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 86011*

[2]*Department of Biological Sciences, University of Rhode Island, 120 Flagg Road, Kingstown, RI, 02881*

*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 from annotated regions (coding regions, introns, untranslated regions, intergenic, and others) of the western painted turtle (*Chrysemys picta bellii*) genome. 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 incomplete lineage sorting and temporal patterns in phylogenomic heterogeneity across turtle evolution. We found 100% 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, and some nodes in the turtle phylogeny were well-explained by incomplete lineage sorting. There was no clear pattern between site concordance, node age, and DNA substitution rate across most annotated genomic regions, suggesting a relatively uniform proportion of informative sites drive phylogenetic inference across the evolution of turtles. We found more gene concordance at older nodes in the turtle phylogeny, and suggest that, in addition to incomplete lineage sorting, an overall lack of gene informativeness stemming from a slow rate of evolution can confound inferred patterns in turtle phylogenomics, particularly at more recent divergences. 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.

Keywords: turtles, genomes, phylogeny, discordance

41 **Introduction**

42 Next-generation sequencing has revolutionized the understanding of evolutionary relationships
43 by greatly increasing the numbers of sampled loci (Lemmon et al. 2012; Faircloth et al. 2013;
44 Edwards et al. 2016), shedding new light on the branching order of diversification across the
45 history of life. However, phylogenetic studies based on genome-scale data often contain a great
46 deal of heterogeneity in the inferred patterns (Hime et al. 2021; Lopes et al. 2021; Morales-
47 Briones et al. 2021; Singhal et al. 2021), particularly in the form of discordance among sites and
48 gene trees (Kumar et al. 2012). Gene tree-species tree discordance occurs when gene trees are in
49 conflict with the underlying species relationships, while site discordance occurs when a subset of
50 sites support different bifurcations in a species tree, and both factors can lead to biased
51 phylogenetic estimates. Biological factors driving this heterogeneity include: differing
52 evolutionary rates across the genome, incomplete lineage sorting, gene duplication or loss, and
53 hidden paralogy (Pamilo and Nei 1988; Galtier and Daubin 2008). Artifactual factors driving
54 heterogeneity include sequencing errors, as well as substitution model violations such as
55 saturation or multiple hits (Foster 2004; Cooper 2014; Cox et al. 2014). Accounting for gene and
56 site discordance often requires the application of multiple tools in phylogenetic analysis,
57 including partitioning schemes which try to account for differing evolutionary rates and model
58 parameters (Kubatko and Degnan 2007), and multispecies coalescent methods accounting for
59 population-level processes such as incomplete lineage sorting (Edwards et al. 2016). The
60 application of hundreds or even thousands of loci is often an attempt to overcome errors
61 stemming from or driving phylogenomic conflicts, with the goal of reconstructing better-
62 resolved phylogenies by reducing the overall sampling variance (Minh et al. 2020a).

63 An emerging advantage of the genomic era is the theoretical ability to access the whole
64 genome in order to sample loci representing a diversity of evolutionary rates and coalescent
65 histories, potentially overcoming problems from discordance in phylogenomics (Wolf et al.
66 2002; Rokas et al. 2003). The most widely used sampling methods include the use of large
67 numbers of coding sequences (Singhal et al. 2021), anchored hybrid enrichment loci (Lemmon et
68 al. 2012), ultra-conserved elements (Faircloth et al. 2013), and transcriptomes (Irisarri et al.
69 2017), which encompass hundreds or thousands of markers yet are still reduced representations
70 of a small proportion of the genome (Lynch 2007). The development of bioinformatics tools and
71 sequencing technologies have enabled chromosome-scale genome assemblies for non-model
72 organisms (see http://dnazoo.org, Dudchenko et al. 2017), facilitating the understanding of not
73 only which diversification events are associated with discordance, but also the genomic regions
74 or types of loci driving heterogeneous phylogenetic results.

75 Only recently have investigators looked closely at patterns of heterogeneity in order to
76 understand how molecular evolution across the genome affects inferred species relationships,
77 mainly in clades where there has been considerable phylogenetic and taxonomic debate such as
78 squamate reptiles (Burbrink et al. 2020; Singhal et al. 2021), birds (Jarvis et al. 2014; Prum et al.
79 2015), amphibians (Hime et al. 2021), and land plants (Morales-Briones et al. 2021). Meanwhile,
80 discordance in well-resolved phylogenies with high statistical support among inferred branches
81 and agreement across studies has been largely overlooked, leaving a gap in knowledge about the
82 expected range of signals and amount of disagreement between loci when there are differing
83 levels of phylogenetic uncertainty. This would be critical for informing the designs of future
84 phylogenomic studies.

85 Here, we examined the importance of heterogeneity in the phylogenomic estimation of
86 turtles (Order Testudines). Turtles are shelled reptiles with a rich fossil history dating from the
87 Triassic Period (Joyce 2007; Joyce et al. 2021). The approximately 350 extant species of turtles
88 are classified into two main groups, Pleurodira and Cryptodira, which are further divided into 8
89 superfamilies and 14 families (Uetz et al. 2021). Early molecular phylogenetic results disagreed
90 in the placement of some turtle taxa; in particular, the position of the monotypic big-headed
91 turtle (*Platysternon*) and its relationships with or within Testudinoidea (Shaffer et al. 1997;
92 Krenz et al. 2005). Other questions about deeper turtle relationships such as the position of soft-
93 shell turtles (Trionychoidea), the Americhelydiae, and the position of turtles in the amniote
94 phylogeny persisted into the next-generation sequencing era (Crawford et al. 2012, 2015). Since
95 then, phylogenomic studies of turtles have produced fully resolved species trees with 100%
96 statistical support at every node and near complete agreement across studies (Crawford et al.
97 2015; Shaffer et al. 2017).
98 Despite the harmony among phylogenomic studies of turtles using next-generation
99 sequencing data, the amount of underlying heterogeneity in these datasets has not been
100 examined. Also, while complete genome assemblies are available for a wide range of turtle
101 species (Shaffer et al. 2013; Wang et al. 2013; Tollis et al. 2017; Quesada et al. 2019),
102 phylogenomic studies have not taken advantage of this genome-wide information. Our goals
103 were threefold: (1) to revisit discoveries about higher turtle systematics using more substantial
104 proportions of the genome than previous studies; (2) to measure heterogeneity at genes and sites
105 in important events in turtle evolution; and (3) determine what drives heterogeneity in turtle
106 phylogenomics in terms of mechanisms and locus type. We show that heterogeneity in
107 phylogenomic datasets is to be expected, even in well-resolved clades such as turtles, and

108 suggest that a combination of incomplete lineage sorting and lack of gene informativeness due to
109 an overall slow rate of evolution may be driving incongruence between studies of turtle
110 relationships.
111

112 **Methods and Materials**

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

114 We downloaded the complete genome assemblies of 19 turtle species plus three outgroup taxa
115 from NCBI. The species, assembly accession numbers, assembly lengths, and assembly
116 contiguities measured by scaffold N50 are shown in Table 1. Using Benchmarking Universal
117 Single-Copy Orthologs (BUSCO) v3 (Waterhouse et al. 2018), we extracted the nucleotide
118 sequences of 5,310 OrthoDB v9 tetrapod orthologs (Zdobnov et al. 2017) from each genome
119 assembly. BUSCO orthologs are high-confidence genes that persist in eukaryotic genomes as
120 single-copy, which precludes downstream problems in phylogenetics stemming from gene
121 duplication and hidden paralogy (Waterhouse et al. 2018). We aligned each set of orthologs with
122 MAFFT v7.475 (Katoh and Standley 2013), and removed erroneous columns in the alignments
123 with the heuristic method (*automated 1*) in TrimAL v1.4.1 (Capella-Gutierrez et al. 2009).
124 Summary statistics including number of taxa, alignment length, missing data percentage, number
125 and proportion of variable sites, number and proportion of parsimony-informative sites, and GC
126 content for each alignment were estimated with AMAS v3.04 (Borowiec 2016). Trimmed
127 alignments were filtered at a cut-off length ≥1,500 bp and a minimum of 50% taxa
128 representation. We estimated the average pairwise Kimura 2-parameter corrected distance for
129 each alignment with MEGAX (Kumar et al. 2018).
130

131 **Table 1: Turtle genome assemblies analyzed in this study.**

Suborder	Superfamily	Family	Species Name	Accession Number	Assembly length (Gb)	Scaffold N50 (Mb)
Cryptodira	Chelonioidea	Cheloniidae	*Chelonia mydas*	GCF_000344595.1	2.2	3.9
		Dermochelyidae	*Dermochelys coriacea*	GCA_006547105.1	2.1	0.12
	Chelydroidea	Chelydridae	*Chelydra serpentina*	GCA_007922165.1	2.5	16
	Kinosternoidea	Dermatemydidae	*Dermatemys mawii*	GCA_007922305.1	1.9	34.4
	Testudinoidea	Emydidae	*Chrysemys picta*	GCA_000241765.3	2.4	0.44
			Malaclemys terrapin	GCA_001728815.2	2.4	0.44
			Terrapene mexicana	GCF_002925995.2	2.6	24.2
		Geoemydidae	*Cuora amboinensis*	GCA_004028625.2	2.1	0.47
			Cuora mccordi	GCA_003846335.1	2.4	32.6
		Platysternidae	*Platysternon megacephalum*	GCA_003942145.1	2.3	7.2
		Testudinidae	*Chelonoidis abingdonii*	GCA_003597395.1	2.3	1.3
			Gopherus agassizii	GCA_002896415.1	2.2	0.228
	Trionychoidea	Carettochelyidae	*Carettochelys insculpta*	GCA_007922185.1	2.3	45.9
		Trionychidae	*Apalone spinifera*	GCA_000385615.1	1.9	2.3
			Pelodiscus sinensis	GCF_000230535.1	2.2	3.4
Pleurodira	Chelidoidea	Chelidae	*Emydura subglobosa*	GCA_007922225.1	2	44.8
			Mesoclemmys tuberculata	GCA_007922155.1	2	46.4
	Pelomedusoidea	Pelomedusidae	*Pelusios castaneus*	GCA_007922175.1	2	14.1
		Podocnemididae	*Podocnemis expansa*	GCA_007922195.1	2.4	37.1

132 Gb=gigabases; Mb=megabases

133

134 Gene trees were reconstructed for each locus using RAxML v1.8 (Stamatakis 2014). For
135 each ortholog, we generated 10 maximum likelihood trees with the GTRGAMMA substitution
136 model, and performed 100 rapid bootstrap replicates on the best tree. We then collected the best
137 tree for each ortholog to use as evidence to construct a species tree using ASTRAL-III (Zhang et

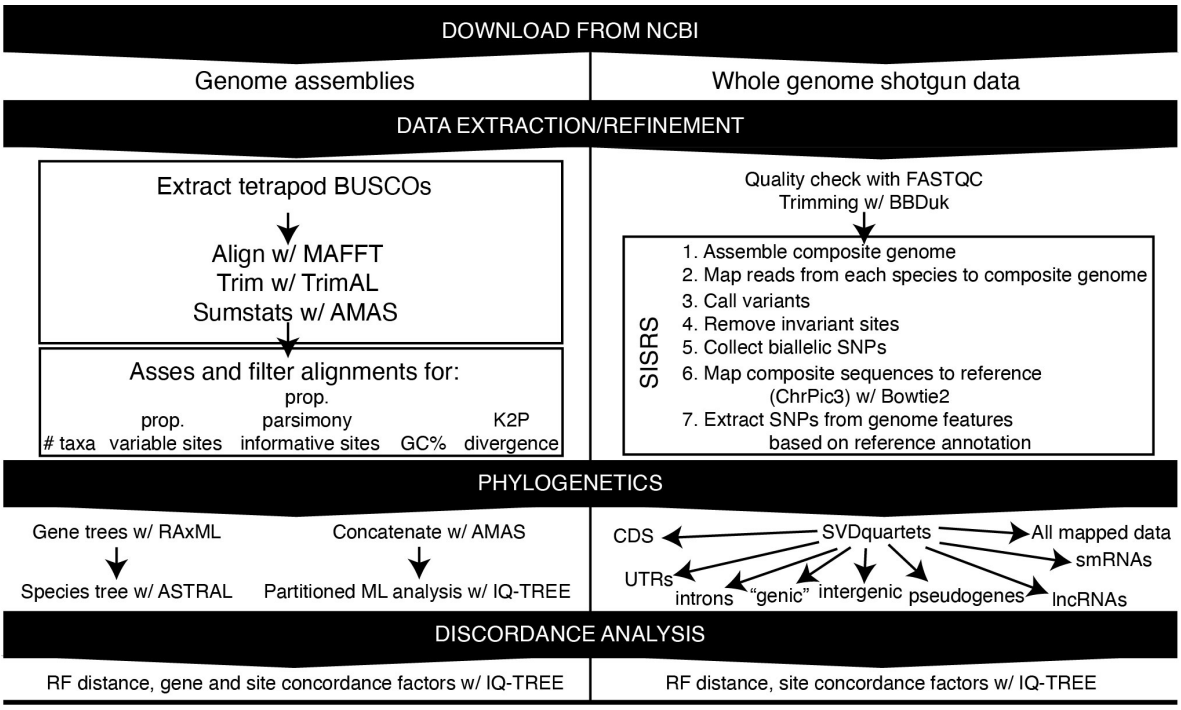
138 al. 2018), assessing branch support with local posterior probabilities (Sayyari and Mirarab 2016).
139 We used Snakemake (Mölder et al. 2021) to automate a reproducible bioinformatic pipeline for
140 BUSCO extraction, filtering, alignment, gene tree and species tree estimation, available at
141 https://github.com/mibyars/busco_phylo_pipeline. We also concatenated the single copy
142 orthologs into a supermatrix with AMAS and performed a partitioned maximum likelihood
143 phylogenetic reconstruction with IQ-TREE v2.1.2 (Minh et al. 2020b). We used ModelFinder
144 (Kalyaanamoorthy et al. 2017) set to TESTMERGE to combine partitions where necessary. We
145 assessed branch support on the concatenated partitioned maximum likelihood tree using 1,000
146 ultrafast bootstrap replicates and the Shimodaira-Hasegawa-like (SH-like) approximate
147 likelihood ratio test (aLRT, Shimodaira 2002).
148
149 *Phylogenomic analysis of biallelic markers*
150 In addition to single-copy orthologs from turtle and outgroup genome assemblies, we constructed
151 sets of biallelic single nucleotide polymorphisms (SNPs) across annotated features of the turtle
152 genome based on paired-end sequence data downloaded from the NCBI Short Read Archive
153 (SRA, Supplementary Data). We assessed each short read dataset with FastQC v0.115 (S.
154 Andrews - http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and trimmed the raw
155 reads with BBDuk v.37.41 (B. Bushnell - sourceforge.net/projects/bbmap/, last accessed
156 December 2020). We used SISRS to generate de novo orthologs from the processed reads
157 (Schwartz et al. 2015; Literman and Schwartz 2021). After subsampling all sequences to a target
158 depth of ~10X per taxon (except for two low-coverage species, Supplementary Data), we
159 assembled a composite genome using Ray v2.2.3-devel (Boisvert et al. 2010) with default
160 parameters and a k-value of 31. We then used SISRS to map the trimmed reads from each

161 species to the composite genome, retaining only uniquely mapped reads to avoid false positives
162 from duplicated or repetitive regions. Bases in the composite genome were replaced according to
163 mapped data from each species if sites were covered by at least three reads and if there was no
164 within-taxon variation; sites not meeting these criteria were masked as "N".

165 For downstream ortholog annotation, we mapped the composite sequences from each
166 species to the reference genome for the western painted turtle (*Chrysemys picta bellii,*
167 GCA_000241765.3) with Bowtie2 v2.3.4 (Langmead and Salzberg 2012), removing contigs that
168 either failed to map or mapped to multiple positions, and removing sites resulting from
169 overlapping coverage of independent scaffolds. We then annotated the mapped composite
170 genome sites parsed from the painted turtle genome as belonging to the following genomic
171 regions (locus types) with BEDTools v2.2.6 (Quinlan and Hall 2010): (1) coding sequences
172 (CDS); (2) 3'-UTR; (3) 5'-UTR; (4) introns; (5) other genic regions not annotated as CDS, UTR,
173 or intronic; (6) long-noncoding RNAs (lncRNAs); (7) pseudogenes; (8) smRNAs (including
174 miRNAs, ncRNAs, rRNAs, scRNAs, smRNAs, snoRNAs, snRNAs, tRNAs, and vaultRNAs).;
175 and (9) intergenic regions. We defined "intergenic" as all regions not included within the other
176 locus types.

177 We then used SISRS to identify parsimony-informative sites along the mapped composite
178 contigs and to remove invariant sites and singletons, as well as any sites containing "Ns" (i.e.
179 polymorphic within species) and overlapping indels or gaps. For the nine annotated genomic
180 regions as well as a set of the complete mapped data (=10 locus types), we concatenated biallelic
181 parsimony-informative sites and inferred locus type phylogenies via SVDQuartets (Chifman and
182 Kubatko 2014) in PAUP* v4.0a (build 168) (Swofford 2003) using the default 100 bootstrap
183 replicates to estimate a majority consensus tree, and maximum likelihood inference in IQ-TREE.

184 A schematic of all bioinformatic analyses used to extract homologous markers is found in Figure
185 1.
186



187

188 **Figure 1. Bioinformatics pipelines.** Schematic representing the workflows used in this study to extract homologous
189 markers from genome assemblies and whole genome shotgun sequences for phylogenomic inference of turtles.
190 BUSCO = Benchmarking Single Copy Orthologs; ML=maximum likelihood; RF = Robinson-Fould's.
191

192 *Heterogeneity analyses*

193 To measure similarity across gene genealogies, we calculated Robinson-Foulds (RF) distances
194 (Robinson and Foulds 1981) with PhyKIT (Steenwyk et al. 2021). First, we measured the RF
195 distance between each inferred single-copy ortholog gene tree and the inferred species tree. In
196 order to take into account missing species across some of the gene trees, we normalized the RF
197 distances by taking the raw RF distance and dividing it by 2(n-3) where n is the number of tips in
198 the phylogeny. We also calculated multiple pairwise RF distances between the locus type

199 phylogenies inferred from data mapped to each genome annotation feature with SISRS (CDS,
200 introns, intergenic, etc.). Finally, we calculated the RF distances of each locus type phylogeny
201 from the topology inferred with all mapped sites across the turtle genome.

202 We computed gene and site concordance factors for each branch in the turtle phylogeny
203 using the trimmed and filtered single-copy ortholog alignments, their gene trees, and the
204 coalescent-based species tree in IQ-TREE (Minh et al. 2020a). A branch's gene concordance
205 factor represents the percentage of decisive gene trees that also contain that branch, and a
206 branch's site concordance factor is the percentage of decisive alignment sites supporting that
207 branch. We also computed gene concordance factor and site concordance factor for each branch
208 in the concatenated single-copy ortholog tree of alignments >1,500 bp by computing individual
209 locus trees based on the partitions in IQ-TREE. Locus trees were inferred with default settings,
210 and the species tree was inferred using an edge-linked proportional partition model with an SH-
211 like approximate likelihood ratio test and ultrafast bootstrap analysis of 10,000 replicates each (-
212 alrt 10000 -B 10000). The greedy algorithm of PartitionFinder (Lanfear et al. 2012) was applied
213 to find the best-fit partitioning scheme which was then used in the subsequent tree reconstruction
214 step (-m TESTMERGE). We also calculated site concordance factor for every branch in the
215 maximum likelihood trees inferred from biallelic parsimony informative sites within locus types.

216 Concordance factors for genes and sites allow for complementary approaches to measure
217 the consistency of loci that agree with particular branches in a phylogeny, as well as
218 opportunities to test for incomplete lineage sorting, which we applied to each branch in the
219 single-copy ortholog species tree. One expectation of incomplete lineage sorting is a roughly
220 equal proportion of topologies supporting alternative relationships (Green et al. 2010). We used a
221 chi-square test to determine if similar proportions of gene trees supported conflicting

222 relationships at each branch, where a failure to reject the null hypothesis supported incomplete
223 lineage sorting at a given branch.
224
225 *Divergence Time Estimation*
226 We estimated divergence times with single-copy ortholog alignments using Bayesian modeling
227 implemented in BEAST v.2.6.3 (Bouckaert et al. 2019). We subsampled 685 alignments with
228 complete taxon sampling, from which we further randomly sampled three replicate sets of 10
229 genes. Based on model testing with ModelFinder, we applied the HKY+Gamma+Invariant site
230 model to all partitions, and utilized a relaxed lognormal molecular clock and Calibrated Yule tree
231 prior. To calibrate the time trees, we used 11 calibration priors from the literature (Joyce et al.
232 2013) which overlapped with our taxon sampling (Supplementary Data). To account for
233 uncertainty in fossil dating, we used hard minimum constraints and set soft maximum constraints
234 by placing the maximum ages in the 97.5% quantile of a lognormal prior distribution. Markov
235 chain Monte Carlo (MCMC) analyses were sampled every 50,000 generations. For each
236 replicate, we assessed convergence of parameter estimates across the MCMC by monitoring
237 effective sample sizes with Tracer v.1.7.2 (Rambaut et al. 2018), and we estimated a maximum
238 clade credibility tree with TreeAnnotator v2.6.3.
239 We also estimated divergence times in turtle evolution using penalized likelihood with
240 r8s v1.8.1 (Sanderson 1997, 2002, 2003), using the same fossil calibrations as above for
241 minimum and maximum node ages; the main difference in this analysis was we fixed the time to
242 most recent common ancestor (TMRCA) for amniotes at 312 million years ago (Ma) (Donoghue
243 and Benton 2007) and the TMRCA for Testudines at 220 Ma (Thomson et al. 2021). First, we
244 concatenated the 685 complete partitions and inferred a maximum likelihood tree with model

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

249 We used linear regression to determine the relationship between the estimated divergence
250 time for each node and the gene and site concordance factors in the inferred trees from each
251 locus type in the SISRS analysis, the gene trees and the species tree based on single-copy
252 orthologs, and the locus trees and partitioned maximum likelihood tree based on single-copy
253 orthologs. We also tested for a correlation between the concordance factors and the estimated
254 rates of molecular evolution at each branch. We repeated these regressions using estimated
255 divergence times from Thomson et al. (2020).

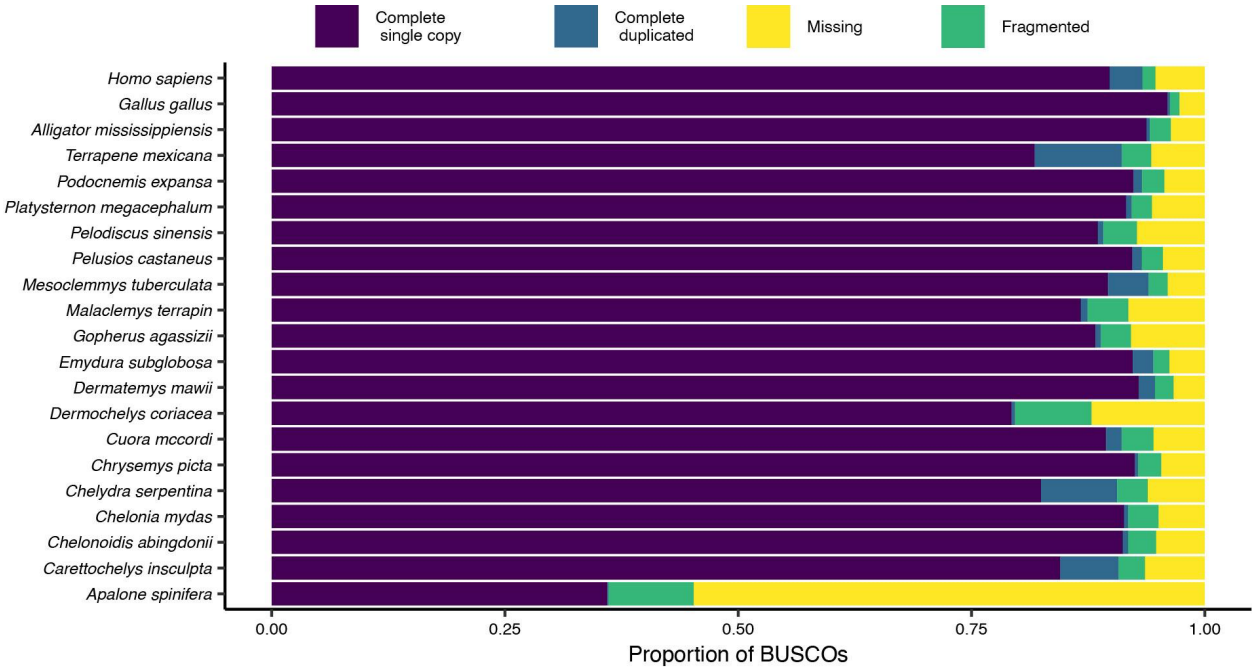
256

257 **Results**

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

259 We analyzed publicly available genome assemblies to extract single copy orthologs from 19
260 turtles representing 13 out of 14 extant families and 7 of 8 superfamilies, plus 3 outgroup taxa
261 (Table 1). On average, 4,609 (83%, range 1,911-5,099) conserved tetrapod orthologs were
262 complete in the genome assemblies, 112 (2.1%, range 7-496) were duplicated, 170 (3.2%, range
263 55-485) were fragmented, and 419 (7.9%, range 143-2,907) were missing. Fifty-five percent of
264 conserved tetrapod orthologs were missing from the genome assembly of the trionychid *Apalone*
265 *spinifera*; therefore, we omitted this species from downstream analyses of single-copy orthologs
266 due to concerns about assembly quality, of which BUSCO is known to be a reliable indicator
267 (Waterhouse et al. 2018). Each single-copy ortholog was present in an average of 19 out of 21

268 species, averaged 1,902 bp in length, and contained an average of 53% variable sites and 26%
269 parsimony-informative sites. Alignment length was correlated with the number of parsimony
270 informative sites (R^2=0.69, P=2.2e-16, Fig. 3a). GC content was consistent across the single-
271 copy orthologs and averaged 48% (standard deviation 6.2%, Fig. 3b), as was average pairwise
272 Kimura 2-parameter distance (mean 0.154, median 01.39, standard deviation 0.072, Fig. 3c). The
273 subset of 2,513 alignments ≥1,500 bp in length had very similar characteristics to the complete
274 set of loci; the average large locus length was 1,903 bp and the relationship between alignment
275 length and the number of parsimony-informative sites held (R^2=0.73, P=2.2e-16).
276



277

278 **Figure 2. Summary of single-copy orthologs.** Percent of tetrapod orthologs from orthoDBv9 present in each
279 genome assembly as complete single copy, complete duplicated, fragmented, or missing.
280

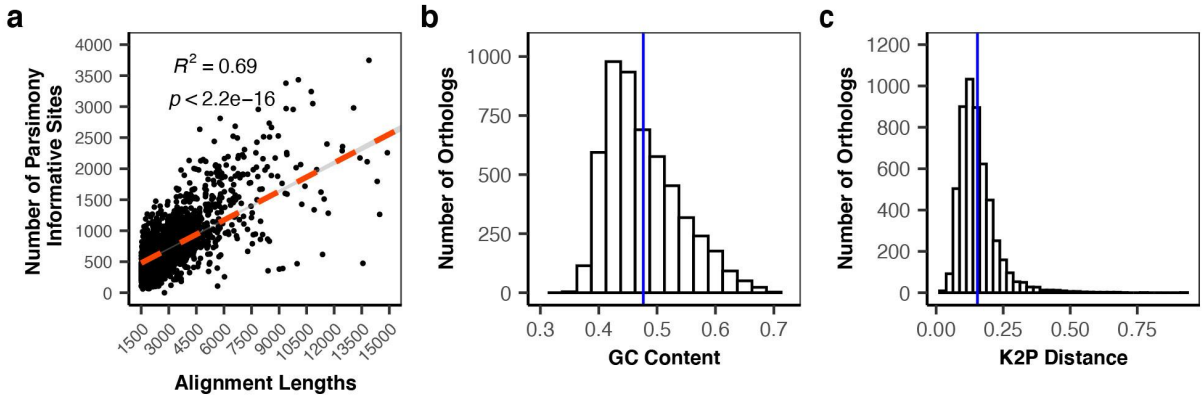


Figure 3. Description of single copy ortholog alignments. (a) Scatterplot showing the relationship between alignment length and the number of parsimony informative sites for trimmed alignments ≥1,500 bp. (b) Histogram showing the distribution of GC content in all single copy orthologs after trimming. Solid vertical line represents the mean value (0.476). (c) Histogram showing the distribution of average pairwise Kimura 2-parameter (K2P) distances for single copy orthologs after trimming. Solid vertical line represents the mean value (0.154).

We downloaded 38 SRA datasets from 26 turtle species representing 7 of 8 superfamilies and 14 families (Supplementary Data). Based on FastQC results, we selected 32 of the datasets for further analysis ranging from ~1–70X post-trimming coverage. The composite genome consisted of 5,555,666 contigs with an N50 of 2.4 Mb. We successfully mapped 302,355,876 bases to 2,110,354 contigs in the composite genome which covered 80% of the painted turtle genome (Table 2), and called 1,655,675 parsimony-informative biallelic SNPs when allowing missing data for up to two species at a given site. The number of parsimony-informative biallelic SNPs was much larger in mapped datasets containing only cryptodires (2,787,072 bases) and testudinoideans (4,696,637 bases) (Supplementary Data), as expected based on Literman et al. (2021). We collected the highest number of biallelic SNPs from intergenic regions (746,724 bases), followed by introns (533,336 bases) and CDS (136,823 bases). The relatively low

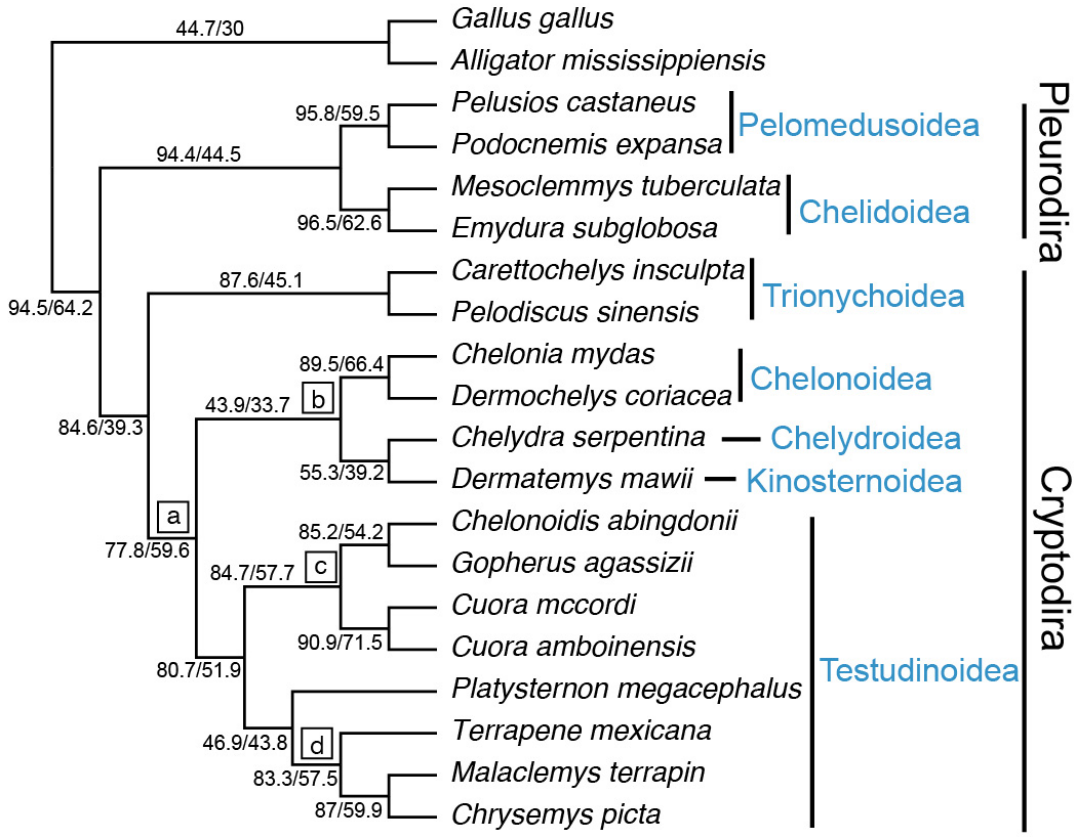
299 number of parsimony-informative biallelic SNPs in pseudogenic (427) and smRNA regions
300 (137) is likely due to a small number of annotations for these features in the western painted
301 turtle genome.
302
303 **Table 2. Summary of data mapped to the western painted turtle (*Chrysemys picta*) genome annotations.**

Locus Type	Annotated Bases in the Genome	Based Mapped by SISRS	Percent of Annotated Bases Mapped
All Mapped	2,365,766,571	302,355,876	12.78%
Intergenic	1,382,145,112	161,445,195	11.68%
Intronic	812,010,672	115,297,781	14.20%
lncRNA	64,450,450	9,451,244	14.66%
Genic Other	65,764,053	9,052,113	13.76%
CDS	33,486,790	5,718,074	17.08%
3'-UTR	8,034,514	1,510,783	18.80%
5'-UTR	5,096,255	708,214	13.90%
Pseudogenic	670,023	59,058	8.81%
smRNA	133,894	20,076	14.99%

304
305 *Bioinformatic analysis of turtle genomes yielded well-resolved species relationships*
306 The inferred relationships among turtle lineages based on single-copy orthologs were consistent
307 with previous studies using phylogenomic markers (Crawford et al. 2015; Shaffer et al. 2017),
308 with full statistical support for all resulting branches based on posterior probabilities in the
309 multilocus coalescent tree, and bootstrap replicates and SH-like approximate likelihood in the

supermatrix tree (Fig. 4, Supplementary Fig. 1, Supplementary Fig. 2). Hereafter we refer to these results collectively as the turtle species tree. The turtle species tree supports the reciprocal monophyly of Pleurodira and Cryptodira. Within cryptodires, there is a split between the Trionychoidea (soft-shelled turtles) and the Durocryptodira (Crawford et al. 2015), and a further subdivision of Durocryptodira into the Testudinoidea (pond turtles, big-headed turtles, geoemyds, and tortoises) and Americhelydia (sea turtles, snapping turtles, mud turtles; Joyce et al. 2013). Within the Testudinoidea, the turtle species tree supports Testuguria (geoemyds and tortoises; Joyce et al. 2004), and a *Platysternon* + Emydidae clade ("Platyemidae"; Crawford et al. 2015).

320



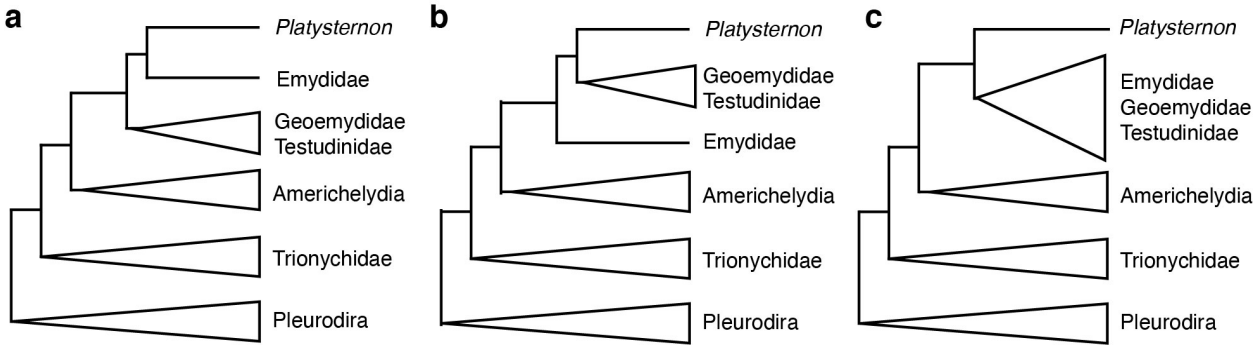
321

322 **Figure 4. Multilocus coalescent species tree of turtles based on high-confidence single-copy orthologs.**
323 Topology of the species tree inferred from 2,513 single-copy orthologs ≥1,500 bp in length. All branches received
324 100% local posterior support. Numbers on branches indicate gene and site concordance factors before and after the
325 slash, respectively. Superfamilies are indicated by vertical text. Key clades are labelled with letters in boxes. (a)
326 Durocryptodira, (b) Americhelydia, (c) Testuguria, (d) Emydidae. Additional outgroup *Homo sapiens* not shown.

327

328 Maximum likelihood and quartets trees from the mapped data, particularly for all-
329 mapped, CDS, introns, and intergenic regions, were consistent with the species tree based on
330 single-copy orthologs (Supplementary Fig. 3-12) although this was highly dependent on the
331 amount of allowed missing data and gapless sites, and was variable between the taxonomic
332 specificity of the data used to map sites (Supplementary Data). When there was disagreement

333 between trees from mapped regions and the species tree, the difference was often in the
334 placement of *Platysternon* (Fig. 5). For instance, when allowing zero missing gapless parsimony-
335 informative sites from the all-turtle data, 5' UTR placed *Platysternon* as the outgroup of
336 Geoemydidae + Testudinidae. When allowing up to two missing gapless parsimony-informative
337 sites from the all-turtle data, pseudogenic regions (426 sites) also placed *Platysternon* as the
338 outgroup of Geoemydidae + Testudinidae, while smRNA (137 sites) agreed with the species tree
339 and all-mapped topologies. Meanwhile, at multiple levels of allowed missing data the
340 Testudinoidea-only mapped sites weakly supported *Platysternon* as the sister taxon to
341 Geoemydidae + Testudinidae + Emydidae (Supplementary Data).
342



343

344 **Figure 5. Alternative topologies in turtle phylogenetics based on markers from different regions of the**
345 **genome.** (a) *Platysternon* as the sister taxon to Emydidae to the exclusion of Testuguria (Geoemydidae +
346 Testudinidae), consistent with the turtle species tree (current study), Crawford et al. (2015), and Shaffer et al.
347 (2017). (b) *Platysternon* as the sister taxon to Testuguria with the exclusion of Emydidae, consistent with 17% of the
348 single-copy ortholog genes trees, mapped 5'-UTR parsimony-informative sites data with zero missing data, mapped
349 pseudogene parsimony-informative sites when allowing 2 missing taxa per site, and Krenz et al. (2005). (c)
350 *Platysternon* as the sister taxon to Geoemydidae + Testudinidae + Emydidae, consistent with 20% of single-copy
351 ortholog gene trees and Shaffer et al. (1997).

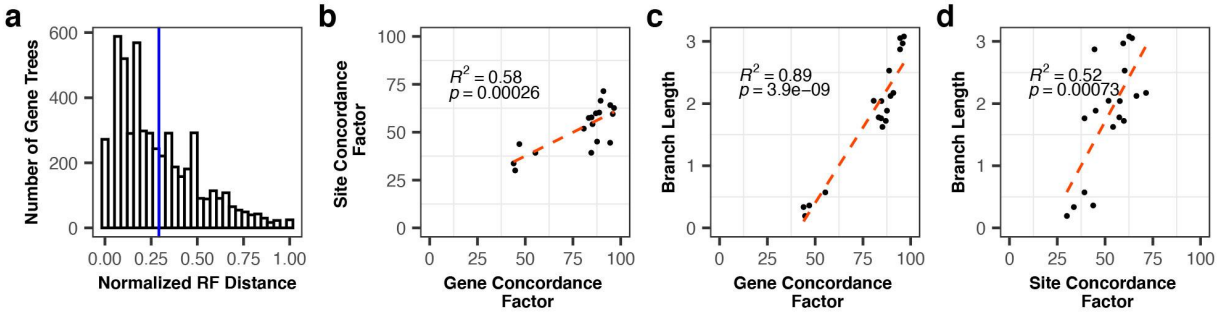
352

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

354 While results from our multilocus and biallelic sites analyses of turtles are consistent and largely
355 in agreement with other studies based on 100% branch support, some genomic regions, either
356 from single-copy orthologs or from genome feature annotations, yielded genealogies conflicting
357 with the turtle species tree. While most gene trees from single-copy orthologs matched the
358 species tree based on normalized RF distances, there was a long tail to this distribution indicating
359 gene trees with greater dissimilarity (Fig. 6a). Quartet trees using mapped data from CDS,
360 introns, and intergenic regions were consistent with both the single-copy ortholog species tree
361 and a quartet tree derived from the all-mapped data, except for pseudogenic DNA, 5' UTR, and
362 smRNA which deviated from the all-mapped data in terms of RF distances (Supplementary Fig.
363 13).

364 We also found a high degree of heterogeneity based on gene and site concordance factors
365 throughout the turtle phylogeny (Fig. 1, Supplementary Fig. 1, Supplementary Fig. 2). For
366 instance, while local posterior probabilities were uniformly 100% across branches of the
367 multilocus coalescent species tree, gene concordance factors ranged from 44 to 97 and site
368 concordance factors ranged from 30 to 71 (Fig. 6b-d). Gene and site concordance factors based
369 on the multilocus species tree were correlated with each other (Figure 6b; R^2=0.55, P=0.0002)
370 and with branch lengths (Fig. 6c-d; R^2=0.88, 3.88e-09 and R^2=0.49, P=0.0007, respectively). We
371 also estimated gene and site concordance factors based on the locus trees with respect to the
372 maximum likelihood tree inferred from the concatenated supermatrix, and found a similar range
373 of gene concordance factors (34 to 92) and site concordance factors (30 to 71) despite 100%
374 bootstrap support, as well as a strong correlation between the two concordance measures

375 (R^2=0.60, P=0.0001). Gene and site concordance factors based on the concatenated partitioned
376 maximum likelihood tree were also correlated with branch lengths (R^2=0.90, 1.02e-09 and
377 R^2=0.58, P=0.0001, respectively).



378

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

385 The root of the turtle species tree was assigned a relatively large gene concordance factor
386 (94.5) and site concordance factor (64.2), suggesting strong support for reciprocally
387 monophyletic Pleurodira and Cryptodira across genes and sites. We rejected incomplete lineage
388 sorting at the root of the turtle phylogeny and the roots of Cryptodira (P=8.05E-04) and
389 Pleurodira (P=2.32E-04). The placement of Trionychoidea as sister taxon to all other cryptodires
390 was assigned relatively large gene concordance factor (84.56) but small site concordance factor
391 (39.26), suggesting that there are a large number of sites supporting alternative bifurcations. One
392 area with particularly high discordance (i.e. relatively small gene concordance factor=46.9 and
393 site concordance factor=43.8) was the placement of the big-headed turtle (*Platysternon*) with

394 respect to Emydidae, for which we could not reject incomplete lineage sorting (see Table 3, Fig.
395 5 and Discussion).
396
397 **Table 3. Gene concordance factors for nodes in the species tree phylogeny where incomplete lineage sorting**
398 **was not rejected.**

Clade description	Node	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	28	2387	88.73	0.54	0.21
Chelonoidea	30	1755	89.46	1.48	0.8
Testudinoidea	32	2435	80.66	0.57	0.7
Testuguria	33	2255	84.7	0.62	0.8
Testudinidae	34	1989	85.22	4.17	5.38
Cuora	35	1932	90.94	2.69	2.07
Emydidae + Platysternidae	36	2270	46.92	17.05	19.52
Malachlemys + *Chrysemys*	38	1730	86.99	5.03	4.39
Pelomedusoidea	40	2195	95.81	1.32	1.14
Chelidoidea	41	2138	96.49	0.8	1.08

399
400 *Heterogeneity at different timescales in turtle evolution*
401 Results from the replicate BEAST analyses consistently underestimated turtle divergence times
402 with respect to results from the penalized likelihood analysis as well as those of Thomson et al.
403 (2020), even though the TMRCA for all taxa (i.e. the time of the mammal-reptile split) was
404 consistent (Supplementary Data). Across locus types from the SISRS analysis, we found that site
405 concordance factors were not strongly correlated with estimated node ages, regardless of whether
406 or not we used our estimated divergence times or those from Thomson et al. (2021) (Fig. 7,
407 Supplementary Fig. 14). For most locus types the relationship between site concordance factor
408 and divergence time at a given node trended slightly negatively, suggesting a pattern of higher
409 concordance at younger nodes, although these comparisons were non-significant (Fig. 7b,

410 P>0.1). The one exception to this trend was for pseudogenes, for which the relationship between
411 site concordance factors and estimated divergence times trended positive albeit without
412 significance or with only moderate significance (P=0.072 using Thomson et al. 2020's
413 divergence time estimates). The relationship between site concordance factors and estimated
414 DNA substitution rates trended slightly positive for all locus types from the SISRS analysis, but
415 was non-significant in all comparisons (Supplementary Fig. 15).

416 We also found a negative trend in the relationship between site concordance factor and
417 divergence time at a given node based on both the multilocus species tree and the supermatrix
418 tree based on single-copy orthologs, similar to the patterns observed using the locus types from
419 SISRS (Fig. 8a-c; Supplementary Fig. 16). However, we found a positive (albeit non-significant)
420 trend in the relationship between the estimated divergence time at a given node and its gene
421 concordance factor when using both trees derived from single-copy orthologs (Fig. 8e-g;
422 Supplementary Fig. 16). There was no clear relationship between the site concordance factor of a
423 given node and the estimated substitution rate at the preceding branch (Fig. 8d), and a
424 moderately significant positive trend (P≤0.05) for the relationship between the gene concordance
425 factor of a given node and the estimated substitution rate at the preceding branch (Fig. 8h;
426 Supplementary Fig. 16).

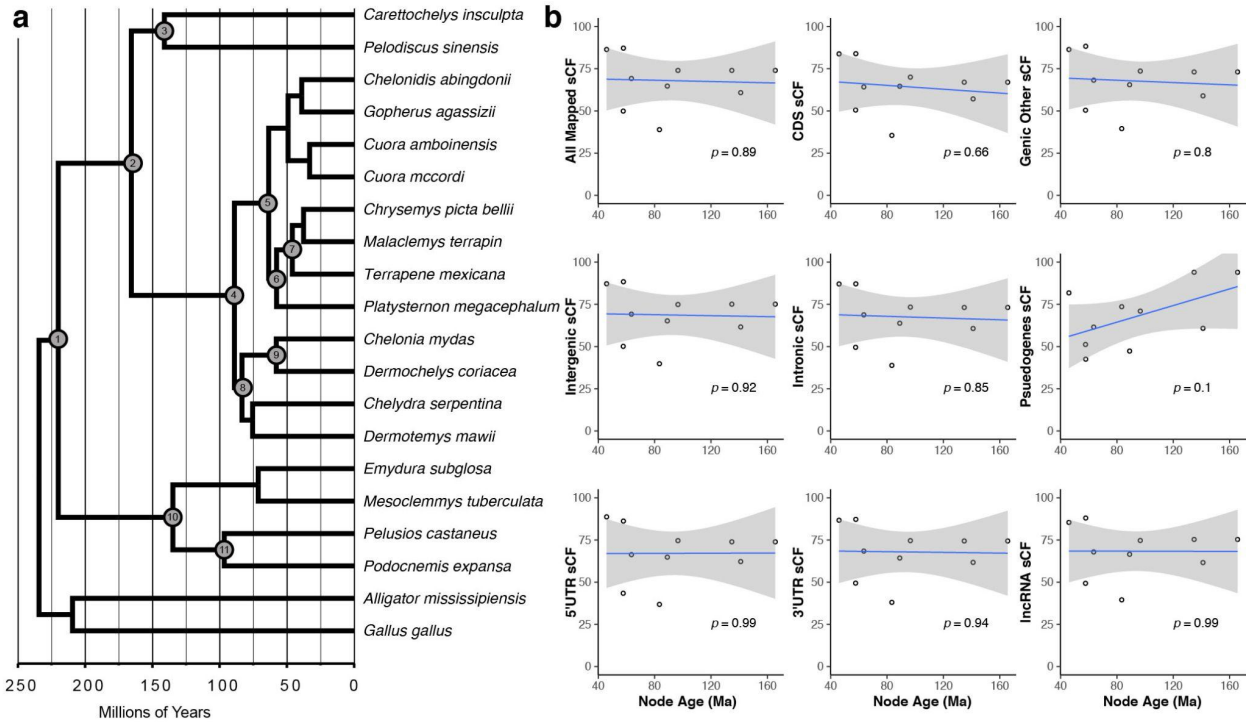


Figure 7. Site concordance and estimated node ages in turtle evolution. (a) Chronogram showing estimated divergence times in millions of years based on penalized likelihood and fossil calibrations of labelled nodes. 1. Testudines, 2. Cryptodira, 3. Trionychoidea, 4. Durocryptodira, 5. Geoemydidae, 6. 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.

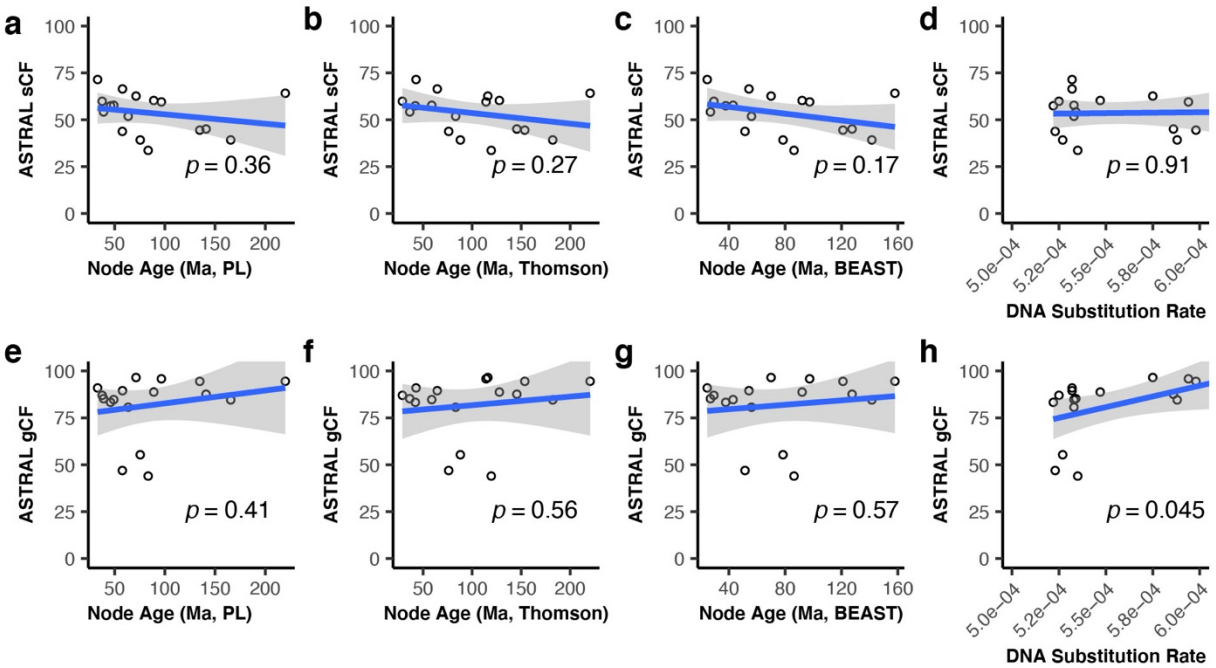


Figure 8. Concordance at genes and sites, estimated node ages, and DNA substitution rates in turtle evolution. Results shown are from comparisons of single-copy ortholog alignments and their gene trees to the multilocus coalescent species tree inferred using ASTRAL. There was a negative trend between 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 single-copy ortholog alignments, locus trees, and the maximum likelihood tree inferred from the concatenated supermatrix are shown in Supplementary Materials.

Discussion

While 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.

451 2012; Minh et al. 2020a), to our knowledge there has not been an assessment of phylogenomic
452 heterogeneity in turtles until now. Here, we have constructed a robust dataset based on whole
453 genomes with the goal of disentangling conflicting hypotheses about turtle relationships. Single-
454 copy orthologs yielded more than double the variation and triple the number of parsimony
455 informative sites for turtles compared to other high-throughput methods (Crawford et al. 2015;
456 Shaffer et al. 2017). In addition, the mapped data included millions of parsimony informative
457 sites that were further parsed by locus type. We identified splits in the turtle phylogeny and
458 regions of the turtle genome associated with gene tree and site heterogeneity, and investigated
459 potential sources of this discordance. We found that heterogeneity in turtle phylogenomics can
460 be explained by a combination of factors, such as incomplete lineage sorting and lack of gene
461 informativeness, and has contributed to conflicting results in phylogenetic studies of turtles. The
462 results of our study demonstrate that heterogeneity can be rampant, even in phylogenomic
463 datasets which have yielded fully resolved species trees.

464 Gene tree-species tree discordance in turtles may be partly due to an overall slow
465 substitution rate compared to other amniotes (Tollis et al. 2018; Gemmell et al. 2020), resulting
466 in a lack of gene informativeness at more recent divergences. We found no significant
467 relationship between site concordance and node age across most locus types, a uniform
468 distribution of site concordances across the range of estimated DNA substitution rates across
469 most annotated genomic regions, and that the average pairwise Kimura 2-parameter distances
470 across single-copy orthologs were also fairly uniform. This suggests relatively consistent
471 substitution rates across the turtle genome, as well as across turtle evolution. Where we did
472 estimate elevated DNA substitution rates along branches in the turtle phylogeny, we found that
473 gene concordance factors were significantly larger. This suggests that shifts in the rate of

474 evolution may lead to a comparatively rapid accumulation of mutations during certain time
475 periods, resulting in more gene trees in agreement with the species phylogeny. Meanwhile, the
476 gene concordances at older nodes were larger (albeit non-significantly), suggesting that
477 heterogeneity at more recent nodes may be more driven by stochastic error due to limited
478 information at slowly evolving loci. Our results are consistent with the longstanding prediction
479 that slowly evolving loci, which includes much of the turtle genome, provide disproportionally
480 greater phylogenetic information for older splits (Townsend et al. 2008).

481 Interestingly, site concordance in pseudogenes was positively associated (albeit non-
482 significantly) with node age, suggesting that even slightly elevated substitution rates in some
483 noncoding regions of the turtle genome can lead to more informative sites supporting the species
484 tree topology at older nodes. This is somewhat different from what was reported in Literman et
485 al. (2021), who showed that noncoding regions such as introns and intergenic regions supported
486 more recent splits in the mammalian phylogeny, and may be again related to the fact that
487 estimated substitution rates in mammals are consistently higher than for other amniotes such as
488 turtle and crocodilians. However, our inferences from pseudogenes are based on a relatively
489 small number of parsimony informative sites (427). More comprehensive annotations for
490 noncoding regions of turtle genomes are needed to investigate this further.

491 Our tests for incomplete lineage sorting reveal that population-level processes likely also
492 have played a key role at multiple stages of turtle diversification. In particular, the phylogenetic
493 position of *Platysternon* has been highly contentious (Joyce et al. 2004, 2021), with early
494 molecular studies in disagreement (Shaffer et al. 1997; Krenz et al. 2005; Parham et al. 2006;
495 Barley et al. 2010). Our turtle species tree results mirror Crawford et al. (2015) and Shaffer et al.
496 (2017), fully supporting *Platysternon* as the sister taxon to Emydidae. However, only 47% of

single-copy ortholog gene trees are in agreement with a *Platysternon* + Emydidae clade, and 37% of 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-square test (Table 3), and supports incomplete lineage sorting rather than introgression or statistical noise (Green et al. 2010). As these loci support published alternative hypotheses concerning the relationship of *Platysternon* to other Testudinoidea, we suggest that incomplete lineage sorting is a previously unrecognized driver of conflicting phylogenetic results in the higher classification of turtles.

Acknowledgements

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

Data Availability

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

520

521 **Literature Cited**

522 Barley A.J., Spinks P.Q., Thomson R.C., Shaffer H.B. 2010. Fourteen nuclear genes provide
523 phylogenetic resolution for difficult nodes in the turtle tree of life. Mol. Phylogenet. Evol.
524 55:1189–1194.

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

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

529 Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A.,
530 Heled J., Jones G., Kühnert D., Maio N.D., Matschiner M., Mendes F.K., Müller N.F.,
531 Ogilvie H.A., Plessis L. du, Popinga A., Rambaut A., Rasmussen D., Siveroni I., Suchard
532 M.A., Wu C.-H., Xie D., Zhang C., Stadler T., Drummond A.J. 2019. BEAST 2.5: An
533 advanced software platform for Bayesian evolutionary analysis. PLOS Comput. Biol.
534 15:e1006650.

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

540 Chifman J., Kubatko L. 2014. Quartet Inference from SNP Data Under the Coalescent Model.
541 Bioinformatics. 30:3317–3324.

542 Cooper E.D. 2014. Overly simplistic substitution models obscure green plant phylogeny. Trends

543 Plant Sci. 19:576–582.

544 Cox C.J., Li B., Foster P.G., Embley T.M., Civáň P. 2014. Conflicting Phylogenies for Early
545 Land Plants are Caused by Composition Biases among Synonymous Substitutions. Syst.
546 Biol. 63:272–279.

547 Crawford N.G., Faircloth B.C., McCormack J.E., Brumfield R.T., Winker K., Glenn T.C. 2012.
548 More than 1000 ultraconserved elements provide evidence that turtles are the sister group
549 of archosaurs. Biol. Lett. 8:783–786.

550 Crawford N.G., Parham J.F., Sellas A.B., Faircloth B.C., Glenn T.C., Papenfuss T.J., Henderson
551 J.B., Hansen M.H., Simison W.B. 2015. A phylogenomic analysis of turtles. Mol.
552 Phylogenet. Evol. 83:250–257.

553 Donoghue P.C.J., Benton M.J. 2007. Rocks and clocks: calibrating the Tree of Life using fossils
554 and molecules. Trends Ecol. Evol. 22:424–431.

555 Dudchenko O., Batra S.S., Omer A.D., Nyquist S.K., Hoeger M., Durand N.C., Shamim M.S.,
556 Machol I., Lander E.S., Aiden A.P., Aiden E.L. 2017. De novo assembly of the *Aedes*
557 *aegypti* genome using Hi-C yields chromosome-length scaffolds. Science. 356:92–95.

558 Edwards S.V., Xi Z., Janke A., Faircloth B.C., McCormack J.E., Glenn T.C., Zhong B., Wu S.,
559 Lemmon E.M., Lemmon A.R., Leaché A.D., Liu L., Davis C.C. 2016. Implementing and
560 testing the multispecies coalescent model: A valuable paradigm for phylogenomics. Mol.
561 Phylogenet. Evol. 94:447–462.

562 Faircloth B.C., Sorenson L., Santini F., Alfaro M.E. 2013. A Phylogenomic Perspective on the
563 Radiation of Ray-Finned Fishes Based upon Targeted Sequencing of Ultraconserved
564 Elements (UCEs). PLOS ONE. 8:e65923.

565 Foster P.G. 2004. Modeling Compositional Heterogeneity. Syst. Biol. 53:485–495.

566 Galtier N., Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philos. Trans.
567 R. Soc. B Biol. Sci. 363:4023–4029.
568 Gemmell N.J., Rutherford K., Prost S., Tollis M., Winter D., Macey J.R., Adelson D.L., Suh A.,
569 Bertozzi T., Grau J.H., Organ C., Gardner P.P., Muffato M., Patricio M., Billis K., Martin
570 F.J., Flicek P., Petersen B., Kang L., Michalak P., Buckley T.R., Wilson M., Cheng Y.,
571 Miller H., Schott R.K., Jordan M.D., Newcomb R.D., Arroyo J.I., Valenzuela N., Hore
572 T.A., Renart J., Peona V., Peart C.R., Warmuth V.M., Zeng L., Kortschak R.D., Raison
573 J.M., Zapata V.V., Wu Z., Santesmasses D., Mariotti M., Guigó R., Rupp S.M., Twort
574 V.G., Dussex N., Taylor H., Abe H., Bond D.M., Paterson J.M., Mulcahy D.G., Gonzalez
575 V.L., Barbieri C.G., DeMeo D.P., Pabinger S., Van Stijn T., Clarke S., Ryder O.,
576 Edwards S.V., Salzberg S.L., Anderson L., Nelson N., Stone C., Ngatiwai Trust Board.
577 2020. The tuatara genome reveals ancient features of amniote evolution. Nature.
578 584:403–409.
579 Green R.E., Krause J., Briggs A.W., Maricic T., Stenzel U., Kircher M., Patterson N., Li H., Zhai
580 W., Fritz M.H.-Y., Hansen N.F., Durand E.Y., Malaspinas A.-S., Jensen J.D., Marques-
581 Bonet T., Alkan C., Prüfer K., Meyer M., Burbano H.A., Good J.M., Schultz R., Aximu-
582 Petri A., Butthof A., Höber B., Höffner B., Siegemund M., Weihmann A., Nusbaum C.,
583 Lander E.S., Russ C., Novod N., Affourtit J., Egholm M., Verna C., Rudan P., Brajkovic
584 D., Kucan Ž., Gušic I., Doronichev V.B., Golovanova L.V., Lalueza-Fox C., Rasilla M.
585 de la, Fortea J., Rosas A., Schmitz R.W., Johnson P.L.F., Eichler E.E., Falush D., Birney
586 E., Mullikin J.C., Slatkin M., Nielsen R., Kelso J., Lachmann M., Reich D., Pääbo S.
587 2010. A Draft Sequence of the Neandertal Genome. Science. 328:710–722.
588 Hime P.M., Lemmon A.R., Lemmon E.C.M., Prendini E., Brown J.M., Thomson R.C., Kratovil

589 J.D., Noonan B.P., Pyron R.A., Peloso P.L.V., Kortyna M.L., Keogh J.S., Donnellan
590 S.C., Mueller R.L., Raxworthy C.J., Kunte K., Ron S.R., Das S., Gaitonde N., Green
591 D.M., Labisko J., Che J., Weisrock D.W. 2021. Phylogenomics Reveals Ancient Gene
592 Tree Discordance in the Amphibian Tree of Life. Syst. Biol. 70:48-66.
593 Irisarri I., Baurain D., Brinkmann H., Delsuc F., Sire J.-Y., Kupfer A., Petersen J., Jarek M.,
594 Meyer A., Vences M., Philippe H. 2017. Phylotranscriptomic consolidation of the jawed
595 vertebrate timetree. Nat. Ecol. Evol. 1:1370–1378.
596 Jarvis E.D., Mirarab S., Aberer A.J., Li B., Houde P., Li C., Ho S.Y.W., Faircloth B.C., Nabholz
597 B., Howard J.T., Suh A., Weber C.C., da Fonseca R.R., Li J., Zhang F., Li H., Zhou L.,
598 Narula N., Liu L., Ganapathy G., Boussau B., Bayzid M.S., Zavidovych V., Subramanian
599 S., Gabaldón T., Capella-Gutiérrez S., Huerta-Cepas J., Rekepalli B., Munch K.,
600 Schierup M., Lindow B., Warren W.C., Ray D., Green R.E., Bruford M.W., Zhan X.,
601 Dixon A., Li S., Li N., Huang Y., Derryberry E.P., Bertelsen M.F., Sheldon F.H.,
602 Brumfield R.T., Mello C.V., Lovell P.V., Wirthlin M., Schneider M.P.C., Prosdocimi F.,
603 Samaniego J.A., Velazquez A.M.V., Alfaro-Núñez A., Campos P.F., Petersen B.,
604 Sicheritz-Ponten T., Pas A., Bailey T., Scofield P., Bunce M., Lambert D.M., Zhou Q.,
605 Perelman P., Driskell A.C., Shapiro B., Xiong Z., Zeng Y., Liu S., Li Z., Liu B., Wu K.,
606 Xiao J., Yinqi X., Zheng Q., Zhang Y., Yang H., Wang J., Smeds L., Rheindt F.E., Braun
607 M., Fjeldså J., Orlando L., Barker F.K., Jønsson K.A., Johnson W., Koepfli K.-P.,
608 O'Brien S., Haussler D., Ryder O.A., Rahbek C., Willerslev E., Graves G.R., Glenn T.C.,
609 McCormack J., Burt D., Ellegren H., Alström P., Edwards S.V., Stamatakis A., Mindell
610 D.P., Cracraft J., Braun E.L., Warnow T., Jun W., Gilbert M.T.P., Zhang G. 2014.
611 Whole-genome analyses resolve early branches in the tree of life of modern birds.

Science. 346:1320–1331.

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.

Katoh K., Standley D.M. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol. Biol. Evol. 30:772–780.

Krenz J.G., Naylor G.J.P., Shaffer H.B., Janzen F.J. 2005. Molecular phylogenetics and evolution of turtles. Mol. Phylogenet. Evol. 37:178–191.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of Phylogenetic Estimates from Concatenated Data under Coalescence. Syst. Biol. 56:17–24.

Kumar S., Filipski A.J., Battistuzzi F.U., Kosakovsky Pond S.L., Tamura K. 2012. Statistics and

635 Truth in Phylogenomics. Mol. Biol. Evol. 29:457–472.

636 Kumar S., Stecher G., Li M., Knyaz C., Tamura K. 2018. MEGA X: Molecular Evolutionary
637 Genetics Analysis across Computing Platforms. Mol. Biol. Evol. 35:1547–1549.

638 Lanfear R., Calcott B., Ho S.Y.W., Guindon S. 2012. Partitionfinder: combined selection of
639 partitioning schemes and substitution models for phylogenetic analyses. Mol. Biol. Evol.
640 29:1695–1701.

641 Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods.
642 9:357–359.

643 Lemmon A.R., Emme S.A., Lemmon E.M. 2012. Anchored Hybrid Enrichment for Massively
644 High-Throughput Phylogenomics. Syst. Biol. 61:727–744.

645 Literman R., Schwartz R. 2021. Genome-Scale Profiling Reveals Noncoding Loci Carry Higher
646 Proportions of Concordant Data. Mol. Biol. Evol.

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

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

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

656 Minh B.Q., Schmidt H.A., Chernomor O., Schrempf D., Woodhams M.D., von Haeseler A.,
657 Lanfear R. 2020b. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic

658 Inference in the Genomic Era. Mol. Biol. Evol. 37:1530–1534.
659 Mölder F., Jablonski K.P., Letcher B., Hall M.B., Tomkins-Tinch C.H., Sochat V., Forster J.,
660 Lee S., Twardziok S.O., Kanitz A., Wilm A., Holtgrewe M., Rahmann S., Nahnsen S.,
661 Köster J. 2021. Sustainable data analysis with Snakemake.
662 https://f1000research.com/articles/10-33.
663 Morales-Briones D.F., Kadereit G., Tefarikis D.T., Moore M.J., Smith S.A., Brockington S.F.,
664 Timoneda A., Yim W.C., Cushman J.C., Yang Y. 2021. Disentangling Sources of Gene
665 Tree Discordance in Phylogenomic Data Sets: Testing Ancient Hybridizations in
666 Amaranthaceae s.l. Syst. Biol. 70:219-235.
667 Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol.
668 5:568–583.
669 Parham J.F., Feldman C.R., Boore J.L. 2006. The complete mitochondrial genome of the
670 enigmatic bigheaded turtle (Platysternon): description of unusual genomic features and
671 the reconciliation of phylogenetic hypotheses based on mitochondrial and nuclear DNA.
672 BMC Evol. Biol. 6:11.
673 Prum R.O., Berv J.S., Dornburg A., Field D.J., Townsend J.P., Lemmon E.M., Lemmon A.R.
674 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA
675 sequencing. Nature. 526:569–573.
676 Quesada V., Freitas-Rodríguez S., Miller J., Pérez-Silva J.G., Jiang Z.-F., Tapia W., Santiago-
677 Fernández O., Campos-Iglesias D., Kuderna L.F.K., Quinzin M., Álvarez M.G., Carrero
678 D., Beheregaray L.B., Gibbs J.P., Chiari Y., Glaberman S., Ciofi C., Araujo-Voces M.,
679 Mayoral P., Arango J.R., Tamargo-Gómez I., Roiz-Valle D., Pascual-Torner M., Evans
680 B.R., Edwards D.L., Garrick R.C., Russello M.A., Poulakakis N., Gaughran S.J., Rueda

681 D.O., Bretones G., Marquès-Bonet T., White K.P., Caccone A., López-Otín C. 2019.
682 Giant tortoise genomes provide insights into longevity and age-related disease. Nat. Ecol.
683 Evol. 3:87–95.
684 Quinlan A.R., Hall I.M. 2010. BEDTools: a flexible suite of utilities for comparing genomic
685 features. Bioinformatics. 26:841–842.
686 Rambaut A., Drummond A.J., Xie D., Baele G., Suchard M.A. 2018. Posterior summarization in
687 Bayesian phylogenetics using Tracer 1.7. Syst. Biol. 67:901–904.
688 Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
689 Rokas A., Williams B.L., King N., Carroll S.B. 2003. Genome-scale approaches to resolving
690 incongruence in molecular phylogenies. Nature. 425:798–804.
691 Sanderson M.J. 1997. A nonparametric approach to estimating divergence times in the absence
692 of rate constancy. Mol. Biol. Evol.
693 Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: a
694 penalized likelihood approach. Mol. Biol. Evol. 19:101–109.
695 Sanderson M.J. 2003. r8s: inferring absolute rates of molecular evolution and divergence times
696 in the absence of a molecular clock. Bioinformatics. 19:301–302.
697 Sayyari E., Mirarab S. 2016. Fast Coalescent-Based Computation of Local Branch Support from
698 Quartet Frequencies. Mol. Biol. Evol. 33:1654–1668.
699 Schwartz R.S., Harkins K.M., Stone A.C., Cartwright R.A. 2015. A composite genome approach
700 to identify phylogenetically informative data from next-generation sequencing. BMC
701 Bioinformatics. 16:193.
702 Shaffer H.B., McCartney-Melstad E., Near T.J., Mount G.G., Spinks P.Q. 2017. Phylogenomic
703 analyses of 539 highly informative loci dates a fully resolved time tree for the major

704 clades of living turtles (Testudines). Mol. Phylogenet. Evol. 115:7–15.
705 Shaffer H.B., Meylan P., McKnight M.L. 1997. Tests of turtle phylogeny: molecular,
706 morphological, and paleontological approaches. Syst. Biol. 46:235–268.
707 Shaffer H.B., Minx P., Warren D.E., Shedlock A.M., Thomson R.C., Valenzuela N., Abramyan
708 J., Amemiya C.T., Badenhorst D., Biggar K.K., Borchert G.M., Botka C.W., Bowden
709 R.M., Braun E.L., Bronikowski A.M., Bruneau B.G., Buck L.T., Capel B., Castoe T.A.,
710 Czerwinski M., Delehaunty K.D., Edwards S.V., Fronick C.C., Fujita M.K., Fulton L.,
711 Graves T.A., Green R.E., Haerty W., Hariharan R., Hernandez O., Hillier L.W.,
712 Holloway A.K., Janes D., Janzen F.J., Kandoth C., Kong L., de Koning A.P.J., Li Y.,
713 Literman R., McGaugh S.E., Mork L., O'Laughlin M., Paitz R.T., Pollock D.D., Ponting
714 C.P., Radhakrishnan S., Raney B.J., Richman J.M., St John J., Schwartz T., Sethuraman
715 A., Spinks P.Q., Storey K.B., Thane N., Vinař T., Zimmerman L.M., Warren W.C.,
716 Mardis E.R., Wilson R.K. 2013. The western painted turtle genome, a model for the
717 evolution of extreme physiological adaptations in a slowly evolving lineage. Genome
718 Biol. 14:R28.
719 Shimodaira H. 2002. An Approximately Unbiased Test of Phylogenetic Tree Selection. Syst.
720 Biol. 51:492–508.
721 Singhal S., Colston T.J., Grundler M.R., Smith S.A., Costa G.C., Colli G.R., Moritz C., Pyron
722 R.A., Rabosky D.L. 2021. Congruence and Conflict in the Higher-Level Phylogenetics of
723 Squamate Reptiles: An Expanded Phylogenomic Perspective. Syst. Biol. 70:542–557.
724 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
725 large phylogenies. Bioinformatics. 30:1312–1313.
726 Steenwyk J.L., Buida T.J. III, Labella A.L., Li Y., Shen X.-X., Rokas A. 2021. PhyKIT: a

727 broadly applicable UNIX shell toolkit for processing and analyzing phylogenomic data.
728 Bioinformatics. btab096.
729 Swofford D.L. 2003. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods).
730 Sinauer associates Sunderland, MA.
731 Thomson R.C., Spinks P.Q., Shaffer H.B. 2021. A global phylogeny of turtles reveals a burst of
732 climate-associated diversification on continental margins. Proc. Natl. Acad. Sci. 118.
733 Tollis M., Denardo D.F., Cornelius J.A., Dolby G.A., Edwards T., Henen B.T., Karl A.E.,
734 Murphy R.W., Kusumi K. 2017. The Agassiz's desert tortoise genome provides a
735 resource for the conservation of a threatened species. PloS One. 12:e0177708.
736 Tollis M., Hutchins E.D., Stapley J., Rupp S.M., Eckalbar W.L., Maayan I., Lasku E., Infante
737 C.R., Dennis S.R., Robertson J.A., May C.M., Crusoe M.R., Bermingham E., Denardo
738 D.F., Hsieh S.T., Kulathinal R.J., McMillan W.O., Menke D.B., Pratt S.C., Rawls J.A.,
739 Sanjur O., Wilson Rawls J., Wilson Sayres M.A., Fisher R.E., Kusumi K. 2018.
740 Comparative Genomics Reveals Accelerated Evolution in Conserved Pathways during
741 the Diversification of Anole Lizards. Genome Biol. Evol. 10:489–506.
742 Townsend J.P., López-Giráldez F, Friedman R. 2008. The phylogenetic informativeness of
743 nucleotide and amino acid sequences for reconstructing the vertebrate tree. J. Mol. Evol.
744 67:437-447.
745 Uetz P., Freed P., Aguilar R., Hošek J. 2021. The Reptile Database. (last accessed August 2021)
746 Wang Z., Pascual-Anaya J., Zadissa A., Li W., Niimura Y., Wang J., Huang Z., Li C., White S.,
747 Xiong Z., Fang D., Wang B., Ming Y., Chen Y., Zheng Y., Kuraku S., Pignatelli M.,
748 Herrero J., Beal K., Nozawa M., Li Q., Wang J., Zhang H., Yu L., Shigenobu S., Wang
749 J., Liu J., Flicek P., Searle S., Wang J., Kuratani S., Yin Y., Aken B., Zhang G., Irie N.

750 2013. The draft genomes of soft-shell turtle and green sea turtle yield insights into the
751 development and evolution of the turtle-specific body plan. Nat. Genet. 45:701–706.
752 Waterhouse R.M., Seppey M., Simão F.A., Manni M., Ioannidis P., Klioutchnikov G.,
753 Kriventseva E.V., Zdobnov E.M. 2018. BUSCO Applications from Quality Assessments
754 to Gene Prediction and Phylogenomics. Mol. Biol. Evol. 35:543–548.
755 Wolf Y.I., Rogozin I.B., Grishin N.V., Koonin E.V. 2002. Genome trees and the tree of life.
756 Trends Genet. TIG. 18:472–479.
757 Zdobnov E.M., Tegenfeldt F., Kuznetsov D., Waterhouse R.M., Simão F.A., Ioannidis P.,
758 Seppey M., Loetscher A., Kriventseva E.V. 2017. OrthoDB v9.1: cataloging evolutionary
759 and functional annotations for animal, fungal, plant, archaeal, bacterial and viral
760 orthologs. Nucleic Acids Res. 45:D744–D749.
761 Zhang C., Rabiee M., Sayyari E., Mirarab S. 2018. ASTRAL-III: polynomial time species tree
762 reconstruction from partially resolved gene trees. BMC Bioinformatics. 19:153.
763