A Genomic Perspective on the Evolutionary Diversification of Turtles

Simone M. Gable[1*], Michael I. Byars[1*], Robert Literman[2], Marc Tollis[1]

[1]*School of Informatics, Computing, and Cyber Systems, Northern Arizona University, PO Box 5693, Flagstaff, AZ 8601, USA*

[2]*Department of Biological Sciences, University of Rhode Island, 120 Flagg Road, Kingstown, RI, 0288, USA*

*These authors contributed equally.

Corresponding author: Marc Tollis, School of Informatics, Computing, and Cyber Systems, Northern Arizona University, PO Box 5693, Flagstaff, AZ, 86011; marc.tollis@nau.edu

Running header: Phylogenomic Heterogeneity in Turtles

ABSTRACT

To examine phylogenetic heterogeneity in turtle evolution, we collected thousands of high-confidence single-copy orthologs from 19 genome assemblies representative of extant turtle diversity and estimated a phylogeny with multispecies coalescent and concatenated partitioned methods. We also collected next-generation sequences from 26 turtle species and assembled millions of biallelic markers to reconstruct phylogenies based on annotated regions from the western painted turtle (*Chrysemys picta bellii*) genome (coding regions, introns, untranslated regions, intergenic, and others). We then measured gene tree-species tree discordance, as well as gene and site heterogeneity at each node in the inferred trees, and tested for temporal patterns in phylogenomic conflict across turtle evolution. We found strong and consistent support for all bifurcations in the inferred turtle species phylogenies. However, a number of genes, sites, and genomic features supported alternate relationships between turtle taxa. Our results suggest that gene tree-species tree discordance in these datasets is likely driven by population-level processes such as incomplete lineage sorting. We found very little effect of substitutional saturation on species tree topologies, and no clear phylogenetic patterns in codon usage bias and compositional heterogeneity. There was no correlation between gene and site concordance, node age, and DNA substitution rate across most annotated genomic regions. Our study demonstrates that heterogeneity is to be expected even in well resolved clades such as turtles, and that future phylogenomic studies should aim to sample as much of the genome as possible in order to obtain accurate phylogenies for assessing conservation priorities in turtles.

Keywords: turtles, genomes, phylogeny, discordance

40 **Introduction**

41 Next-generation sequencing has greatly increased the numbers of sampled loci for phylogenomic
42 analyses (Lemmon et al. 2012; Faircloth et al. 2013; Edwards et al. 2016), shedding new light on
43 the branching order of diversification across the history of life. However, phylogenetic studies
44 based on genome-scale data often contain a great deal of heterogeneity in the inferred patterns
45 (Hime et al. 2021; Lopes et al. 2021; Morales-Briones et al. 2021; Singhal et al. 2021),
46 particularly in the form of discordance among sites and gene trees (Kumar et al. 2012). Gene
47 tree-species tree discordance occurs when gene trees are in conflict with the underlying species
48 relationships, while site discordance occurs when a subset of sites support different bifurcations
49 in a species tree, and both factors can lead to biased phylogenetic estimates. Understanding
50 patterns and potential sources of phylogenomic heterogeneity will not only settle phylogenetic
51 arguments and reduce bias in tree building, but also provide better models with which to
52 understand trait evolution and help define conservation priorities through comparative methods
53 (Garland et al. 2005; Colston et al. 2021).

54 Biological factors driving phylogenomic heterogeneity include: differing evolutionary
55 rates across the genome, incomplete lineage sorting, and hidden paralogy (Pamilo and Nei 1988;
56 Galtier and Daubin 2008). Artifacts that drive heterogeneity include sequencing errors and
57 substitution model violations such as genetic saturation, compositional heterogeneity, and codon
58 usage bias (Foster 2004; Cooper 2014; Cox et al. 2014). Accounting for gene and site
59 discordance in phylogenomic estimation often requires partitioning schemes which try to account
60 for differing evolutionary rates and model parameters (Kubatko and Degnan 2007), methods that
61 are consistent with the multispecies coalescent accounting for population-level processes such as
62 incomplete lineage sorting (Edwards et al. 2016), as well as strict filtering of sequence

63 alignments that aim to reduce potential errors. The application of hundreds or even thousands of
64 loci is often an attempt to overcome biases stemming from or driving phylogenomic conflict, and
65 these studies often boast strong branch support as measured by bootstrap replicates or posterior
66 probabilities, likely as a result of reduced sampling variance in large datasets (Minh et al. 2020a).
67 However, competing phylogenomic studies of the same clades that use different datasets and tree
68 building methods can produce conflicting results, despite high levels of statistical support (see
69 modern birds: Jarvis et al. 2014; Prum et al. 2015).

70 An emerging advantage of the genomic era is the theoretical ability to access the whole
71 genome in order to sample loci representing a diversity of evolutionary rates and coalescent
72 histories (Wolf et al. 2002; Rokas et al. 2003). The most widely used sampling methods include
73 the use of large numbers of coding sequences (Singhal et al. 2021), anchored hybrid enrichment
74 loci (AHE; Lemmon et al. 2012), ultra-conserved elements (UCEs; Faircloth et al. 2013), and
75 transcriptomes (Irisarri et al. 2017), but these are each still reduced representations of a relatively
76 small proportion of the genome (Lynch 2007). Meanwhile, the development of bioinformatics
77 tools and sequencing technologies have enabled chromosome-scale genome assemblies for non-
78 model organisms (see http://dnazoo.org, Dudchenko et al. 2017), potentially providing insights
79 to the genomic regions or types of loci that drive heterogeneous phylogenetic results. Thus, by
80 accessing complete genomes and studying heterogeneity closely, we can more fully comprehend
81 conflicting results across gene trees, obtain more accurate perspectives of species relationships,
82 and gain a fuller understanding of molecular evolution across the genome.

83 Here, we examined the importance of genome-wide heterogeneity in the phylogenomic
84 estimation of turtles (Order Testudines). Turtles are a near-globally distributed and
85 morphologically distinct clade of shelled reptiles, with a rich fossil history dating from the

86 Triassic Period (Gaffney 1980; Joyce 2007; Joyce et al. 2021). Despite their persistence in the
87 fossil record and across modern ecosystems, almost half of turtle species are listed in the
88 International Union for Conservation of Nature (IUCN) as endangered, critically endangered, or
89 vulnerable to extinction (IUCN, 2021). The approximately 350 extant species of turtles are
90 classified into two main groups, Pleurodira ("side-necked") and Cryptodira ("hidden-necked"),
91 which are further divided into 8 superfamilies and 14 families (Uetz et al. 2021). Early molecular
92 phylogenetic results disagreed in the placement of some turtle taxa; in particular, the position of
93 the monotypic big-headed turtle (*Platysternon*) and its relationships with or within Testudinoidea
94 (Shaffer et al. 1997; Krenz et al. 2005). Other questions about deeper turtle relationships such as
95 the position of soft-shell turtles (Trionychoidea), the Americhelydiae, and the position of turtles
96 in the amniote phylogeny persisted into the next-generation sequencing era (Chiari et al. 2012;
97 Crawford et al. 2012, 2015; Brown and Thomson 2017). Since then, phylogenomic studies of
98 turtles have produced fully resolved species trees with 100% statistical support at every node and
99 near complete agreement across studies as measured by bootstrap replicates or posterior
100 probabilities (Crawford et al. 2015; Shaffer et al. 2017).

101 Phylogenomic heterogeneity has been examined in vertebrate groups such as mammals
102 (Tarver et al. 2016; Liu et al. 2017) squamates (Burbrink et al. 2020; Singhal et al. 2021), birds
103 (Jarvis et al. 2014; Prum et al. 2015), and amphibians (Hime et al. 2021), but there has not yet
104 been an investigation into how gene tree-species tree discordance across the genome affects the
105 inference of relationships between the major turtle lineages. While phylogenetic conflict often
106 belies more traditional means of measuring certainty via bootstrap support or posterior
107 probabilities (Minh et al. 2020a), the study of turtle systematics has not yet truly taken advantage
108 of the genomic era; this is despite ample genomic resources being available for a wide range of

109 turtle species (Shaffer et al. 2013; Wang et al. 2013; Tollis et al. 2017; Quesada et al. 2019).
110 Since evolutionary distinctiveness is linked to conservation status in turtles (Colston et al. 2020),
111 and many turtle species have adaptations of physiological and developmental importance such as
112 longevity, resistance to hypoxia, cold tolerance, and the anatomical changes responsible for the
113 turtle shell, a study using genome-scale data to measure phylogenomic heterogeneity is needed in
114 order to obtain the most accurate models of turtle evolution. Our goals in this study were
115 threefold: (1) to revisit discoveries about higher turtle systematics using substantial proportions
116 of the genome; (2) to measure heterogeneity at genes and sites in important events in turtle
117 evolution; and (3) determine what drives heterogeneity in turtle phylogenomics in terms of
118 mechanisms and locus type.

119 To accomplish these goals, we generated two distinct yet overlapping genomic datasets
120 for turtles. The first comprises 5,310 high-confidence single-copy orthologs bioinformatically
121 extracted from the genome assemblies of 19 turtles plus three outgroups, aligned and
122 phylogenetically analyzed using multilocus coalescent-consistent and concatenated partition-
123 based methods. The second dataset (26 species, including all turtle species from the first method)
124 consisted of 1,655,675 biallelic parsimony-informative sites that we extracted using mapped
125 Illumina sequence reads. We assembled the reads into contigs that were further mapped to a
126 reference turtle genome (the western painted turtle, *Chrysemys picta bellii*). Based on genome
127 annotations, we partitioned sites by locus type including coding regions (CDS), introns, 5'-UTR,
128 3'-UTR, intergenic, pseudogenes, lncRNA, and smRNA, and reconstructed separate phylogenies
129 using each locus type. With the first dataset we assessed patterns and sources of heterogeneity in
130 turtle phylogenomics with a large number of DNA sequence alignments, while with the second
131 dataset we compared topologies reconstructed with parsimony-informative sites extracted from

132 different annotated regions across the turtle genome. Taken together, this scope of analyses is
133 unprecedented in turtle phylogenomics. We confirm that heterogeneity in phylogenomic datasets
134 is to be expected, even in well-resolved clades such as turtles, and suggest that a combination of
135 processes is driving the incongruence between previous studies of turtle relationships.
136

137 **Methods and Materials**

138 *Phylogenomic analysis of high-confidence single-copy orthologs*

139 We downloaded the complete genome assemblies of 19 turtle species plus three outgroup taxa
140 from NCBI. The species, assembly accession numbers, assembly lengths, and assembly
141 contiguities measured by scaffold N50 are shown in Table 1. Using BUSCO v3 (Waterhouse et
142 al. 2018), we extracted the nucleotide sequences of 5,310 OrthoDB v9 tetrapod orthologs
143 (Zdobnov et al. 2017) from each genome assembly. BUSCO orthologs are high-confidence genes
144 that persist in eukaryotic genomes as single-copy, which precludes downstream problems in
145 phylogenetics stemming from gene duplication such as hidden paralogy (Waterhouse et al.
146 2018). We aligned each set of orthologs with MAFFT v7.475 (Katoh and Standley 2013), and
147 removed erroneous columns in the alignments with the heuristic method (*automated 1*) in
148 TrimAL v1.4.1 (Capella-Gutierrez et al. 2009). Summary statistics including number of taxa,
149 alignment length, missing data percentage, number and proportion of variable sites, number and
150 proportion of parsimony-informative sites, and GC content for each alignment were estimated
151 with AMAS v3.04 (Borowiec 2016). Trimmed alignments were filtered at a cut-off length
152 ≥1,500 bp and a minimum of 50% taxa representation. We estimated the average pairwise
153 Kimura 2-parameter corrected distance for each alignment with MEGAX (Kumar et al. 2018).

154 **Table 1. Genomic data for turtles analyzed in this study.**

Suborder	Superfamily/Clade	Family	Species Name	Assembly Accession	Assembly Length (Gb)	Scaffold N50 (Mb)	Trimmed Read Depth*
Cryptodira	Chelonioidea	Cheloniidae	*Chelonia mydas*	GCF_000344595.1	2.2	3.9	12.8X
		Dermochelyidae	*Dermochelys coriacea*	GCA_006547105.1	2.1	0.12	31.3X
	Chelydridae	Chelydridae	*Chelydra serpentina*	GCA_007922165.1	2.5	21	16.1X
	Kinosternoidea	Dermatemydidae	*Dermatemys mawii*	GCA_007922305.1	1.9	34.4	39.7X
		Kinosternidae	*Sternotherus carinatus*	NA	NA	NA	45.9X
	Testudinoidea	Emydidae	*Chrysemys picta*	GCA_000241765.3	2.4	16	10.3X
			Malaclemys terrapin	GCA_001728815.2	2.4	0.44	9.3X
			Terrapene mexicana	GCF_002925995.2	2.6	24.2	12.4X
			Emys orbicularis	NA	NA	NA	1.3X
			Trachemys scripta	NA	NA	NA	40.2X
			Actinemys marmorata	NA	NA	NA	41.4X
		Geoemydidae	*Cuora amboinensis*	GCA_004028625.2	2.1	0.25	23.1X
			Cuora mccordi	GCA_003846335.1	2.4	32.6	70.6X
			Mauremys reevesii	NA	NA	NA	34.1X
			Batagur trivittata	NA	NA	NA	1.1X
		Platysternidae	*Platysternon megacephalum*	GCA_003942145.1	2.3	7.2	16.7X
		Testudinidae	*Chelonoidis abingdonii*	GCA_003597395.1	2.3	1.3	17.3X
			Gopherus agassizii	GCA_002896415.1	2.2	0.228	40.5X
			Aldabrachelys gigantea	NA	NA	NA	21.9X
	Trionychoidea	Carettochelyidae	*Carettochelys insculpta*	GCA_007922185.1	2.3	45.9	49.7X
		Trionychidae	*Apalone spinifera*	GCA_000385615.1	1.9	2.3	10.3X
			Pelodiscus sinensis	GCF_000230535.1	2.2	3.4	21.1X

Pleurodira	Chelidoidea	Chelidae	*Emydura subglobosa*	GCA_007922225.1	2	44.8	40.6X
			Mesoclemmys tuberculata	GCA_007922155.1	2	46.4	49.0X
	Pelomedusoidea	Pelomedusidae	*Pelusios castaneus*	GCA_007922175.1	2	14.1	47.5X
		Podocnemididae	*Podocnemis expansa*	GCA_007922195.1	2.4	37.1	10.5X

Gb=gigabases; Mb=megabases; *Short Read Archive Accession numbers are in Supplementary Data; NA=no genome assembly was analyzed in the current study

Gene trees were reconstructed for each locus using RAxML v1.8 (Stamatakis 2014). For each ortholog, we generated 10 maximum likelihood trees with the GTRGAMMA substitution model, and performed 100 rapid bootstrap replicates on the best tree. We then collected the best tree for each ortholog to use as evidence to construct a species tree with ASTRAL-III (Zhang et al. 2018), assessing branch support with local posterior probabilities (Sayyari and Mirarab 2016). We used Snakemake (Mölder et al. 2021) to automate a reproducible bioinformatics pipeline for BUSCO extraction, filtering, alignment, gene tree and species tree estimation, available at https://github.com/mibyars/busco_phylo_pipeline. We also concatenated the single copy orthologs into a supermatrix with AMAS and performed a partitioned maximum likelihood phylogenetic reconstruction with IQ-TREE v2.1.2 (Minh et al. 2020b). We used ModelFinder (Kalyaanamoorthy et al. 2017) set to TESTMERGE to combine partitions where necessary. We assessed branch support on the concatenated partitioned maximum likelihood tree using 1,000 ultrafast bootstrap replicates and the Shimodaira-Hasegawa-like (SH-like) approximate likelihood ratio test (aLRT, Shimodaira 2002).

173 *Phylogenomic analysis of biallelic markers*

174 In addition to single-copy orthologs from turtle and outgroup genome assemblies, we constructed
175 sets of biallelic single nucleotide polymorphisms (SNPs) using paired-end sequence data from
176 multiple turtle species downloaded from the NCBI Short Read Archive (SRA, Table 1,
177 Supplementary Data). We assessed each SRA dataset with FastQC v0.115 (S. Andrews -
178 http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed the raw reads with
179 BBDuk v.37.41 (B. Bushnell - sourceforge.net/projects/bbmap/, last accessed December 2020).
180 We used SISRS to generate de novo orthologs from the processed reads (Schwartz et al. 2015;
181 Literman and Schwartz 2021). After subsampling all sequences to a target depth of ~10X per
182 taxon (except for two low-coverage species, Table 1, Supplementary Data), we assembled a
183 composite genome using Ray v2.2.3-devel (Boisvert et al. 2010) with default parameters and a k-
184 value of 31. We then used SISRS to map the trimmed reads from each species to the composite
185 genome, retaining only uniquely mapped reads to avoid false positives from duplicated or
186 repetitive regions. Bases in the composite genome were replaced according to mapped data from
187 each species if sites were covered by at least three reads and if there was no within-taxon
188 variation; sites not meeting these criteria were masked as "N".

189 For downstream ortholog annotation, we mapped each scaffold from the composite
190 sequences against the reference genome for the western painted turtle (*Chrysemys picta bellii,*
191 GCA_000241765.3) with Bowtie2 v2.3.4 (Langmead and Salzberg 2012), removing contigs that
192 either failed to map or mapped to multiple positions, and removing sites resulting from
193 overlapping coverage of independent scaffolds. We then annotated the mapped composite
194 genome sites parsed from the painted turtle genome as belonging to the following genomic
195 regions (locus types) with BEDTools v2.2.6 (Quinlan and Hall 2010): (1) coding sequences

196 (CDS); (2) 3'-UTR; (3) 5'-UTR; (4) introns; (5) other genic regions not annotated as CDS, UTR,
197 or intronic; (6) long-noncoding RNAs (lncRNAs); (7) pseudogenes; (8) smRNAs (including
198 miRNAs, ncRNAs, rRNAs, scRNAs, smRNAs, snoRNAs, snRNAs, tRNAs, and vaultRNAs).;
199 and (9) intergenic regions. We defined "intergenic" as all regions not included within the other
200 locus types.

201 We used SISRS to identify parsimony-informative sites along the mapped composite
202 contigs and to remove invariant sites and singletons, as well as any sites containing "Ns" (i.e.
203 polymorphic within species) and overlapping indels or gaps. For the nine annotated genomic
204 regions as well as a set of the complete mapped data (=10 locus types), we concatenated biallelic
205 parsimony-informative sites and inferred locus type phylogenies via SVDquartets (Chifman and
206 Kubatko 2014) in PAUP* v4.0a (build 168) (Swofford 2003) using 1,000 bootstrap replicates
207 and estimated majority consensus trees.

208

209 *Heterogeneity analyses*

210 To measure similarity across genealogies, we calculated Robinson-Foulds (RF) distances
211 (Robinson and Foulds 1981) with PhyKIT (Steenwyk et al. 2021). First, we measured the RF
212 distance between each inferred single-copy ortholog gene tree and the inferred species tree. In
213 order to take into account missing species across some of the gene trees, we normalized the RF
214 distances by taking the raw RF distance and dividing it by $2(n\text{-}3)$ where n is the number of tips in
215 the phylogeny. We also calculated multiple pairwise RF distances between the trees inferred
216 from SISRS data mapped to each genome annotation (CDS, introns, intergenic, etc.). Finally, we
217 calculated the RF distances of each SISRS phylogeny from the topology inferred with all
218 mapped sites across the turtle genome.

219 We computed gene and site concordance factors for each branch in the turtle phylogeny
220 using the trimmed and length-filtered single-copy ortholog alignments, their gene trees, and the
221 ASTRAL tree with IQ-TREE (Minh et al. 2020a). A branch's gene concordance factor
222 represents the percentage of decisive gene trees that also contain that branch, and a branch's site
223 concordance factor is the percentage of decisive alignment sites supporting that branch. We also
224 computed gene concordance factor and site concordance factor for each branch in the
225 concatenated single-copy ortholog tree by computing individual locus trees based on the
226 partitions in IQ-TREE. Locus trees were inferred with default settings, and the species tree was
227 inferred using an edge-linked proportional partition model with an SH-like approximate
228 likelihood ratio test and ultrafast bootstrap analysis of 10,000 replicates each (-alrt 10000 -B
229 10000). The greedy algorithm of PartitionFinder (Lanfear et al. 2012) was applied to find the
230 best-fit partitioning scheme which was then used in the subsequent tree reconstruction step (-m
231 TESTMERGE). We also calculated site concordance factors for every branch in the
232 SVDquartets trees inferred from biallelic parsimony-informative sites within locus types.
233 Discordance between gene trees and species trees can be caused by introgression,
234 incomplete lineage sorting, erroneously inferred gene trees stemming from model violations
235 and/or other artifacts, or combinations of all of these factors, and we aimed to determine how
236 these issues affect phylogenomic inference in turtles. One expectation of incomplete lineage
237 sorting is a roughly equal proportion of topologies supporting alternative relationships, while
238 introgression would cause some minor gene trees to be more frequent than others, reflecting the
239 direction of gene flow (Huson et al. 2005; Green et al. 2010). We collected the two most
240 common minor topologies and each node (gDF1 and gDF2 resulting from the concordance factor
241 analysis in IQ-TREE) and used a chi-squared test to determine if they were similar in terms of

242 their proportions of all gene trees, where a rejection of the null hypothesis was interpreted as the
243 presence of introgression in addition to incomplete lineage sorting and potential model
244 violations. Nodes where we failed to reject the null hypothesis were interpreted as divergence
245 patterns in turtle evolution driven primarily by either incomplete lineage sorting or model
246 violations.
247
248 *Analysis of potential model violations via substitutional saturation, compositional heterogeneity,*
249 *and codon usage bias*
250 Phylogenetic signal can be obscured by the accumulation of multiple substitutions at the same
251 site over time, leading to model violations and increased gene tree-species tree discordance,
252 particularly at deep evolutionary timescales (Jeffroy et al. 2006; Philippe et al. 2011). Therefore,
253 to further distinguish between the effects of incomplete lineage sorting and model violations at
254 nodes in the turtle species tree, we estimated levels of substitutional saturation in the single-copy
255 ortholog dataset. First, we calculated the slope of a linear regression between the computed raw
256 pairwise distances and corrected pairwise distances under a TN93 substitution model for 685
257 single-copy ortholog alignments with complete taxon sampling using the ape 5.5 R package
258 (Paradis and Schliep 2019). For genes with high levels of saturation, the slope of this regression
259 will be closer to 0 while less saturated genes will have a slope approaching 1 (Philippe et al.
260 1994). We then plotted a histogram of all slopes and characterized each alignment as having
261 slopes above (“unsaturated”) and below (“saturated”) the mean. To determine the effect of
262 substitutional saturation on phylogenomic inference in turtles, we compared ASTRAL species
263 trees inferred from the RAxML trees inferred from unsaturated and saturated genes.

264 To analyze substitutional saturation at codon positions and assess their effects on
265 phylogenomic inference in turtles, we realigned the 685 single-copy orthologs with complete
266 taxon sampling using MACSE (Ranwez et al. 2011), and concatenated alignments that were
267 partitioned based on (1) first and second codon positions and (2) third codon positions with
268 AMAS. We then computed the corrected and uncorrected pairwise distances and estimated the
269 regression slopes for the two codon-based partitions in R as above. We also further separated
270 these codon partitions according to whether they belonged to the unsaturated or saturated gene
271 sets, inferred maximum likelihood phylogenies on the four resulting datasets in IQ-TREE with
272 model testing and 1,000 ultrafast bootstrap replicates, and calculated site concordance factors for
273 each inferred tree from codon partitions. Finally, to account for substitutional saturation we
274 computed a maximum likelihood tree from the concatenated and partitioned MACSE amino acid
275 alignments of the 685 single-copy orthologs using model testing in IQ-TREE. Maximum
276 likelihood tree topologies, branch support, and concordance factors from codon-partitioned
277 saturated and unsaturated alignments as well as the amino acids were compared to the ASTRAL
278 and concatenated supermatrix trees based on single-copy orthologs ≥1,500bp.

279 A chi-squared test for lineage-specific compositional heterogeneity is computed before
280 each IQ-TREE run, where the frequency of each nucleotide in a given taxon is compared to the
281 overall frequency of each nucleotide in the entire dataset. We compared the results of this test
282 across IQ-TREE analyses of the saturated and unsaturated concatenated partitions based on
283 codon position. To analyze codon usage bias in our single copy ortholog dataset, we estimated
284 how often a codon is observed relative to its predicted occurrence in the absence of codon usage
285 bias, by computing the Relative Synonymous Codon Usage (RSCU) of the concatenated 685

286 complete-taxon MACSE alignments using CodonW v1.4.4 (Peden 1999) followed by factor
287 analysis in FactoMineR (Lê et al. 2008) in R.
288
289 *Divergence time estimation*
290 We estimated divergence times with single-copy ortholog alignments using Bayesian modeling
291 implemented in BEAST v.2.6.3 (Bouckaert et al. 2019). We subsampled 685 alignments with
292 complete taxon sampling, from which we further randomly sampled three replicate sets of 10
293 genes. Based on model testing with ModelFinder, we applied the HKY+Gamma+Invariant site
294 model to all partitions, and utilized a relaxed lognormal molecular clock and Calibrated Yule tree
295 prior. To calibrate the time trees, we used 11 calibration priors from the literature (Joyce et al.
296 2013) which overlapped with our taxon sampling (Supplementary Data). To account for
297 uncertainty in fossil dating, we used hard minimum constraints and set soft maximum constraints
298 by placing the maximum ages in the 97.5% quantile of a lognormal prior distribution. Markov
299 chain Monte Carlo (MCMC) analyses were sampled every 50,000 generations. For each
300 replicate, we assessed convergence of parameter estimates across the MCMC by monitoring
301 effective sample sizes with Tracer v.1.7.2 (Rambaut et al. 2018), and we estimated a maximum
302 clade credibility tree with TreeAnnotator v2.6.3.
303 We also estimated divergence times in turtle evolution using penalized likelihood with
304 r8s v1.8.1 (Sanderson 1997, 2002, 2003), using the same fossil calibrations as above for
305 minimum and maximum node ages; the main difference in this analysis was we fixed the time to
306 most recent common ancestor (TMRCA) for amniotes at 312 million years ago (Ma) (Donoghue
307 and Benton 2007) and the TMRCA for Testudines at 220 Ma (Thomson et al. 2021). First, we
308 concatenated the 685 complete partitions and inferred a maximum likelihood tree with model

309 testing for each partition as above in IQ-TREE. We then used the inferred phylogeny with
310 branches in terms of substitutions per site and the 1,836,182 concatenated bases for penalized
311 likelihood estimation, with cross validation to optimize the smoothing parameter which
312 quantifies the deviation from the molecular clock.

313 We used linear regression to determine the relationship between the estimated divergence
314 time for each node and the concordance factors computed with the inferred trees from each locus
315 type in the SISRS analysis, the gene trees and the species tree based on single-copy orthologs,
316 and the locus trees and partitioned maximum likelihood tree based on single-copy orthologs. We
317 also tested for a correlation between the concordance factors and the estimated rates of molecular
318 evolution at each branch. We repeated these regressions using estimated divergence times from
319 Thomson et al. (2020).

320

321 **Results**

322 *A very large number of informative sites for turtle phylogenomics*

323 We analyzed publicly available genome assemblies to extract single copy orthologs from 19
324 turtles representing 13 out of 14 extant families and 7 of 8 superfamilies, plus 3 outgroup taxa
325 (Table 1). On average, 4,609 (83%, range 1,911-5,099) conserved tetrapod orthologs were
326 complete in the genome assemblies, 112 (2.1%, range 7-496) were duplicated, 170 (3.2%, range
327 55-485) were fragmented, and 419 (7.9%, range 143-2,907) were missing (Fig. 1). Fifty-five
328 percent of conserved tetrapod orthologs were missing from the genome assembly of the
329 trionychid *Apalone spinifera*; therefore, we omitted this species from downstream analyses of
330 single-copy orthologs due to concerns about assembly quality (Waterhouse et al. 2018). Each
331 single-copy ortholog was present in an average of 19 out of 21 species, averaged 1,902 bp in

length, and contained an average of 53% variable sites and 26% parsimony-informative sites. Alignment length was correlated with the number of parsimony-informative sites (R^2=0.69, P=2.2e-16, Fig. 2a). GC content was consistent across the single-copy orthologs and averaged 48% (standard deviation 6.2%, Fig. 2b), as was average pairwise Kimura 2-parameter distance (mean 0.154, median 01.39, standard deviation 0.072, Fig. 2c). The subset of 2,513 alignments ≥1,500 bp in length had very similar characteristics to the complete set of loci; the average large locus length was 1,903 bp and the relationship between alignment length and the number of parsimony-informative sites held (R^2=0.73, P=2.2e-16).

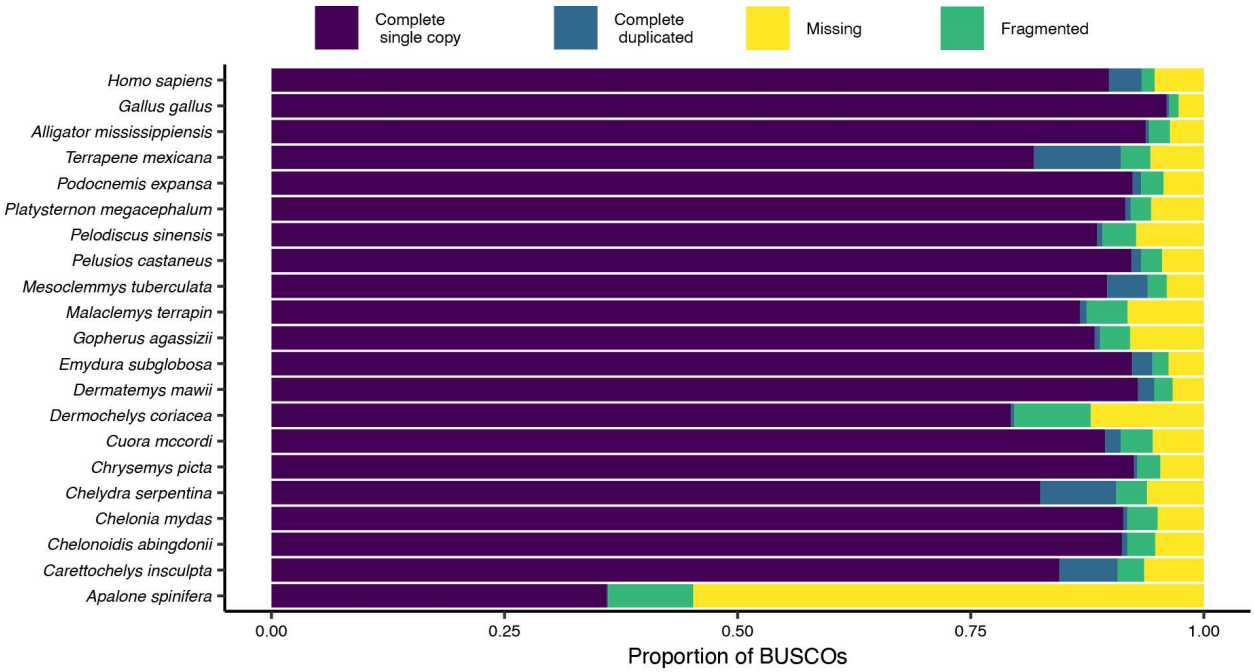
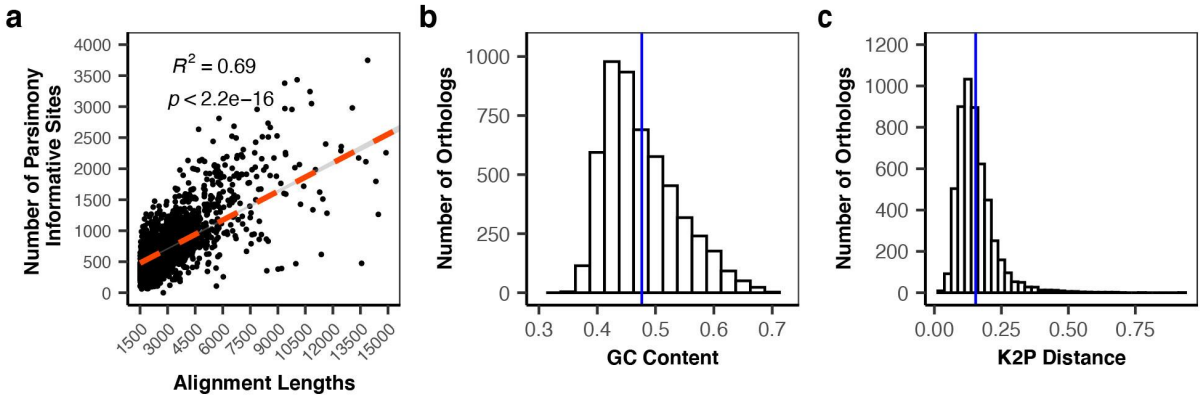


Figure 1. Summary of single-copy orthologs. Percent of tetrapod orthologs from orthoDBv9 present in each genome assembly as complete single copy, complete duplicated, fragmented, or missing.

345



346

347 **Figure 2. Description of single copy ortholog alignments.** (a) Scatterplot showing the relationship between
348 alignment length and the number of parsimony-informative sites for trimmed alignments ≥1,500 bp. (b) Histogram
349 showing the distribution of GC content in all single copy orthologs after trimming. Solid vertical line represents the
350 mean value (0.476). (c) Histogram showing the distribution of average pairwise Kimura 2-parameter (K2P)
351 distances for single copy orthologs after trimming. Solid vertical line represents the mean value (0.154).

352

353 We downloaded 38 SRA datasets from 26 turtle species representing 7 of 8 superfamilies
354 and 14 families (Table 1; Supplementary Data), including the 19 species with complete genome
355 assemblies from the first dataset. Based on FastQC results, we selected 32 of the SRA datasets
356 for further analysis ranging from ~1–70X post-trimming coverage. The composite genome
357 consisted of 5,555,666 contigs with an N50 of 2.4 Mb. We successfully mapped 302,355,876
358 bases to 2,110,354 contigs in the composite genome which covered 80% of the painted turtle
359 genome (Table 2), and called 1,655,675 parsimony-informative biallelic SNPs when allowing
360 missing data for up to two species at a given site. The number of parsimony-informative biallelic
361 SNPs was much larger in mapped datasets containing only cryptodires (2,787,072 bases) and
362 testudinoideans (4,696,637 bases) (Supplementary Data), as expected based on Literman et al.
363 (2021). We collected the highest number of biallelic SNPs from intergenic regions (746,724

364 bases), followed by introns (533,336 bases) and CDS (136,823 bases). The relatively low
365 number of parsimony-informative biallelic SNPs in pseudogenic (427) and smRNA regions
366 (137) is likely due to a small number of annotations for these features in the western painted
367 turtle genome.
368
369 **Table 2. Summary of short read data from 26 turtle species mapped to the western painted turtle (*Chrysemys*
370 *picta*) genome annotations.**

Locus Type	Annotated Bases in the Genome	Bases Mapped by SISRS	Number of biallelic parsimony-informative sites*	Percent of Annotated Bases Mapped
All Mapped	2,365,766,571	302,355,876	1,655,675	12.78%
Intergenic	1,382,145,112	161,445,195	746,724	11.68%
Intronic	812,010,672	115,297,781	533,336	14.20%
lncRNA	64,450,450	9,451,244	52,207	14.66%
Genic Other	65,764,053	9,052,113	36,604	13.76%
CDS	33,486,790	5,718,074	136,823	17.08%
3'-UTR	8,034,514	1,510,783	15,150	18.80%
5'-UTR	5,096,255	708,214	6,519	13.90%
Pseudogenic	670,023	59,058	427	8.81%
smRNA	133,894	20,076	137	14.99%

371 *When allowing up to two missing species per site, see Supplementary Data for numbers of sites with varying levels
372 of missing data.
373

374 *Bioinformatic analysis of turtle genomes yielded well-resolved species relationships*

375 The inferred relationships among turtle lineages based on single-copy orthologs were consistent
376 with previous studies using phylogenomic markers (Crawford et al. 2015; Shaffer et al. 2017),
377 with full statistical support for all resulting branches based on posterior probabilities in the
378 ASTRAL tree, and bootstrap replicates and SH-like approximate likelihood in the supermatrix
379 tree (Fig. 3, Supplementary Fig. 1, Supplementary Fig. 2). Hereafter we refer to these results
380 collectively as the turtle species tree. The turtle species tree supports the reciprocal monophyly
381 of Pleurodira and Cryptodira (Fig. 4). Within cryptodires, there is a split between the
382 Trionychoidea (soft-shelled turtles) and the Durocryptodira (Crawford et al. 2015), and a further
383 subdivision of Durocryptodira into the Testudinoidea (pond turtles, big-headed turtles,
384 geoemyds, and tortoises) and Americhelydia (sea turtles, snapping turtles, mud turtles, and
385 *Dermatemys*; Joyce et al. 2013). Within the Testudinoidea, the turtle species tree supports
386 Testuguria (geoemyds and tortoises; Joyce et al. 2004), and a *Platysternon* + Emydidae clade
387 ("Platyemidae"; Barley et al. 2010; Crawford et al. 2015).
388

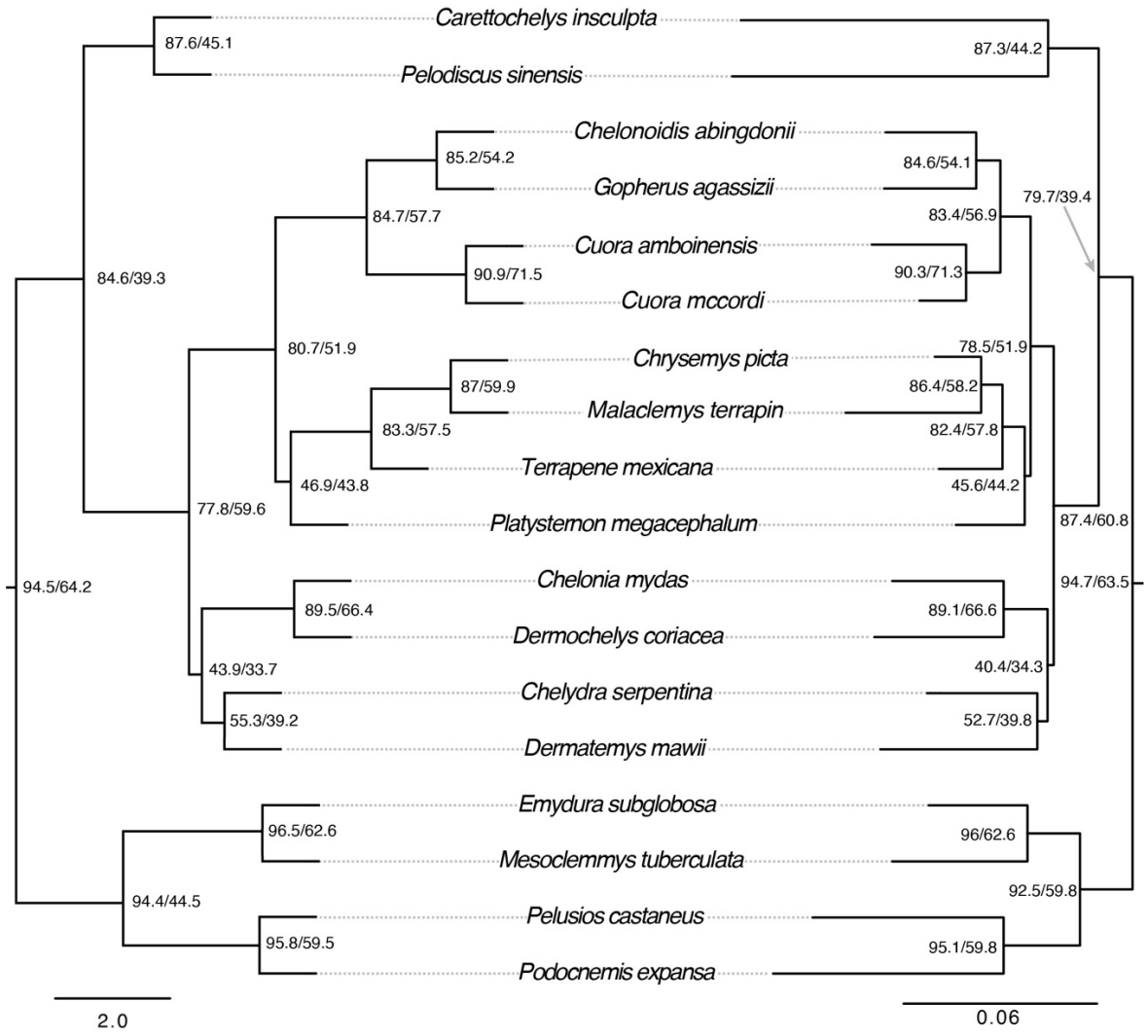


Figure 3. Species trees for turtles based on high-confidence single-copy orthologs. We used 2,513 single-copy orthologs ≥1,500 bp in length to infer a multilocus coalescent-consistent tree based on gene genealogies (ASTRAL, left) and a maximum likelihood tree (ML, right) from a concatenated and partitioned supermatrix in IQ-TREE. All branches received 100% support from local posterior probabilities (ASTRAL), 1,000 bootstrap replicates (IQ-TREE), and approximate likelihood ratio test (IQ-TREE). Node labels indicate gene and site concordance factors before and after the slash, respectively. Branch lengths are in terms of coalescent units (ASTRAL) and substitutions per site (ML). Outgroups (*Gallus, Alligator, Homo*) are not shown.

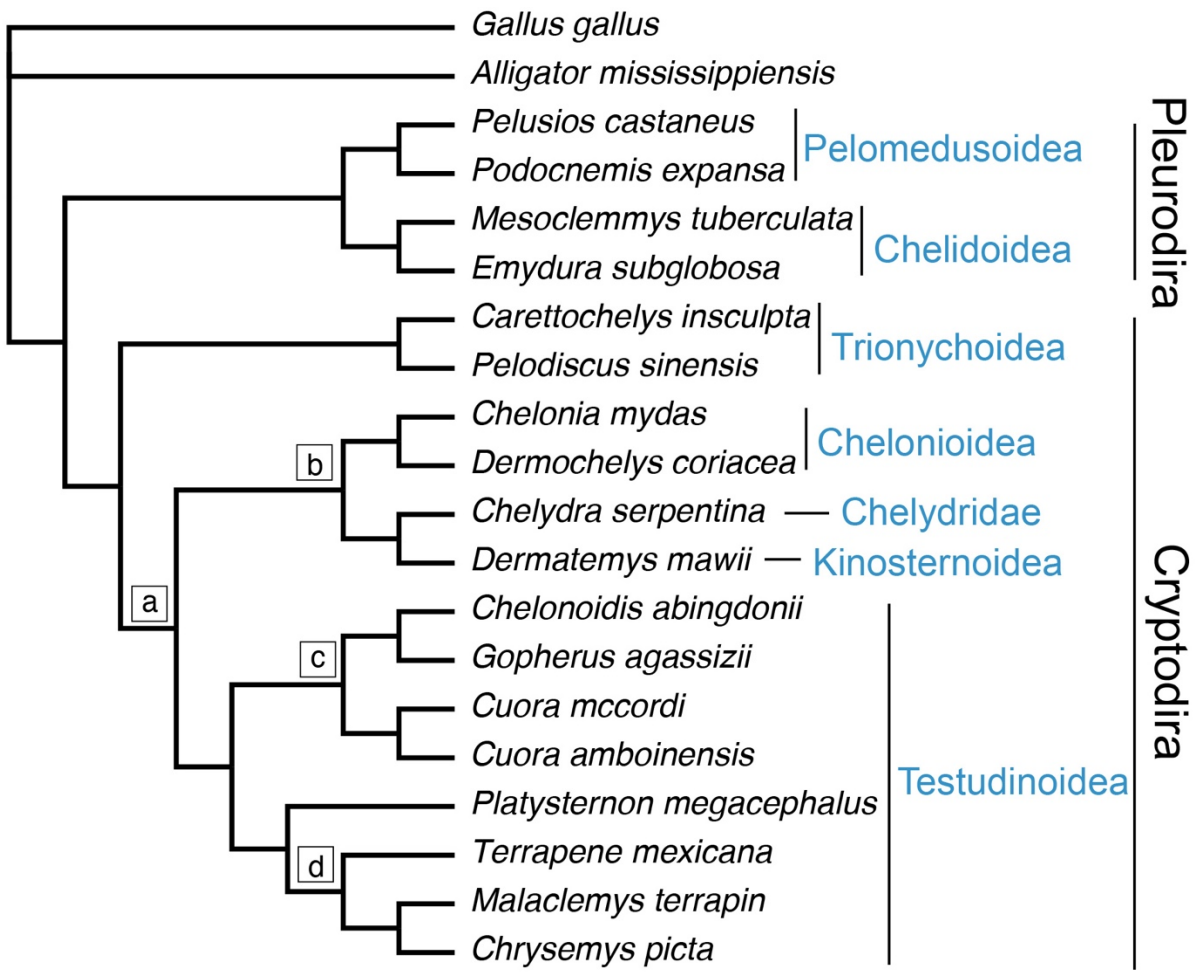


Figure 4. The phylogenetic relationships of higher turtle taxa. Cladogram depicting the turtle species tree in agreement with results from ASTRAL and concatenated partitioned maximum likelihood analyses as well as most SVDquartets analyses using the mapped data (although these analyses lacked outgroups). Superfamilies or clades are indicated by horizontal shaded text. Additional key clades are labelled with letters in boxes. (a) Durocryptodira, (b) Americhelydia, (c) Testuguria, (d) Emydidae. Additional outgroup *Homo sapiens* not shown.

SVDquartets trees from the genome-wide, all-mapped data as well as from CDS, introns, and intergenic regions were consistent with the species tree based on single-copy orthologs (Fig. 4, Supplementary Fig. 3-11; Supplementary Data). When there was disagreement between trees from mapped regions and the species tree, the difference was often in the placement of *Platysternon* (Fig. 5). For instance, when allowing zero missing gapless parsimony-informative

sites from the all-turtle data, 5' UTR placed *Platysternon* as the outgroup of Geoemydidae + Testudinidae (435 sites). When allowing up to two missing gapless parsimony-informative sites from the all-turtle data, pseudogenic regions (426 sites) also placed *Platysternon* as the outgroup of Geoemydidae + Testudinidae, while smRNA (137 sites) agreed with the species tree and all-mapped topologies. Meanwhile, at multiple levels of allowed missing data the Testudinoidea-only mapped sites weakly supported *Platysternon* as the sister taxon to Geoemydidae + Testudinidae + Emydidae (Supplementary Data).

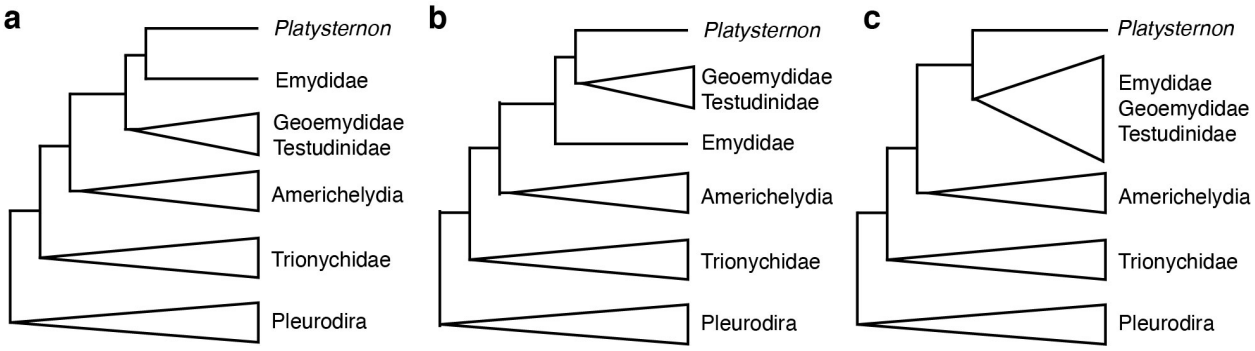


Figure 5. Alternative topologies in turtle phylogenetics based on markers from different regions of the genome. (a) *Platysternon* as the sister taxon to Emydidae to the exclusion of Testuguria (Geoemydidae + Testudinidae), consistent with the turtle species tree (current study), Crawford et al. (2015), and Shaffer et al. (2017). (b) *Platysternon* as the sister taxon to Testuguria with the exclusion of Emydidae, consistent with 17% of the single-copy ortholog genes trees, mapped 5'-UTR parsimony-informative sites data with zero missing data, mapped pseudogene parsimony-informative sites when allowing 2 missing taxa per site, and Krenz et al. (2005). (c) *Platysternon* as the sister taxon to Geoemydidae + Testudinidae + Emydidae, consistent with 20% of single-copy ortholog gene trees and Shaffer et al. (1997).

428 *Heterogeneity is prevalent at well resolved nodes in the turtle phylogeny*

429 While results from our single-copy orthologs and biallelic sites analyses of turtles are consistent
430 and largely in agreement with other studies based on 100% branch support (Crawford et al. 2015;
431 Shaffer et al. 2017), some genomic regions, either from single-copy orthologs or from genome
432 feature annotations, yielded genealogies conflicting with the turtle species tree. For instance,
433 based on normalized RF distances most gene trees from single-copy orthologs were similar to the
434 species tree, although there was a long tail to this distribution indicating gene trees with greater
435 dissimilarity (Fig. 6a). SVDquartets trees using mapped data from CDS, introns, and intergenic
436 regions were consistent with both the single-copy ortholog species tree and a SVDquartets tree
437 derived from the all-mapped data. In contrast, mapped pseudogenic DNA, 5' UTR, and smRNA
438 deviated from the all-mapped data in terms of their RF distances (Supplementary Fig. 12).

439 We also found a high degree of heterogeneity based on gene and site concordance factors
440 throughout the turtle phylogeny (Fig. 3, Fig. 6, Supplementary Fig. 1, Supplementary Fig. 2). For
441 instance, while local posterior probabilities were uniformly 100% across branches of the
442 ASTRAL tree, gene concordance factors ranged from 44 to 97 and site concordance factors
443 ranged from 30 to 71 (Fig. 6b-d). Gene and site concordance factors based on the ASTRAL tree
444 were correlated with each other (Figure 6b; R^2=0.55, P=0.0002) and with branch lengths (Fig.
445 5c-d; R^2=0.88, 3.88e-09 and R^2=0.49, P=0.0007, respectively). We also estimated gene and site
446 concordance factors based on the locus trees with respect to the maximum likelihood tree
447 inferred from the concatenated supermatrix, and found a similar range of gene concordance
448 factors (34 to 92) and site concordance factors (30 to 71) despite 100% bootstrap support, as well
449 as a strong correlation between the two concordance measures (R^2=0.60, P=0.0001). Gene and

site concordance factors based on the concatenated partitioned maximum likelihood tree were also correlated with branch lengths (R^2=0.90, 1.02e-09 and R^2=0.58, P=0.0001, respectively).

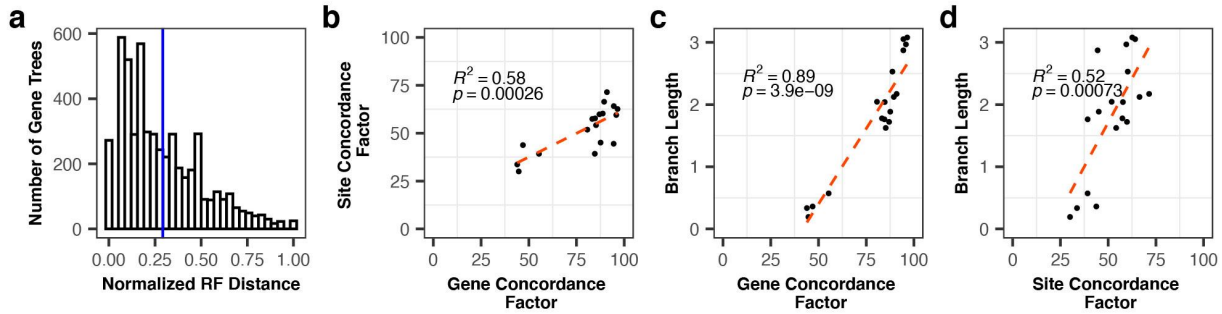


Figure 6. Phylogenomic heterogeneity in turtles. (a) Histogram of RF distances between gene trees and species tree. (b-c) Concordance factors calculated using single-copy ortholog alignments, their gene trees, and the ASTRAL species tree. Results based on locus trees and concatenated supermatrix maximum likelihood tree are in the supplement. (b) Scatterplot depicting the relationship between gene and site concordance factors at each node. (c) Scatterplot depicting the relationship between gene concordance factor and branch length. (d) Scatterplot depicting the relationship between site concordance factor and branch length.

The ancestral branch leading to turtles was assigned a relatively large gene concordance factor and site concordance factor across analyses based on the ASTRAL, supermatrix, and SVDquartets trees, and there was high concordance for reciprocally monophyletic Pleurodira and Cryptodira across genes and sites (Fig. 3, Fig. 4, Supplementary Figs. 3-10). The placement of Trionychoidea as sister taxon to all other cryptodires was assigned a relatively large gene concordance factor but small site concordance factor, suggesting that there are a large number of sites supporting alternative bifurcations. One area with particularly high discordance (i.e. relatively small gene concordance factor and site concordance factor across all datasets) was the placement of the big-headed turtle (*Platysternon*) with respect to Emydidae. Our chi-squared test rejected introgression at this node (Table 3), favoring incomplete lineage sorting or potential

469 model violations as sources of discordance in the position of *Platysternon*. There was also a high
470 degree of discordance associated with the americhelydian taxa (Chelonioidea + Chelydridae +
471 Kinosternoidea), particularly with the ancestral branch leading to Americhelydia as well as with
472 the sister taxon relationship between *Dermatemys* (Kinosternoidea) and *Chelydra* (Chelydridae).
473 Based on single-copy orthologs, minor gene tree frequencies were significantly different at these
474 nodes (Table 4), suggesting that introgression in addition to incomplete lineage sorting and/or
475 model violations may be driving phylogenomic discordance in these groups.
476

477 **Table 3. Gene concordance factors for clades represented by nodes in the ASTRAL species tree phylogeny**
478 **where the minor gene tree frequencies were similar according to a chi-squared test.**

Clade description	Number of possible supporting gene trees	Gene concordance factor	Percent of gene trees supporting first alternative topology	Percent of gene trees supporting second alternative topology
Durocryptodira	2,387	88.73	0.54	0.21
Chelonioidea	1,755	89.46	1.48	0.8
Testudinoidea	2,435	80.66	0.57	0.7
Testuguria	2,255	84.7	0.62	0.8
Testudinidae	1,989	85.22	4.17	5.38
Cuora	1,932	90.94	2.69	2.07
Emydidae + Platysternidae	2,270	46.92	17.05	19.52
Malachlemys + *Chrysemys*	1,730	86.99	5.03	4.39
Pelomedusoidea	2,195	95.81	1.32	1.14
Chelidoidea	2,138	96.49	0.8	1.08

479

480 **Table 4. Gene concordance factors for clades represented by nodes in the ASTRAL species tree phylogeny**
481 **where the minor gene tree frequencies were statistically different according to a chi-squared test.**

Clade description	Number of possible supporting gene trees	Gene concordance factor	Percent of gene trees supporting first alternative topology	Percent of gene trees supporting second alternative topology
Archosauria	2,144	44.73	10.21	43.38
Testudines	2,254	94.5	0.35	1.24

Cryptodira	2,390	84.56	1.63	5.98
Trionychoidea	1,920	87.6	2.24	3.85
Americhelydia	2,335	43.94	17.69	13.83
Kinosternoidea + Chelydridae	1,970	55.33	14.01	9.95
Emydidae	1,924	83.26	3.12	1.87
Pleurodira	2,396	94.45	0.29	1.34

Patterns of substitutional saturation, compositional heterogeneity, and codon usage bias in turtle phylogenomics

We measured the slope of the line in a regression between uncorrected and corrected genetic distances for 685 single-copy ortholog alignments with complete taxon sampling. The slopes ranged from y=0.47x (most saturated) to y=0.91x (least saturated), but were skewed towards low levels of saturation, with a mean of y=0.76x, a first quartile of y=0.72x, and a third quartile of y=0.81x (Supplementary Fig. 13). When analyzed by codon position, we found that first and second codon positions were unsaturated (y=0.86x) while the slope of the line between corrected and uncorrected distances at third codon positions were consistent with saturation (y=0.57x) (Supplementary Fig. 14). Saturation did not have an effect on the topology of the species tree, as ASTRAL trees derived from the most and least saturated sets of genes were identical (Supplementary Data). In addition, all codon partitions regardless of their saturation levels produced maximum likelihood trees with intra-Testudines relationships that were identical in topology to the species tree, and the maximum likelihood tree estimated based on amino acid partitions was also identical in topology to the species tree (Supplementary Data). There was some disagreement in the placement of turtles in the amniote phylogeny that appeared to be driven by substitutional saturation; for instance, the maximum likelihood trees based on first and

500 second codon positions from saturated genes and all sets of third codon positions supported
501 turtles as the sister taxon to crocodilians (Supplementary Data; see Discussion).
502 We found that compositional heterogeneity was most prominent in the saturated third
503 position and first and second position datasets, as well as in the saturated first and second
504 position dataset (Supplementary Data). The unsaturated alignment of first and second codon
505 positions exhibited the least amount of compositional heterogeneity, with five species differing
506 significantly from the rest of the alignment; however, these species were distributed randomly
507 across the turtle species tree. Correspondence analyses of the RSCU revealed that some codons,
508 particularly CCG (Proline), ACG (Threonine), and GCG (Alanine) were favored in different
509 species (Supplementary Fig. 15). However, there was little evidence for a phylogenetic signal in
510 this pattern within turtles since codon usage bias was dispersed randomly among species, and the
511 only outlier along the second dimension was an outgroup taxon (*Gallus*).
512
513 *Heterogeneity at different timescales in turtle evolution*
514 Results from the replicate BEAST analyses consistently underestimated turtle divergence times
515 with respect to results from the penalized likelihood analysis as well as those of Thomson et al.
516 (2020) (Supplementary Data), although are closer to estimates from Chiari et al. (2012) and
517 Pereira et al. (2017). Across locus types from the SISRS analyses, we found that site
518 concordance factors were not strongly correlated with estimated node ages (Fig. 7,
519 Supplementary Fig. 16). For most locus types the relationship between site concordance factor
520 and divergence time at a given node trended slightly negatively, suggesting a pattern of higher
521 concordance at younger nodes, although these comparisons were non-significant (Fig. 7b,
522 $P>0.1$). The one exception to this trend was for pseudogenes, for which the relationship between

site concordance factors and estimated divergence times trended positive albeit without significance or with only moderate significance (P=0.072 using Thomson et al. 2020's divergence time estimates). The relationship between site concordance factors and estimated DNA substitution rates trended slightly positive for all locus types from the SISRS analysis, but was non-significant in all comparisons (Supplementary Fig. 17).

We also found a nonsignificant negative trend in the relationship between site concordance factor and divergence time at a given node based on both the ASTRAL and supermatrix trees based on single-copy orthologs (Fig. 8a-c; Supplementary Fig. 18. However, we found a positive (albeit non-significant) trend in the relationship between the estimated divergence time at a given node and its gene concordance factor when using both trees derived from single-copy orthologs (Fig. 8e-g; Supplementary Fig. 18). There was no clear relationship between the site concordance factor of a given node and the estimated substitution rate at the preceding branch (Fig. 8d), and a moderately significant positive trend ($P \leq 0.05$) for the relationship between the gene concordance factor of a given node and the estimated substitution rate at the preceding branch (Fig. 8h; Supplementary Fig. 18).

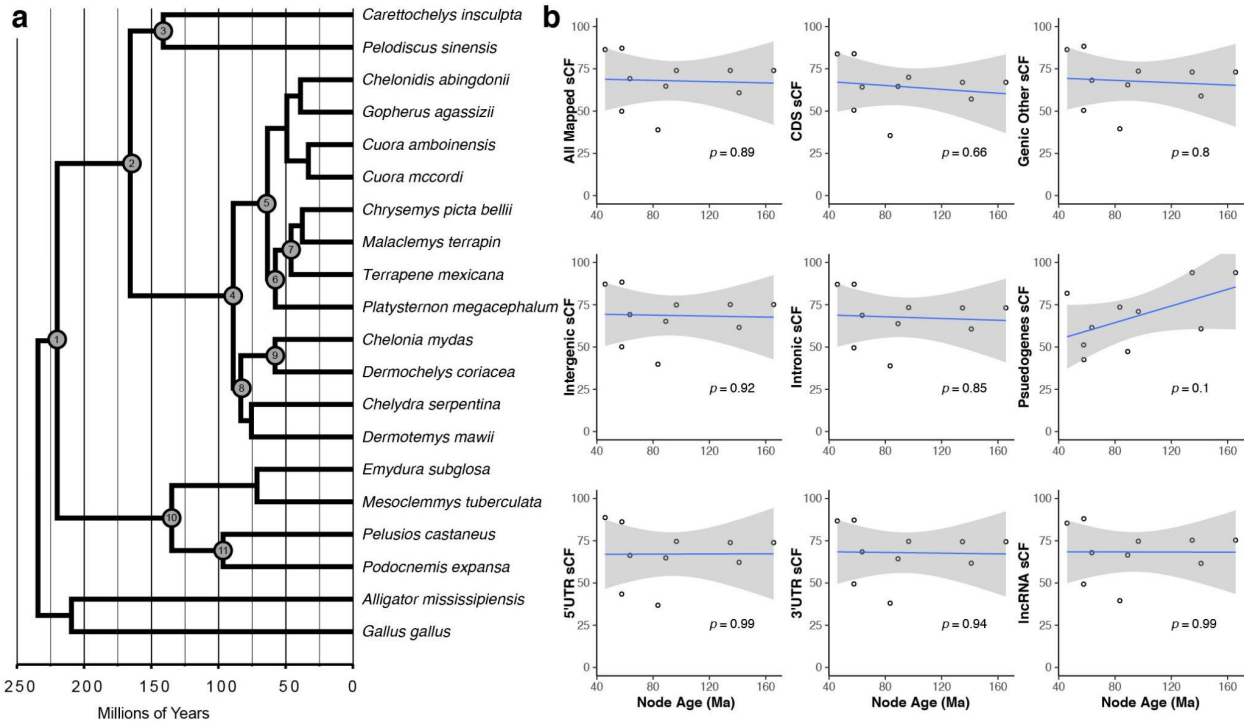


Figure 7. Site concordance and estimated node ages in turtle evolution. (a) Chronogram showing estimated divergence times in millions of years based on penalized likelihood and fossil calibrations of labelled nodes. 1. Testudines, 2. Cryptodira, 3. Trionychoidea, 4. Durocryptodira, 5. Geoemydidae, 6. Platyemidae, 7. Emydidae, 8. Americhelydia, 9. Chelonioidea, 10. Pleurodira, 11. Pelomedusoidea. (b) Site concordance factors (sCF) for each node in phylogenies inferred from nine locus types and the estimated node age in millions of years ago (Ma) based on penalized likelihood.

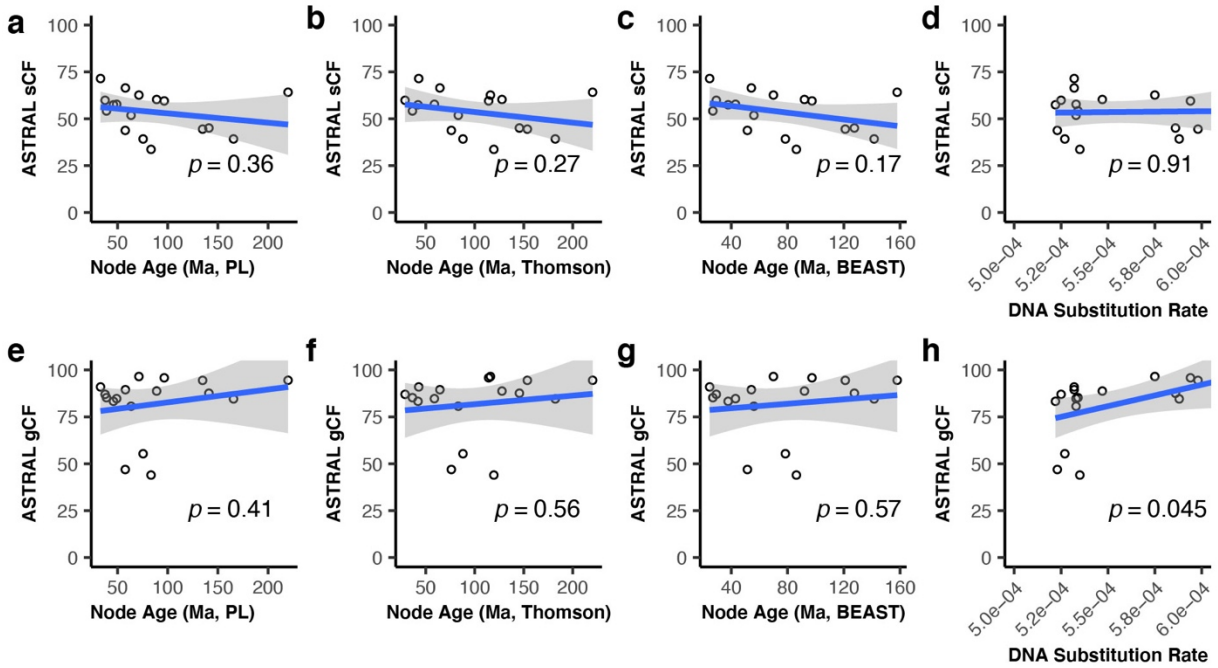


Figure 8. Concordance at genes and sites, estimated node ages, and DNA substitution rates in turtle evolution. Results shown are from comparisons of single-copy ortholog alignments and their gene trees to the species tree inferred using ASTRAL. There was a negative trend between site concordance factors (sCF) and estimated node age (in millions of years or Ma) based on penalized likelihood (PL) from the current study (a), estimated node ages in Ma from Thomson et al. (2020) (b), estimated node ages in Ma from the BEAST analysis in the current study (c), and no discernable trend between sCF and estimated DNA substitution rates in terms of substitutions per site per Ma (d). There was a positive trend between gene concordance factors (gCF) and estimated node ages (e, f, g) and a moderately significant positive relationship between gene concordance factors and estimated DNA substitution rates (h). Results from comparisons of the locus trees and the maximum likelihood tree inferred from the concatenated supermatrix are shown in Supplementary Materials.

Discussion

Heterogeneity has been recognized as an important feature of phylogenomic datasets, and the exploration of the patterns and sources of gene tree and site discordance has become an essential step when reconstructing the branching order of life with molecular data (Kumar et al. 2012;

562 Minh et al. 2020a). To our knowledge, while the role of gene tree-species tree discordance in
563 estimating the position of turtles within the amniote phylogeny has been the focus of a few
564 studies (Chiari et al. 2012), there has not been an assessment of phylogenomic heterogeneity and
565 its effects on the assessment of intra-turtle relationships until now. Here, we have constructed a
566 robust phylogenomic dataset for turtles based on whole genomes in order to sample the major
567 turtle lineages. We identified splits in the turtle phylogeny and regions of the turtle genome
568 associated with gene tree and site heterogeneity, and investigated how potential sources of this
569 discordance such as introgression, incomplete lineage sorting, and technical issues leading to
570 model violations may have contributed to conflicting results in phylogenetic estimation of
571 turtles. The results of our study demonstrate that heterogeneity can be rampant, even in
572 phylogenomic datasets which have yielded fully resolved species trees.

573 Protein-coding genes such as the single-copy orthologs used in the current study are
574 common markers used for phylogenetics, due to their straightforward amplification via PCR or
575 RNA-Seq, as well as relatively straightforward methods for their orthology assignment and
576 alignment. However, the use of coding sequences in phylogenetics is complicated by difficulties
577 with accurate evolutionary modeling due to multiple hits at ancient divergences, a diversity of
578 substitution rates across the genome, and selective constraint potentially limiting the number of
579 variant sites. These problems often scale with the size of phylogenomic datasets and can lead to
580 biased results (Philippe et al. 2011). Therefore, in addition to the single-copy orthologs we also
581 generated a complementary dataset for turtle phylogenomics consisting of millions of
582 parsimony-informative biallelic sites using SISRS. Employment of these sites has multiple
583 advantages: (1) they are easily extracted from multiple sequence alignments; (2) they do not
584 require complicated evolutionary modeling (i.e. they either support or do not support a splitting

585 event); and (3) the sheer number of potential sites allows strict quality filtering without loss of
586 signal (Literman et al. 2021). SISRS resulted in a considerable number of sites collected from
587 both coding, noncoding, and regulatory parts of the genome, including intergenic regions which
588 yielded almost 750,000 parsimony-informative sites for turtles. These locus types sample a wide
589 range of substitution rates, whereas many other phylogenomic methods focus on a much smaller
590 range of locus types (such as UCEs, AHE, or coding genes). Our results show that many
591 different regions of the turtle genome yield consistent support for the turtle species tree, leaving a
592 solid phylogenetic hypothesis for higher turtle systematics. However, inferences from some
593 regions such as pseudogenes and smRNA were based on a relatively small number of parsimony-
594 informative sites, highlighting the need for more comprehensive annotations of noncoding
595 portions of turtle genomes.

596 The relationships between turtles and other reptiles has been a longstanding debate in
597 evolutionary biology, due to conflicting results between studies that use fossils and molecules
598 (Benton et al. 2005; Chiari et al. 2012; Brown and Thomson 2017). Paleontological studies have
599 placed turtles either at the root of the amniote phylogeny, or as sister taxa with various other
600 reptilian groups, some of which are long extinct (Gaffney 1980). In the genomic era, analyses of
601 nucleotides and protein sequences have been mostly in agreement, placing turtles as the sister
602 taxon to archosaurs (birds and crocodilians) (Shaffer et al. 2013; Wang et al. 2013). Based on
603 2,513 long single-copy orthologs, our ASTRAL tree results supported a turtles + archosaurs
604 clade, while the maximum likelihood tree based on the concatenated and partitioned supermatrix
605 supported a turtles + crocodilians clade that excludes birds. Several lines of evidence suggest that
606 the latter result is due to poorly fit evolutionary models. First, the maximum likelihood
607 phylogeny inferred using partitioned amino acids, which accounts for saturation since proteins

608 are less degenerate than nucleotides, yielded the turtles + archosaurs relationship. Second, we
609 found that first and second codon positions from non-saturated genes supported turtles +
610 archosaurs, while first and second codon positions from saturated genes as well as all third codon
611 positions agreed with turtles + crocodilians (Supplementary Data). Brown and Thomson (2017)
612 found that duplicated genes in the dataset published by Chiari et al. (2012) dataset were more
613 likely to reject a turtles + archosaurs clade. Meanwhile, our reliance on single-copy orthologs
614 from OrthoDB greatly reduces the role of paralogy in the conflict at this node. We suggest that
615 the correct placement of turtles within the amniote phylogeny using genomic data has been
616 hampered by substitutional saturation at such ancient divergences, and that when accounting for
617 these artifacts the data support a turtles + archosaurs clade (Archelosauria; see Crawford et al.
618 2015).

619 When gene tree-species tree discordance is driven largely by technical errors such as
620 saturation, it should be most apparent at older nodes in phylogeny, which are more affected by
621 homoplasy, poor alignments, or model misspecifications. However, with the exception of the
622 turtle-archosaur split, across both single-copy ortholog and SISRS datasets we found no
623 significant correlation between concordance factors and estimated node ages. A slightly negative
624 relationship between site concordance factors calculated from the single-copy ortholog dataset
625 and node age suggests that these deeper splits in turtle evolution might be affected by multiple
626 substitutions; however, this pattern is obscured a slightly positive relationship between gene
627 concordance factors and node age using the same data. In addition, inferred turtle relationships
628 were identical when analyzing different sets of genes separated by levels of substitutional
629 saturation and partitioned by codon position. Thus, while it may be possible that multiple
630 substitutions or other model violations are driving discordance at ancient nodes in the turtle

phylogeny, their effects are either extremely weak or nonexistent within and among most turtle lineages, and the observed patterns may have more to do with errors in rate estimation using molecular data and fossil calibrations.

Our results suggest that population-level processes such as introgression and/or incomplete lineage sorting have played a key role at multiple stages of turtle diversification, leading to confusion in the literature about the placement of at least some taxa. In particular, the phylogenetic position of *Platysternon* has been highly contentious (Joyce et al. 2004, 2021), with early molecular studies in disagreement (Shaffer et al. 1997; Krenz et al. 2005; Parham et al. 2006; Barley et al. 2010). Our turtle species tree results mirror Crawford et al. (2015) and Shaffer et al. (2017), fully supporting *Platysternon* as the sister taxon to Emydidae. However, we found that only 47% of single-copy ortholog gene trees are in agreement with a *Platysternon* + Emydidae clade, and 37% of the remaining gene trees support only two alternative relationships: of these, 17% of gene trees support *Platysternon* as the sister taxon to Testuguria (Geoemydidae + Testudinidae) following Krenz et al. (2005), and 20% support *Platysternon* as the sister taxon to Testudinoidea (Emydidae + Geoemydidae + Testudinidae) following Shaffer et al. (1997). The proportion of gene trees supporting these alternative topologies was similar according to our chi-squared test (Table 3), and when we analyzed gene trees according to their levels of substitutional saturation, we saw no effect on the placement of *Platysternon*. As the most frequent alternate gene trees mirror published alternative hypotheses concerning the relationship of *Platysternon* to other testudinoideans, we suggest that incomplete lineage sorting is a previously unrecognized driver of conflicting phylogenetic results in the higher classification of these groups.

653 Another subject of conflict across previous phylogenetic studies of turtles is the
654 relationship between Chelonioidea, Kinosternoidea, and the Chelydridae, all of which are now
655 considered to be in the monophyletic Americhelydia (Barley et al. 2010; Joyce et al. 2013;
656 Crawford et al. 2015). We found consistent support for Americhelydia across datasets and
657 methods, as well as a sister taxon relationship between Chelydridae and Kinosternoidea, yet with
658 considerable gene tree-species tree discordance with single-copy orthologs (Fig. 3) and site
659 discordance with SISRS data (Supplementary Fig. 3-11). For instance, only 44% of decisive
660 gene trees from single-copy orthologs contained an Americhelydia clade, and only 55% of
661 decisive gene trees supported Chelydridae + Kinosternoidea. As with *Platysternon*, comparison
662 of trees when accounting for saturation or codon position did not affect the placement of
663 americhelydians. The frequencies of minor gene trees disagreeing with these splits were
664 statistically different (Table 4), suggesting both introgression and incomplete lineage sorting may
665 have played a role in the diversification of these taxa. To our knowledge, the role of
666 hybridization during the speciation events at the deeper splits of the turtle phylogeny has not
667 been investigated, although contemporary hybridization between sea turtle genera with up to ~65
668 Ma of divergence is well documented (Karl et al. 1995; Arantes et al. 2020). As these
669 occurrences are associated with low reproductive success (Arantes et al. 2020), it will be
670 important to determine if episodic introgression is a feature of the long-term evolution of turtles.
671 Some recent molecular analyses of turtle relationships are limited in terms of the size of
672 the molecular dataset, yet have more complete taxonomic sampling than our study, including up
673 to 85% of extant species (Pereira et al. 2017; Colston et al. 2020; Thomson et al. 2021). A denser
674 sampling of lineages maximizes the molecular signal for testing hypotheses about past
675 diversification events as well as the relationship between evolutionary distinctiveness and

676 extinction risk. In contrast, our study is limited in terms of taxonomic sampling due to our
677 reliance on complete genomic sequences and assemblies for turtles, which are lagging compared
678 to other amniote groups. As our results demonstrate widespread discordance in turtle
679 phylogenomic datasets, additional lineages should be targeted for genomic sequencing in order
680 to improve our ability to accurately estimate the evolutionary distinctiveness of certain species.
681 These studies will be crucial in setting future conservation priorities for turtles.
682

683 **Acknowledgements**

684 This research was funded by the National Science Foundation (#1812291 to RL) and by startup
685 funds provided to MT by the School of Informatics, Computing, and Cyber Systems at Northern
686 Arizona University. We would like to acknowledge the Monsoon computing cluster at Northern
687 Arizona University (https://nau.edu/high-performance-computing/) for providing the
688 computational resources necessary to carry out this study.
689

690 **Data Availability**

691 The bioinformatic workflows used to extract and analyze single-copy orthologs from genome
692 assemblies are available at https://github.com/mibyars/busco_phylo_pipeline. Data generated for
693 this study including alignments and trees are available at the Harvard Dataverse
694 (https://doi.org/10.7910/DVN/WYQRPC).
695

696 **Literature Cited**

697 Arantes L.S., Ferreira L.C.L., Driller M., Repinaldo Filho F.P.M., Mazzoni C.J. and Santos F.R.
698 2020. Genomic evidence of recent hybridization between sea turtles at Abrolhos

699 Archipelago and its association to low reproductive output. Sci. Rep. 10:12847.

700 Barley A.J., Spinks P.Q., Thomson R.C., Shaffer H.B. 2010. Fourteen nuclear genes provide
701 phylogenetic resolution for difficult nodes in the turtle tree of life. Mol. Phylogenet. Evol.
702 55:1189–1194.

703 Benton M. 2005. Vertebrate Paleontology. Oxford, UK: Blackwell Science Ltd.

704 Boisvert S., Laviolette F., Corbeil J. 2010. Ray: Simultaneous Assembly of Reads from a Mix of
705 High-Throughput Sequencing Technologies. J. Comput. Biol. 17:1519–1533.

706 Borowiec M.L. 2016. AMAS: a fast tool for alignment manipulation and computing of summary
707 statistics. PeerJ. 4:e1660.

708 Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A.,
709 Heled J., Jones G., Kühnert D., Maio N.D., Matschiner M., Mendes F.K., Müller N.F.,
710 Ogilvie H.A., Plessis L. du, Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard
711 M.A., Wu C.-H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. BEAST 2.5: An
712 advanced software platform for Bayesian evolutionary analysis. PLOS Comput. Biol.
713 15:e1006650.

714 Brown J.M., Thomson R.C. 2017. Bayes Factors unmask highly variable information content,
715 bias and extreme influence in phylogenomic analyses. Syst. Biol. 66:517–530.

716 Burbrink F.T., Grazziotin F.G., Pyron R.A., Cundall D., Donnellan S., Irish F., Keogh J.S.,
717 Kraus F., Murphy R.W., Noonan B., Raxworthy C.J., Ruane S., Lemmon A.R., Lemmon
718 E.M., Zaher H. 2020. Interrogating Genomic-Scale Data for Squamata (Lizards, Snakes,
719 and Amphisbaenians) Shows no Support for Key Traditional Morphological
720 Relationships. Syst. Biol. 69:502-520.

721 Chiari Y., Cahais V., Galtier N., Delsuc F. 2012. Phylogenomic analyses support the position of

722 turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 65 (2012).
723 Chifman J., Kubatko L. 2014. Quartet Inference from SNP Data Under the Coalescent Model.
724 Bioinformatics. 30:3317–3324.
725 Cooper E.D. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends
726 Plant Sci. 19:576–582.
727 Colston T.J., Kulkarni P., Jetz W., Pyron R.A. 2020. Phylogenetic and spatial distribution of
728 evolutionary diversification, isolation, and threat in turtles and crocodilians (non-avian
729 archosauromorphs). BMC Evol. Biol. 20:81.
730 Cox C.J., Li B., Foster P.G., Embley T.M., Civáň P. 2014. Conflicting Phylogenies for Early
731 Land Plants are Caused by Composition Biases among Synonymous Substitutions. Syst.
732 Biol. 63:272–279.
733 Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012.
734 More than 1000 ultraconserved elements provide evidence that turtles are the sister group
735 of archosaurs. Biol. Lett. 8:783–786.
736 Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J., Henderson
737 J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of turtles. Mol.
738 Phylogenet. Evol. 83:250–257.
739 Donoghue P.C.J., Benton M.J. 2007. Rocks and clocks: calibrating the Tree of Life using fossils
740 and molecules. Trends Ecol. Evol. 22:424–431.
741 Dudchenko O., Batra S.S., Omer A.D., Nyquist S.K., Hoeger M., Durand N.C., Shamim M.S.,
742 Machol I., Lander E.S., Aiden A.P., Aiden E.L. 2017. De novo assembly of the *Aedes*
743 *aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 356:92–95.
744 Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S.,

745 Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and
746 testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol.
747 Phylogenet. Evol. 94:447–462.
748 Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A Phylogenomic Perspective on the
749 Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved
750 Elements (UCEs). PLOS ONE. 8:e65923.
751 Foster P.G. 2004. Modeling Compositional Heterogeneity. Syst. Biol. 53:485–495.
752 Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos. Trans.
753 R. Soc. B Biol. Sci. 363:4023–4029.
754 Garland T., Bennett A.F., Rezende E.L. 2005. Phylogenetic approaches in comparative
755 physiology. J. Exp. Biol. 208:3015-3035.
756 Gaffney E. 1980. Phylogenetic relationships of the major groups of amniotes. In: Panchen A.
757 editor. The terrestrial environment and the origin of land vertebrates. London: Academic
758 Press. P. 593-610.
759 Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai
760 W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-
761 Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-
762 Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C.,
763 Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic
764 D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., Rasilla M.
765 de la, Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney
766 E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S.
767 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.

768 Hime P.M., Lemmon A.R., Lemmon E.C.M., Prendini E., Brown J.M., Thomson R.C., Kratovil
769 J.D., Noonan B.P., Pyron R.A., Peloso P.L.V., Kortyna M.L., Keogh J.S., Donnellan
770 S.C., Mueller R.L., Raxworthy C.J., Kunte K., Ron S.R., Das S., Gaitonde N., Green
771 D.M., Labisko J., Che J., Weisrock D.W. 2021. Phylogenomics Reveals Ancient Gene
772 Tree Discordance in the Amphibian Tree of Life. Syst. Biol. 70:48-66.
773 Huson D.H., Klöpper T., Lockhart P.J., Steel M.A. 2005. Reconstruction of reticulate networks
774 from gene trees. Springer, Berlin, Heidelberg. P. 233-249.
775 Irisarri I., Baurain D., Brinkmann H., Delsuc F., Sire J.-Y., Kupfer A., Petersen J., Jarek M.,
776 Meyer A., Vences M., Philippe H. 2017. Phylotranscriptomic consolidation of the jawed
777 vertebrate timetree. Nat. Ecol. Evol. 1:1370–1378.
778 IUCN 2021. The IUCN Red List of Threatened Species. Version 2021-2.
779 https://www.iucnredlist.org. Downloaded on November 1, 2021.
780 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz
781 B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L.,
782 Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian
783 S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K.,
784 Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X.,
785 Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H.,
786 Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F.,
787 Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B.,
788 Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q.,
789 Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K.,
790 Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun

M., Fjeldså J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.-P., O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C., McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. Science. 346:1320–1331.

Jeffroy O., Brinkmann H., Delsuc F., Philippe H. 2006. Phylogenomics: the beginning of incongruence? 22:225-231.

Joyce W.G. 2007. Phylogenetic Relationships of Mesozoic Turtles. Bull. Peabody Mus. Nat. Hist. 48:3–102.

Joyce W.G., Anquetin J., Cadena E.-A., Claude J., Danilov I.G., Evers S.W., Ferreira G.S., Gentry A.D., Georgalis G.L., Lyson T.R., Pérez-García A., Rabi M., Sterli J., Vitek N.S., Parham J.F. 2021. A nomenclature for fossil and living turtles using phylogenetically defined clade names. Swiss J. Palaeontol. 140:5.

Joyce W.G., Parham J.F., Gauthier J.A. 2004. Developing a protocol for the conversion of rank-based taxon names to phylogenetically defined clade names, as exemplified by turtles. J. Paleontol. 78:989–1013.

Joyce W.G., Parham J.F., Lyson T.R., Warnock R.C.M., Donoghue P.C.J. 2013. A divergence dating analysis of turtles using fossil calibrations: an example of best practices. J. Paleontol. 87:612–634.

Kalyaanamoorthy S., Minh B.Q., Wong T.K.F., von Haeseler A., Jermiin L.S. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat. Methods. 14:587–589.

814 Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7:
815 Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.
816 Karl S.A., Bowen B.W., Avise J.C. 1995. Hybridization among the ancient mariners:
817 Characterization of marine turtle hybrids with molecular genetic assays. J. Hered.
818 86:262–268.
819 Krenz J.G., Naylor G.J.P., Shaffer H.B., Janzen F.J. 2005. Molecular phylogenetics and
820 evolution of turtles. Mol. Phylogenet. Evol. 37:178–191.
821 Kubatko L.S., Degnan J.H. 2007. Inconsistency of Phylogenetic Estimates from Concatenated
822 Data under Coalescence. Syst. Biol. 56:17–24.
823 Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and
824 Truth in Phylogenomics. Mol. Biol. Evol. 29:457–472.
825 Kumar S., Stecher G., Li M., Knyaz C., Tamura K. 2018. MEGA X: Molecular Evolutionary
826 Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35:1547–1549.
827 Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. Partitionfinder: combined selection of
828 partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol.
829 29:1695–1701.
830 Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods.
831 9:357–359.
832 Lê S., Josse J., Husson F. 2008. FactoMineR: An R Package for Multivariate Analysis. J. Stat.
833 Softw. 25: 1–18.
834 Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored Hybrid Enrichment for Massively
835 High-Throughput Phylogenomics. Syst. Biol. 61:727–744.
836 Literman R., Schwartz R. 2021. Genome-Scale Profiling Reveals Noncoding Loci Carry Higher

Proportions of Concordant Data. Mol. Biol. Evol.

Liu L., Zhang J., Rheindt F.E., Lei F., Qu Y., Wang Y., Zhang Y., Sullivan C., Nie W., Wang J., Yang F., Chen J., Edwards S.V., Meng J., Wu S. 2017. Genomic evidence reveals a radiation of placental mammals uninterrupted by the KPg boundary. Proc. Nat. Acad. Sci. E7282-E7290.

Lopes F., Oliveira L.R., Kessler A., Beux Y., Crespo E., Cárdenas-Alayza S., Majluf P., Sepúlveda M., Brownell R.L. Jr., Franco-Trecu V., Páez-Rosas D., Chaves J., Loch C., Robertson B.C., Acevedo-Whitehouse K., Elorriaga-Verplancken F.R., Kirkman S.P., Peart C.R., Wolf J.B.W., Bonatto S.L. 2021. Phylogenomic Discordance in the Eared Seals is best explained by Incomplete Lineage Sorting following Explosive Radiation in the Southern Hemisphere. Syst. Biol. 70:786–802.

Lynch M. 2007. The Origins of Genome Architecture. Sinauer Associates Sunderland, MA.

Minh B.Q., Hahn M.W., Lanfear R. 2020a. New Methods to Calculate Concordance Factors for Phylogenomic Datasets. Mol. Biol. Evol. 37:2727–2733.

Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A., Lanfear R. 2020b. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol. Biol. Evol. 37:1530–1534.

Mölder F., Jablonski K.P., Letcher B., Hall M.B., Tomkins-Tinch C.H., Sochat V., Forster J., Lee S., Twardziok S.O., Kanitz A., Wilm A., Holtgrewe M., Rahmann S., Nahnsen S., Köster J. 2021. Sustainable data analysis with Snakemake. https://f1000research.com/articles/10-33.

Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F., Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021. Disentangling Sources of Gene

860 Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in
861 Amaranthaceae s.l. Syst. Biol. 70:219-235.
862 Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol.
863 5:568–583.
864 Parham J.F., Feldman C.R., Boore J.L. 2006. The complete mitochondrial genome of the
865 enigmatic bigheaded turtle (Platysternon): description of unusual genomic features and
866 the reconciliation of phylogenetic hypotheses based on mitochondrial and nuclear DNA.
867 BMC Evol. Biol. 6:11.
868 Paradis E., Schliep K. 2019. ape 5.0: an environment for modern phylogenetics and evolutionary
869 analyses in R. Bioinformatics. 35:526-528.
870 Peden J. 1999. Analysis of Codon Usage. PhD diss., University of Nottingham.
871 Pereira A.G., Sterli J., Moreira F.R.R., Schrago C.G. 2017. Multilocus phylogeny and statistical
872 biogeography clarify the evolutionary history or major lineages of turtles. Syst. Biol.
873 113:59-66.
874 Philippe H., Brinkmann H., Lavrov D.V., Littlewood D.T.J., Manuel M., Wörheide G., Baurain
875 D. 2011. Resolving difficult phylogenetic questions: why more sequences are not enough.
876 PLOS Biology. 9:e1000602.
877 Philippe H., Sörhannus U., Baroin A., Perasso R., Gasse F., Adoutte A. 1994. Comparison of
878 molecular and paleontological data in diatoms suggests a major gap in the fossil record. J.
879 Evol. Biol. 7:247-265.
880 Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R.
881 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA
882 sequencing. Nature. 526:569–573.

883 Quesada V., Freitas-Rodríguez S., Miller J., Pérez-Silva J.G., Jiang Z.-F., Tapia W., Santiago-
884 Fernández O., Campos-Iglesias D., Kuderna L.F.K., Quinzin M., Álvarez M.G., Carrero
885 D., Beheregaray L.B., Gibbs J.P., Chiari Y., Glaberman S., Ciofi C., Araujo-Voces M.,
886 Mayoral P., Arango J.R., Tamargo-Gómez I., Roiz-Valle D., Pascual-Torner M., Evans
887 B.R., Edwards D.L., Garrick R.C., Russello M.A., Poulakakis N., Gaughran S.J., Rueda
888 D.O., Bretones G., Marquès-Bonet T., White K.P., Caccone A., López-Otín C. 2019.
889 Giant tortoise genomes provide insights into longevity and age-related disease. Nat. Ecol.
890 Evol. 3:87–95.
891 Quinlan A.R., Hall I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic
892 features. Bioinformatics. 26:841–842.
893 Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in
894 Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.
895 Ranwez V., Harispe S., Delsuc F., Douzery E.J.P. 2011. MACSE: Multiple Alignment of Coding
896 SEquences accounting for frameshifts and stop codons. PLOS ONE. 6:e22594.
897 Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
898 Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving
899 incongruence in molecular phylogenies. Nature. 425:798–804.
900 Sanderson M.J. 1997. A nonparametric approach to estimating divergence times in the absence
901 of rate constancy. Mol. Biol. Evol.
902 Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: a
903 penalized likelihood approach. Mol. Biol. Evol. 19:101–109.
904 Sanderson M.J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times
905 in the absence of a molecular clock. Bioinformatics. 19:301–302.

906 Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from
907 Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.
908 Schwartz R.S., Harkins K.M., Stone A.C., Cartwright R.A. 2015. A composite genome approach
909 to identify phylogenetically informative data from next-generation sequencing. BMC
910 Bioinformatics. 16:193.
911 Shaffer H.B., McCartney-Melstad E., Near T.J., Mount G.G., Spinks P.Q. 2017. Phylogenomic
912 analyses of 539 highly informative loci dates a fully resolved time tree for the major
913 clades of living turtles (Testudines). Mol. Phylogenet. Evol. 115:7–15.
914 Shaffer H.B., Meylan P., McKnight M.L. 1997. Tests of turtle phylogeny: molecular,
915 morphological, and paleontological approaches. Syst. Biol. 46:235–268.
916 Shaffer H.B., Minx P., Warren D.E., Shedlock A.M., Thomson R.C., Valenzuela N., Abramyan
917 J., Amemiya C.T., Badenhorst D., Biggar K.K., Borchert G.M., Botka C.W., Bowden
918 R.M., Braun E.L., Bronikowski A.M., Bruneau B.G., Buck L.T., Capel B., Castoe T.A.,
919 Czerwinski M., Delehaunty K.D., Edwards S.V., Fronick C.C., Fujita M.K., Fulton L.,
920 Graves T.A., Green R.E., Haerty W., Hariharan R., Hernandez O., Hillier L.W.,
921 Holloway A.K., Janes D., Janzen F.J., Kandoth C., Kong L., de Koning A.P.J., Li Y.,
922 Literman R., McGaugh S.E., Mork L., O'Laughlin M., Paitz R.T., Pollock D.D., Ponting
923 C.P., Radhakrishnan S., Raney B.J., Richman J.M., St John J., Schwartz T., Sethuraman
924 A., Spinks P.Q., Storey K.B., Thane N., Vinař T., Zimmerman L.M., Warren W.C.,
925 Mardis E.R., Wilson R.K. 2013. The western painted turtle genome, a model for the
926 evolution of extreme physiological adaptations in a slowly evolving lineage. Genome
927 Biol. 14:R28.
928 Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst.

929 Biol. 51:492–508.

930 Singhal S., Colston T.J., Grundler M.R., Smith S.A., Costa G.C., Colli G.R., Moritz C., Pyron
931 R.A., Rabosky D.L. 2021. Congruence and Conflict in the Higher-Level Phylogenetics of
932 Squamate Reptiles: An Expanded Phylogenomic Perspective. Syst. Biol. 70:542–557.

933 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
934 large phylogenies. Bioinformatics. 30:1312–1313.

935 Steenwyk J.L., Buida T.J. III, Labella A.L., Li Y., Shen X.-X., Rokas A. 2021. PhyKIT: a
936 broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data.
937 Bioinformatics. btab096.

938 Swofford D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).
939 Sinauer associates Sunderland, MA.

940 Tarver J.E., dos Reis M., Mirarab S., Moran R.J., Parker S., O'Reilly J.E., King B.L., O'Connell
941 M.J., Asher R.J., Warnow T., Peterson K.J., Donoghue P.C.J., Pisani D. 2016. The
942 interrelationships of placental mammals and the limits of phylogenetic inference.
943 Genome Biol. Evol. 8:330-344.

944 Thomson R.C., Spinks P.Q., Shaffer H.B. 2021. A global phylogeny of turtles reveals a burst of
945 climate-associated diversification on continental margins. Proc. Natl. Acad. Sci. 118.

946 Tollis M., Denardo D.F., Cornelius J.A., Dolby G.A., Edwards T., Henen B.T., Karl A.E.,
947 Murphy R.W., Kusumi K. 2017. The Agassiz's desert tortoise genome provides a
948 resource for the conservation of a threatened species. PloS One. 12:e0177708.

949 Townsend J.P., López-Giráldez F, Friedman R. 2008. The phylogenetic informativeness of
950 nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol.
951 67:437-447.

952 Uetz P., Freed P., Aguilar R., Hošek J. 2021. The Reptile Database. (last accessed August 2021)
953 Wang Z., Pascual-Anaya J., Zadissa A., Li W., Niimura Y., Wang J., Huang Z., Li C., White S.,
954 Xiong Z., Fang D., Wang B., Ming Y., Chen Y., Zheng Y., Kuraku S., Pignatelli M.,
955 Herrero J., Beal K., Nozawa M., Li Q., Wang J., Zhang H., Yu L., Shigenobu S., Wang
956 J., Liu J., Flicek P., Searle S., Wang J., Kuratani S., Yin Y., Aken B., Zhang G., Irie N.
957 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the
958 development and evolution of the turtle-specific body plan. Nat. Genet. 45:701–706.
959 Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G.,
960 Kriventseva E.V., Zdobnov E.M. 2018. BUSCO Applications from Quality Assessments
961 to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35:543–548.
962 Wolf Y.I., Rogozin I.B., Grishin N.V., Koonin E.V. 2002. Genome trees and the tree of life.
963 Trends Genet. TIG. 18:472–479.
964 Zdobnov E.M., Tegenfeldt F., Kuznetsov D., Waterhouse R.M., Simão F.A., Ioannidis P.,
965 Seppey M., Loetscher A., Kriventseva E.V. 2017. OrthoDB v9.1: cataloging evolutionary
966 and functional annotations for animal, fungal, plant, archaeal, bacterial and viral
967 orthologs. Nucleic Acids Res. 45:D744–D749.
968 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree
969 reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.
970