Multiple Genome Alignments Facilitate Development of NPCL Markers: A Case Study of Tetrapod Phylogeny Focusing on the Position of Turtles

Xing-Xing Shen,[1] Dan Liang,[1] Jun-Zhi Wen,[1] and Peng Zhang*[,1]

[1]State Key Laboratory of Biocontrol, Key Laboratory of Gene Engineering of the Ministry of Education, School of Life Sciences, Sun Yat-Sen University, Guangzhou, People's Republic of China

***Corresponding author:** E-mail: alarzhang@gmail.com.

Associate editor: Barbara Holland

Abstract

In recent years, the increasing availability of genomic resources has provided an opportunity to develop phylogenetic markers for phylogenomics. Efficient methods to search for candidate markers from the huge number of genes within genomic data are particularly needed in the era of phylogenomics. Here, rather than using the traditional approach of comparing genomes of two distantly related taxa to develop conserved primers, we take advantage of the multiple genome alignment resources from the the University of California–San Cruz Genome Browser and present a simple and straightforward bioinformatic approach to automatically screen for candidate nuclear protein–coding locus (NPCL) markers. We tested our protocol in tetrapods and successfully obtained 21 new NPCL markers with high success rates of polymerase chain reaction amplification (mostly over 80%) in 16 diverse tetrapod taxa. These 21 newly developed markers together with two reference genes (RAG1 and mitochondrial 12S–16S) are used to infer the higher level relationships of tetrapods, with emphasis on the debated position of turtles. Both maximum likelihood (ML) and Bayesian analyses on the concatenated data combining the 23 markers (21,137 bp) yield the same tree, with ML bootstrap values over 95% and Bayesian posterior probability equaling 1.0 for most nodes. Species tree estimation using the program BEST without data concatenation produces similar results. In all analyses, turtles are robustly recovered as the sister group of Archosauria (birds and crocodilians). The jackknife analysis on the concatenated data showed that the minimum sequence length needed to robustly resolve the position of turtles is 13–14 kb. Based on the large 23-gene data set and the well-resolved tree, we also estimated evolutionary timescales for tetrapods with the popular Bayesian method MultiDivTime. Most of the estimated ages among tetrapods are similar to the average estimates of the previous dating studies summarized by the book *The Timetree of Life*.

Key words: phylogenomics, comparative genomics, data mining, tree of life, timetree, molecular dating.

Introduction

In recent years, molecular markers, primarily DNA and derived protein sequences, have become a fundamental means to reconstruct many parts of the "Tree of Life." However, phylogenetic inference based on a single gene or a few genes is rarely robust and often leads to conflicting results (Rokas et al. 2003). This is partly because small data sets contain fewer characters and often suffer from stochastic errors related to the length of the data. Moreover, individual gene genealogies may differ from each other and from the true organismal phylogeny due to mechanisms such as gene duplication, horizontal gene transfer, incomplete lineage sorting, and convergent evolution (the "gene-tree vs. species-tree" issue) (Pamilo and Nei 1988; Leaché and Rannala 2010), resulting in systematic incongruence between studies. One tempting and effective solution to these problems is to conduct phylogenetic inference by combining many independent nuclear loci for many species, that is, phylogenomic analysis (Delsuc et al. 2005; Philippe et al. 2005). Adopting a genome-scale approach theoretically increases the probability of obtaining a well-resolved and accurate tree by increasing the number of phylogenetically informative characters used for an analysis. More importantly, the systematic errors caused by the "gene-tree vs. species-tree" issue will probably be buffered in a multigene analysis. In theory, every gene sampled may bring systematic errors to a tree, but the occurrence of these errors is randomly distributed in the whole tree; stochastic error naturally diminishes when more and more genes are considered, thus the overall answer is still likely to be reliable.

Resolving the relationships among major tetrapod lineages is critical if we are to understand early land vertebrate evolution. To date, one of the challenges in reconstructing the tetrapod tree of life (see the review of Meyer and Zardoya 2003) is the phylogenetic position of turtles that has not yet been resolved and is widely debated based on morphological and molecular data. Currently, four main hypotheses concerning the phylogenetic position of turtles have been proposed (illustrated in fig. 1): "Hypothesis 1": Turtles are placed as the sister group to Diapsida (Gauthier et al. 1988; Lee 1997). This is the traditional view

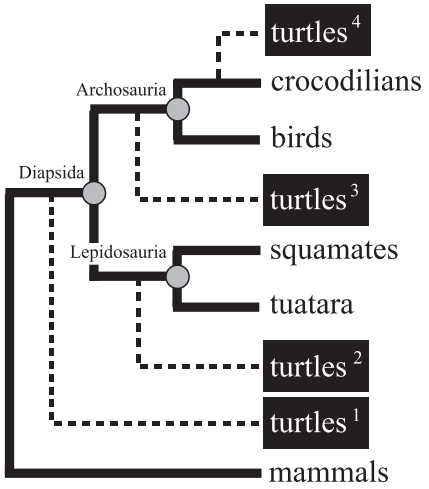


Fig. 1. The phylogenetic position of turtles in the amniote tree of life. Dashed lines separately indicate four possible hypotheses. (1) Turtles are the only survivors of anapsid reptiles and placed as the sister group of diapsid reptiles. (2) Turtles as the sister group of the Lepidosauria. (3) Turtles as the sister group of the Archosauria. (4) Turtles as the sister group of the crocodilians.

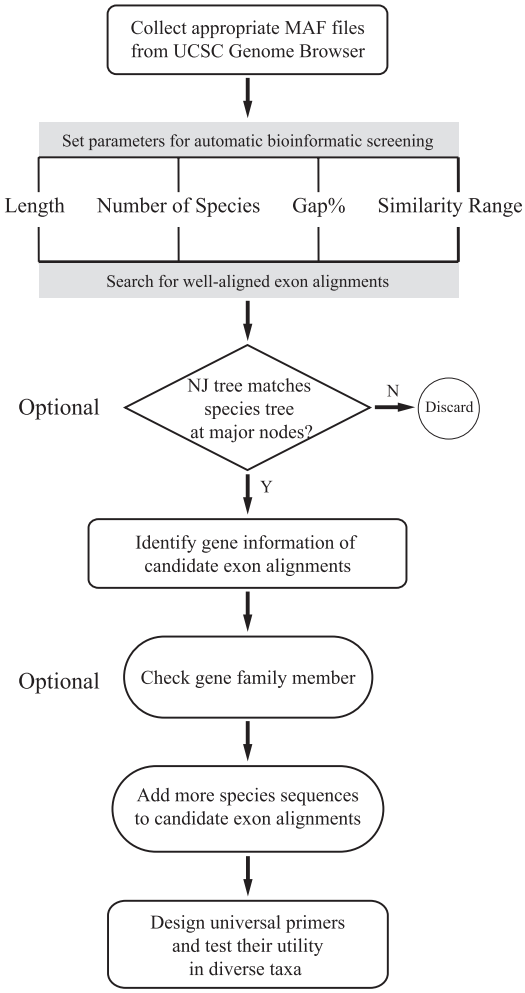
of turtles' placement, and it was recently supported by developmental biology evidence about the timing of organogenesis in turtles (Werneburg and Sánchez-Villagra 2009). "Hypothesis 2": Some other morphological studies, however, did not support the traditional view and placed turtles as the sister group to Lepidosauria (deBraga and Rieppel 1997; Lyson et al. 2010; Becker et al. 2011). "Hypothesis 3": In contrast to the morphological views, recent molecular phylogenetic studies tend to support a relationship of turtles as the sister group of archosaurs based on nuclear data (Iwabe et al. 2004; Hugall et al. 2007) and complete mitochondrial genomes (Zardoya and Meyer 1998; Kumazawa and Nishida 1999; Rest et al. 2003). "Hypothesis 4": Even under the assumption that turtles are close to Archosauria, they are sometimes placed as the sister group of crocodilians. This hypothesis was independently favored by DNA–DNA hybridization data (Kirsch and Mayer 1998), mitochondrial and nuclear genes (Hedges and Poling 1999; Cao et al. 2000), and genomic signatures (Shedlock et al. 2007). Considering that previous studies generally used only a limited number of independent markers, it is worthwhile to perform a phylogenomic analysis to see whether substantial amounts of data and a large number of independent markers can help to resolve the position of turtles.

The question about the position of turtles must be addressed in the context of the amniote tree of life, or more desirably, the tetrapod tree of life. Such a deep phylogenetic application requires highly conserved sequences, normally nuclear protein–coding loci (NPCL markers). Although many NPCL markers have been developed for phylogenomic studies in the past decade, they are normally designed for specific animal groups, such as mammals, birds, squamates, and teleosts (Murphy et al. 2001; Li et al. 2007; Townsend et al. 2008; Wright et al. 2008). Indeed, group-specific markers might also work on other animal groups, but it is normally necessary to redesign primers and optimize amplification conditions, which can be rather time consuming. Currently, universal nuclear sequence markers that work across diverse tetrapod taxa have been limited to a few "stock" genes, such as RAG1, C-mos, and POMC. Therefore, developing more universal nuclear sequence markers suitable for all major tetrapod groups has become increasingly important.

With the advances in genomic biology, genomic resources are deposited at an increasingly rapid pace and become easier and easier to access. Recently, there have emerged some reports about mining genomic data to obtain candidate NPCL markers: Li et al. (2007) used automated Blast comparisons of whole-genome sequences of two fish, zebrafish (*Danio rerio*) and pufferfish (*Takifugu rubripes*), to identify homologous exon regions. They were able to develop primers for ten relatively conserved NPCL markers useful for ray-finned fish systematics. Townsend et al. (2008) employed a similar approach in squamate reptiles, using the pufferfish and human genomes as comparative data for primer design. In order to increase the probability of successful amplification in squamates, Townsend et al. performed a second round of Blast searches to seek homologs of their candidate markers in the chicken genome and succeeded in identifying 26 NPCL markers applicable in ten tested squamate taxa and several additional vertebrates.

The above approaches of developing NPCL markers by mining genomic resources are basically through pairwise comparison of available genome data. The drawback of this method is the unpredictable polymerase chain reaction (PCR) success rate (PSR) of the newly developed markers for the target animal group. This is because the conserved exon regions used for primer design, which are identified by two-species alignments, may be variable or simply not exist in other distantly related species. This is why Townsend et al. (2008) manually added the chicken sequences into their alignments. A straightforward solution to this problem is to increase the taxon sampling density of genome alignments, that is, using multiple genome alignments (MGAs) rather than pairwise genome alignments (PGAs) for the development of NPCLs. Because of the tremendous computational load, aligning multiple animal genomes in personal desktop computers is still impractical. Nevertheless, recent advances in comparative genomics give us an opportunity to implement the idea: the University of California–San Cruz (UCSC) Genome Browser now provides a great number of well-aligned MGAs with different animal species combinations, which are free to download from its website (http://genome.ucsc.edu/). This resource is continuously updated with more and more animal genomes sequenced and provides a gold mine for identifying new NPCL markers to further resolve the vertebrate tree at different taxonomic levels.

In this study, we took advantage of the UCSC genome alignment resources and presented a simple and straightforward bioinformatic approach to automatically search for candidate NPCL markers. Our method incorporates two improvements compared with the previous studies of



Step 1: • Download compressed multiple alignment format (MAF) files from UCSC Genome Browser
• MAF files contain many genome fragment alignments

Step 2: • Screen genome fragment alignments in MAF files
criteria are (for example):
0.7kb ⩽ Alignment length ⩽ 10kb
Number of species in an alignment ⩾ 5
Gap% in an alignment ⩽ 2%
60%⩽ Similarity Range within an alignment ⩽ 90%

Step 3: • Build a NJ tree for each selected alignments
check if the NJ tree matches the expected species tree at most major nodes.
• Avoid potential aligning mistakes in MAF files
Avoid the ortholog-paralog problem

Step 4: • Batch BLAST candidate alignments to the Human genome
Obtain gene information for each selected alignment
e.g. gene name

Step 5: • Check gene family member for each selected alignments
e.g. through the Genecards website
• Discard genes with too many family members
Avoid troubles of paralog sequences

Step 6: • Lizard, Platypus sequences are normally missing in candidate alignments
• Use Ensembl to get more species sequences for candidate alignments

Step 7: • Locate less-conserved regions in an alignment
Design primers in the flanking highly conserved parts
Use amino acid sequences as guidance
• Test primers in target groups, redesign them if needed

FIG. 2. The workflow of the method described in this study for the development of new NPCL markers. A Python script was written to automatically perform Step 2.

using pairwise Blast searches to identify conserved exon regions. First, our method does not require aligning genomes locally and is more practical, especially for researchers with little experience in processing genome data. Second, our method involves more species in the initial genome alignments; thus, it is easier to identify "shared and conserved" exon regions across taxa, increasing the probability of successful PCR amplifications. We applied this method to tetrapods (land vertebrates) with the aim of determining the phylogenetic placement of turtles. With the 21 newly developed markers, we tried to address the question—how much data is needed to robustly resolve the phylogenetic position of turtles? Furthermore, based on substantial amounts of data and well-resolved trees, we estimated divergence times for major split events along tetrapod evolutionary history and provided more information about the tetrapod timetree of life.

Materials and Methods

Development of NPCL Markers

The workflow to search for NPCL markers in our study can be divided into seven steps and is illustrated in figure 2. The first step is to retrieve MGAs (in MAF format) from the UCSC Genome Browser (http://genome.ucsc.edu/). The MAF file is an aggregate of a huge number of small genome fragment alignments. To ensure candidate NPCL markers can be amplified in tetrapods, we only use those MAF files that have species coverage from ray-finned fishes to mammals. For example, the multiple alignment of four vertebrate genomes with *Xenopus tropicalis* contains five species: zebrafish, frog, chicken, mouse, and human, thus it meets our requirements. The second step is to search for well-aligned alignments that meet certain criteria in the MAF files. For the aforementioned five-species MAF file, the following criteria are used: the length of an alignment ranging from 0.7 to 10 kb, minimum number of species of an alignment no less than 5, the percentage of gap sites in an alignment no more than 2%, and the sequence similarity within an alignment ranging from 60% to 90%. The screening procedure is automatically carried out with a Python script (source code available upon request). The third step is to build a simple neighbor joining (NJ) tree for each selected alignment and check if it agrees with the expected species tree at most major nodes. Those alignments whose NJ trees differ from the expected species tree at

Table 1. List of all Species Used in This Study; Species with Genome Data Available Are Shaded.

Taxonomy		Species	Common Name (short)	Source or Collection Locality
Mammalia	Primates	*Homo sapiens*	Human	Public Genome Project
	Proboscidea	*Loxodonta africana*	Elephant	Public Genome Project
	Metatheria	*Monodelphis domestica*	Opossum	Public Genome Project
	Prototheria	*Ornithorhynchus anatinus*	Platypus	Public Genome Project
Aves	Paleognathae	*Struthio camelus*	Ostrich	Commercial food source
	Anseriformes	*Anas platyrhynchos*	Duck	Commercial food source
	Galliformes	*Gallus gallus*	Chicken	Public Genome Project
Crocodylia	Alligatoridae	*Alligator sinensis*	Alligator	Alligator breeding center, Xuancheng, China
	Crocodylinae	*Crocodylus siamensis*	Crocodile	Commercial food source
Testudines	Podocnemididae	*Podocnemis unifilis*	Side-necked turtle	Private captivity
	Emydidae	*Trachemys scripta*	Pond turtle	Commercial food source
	Carettochelyidae	*Carettochelys insculpta*	Pig-nosed turtle	Private captivity
	Trionychidae	*Pelodiscus sinensis*	Softshell turtle	Commercial food source
Squamata	Dibamidae	*Dibamus bourreti*	Dibamid	Hongkong, China
	Gekkonidae	*Hemidactylus bowringii*	Gecko	Guangzhou, Guangdong, China
	Scincidae	*Scincella reevesii*	Skink	Guangzhou, Guangdong, China
	Serpentes	*Naja naja atra*	Snake	Shaoguan, Guangdong, China
	Iguania	*Anolis carolinensis*	Iguanian	Public Genome Project
Lissamphibia	Gymnophiona	*Ichthyophis bannanicus*[a]	Caecilian	Beiliu, Guangxi, China
	Caudata	*Batrachuperus yenyuanensis*[a]	Salamander	Xichang, Sichuan, China
	Anura	*Silurana tropicalis*	Clawed frog	Public Genome Project
		Rana nigromaculata[a]	Pond frog	Guilin, Guangxi, China
Dipnoi	Protopteridae	*Protopterus annectens*	Lungfish	Private captivity
Actinopterygii	Teleostei	*Takifugu rubripes*	Fugu	Public Genome Project
		Tetraodon nigroviridis	Tetraodon	Public Genome Project
		Gasterosteus aculeatus	Stickleback	Public Genome Project
		Oryzias latipes	Medaka	Public Genome Project
		Danio rerio	Zebrafish	Public Genome Project

[a] A few markers were not able to be amplified from this species; to reduce missing data, other related species were used for supplementary PCR amplifications (for details, see supplementary table S1, Supplementary Material online).

most major nodes are discarded because they may contain paralog genes or aligning errors (this step is optional; see the later Discussion). The fourth step is to identify the gene name for each of the selected alignments by batch Blast searches against the human genome. The fifth step is to check the member number of the gene family that a candidate gene belongs to (through the Genecards website and HomoloGene in NCBI). We tried to avoid using genes with many similar family members as candidate markers because misamplified paralog sequences often interfere with phylogenetic inference in practical applications. We empirically set the cutoff number as 4. The sixth step is to manually add more sequences of other species to selected alignments. Because some species (e.g., lizard and platypus) are normally not included in the MAF files retrieved from UCSC, we collected relevant sequences of these species from the ENSEMBL database and aligned them to the corresponding UCSC alignments by ClustalW (Thompson et al. 1997). These rebuilt alignments with more species (often no less than eight) can provide conserved regions for designing universal primers without bias toward certain tetrapod groups. The last step is to design universal primers and test their utility in tetrapods. We translated all candidate DNA alignments into amino acid alignments and manually located the less-conserved regions for marker development in order to increase the informativeness of our markers. Primers were designed on highly conserved blocks in the flanking regions. To reduce primer degeneracy, whenever possible, we tried to design the primers on conserved blocks without residues of high degeneracy (e.g., L, R, S).

Taxon Sampling and Experimental Procedures

We selected 28 taxa for our study, representing six major tetrapod lineages (amphibians, squamates, turtles, birds, crocodilians, and mammals) and two outgroup lineages (ray-finned fishes and lobe-finned fishes). We included at least two taxa for each major tetrapod lineage and, in order to reduce the long-branch attraction (LBA) artifact (Bergsten 2005) and, date phylogenetic events more accurately, the selected taxa usually spanned the basal split of each group. Among the 28 selected taxa, 12 taxa had public genome data, whereas sequences for the remaining 16 taxa needed to be generated. Detailed information on all taxa used in this study is listed in table 1.

Total genomic DNA was extracted from ethanol-preserved tissues (liver or muscle) using the standard salt extraction protocol. A total of 23 markers were amplified, including 22 nuclear genes and 1 mitochondrial DNA (mtDNA) fragment (table 2). Each pair of primers was initially tested in 25 μl reaction volumes with Ex*Taq* DNA polymerase (Takara, Dalian) for 16 phylogenetically diverse taxa, using the following cycling settings: an initial denaturation step of 4 min at 94 °C, followed by 35 cycles of a 45 s

Table 2. PCR Primers Used to Amplify the 1 mtDNA and 22 NPCL Markers.

	Forward Primer		Reverse Primer		Fragment	Genomic
Gene	Name	Sequence (5′→3′)	Name	Sequence (5′→3′)	Size (bp)≈	Location[a]
BCHE	BCHE-10F	GARATGTGGAAYCCNAANAC	BCHE-750R	CCTTCATCTTTRTTNACNCC	750	Chr.3
BPTF	BPTF-230F	GARCARTGCACNCTNATGGCNGA	BPTF-820R	CKYCKGTTNARRAACCARTAYTT	600	Chr.17
CAND1	CAND1-160F	TGTGTKGGWGAYCCNTTYTAYAA	CAND1 -1370R	CCARATGTTYTCNACATANGGYTT	1230	Chr.12
CHAD	CHAD-30F	GACCTNCARCAYGTCATHTGYGA	CHAD-670R	TAKCGMCCAAARGWCTGGAANGC	650	Chr.17
DOLK	DOLK-10F	CGMTGCTTYACHCCYGGNGARGC	DOLK-870R	GTYTTYTTDGTNCCNGGCCA	860	Chr.9
FAT4	FAT4-56F	GTSBTGGAYACNCARGAYAAYCC	FAT4-817R	TGVCCATCNGGRAADATNCCRAA	800	Chr.4
FICD	FICD-130F	TACTAYCAYCAYATHTAYCAYAC	FICD-700R	AARGGCCKVACRTCNCCYTCRTT	580	Chr.12
GLCE	GLCE-200F	GTGGTNCTRGAGACNACNGARAA	GLCE-1030R	ATGTGNGTSGTRTGRTARTCCCA	830	Chr.15
GPER	GPER-100F	ATGACCATYCCNGAYCTKTAYTT	GPER-660R	ATGAAGACRTTYTCNGGNAGCCA	570	Chr.7
KIAA1239	KIAA1239-10F	CARCCTTGGGTNTTYCARTGYAA	KIAA1239-1000R	TTCACRAANCCMCCNGAAAAYTC	1020	Chr.4
LIG4	LIG4-10F	AGRATGGCBTAYGGMATHAARGA	LIG4-1100R	GTTCMCCDCKTTTRTCYGGYTTGTA	1100	Chr.13
			LRRN1-1040R	GTKAGGTTRTAYTCRTGNACRTC	900	Chr.3
LRRN1	LRRN1-150F	AAAGARCTKGGNATHAAYAAYATG	LRRN1-820R	ATRTGWGGRTTRTCDATYTTCAT	840	
MACF1	MACF1-10F	CARTTCCAGCANATGTTYGAYGA	MACF1-1020R	TCYGCCARYTGNGARAACATYTC	1030	Chr.1
PANX2	PANX2-10F	GAGGARCCHATHTAYTGYTAYAC	PANX2-400R	CTYTTYTCCTTCTCNGCRTTYTC	390	Chr.22
PDP1	PDP1-10F	GTKCCHGARTTYGATGGNAARAA	PDP1-1270R	ATRGTDATGTCRTCNCTRTACAT	1270	Chr.8
PLCL2	PLCL2-714F	GTNCARTTYTCNAGYAAYAARGA	PLCL2-1708R	TTCCARAARTCYTGNGGRTTCAT	1040	Chr.3
PPL	PPL-400F	GTSAARGAGGTSCTRCGSATHGA	PPL-2470R	ATYTCYTCCCAGTCGCAYTCYTG	2120	Chr.16
	PPL-740F	AAGGARGTGYTKAARGTNGARAA			1750	
HYP[b]	HYP-198F	GARTGGYTNAARAARTTYTGGTT	HYP-1126R	ACCTTNKGYTCNCCDATDATCCA	1000	Chr.11
RERE	RERE-10F	GAGTACGCYCGKCCYCAYGTNATG	RERE-510R	TGRTGNGGMGTSACRTTRAACAT	510	Chr.6
	SACS-3341F	ATGGAYCCNATGAAYGTNTTYTA	SACS-4555R	ATDATRCANGCNGTRCAYTCCAT	2000	Chr.13
SACS	SACS-60F	TAYCARCCAACWTAYACNTAYGC	SACS-940R	CATTTRAAGCANACCCAYTCRTT	850	
TTN	TTN-3680F	GATGGNMGKTGGYTNAARTGYAA	TTN-4573R	AGRTCRTANACNGGYTTYTTRTT	940	Chr.2
	RAG1-2900F	AGCTGCAGYCARTACCAYAARATGTA	RAG1-3300R	AACTCAGCTGCATTKCCAATRTCACA	980	Chr.11
RAG1[c]	RAG1-3000F	ACAGGATATGATGARAAGCTTGT	RAG1-3900R	TTRGAGGTGTAGAGCCARTGRTGYTT	890	
12S-16S[c]	12SAL[d]	AAACTGGGATTAGATACCCCACTAT	16S2000H[d]	GTGATTAYGCTACCTTTGCACGGT	1500	Mitochondrion
	LX12SN1[d]	TACACACCGCCCGTCA	LX16S1R[d]	GACCTGGATTACTCCGGTCTGAACTC	1600	

[a] Human genome as reference.
[b] Hypothetical protein.
[c] Markers were amplified in two overlapping fragments.
[d] Zhang et al. (2008).

denaturation at 94 °C, a 40 s annealing at 50 °C, and a 1.5 min elongation at 72 °C, followed by a final ten min at 72 °C. Sometimes the universal primers failed to work for a particular species even after PCR optimization. In these cases, we chose a phylogenetically close species to replace the original species and redid the PCR amplification, for example, used *Ranodon sibiricus* (Siberian salamander) to replace *Batrachuperus yenyuanensis* (Yenyuan stream salamander) (for details, see supplementary table S1, Supplementary Material online). Because our concerns are high-level relationships of tetrapods, a small number of mosaic sequences at terminal nodes should not influence phylogenetic inferences. All PCR products were purified by gel cutting and then cloned into a PMD19-T vector (Takara, Dalian). Positive recombinant clones were identified by colony PCR, and the PCR products (at least two) were cleaned with ExoSap treatment and sequenced on an ABI3730 DNA sequencer. All sequences were examined by performing a Blast search against the human genome to make sure they are our target genes.

Phylogenetic Analyses

Nuclear and mitochondrial sequences were aligned using ClustalW (Thompson et al. 1997) with default settings. Ambiguous alignment regions were removed using Gblocks (Castresana 2000) with minimum length of a block set to 8 and no gaps allowed; otherwise, default settings were assumed. Finally, three DNA data sets were generated for phylogenetic analyses: Data set I (mtDNA; 1,289 bp), Data set II (22 nuclear genes; 19,848 bp), and Data set III (mtDNA +22 nuclear genes; 21,137 bp).

The three DNA data sets were separately analyzed with both maximum likelihood (ML) and Bayesian inference (BI) methods under a partitioned scheme (by genes). Partitioned ML analyses were implemented using RAxML 7.0.3 (Stamatakis 2006) with 100 inferences and with GTR+I+Γ models assigned to each partition (-q option). Branch support was assessed with 1,000 rapid bootstrap replicates implemented in RAxML. Partitioned Bayesian analyses were performed in MrBayes 3.1.2 (Huelsenbeck and Ronquist 2001). The best-fitting model for each gene was separately selected by Akaike information criterion, implemented in MrModelTest2.3 (Nylander 2004). As a result, the GTR+Γ model was chosen for the DOLK gene, and the GTR+I+Γ model was chosen for the remaining 22 markers. Two Markov chain Monte Carlo (MCMC) runs (Unlink Revmat = (all) Statefreq = (all) Shape = (all) Pinvar = (all)) were performed with one cold and three heated

chains (temperature set to 0.2) for 30 million generations and sampled every 1,000 generations. The chain stationarity was visualized by plotting *-lnL* against the generation number using Tracer version 1.4 (http://evolve.zoo.ox.ac.uk/beast/help/Tracer), and the first 15–50% generations were discarded. Topologies and posterior clade probabilities from the two runs were compared for congruence.

Species trees estimations were implemented using BEST (Bayesian estimation of species trees) version 2.3 (Liu 2008). This method can estimate species tree from multiple genes without data concatenation. Both Data set II and Data set III were analyzed. We applied the same gene-specific substitution models for each data set as we did in the partitioned Bayesian analyses. In the BEST analysis, the prior for the theta is defined as invgamma (α, β) and is the most important parameter. When $\alpha > 1$, the mean of theta $= \beta/(\alpha - 1)$. For Data set II and Data set III, the mean of theta is 0.6 and 0.7, respectively. We separately tried different invgamma (α, β) combinations for Data set II and Data set III. When $\alpha = 2$ or 3, the MCMC chains hardly reached convergence after 30 million generations. When $1 < \alpha < 2$, the chains converged after ten million generations. We had tried $\alpha = 1.2$, $\alpha = 1.5$, $\alpha = 1.8$ (β is set accordingly) and found that when $\alpha = 1.2$, the chains converged most quickly. Therefore, we chose $\alpha = 1.2$, then β is 0.12 and 0.14 for Data set II and Data set III, respectively. The prior of the GeneMu was set to uniform (0.5, 1.5) for both Data set II and Data set III. Other priors were default settings and all parameters were unlinked. For each data set, two independent runs were conducted, each with one cold and three heated chains (temperature set to 0.2) for 40 million generations and sampled every 1,000 generations. Checking convergence of the MCMC chains and summarizing topologies and posterior probabilities followed the same methodology as we did in the MrBayes analyses.

Estimating the Minimum Data Needed to Robustly Resolve the Position of Turtles

In order to explore the cause of previous conflicting results about the placement of turtles, we considered investigating the minimum amount of data necessary to resolve the position of turtles with strong support. To this end, we generated subsamples with different length from the concatenated 23-marker DNA data set (21,137 bp) by the Jackknife method implemented in the Seqboot program of the Phylip 3.68 package (Felsenstein 2005). For a given length, we randomly generated 200 subsamples. Ultimately, 20 data sets of different length were constructed, ranging from 1 to 20 kb, each containing 200 subsamples. These data sets were subjected to ML analyses by both PhyML version 3.0 (Guindon and Gascuel 2003) and RAxML 7.0.3 (Stamatakis 2006). Branch supports were estimated by either aLRT (an approximation to bootstrap but using the likelihood ratio test; Guindon and Gascuel 2003) or nonparametric bootstrapping. The aLRT analyses were performed in PhyML with a GTR+Γ+I model, where base frequencies, proportion of invariable sites, and gamma shape distribution (six categories) parameters were estimated from the data. The bootstrap analyses were implemented in RAxML (200 rapid bootstrap inferences) with a GTR+I+Γ substitution model. Eight data sets (1, 4, 7, 10, 13, 16, 19, and 20 kb; 200 replicates for each level, a total number of 1,600 subsamples) were analyzed. The flow chart of the aforementioned procedures is illustrated in the supplementary figure S1 (Supplementary Material online).

Molecular Dating

It has been shown that using a single or a few points to calibrate a large phylogeny can result in high estimation errors for the divergence times of distantly related nodes (Müller and Reisz 2005). Here, we used a total of eleven calibration nodes. Three well-estimated nodes were used as hard constraints with lower and upper bounds: the lungfish–tetrapod split (408–419 Ma; Müller and Reisz 2005), the bird–mammal split (312–330 Ma; Benton and Donoghue 2007), and the Monotremata–Theria split (163–191Ma; Benton and Donoghue 2007). Eight additional nodes were constrained with minimal bounds only, based on known fossil records: the origin of living tetrapods was constrained to be at least 330 Ma (*Lethiscus stocki*, Ruta et al. 2003), the split between bird and lizard at least 252 Ma (*Protorosaurus speneri*, Evans and King 1993), the common ancestor of archosaurians at least 235 Ma (*Vjushkovisaurus*, Benton 1993), the split between marsupials and placentals at least 124 Ma (*Eomaia scansoria*, Ji et al. 2002), the split between human and elephant at least 71.2 Ma (*Phosphatherium* and *Daouitherium*, Gheerbrandt et al. 2005), the origin of living turtles at least 193 Ma (*Proterochersis*, Gaffney 1986), the origin of living birds at least 66 Ma (*Vegavis*, Clarke et al. 2005), the split between alligator and crocodile at least 66 Ma (*Stangerochampsa*, Brochu 1999).

Molecular dating under a relaxed molecular clock Bayesian method was implemented in MultiDivTime (Thorne and Kishino 2002). The ML tree from the concatenated DNA alignment (21,137 bp) was used as the reference tree. Using the five ray-finned fishes as outgroup, the lungfish–tetrapod split was regarded as the ingroup root. Because the lungfish sequences were absent for some genes, the ingroup root age could not be applied to each gene, thus all 23 genes were combined as a single "super gene." The model parameters were calculated with an F84 + G model, using the Baseml program in PAML package (Yang 1997). Optimized branch lengths with their variance–covariance matrices of the DNA data set were estimated with the program Estbranches_dna, a component of MultiDivTime. The priors for the mean and standard deviation of the ingroup root age, rttm and rttmsd were set to 4.13 and 0.05, respectively. The prior mean and standard deviation for the gamma distribution describing the rate at the root node (rtrate and rtratesd) were both set to 0.15. These values were based on the median of the substitution path lengths between the ingroup root and each terminal, divided by rttm. The prior mean and standard deviation for the gamma distribution of the parameter controlling rate variation over time (i.e., brownmean and brownsd) were both set to 0.5. After ignoring 300,000 initial cycles (to ensure

Table 3. .Summary Information of the 23 Gene Markers Amplified in 16 Taxa.

Gene	Length of Refined Alignments	Var. Sites	PI Sites	Alpha	Pinvar	TL	Sub. Rate	Treeness	RCV	Treeness /RCV	Topological Similarity to the Final ML Tree (%)	No. of Successful PCR (% of 16 total)	The Phylogenetic Position of Turtles
BCHE	702	518	451	1.519	0.203	4.915	0.669	0.297	0.060	4.947	84.2	15 (94)	Hypothesis 4
BPTF	552	277	232	1.229	0.420	3.487	0.626	0.194	0.218	0.889	79.6	15 (94)	Hypothesis 3
CAND1	1173	597	536	1.283	0.436	3.645	1.263	0.262	0.257	1.020	88.4	14 (88)	Hypothesis 3
CHAD	594	402	363	1.450	0.270	4.586	1.340	0.234	0.138	1.697	73.5	13 (81)	Turtle + bird
DOLK	813	642	489	0.607	0.000	6.258	0.903	0.223	0.151	1.473	76.7	16 (100)	Hypothesis 3
FAT4	762	503	424	1.073	0.237	4.327	1.126	0.250	0.120	2.084	83.1	15 (94)	Hypothesis 4
FICD	531	284	242	1.218	0.401	3.739	1.224	0.234	0.255	0.920	68.2	11 (69)	Turtle + bird
GLCE	780	460	384	0.891	0.331	5.161	0.808	0.229	0.229	0.999	65.3	15 (94)	Hypothesis 1
GPER	522	294	253	1.172	0.375	4.362	1.305	0.220	0.267	0.826	70.4	15 (94)	Hypothesis 4
KIAA1239	966	525	456	1.236	0.384	3.583	0.728	0.277	0.126	2.205	90.4	15 (94)	Hypothesis 3
LIG4	1056	676	573	1.117	0.278	4.428	1.018	0.252	0.177	1.427	81.2	15 (94)	Hypothesis 3
LRRN1	840	482	405	0.886	0.340	4.918	1.031	0.221	0.146	1.508	77.7	16 (100)	Hypothesis 3
MACF1	978	754	669	0.893	0.116	5.898	1.394	0.313	0.128	2.433	86.7	12 (75)	Hypothesis 3
PANX2	342	159	138	1.850	0.476	3.263	0.685	0.268	0.210	1.277	71.1	16 (100)	Polytomy
PDP1	1218	638	565	1.093	0.412	4.073	0.928	0.258	0.198	1.308	85.2	15 (94)	Hypothesis 3
PLCL2	993	575	466	0.940	0.318	4.080	0.970	0.205	0.167	1.226	93.6	13 (81)	Hypothesis 3
PPL	2070	1445	1280	1.419	0.243	4.897	1.239	0.252	0.128	1.973	93.5	15 (94)	Hypothesis 3
HYP	936	491	437	1.105	0.403	3.727	0.875	0.248	0.176	1.409	71.7	15 (94)	Polytomy
RERE	462	235	197	0.963	0.397	3.835	1.056	0.235	0.172	1.364	70.1	15 (94)	Turtle + bird
SACS	1152	643	509	1.414	0.398	4.258	0.956	0.184	0.135	1.367	90.9	16 (100)	Hypothesis 3
TTN	894	530	448	1.234	0.318	3.540	0.849	0.265	0.072	3.698	94.2	16 (100)	Hypothesis 3
RAG1	1512	879	791	1.194	0.359	4.452	0.891	0.230	0.115	1.991	91.6	16 (100)	Hypothesis 3
mtDNA 12S-16S	1289	748	594	0.728	0.254	3.935	0.934	0.165	0.081	2.046	80.9	16 (100)	Hypothesis 3

PI sites, parsimony informative sites; Alpha, shape parameter of the gamma distribution; Pinvar, proportion of invariable sites; TL, total tree length; Sub. Rate, relative substitution rate estimated using Bayesian approach; Treeness, proportion of tree distance on internal branches; RCV, relative composition variability; HYP, hypothetical protein.

that the Markov chain reached stationarity), the Markov chain was sampled every 100 cycles until a total of 15,000 samples was collected. To check for convergence, three independent runs were performed and similar results were observed.

Results and Discussion

New NPCL Markers

The main MGA file used in this study was downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/xenTro1/multiz5way/. The decompressed MAF file is 1.04 gb in size includes five species: zebrafish, frog, chicken, mouse, and human and contains 1,901,257 genome fragment alignments. The prefiltered process (0.7 kb $<$ length $<$ 10 kb, species $\geq$ 5, gap% $<$ 2%, 0.6 $<$ similarity $<$ 0.9) selected 141 fragment alignments. Among them, 21 alignments were discarded due to apparent incongruence between their NJ trees and the expected species tree (zebrafish, (frog, (chicken, (mouse, and human)))). This step should not bias the results for the placement of turtles because the five-species tree is well established and unrelated to turtles. The remaining 120 alignments were Blast to the human genome and all identified as NPCLs. Among them, 73 genes belonged to gene families with over four members and were discarded according to our criterion. The remaining 47 alignments were finally chosen as NPCL marker candidates.

As a positive control of the utility of our searching method, the widely used nuclear marker RAG1 was within our candidate data set. Because previous studies have placed the recommended number of independent genes for robust inference at about 20 (Rokas et al. 2003), we randomly selected 21 novel NPCLs from the 47 NPCL marker candidates. For each selected NPCL alignment, PCR primers were designed on conserved blocks flanking less-conserved regions, with the size of the amplified fragments ranging from 400 to 2000 bp (table 2). For comparative purposes, the RAG1 gene and an mtDNA fragment (from 12S to 16S) were also included in this study. We tested the utility of our universal primers on 16 sarcopterygian species (one lobe-finned fish, three amphibians, four squamates, four turtles, two crocodians, and two birds; see table 1). The success rate of obtaining target products for each markers are normally 80–100%, with two exceptions of FICD (69%) and MACF1 (75%), which indicates that our newly developed markers work for most tetrapod taxa (table 3). The newly generated sequences are deposited in GenBank under accession numbers HQ902269–HQ902589 (supplementary table S1, Supplementary Material online).

To describe the characteristics of the 23 markers, we performed ML and Bayesian analyses (under a GTR+I+Γ model) for each marker and summarized parameters such as gamma shape parameter (α), proportion of invariable sites (Pinvar), total tree length, relative substitution rate, relative composition variability (RCV), and proportion of internal branch length (Treeness). Both RCV and Treeness were calculated following Phillips and Penny (2003). In addition, we also calculated the topological similarity (TS)

between the ML tree of each marker and the final ML tree from the concatenated 23 marker data set and used it as an indicator of phylogenetic performance for each marker. The TS was estimated following Nye et al. (2006). All the abovementioned parameter values for the 23 markers were listed in table 3.

The range of relative substitution rate is nearly 2-fold among the 23 markers used here (table 3), whereas the two commonly used phylogenetic markers (RAG1 and mtDNA 12S–16S) have moderate substitution rates (0.891 and 0.934, respectively; table 3). This result suggests that evolutionary rates of our 21 newly developed nuclear markers are not biased toward one direction. According to table 3, none of the parameters appear strongly correlated with the TS values (phylogenetic performance). We empirically use the TS (phylogenetic performance) and the PSR (experimental performance) to classify all 23 markers. Three levels for the TS: <80%, 80–90%, >90%, and two levels for the PSR: <80%, and >80% are set. Based on this criterion, the widely used RAG1 gene (TS = 91.6% and PSR = 100%) is our "recommended" marker, which means it is easy to amplify and has a high chance (~0.92) of obtaining correct results. The mitochondrial 12S–16S ribosomal RNA (rRNA) gene is a good marker because although it is easy to use (PSR = 100%), it has only a moderate chance (TS = 81%) to get the right tree. Following the above classification, five markers (TTN, SACS, PPL, PLCL2, and KIAA1239) are at the recommended level, comparable to RAG1; five markers (BCHE, CAND1, FAT4, LIG4, and PDP1) are at the good level, comparable to mitochondrial rRNA genes; the remaining eleven markers are at the "ordinary" level, which means they are either not so easy to use or have only a fair chance to recover the right tree. It should be noted that the classification is based on deep tetrapod phylogeny; the phylogenetic performance of each marker to a given taxonomic group (amphibians, turtles . . .) needs to be tested in future studies.

Practicality of Developing NPCL Markers Based on MGAs

In this study, we reported an automatic approach to develop NPCL markers based on MGAs available in the UCSC genome browser (see fig. 2). In theory, our approach is similar to that of Li et al. (2007) and that of Townsend et al. (2008). All methods aim to seek for "shared and conserved" regions by aligning genome data. However, there is one important difference. The previous two studies firstly identified putative homologs between two reference species and then proceeded to seek for continuous open reading frames within these genes. Using this strategy, the success rate of the newly developed markers cannot be guaranteed for the target animal group. This is because the conserved exon regions used for primer design, which are identified by two-species alignments, may be variable or simply not exist in other distantly related species. To overcome this drawback, Townsend et al. (2008) manually added the chicken sequences into their alignments to develop markers workable for squamate reptiles. Our method simplified the overall process of identifying conserved regions by using MGAs as the input data. This can ensure locating "shared and conserved" regions, and more importantly, facilitate automation of the process.

One important reason why the UCSC MGA data can be used for NPCL marker development is that the MGA is not a continuous huge alignment but an aggregate of many subalignments of alignable regions of the genomes. Only those shared and conserved regions across given taxa can be aligned into an MGA. More importantly, a genome region that can be aligned into a subalignment within an MGA is usually a continuous coding region. So it is not necessary to identify whether a selected conserved region is split by noncoding sequences within the group of interest. To some extent, an MGA has done the job of screening conserved and continuous coding regions from the genome data, which is the most complicated step of the previous strategy, and thus greatly simplifies the overall analysis.

By choosing the appropriate MGA data, our method is easy to apply to other taxonomic groups of interest. For example, we can use the MGA (http://hgdownload.cse.ucsc.edu/goldenPath/oryLat2/multiz5way/) which contains five fishes: medaka, stickleback, tetraodon, fugu, and zebrafish to develop NPCL markers for teleosts only; or use the MGA (http://hgdownload.cse.ucsc.edu/goldenPath/ornAna1/multiz6way/) which includes six amniotes: platypus, opossum, human, mouse, chicken, and lizard to develop NPCL markers for amniotes (or to say reptiles). Besides MGAs, the UCSC Genome Browser also provides many PGAs, which in some cases, can serve as a supplementary data source for NPCL development in a very specific group. If we use PGAs, our method is almost identical to the previous studies (Li et al. 2007; Townsend et al. 2008), but the whole data processing procedure will still be greatly simplified.

As in previous methods, our method also searches for marker candidates under given criteria. In our experience, the most influential screening parameter in our method is the similarity range. The minimum and maximum bounds must be set according to the species within an MGA. Taking the aforementioned five-species MGA (zebrafish, frog, chicken, mouse, and human) as paradigm, the maximum bound actually refers to the similarity between human and mouse. If it is set to 95% instead of 90% (as in the current study), more candidate alignments will be obtained but the additional part will be slowly evolving genes. Decreasing the minimum bound from 60% (as in the current study) to 50% will retrieve more rapidly evolving genes, many of which are sometimes too variable to design universal primers. Therefore, the similarity range parameter needs to be optimized according to the different MGAs being analyzed. Comparing NJ trees of candidate alignments with the expected species tree is a strict criterion used in this study. This step is done to avoid possible aligning errors within MGAs (align paralog genes with ortholog genes). For example, if the NJ tree of a selected alignment is (human, (zebrafish, (frog, (mouse, and chicken)))), the human sequence may be

a paralog gene. However, if the expected species tree is not a well-established one, we suggest skipping this step to avoid introducing possible bias in the marker selection. Likewise, eliminating gene candidates with over four gene family members is to avoid the paralog problem but is totally empirical. Skipping this step normally saves 40–50% gene candidates, but we have no experience on developing markers from these candidates. More experimental explorations may be needed to determine the optimal cutoff values of this parameter or to clarify whether the whole step is necessary.

The Multilocus Data Set Places Turtles as the Sister Group to a Monophyletic Cluster of Archosaurs (Birds and Crocodilians)

Partitioned ML and BI on the Data set II (22 nuclear genes) and Data set III (22 nuclear genes + one mtDNA gene) yielded identical tree and similar branch supports (fig. 3). This result indicates that mitochondrial phylogenetic signals do not overwhelm nuclear signals thus analyses of combined mitochondrial and nuclear sequences are appropriate. ML and Bayesian analyses on the mtDNA gene alone show some differences with the multilocus analyses. However, many branches of the mtDNA tree are only weakly supported (ML bootstrap $< 60\%$) thus the tree topology is not presented. The species tree estimated from the Data set II and Data set III using BEST (fig. 4) is very similar to those estimated from the concatenated analyses (fig. 3). Only within living squamates, relationships are somewhat different from the concatenated analyses but only weakly supported. Overall, these results strongly suggest that the results from the concatenated analyses are not compromised by discordance between gene trees and the species trees.

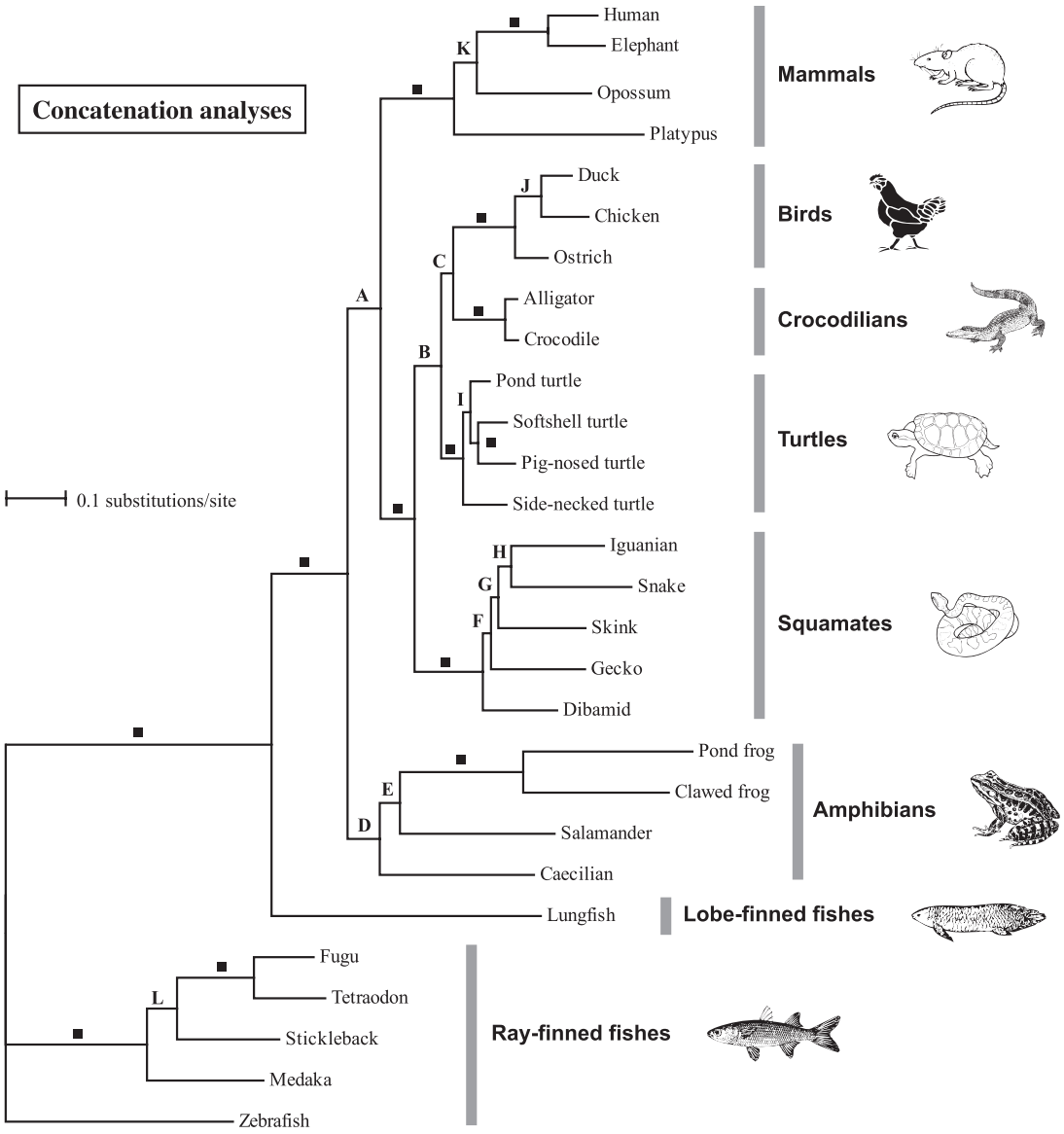
In the recovered trees (figs. 3 and 4), all acknowledged natural groups (amphibians, squamates, turtles, crocodilians, birds, and mammals) are well supported. Although we aimed to address the position of turtles, before that, we may firstly evaluate certain debatable interrelationships of other tetrapod groups to test the reliability of our trees. For example, within the amphibians, the interrelationship among the three living groups (frogs, salamanders, and caecilians) is recovered as (caecilians, (salamanders, frogs)) (figs. 3 and 4), favoring the Batrachia hypothesis (a salamander–frog clade), in agreement with most recent molecular studies (e.g., Zhang et al. 2005; Zhang and Wake 2009; Hugall et al. 2007; Roelants et al. 2007). As for the squamates, the concatenated analyses showed that the dibamids (Dibamidae) diverged early in squamate evolutionary history, followed by geckos (Gekkota), skinks (Scinciformata), snakes (Serpentes), and iguanians (Iguania), ordinally (fig. 3). The branching order is identical to the current view of higher level squamate relationships (Townsend et al. 2004; Vidal and Hedges 2004, 2005; Wiens et al. 2010). However, relationships within squamates from the BEST analysis (fig. 4; but not strongly supported) show some differences with the concatenated analyses, indicating that there is still discordance between gene and species trees in this part and more markers should be used in the future. Overall, consistency between our results and current opinions on controversial nodes of tetrapod phylogeny has raised confidence in our phylogenomic reconstruction and demonstrated the utility of our newly developed NPCL markers.

So far, molecular studies have not completely clarified the placement of turtles among the amniote tree. Although most of the previous molecular studies have favored neither Hypothesis 1 nor Hypothesis 2 (see fig. 1), different genes and sampling schemes have in some cases suggested that turtles are the closest relatives of crocodilians to the exclusion of birds (Hypothesis 4; fig. 1; Hedges and Poling 1999; Cao et al. 2000; Shedlock et al. 2007), and in other cases, turtles are the closest relatives of the whole archosaur clade (Hypothesis 3; fig. 1; Zardoya and Meyer 1998; Rest et al. 2003; Iwabe et al. 2004; Hugall et al. 2007). In particular, although these previous studies generally used substantial amounts of data ($>$10 kb), the number of independent markers were no more than three (all mitochondrial genes should be considered as only one marker because they are genetically linked in the mitochondrial genome). The only exception is that of Hedges and Poling (1999), which used 23 nuclear and 2 mitochondrial genes but with limited taxon sampling. The incongruence on the phylogenetic position of turtles between different studies is likely due to insufficient marker sampling or limited taxon sampling. As a phylogenomic effort to resolve the position of turtles, our phylogenetic analyses based on 23 independent markers (figs. 3 and 4) clearly shows that turtles are the sister group of the whole archosaurs (birds + crocodilians) but not the crocodilians only, and the monophyly of archosaurs is strongly supported.

Recently, Koshiba-Takeuchi et al. (2009) reported that fully septated ventricles that occur in mammals, birds, and crocodilians are resulted from regional expression of Tbx5 restricted to left ventricle precursors. Squamates and turtles initially express the Tbx5 homogenously in their ventricular chambers. However, in later stages, Tbx5 expression in the turtle (but not squamate) heart is gradually restricted to a distinct left ventricle, forming a left–right gradient. This suggests that turtles hold an intermediate position during evolution from partially to fully septated ventricles. Only Hypothesis 3 about the position of turtles, which gains decisive support from this study, can most parsimoniously explain the finding of reptilian heart development mechanism.

How Much Data Do We Need to Resolve the Phylogenetic Position of Turtles?

In figure 5, it is clearly shown that increasing the amount of data has a profound effect on branch support values at two nodes that determine the position of turtles. Although the means of the support values for given nodes go up as the amount of data increased, both the standard deviations and 95% confidence intervals of the means decrease (fig. 5). This trend visually shows that the random error caused by limited sample size (number of bases) and/or mismatches between gene trees and species trees diminishes when more and



Nodes	MtDNA (1, 289 base pairs)		22 nuclear genes (19, 848 base pairs)		All data (21, 137 base pairs)	
	ML BP	BI PP	ML BP	BI PP	ML BP	BI PP
A	93	1.00	97	1.00	100	1.00
B	92	1.00	100	1.00	100	1.00
C	51	0.68	99	1.00	99	1.00
D	—	—	100	1.00	100	1.00
E	49	0.55	94	1.00	95	1.00
F	—	—	95	1.00	86	1.00
G	—	—	89	1.00	87	1.00
H	—	—	92	1.00	94	1.00
I	82	1.00	100	1.00	100	1.00
J	80	1.00	100	1.00	100	1.00
K	—	—	99	1.00	99	1.00
L	—	—	100	1.00	100	1.00

Fig. 3. Higher level phylogenetic relationships of tetrapods inferred from analyses of 1 mtDNA and 22 nuclear genes. Partitioned ML and BI were conducted for three DNA data sets (mtDNA, 22 nuclear genes, and 1 mtDNA + 22 nuclear genes), respectively. Ray-finned fishes are used as outgroup. Branches with letters have branch support values given below the tree for ML bootstrapping (ML BP) and Bayesian posterior probabilities (BI PP). Hyphens indicate nodes that are not present (or ML BP $<$ 50%) in the corresponding analyses. Branches with ML bootstrap support $\geq$95% and Bayesian posterior probability = 1.0 in all the three data sets are indicated as filled squares. Branch lengths were from the partitioned ML analysis on all data combining 1 mtDNA and 22 nuclear genes.

more markers are used, which once again indicates that large-scale and multigene analyses are indispensable to resolve difficult nodes.

According to figure 5*A*,*C*, when the sequence length reaches 4 kb, a clade comprising turtles, birds, and crocodilians can be recovered with strong branch support values (aLRT

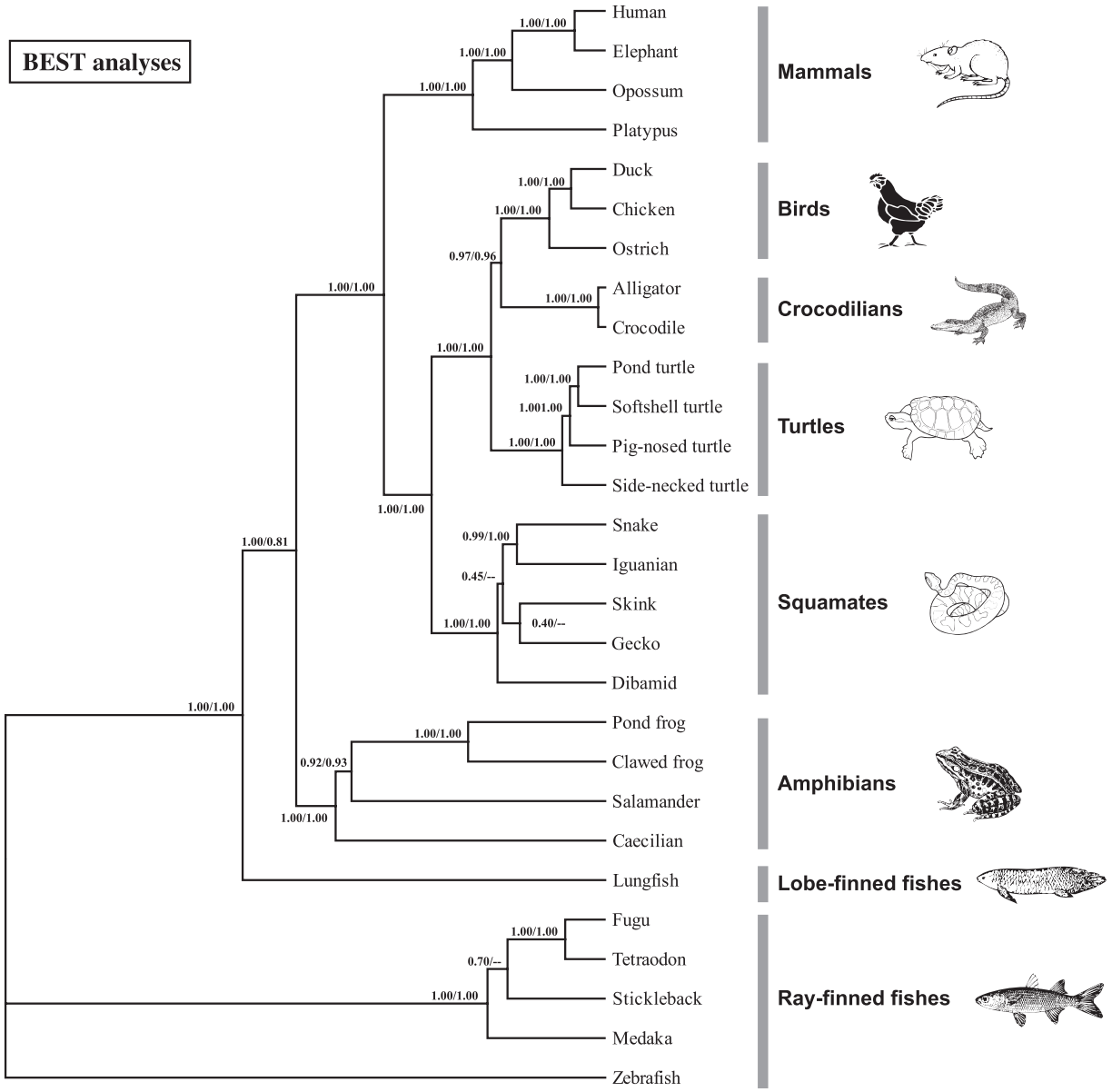


FIG. 4. Species tree estimation of tetrapods based on the Data set II (22 nuclear genes) and Data set III (1 mtDNA + 22 nuclear genes) using the method of BEST. Leftmost numbers along branches represent PP from the Data set II and rightmost numbers represent PP from the Data set III. Hyphens indicate nodes that are not present (or ML BP < 50%) in the corresponding analyses.

support = 1.00; bootstrap = 100). In fact, according to our statistics, when the sequence length is 1 kb, ~2% of subsamples support Hypothesis 1 and ~2.8% of subsamples support Hypothesis 2; when the sequence length reaches 3 kb, the probability drops to zero (results not shown). These results suggest that the minimum amount of data needed to determine the affinity between turtles and archosaurians is only 4 kb, explaining why most molecular studies repeatedly rejected Hypothesis 1 and Hypothesis 2 (fig. 1). However, if we used small data (<2 kb), it is still possible to favor Hypothesis 1 or Hypothesis 2 due to stochastic errors. As a piece of evidence, Becker et al. (2011) recently used the full-length cDNA of POMC (~1.2 kb) to address the phylogenetic position of turtles and pointed out that Hypothesis 2 is the preferred inference, which is in line with our prediction.

The difference between Hypothesis 3 and Hypothesis 4 is whether the archosaurians (birds and crocodilians) are monophyletic. As shown in our jackknife analyses (fig. 5*B*,*D*), increasing sequence length yields a progressive increasement in branch support values for a bird–crocodilian clade therefore denies turtles as the sister group of crocodilians proposed by some studies (e.g., Hedges and Poling 1999; Cao et al. 2000; Shedlock et al. 2007). When the amount of DNA data sampled is 13 kb, the mean of aLRT support values for a bird–crocodilian clade inferred from 200 jackknife subsamples reaches 0.95, a statistically significant threshold (fig. 5*B*). Similarly, approximate 14 kb DNA data would result in a robust affinity between birds and crocodilians with bootstrap support values at 90% (fig. 5*D*). All the aforementioned results suggest that the minimal amount of data needed to resolve the position of turtles

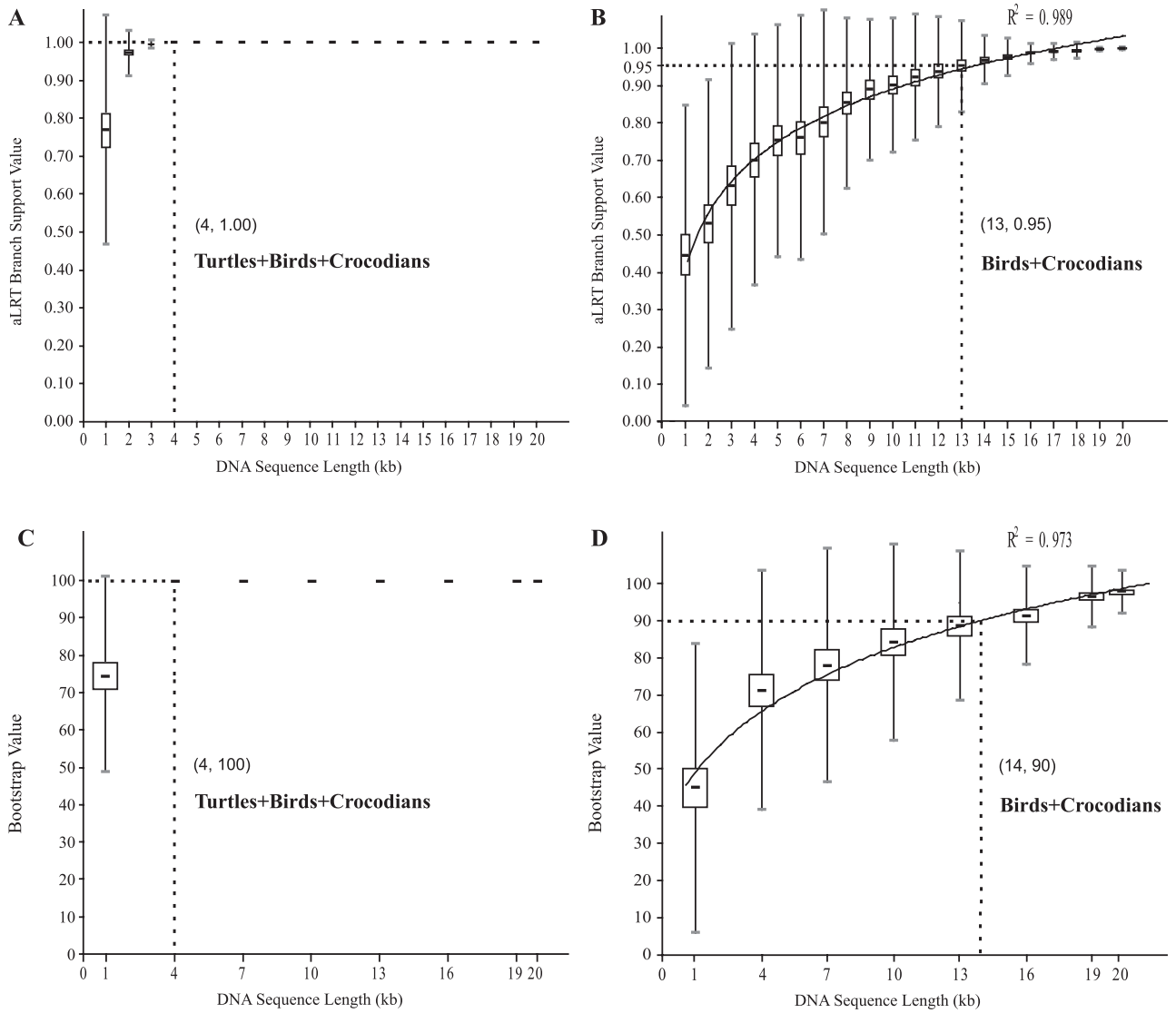


FIG. 5. The effect of increasing sequence length on reconstructing phylogenetic position of turtles. Each data point (indicated by bold horizontal lines) represents the mean of support values estimated from 200 replicate data sets. Error bars show standard deviation, and rectangle boxes represent 95% confidence intervals. The statistical plots show that the minimum data needed to robustly recover a turtle–bird–crocodile clade is 4 kb (both aLRT and bootstrap supports equal 1.0). The minimum data needed to distinguish between (turtles, (birds, crocodilians)) and (birds, (turtles, crocodilians)) is 13 kb under aLRT analyses (cutoff values = 0.95) and 14 kb under ML bootstrap (cutoff values = 90%), respectively.

is about 13–14 kb. Taking into consideration that the average size of our markers is about 900 bp, the minimum number of independent markers to determine the position of turtles should be around 14–16.

The Timetree of Tetrapod Evolution

In this study, we generated a large and comprehensive multilocus data set combining 22 nuclear genes and one mitochondrial fragment (21,137 bp in total) for a comprehensive taxon sampling with the key major lineages of living tetrapods, particularly covering the basal splits. The commonly used relaxed clock method (MultiDivTime; Thorne and Kishino 2002) was used to infer the new timetree of tetrapods, calibrated by multiple recent, reasonably accurate, and conservative calibration points (see Section of Molecular dating). To test the effect of our calibration choices, we estimated times in MultiDivTime with or without the sequence data (table 4). By comparing the differences of the means and 95% confidence intervals between the two analyses (with/without data) for each node, we can find out how much contribution comes from the data to the time estimate for each node.

In general, our new time estimates for many nodes are similar to the average estimates of previous dating studies summarized by the book of *The Timetree of Life* (Hedges and Kumar 2009) (see table 4). The major incongruence occurred within the mammalian part (table 4), largely because we used a maximum constraint (<191 Ma) at the root of mammals. To test the effect of this maximum bound, we redid the dating analysis without this constraint

Table 4. Detailed Results of Bayesian Molecular Dating Using MultiDivTime.

Nodes	With Maximal Constraint on Origin of Living Mammals		Without Maximal Constraint on Origin of Living Mammals		Average Estimates in Timetree of Life (2009)
	No Sequence Data	With Sequence Data	No Sequence Data	With Sequence Data	
1: Lungfish–tetrapod split[a]	413 (408–419)	414 (409–419)	413 (408–419)	414 (409–419)	430
2: Amphibia–Amniote split[a]	372 (332–412)	344 (336–355)	372 (332–412)	341 (332–351)	361
3: Bird–mammal split[a]	322 (313–330)	317 (312–328)	322 (313–330)	319 (312–329)	324.5
4: Origin of living amphibians	279 (109–390)	320 (308–332)	278 (107–389)	315 (303–328)	294
5: Anura–Caudata split	185 (36–339)	295 (279–310)	186 (34–344)	290 (273–305)	264
6: Pond frog–clawed frog split	93 (3–261)	173 (130–200)	91 (3–264)	165 (123–193)	229
7: Bird–lizard split[a]	302 (265–325)	285 (276–295)	303 (265–326)	294 (284–306)	274.9
8: Turtle–Archosauria split	283 (246–317)	257 (249–270)	283 (246–317)	264 (252–280)	230.7
9: Origin of living squamates	243 (121–312)	205 (180–228)	243 (119–343)	209 (182–237)	209.4
10: Gecko–skink split	183 (61–288)	192 (166–216)	183 (58–288)	196 (167–224)	197.9
11: Skink–iguanian split	122 (20–248)	181 (154–205)	184 (155–213)	184 (155–213)	188.3
12: Iguanian–snake split	62 (2–185)	163 (134–187)	121 (20–247)	164 (134–194)	166.4
13: Bird–alligator split[a]	261 (236–302)	241 (235–255)	261 (236–302)	245 (235–263)	219.2
14: Origin of living turtles[a]	244 (196–298)	211 (195–231)	244 (196–299)	211 (195–236)	207
15: Testudinoidea–Trionychoidea split	163 (38–265)	181 (158–203)	163 (38–266)	179 (155–206)	175
16: Softshell turtle–pig–nose turtle split	82 (3–212)	145 (117–170)	82 (3–213)	142 (115–170)	155
17: Alligator–crocodile split[a]	163 (71–267)	110 (71–157)	164 (71–265)	101 (69–150)	102 .6
18: Origin of living birds[a]	182 (76–273)	122 (91–156)	183 (77–271)	117 (89–152)	119
19: Duck–chicken split	91 (4–224)	69 (45–96)	92 (4–223)	65 (44–92)	106.9
20: Origin of living mammals[a]	179 (164–191)	186 (172–191)	268 (178–323)	253 (225–276)	220.2
21: Marsupials–placentals split[a]	155 (126–184)	158 (144–169)	209 (130–300)	228 (193–255)	176.1
22: Elephant–human split[a]	112 (73–165)	81 (71–99)	140 (74–251)	121 (82–153)	104.7

Note.—Serial numbers for nodes are corresponding to the node numbers in figure 6. Numbers and numbers in parentheses indicate divergence time mean and 95% credibility intervals (Ma), respectively.

[a] With calibration constraints.

and found that the changes of time estimates are very slight except within the mammalian part (table 4). Because mammals are not of great concern in our study, we regard the time estimates calculated with all constraints as our preferred dating results as illustrated in figure 6. In particular, the means and 95% confidence intervals of time estimates for the amphibia–amniota split, the lizard–bird split, and the bird–crocodile split were 344 (336–355), 285 (276–295), and 241 (235–255) Ma, respectively, which are in close agreement with fossil-based estimates (330–350, 260–300, and 235–250 Ma, respectively; Benton and Donoghue 2007). Consistency between our estimates and the fossil recommendations has enhanced the credibility of our timetree for tetrapod evolution.

Many molecular clock studies have recently been done to address the question of the origin of living amphibians and dated the caecilian–frog split during either the Late Devonian (~367 Ma; San Mauro et al. 2005; Roelants et al. 2007), the Carboniferous (~337 Ma; Zhang et al. 2005), the Early Permian (~294 Ma; Hugall et al. 2007; Zhang and Wake 2009), or the Late Permian (~267 Ma; Marjanović and Laurin 2007). However, none of the above time estimates are based on more than five independent molecular markers. Recently, San Mauro (2010) assembled a large multilocus data set combining mitogenome and eight nuclear genes (9,133 bp in total) and suggested that extant amphibians originated in the Late Carboniferous, around 315 Ma, and the frog–salamander split occurred in the Early Permian, around 290 Ma. Notably, our time estimates for these two nodes (320 and 295 Ma, respectively), based on our large multilocus data set, are very similar to his results. Using large multilocus data sets is a promising direction for future efforts to settle the debate of the origin of living amphibians.

Vidal and Hedges (2005) used nine nuclear genes to estimate divergence times within squamates and argued that the basal split of living squamates (dibamidae-other) occurred in the Triassic about 240 Ma. Two other studies dated the basal split of living squamates (gekkota-other) in the Jurassic about 180 Ma (Wiens et al. 2006) or 190 Ma (Hugall et al. 2007). However, the phylogenies that the three studies used are somewhat different therefore their results may not be comparable. Our time estimate for the origin of living squamates (205 Ma; fig. 6) is close to the average of the previous results (Hedges and Vidal 2009). Currently, dating data about this node are still very limited and the origin of living squamates will continue to be an ongoing debate.

Our time estimate for the split between the side-necked turtles (Pleurodira) and the hidden-neck turtles (Cryptodira), or to say the origin of extant turtles, is 211 Ma, very similar to another recent molecular result (207 Ma; Hugall et al. 2007) and is in agreement with the fossil record (*Proterochersis*, 210 Ma; Gaffney 1986). In addition, the Testudinoidea–Trionychoidea split took place around 181 Ma in our timetree (fig. 6), which is also very similar to the fossil-based estimate (175 Ma; Near et al. 2005). The consistency shows that the molecular and fossil data tend to reach agreement in dating the evolutionary history of living turtles.

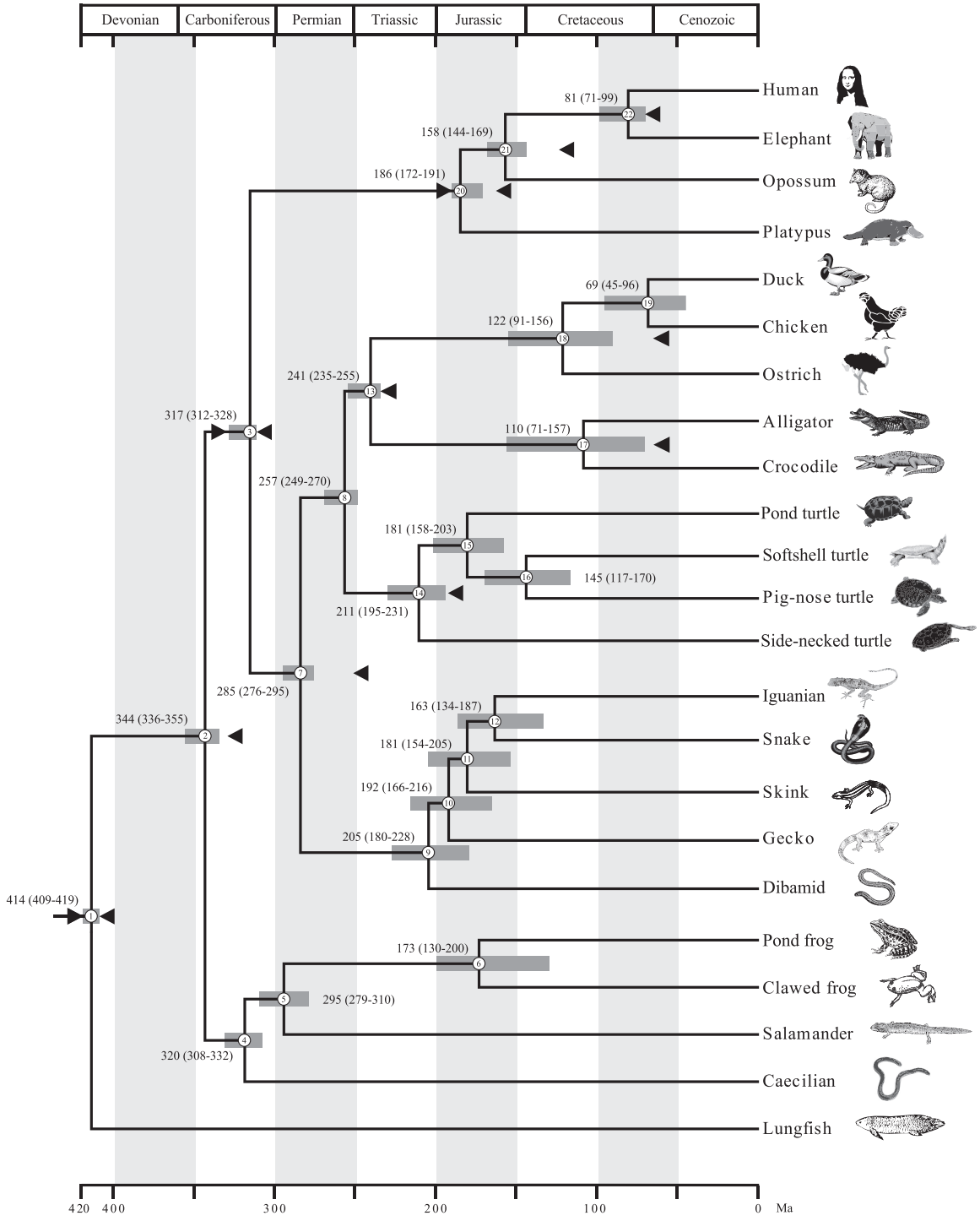


Fig. 6. Timetree of tetrapods inferred from the relaxed molecular clock method implemented in MultiDivTime. A total of 14 time constraints (indicated by arrowheads) are used to calibrate the relaxed clock (see Materials and Methods). Numbers beside the nodes are the mean estimated divergence time (in Ma), and numbers in parentheses represent 95% credibility intervals (also represented by gray horizontal bars). More time estimates can be found in table 4 for nodes with numbered circles above them.

The basal divergence of living crocodilians occurred between alligators and crocodiles, and our time estimate for this split is 110 Ma (fig. 6), slightly older than the average estimate (102.6 Ma) summarized from six studies by Brochu (2009). The origin of extant birds is estimated around 122 Ma (fig. 6), once again close to the current average estimate (119 Ma; van Tuinen 2009). With regard to the origin time of extant mammals, when we removed the maximal constraint for the root of living mammals in our timetree, the time estimate is around 253 Ma (table 4), much older than the current average estimate (220 Ma; Madsen 2009). The cause of the inconsistency is not clear yet and deserves further exploration.

Supplementary Material

Supplementary table S1 and figure S1 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Ke Jiang for kindly providing the *Dibamus bourreti specimen* and Holland Barbara for the great help in English improvement. This work was supported by National Natural Science Foundation of China (No. 30900136), Natural Science Foundation of Guangdong Province (9151027501000080), and Grant of Excellent PhD thesis of Guangdong Province (sybzzxm201012).

References

Becker RE, Valverde RA, Crother BI. 2011. Proopiomelanocortin (POMC) and testing the phylogenetic position of turtles (Testudines). *J Zool Syst Evol Res*. 49:148–159.

Benton MJ. 1993. Reptilia. In: Benton MJ, editor. The fossil record 2. London: Chapman & Hall. p. 681–715.

Benton MJ, Donoghue PC. 2007. Paleontological evidence to date the tree of life. *Mol Biol Evol*. 24:26–53.

Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21(2):163–193.

Brochu CA. 1999. Phylogeny, systematics, and historical biogeography of Alligatoroidea. *Soc Vertebr Paleontol Mem*. 6:9–100.

Brochu CA. 2009. Crocodylians (Crocodila). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 402–406.

Cao Y, Sorenson MD, Kumazawa Y, Mindell DP, Hasegawa M. 2000. Phylogenetic position of turtles among amniotes: evidence from mitochondrial and nuclear genes. *Gene* 259:139–148.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol*. 17:540–552.

Clarke JA, Tambussi CP, Noriega JI, Erickson GM, Ketcham RA. 2005. Definitive fossil evidence for the extant avian radiation in the Cretaceous. *Nature* 433:305–308.

deBraga M, Rieppel O. 1997. Reptile phylogeny and the interrelationships of turtles. *Zool J Linn Soc*. 120:281–354.

Delsuc F, Brinkmann H, Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genet*. 6:361–375.

Evans SE, King ME. 1993. A new specimens of *Protorosaurus* (Reptilia: Diapsida) from the Marl Slate (Late Permian) of Britain. *Proc Yorkshire Geol Soc*. 49:229–234.

Felsenstein J. 2005. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington.

Gaffney ES. 1986. Triassic and early Jurassic turtles. In: Padian K, editor. The beginnings of the age of dinosaurs. Cambridge: Cambridge University Press. p. 183–187.

Gauthier J, Kluge AG, Rowe T. 1988. Amniote phylogeny and the importance of fossils. *Cladistics* 4:105–209.

Gheerbrandt E, Domning DP, Tassy P. 2005. Paenungulata (Sirenia, Proboscidea, Hyracoidea, and relatives). In: Rose KD, Archibald JD, editors. The rise of placental mammals: origins and relationships of the major extant clades. Baltimore (MD): Johns Hopkins University Press. p. 84–105.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol*. 52:696–704.

Hedges SB, Kumar S. 2009. The timetree of life. New York: Oxford University Press.

Hedges SB, Poling LL. 1999. A molecular phylogeny of reptiles. *Science* 283:998–1001.

Hedges SB, Vidal N. 2009. Lizards, snakes, and amphisbaenians (Squamata). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 383–389.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* 17:754–755.

Hugall AF, Foster R, Lee MSY. 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. *Syst Biol*. 56:543–563.

Iwabe N, Hara Y, Kumazawa Y, Shibamoto K, Saito Y, Miyata T, Katoh K. 2004. Sister group relationship of turtles to the bird-crocodilian clade revealed by nuclear DNA-coded proteins. *Mol Biol Evol*. 22(4):810–813.

Ji Q, Luo XZ, Yuan CX, Wible JR, Zhang JP, Georgi JA. 2002. The earliest known eutherian mammal. *Nature* 416:816–822.

Kirsch J, Mayer G. 1998. The platypus is not a rodent: DNA hybridization, amniote phylogeny and the palimpsest theory. *Philos Trans R Soc Lond Ser B*. 353:1221–1237.

Koshiba-Takeuchi K, Mori AD, Kaynak BL, et al. (16 co-authors). 2009. Reptilian heart development and the molecular basis of cardiac chamber evolution. *Nature* 461:95–98.

Kumazawa Y, Nishida M. 1999. Complete mitochondrial DNA sequences of the green turtle and blue-tailed mole skink: statistical evidence for archosaurian affinity of turtles. *Mol Biol Evol*. 16:784–792.

Leaché AD, Rannala B. 2010. The accuracy of species tree estimation under simulation: a comparison of methods. *Syst Biol*. 60(2):126–137.

Lee MSY. 1997. Pareiasaur phylogeny and the origin of turtles. *Zool J Linn Soc*. 120:197–280.

Li C, Ortí G, Zhang G, Lu G. 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. *BMC Evol Biol*. 7:44.

Liu L. 2008. BEST: Bayesian estimation of species trees under the coalescent model. *Bioinformatics* 24:2542–2543.

Lyson TR, Bever GS, Bhullar BA, Joyce WG, Gauthier JA. 2010. Transitional fossils and the origin of turtles. *Biol Lett*. 6:830–833.

Madsen O. 2009. Mammals (Mammalia). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 459–461.

Marjanović D, Laurin M. 2007. Fossils, molecules, divergence times, and the origin of lissamphibians. *Syst Biol*. 56:369–388.

Meyer A, Zardoya R. 2003. Recent advances in the (molecular) phylogeny of vertebrates. *Annu Rev Ecol Evol Syst*. 34:311–338.

Müller J, Reisz RR. 2005. Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. *Bioessays* 10:1069–1075.

Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, O'Brien SJ. 2001. Molecular phylogenetics and the origins of placental mammals. *Nature* 409:614–618.

Near TJ, Meylan PA, Shaffer HB. 2005. Assessing concordance of fossil calibration points in molecular clock studies: an example using turtles. *Am Nat*. 165:137–146.

Nye TM, Lio WP, Gilks WR. 2006. A novel algorithm and web-based tool for comparing two alternative phylogenetic trees. *Bioinformatics* 22:117–119.

Nylander JAA. 2004. MrModeltest v2. Program distributed by the author. Uppsala, Sweden: Evolutionary Biology Centre, Uppsala University.

Pamilo P, Nei M. 1988. Relationships between gene trees and species trees. *Mol Biol Evol*. 5:568–583.

Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu Rev Ecol Evol Syst*. 36:541–562.

Phillips MJ, Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. *Mol Phylogenet Evol*. 28:171–185.

Rest JS, Ast JC, Austin CC, Waddell PJ, Tibbetts EA, Hay JM, Mindell DP. 2003. Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Mol Phylogenet Evol*. 29:289–297.

Roelants K, Gower DJ, Wilkinson M, Loader SP, Biju SD, Guillaume K, Moriau L, Bossuyt F. 2007. Global patterns of diversification in the history of modern amphibians. *Proc Natl Acad Sci U S A*. 104:887–892.

Rokas A, Williams BL, King N, Carroll SB. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425(6960):798–804.

Ruta M, Coates MI, Quicke DLJ. 2003. Early tetrapod relationships revisited. *Biol Rev*. 78:251–345.

San Mauro D. 2010. A multilocus timescale for the origin of extant amphibians. *Mol Phylogenet Evol*. 56:554–561.

San Mauro D, Vences M, Alcobendas M, Zardoya R, Meyer A. 2005. Initial diversification of living amphibians predated the breakup of Pangaea. *Am Nat*. 165:590–599.

Shedlock AM, Botka CW, Zhao S, Shetty J, Zhang T, Liu JS, Deschavanne PJ, Edwards SV. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proc Natl Acad Sci U S A*. 104:2767–2772.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*. 24:4876–4882.

Thorne JL, Kishino H. 2002. Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol*. 51:689–702.

Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW. 2008. Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. *Mol Phylogenet Evol*. 47:129–142.

Townsend TM, Larson A, Louis E, Macey RJ. 2004. Molecular phylogenetics of Squamata: the position of snakes, amphisbaenians, and dibamids, and the root of the squamate tree. *Syst Biol*. 53:735–757.

van Tuinen M. 2009. Birds (Aves). In: Hedges SB, Kumar S, editors. The timetree of life. New York: Oxford University Press. p. 409–411.

Vidal N, Hedges SB. 2004. Molecular evidence for a terrestrial origin of snakes. *Proc R Soc Lond B Suppl*. 271:S226–S229.

Vidal N, Hedges SB. 2005. The phylogeny of squamate reptiles (lizards, snakes, and amphisbaenians) inferred from nine nuclear protein-coding genes. *C R Biol*. 328:1000–1008.

Werneburg I, Sánchez-Villagra MR. 2009. Timing of organogenesis support basal position of turtles in the amniote tree of life. *BMC Evol Biol*. 9:82.

Wiens JJ, Brandley MC, Reeder TW. 2006. Why does a trait evolve multiple times within a clade? Repeated evolution of snake-like body form in squamate reptiles. *Evolution* 60:123–141.

Wiens JJ, Kuczynski CA, Townsend T, Reeder TW, Mulcahy DG, Sites JW. 2010. Combining phylogenomics and fossils in higher level squamate reptile phylogeny: molecular data change the placement of fossil taxa. *Syst Biol*. 59:674–688.

Wright TF, Schirtzinger EE, Matsumoto T, et al. (11 co-authors). 2008. A multilocus molecular phylogeny of the parrots (Psittaciformes): support for a Gondwanan Origin during the Cretaceous. *Mol Biol Evol*. 25:2141–2156.

Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13:555–556.

Zardoya R, Meyer A. 1998. Complete mitochondrial genome suggests diapsid affinities of turtles. *Proc Natl Acad Sci U S A*. 95:14226–14231.

Zhang P, Papenfuss TJ, Wake MH, Qu LH, Wake DB. 2008. Phylogeny and biogeography of the family Salamandridae (Amphibia: Caudata) inferred from complete mitochondrial genomes. *Mol Phylogenet Evol*. 49:586–597.

Zhang P, Wake DB. 2009. Higher-level salamander relationships and divergence dates inferred from complete mitochondrial genomes. *Mol Phylogenet Evol*. 53:492–508.

Zhang P, Zhou H, Chen YQ, Liu YF, Qu LH. 2005. Mitogenomic perspectives on the origin and phylogeny of living amphibians. *Syst Biol*. 54:391–400.