factors unmask highly variable information content, bias, and extreme influence in phylogenomic analyses. *Syst. Biol.* in press (2017) https://doi.org/10.1093/sysbio/syw101.

2 Jeffroy, O., Brinkmann, H., Delsuc, F. & Hervé, P. Phylogenomics: The beginning of incongruence? *Trends Genet.* **22**, 225-231 (2006).

3 Zardoya, R. & Meyer, A. The evolutionary position of turtles revised. *Naturwissenschaften* **88**, 193-200 (2001).

4 Janke, A., Erpenbeck, D., Nilsson, M. & Arnason, U. The mitochondrial genomes of the iguana (*Iguana iguana*) and the caiman (*Caiman crocodylus*): Implications for amniote phylogeny. *Proc. R. Soc. B-Biol. Sci.* **268**, 623-631 (2001).

5 Near, T. J. *et al.* Phylogeny and tempo of diversification in the superradiation of spiny-rayed fishes. *Proc. Natl. Acad. Sci. USA* **110**, 12738-12743 (2013).

6 Irisarri, I. *et al.* The origin of modern frogs (Neobatrachia) was accompanied by acceleration in mitochondrial and nuclear substitution rates. *BMC Genomics* **13**, 626 (2012).

7 Pyron, R. A., Burbrink, F. T. & Wiens, J. J. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol. Biol.* **13**, 93 (2013).

8 Meredith, R. W. *et al.* Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. *Science* **334**, 521-524 (2011).

9 Jetz, W., Thomas, G. H., Joy, J. B., Hartmann, K. & Mooers, A. O. The global diversity of birds in space and time. *Nature* **491**, 444-448 (2012).

10 Chiari, Y., Cahais, V., Galtier, N. & Delsuc, F. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biology* **10**, 65 (2012).

11 Crawford, N. G. *et al.* More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol. Lett.* **8**, 783-786 (2012).

12 Chen, M. Y., Liang, D. & Zhang, P. Selecting question-specific genes to reduce incongruence in phylogenomics: A case study of jawed vertebrate backbone phylogeny. *Syst. Biol.* **64**, 1104-1120 (2015).

13 Tarver, J. E. *et al.* The interrelationships of placental mammals and the limits of phylogenetic inference. *Genome Biol. Evol.* **8**, 330-344 (2016).

14 Romiguier, J., Ranwez, V., Delsuc, F., Galtier, N. & Douzery, E. J. P. Less is more in mammalian phylogenomics: AT-rich genes minimize tree conflicts and unravel the root of placental mammals. *Mol. Biol. Evol.* **30**, 2134-2144 (2013).

15 Jarvis, E. D. *et al*. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320-1331 (2014).

16 Song, S., Liu, L., Edwards, S. V. & Wu, S. Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. USA* **109**, 14942-14947 (2012).

17 Morgan, C. C. *et al*. Heterogeneous models place the root of the placental mammal phylogeny. *Mol. Biol. Evol.* **30**, 2145-2156 (2013).

18 Fong, J. J., Brown, J. M., Fujita, M. K. & Boussau, B. A phylogenomic approach to vertebrate phylogeny supports a turtle-archosaur affinity and a possible paraphyletic Lissamphibia. *PLoS ONE* **7**, e48990 (2012).

19 Lyson, T. R. *et al*. MicroRNAs support a turtle + lizard clade. *Biol. Lett.* **8**, 104-107 (2012).

20 Gauthier, J. A., Kearney, M., Maisano, J. A., Rieppel, O. & Behlke, A. D. B. Assembling the squamate tree of life: Perspectives from the phenotype and the fossil record. *Bull. Peabody Mus. Nat. Hist.* **53**, 3-308 (2012).

21 Townsend, T. M., Larson, A., Louis, E., Macey, J. R. & Sites, J. Molecular phylogenetics of Squamata: The position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. *Syst. Biol.* **53**, 735-757 (2004).

22 Zheng, Y. & Wiens, J. J. Combining phylogenomic and supermatrix approaches, and a time-calibrated phylogeny for squamate reptiles (lizards and snakes) based on 52 genes and 4162 species. *Mol. Phylogenet. Evol. Part B* **94**, 537-547 (2016).

23 Irisarri, I., Vences, M., San Mauro, D., Glaw, F. & Zardoya, R. Reversal to air-driven sound production revealed by a molecular phylogeny of tongueless frogs, family Pipidae. *BMC Evol. Biol.* **11**, 114 (2011).

24 Bewick, A. J., Chain, F. J. J., Heled, J. & Evans, B. J. The pipid root. *Syst. Biol.* **61**, 913-926 (2012).

25 Philippe, H. *et al*. Resolving difficult phylogenetic questions: Why more sequences are not enough. *PLoS Biol.* **9**, e1000602 (2011).

26 Bininda-Emonds, O. R. P. *et al*. The delayed rise of present-day mammals. *Nature* **446**, 507-512 (2007).

27 Springer, M. S. *et al*. Waking the undead: Implications of a soft explosive model for the timing of placental mammal diversification. *Mol*. *Phylogenet*. *Evol*. **106**, 86-102 (2017).
28 Phillips, M. J. Geomolecular dating and the origin of placental mammals. *Syst*. *Biol*. **65**, 546-557 (2016).
29 Hara, Y. *et al*. Optimizing and benchmarking de novo transcriptome sequencing: from library preparation to assembly evaluation. *BMC Genomics* **16**, 977 (2015).
30 Roure, B., Rodriguez-Ezpeleta, N. & Philippe, H. SCaFoS: A tool for selection, concatenation and fusion of sequences for phylogenomics. *BMC Evol*. *Biol*. **7**, S2 (2007).
31 Delsuc, F., Tsagkogeorga, G., Lartillot, N. & Philippe, H. Additional molecular support for the new chordate phylogeny. *Genesis* **46**, 592-604 (2008).
32 Irisarri, I. & Meyer, A. The identification of the closest living relative(s) of tetrapods: Phylogenomic lessons for resolving short ancient internodes. *Syst*. *Biol*. **65**, 1057-1075 (2016).
33 Zhu, M. & Schultze, H.-P. in *Major Events in Early Vertebrate Evolution: Paleontology, Phylogeny and Development* (ed P. E. Ahlberg) 289-314 (Taylor & Francis, 2001).
34 Pyron, R. A. & Wiens, J. J. A large-scale phylogeny of Amphibia including over 2800 species, and a revised classification of extant frogs, salamanders, and caecilians. *Mol*. *Phylogenet*. *Evol*. **61**, 543-583 (2011).
35 San Mauro, D., Vences, M., Alcobendas, M., Zardoya, R. & Meyer, A. Initial diversification of living amphibians predated the breakup of Pangaea. *Am*. *Nat*. **165**, 590-599 (2005).
36 Schoch, R. R. & Sues, H.-D. A Middle Triassic stem-turtle and the evolution of the turtle body plan. *Nature* **523**, 584-587 (2015).
37 Zhang, P. & Wake, D. B. Higher-level salamander relationships and divergence dates inferred from complete mitochondrial genomes *Mol*. *Phylogenet*. *Evol*. **53**, 492-508 (2009).
38 Gregory, T. R. Insertion–deletion biases and the evolution of genome size. *Gene* **324**, 15-34 (2004).
39 Dos Reis, M. *et al*. Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc*. *R*. *Soc*. *B-Biol*. *Sci*. **279**, 3491-3500 (2012).

40 Lepage, T., Bryant, D., Philippe, H. & Lartillot, N. A general comparison of relaxed molecular clock models. *Mol*. *Biol*. *Evol*. **24**, 2669-2680 (2007).

41 Near, T. J., Meylan, Peter A. & Shaffer, H. B. Assessing concordance of fossil calibration points in molecular clock studies: An example using turtles. *Am*. *Nat*. **165**, 137-146 (2005).

42 Sanders, K. L. & Lee, M. S. Y. Evaluating molecular clock calibrations using Bayesian analyses with soft and hard bounds. *Biol*. *Lett*. **3**, 275-279 (2007).

43 Warnock, R. C. M., Parham, J. F., Joyce, W. G., Lyson, T. R. & Donoghue, P. C. J. Calibration uncertainty in molecular dating analyses: There is no substitute for the prior evaluation of time priors. *Proc*. *R*. *Soc*. *B-Biol*. *Sci*. **282** (2014).

44 King, B. L. Bayesian morphological clock methods resurrect placoderm monophyly and reveal rapid early evolution in jawed vertebrates. *Syst*. *Biol*. in press (2017) https://doi.org/10.1093/sysbio/syw107.

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

46 Jones, M. *et al*. Integration of molecules and new fossils supports a Triassic origin for Lepidosauria (lizards, snakes, and tuatara). *BMC Evol*. *Biol*. **13**, 208 (2013).

47 Hsiang, A. Y. *et al*. The origin of snakes: Revealing the ecology, behavior, and evolutionary history of early snakes using genomics, phenomics, and the fossil record. *BMC Evol*. *Biol*. **15**, 1-22 (2015).

48 Inoue, J. G. *et al*. Evolutionary origin and phylogeny of the modern holocephalans (Chondrichthyes: Chimaeriformes): a mitogenomic perspective. *Mol Biol Evol* **27**, 2576-2586 (2010).

49 Dornburg, A., Townsend, J., Friedman, M. & Near, T. J. Phylogenetic informativeness reconciles ray-finned fish molecular divergence times. *BMC Evol*. *Biol*. **14**, 169 (2014).

50 Coates, M. I., Gess, R. W., Finarelli, J. A., Criswell, K. E. & Tietjen, K. A symmoriiform chondrichthyan braincase and the origin of chimaeroid fishes. *Nature* **541**, 208-211 (2017).

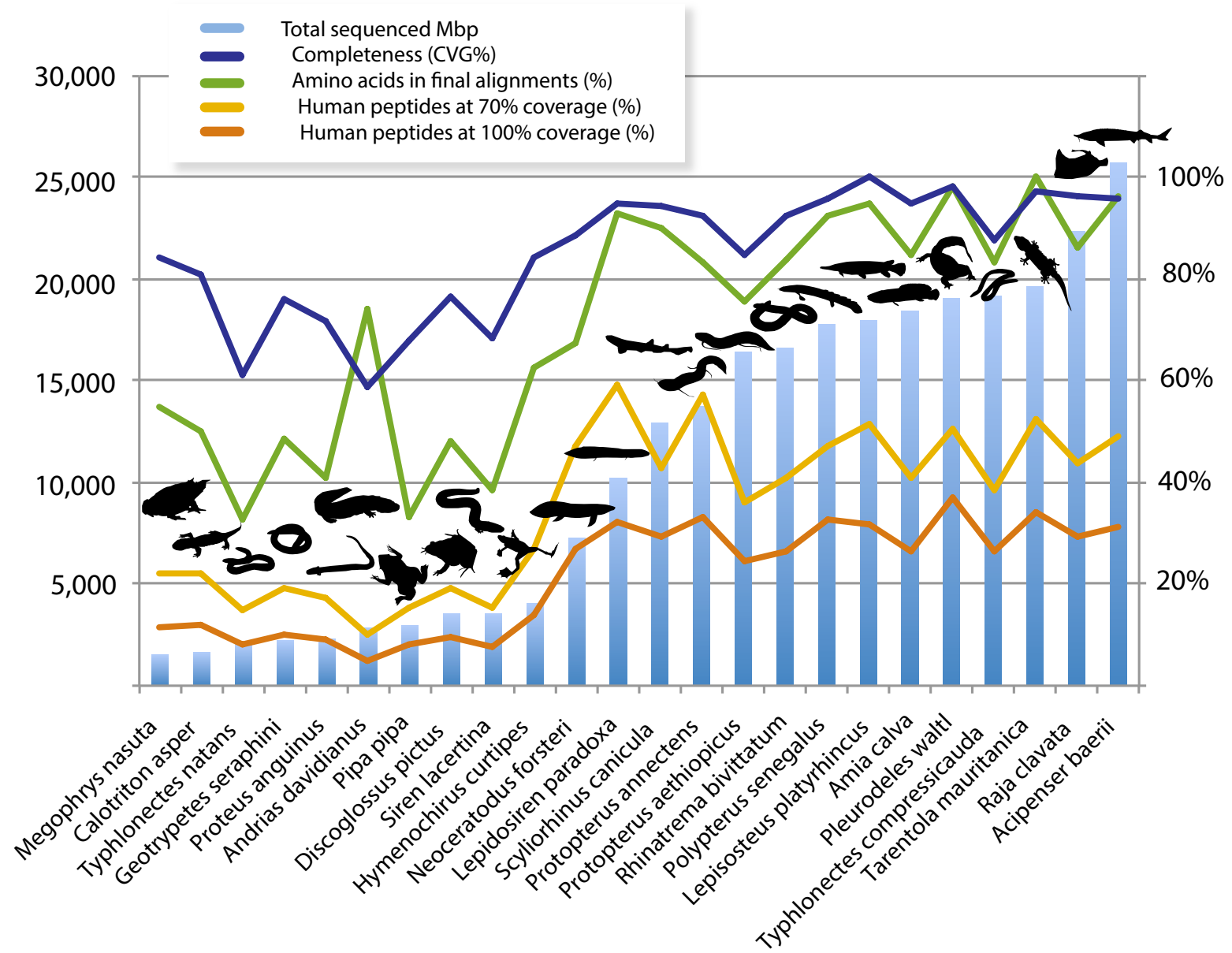
51 Lourenço, J. M., Claude, J., Galtier, N. & Chiari, Y. Dating cryptodiran nodes: Origin and diversification of the turtle superfamily Testudinoidea. *Mol*. *Phylogenet*. *Evol*. **62**, 496-507 (2012).

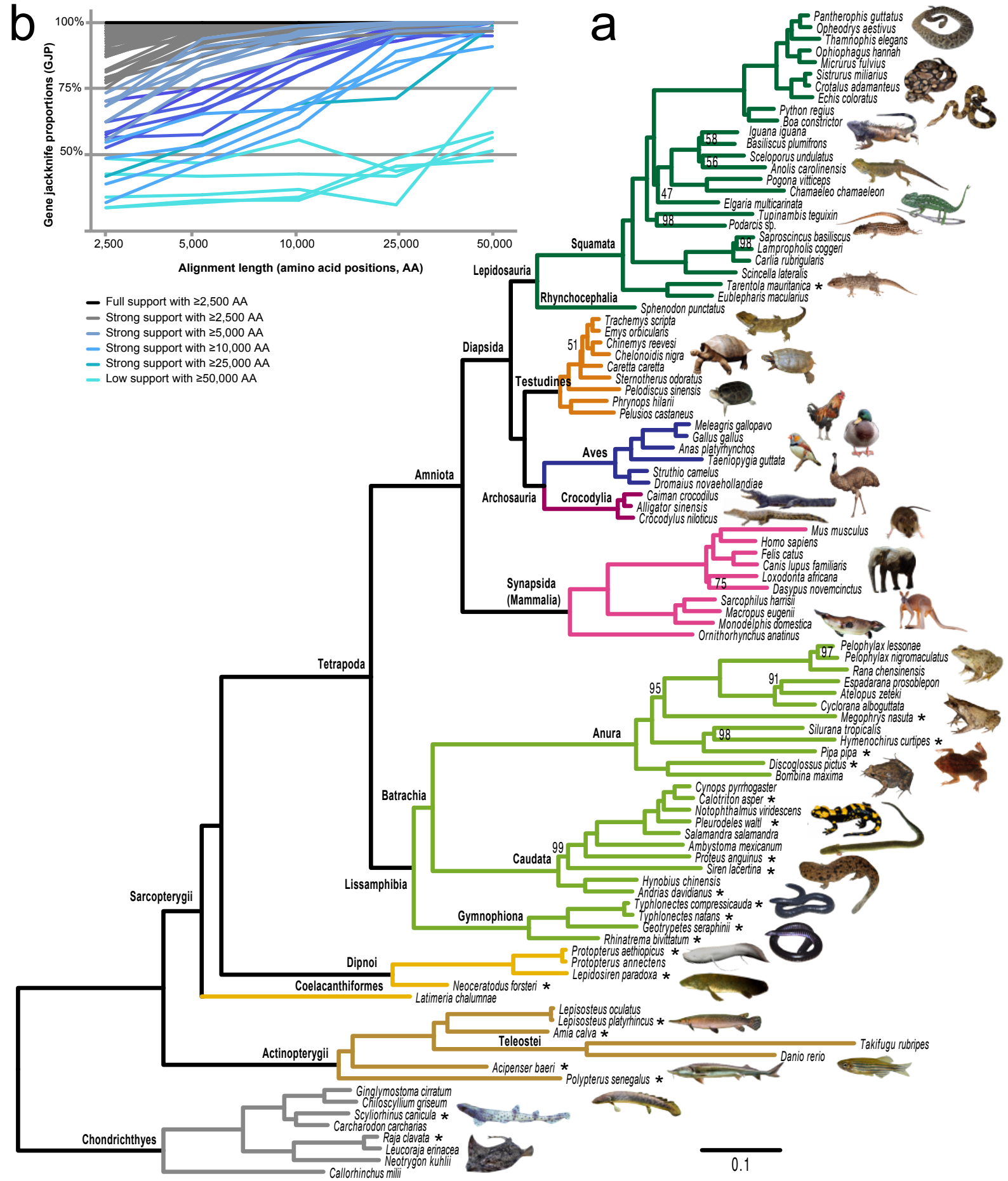
52 Wang, M. *et al*. The oldest record of ornithuromorpha from the early cretaceous of China. *Nat*. *Commun*. **6**, 6987 (2015).

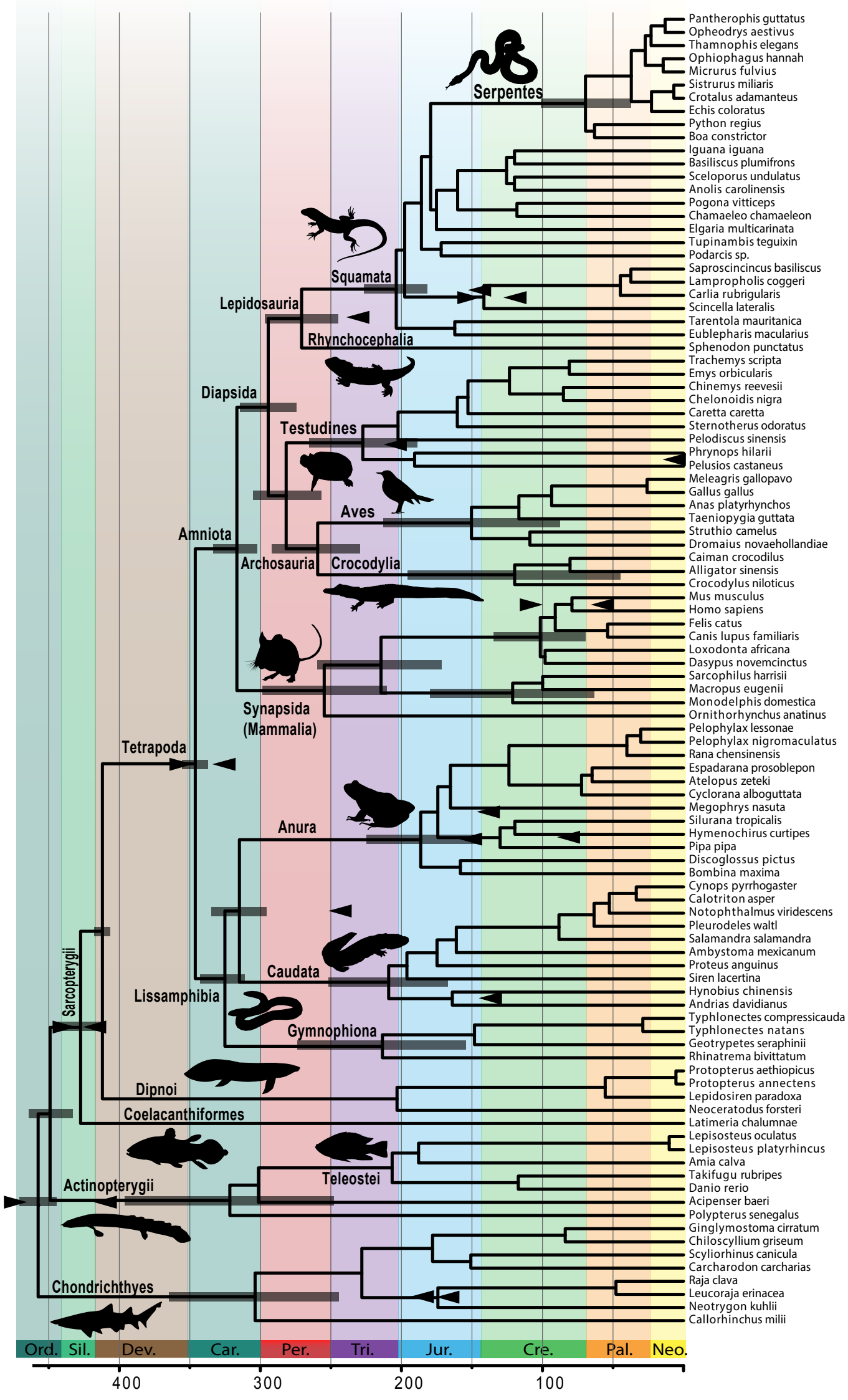
53 Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313 (2014).

54 Lartillot, N., Rodrigue, N., Stubbs, D. & Richer, J. PhyloBayes MPI: Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment. *Syst*. *Biol*. **62**, 611-615 (2013).

644 **Figure 1.** Transcriptome sequencing effort and performance in phylogenomic dataset
645 assembly. Histogram represent sequencing effort as total number of sequenced (clean) Mbp
646 (million bp). Transcriptome completeness is measured as the proportion of recovered core
647 vertebrate genes (233 CVG; Hara et al.[29]). Genes effectively usable for phylogenomics are
648 approximated by the proportion of human peptides reconstructed at full (100%) and nearly
649 full (>70%) lengths (in proportion to a total of 22,964 human genes). The completeness the
650 relevant species in our final phylogenomic dataset is shown as the proportion of amino acids
651 across all 7,189 genes (3,791,500 aligned amino acids in total).
652
653 **Figure 2.** Backbone phylogeny of jawed vertebrates. (a) Bayesian majority-rule consensus
654 tree from 100 independent MCMC chains derived from gene jackknife replicates (~50,000
655 amino acid positions each) of the NoDP nuclear dataset, estimated by PhyloBayes under the
656 CAT+Γ model. All nodes received full gene jackknife support (100%), except those
657 displaying the actual value. The scale bar corresponds to the expected number of substitutions
658 per site. Asterisks denote new transcriptomic data generated in this study. (b) Effect of
659 alignment length on the recovery of single nodes in the phylogeny assessed by gene jackknife
660 proportions derived from the NoDP dataset.
661
662 **Figure 3.** Time-calibrated phylogeny of jawed vertebrates. Divergences have been averaged
663 across 100 timetrees estimated from independent gene jackknife replicates in PhyloBayes,
664 using the subset of most congruent calibrations (C16; marked by arrows) and best-fit
665 evolutionary (CAT-GTR+Γ) and relaxed clock (autocorrelated lognormal) models. Credibility
666 intervals (CrI) are calculated as the absolute maximum and minimum values of 95%
667 confidence intervals across 100 timetrees (only displayed for key nodes; see Supplementary
668 Table 9 for detailed results). The dimensions of the scale is given in million years and main
669 geological periods are highlighted.


Total sequenced Mbp
Completeness (CVG%)
Amino acids in final alignments (%)
Human peptides at 70% coverage (%)
Human peptides at 100% coverage (%)
30,000
25,000
20,000
15,000
10,000
5,000
100%
80%
60%
40%
20%
Megophrys nasuta
Calotriton asper
Typhlonectes natans
Geotrypetes seraphini
Proteus anguinus
Andrias davidianus
Pipa pipa
Discoglossus pictus
Siren lacertina
Hymenochirus curtipes
Neoceratodus forsteri
Lepidosiren paradoxa
Scyliorhinus canicula
Protopterus annectens
Protopterus aethiopicus
Rhinatrema bivittatum
Polypterus senegalus
Lepisosteus platyrhincus
Amia calva
Pleurodeles waltl
Typhlonectes compressicauda
Tarentola mauritanica
Raja clavata
Acipenser baerii


b
100%
75%
50%
Gene jackknife proportions (GJP)
2,500
5,000
10,000
25,000
50,000
Alignment length (amino acid positions, AA)
Full support with ≥2,500 AA
Strong support with ≥2,500 AA
Strong support with ≥5,000 AA
Strong support with ≥10,000 AA
Strong support with ≥25,000 AA
Low support with ≥50,000 AA
a
Pantherophis guttatus
Opheodrys aestivus
Thamnophis elegans
Ophiophagus hannah
Micrurus fulvius
Sistrurus miliarius
Crotalus adamanteus
Echis coloratus
Python regius
Boa constrictor
Iguana iguana
Basiliscus plumifrons
Sceloporus undulatus
Anolis carolinensis
Pogona vitticeps
Chamaeleo chamaeleon
Elgaria multicarinata
Tupinambis teguixin
Podarcis sp.
Saproscincus basiliscus
Lamprophonis coggeri
Carlia rubrigularis
Scincella lateralis
Tarentola mauritanica *
Eublepharis macularius
Sphenodon punctatus
Trachemys scripta
Emys orbicularis
Chinemys reevesi
Chelonoidis nigra
Caretta caretta
Sternotherus odoratus
Pelodiscus sinensis
Phrynops hilarii
Pelusios castaneus
Meleagris gallopavo
Gallus gallus
Anas platyrhynchos
Taeniopygia guttata
Struthio camelus
Dromaius novaehollandiae
Caiman crocodilus
Alligator sinensis
Crocodylus niloticus
Mus musculus
Homo sapiens
Felis catus
Canis lupus familiaris
Loxodonta africana
Dasypus novemcinctus
Sarcophilus harrisii
Macropus eugenii
Monodelphis domestica
Ornithorhynchus anatinus
Pelophylax lessonae
Pelophylax nigromaculatus
Rana chensinensis
Espadarana prosoblepon
Atelopus zeteki
Cyclorana alboguttata
Megophrys nasuta *
Silurana tropicalis
Hymenochirus curtipes *
Pipa pipa *
Discoglossus pictus *
Bombina maxima
Cynops pyrrhogaster
Calotriton asper *
Notophthalmus viridescens
Pleurodeles waltl *
Salamandra salamandra
Ambystoma mexicanum
Proteus anguinus *
Siren lacertina *
Hynobius chinensis
Andrias davidianus *
Typhlonectes compressicauda *
Typhlonectes natans *
Geotrypetes seraphinii *
Rhinatrema bivittatum *
Protopterus aethiopicus *
Protopterus annectens
Lepidosiren paradoxa *
Neoceratodus forsteri *
Latimeria chalumnae
Lepisosteus oculatus
Lepisosteus platyrhincus *
Amia calva *
Takifugu rubripes
Danio rerio
Acipenser baeri *
Polypterus senegalus *
Ginglymostoma cirratum
Chiloscyllium griseum
Scyliorhinus canicula *
Carcharodon carcharias
Raja clavata *
Leucoraja erinacea
Neotrygon kuhlii
Callorhinchus milii
Squamata
Lepidosauria
Rhynchocephala
Diapsida
Testudines
Aves
Crocodylia
Archosauria
Amniota
Synapsida (Mammalia)
Tetrapoda
Anura
Batrachia
Caudata
Gymnophiona
Lissamphibia
Sarcopterygii
Dipnoi
Coelacanthiformes
Teleostei
Actinopterygii
Chondrichthyes
58
56
47
98
98
51
75
97
91
95
98
99
0.1


Pantherophis guttatus
Opheodrys aestivus
Thamnophis elegans
Ophiophagus hannah
Micrurus fulvius
Sistrurus miliaris
Crotalus adamanteus
Echis coloratus
Python regius
Boa constrictor
Iguana iguana
Basiliscus plumifrons
Sceloporus undulatus
Anolis carolinensis
Pogona vitticeps
Chamaeleo chamaeleon
Elgaria multicarinata
Tupinambis teguixin
Podarcis sp.
Saproscincus basiliscus
Lampropholis coggeri
Carlia rubrigularis
Scincella lateralis
Tarentola mauritanica
Eublepharis macularius
Sphenodon punctatus
Trachemys scripta
Emys orbicularis
Chinemys reevesii
Chelonoidis nigra
Caretta caretta
Sternotherus odoratus
Pelodiscus sinensis
Phrynops hilarii
Pelusios castaneus
Meleagris gallopavo
Gallus gallus
Anas platyrhynchos
Taeniopygia guttata
Struthio camelus
Dromaius novaehollandiae
Caiman crocodilus
Alligator sinensis
Crocodylus niloticus
Mus musculus
Homo sapiens
Felis catus
Canis lupus familiaris
Loxodonta africana
Dasypus novemcinctus
Sarcophilus harrisii
Macropus eugenii
Monodelphis domestica
Ornithorhynchus anatinus
Pelophylax lessonae
Pelophylax nigromaculatus
Rana chensinensis
Espadarana prosoblepon
Atelopus zeteki
Cyclorana alboguttata
Megophrys nasuta
Silurana tropicalis
Hymenochirus curtipes
Pipa pipa
Discoglossus pictus
Bombina maxima
Cynops pyrrhogaster
Calotriton asper
Notophthalmus viridescens
Pleurodeles waltl
Salamandra salamandra
Ambystoma mexicanum
Proteus anguinus
Siren lacertina
Hynobius chinensis
Andrias davidianus
Typhlonectes compressicauda
Typhlonectes natans
Geotrypetes seraphinii
Rhinatrema bivittatum
Protopterus aethiopicus
Protopterus annectens
Lepidosiren paradoxa
Neoceratodus forsteri
Latimeria chalumnae
Lepisosteus oculatus
Lepisosteus platyrhincus
Amia calva
Takifugu rubripes
Danio rerio
Acipenser baeri
Polypterus senegalus
Ginglymostoma cirratum
Chiloscyllium griseum
Scyliorhinus canicula
Carcharodon carcharias
Raja clava
Leucoraja erinacea
Neotrygon kuhlii
Callorhinchus milii
Serpentes
Squamata
Lepidosauria
Rhynchocephalia
Diapsida
Testudines
Aves
Amniota
Archosauria
Crocodylia
Synapsida
(Mammalia)
Tetrapoda
Anura
Caudata
Lissamphibia
Gymnophiona
Sarcopterygii
Dipnoi
Coelacanthiformes
Actinopterygii
Teleostei
Chondrichthyes
Ord.
Sil.
Dev.
Car.
Per.
Tri.
Jur.
Cre.
Pal.
Neo.
400
300
200
100