Supplementary Materials for

Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs

Richard E. Green,* Edward L. Braun, Joel Armstrong, Dent Earl, Ngan Nguyen, Glenn Hickey, Michael W. Vandewege, John A. St. John, Salvador Capella-Gutiérrez, Todd A. Castoe, Colin Kern, Matthew K. Fujita, Juan C. Opazo, Jerzy Jurka, Kenji K. Kojima, Juan Caballero, Robert M. Hubley, Arian F. Smit, Roy N. Platt, Christine A. Lavoie, Meganathan P. Ramakodi, John W. Finger Jr., Alexander Suh, Sally R. Isberg, Lee Miles, Amanda Y. Chong, Weerachai Jaratlerdsiri, Jaime Gongora, Christopher Moran, Andrés Iriarte, John McCormack, Shane C. Burgess, Scott V. Edwards, Eric Lyons, Christina Williams, Matthew Breen, Jason T. Howard, Cathy R. Gresham, Daniel G. Peterson, Jürgen Schmitz, David D. Pollock, David Haussler, Eric W. Triplett, Guojie Zhang, Naoki Irie, Erich D. Jarvis, Christopher A. Brochu, Carl J. Schmidt, Fiona M. McCarthy, Brant C. Faircloth, Federico G. Hoffmann, Travis C. Glenn, Toni Gabaldón, Benedict Paten, David A. Ray*

*Corresponding author. E-mail: ed@soe.ucsc.edu (R.E.G.); david.a.ray@ttu.edu (D.A.R.)

This PDF file includes:

Materials and Methods

Figs. S1 to S29

Tables S1 to S24

References

Contents:

Materials and Methods

Materials and Methods

SM 1 – Genome sequencing and assembly

John A. St. John and Richard E. Green*

* To whom correspondence should be addressed (ed@soe.ucsc.edu)

Genome Sequencing. Genomic DNA was isolated using blood from four individuals. Details for each are in the sections below. In all cases, next-generation sequencing was performed using an Illumina GAIIx for early libraries and a HiSeq 2000 for libraries constructed later in the project.

Genome assembly of *Alligator mississippiensis.* Genomic DNA was isolated from the blood of two wild caught individuals of unknown sex from the Kennedy Space Center, Florida, USA. From these DNA preps, we prepared libraries and sequencing data as described in Table S1.

The mate pair libraries were adapter trimmed using ea-utils to remove the non-genomic adapter sequence used to construct these libraries.

We generated an initial assembly using R41313 of Allpaths-LG (*53*) with default parameters on the input data described above. We ran the bundled allpaths NCBI submission prep tool to remove short contigs from the assembly. To the assembly, we manually merged the 21 previously assembled alligator BAC sequences available on GenBank (AC148578.2, AC148923.3, AC148964.2, AC149025.3, AC149026.2, AC149027.1, AC149028.2, AC149029.2, AC154087.3, AC154088.2, AC154169.2, AC154170.2, AC154945.2, AC155799.2, AC155800.2, AC155801.3, AC155802.2, AC161341.3, AC162159.2, AC164519.3, AC165215.2). We used nucmer version 3, from the mummer package, to generate alignments of these assembled BAC sequences against our assembly scaffolds for visual inspection to generate consensus sequences. Of the 21 BAC sequences, 8 were contained within a single scaffold in our original assembly. 12 BAC sequences facilitated scaffold merges. The remaining BAC sequence was omitted because its contigs were marked as being unordered.

After using the assembled BAC sequences, we mapped the 1,309 Sanger BAC-end read pairs also available on GenBank (*54*). We used BLAT to map each read individually to the alligator assembly. We filtered non-primary and non-unique hits of each BAC-end sequence using pslCdnaFilter (http://www.soe.ucsc.edu/~kent). We collected reads whose pairs were both uniquely mapped in the alligator assembly. These comprised 1,195 (91%) of the input BAC read pairs. We converted these alignments into paired-end SAM format using customs scripts. Using this SAM file as input, we ran the ABYSS module, abuss-fixmate, to generate an insert size histogram for the pairs the pairs that mapped to a single scaffold. Finally, we used the BAC-end pairs that uniquely mapped to different scaffolds with SGA to further scaffold the genome. This operation resulted in 465 scaffold joins supported by the BAC-end sequence pairs.

Next, we performed gene-assisted genome scaffolding. To do this, we found ORFs from all assembled transcripts and filtered only those with recognizable similarity to a protein sequence in Swiss-Prot. We then filtered these ORFs for redundancy and mapped these to the genome assembly. We used pslReps from the UCSC genme browser source to identify the highest quality alignment segments for each gene. We generated a distanceEst format file for use with the SGA scaffolder, labelling as UNKNOWN the length of gaps between segments since these come from introns of unknown length. This procedure joined 1,694 scaffolds.

We submitted this final scaffolded assembly to NCBI where it was further filtered by the NCBI quality-control procedures. It is assigned the accession number AKHW01000000.

Genome assembly of *Crocodylus porosus.* We collected blood from a single male *C. porosus*, Errol, a wild caught individual originally collected in the Northern Territory of Australia and now housed at the Fort Worth Zoo. DNA preps were used to generate the sequencing libraries described in Table S2.

These data were used as input to Allpaths-LG with default parameters for assembly. This assembly was filtered for mitochondria and vector contamination and short contigs (less than 1kb) upon submission to NCBI. Using the filtered assembly, we used Jaime Gongora's MHC BAC assemblies (*84*) to further scaffold the assembly similar to the procedure described for the alligator BACs. These 360 MHC region BAC assemblies resulted in 13 new scaffold joins. The remaining BACs all existed within single scaffolds within the crocodile genome assembly.

We performed gene-based scaffolding using assembled and validated ORF sequences from both crocodile and alligator RNAseq data as described above for alligator. This procedure resulted in 137 scaffold joins from the crocodile ORFs and 1,578 scaffold joins from the alligator ORF sequences. This assembly was submitted to NCBI and assigned accession number JRXG00000000.

Genome assembly of *Gavialis gangeticus.* Genomic DNA was prepared from a blood sample of a single female gharial from the Ft. Worth zoo. This DNA prep was used to generate the sequencing libraries described in Table S3. The gharial genome was assembled from trimmed paired-end and mate-pair reads using SOAPdenovo version 2.04.3 (*55*) and pregraph_sparse version 5.1. We used kmer size 39. Following assembly, we used GapCloser version 1.12. This step reduced the gapped percentage of the genome assembly length from 10.9% to 1.7%.

From this initial assembly, we removed contigs less than 200bp and scaffolds less than 1kb. Noting that this assembly contained 120,245 single N positions after the GapCloser step, we manually inspected a random subset of these. We noted that for many, no reads spanned these single N positions. Therefore, we expanded all instances of single N positions into gaps of unknown length.

Following vector and mitochondrial contamination removal via the NCBI screening procedure, this assembly was submitted to NCBI and assigned accession number JRWT00000000.

Final genome assembly statistics are available in Table S4

SM 2 – Transcriptome sequencing and sequence annotation

Fiona M. McCarthy*, Cathy Gresham, Colin Kern, Toni Gabaldón, Carl J. Schmidt, Todd A. Castoe, Travis C. Glenn, Jerzy Jurka, Kenji Kojima, Arian Smit, Juan Caballero and David A. Ray*

* To whom correspondence should be addressed (david.a.ray@ttu.edu)

Alligator mississippiensis individuals were originally collected from Rockefeller Wildlife Refuge, Grand Chenier, Louisiana. Alligators (~75 cm) were transported to the Savannah River Ecology Laboratory and kept on a diet of gator chow for ~ 2 months. Immediately prior to sacrifice, animals were kept in the dark for ~1 hr and then a recording of various alligator vocalizations were played. Animals were injected with propofol and killed by vertebral dislocation. Tissues were harvested immediately and frozen whole or in RNAlater by placing on dry ice. *G. gangeticus* blood was collected by Kent Vliet of the St. Augustine Alligator Farm. Total RNA was extracted from multiple *A. mississippiensis* and *C. porosus* tissues (Table S5) and *G. gangeticus* blood. For all tissues, RNA was extracted using standard Tri-reagent Trizol (Life Technologies, Grand Island, NY, USA) or blood-specific tri-reagent for the whole blood RNA extractions). Total RNA quality was determined using a Bioanalyzer 2100 (Agilent, Santa Clara, CA, USA). The Illumina TruSeq RNA Sample preparation kit (Catalog # RS-930-20 01; Illumina, San Diego, CA, USA) was used to create cDNA libraries from whole RNA, using the low-throughput version of the protocol, except *C. porosus* where MINT kits (Evrogen, Moscow, Russia) were used for cDNA synthesis. Custom barcoded adapters were used in the place of the kit-provided adapters. Most cDNA libraries had a mode size of 300-350bp, including the adapters. After pooling barcoded samples approximately at equimolar ratios, pooled libraries were further size-selected using the pippin prep to remove small fragments < 200 bp. Libraries were sequenced in December, 2011 on an Illumina GAIIx genome analyzer the USDA facility in Beltsville, MD. Reads were assembled using Trinity (*85*) under the default parameters for each tissue independently and for a concatenated data consisting of all RNASeq data. In total, we generated ~11Gb of high quality sequence for all tissues (Table S5).

Gene annotation. Gene predictions for the three crocodilian species were made using Augustus (version 2.5.5) (*56*). Transcriptome data from *A. mississippiensis* was aligned to the genome draft of American alligator using Tophat version 2.0.6 and Bowtie version 2.0.5. Augustus used these alignments to improve its gene predictions. Because the alligator assembly was a higher quality than the crocodile and gharial assemblies, the proteins predicted for alligator were aligned to the other crocodilian assemblies using Genblastg version 1.38, and those alignments were used by Augustus to improve the gene predictions for those species.

Functional annotation. Functional annotation was done by assigning gene nomenclature, Gene Ontology (GO) and pathway information. Gene names are assigned based upon orthology or homology to species with a gene nomenclature project by transferring names to the crocodilian genes. Following this strategy, 15,386, 10,366 and 10,489 gene symbols and names were assigned to American alligator, crocodile and gharial, respectively. This nomenclature was manually reviewed to ensure it was species appropriate (e.g. human genes referencing "HLA" were manually edited to "MHC"). GO was assigned to predicted proteins based upon a combinatorial approach. Predicted proteins were mapped InterPro identifiers and GO with a GO evidence code of "IEA" or Inferred from Electronic Annotation. We also transferred GO from orthologous vertebrate genes using both orthology (assigned the GO evidence code "ISO" or Inferred from Sequence Orthology) and reciprocal Blast (assigned the GO evidence code "ISA" or Inferred from Sequence Alignment). GO annotations from these three sources were merged, duplicates removed, GO terms manually reviewed to ensure it was species appropriate (e.g. annotations terms such as GO:0000803 sex chromosome and GO:0033333 fin development removed) and the number of annotations for each species is shown in Table S6. Pathway information was assigned based upon reciprocal Blast. Annotated genes, gene products and genome assemblies are available at CrocBase (http://crocgenome.hpc.msstate.edu/crocbase/gene.php) and genomes are also available via the Comparative Genomics (CoGe) browser (http://genomevolution.org/CoGe/).

TE discovery and annotation. Repeat discovery was performed semi-independently in the laboratories of D. Ray, J. Jurka and A. Smit. Two primary methods were employed. In the Ray laboratory, each genome draft independently investigated RepeatModeler (*59*). RepeatModeler utilizes RECON (a graphical, multiple alignment approach (*86*)) and RepeatScout (a string expansion approach (*61*)) to identify possible TE consensus sequences in a genome. Output from RepeatModeler consists of putative TE consensus sequences. Using the RepeatModeler output from *A.*

mississippiensis, we went on to confirm or extend the consensus sequences for each repeat. This was accomplished first by querying the entire *A. mississippiensis* draft using BLAST v2.2.23 (*87*). Up to fifty of the top hits for each putative consensus were extracted along with up to 1,000 bases of flanking sequence. We aligned the extracted sequences with their respective RepeatModeler consensus using MUSCLE 4.0 and generated 50% majority rule consensus sequences. Consensus sequences were considered 'complete' when single copy sequence could be identified at the 5' and 3' ends in each component sequence. If this condition was not met, the process was repeated until single copy DNA sequence was identifiable at both ends. This was followed by RepeatModeler analysis of *C. porosus*, after which, the initial RepeatModeler output from *C. porosus* was compared to the repeat complement of *A. mississippiensis* to identify elements predicted from both genomes. Any unique repeats were used to query the *C. porosus* assembly and the Blast/Extract/Align process was repeated until consensus sequences were produced. Finally, the process was repeated for *G. gangeticus* elements. The libraries from each taxon were combined and the resulting library was submitted to CENSOR (*88*), BLASTN and BLASTX to classify the elements. Manual examination of consensus sequences were also used to classify the putative TEs. Hallmarks of TE superfamilies used for classification included, target site duplications, terminal inverted repeats, putative ORFs, etc.

In the Jurka laboratory, *de novo* repeat identification independently applied to each assembly. In the first step, Repeatmasker (*89*) was applied to each draft to mask regions with high similarity to known repeat families. The remaining data were analysed using PILER (*60*), RepeatScout and LTRharvest (*62*). Tandem repeats and low complexity regions were removed and the output from each analysis was clustered, producing a merged list of putative *de novo* repeat families. UCLUST was used to group potential families at 95% identity thresholds and consensus regions were compared using TBLASTX against the non-redundant (NR) NCBI protein sequence database to remove known peptides. The remaining sequences were then classified by REPCLASS (*90*), followed by manual annotations.

The Smit laboratory focused on building consensus sequences for older elements and added curation of all families identified using the methods above. Curation was as described in for the platypus (*22*) and zebra finch (*21*) genomes. Libraries from all taxa and all methods were combined and duplicate consensus sequences were removed. This resulted in a custom library of elements, which served as our library for subsequent analyses. The library of TEs was passed through a locally implemented version of RepeatMasker to estimate the TE content of each genome (Table S7). All new consensus sequences were uploaded to RepBase (*88*).

SM 3 – Identification and analysis of ultraconserved elements (UCEs)

Brant C. Faircloth*, Edward L. Braun, John E. McCormack, and Travis C. Glenn

* To whom correspondence should be addressed (brant@faircloth-lab.org)

UCE Identification and Phylogenetic Analyses. To create a large set of ultraconserved element loci (UCEs), we combined two sets of ultraconserved elements (*25, 63*) and kept unique and non-duplicate loci in the set (n=8,047 UCE loci). Using the positions of these loci in the chicken genome (galGal3), we designed capture probes (n=12,237; Supplemental File 3.1 - UCEFi1) for each locus to use for *in silico* identification of orthologous UCEs in other tetrapods. We aligned each capture probe to genome-enabled tetrapods (Table S8, Supplemental File 3.1 - UCEFi2) using a wrapper around LASTZ from the PHYLUCE package (v1.0; https://github.com/faircloth-lab/phyluce). The wrapper implements LASTZ (*91*) searches in parallel and ensures matches have >93% sequence identity over ≥83% of the probe length (minimum match length ≥ 100 nucleotides). We searched for probes designed from UCE loci rather than entire UCE loci to standardize the specificity of search parameters across probes (which are identical lengths) versus UCE loci (which are different lengths) and also to increase the number of loci we detected, because conserved loci are sometimes shorter in other taxa than they are in chicken.

Following identification of putative UCE loci in each genome (Supplemental File 3.3 - UCEFi3, Supplemental File 3.4 - UCEFi4), we sliced the match location of all probes ± 2000 bp from each genome assembly (Supplemental File 3.5 - UCEFi5). Where we recovered slices derived from multiple probes targeting the same locus, we re-assembled sequences back into full UCE loci (Supplemental File 3.6 - UCEFi6) using computer code available in the PHYLUCE package and MAFFT (*92*), and we trimmed all slices to approximately the length of the UCE locus ± 1000 bp. Following locus re-assembly and trimming, we ran contigs through an additional screening process to remove apparent duplicate loci. Based on this filtered set of matches, we created a relational database (Supplemental File 3.6 - UCEFi6) containing a Table of UCE loci identified in each taxon, and we used this Table to identify the set of all loci found in all taxa (a complete matrix) from two different taxon samples (Table S8, Supplemental File 3.7 - UCEFi7). We named these taxon-set-1 (n=633 UCE loci) and taxon-set-2 (n=1,010 UCE loci). Taxon-set-1 includes African clawed frog (*Xenopus tropicalis*), and as a result, contains fewer orthologous loci in a complete matrix.

Using the complete data matrix for each taxon set, we aligned FASTA data corresponding to each re-assembled UCE locus for each taxon by running a program in the PHYLUCE package (seqcap_align_2.py) that uses MAFFT to align sequences and trims resulting alignments for misaligned regions at the edges. Following alignment and trimming, we removed any loci containing ambiguous base calls (encoded by "N" or "X") present in the source genome assemblies ($n_{taxon-set-1}$=28; $n_{taxon-set-2}$=46). The remaining alignment data for taxon-set-1 (Supplemental File 3.8 - UCEFi8) contained 604 loci totaling 495,744 characters and 93,374 alignment patterns (mean locus length=820 bp; 95 CI = 47 bp). The remaining alignment data for taxon-set-2 (Supplemental File 3.9 - UCEFi9) contained 965 loci totaling 878,786 characters and 172,112 alignment patterns (mean locus length=911 bp; 95 CI = 40 bp). We concatenated all loci in each set, and we analyzed the resulting, concatenated alignments using RAxML 7.3.4 (GTRGAMMA model) (*64*), conducting 20 maximum likelihood (ML) tree searches and 500 bootstrap replicates for each data set. Using RAxML, we checked for bootstrap replicate convergence using the "autoMRE" function. Both data sets converged after 50 replicates, and we used RAxML to reconcile each best, ML tree with each set of 500 bootstrap replicates.

To evaluate the effects of partitioning on the topology and branch lengths we present in Fig. 1, we partitioned the alignment data for taxon-set-2 using PartitionFinder (*93*) and the strict hierarchical clustering algorithm (*94*); RAxML; model selection by Bayesian Information Criterion; and equal weighting for overall rates, base frequencies, model parameters, and the alpha parameter. Following partitioning, we input the concatenated taxon-set-2 alignment from above to RAxML, along with the optimal partitioning scheme suggested by PartitionFinder, and we analyzed these data using RAxML 7.3.4 (*64*), conducting 20 maximum likelihood (ML) tree searches and boostrapping using the "autoMRE" function. Bootstrapping converged after 50 replicates, and we used RAxML to reconcile each best, ML tree with the set of 50 bootstrap replicates.

To estimate rates of molecular evolution, we collected divergence time estimates from existing literature on the subject (SM 4). We created four alternative calibration schemes that allowed us to evaluate the effects of different age estimates on our overall estimates of evolutionary rate. We estimated rates of molecular evolution with the r8s program (*95*) (Supplemental File 3.10 - UCEFi10, Supplemental File 3.11 - UCEFi11, Supplemental File 3.12 - UCEFi12, Supplemental File 3.13 - UCEFi13) using these divergence-time estimates and the best ML tree for each UCE data set (taxon-set-1; taxon-set-2; taxon-set-2-partitioned).

We present the best ML tree and r8s estimates for taxon-set-2 in the main text, Figure S1 because this data set includes a larger sample of UCE loci from the taxa listed in Table S8 (excluding *Xenopus tropicalis*). R8s estimates for taxon-set-2 are provided in Figure S2. Here, we also provide the best ML tree and r8s estimates for taxon-set-1 (Figure S3 and Figure S4) and taxon-set-2-partitioned (Figure S5 and Figure S7). The topology, bootstrap support values, and pattern of estimated rates for taxon-set-1 do not differ from taxon-set-2. The topology of and support values across the partitioned tree for taxon-set-2 were not different from the unpartitioned tree (Figure S5), branch lengths of the partitioned versus unpartitioned tree were nearly identical (Figure S6), and r8s estimates from the partitioned tree showed the same patterns as we observed for the unpartitioned tree (Figure S7). Additional Supplemental Files 3.14-3.15 (UCEFi12-14) describe the steps used to generate and analyze these data and Supplemental File 3.15 - UCEFi15 provides a list of all files described above with checksums and additional descriptive information.

SM 4 – Establishing a temporal framework for vertebrate evolution

Edward L. Braun*, Brant C. Faircloth, John E. McCormack, and Christopher A. Brochu

* To whom correspondence should be addressed (ebraun68@ufl.edu)

The estimation of evolutionary rates requires a both well-supported phylogenetic framework and information from the fossil record. A consensus regarding vertebrate phylogeny has emerged (Figure S8), making it possible to constrain the timing for a number of divergences using fossil data. The majority of these divergence times, shown in Table S9, were obtained from Benton and Donoghue (*96*).

We add several additional calibrations based upon the fossil record, described below. Throughout this discussion we will define *crown group* as the least inclusive clade that does include all extant members of a clade (the common usage in paleontology). We define a *ghost lineage* as an evolutionary lineage that must exist based upon the best available phylogeny that nonetheless cannot be found in the fossil record for the appropriate times. Ksepka & Boyd (*97*) provide an excellent illustration of this for birds in their Figure 5.

There are several key divergence times needed to estimating rates of evolution for archosaurs that cannot be constrained using information obtained directly from the fossil record. Those dates require estimates of divergence times based upon other analyses that assume a relaxed molecular clock. We have chosen several different analyses to avoid biasing results as much as possible. However, we believe the 'young-young' calibration, which uses the fossil record for crocodilians (described below) and the molecular clock analysis from (*76*) for birds, is likely to be the most accurate.

General considerations for estimating evolutionary rates. A major goal of this study is estimating the absolute rate of genome evolution for the extant archosaurs (crocodilians and birds). This goal is related to analyses of divergence times based upon the relaxed molecular clock. The idea that estimating divergence times using molecular clock methods with incorrect fossil calibrations will have be problematic is not controversial. This is exacerbated by the absence of rigorous protocols for assigning molecular clock calibrations using fossils raises serious questions about the credibility of divergence dating analyses, and this has led to efforts standardize the identification of calibrations and build databases to serve this effort (*98-100*).

It is unclear, however, whether the estimation of rates (as opposed to divergence dates) is as sensitive to incorrect calibration times. Here we take an empirical approach to assess this sensitivity, anchoring our rate estimates to a rigorous assessment of the fossil record as possible and examining several different dates to assess the sensitivity of rate estimates. Although we limited their use, we did use some 'secondary' calibrations (based upon relaxed molecular clock analyses) where we were unable to identify appropriate primary calibrations. Since those secondary calibrations are based on other taxon samples, calibrations, and molecular data they do provide information that is independent of our genomic data.

The ideal primary calibration would involve a fossil in an accurately dated stratigraphic layer with sufficient preservation to allow multiple characters to be scored for a cladistic analysis. Those character data would then be used in a rigorous analysis that provides strong support for the placement of the fossil. Moreover, the topology extant taxa in a total evidence analysis using both molecular and morphological data would be for congruent with analyses of large molecular datasets. This criterion is seldom met. For example, analyses of crocodilians morphology alone typically place the root of extant crocodilians between Gavialidae and other crocodilians [e.g., Brochu (*101*)]. Analyses of molecular data alone and combined evidence place the root between Alligatoridae and a strongly supported Longirostres (the Crocodylidae+Gavialidae clade) (*8, 102*).

Some authors (*98*) have suggested that it is desirable to have fossils of the sister group, placed by rigorous phylogenetic analyses, with a similar temporal distribution. The goal of this criterion is to test whether the evolutionary divergence occurred close to the oldest record of the indicator taxon. The value of this criterion is debatable, since there may be a long 'ghost' range for the sister group and the sister group may persist for a long time. However, this criterion will typically be met when the fossil record is especially rich and, therefore, particularly useful. Therefore, it may prove to be a useful criterion.

Some of this difficulty could reflect specific difficulties such as relatively poor record for close crocodylian outgroups. This would be expected to result in a long branch at the base of extant crocodilians, which could be problematic. Few Campanian-Maastrichtian (upper Cretaceous) forms are known (e.g. *Massialasuchus, Allodaposuchus*) and there are some lineages that first appear during the Early Cretaceous and persist until later in the period (e.g. paralligatorids, hylaeochampsids, aegyptosuchids). Only one hylaeochampsid was known until comparatively recently, and other related groups, such as the susisuchids, were discovered within the past 10 years. Nonetheless, we are learning much about the closest relatives of Crocodylia (*103-110*) and current work is consistent with a Santonian-Campanian origin for Crocodylia, but there is more to be done.

There are other putative Campanian crocodylians, including the thoracosaurs and *Borealosuchus* Brochu 1997. Phylogenetic analyses support a gavialoid affinity for the Late Cretaceous-Paleocene thoracosaurs, which conflicts with molecular evidence for a Cenozoic divergence between *Gavialis* and *Tomistoma*. *Borealosuchus* and the thoracosaurs share morphological features currently optimized as plesiomorphic for Crocodylia, but they might instead be derived traits being mischaracterized by a rooting problem. Either way, these groups had a widespread geographic range and tend to be common in nonmarine, estuarine, and even shallow marine deposits when they occur; their rarity or absence prior to the Campanian is consistent with a radiation of derived eusuchians, including crocodylians, at or near the Santonian-Campanian boundary.

Calibration for the *Gavialis-Crocodylus* divergence (node P, 69-72 MYA). This divergence time for this node is based upon the existence of upper Cretaceous (early Maastrichtian) fossil gavialoid *Eothoracosaurus mississippiensis*, found in the Ripley formation of Mississippi (*111*). Cladistic analysis of morphological characters places this taxon sister to all other Gavialoidea. As stated above, analyses of morphology alone place deepest branch within the crocodilians between Gavialoidea and a clade comprising Alligatoridae and Crocodylidae. However, because combined analyses of molecular and morphological data (*102*) yield a strongly supported Longirostres (the crocodile + gharial clade) that includes fossil gavialoids we believe this fossil is likely to be member of crown Longirostres.

Calibration for the *Alligator-Crocodylus* divergence (node O, 77.9-83.6 MYA based on fossil data; also see Table S9). The minimum divergence time for this node is based on the oldest known alligatoroid, *Brachychampsa sealeyi* Williamson 1996 from the Allison Member of the Menefee Formation in New Mexico. The Menefee Formation is of Campanian age, and the Allison Member is from the lower part of the unit. This places *B. sealeyi* in the first half of the Campanian (77.9-83.6 MYA). Below we review some additional information on crocodilian paleontology that provide additional context and discuss our 'older' calibration for this node.

Two alligatoroids, *Leidyosuchus canadensis* Lambe 1907 and *Albertochampsa langstoni* Erickson 1972, are known from the Dinosaur Park Formation of Alberta (*112, 113*). Alligatoroids are also known from the Kaipirowits Formation of Utah (*114*) and the El Gallo Formation of Baja California (*115, 116*). These units are all within the Campanian, but are younger than the Allison Member (*117-120*). Likewise, the giant basal alligatoroid *Deinosuchus* Holland 1909 (which may have exceeded 12 m total length) had a widespread distribution throughout North America (*114, 121-125*), but is primarily known from the later Campanian.

We have reason to suspect an unsampled history for the group at least into the earliest part of the Campanian and perhaps into earlier stages. The oldest alligatoroid (*B. sealeyi*) is not the deepest-branching that honor belongs to *Leidyosuchus*. Most other known Campanian alligatoroids are globidontans more closely related to crown Alligatoridae than to *Leidyosuchus*, *Deinosuchus*, or *Diplocynodon* Pomel 1847, a lineage known only from the Cenozoic (Paleocene/Eocene through Miocene) of Europe (*126-129*), but which must have been present minimally by the Campanian.

The deepest-branch crocodyloid is *Prodiplocynodon langi* Mook 1941. The holotype is a skull from the Maastrichtian Lance Formation of Wyoming (*130*).

Several Late Cretaceous eusuchians from Europe have been referred to either Alligatoroidea (e.g. *Acynodon, Allodaposuchus, Masiliasuchus*; (*131-135*) or Crocodyloidea (*Arenysuchus* (*136*)). More recent analyses reject these referrals (*106, 108, 137-139*). At present, the oldest Eurasian fossils that can be robustly referred to Alligatoroidea or Crocodyloidea are of Cenozoic age.

Given the importance of this node for determining rates within crown crocodilians we also examined the impact of a secondary calibration from Oaks (*9*). These date push the origin of crown crocodilians back somewhat further, but they do correspond to a secondary calibration. This calibration is included to test the sensitivity of our conclusions regarding rates of genomic change to different assumptions regarding critical dates.

Calibration for the Crocodylia-Aves divergence (node N, 247-250 MYA). Ctenosauriscids are the oldest known members of the crocodile line (pseudosuchians) and they are also the oldest crown-group archosaurs. Ctenosauriscids first appear in the Olenekian (*2, 140-142*). The oldest dinosaur-line archosaurs (ornithodirans) known from body fossils are silesaurids from the overlying Anisian stage (*141*). Controversial footprint evidence may draw ornithodirans into the Olenekian (*143*), but they are not universally accepted.

Neither the ctenosauriscids nor the silesaurids are deepest-branching members of their lineages. It would thus be reasonable to expect crown archosaurs in units of earliest Triassic (Induan) or even latest Permian age, but at present, they are unknown prior to the Olenekian.

Calibration for the Archosauria-Testudines divergence (node L, 260- 265 MYA). The oldest known archosauriforms (crown-group archosaurs and close relatives bearing features diagnostic of "thecodonts," such as an antorbital fenestra) are from the latest Permian of Russia, but the date for this node is actually based on *Eunotosaurus* from the late Middle Permian (Capitanian) of South Africa. Recent analyses of this animal (e.g. (*144, 145*) point to very strong derived similarities with the Triassic basal turtle *Odontochelys*.

This node is generally viewed as difficult to calibrate. The phylogenetic relationships of turtles to extinct reptilian lineages are essentially unknown (*146-149*) and the position of turtles among extant vertebrates has been viewed as highly controversial (*150*). However, recent molecular analyses have converged on the placement of turtles shown in Figure S8 with strong support (*13, 28, 29, 151*). Thus, the uncertainty that remains reflects the difficulty of placing turtles based on morphology (necessary for fossil lineages). Some hypotheses would draw the turtle-archosaur split minimally into the Pennsylvanian. Alternatively, if *Eunotosaurus* is not a stem turtle, the date would instead be based on the first appearance of Archosauriformes. Note that Archosauriformes is more inclusive than Archosauria [the crown group that includes crocodilians and birds; cf. (*141*)] as it also includes stem taxa.

Calibration for the Palaeognathae-Neognathae divergence (node Q, see Table S9). The fossil record for the deepest branches for Neornithes (crown birds) are possibly one of the most problematic parts of the tree of life. With the exception of *Vegavis iaai* [viewed as a crown anseriform and therefore an indicator of Neognathae (*152*)] there are no convincing upper Cretaceous neornithine fossils (*153*). This absence has even led to continued defense of the hypothesis that all neornithine divergences occurred in the Paleogene (after the K-Pg mass extinction)(*154*).

However, a very recent radiation would demand truly remarkable acceleration of the rate of genomic change, far beyond that documented for other vertebrate lineages, for the long branches uniting the three major neornithine lineages: Palaeognathae (represented in Figure S8 by the ostrich, *Struthio camelus*), Galloanseres (represented in Figure S8 by the chicken, *Gallus gallus*), and Neoaves (the clade united by node S in Figure S8). The lengths of these branches relative to those at the base of Neoaves are simply incompatible with a radiation that has to be compressed into the early Paleogene [see Figure 1 in the main paper and Figure 1 in Jarvis et al. (*76*) for branch lengths and Figure 5 in Ksepka and Boyd (*97*) for a strict interpretation of the fossil record].

Therefore, we used two sets of dates: the most recent relaxed clock analyses, which was based on whole genome data (*76*) and a morphological clock analysis (*155*). The latter leads to older dates with broader credible intervals, but both push the radiation of Neornithes to the middle of the Cretaceous and represent a very realistic combination of information from the fossil record and molecular data. The use of the Lee et al. (*155*) study as part of our sensitivity analysis is especially important, since it is solely based on analyses of morphology from fossil and extant taxa, avoiding any potential for a bias due to the inclusion of molecular data.

SM 5 – Phylome analysis

Salvador Capella-Gutíerrez and Toni Gabaldón*

* To whom correspondence should be addressed (toni.gabaldon@crg.eu)

To gain insight into the evolutionary dynamics of three Crocodile genomes, representative of the main lineages of Crocodylia, we reconstructed their respective phylomes (*156*) *i.e.* the complete collections of evolutionary histories of all genes encoded in a genome and their homologs across other fully-sequenced species (Table S10).

Phylome reconstruction. Proteins encoded in 22 fully-sequenced vertebrate genomes - including the three crocodile genomes of this study - were downloaded from various sources (Table S10). Phylomes were reconstructed using as a seed each of the three newly-sequenced crocodile genomes. 23,309 unique protein sequences were used for the American alligator phylome, resulting in 19,226 gene trees (82.48% of the predicted proteins). Similarly, the crocodile phylome is based on 13,314 unique protein sequences, resulting in 11,857 gene trees (89.06% of the predicted sequences) and the gharial phylome is based on 14,024 unique protein sequences, allowing reconstruction of 12,644 gene trees, or 90.16% of gharial proteins.

To perform the phylome reconstruction, a Smith-Waterman (*157*) search was used to retrieve homologs using an e-value cut-off of 1e-5, and considering only sequences that aligned with a continuous region representing more than 50% of the query sequence. Then, selected homologous sequences were aligned using three different programs: MUSCLE v3.8 (*158*), MAFFT v6.712b (*159*), and DiAlign-TX (*160*). Alignments were performed in forward and reverse direction (i.e using the Head or Tail approach (*161*)), and the six resulting alignments were combined using M-Coffee (*162*). The resulting combined alignment was subsequently trimmed with trimAl v1.4 (*67*), using a consistency score cutoff of 0.1667 and a gap score cutoff of 0.1, to remove poorly aligned regions.

Phylogenetic trees based on the Maximum Likelihood (ML) approach were inferred from these alignments. ML trees were reconstructed using the best-fitting evolutionary model. The selection of the evolutionary model best fitting each protein family was performed as follows: A phylogenetic tree was reconstructed using a Neighbour Joining (NJ) approach as implemented in BioNJ (*163*); The likelihood of this topology was computed, allowing branch-length optimisation, using eight different models (JTT, LG, WAG, Blosum62, MtREV, VT, DCMut and Dayhoff), as implemented in PhyML v3 (*68*); The two evolutionary models best fitting the data were determined by comparing the likelihood of the used models according to the AIC criterion (*164*). Then, ML trees were derived using these two models, using the default tree topology search method NNI (Nearest Neighbor Interchange), and the one with the best likelihood was used for further analyses. A similar approach based on NJ topologies to select the best-fitting model for a subsequent ML analysis has been shown previously to be highly accurate (*165*). Branch support was computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution, as implemented in PhyML. In all cases, a discrete gamma-distribution with four rate categories plus invariant positions was used, estimating the gamma parameter and the fraction of invariant positions from the data.

Orthology/paralogy predictions. Orthology and paralogy relationships among crocodilian genes and those encoded by the other genomes included in the crocodilian phylomes were inferred using a phylogenetic approach (*69*). In brief, a species-overlap algorithm, as implemented in ETE v2 (*70*), was used to label each node in the phylogenetic tree as duplication or speciation depending on whether the descendant partitions have, at least one, common species or not (i.e. using a Species Overlap Score of 0). The resulting orthology and paralogy predictions can be accessed through phylomeDB.org (*19*). These predictions have been used in subsequent analyses such as orthology-based functional annotation, identification of gene expansions or duplication dating.

Species tree reconstruction. A phylogeny for the species included in the phylome was inferred using two complementary approaches, which rendered almost identical topologies. The only difference was in the internal organization of placental mammalian species used as out-groups. First, a super-tree was inferred from all trees in the three phylomes (43,727 trees) by using a Gene Tree Parsimony approach as implemented in the dup-tree algorithm [16]. This procedure finds the species topology that minimizes the number of total duplications implied by a collection of gene family trees, i.e. the phylomes. Second, 337 gene families with a clear, phylogeny-based, one-to-one orthology present in all species included in the analyses, were used to perform a multi-gene phylogenetic analysis. Gene families were taken from all three phylomes and the overlapping ones were included only once.

Protein sequence alignments were performed as described above and then concatenated into a single alignment. Species relationships (Figure S9) were inferred from this alignment using a Maximum Likelihood (ML) approach as implemented in PhyML v3.0 (*68*), using JTT as evolutionary model since in 319 out of 337 gene families this model was the best-fitting, and the tree topology search method SPR (Subtree pruning and regrafting). Branch supports were computed using an aLRT (approximate likelihood ratio test) parametric test based on a chi-square distribution.

We used ML analysis to contrast 2 additional scenarios (Figure S10). The hypothesis grouping Crocodilian with Birds was significantly more supported than any alternative topology as inferred from eight different statistical tests implemented in CONSEL (*166*). The vast majority of gene trees (6,880; 72%) placed crocodilians and birds together in a clade and only 28% of trees supported alternate topologies (birds + turtles or crocodilians + turtles). The alternative topology found most often grouped crocodilians and turtles (Table S11), possibly reflecting long branch attraction associated with the evolutionary distance between birds and the outgroup. Although the placement of gharial within the crocodilian phylogeny has been contentious over the last several decades (*33*) a clear majority (78.4%, #) of protein coding gene trees supported placing alligator as the sister group of a crocodile + gharial clade, termed Longirostres (*8*).

Phylome support for the species tree. All single gene trees in the phylomes were scanned to evaluate the support to the internal grouping of the main sauria groups obtained for the phylogeny of the species. Individual trees containing the four sauria groups: Squamata, Turtles, Crocodiles and Birds, were considered. It was required also that such groups were monophyletic. Table S11 shows the results obtained for each phylome, it is especially important to see that most of the selected trees support the reconstructed species phylogenies. Similarly we scanned all phylome trees to assess the proportion of trees that supported each of the three possible phylogenetic arrangements of crocodile, gharial and alligator. 78.44% of the trees supported the majority topology placing crocodile as the earliest diverging species ((crocodile, gharial),alligator), whereas 11.19% and 10.37 supported the two alternatives (crocodile,(gharial, alligator)) and ((crocodile,alligator),gharial), respectively.

Comparative genomics in the main saurian lineages. A comparative analysis, in terms of homology relationships, was performed among the 22 species included in the three crocodilian phylomes. Although all species included in the phylomes were used for this analysis, the main focus was on the main Sauria lineages: Crocodylia, Birds, Turtles and Squamata. To carry out this analysis, a BLAST search for each crocodilian protein-coding gene was performed to retrieve the first 200 homologous sequences with a cut-off e-value of 1e-5 and a minimum coverage -aligned region compared to the query sequence length- of 50%. Then, the different patterns of homology for each species were computed and plotted into Figure S11.

Transposable element detection. PFAM domains associated to transposable elements (TEs) activity were used to scan each coding-protein gene set using HMMER3 (*167*). 331 (~1.4% of total proteins), 155 (~1,1%) and 174 (1.2%) proteins from American alligator, crocodile and gharial, respectively, were discarded for downstream analyses since there were associated with at least one TEs domain, at e-value cutoff of 0.10, and had only homologs on the same species indicating a species-specific gene family expansion enriched on those elements.

Dating duplications. We scanned the three crocodilian phylomes to detect and date duplication events using a previously described algorithm (*71*). We focused on events assigned to five different relative evolutionary periods: Age 1) Species specific, Age 2) Basal to all three crocodilians (Crocodiles), Age 3) Basal to Crocodiles and Birds (Archosauria), Age 4) Basal to Crocodiles, Birds and Turtles, Age 5) Basal to Crocodiles, Birds, Turtles and Squamata (Sauria). Individual trees were scanned to date all duplication events which involved the seed protein and other proteins from the same species (paralogs). Table S12 shows the duplication ratios obtained for each relative age and species.

Functional enrichment for dated duplicates in the saurian lineages. Crocodilian proteins duplicated at different relative ages were analyzed looking for any functional enrichment. Enrichment analyses for over-represented GO terms for duplicated proteins as compared to individual species were performed using FatiGO as implemented in Babelomics webserver (*168*). A genome comparison approach was used with a Fisher exact test looking for over-represented terms with a normalized e-value cutoff of 0.01. Then, GO terms redundancy was reduced using REViGO webserver (*169*) with default settings. Table S13 shows those over-represented terms grouped by species, age and ontology.

Species-specific expansions of protein-coding genes in crocodilians. We focused on species-specific expansions of coding-protein genes. Proteins dated as duplicated before the split of the different crocodilian species were discarded. We detected 776 (~3.3% of total proteins), 176 (~1.3%) and 106 (0.75%) proteins mapped to such events in *A. mississippiensis*, *C. porosus* and *G. gangeticus*, respectively. Then, those proteins were grouped into clusters when overlaps of at least 75% among different groups of proteins from single gene trees were detected. More than 44% of the genes were assigned to a unique cluster in the species: ~48%, ~50% and ~44% for *A. mississippiensis*, *C. porosus* and *G. gangeticus*, respectively. Figure S12 shows a plot with the distribution of the cluster sizes frequency for each of the species studied.

Functional enrichment analysis for species-specific expansions in crocodilians. Clusters of protein-coding genes expanded specifically at the species level were analyzed looking for any statistically significant functional enrichment. Enrichment analyses for over-represented GO terms of those expanded protein families were performed by using FatiGO as implemented in the Babelomics webserver (*168*). Over-represented terms were detected by a Fisher exact test for genome comparison at an e-value cutoff of 0.01. Then, GO terms redundancy was reduced using REViGO webserver (*169*) with default parameters. Table S14 contains the non-redundant over-represented GO terms for clusters with 5 or more members in any of the three crocodilian species studied.

SM 6 – Analysis of 4D site rates

Edward L. Braun* and Brant C. Faircloth

* To whom correspondence should be addressed (ebraun68@ufl.edu)

The annotation of novel genomes can be difficult and is typically facilitated by the availability of transcriptome data, which was largely limited to the alligator (*Alligator mississippiensis*) in this study (see SOM 2 above). This suggests that the annotation of the alligator is better than that of the other two crocodilians, consistent with the numbers of genes that could be assigned gene symbols in each taxon as described above. Moreover, the quality of the genome assemblies (and the associated annotations) for taxa that were used for comparisons is expected to vary because different pipelines were used to generate those data. The impact of annotation (e.g., incorrect exon calls) and alignment errors on estimates of phylogeny (and the associated branch lengths) on phylogenomic analyses remain poorly characterized.

The topology for the tree (Figure S5) that resulted from phylome analyses, which used coding regions, was virtually identical to the topology based upon UCEs (Figure 1 of the main paper). The only topological difference involved the earliest divergence among eutherian mammals, where the UCE tree supported a clade comprising the armadillo (*Dasypus novemcinctus*) and elephant (*Loxodonta africana*) while the phylome tree placed the armadillo sister to all other eutherians. This relationship is extremely controversial (*26, 170, 171*) and has even been suggested to represent a hard polytomy (*172*). This degree of similarity strongly suggests that problems associated with coding region annotation have a minimal impact upon the estimation of tree topology, at least for the taxa examined here.

Despite the topological congruence between the UCE tree (Figure 1 of the main paper) and the phylome tree (Figure S5), the branches for the crocodile (*Crocodylus porosus*) and gharial (*Gavialis gangeticus*) were substantially longer in the phylome tree. UCEs represent contiguous regions that are straightforward to extract from genomes but analyses of coding regions are dependent upon the proper annotation of exons. This led us to hypothesize that the observed branch length differences might reflect differences in the annotation of these genomes. The alternative hypothesis (that there has been an acceleration of protein evolution in the crocodile and gharial) remains possible, but is less consistent with other lines of evidence (e.g., transposable element visibility) indicating slow rates of evolution for multiple genomic partitions in crocodilians (see the main text for details).

To test the hypothesis that these apparent rate differences reflect issues with the annotation of protein coding regions we applied two different filtering approaches to 686 one-to-one orthologs identified as part of the phylome analysis. First, we retained codons only if they were present in more than 20 taxa (i.e., up to two missing taxa were allowed). This corresponds to the *conserved fourfold degenerate* status described in more detail below, in SM 7. Second, we imposed the additional condition that the amino acid encoded by the codon of interest be conserved. The first filtering approach resulted in an alignment of 35,092 sites with 5.16% missing data (due to the presence of loci missing for one or two taxa) and 34,462 site patterns. The second filtering strategy is more stringent, reducing the alignment to 22,960 with 5.12% missing data and 22,576 site patterns. These alignments were used to estimate branch lengths in RAxML (*64*) using the UCE tree estimated for taxon-set-1 (Figure S3).

Both of these filtering strategies resulted in greatly reduced branch lengths for the crocodile and gharial (Figures S13 and S14), suggesting that the long branches evident in the unfiltered tree reflect the difficulties associated with identifying and annotating coding regions. Based upon these results we used a similar filtering method to assess the substitution rate for 4D sites in the whole genome alignment.

These analyses of 4D site rates also allowed us to estimate the rate at which these substitutions occur using the same method that was applied to the UCE data. Briefly, we estimated rates of molecular evolution using the r8s program (*95*) and the same divergence-time values used for analyses of the UCE data. The tree used for this analysis had the topology of the ML tree estimated using the UCE data and taxon-set-1. These analyses indicated that substitutions at 4D sites accumulated approximately an order of magnitude faster than those in UCEs, but the relative rates were quite similar, especially within tetrapods (Figures S15 and S16). We also note that the substitution rate estimates obtained using filtering strategy 2, which is expected to eliminate the largest amount of incorrectly aligned data, were slightly more congruent with the results of UCE analyses.

SM 7 – Whole genome alignments and ancestral genome reconstruction

Joel Armstrong, Ngan Nguyen, Edward L. Braun, Dent Earl, David Haussler, Richard E Green and Benedict Paten*

* To whom correspondence should be addressed (benedict@soe.ucsc.edu)

Whole Genome Alignment and Ancestral Genome Reconstruction
The whole genome alignment of 23 taxa was computed using progressive-cactus (github.com/glennhickey/progressiveCactus) using its default parameters and the phylogeny shown in Figure 2 of the main text. The genome assemblies used are listed in Table S15. The topology of the phylogeny was derived by manually merging a subtree of the UCE trees with results from the avian phylogeny sister paper (*76*) along with published phylogenies for passerine birds (*77*), parrots, (*78*), and turtles (*79*).

We used a 512 CPU cluster for running the local alignment jobs and a 1 terabyte shared memory machine with 64 cores for computing the CAF and reconstruction algorithms (*75, 80*). Ancestral reconstruction of all internal nodes was performed as part of this process. To improve the ancestral base calls we used the ancestorsML tool in the HAL tools library (github.com/glennhickey/hal) (*81*) to call bases by maximum likelihood, using the general reversible continuous time nucleotide substitution model implementation from the PHAST package *(60).* To parameterise the model and estimate branch lengths for the topology (show in Figure 5a of the main text) we used the phyloFit tool (*82*) on conserved fourfold degenerate sites in alligator genes, as described below. We also stored the posterior probability of these base calls given the model, and these results were used to calculate the expected accuracy of base calls in the archosaur genome. These results are shown in Figure 7a of the main text.

Whole Genome Alignment Analyses
The following gives technical definitions of the WGA analyses performed. A *whole genome alignment* (WGA) is formally a partitioning of the residues within a set of genomes into a set of aligned *columns*, each of which represents a set (technically an equivalence class) of residues inferred to be homologous. Given a chosen subset of genomes X within the WGA, a *non-duplicated column* for X is a column containing one or zero residues from each of the species in X. Similarly, given a chosen subset of species X within the WGA, a *single copy column* for X is a column containing exactly one residue from each of the species in X.

Percentage Identity. For a pair of genomes their *percentage identity* is the proportion of single copy columns for the pair that do not contain a wildcard character representing either genome and in which the nucleotide from both genomes is identical. The percentage identity value reported therefore includes the maximum number of columns where there is no apparent ambiguity about the ancestry of the residues. Table S16 shows the percentage identity between each of the three crocodilian genomes. Near identical results were produced using Mummer (*173*).

Fourfold Degenerate Codon Substitution Rates. In the WGA, a column is *conserved fourfold degenerate* if the column contains a residue that is a fourfold degenerate site in an annotated transcript in a given reference genome, and the two previous bases (in the opposite direction to the direction of transcription) in every leaf genome are such that each leaf genome site in the column is fourfold degenerate. Coding transcripts in alligator were filtered to only those with all the columns in their coding exons non-duplicated among crocodylia, chicken, zebra finch, and Carolina anole. The sites in alligator that corresponded to conserved fourfold degenerate columns in non-duplicated alligator coding sequences were then extracted. The program halPhyloPTrain (available in the HAL tools library) was used on these sites to estimate substitution rates for every branch in the WGA.

To validate these branch lengths, the non-duplicated alligator coding transcripts described above were also exported in Phylip format and further processed to remove regions that might bias the estimation of neutral rates before analysis in RAxML version 8.0.0 (*64*). We retained codons only if they were present in all taxa and did not have any internal gaps. This was done because the WGA allowed some indels of length one or two in coding regions; this typically occurred in regions where one or more of the sequences were poorly assembled (based upon visual inspection). The rationale for this stringent filtering is provided below in the section on 4-fold degenerate sites. This resulted in a matrix of 144,733 nucleotides and no missing data. Branch lengths were estimated for this alignment using the '-f e' option in RAxML with the GTRGAMMA model and the phylogeny used for other WGA analyses (see above). The tree length was 3.180126 and ML estimates of the parameters describing 4D site evolution were:

Γ distribution shape parameter
    α = 1.658157
GTR rate parameters (normalized to the G-T rate)
    rate A-C = 0.750544
    rate A-G = 4.203959
    rate A-T = 0.591019
    rate C-G = 1.387873
    rate C-T = 2.267481
Equilibrium base frequencies:
    π (A) = 0.280429
    π (C) = 0.252743
    π (G) = 0.167161
    π (T) = 0.299667

The resulting phylogeny with ML estimates of branch lengths is shown in Figure 2 of the main paper. A similar analysis was conducted by after increasing the stringency of the filtering to require that the amino acid encoded by the *conserved fourfold degenerate* column is itself conserved. This reduced the length of the alignment to 114,709 nucleotides but it had a negligible impact upon the branch length or parameter estimates. Estimates of the rates based upon 4D sites are presented in Table S17.

Transposable Element Substitution Rates. Any transposable element Y defines a nonempty subsequence x_i, x_{i+1}, … , x_j of a chosen reference genome (here the common ancestral genome of the extant crocodilian genomes in our analysis). We call Y syntenic with respect to a subset of genomes X, if:

    (1) The residues in x_{i-m}, x_{i-m+1}, …, x_{j+m+1} are all members of non-duplicated columns for X, where m is a flanking parameter (set to 2kb in this analysis; this ensures that the phylogeny is apparently unambiguous across the element).

    (2) For each pair of contiguous residues in x_{i-m}, x_{i-m+1}, … , x_{j+m+1}, if the columns they are contained in both contain residues from another genome in X, then those residues in the other genome are in the same order and orientation as in the chosen reference genome and are separated by no more 100 bases in the other genome (this ensures that no rearrangements other than indels less than 100bp in length and substitutions have been observed to effect the sequences). We use the set of single-copy columns that contain residues from syntenic transposable elements to calculate the substitution rate in transposable elements.

To estimate rates we used the strand symmetric general reversible continuous time substitution model implemented in the Phast package continuous time substitution model (*82*), using the halPhyloPTrain program on the single-copy columns from syntenic transposable elements in the common ancestor of Crocodylia. Table S18 below shows the TE substitution rates.

Micro Insertion and Micro Deletion Rates. For a pair of genomes A and B, a *clean insertion* in A with respect to B is a nonempty subsequence x_i, x_{i+1}, … , x_j of a sequence in A such that:

    (1) The residues in x_i, x_{i+1}, … , x_j are not aligned to any residues in B.

    (2) The residues in x_{i-k-1}, … , x_{i-1} and x_{j+1}, … , x_{j+k+1} are each aligned in single copy columns for { A, B }, where k represents a number of cleanly aligned residues flanking the insertion. This condition ensures no duplications that suggest ambiguity in the phylogeny. After some testing, in this analysis, k=5 though other larger values of k produce similar results.

    (3) The corresponding residues in B aligned to x_{i-k-1}, … , x_{i-1} and x_{j+1}, … , x_{j+k+1} are in the same order and orientation in B as in A. This ensures the structural change is indeed an insertion rather than a more complex rearrangement.

    (4) None of the residues in the alignment columns containing x_{i-k-1}, … , x_{j+k+1} represent the wildcard character. This avoids labeling scaffold gaps as insertions.

A *clean insertion* in A with respect to B is, reversely, a *clean deletion* in B with respect to A. A clean indel (insertion or deletion) is a *micro* event if the inserted or deleted subsequence is less than or equal to 10 residues in length. A *clean adjacency* is either a clean insertion or deletion, or equivalent to a clean insertion or deletion in which the inserted subsequence has zero length; a clean adjacency represents a place where there could have been a clean indel, but potentially none was observed.

Let an induced subtree of the phylogeny connecting a chosen genome A, its sister genome and their closest outgroup genome, be termed a *three-leaf subtree for A*. For a chosen genome A with corresponding three-leaf subtree, a clean insertion, deletion or adjacency with respect to its three-leaf subtree is a clean branch insertion, deletion or adjacency, respectively, in A with respect to both the sister and outgroup genomes of the three-leaf subtree. Note this definition discounts clean insertions, deletions or adjacencies which differ between the outgroup and sister genomes, i.e. the indel subsequence has to have the same length in both other species. Defining events with respect to three-leaf subtrees gives confidence in the categorization of the event as an insertion, deletion or clean adjacency.

The insertion and deletion rates reported are the ratio of clean micro insertion or deletion events per clean adjacency. Normalising by clean adjacency proved necessary to factor in coverage differences between assemblies. Table S19 shows the measured rates of clean insertions and deletions in each of the leaf taxa of the WGA.

Gene Synteny. For a pair of genomes A and B and pair of genes X and Y on A, we say the pair X and Y are *candidates for synteny* if X and Y both:
   (1) map uniquely from A to B (no evidence of duplication in the other species),
   (2) map to the same scaffold on B,
   (3) map at least 90% of their sequence to B,
   (4) reciprocally preserve their structures to B (i.e. X and Y must be preserved from A to B, and their images on B must be preserved back from B to A, see below for technical definition of preservation).

If X and Y are candidates for synteny and they are in the same order and orientation on B as on A, they are *syntenic*, otherwise they are *broken*. If X or Y are not candidates for synteny, they are considered invalid and the pair is neither considered syntenic nor broken. The ratio of broken/syntenic pairs is the *gene pair synteny rate*. This carefully constructed analysis was necessary to unambiguously identify orthologous pairs of gene and minimise assembly differences impacting the results, though it is still likely to be somewhat affected by assembly composition and errors.

Archosaur Reconstruction Analyses
Below we detail analyses used to assess the reconstructed archosaur assembly.

Estimating Potential Missing Sequence in the Archosaur Assembly. Due to the parsimony based simultaneous progressive alignment and reconstruction approach used to construct the WGA, any sequence in alligator that has a homolog outside the crocodilian lineage must have a homolog in the reconstructed archosaur. This implies that alligator sequences without an ortholog in the archosaur do not have an ortholog outside the crocodilian lineage. Misalignment and assembly errors will tend to reduce the quality of the ancestral reconstruction (usually by missing sequence), and thus lower the amount of sequence aligned. To estimate how much sequence should have been included in the archosaur reconstruction but was not, alligator fragments that did not align to archosaur at all were aligned against a selection of other leaf genomes (Table S20) using LastZ (version 1.03.52) with the following parameters: [multiple,unmask], which ensure that any repetitive sequence will not be unaligned due to heuristics that by default avoid the alignment of soft-masked sequence.

While most fragments (87%) mapped to crocodile, the substantial majority do not map to these outgroup genomes (e.g. 91% of unmapped fragments do not align to anywhere in chicken) using LASTZ. This suggests that the reconstructed genome, despite its small size, already approaches the maximum possible size for a reconstructed genome given current alignment techniques. The small minority of regions that do map to out-group genomes are largely repetitive (mapping to many places in the outgroup genomes), suggesting the reconstruction of repetitive elements is an area of future improvement.

Element Categories for Archosaur Analysis. To avoid issues with double-counting elements and bases in the mapping, phyloP, and structure-preservation analyses, the BED files for these categories were pre-processed. Gene and chained-CDS categories were processed to select only the longest transcript where there was overlap between multiple elements on the same strand. All other categories had their elements merged together where overlapping to avoid multiple-counting.

The gene annotations used were as described above for the alligator and, for the chicken, RefSeq gene annotations available from the assembly hub (see below), or at http://hgwdev.cse.ucsc.edu/~jcarmstr/permanent/galGal4_refSeq.bed .

Selection Analysis. The halPhyloP tool, available from the HAL tools library, was used to generate phyloP scores for all columns in the alignment. The input branch lengths were determined by running phyloFit on conserved alligator fourfold degenerate sites, as described above. Using the WGA, each column was lifted over to each genome, creating a phyloP wiggle track for each of the 23 leaf and 22 ancestral genomes in the WGA.

Order Preservation. An element, either exon, UTR, intron, etc., is defined by an interval of a genome. For a pair of genomes A and B and element Y in A, for a pair of successive residues in Y that align to a sequence X in B, we say their *adjacency* is preserved if the corresponding residues aligned in X are in the same order and orientation and separated by less than 100bp. We say the structure of Y is preserved if for all such pairs the adjacency is preserved and at least 25% of bases in Y align to X, and X is the sequence in B where the majority residues in Y align without self alignment (self duplication). This definition ensures that to have preserved structure at least a reasonable minority of bases must align to a single sequence and be organized as in the reference genome. If Y is a coding sequence (CDS), comprising the coding portions of a gene's exons, it is treated as a single element, except that residues in the interstitial introns are ignored, and introns are allowed to change in size by up to 100kb in X.

Extant Mapping Controls. The proportion of elements and adjacencies that were preserved is shown for alignments between alligator to archosaur (see main text, and Figure S17), alligator to chicken (Figure S18) and chicken to archosaur (Figure S19). These comparative controls show we get similar, but uniformly slightly higher results mapping chicken, rather than alligator, genes to the archosaur (presumably due to differences in gene sets, as the evolutionary distance is expected to be greater), and substantially higher mapping and order and orientation preservation results mapping extant annotations (either alligator or chicken) to archosaur than mapping alligator annotations to chicken, or vice versa.

Assembly Hub

A Comparative Assembly Hub for the WGA is available for the UCSC genome browser (*174*) at http://genome.ucsc.edu/cgi-bin/hgHubConnect (locate the "Croc and Bird Hub" link). From it, it is possible to browse the genomes, annotations and alignments, and download (via the Table Browser), portions of the WGA, the sequences of the reconstructed genome as well as the alligator, gharial and crocodile gene sets.

SM 8 – Lineage specificity and 'visibility' of ancient TEs in amniotes.

David A. Ray, Robert Hubley and Arian F. Smit[*]

* To whom correspondence should be addressed (asmit@systemsbiology.org)

<u>Lineage specificity of TE families</u>. The lineage-specificity of each family was determined using several methods. The ancestral nature of several hundred families was predicted by the fact that we had constructed identical consensus sequences from the genomes of two or all three species (or near-identical consensus sequences that upon inspection were found to represent the same element).

The main strategy involved running RepeatMasker with a combined library of all repeats reconstructed in any of the three crocodilian genomes. Repeat families were considered ancestral to two species if the family covered the same number of bases (±10%) in both species' genomes and the average mismatch level was the same (again ±10%). They were considered lineage-specific if one species had a 10-fold higher representation of said repeat. The characteristics of 81 of the 1269 repeat families fell in between these margins and were confirmed by other methods. For some of these families both the base coverage and the average divergence level in one species was considerably higher than in the other species, indicating matches to an undefined lineage-specific repeat related to the repeat for which the consensus was derived. Others of these exceptional families showed a higher base coverage but lower divergence level in one than in the other species. Here the consensus represented a lineage-specific repeat that matches related unidentified ancestral repeats.

To account for the possibility of elements that are present in two species through horizontal transfer rather than vertical inheritance, we checked for and confirmed the presence of copies at orthologous sites for all ancestral families with a very low average divergence level. Finally, we verified the results of this analysis by checking if copies can be found at orthologous sites in different species for a random subset of both lineage-specific and ancestral repeats.

Rates of new TE family evolution in each lineage were calculated by identifying lineage-specific TE families Crocodylia (alligator + crocodile + gharial), Longirostres (crocodile + gharial), alligator alone, crocodile alone, and gharial alone. A few dozen were found to be at orthologous sites in all amniotes. The remainder were mapped to each lineage of the crocodilian tree and the number of TE families was divided by the approximate divergence date for each lineage from its closest relative (Figure S20).

<u>TE visibility</u>. The data in Figure 3 (main text) and Figures S21, S22, and S23 are derived from RepeatMasker output (sensitive settings) for each genome listed, using RepeatMasker version 4.0.5 and RepeatMasker library version 20140131, largely corresponding to the RepBase release 19.01 (a number of ancient consensus sequences appeared in a later release). The assembly versions of the different genomes are given in the Figure legend of each Figure. The 1.86 Gbp reconstructed boreoeutherian genome was derived from release 74 of the EPO (Enredo-Pecan-Orpheus) multiple alignment of 15 boreoeutherian genomes (*36*).

Before and during this study, we had created (partial) consensus sequences for 132 interspersed repeats predating the amniote speciation. We reduced this number to 110 by combining subtypes of the same family of repeats and further to 74 by eliminating families with fewer than 100 copies in more than one species as well as families that show a distorting number of matches to uncharacterized lineage-specific repeats. The latter families included most LINE elements. Some irregularities will still arise from the presence of uncharacterized lineage-specific repeats on the one hand or from characterized related consensus sequence, appropriately present in the repeat library of one species but not the other, but competing with the amniote-wide consensus for matches. The Y axis represents the number of bases seen in a particular species divided by the average number of bases seen in all species displayed. This is the same as (# of species) * (bp of this repeat in this species) / (bp of this repeat in all species). On the X axis the repeat families are ordered alphabetically.

There are many forces that can affect the relative preservation level of the repeats in the genome, so that one does not expect the same ratios of visible repeats for each family, even when there is no interference by lineage-specific matches. These include but are not limited to:

1. In species subject to fast neutral decay, like the mouse, shrew and anole, perhaps all visible copies of 300 milion-year old repeats have been exapted and remained under selective constraint. The difference with slower evolving genomes becomes smaller if a larger fraction of repeat sequences have gained a function in the genome.
2. Very old repeats may be so diverged that they evade detection in all but the slowest evolving genomes unless they have become fixed. This would lead to a larger difference in the number detected in turtles over those in crocodiles, and a smaller difference between crocodiles and mammals.
3. The ratio of GC->AT over AT->GC decay may differ significantly between lineages, so that, for example, GC-rich repeats may be easier to detect in some lineages than others.

Figures 3 (main text), S18, S19, and S20 each attempt to display the relative rate of neutral decay through deletions and point mutations. They are all in log scale, but differ from one another in the following ways. As mentioned above, Figure 3 displays the bases identified in each species relative to the average identified in all 16 genomes the number of TE families illustrated has been reduced for the sake of clarity. Figure S21 includes all of the TE families analyzed. These figures most cleanly separate the turtle and crocodilians from the other species, but fail to display, for example, the notably slower substitution levels observed for ancestral repeats in birds and platypus compared to those in eutherian mammals or the partial restoration of those repeats in the reconstructed ancestral boreoeutherian genome. The latter issue has a straightforward explanation in that only a fraction (about two thirds) of the ancestral genome is reconstructed. To take that into account, Figure S22 shows the relative density (bp of repeat / genome size) for each genome. The density of recognizable ancient repeats in the boreoeutherian genome is, on average over all 74 repeat families, indeed 2.3 times higher than it is in human, while the density in the chicken and platypus genomes is 3.5 and 2.9-fold higher. This comparison is however also unfair, as a larger accumulation of lineage-specific repeats will lower the ancient repeat density given the same neutral decay rate. Figure S23 presents the relative density of repeats using a genome size excluding the lineage-specific repeats. Compared to human, the relative average densities excluding lineage repeats are 8.6 in the painted turtle, 4.7 in crocodilians, 1.95 in chicken, 1.72 in platypus, 1.22 in opossum, 0.64 in mouse, and 0.52 in the anole lizard.

SM 9 – Analysis of selected gene families

Mike W. Vandewege, Andrés Iriarte, and Federico G. Hoffmann[*]

* To whom correspondence should be addressed (federico.g.hoffmann@gmail.com)

Analyses of selected gene families: We conducted bioinformatic searches to characterize the repertoires of Olfactory Receptors (ORs), Vomeronasal Receptors types 1 and 2 (V1R and V2R), Taste Receptors type 1 and 2 (T1R and T2R) and trace amine-associated receptors (TAAR) of the three crocodilians in our study, and compared with representative vertebrates.

Olfactory Receptors:
In the case of ORs (Table S21), we performed TblastN searches with a cutoff of e-10 of the three crocodilian genomes using known vertebrate ORs as queries. The best non-overlapping BLAST hits were extracted using modules in BEDTOOLS (175) and using custom Python scripts. Putative OR genes were declared intact if there was an uninterrupted open reading frame with no gaps longer than five amino acids in any of the seven transmembrane domains or conserved regions and an appropriate stop codon. Newly discovered intact ORs were added to the amino acid query and a new TBlastN search using a cutoff of e-20 was conducted to annotate pseudogenes and truncated genes. ORs were declared pseudogenes if the longest open reading frame was shorter than 250 amino acids, there were gaps of five or more amino acids in the transmembrane domains/conserved regions, or had frame-shift mutations or premature stop codons. ORs were declared truncated if they were located at the end of a scaffold or were interrupted by scaffold gaps and contained a truncated but intact OR amino acid sequence. Truncated ORs were validated by alignment to functional genes using MAFFT 7.127 (176) and visually inspected for premature stop codons and gaps. Putative ORs were annotated to their subfamily by comparing amino acid sequences against a BLASTP database of known OR amino acid sequences.

Phylogenetic analyses were conducted using MEGA v5 (73). We inferred neighbor-joining phylogenies to assess patterns of divergence and diversity of intact crocodilian ORs compared to other vertebrates using a Poisson model of substitution and evaluated support for the nodes with 1,000 pseudoreplicates. In Figure S24, we compared the evolution of ORs for the three crocodilians, chicken and zebra finch (74), and green sea turtle and Asian softshell turtle (16).

Vomeronasal, Taste and Trace amine-associated receptors:
In the case of V1R, V2R, T1R, T2R and TAARs (Table S22), we followed a similar procedure as the one implemented for ORs. Candidate sequences were identified with TblastN using the protein coding sequences of well-characterized genes to seed searches. V1R, V2R, T1R, T2R and TAARs genes were declared intact if the putative coding sequences included the seven transmembrane helices domain as predicted by means of the TMHMM v.2.0 program (177). Candidate receptors were declared pseudogenes if the seven transmembrane helices domain structure was absent or incomplete according to the TMHMM prediction method, or had frame-shift mutations or premature stop codons in any predicted exon. Candidate receptors were declared truncated if they were located at the end of a scaffold or were interrupted by scaffold gaps. The number of multiexon genes, V2R and T1R, was estimated based on the presence of the complete highly conserved seven transmembrane helices domain encoded in the last exon. In these cases, partial genes that were apparently truncated owing to incomplete sequence assembly, were counted as functional intact genes as long as they had an entire seven transmembrane helices domain, other partial genes were counted as truncated genes.

Phylogenetic analyses for V1R, V2R, T1R, T2R and TAARs were as described for ORs.

SM 10 – Population genetic analyses

Michael Vandewege and Richard E. Green*

* To whom correspondence should be addressed (ed@soe.ucsc.edu)

Estimation of crocodilian-lineage mutation rate. We used a phylogenetic approach to estimate the overall mutation rate, , along the crocodilian lineage. From both the whole genome alignment between alligator and crocodile and the multiple sequence alignment that includes alligator and crocodile, we estimate the overall divergence between alligator and crocodile to be 7.1%. Because of the remarkably small divergence between these two, we assume an infinite sites model of evolution and ignore back mutations. We calculate a per generation mutation rate using 90 MY as the time of the most recent common ancestor of alligator and crocodile and an average generation time of 20 years (Table S23) as shown below.

$$\mu = \frac{\dfrac{0.071}{2}}{\dfrac{90,000,000\,years}{20\,years\,/\,generation}} = 7.9 \times 10^{-9}\ subtitutions\,/\,site\,/\,generation$$

Estimation of heterozygosity and N_e. For each genome, we aligned 10 to 14x fold coverage of the paired-end genome reads from a single individual back to the final genome assembly using bwa (*83*). We used tools in the GATK package (http://www.broadinstitute.org/gatk/) to perform indel realignment of each read around possible insertion-deletion positions. We then analysed all genomic positions where the read depth was *exactly* equal to the genome-wide mean, considering only reads with map quality at least 30 and positions of base quality at least 30 (10x for gharial, 14x for alligator and crocodile). For all such positions, we counted the number of reference and alternate alleles at each position annotated as exon, intron, or intergenic (neither exon nor intron). This distribution is shown for the crocodile data in Figure S25.

We analysed the substitution spectra across this distribution to determine a reasonable level of evidence required to call a position a heterozygous site as shown in Table S24, for the crocodile data at 14x genome coverage. As expected, the substitution spectra at genomic positions that have only one or two reads that differ from the consensus are highly asymmetric in the number of each transition (pink elements in Table S24) and in the number of transversions (blue elements in Table S24). Further these positions have transition/transversion ratios that are not typical of genuine molecular sequence evolution. For example, the ratio of transitions to transversions at sites where only a single alternate allele base is seen is 0.42. In contrast, the substitution spectra at positions covered by equal numbers of two alleles appear more consistent with *bona fide* molecular evolution. There are roughly equal numbers of each transition and each transversion. At sites where 7 reference and 7 alternate alleles are seen in the read data, the observed ratio of transitions to transversions is 2.50. From analysis of these data, we required three or more high-quality observations of both the reference and alternate allele to call a site as heterozygous.

We followed a similar strategy for the gharial and alligator data where these trends were similar. From this analysis, we calculated the observed rate of heterozygosity, H, at intergenic sequence in each species: alligator $H = 0.000136$; gharial $H = 0.000217$; crocodile $H = 0.000360$. Using these values as an estimate for theta and the substitution rate, calculated above, we estimated the effective population size for each species as shown in Figure 5 of the main manuscript.

PSMC analysis. Whole genomic Illumina pair-end reads from *A. mississippiensis*, *C. porosus*, and *G. gangeticus* were aligned to the respective reference genomes using BWA. PCR duplicates were removed using Picard and SNPs were called with SAMtools. We only used reads with a >30 map score and base calls with a >20 quality score to infer SNPs. We applied the pairwise sequential Markovian coalescent (*43*) model using 20 years for the generation time (Table S23). We estimated that the MRCA of *C. porosus* and *A. mississippiensis* existed approximately 90 MYA and our analyses indicated these genomes are 92.9% identical. Therefore, given a 20 year generation time, we yielded the expected crocodilian mutation rate of 7.89 x 10^-9/year/site. We conducted bootstrap tests for each of the 3 samples by splitting the scaffolds into smaller segments and randomly sampling the segments with replacement (Figure S26). We used 100 replicates to test the robustness of the returned population demographic history. We also

gathered ancestral Northern Hemisphere air temperature data from (*44*). We took averages for 200k year bins and plotted the ancestral climate history, relative to present, over the past eight million years with the demographic history to infer any climate related demographic changes. We repeated the process to graph climate oscillations for the last million years; however we took averages for 20k year bins.

SM 11 – GC content evolution

Matthew K. Fujita* and Juan C. Opazo*

* To whom correspondence should be addressed (mkfujita@gmail.com, jopazo@gmail.com)

Isochores are regions of the genome with homogeneous GC content that correlates with gene density, repetitive element content, recombination rate, and mutation rate (*178-181*). We used the 337 orthologous protein-coding genes from 22 vertebrate species identified in the Phylome analysis above to analyze the GC content at third-codon positions. With these gene alignments as input, we used the program nhPhyml to calculate ancestral GC3 as well as GC3* assuming the tree topology shown in Figure S27 (*182*). In addition to GC content at third codon positions (GC3), we also examine the inferred GC3 assuming a model at equilibrium (equilibrium GC3*). Finally, we calculated divergence the lineage-specific divergence in GC3 based on the ancestral GC3 as inferred from the ancestral genome reconstructions (*183*). This analysis allows us to identify patterns of GC content enrichment or erosion in the crocodilian lineages.

GC3* represents the GC3 toward which a lineage is evolving (*184*). Based on our analyses we found that GC3 is very stable in crocodilians with an overall trend in all three lineages towards a decreasing GC3 (indicating GC3 is not at equilibrium).

Unlike lizards, mammal and bird genomes exhibit genomic isochores (*178*). Cocodilians exhibit GC content heterogeneity similar to mammals and birds (*55*) but it was unclear whether crocodilians actually exhibited isochore structure. We measured the presence of isochores in a genome by examining the variance in GC content in windows of appropriate size across the genome. Genomes with no spatial heterogeneity in GC content have a small variance because all regions of the genome have a similar average GC percentage.Crocodilian genomes have an overall genomic GC content of 44% (SOM 8), higher than most other amniotes but similar to turtles (*37*). Both crocodilians and turtle genomes demonstrate spatial heterogeneity of GC content at smaller spatial scales (5 kb), similar to mammals and birds, and consistent with the presence of isochores. However, heterogeneity falls more quickly in crocodilian genomes and the painted turtle genome at larger spatial scales (Figure S28). Whether this phenomenon is indicative of a fundamental difference in isochore length or a function of the fragmented genome assemblies of crocodilians and the painted turtle is not clear. In either case, however, it is clear that crocodilian genomes, unlike lizards, do contain GC isochores.

Ancestral genome reconstruction represents an opportunity to explore genomic features of the ancestors of extant clades. In this regard, isochore structure represents an important genome trait that has been associated to key aspects of genome organization. To investigate the evolution of the isochore structure among archosaurs and turtles, we reconstructed the ancestral genomes of birds, archosaurs (birds and crocodiles) and of the clade containing turtles and archosaurs, using a taxonomic sampling as described for the Phylome analysis, and estimated the genomic GC standard deviation at different windows sizes (Figure S29).

SM 12 – Data availability

The data for this project are available in several publically accessible locations. Genome assemblies (including, in some cases BLAST and BLAT searchable versions) are available at NCBI, CoGe, GigaScience and crocgenomes.org. GigaScience is serving as host to the genome assemblies as well as other important files reflecting the many analyses performed. PhylomeDB generated gene trees are also available (see link below).

Genome downloads and data repositories
www.crocgenomes.org - http://www.crocgenomes.org/downloads.html
CrocBase - http://crocgenome.hpc.msstate.edu/crocbase/
GBrowse Databases - http://crocgenome.hpc.msstate.edu/crocbase/gbrowsers.html
Genome Annotation Database – http://crocgenome.hpc.msstate.edu/crocbase/
GigaScience –   http://dx.doi.org/10.5524/100125
                     http://dx.doi.org/10.5524/100126
                     http://dx.doi.org/10.5524/100127
                     http://dx.doi.org/10.5524/100128
National Center for Biotechnology Information - http://www.ncbi.nlm.nih.gov/
                     Alligator mississippiensis accession – AKHW00000000
                     Crocodylus porosus accession – JRXG00000000
                     Gavalis gangeticus – JRWT00000000.
UCSC Genome Browser - https://genome.ucsc.edu/ (currently alligator only)
Comparative Genomics - https://genomevolution.org/CoGe/
Ancestral archosaur reconstructions - http://genome.ucsc.edu/cgi-bin/hgHubConnect
                     Locate the "Croc and Bird Hub" link

Supplemental files
Files below are downloadable from GigaScience data notes (see above)
Tree inferences and alignments
Ultraconserved Elements
        ReadMe.txt
        UCEFi1-uce-12k-probes.fasta.gz
        UCEFi2-genome-sequence-location.conf.gz
        UCEFi3-uce-12k-probes-to-genomes-lastz-files.tar.gz
        UCEFi4-uce-12k-probes-v2.sqlite.gz
        UCEFi5-uce-12k-2kb-fasta-flanks.tar.gz
        UCEFi6-uce-12k-2kb-fasta-flanks-assembled-to-1kb-UCE-loci.tar.gz
        UCEFi7-croc-uce-match-count.conf.gz
        UCEFi8-taxon-group1.tar.gz
        UCEFi9-taxon-group2.tar.gz
        UCEFi10-Fig1-main-panels.tar.gz
        UCEFi11-FigS2-r8s-analysis.tar.gz
        UCEFi12-FigS4-r8s-analysis.tar.gz
        UCEFi13-FigS7-r8s-partitioned-analysis.tar.gz
        UCEFi14-analysis-steps.txt
        UCEFi15-croc-uce-name-map.conf.gz
PhylomeDB
        ReadMe.txt
        337sets.one2one.ConcatenateAlignment.phy
        337sets.one2one.ConcatenateAlignment.SpeciesTree.PhyML.SpCodes.nw
Olfactory Receptors
        ReadMe.txt
        Fig_S24_A_crocs.nwk
        Fig_S24_A_crocs_aln.fas

Fig_S24_B_birds.nwk
Fig_S24_B_birds_aln.fas
Fig_S24_C_turtles.nwk
Fig_S24_C_turtles_aln.fas
4D Sites Analysis
README.4Dsite_analysis.txt
Phylome_4Dsites_filter1.phy
Phylome_4Dsites_filter2.phy
Phylome_species_tree.tre
RAxML_info.Phylome_4Dsites_filter1.brlen
RAxML_info.Phylome_4Dsites_filter2.brlen
RAxML_info.WholeGenome_4Dsites_filter1.brlen
RAxML_info.WholeGenome_4Dsites_filter2.brlen
RAxML_result.Phylome_4Dsites_filter1.brlen
RAxML_result.Phylome_4Dsites_filter2.brlen
RAxML_result.WholeGenome_4Dsites_filter1.brlen
RAxML_result.WholeGenome_4Dsites_filter2.brlen
WholeGenome_4Dsites_filter1.phy
WholeGenome_4Dsites_filter2.phy
WholeGenome_species_tree.tre
Ancestral Archosaur Reconstruction
ReadMe.txt
ancestral_genomes_reestimated.tar.gz
crocPaper_reestimated.hal
Genes and Gene Ontology
Gbrowse files
amis_gbrowse_v1.2.gff3
cpor_gbrowse_v1.2.gff3
ggan_gbrowse_v1.2.gff3
Gene Ontology (GO)
amis_go_annotation_201306.tsv
cpor_go_annotation_201306.tsv
croc_go_annots_20130621.zip
Croc_GO_Stats_201306.txt
ggan_go_annotation_201306.tsv
go_summary_by_gene_id_201306.tsv
go_summary_by_go_id_201306.tsv
interpro_summary_by_genes_201306.tsv
interpro_summary_by_interpros_201306.tsv
Gene Predictions
amiss_cds.fa
amiss_proteins.fa
cpor_cds.fa
cpor_proteins.fa
ggan_cds.fa
ggan_proteins.fa
Nomenclature and Function
tbl_gene_info.zip
tbl_gene_info_alligator.tsv
tbl_gene_info_crocodile.tsv
tbl_gene_info_gharial.tsv
Pathway Predictions

croc-pathway-genes.txt

Transposable elements

All transposable element consensus sequences have been deposited in RepBase - http://www.girinst.org/repbase/index.html. The Boreoeutherian genome reconstruction mentioned in SM 8 is available at crocgenomes.org and GigaScience (link above).

PhylomeDB data

All gene trees, alignments and orthology and paralogy calls for the phylomes generated for this project are available through phylomeDB (http://phylomedb.org).

Supplemental Figures:
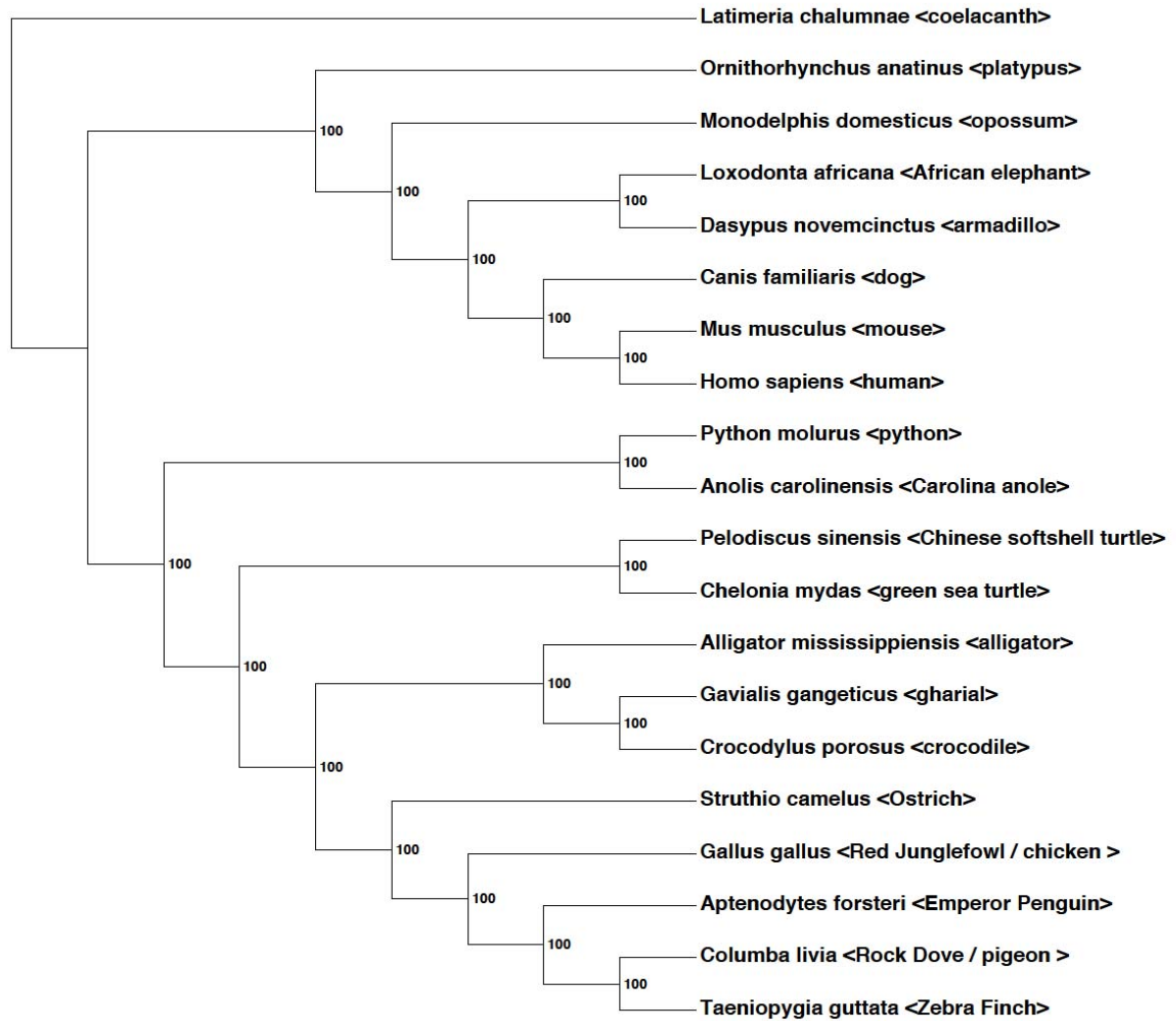
Figure S1. Inferred phylogeny of taxon-set-2 using a maximum likelihood analysis of a complete matrix of 965 ultraconserved element loci and RAxML v7.3.4.
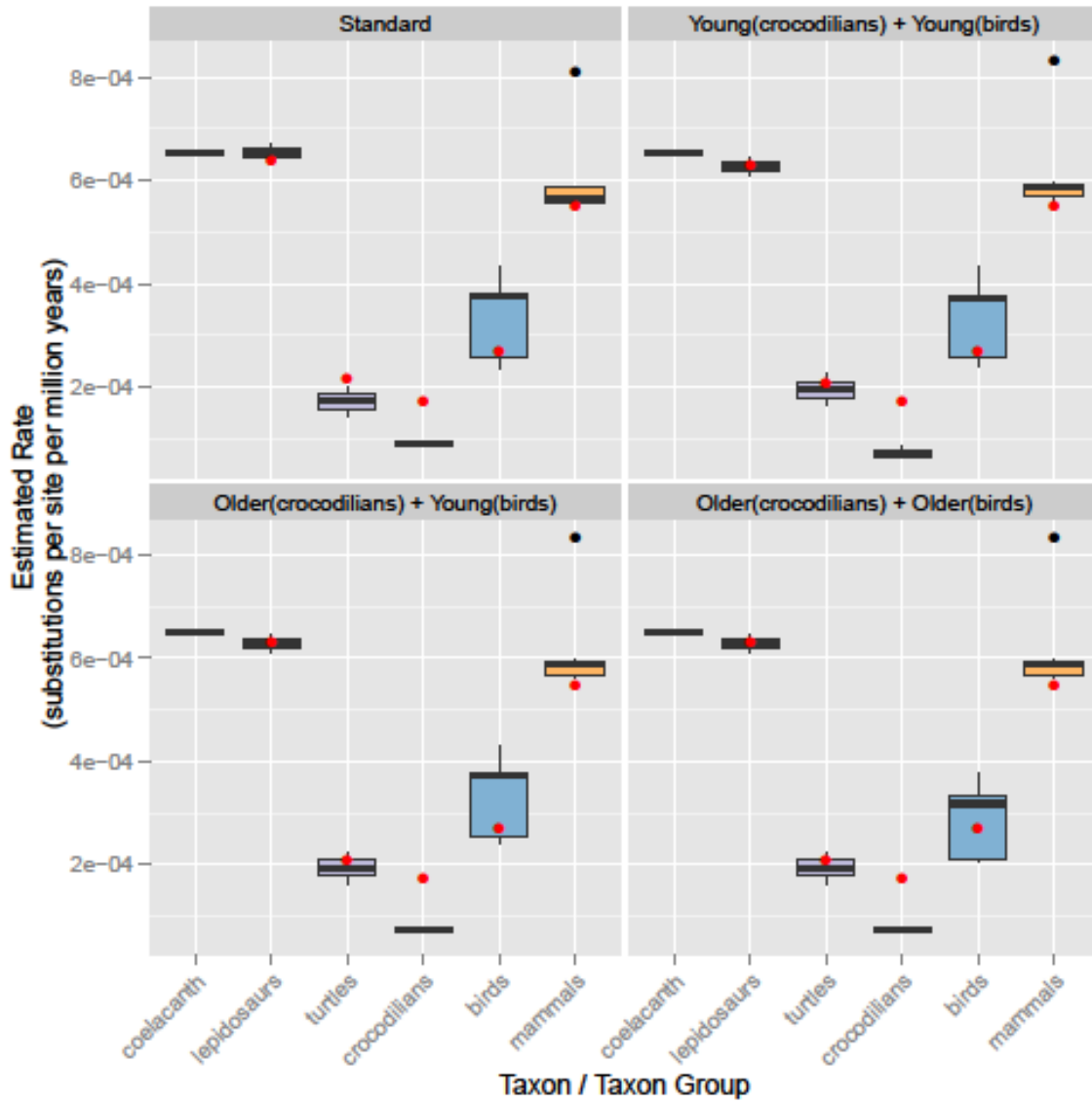
Figure S2. Rate (substitutions per site per million years) of molecular evolution across taxonomic groups comprising taxon-set-2 estimated using the topology from Figure S1 (unpartitioned data) and the four time constraint schemes detailed in SM 4. Red dots indicate the estimated rate on the common ancestor branch for each group.
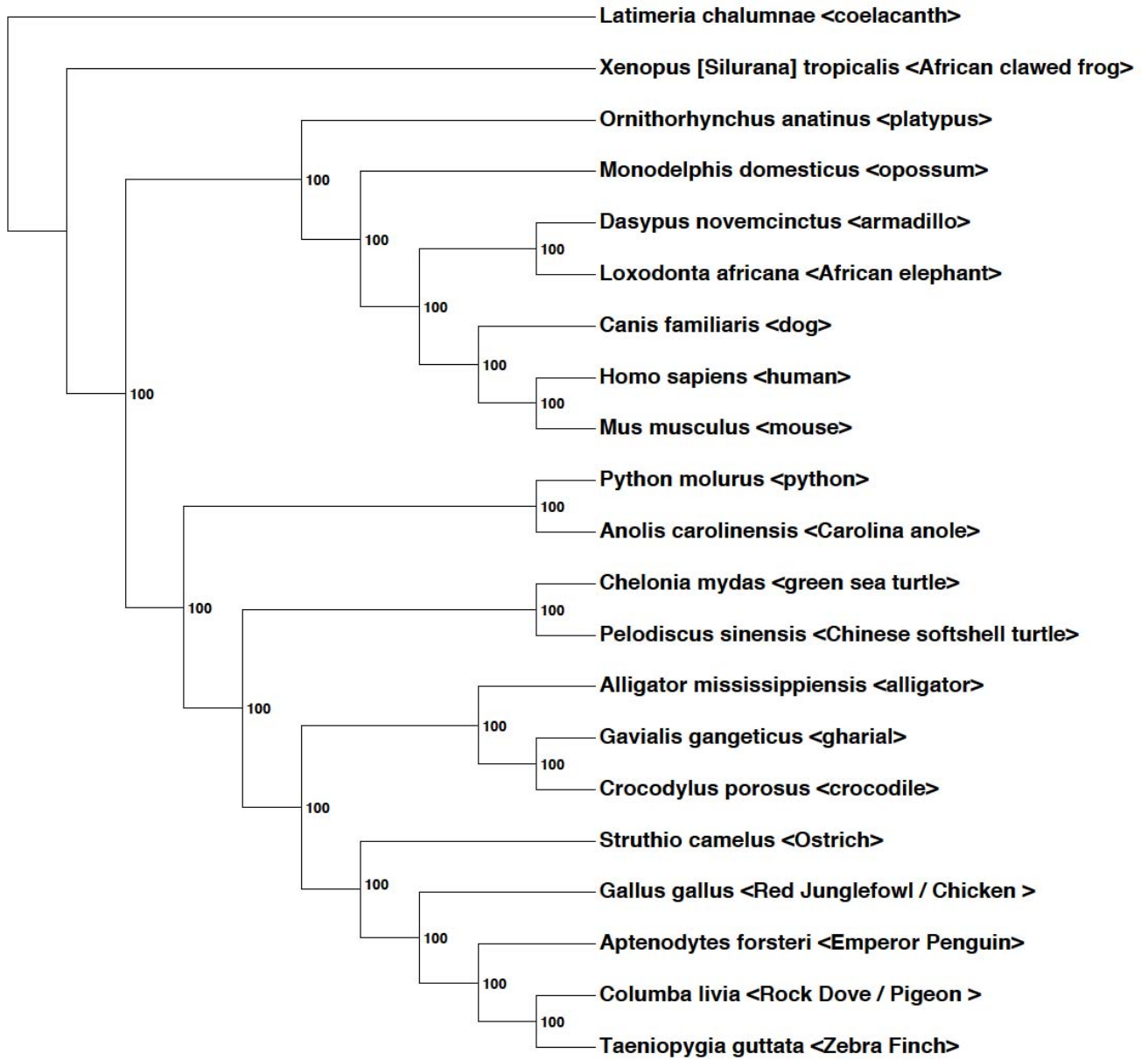
Figure S3. Inferred phylogeny of taxon-set-1 using a maximum likelihood analysis of a complete matrix of 604 ultraconserved element loci and RAxMLv7.3.4.
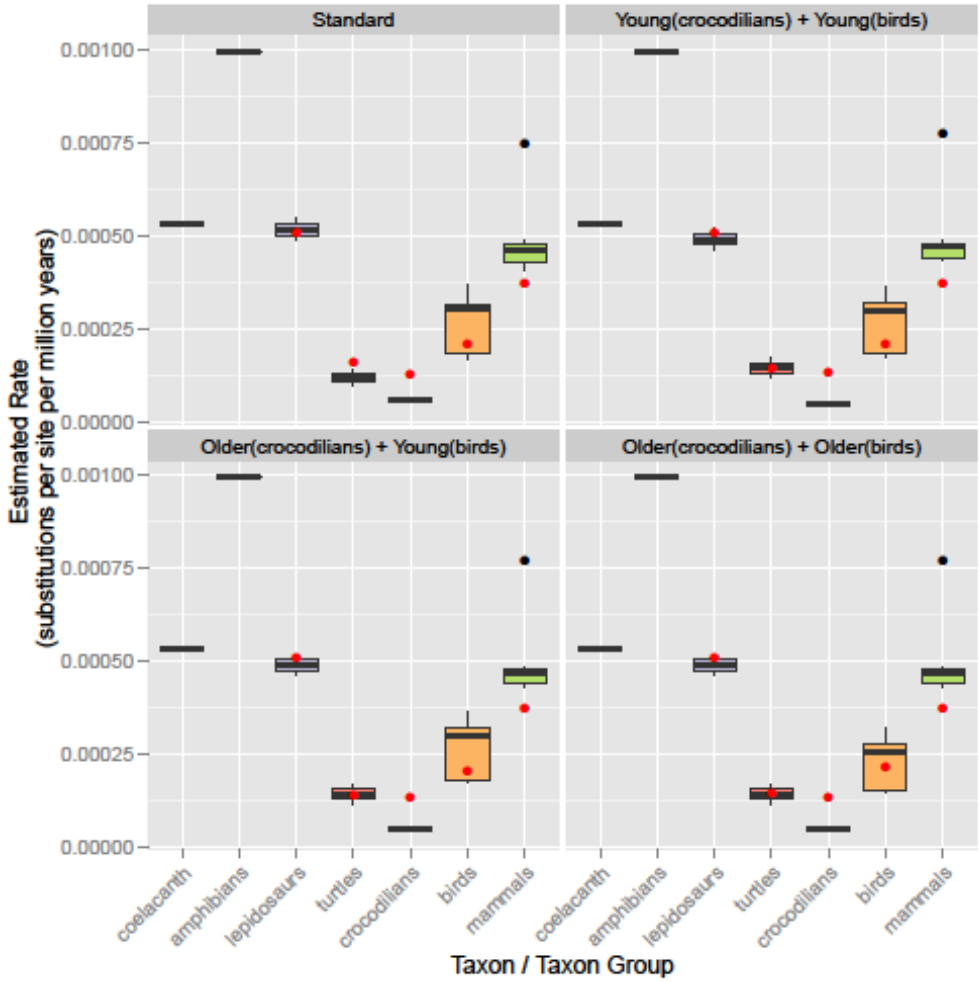
Figure S4. Rate (substitutions per site per million years) of molecular evolution across taxonomic groups comprising taxon-set-1 estimated using the topology from Figure S3 (unpartitioned data) and the four time constraint schemes detailed in SM 4. Red dots indicate the estimated rate on the common ancestor branch for each group.
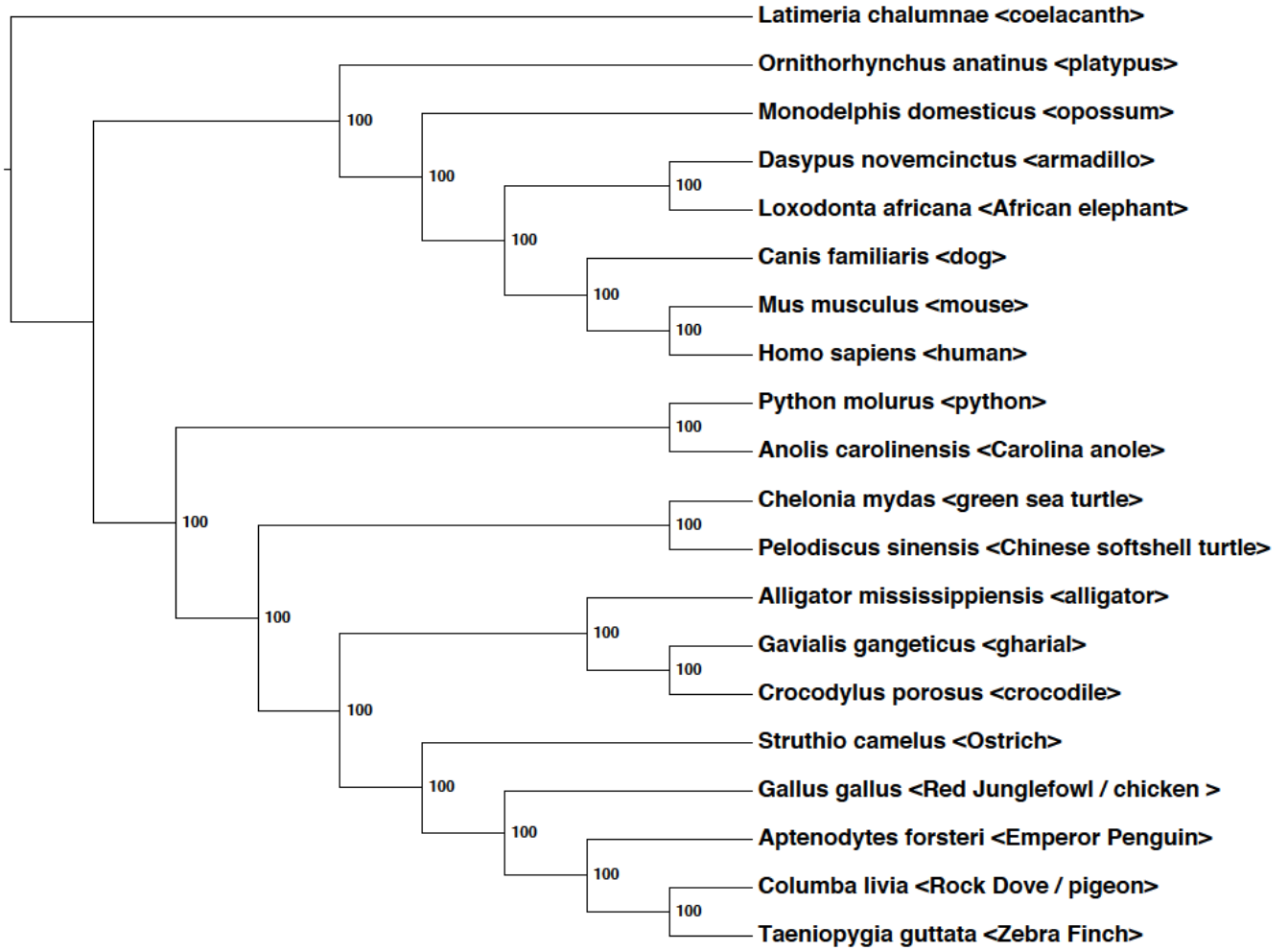
Figure S5. Inferred phylogeny of taxon-set-2 using a maximum likelihood analysis of a partitioned, complete matrix of 965 ultraconserved element loci and RAxML v7.3.4.

$$y = 0.00058 + 0.97 \cdot x, \ r^2 = 1$$
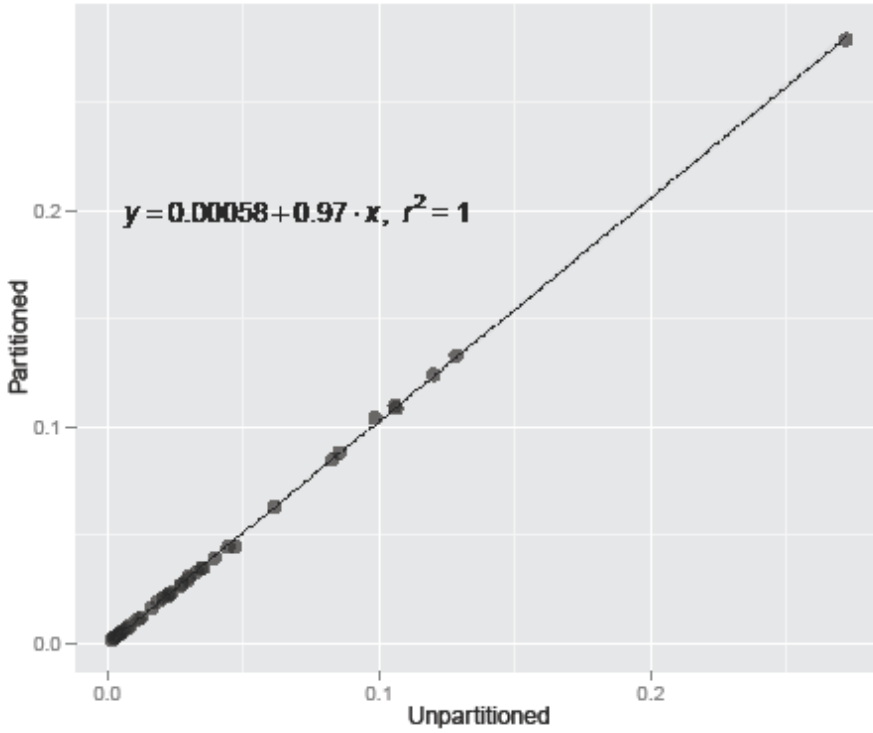
Figure S6. Branch lengths inferred from the partitioned analysis of taxon-set-2 data (complete matrix of 965 ultraconserved element loci) compared to branch lengths inferred from the unpartitioned analysis of taxon-set-2 data.
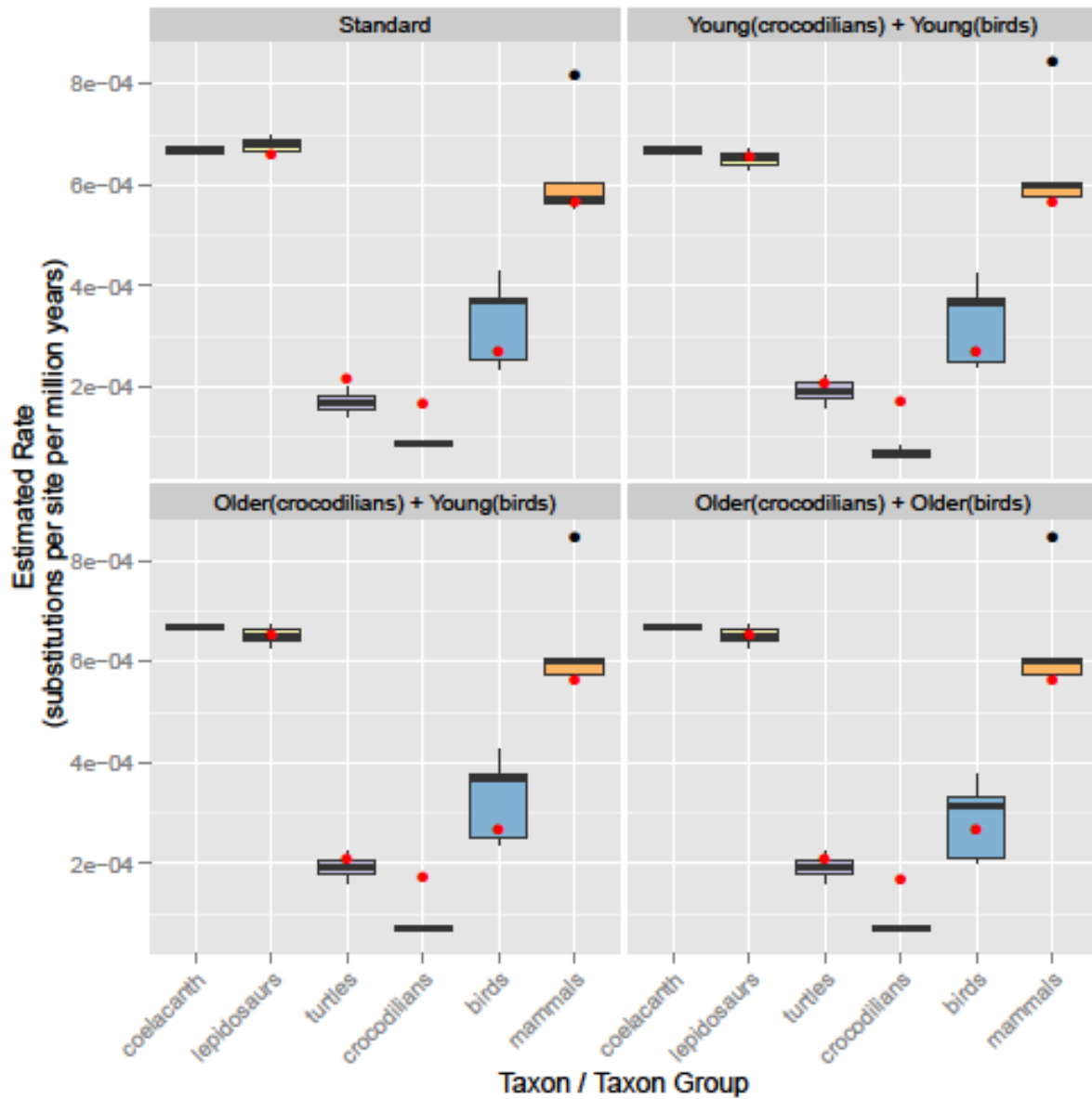
Figure S7. Rate (substitutions per site per million years) of molecular evolution across taxonomic groups comprising taxon-set-1 estimated using the topology from Figure S5 (partitioned data) and the four time constraint schemes detailed in SM 4. Red dots indicate the estimated rate on the common ancestor branch for each group.
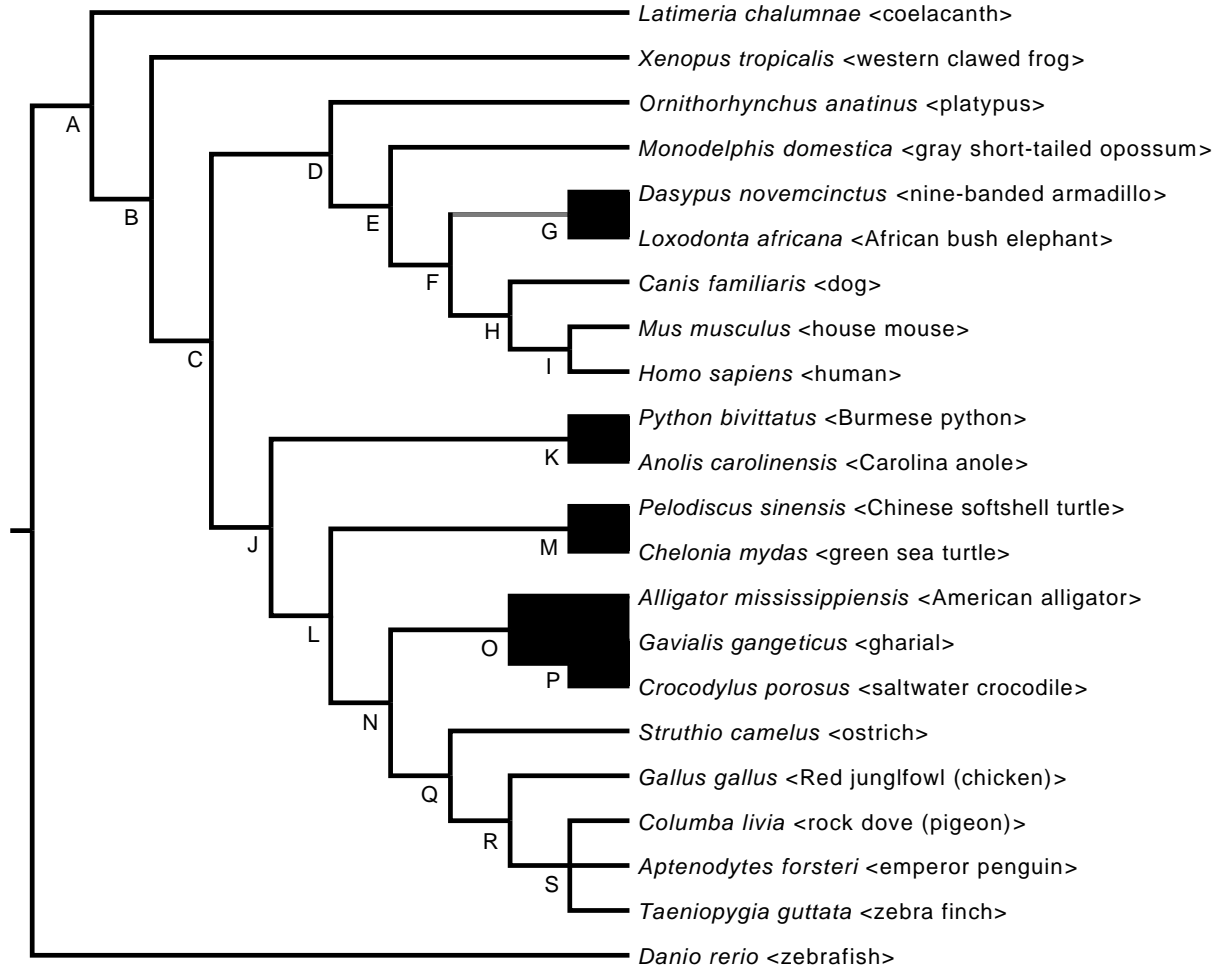
Figure S8. Consensus view of the currently understood vertebrate phylogeny with labeled nodes. Neoaves is shown as unresolved because this tree was based upon data available prior to avian whole genome phylogeny efforts (see Jarvis et al. (*76*)). Divergence times for these nodes are presented in Table S9. One node within mammals (node G, indicated in gray and uniting the putative clade Atlantogenata) remains controversial (*171*). However, this branch is short and unlikely to bias our rate estimates.
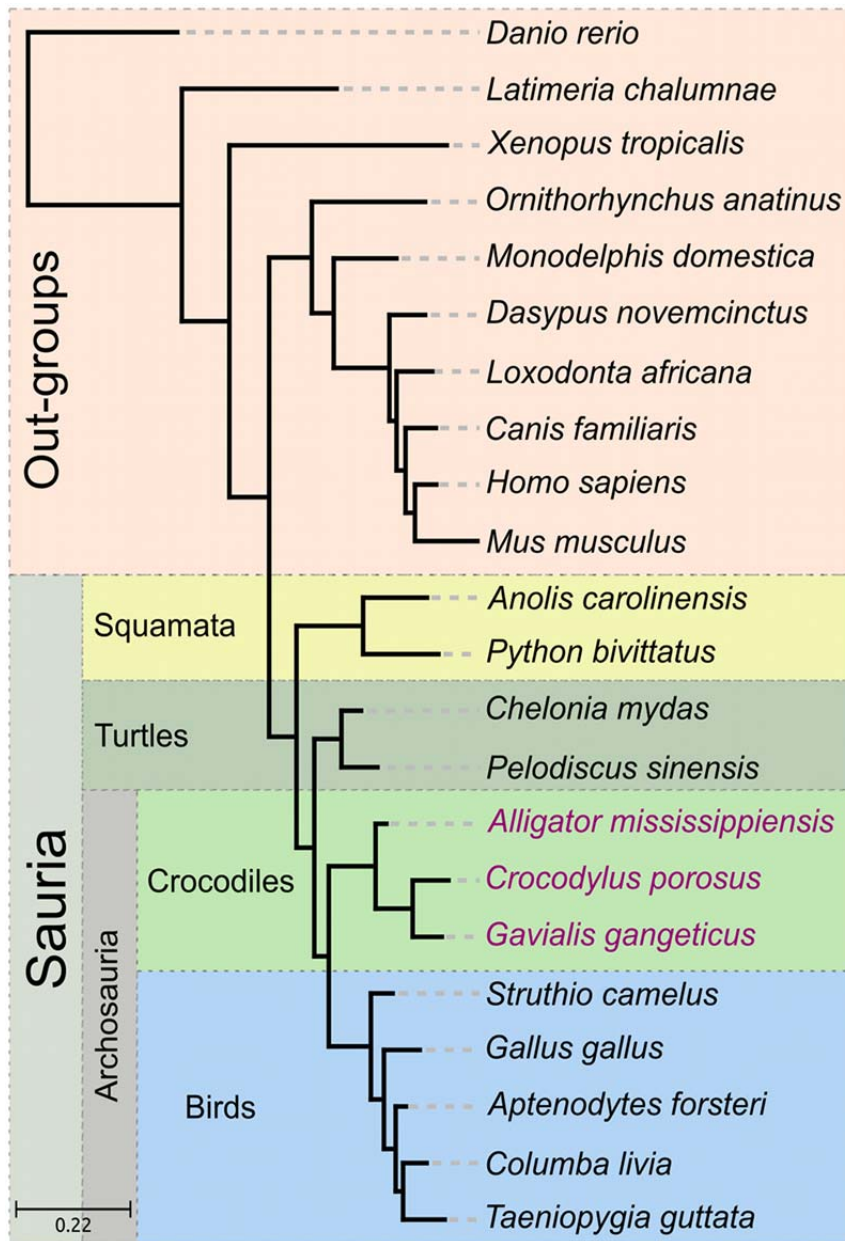
Figure S9. Maximum likelihood estimate of the species phylogeny of the 22 taxa used in the comparative genomics analyses of the three crocodile species (the names of crocodilians are in magenta). Phylogeny is based on the concatenation of 337 one-to-one orthologous genes for the species used in the study and the tree was inferred using PhyML v3.0. All aLRT branch support values are equal to 1.0 and, therefore, not shown in the tree. Different background colors correspond to taxonomic groupings present in this study.
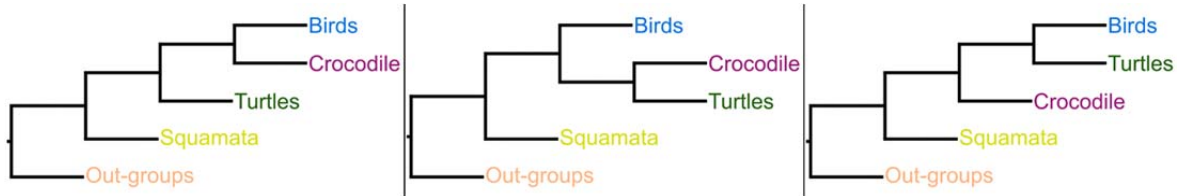
Figure S10. Representations of alternative scenarios evaluated to test the support received by the species phylogeny obtained after concatenating 337 set of marker genes. In all tests, as implemented in CONSEL, the most supported one is on the left, confirming the previously obtained result. This topology was also the most supported by individual gene trees.
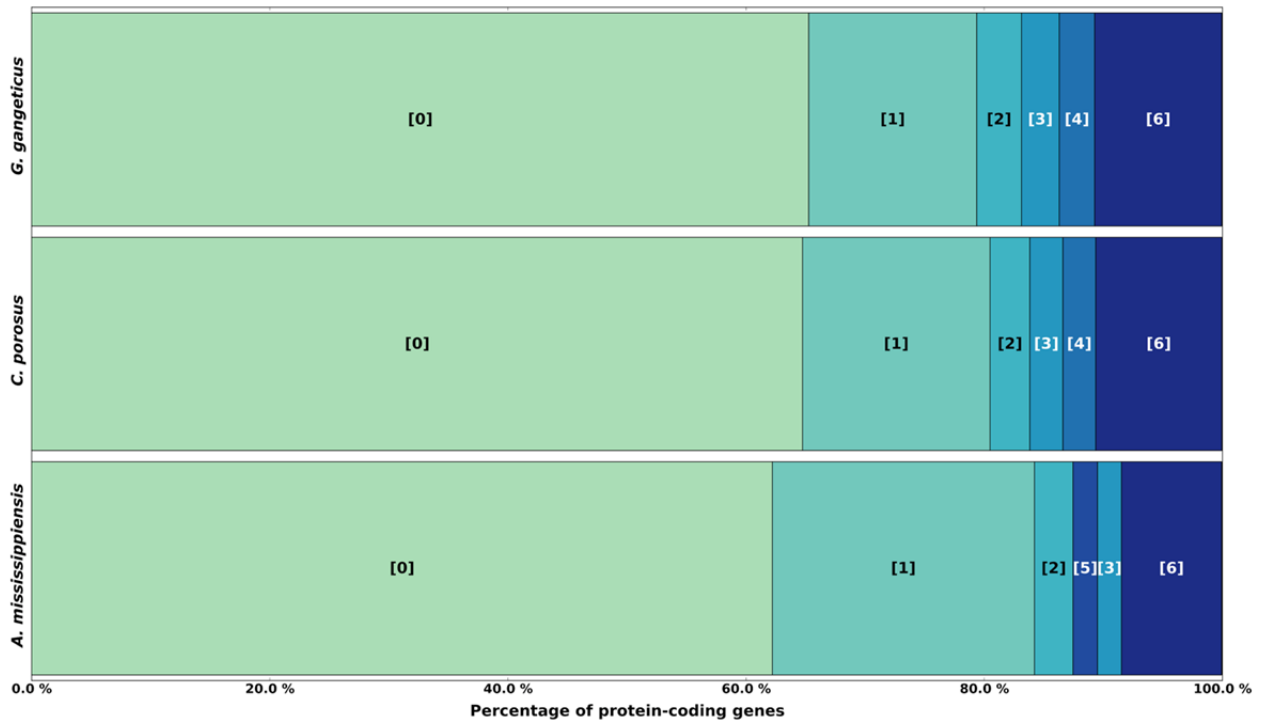
Figure S11. Comparative genomics analysis, in terms of homology relationships, for the three Crocodile species in the context of the 22 species used in this study. Different background colors indicate different taxonomic groupings used in the analysis. Labels correspond to these taxonomic groups:

[0] All main groups: Crocodiles, Birds, Turtles, Squamata, Out-groups.
[1] Crocodile-specific
[2] Crocodile,Turtles, Squamata, Out-groups
[3] Crocodile, Squamata, Out-groups
[4] Crocodile, Out-groups
[5] Crocodile, Turtles
[6] Other patterns.
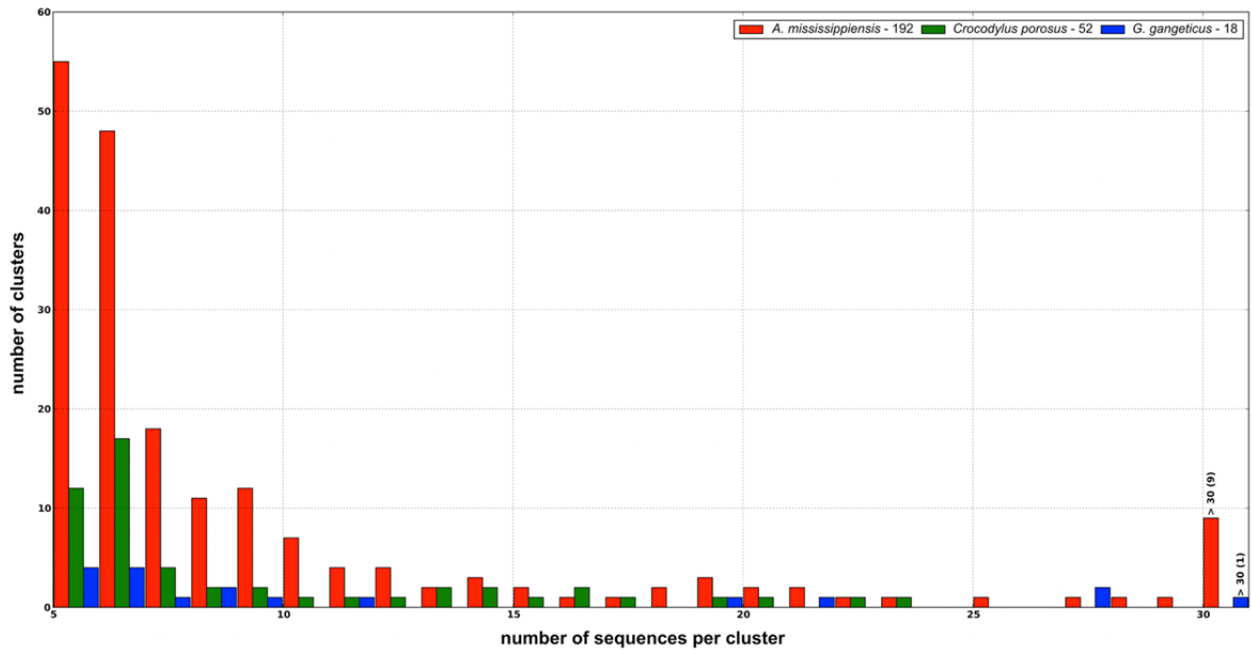
Figure S12. Clusters of specifically expanded proteins in each of the crocodilian species. Different backgrounds represent each species: *A. mississippiensis* (red), *C. porosus* (green) and *G. gangeticus* (blue). Only clusters of proteins with 5 or more members are shown in the plot.
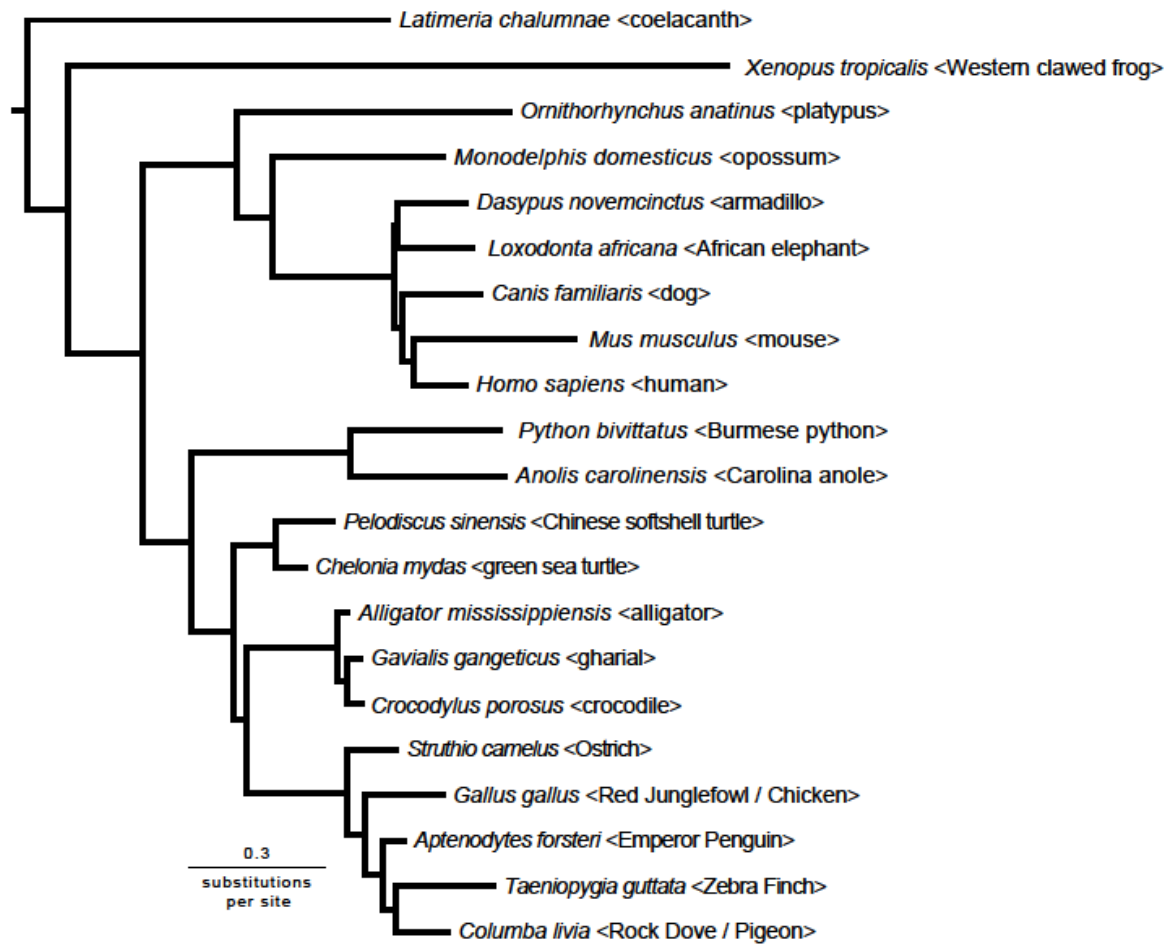
Figure S13. Maximum likelihood estimates of 4D site branch lengths for taxon-set-1 based upon filtering strategy 1 and the topology of Figure 1 (main manuscript). This tree was rooted to zebrafish (*Danio rerio*) and the branch length estimates were obtained using the GTRGAMMA model in RAxML. The tree length was 10.089706 (including the zebrafish outgroup). Parameter estimates: Γ distribution shape parameter [α] = 3.664242; GTR rate parameters, normalized to the G-T rate: A-C = 0.796327; A-G = 3.111488; A-T = 0.793795; C-G = 1.201402; C-T = 1.822809; equilibrium base frequencies π_A = 0.270879; π_C = 0.268545; π_G = 0.157921; π_T = 0.302655).
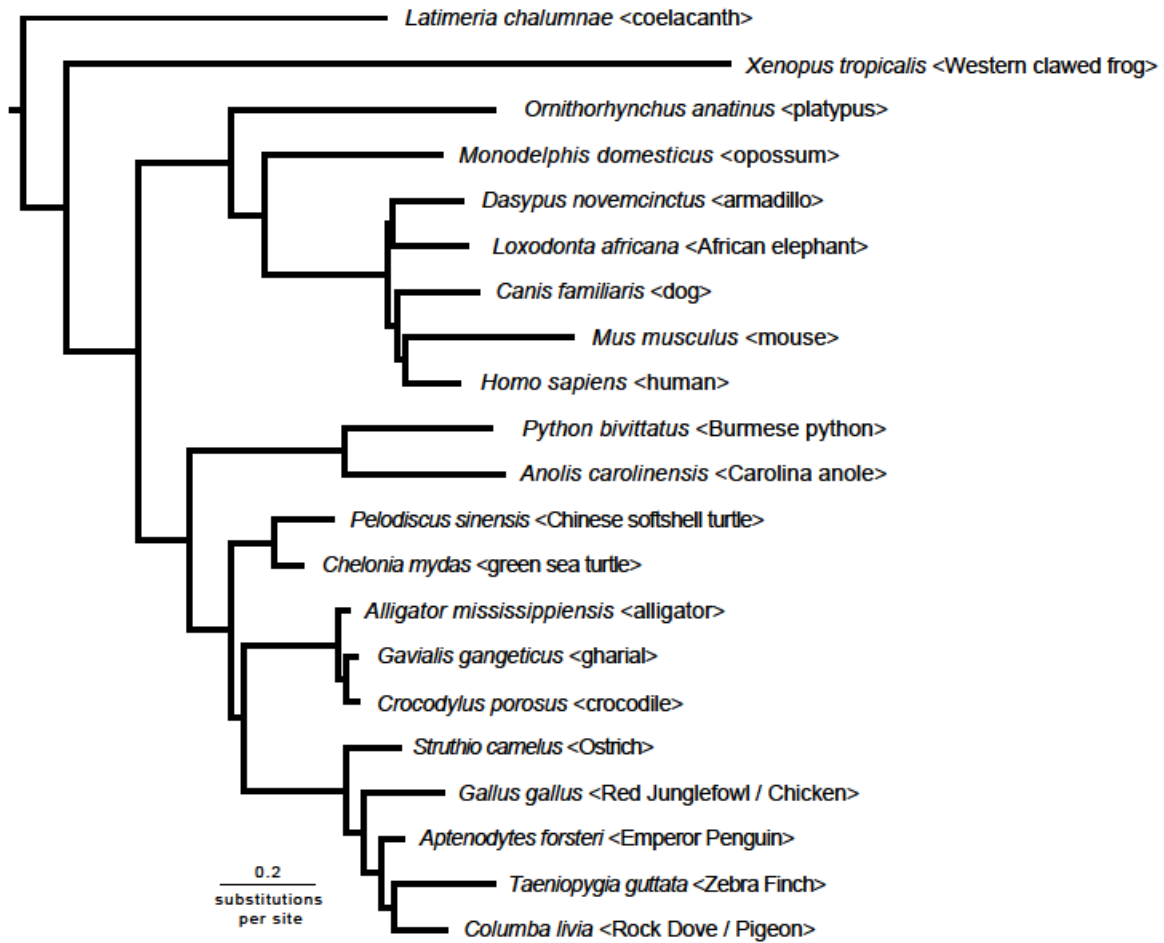
Figure S14. Maximum likelihood estimates of 4D site branch lengths for taxon-set-1 based upon filtering strategy 2 and the topology of Figure 1 (main manuscript). As for Figure S13, this tree was rooted to zebrafish and the GTRGAMMA model was used for analyses in RAxML. The tree length was 9.506303 (including the zebrafish outgroup). Parameter estimates: Γ distribution shape parameter [α] = 3.744828; GTR rate parameters, normalized to the G-T rate: A-C = 0.833754; A-G = 3.025612; A-T = 0.839196; C-G = 1.175113; C-T = 1.916265; equilibrium base frequencies π_A = 0.264055; π_C = 0.268591; π_G = 0.170805; π_T = 0.296549).
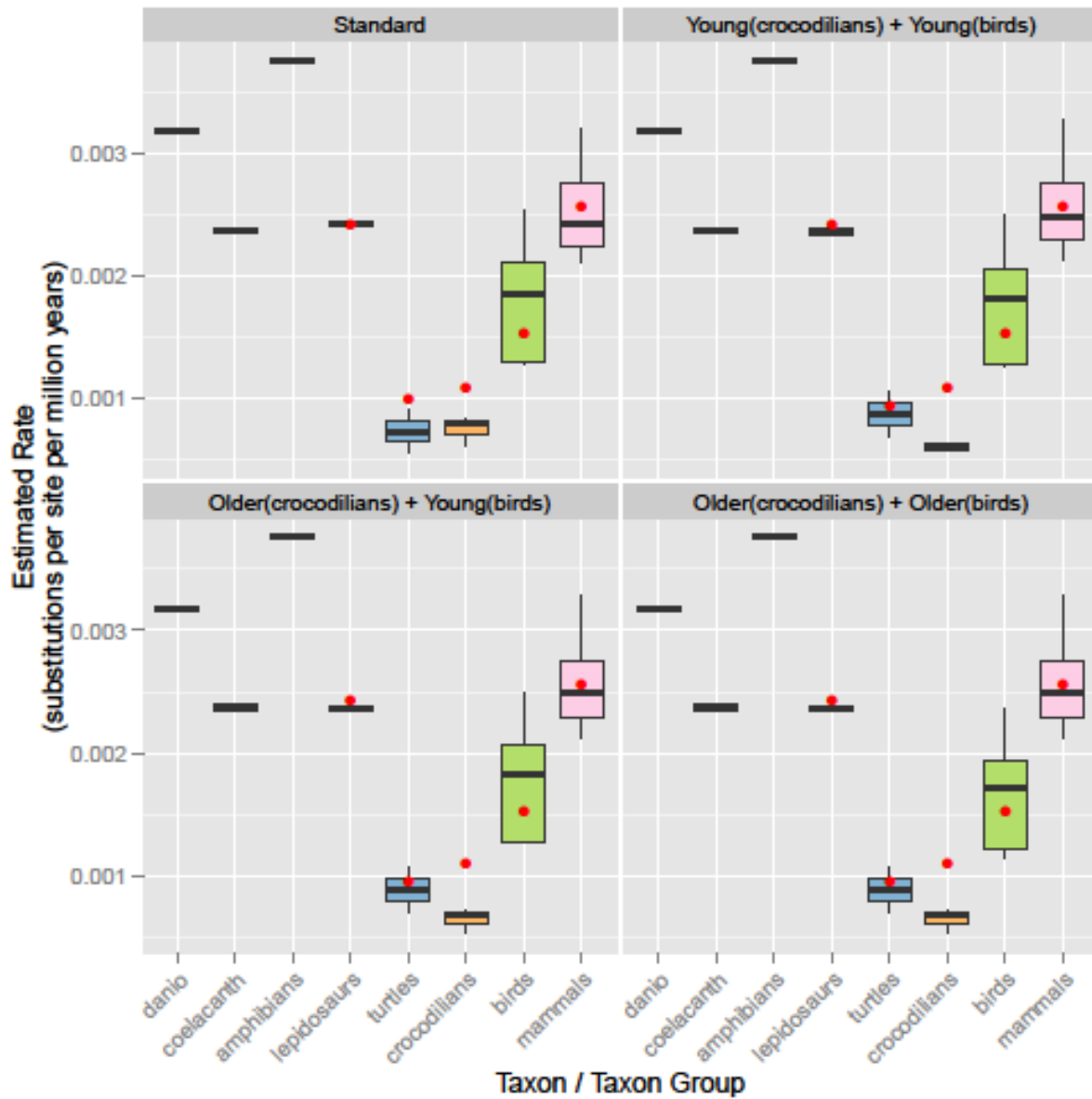
Figure S15. Estimated substitution rates for 4D sites based upon filtering strategy 1. These analyses used the tree and branch lengths shown in Figure S13. Red dots indicate the estimated rate on the common ancestor branch for each group.
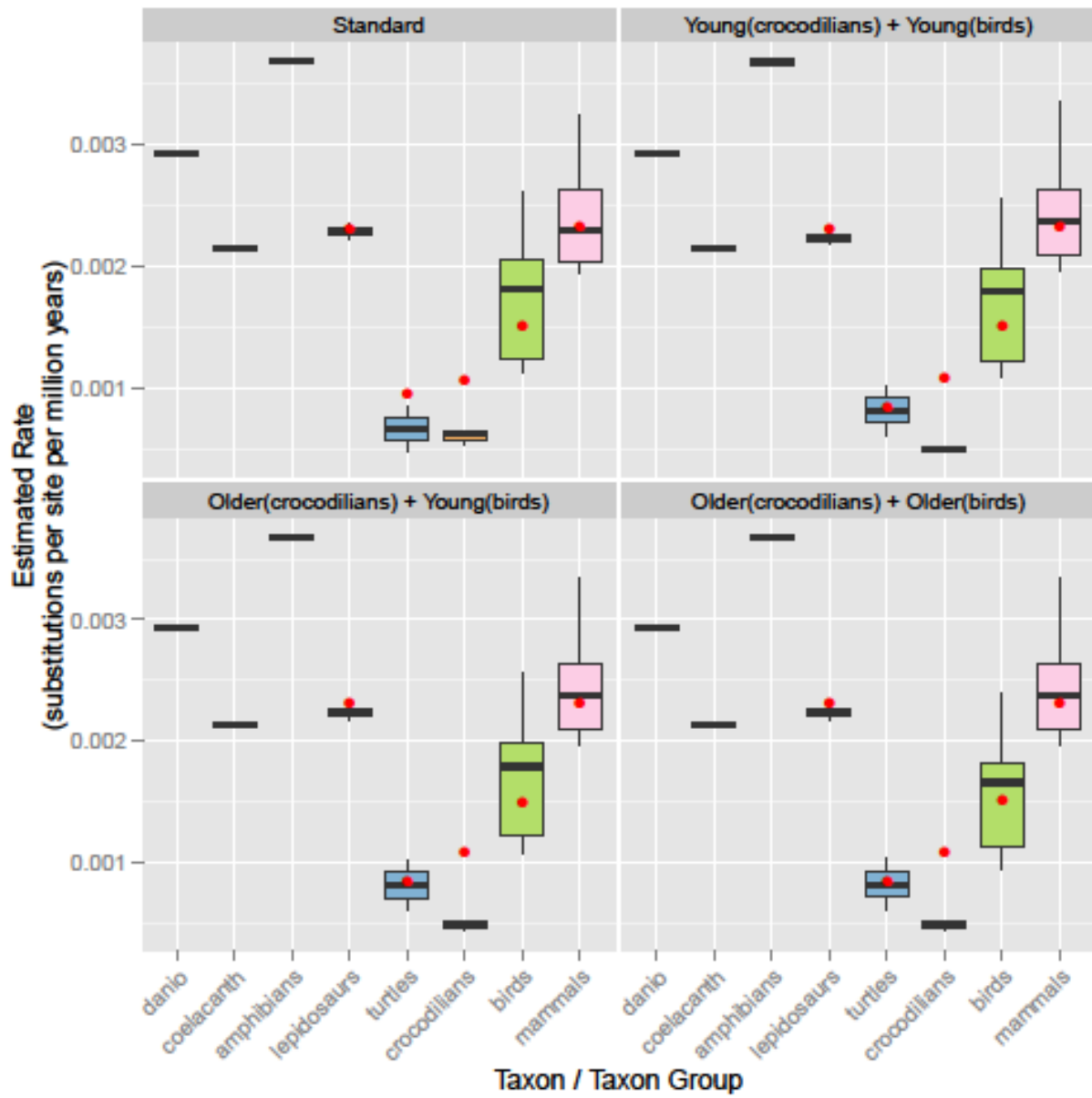
Figure S16. Estimated substitution rates for 4D sites lengths based upon filtering strategy 2. These analyses used the tree and branch lengths shown in Figure S14. Red dots indicate the estimated rate on the common ancestor branch for each group.
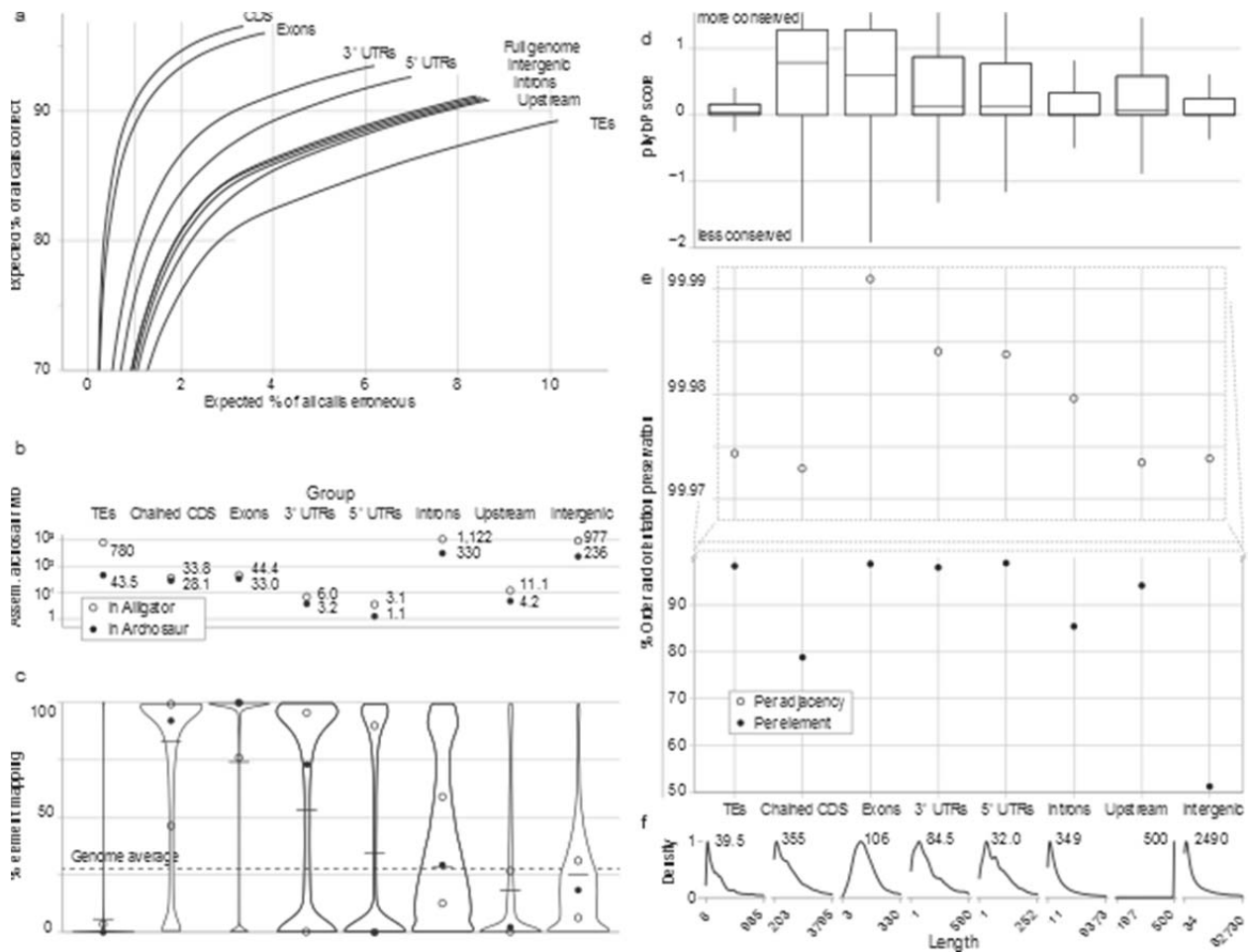
Figure S17. Analyzing the archosaur assembly using projected alligator annotations. a) Expected base reconstruction accuracy. b) Total archosaur bases assembled. c) Proportion of annotations mapping as violin plots, with box plot (circles) and average (line) inset. d) Conservation scores (phyloP). e) Order and orientation preservation of annotated elements. f) For comparison, length distributions of annotations on alligator. Note: 'Chained CDS' category includes complete CDS, with introns spliced out. This is expansion of the analysis presented in Figure 7 of the main text.
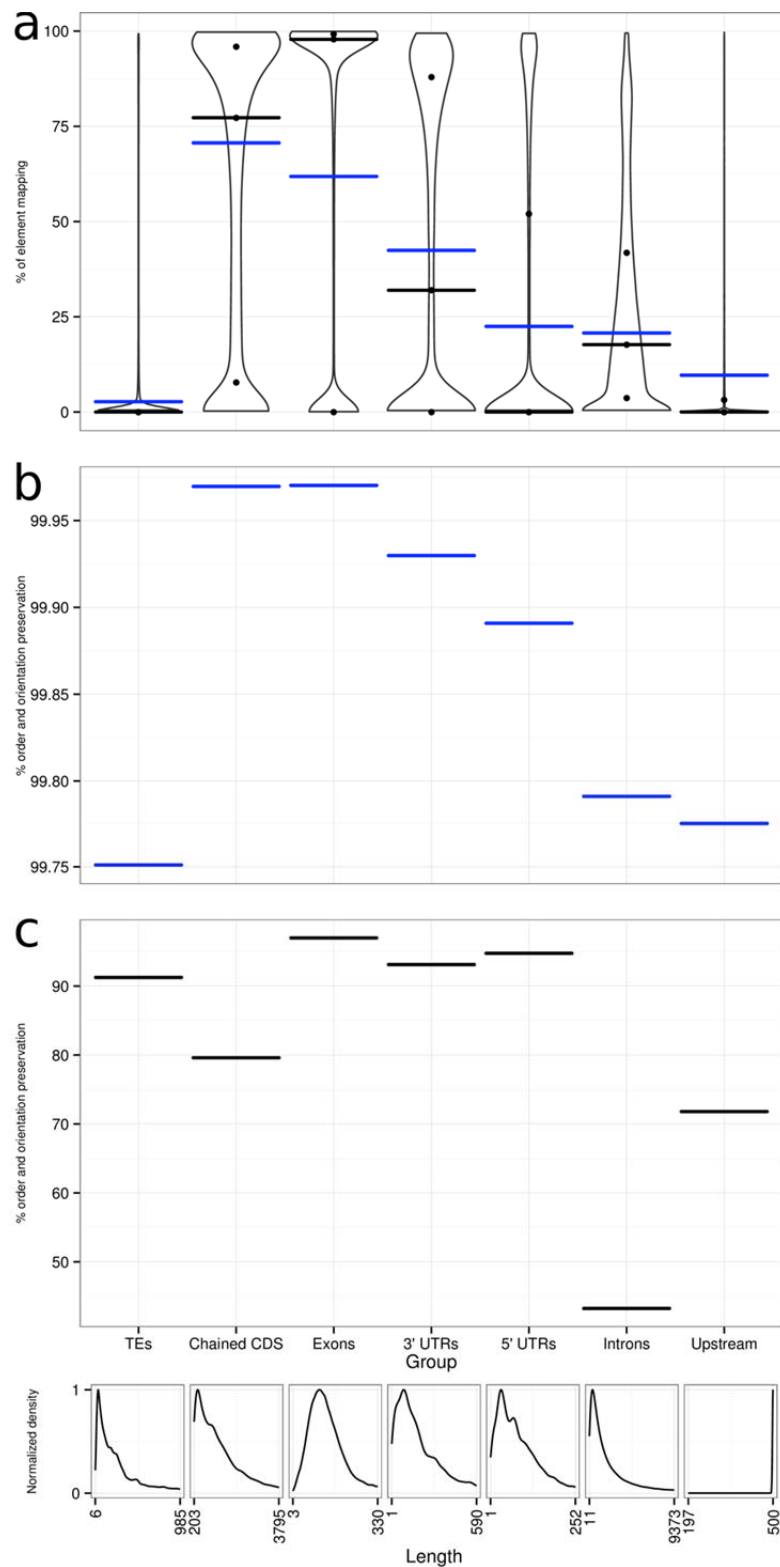
Figure S18. Mapping and order-and-orientation-preserving statistics from alligator to chicken in the alignment. A) Violin plot for the percent of an element that maps to the target genome. Blue lines represent the average mapping for the category, and dots show inner quartiles. B) Percent of adjacencies preserved. C) Percent of elements preserved.
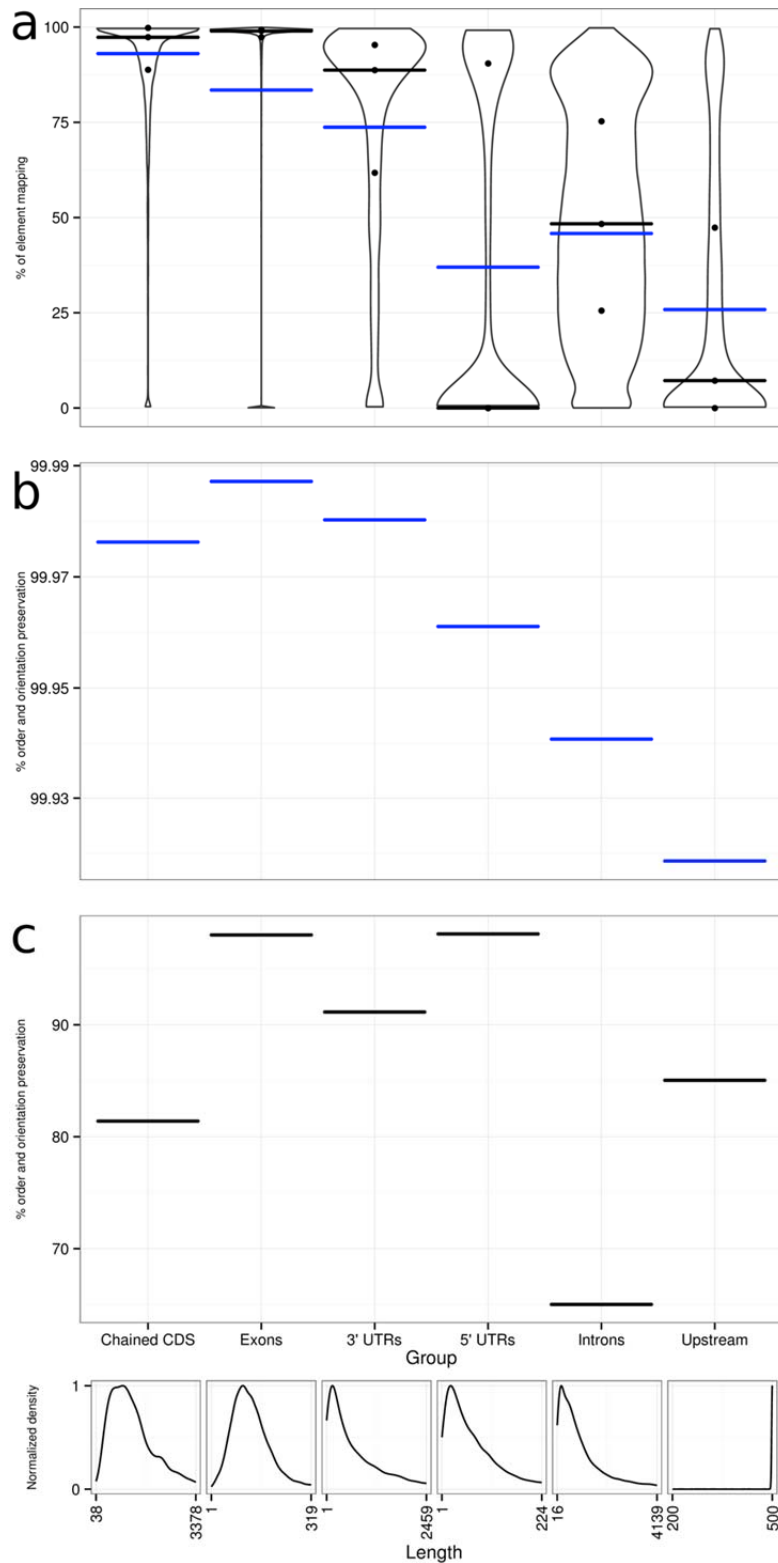
Figure S19. Mapping and order-and-orientation-preserving statistics from chicken to archosaur in the alignment. A) Violin plot for the percent of an element that maps to the target genome. Blue lines represent the average mapping for the category, and dots show inner quartiles. B) Percent of adjacencies preserved. C) Percent of elements preserved.
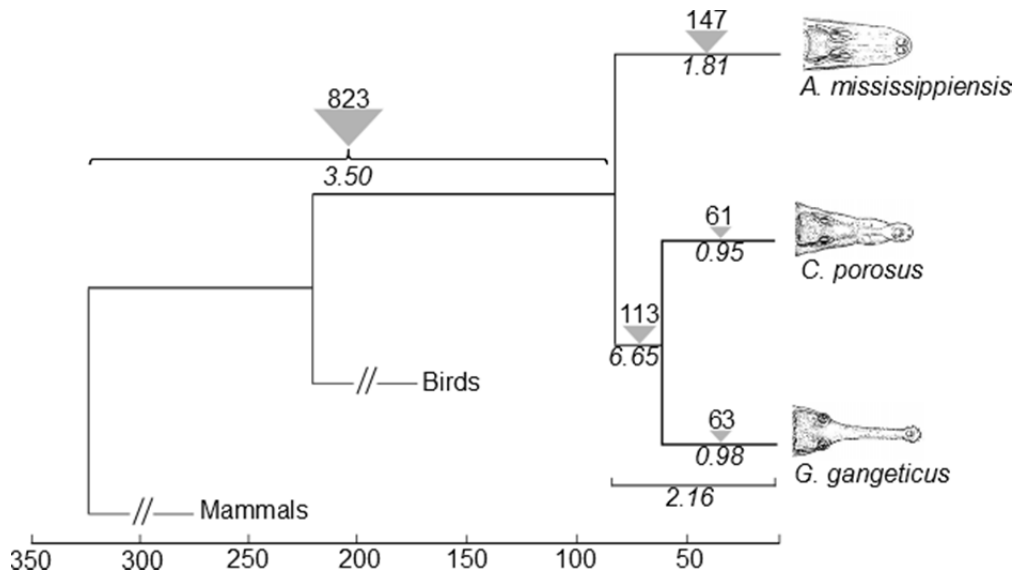
Figure S20. Rates of TE family accumulation along each lineage of this representation of the crocodilian tree.
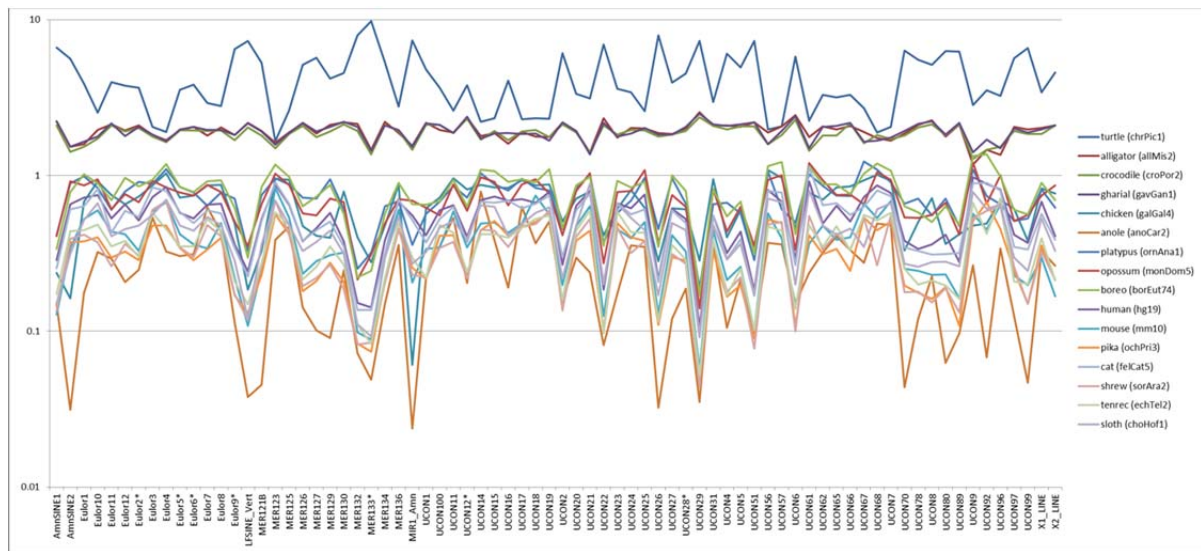
Figure S21. Relative visibility of TE copies that predate the speciation of crocodilians and mammals in 16 amniote genomes. The figure displays, on a log scale and for 74 unrelated TE families present in all amniote genomes, the bases identified in each individual genome relative to the average identified in all 16 genomes.
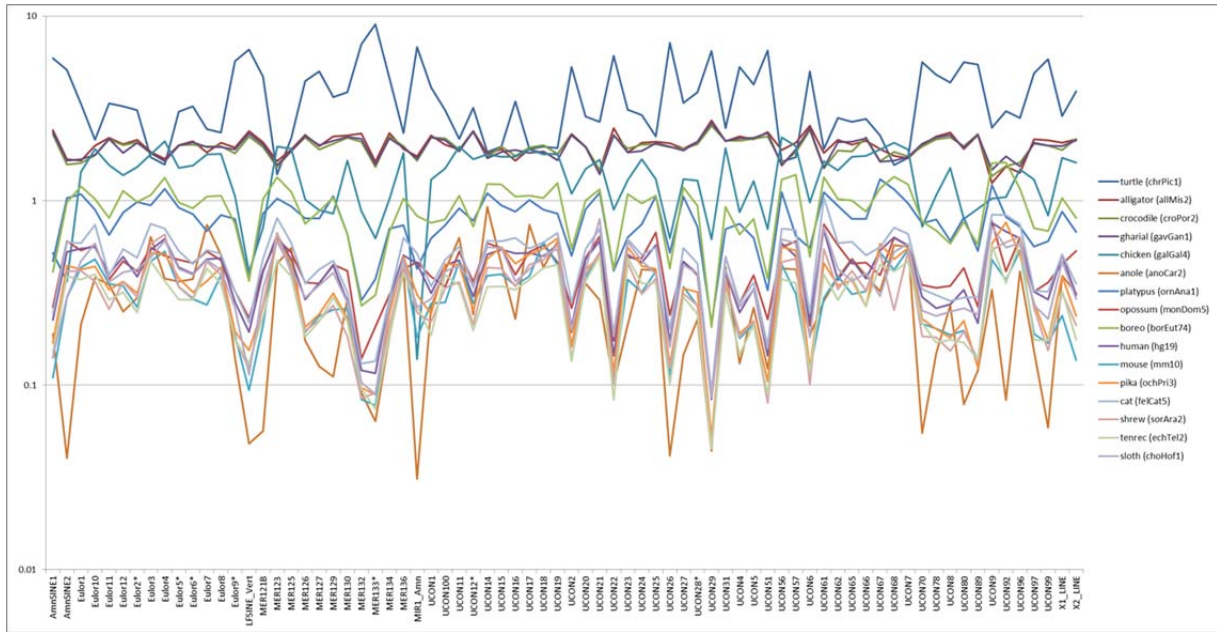
Figure S22. TE visibility as determined by the relative density (bp of repeat / genome size). See text for a full explanation.
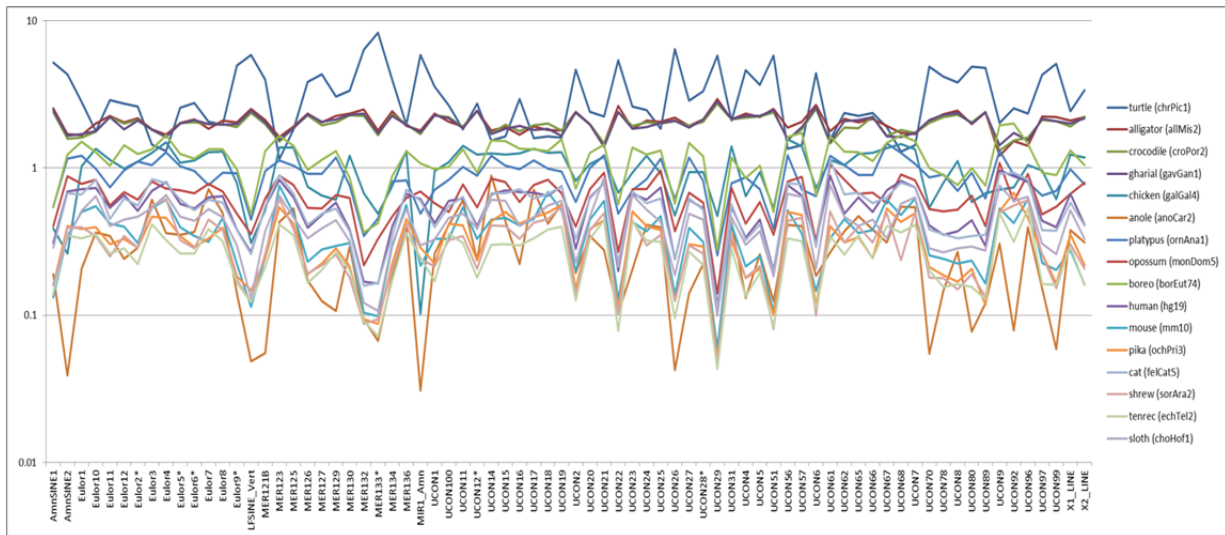
Figure S23. TE visibility as determined by the relative density of repeats using a genome size excluding the lineage-specific repeats.

48
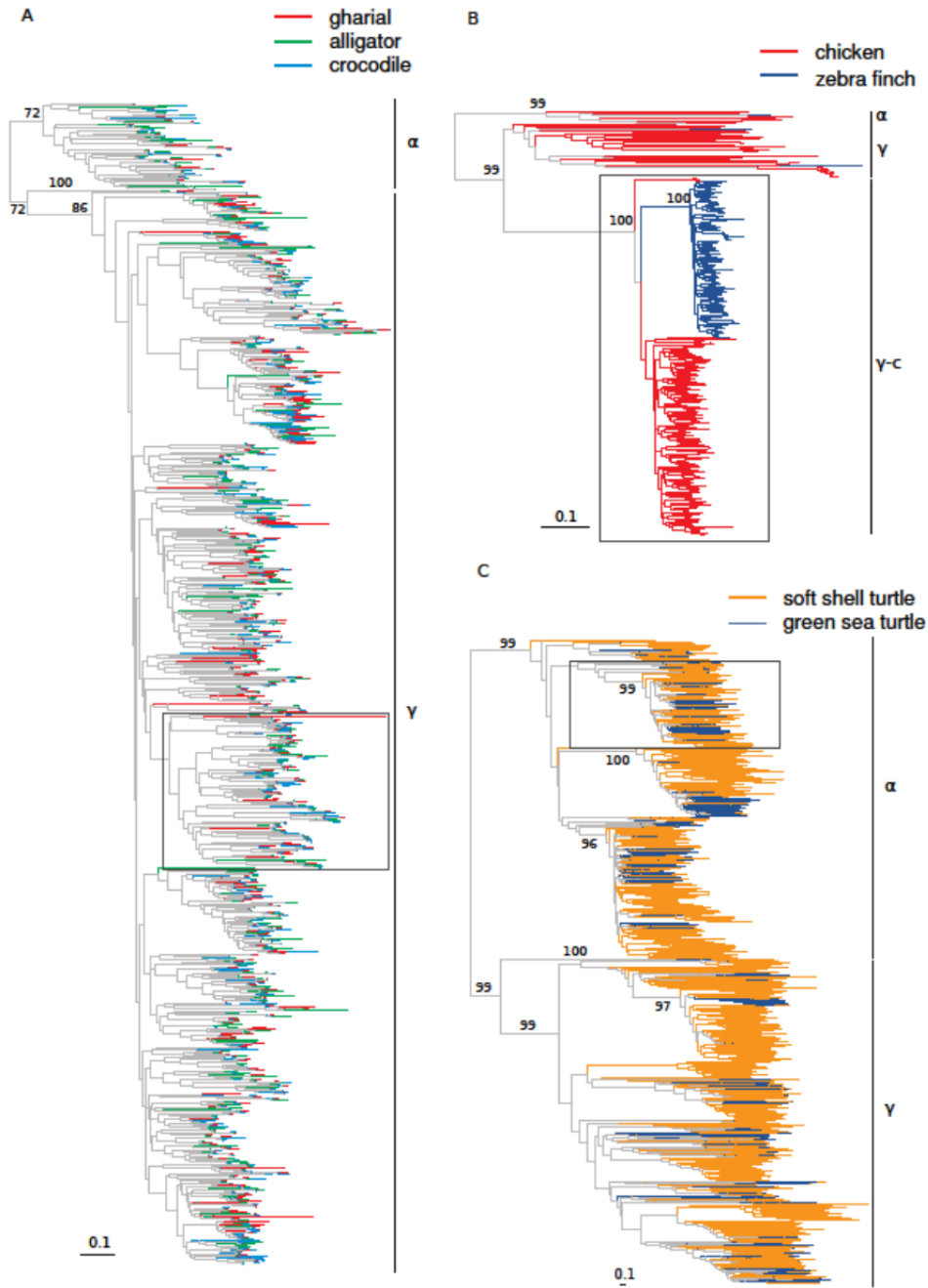
Figure S24. Neighbor-joining phylogenies of the intact crocodilian (A), avian (B) and testudine (C) OR repertoires. Crocodiles are represented by the saltwater crocodile, gharial and American alligator; birds are represented by chicken and zebra finch; and testudines are represented by soft shell and green sea turtle. Note the higher frequency of lineage-specific OR expansions in panels B and C relative to A. Rectangles correspond to the subtrees in the main manuscript. Trees were inferred using MEGA v5, a Poisson model of substitution, and 1000 bootstrap iterations were performed to evaluate support.
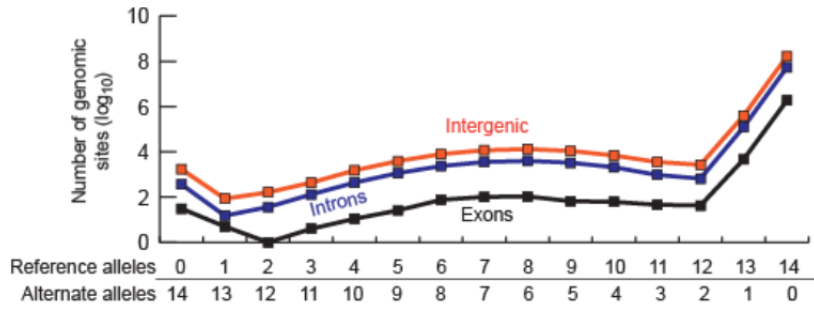
Figure S25. – Distribution of counts for reference and alternate allele at positions. After mapping 14-fold average genomic coverage of reads to the reference crocodile genome assembly, we analysed sites at exactly 14-fold coverage. We counted the number of of reference and alternative bases of high quality (Phred score >= 30) only on uniquely mapping reads (map quality >=10). As expected, for most positions, all 14 bases are the reference base. Heterozygous positions should have an equal number of an equivalent base. Reference mapping bias is evident in that the central distribution is not centered on seven reference and seven alternate alleles.
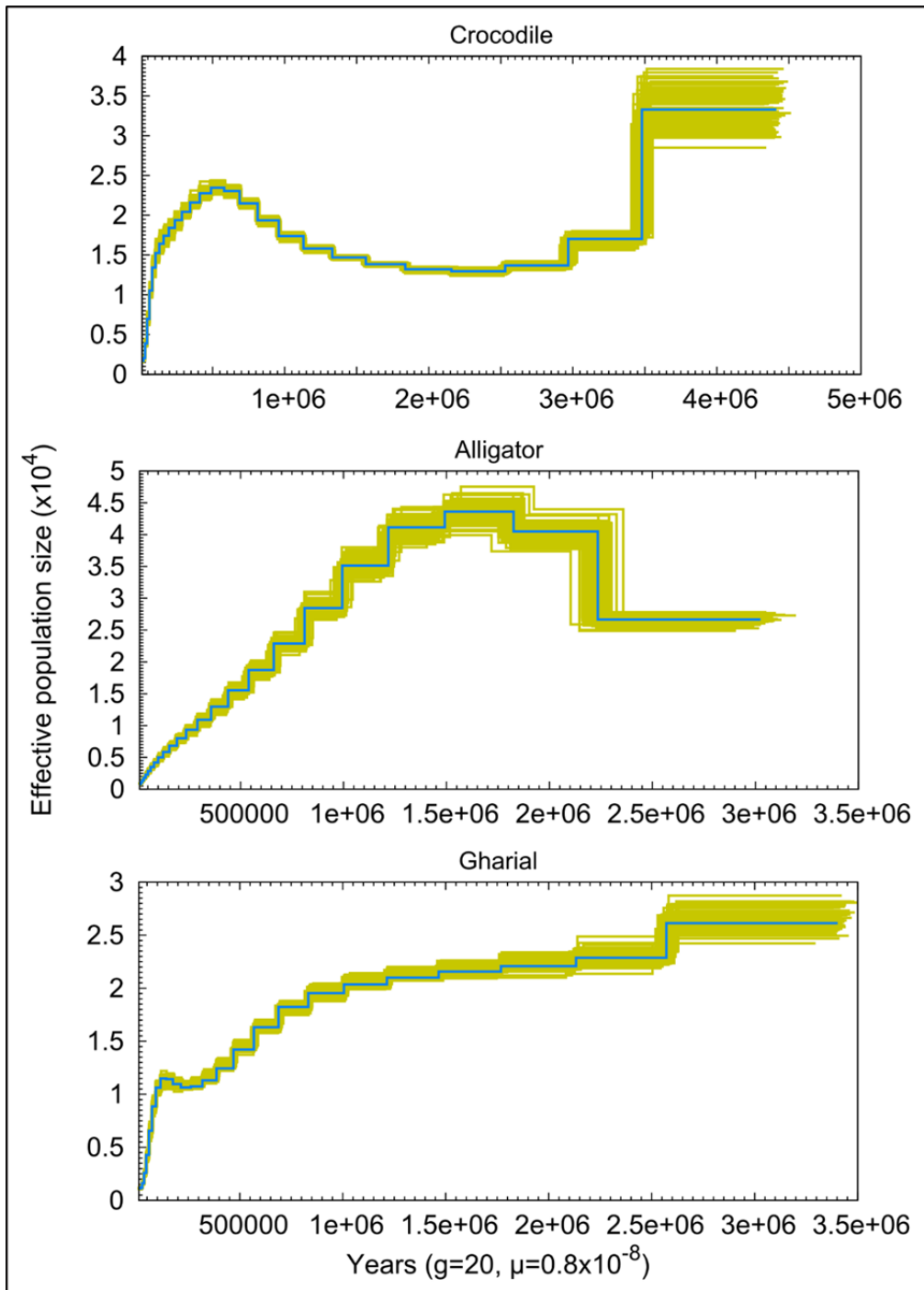
Figure S26. PSMC estimates from each genome. The thin green curves in each graph are estimates for 100 bootstrap replicates from the complete sequence data used.
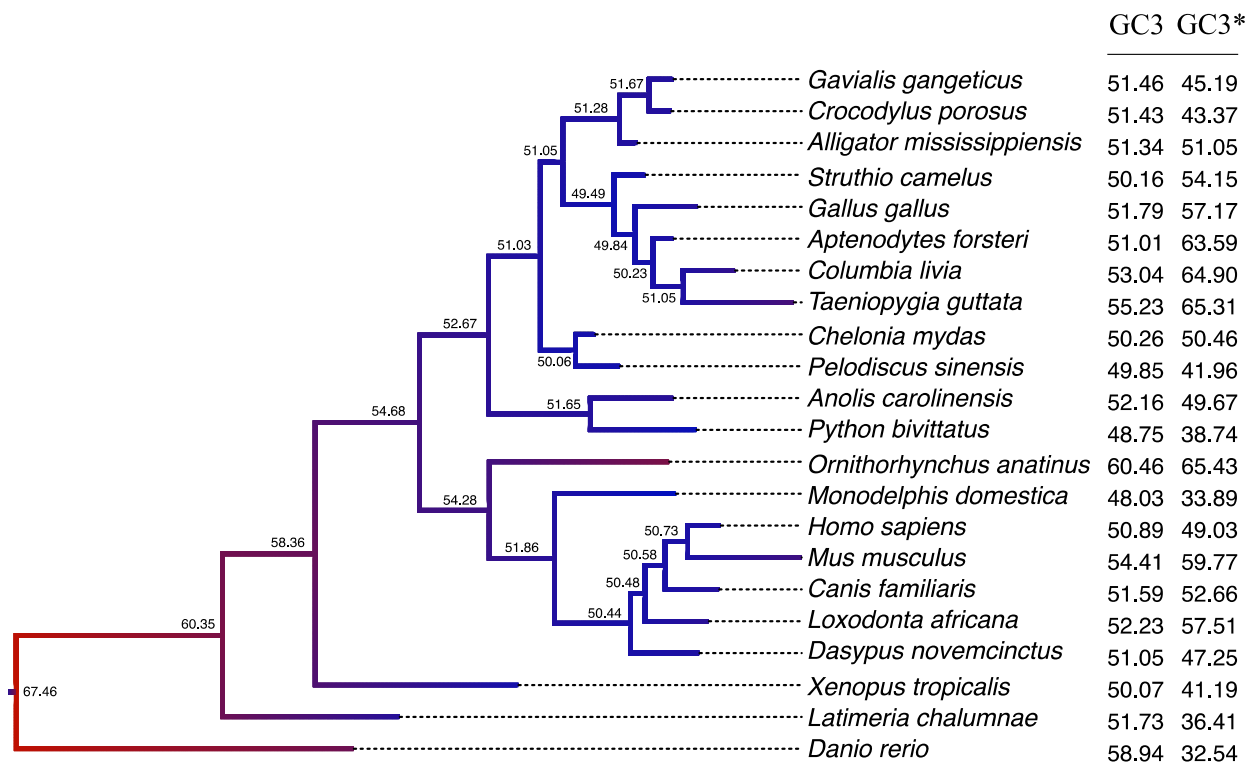
	GC3	GC3*
Gavialis gangeticus	51.46	45.19
Crocodylus porosus	51.43	43.37
Alligator mississippiensis	51.34	51.05
Struthio camelus	50.16	54.15
Gallus gallus	51.79	57.17
Aptenodytes forsteri	51.01	63.59
Columbia livia	53.04	64.90
Taeniopygia guttata	55.23	65.31
Chelonia mydas	50.26	50.46
Pelodiscus sinensis	49.85	41.96
Anolis carolinensis	52.16	49.67
Python bivittatus	48.75	38.74
Ornithorhynchus anatinus	60.46	65.43
Monodelphis domestica	48.03	33.89
Homo sapiens	50.89	49.03
Mus musculus	54.41	59.77
Canis familiaris	51.59	52.66
Loxodonta africana	52.23	57.51
Dasypus novemcinctus	51.05	47.25
Xenopus tropicalis	50.07	41.19
Latimeria chalumnae	51.73	36.41
Danio rerio	58.94	32.54

Figure S27. Evolution of GC3 across vertebrates overlayed on the tree inferred in Figure S9. The branch lengths represent changes in GC3 (red = high GC3, blue = low GC3). Branch lengths are proportional to divergence in GC3.
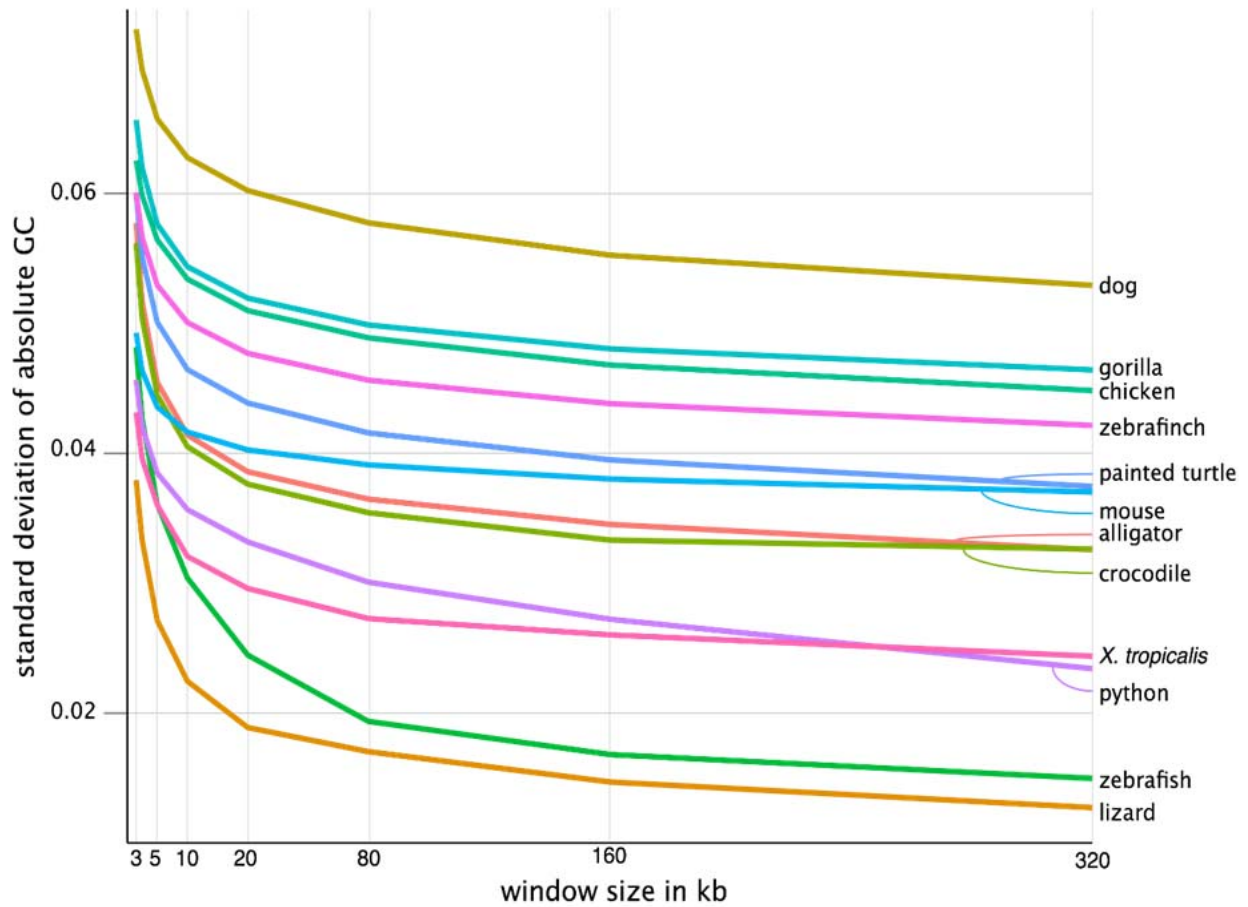
Figure S28. Variation in GC content declines at different spatial scales based on sliding window analysis of different window sizes. The Y-axis is standard deviation of GC content, and the X-axis is window size. Much of the GC content heterogeneity in crocodilians occurs at smaller spatial scales: alligator and crocodile resemble birds/turtle at smaller window sizes, but the decline of GC content variation for larger window sizes indicates a more homogeneous genome on the scale of isochores.
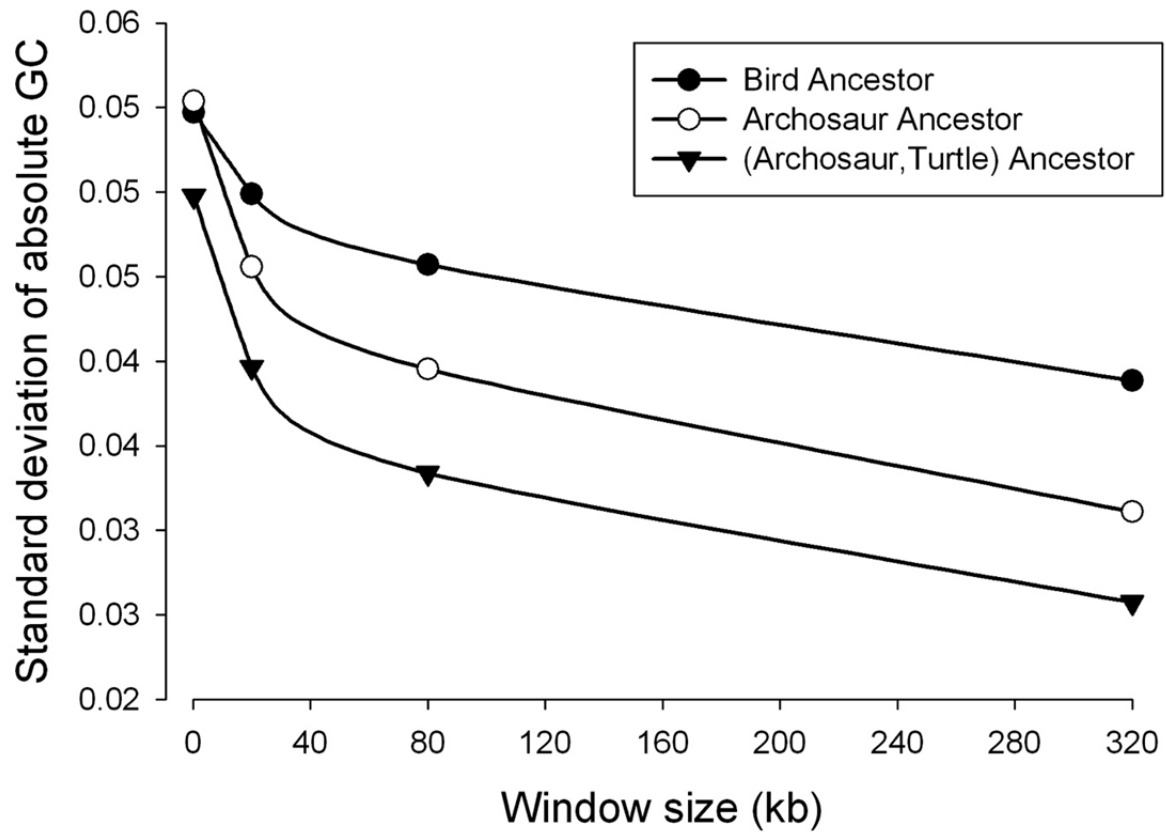
Figure S29. Standard deviation of absolute GC content as a function of window size in three ancestral reconstructed genomes.

Supplemental Tables:

Table S1. Alligator sequencing libraries

Library type	Illumina read	Physical coverage
180 paired end	2x100	5.5 x
200 paired end	2x150	21.6 x
320 paired end	2x100	33.9 x
1.6 kb mate pair	2x50	88.7 x
3 kb mate pair	2x50	6.5 x

Table S2. Crocodile sequencing libraries.

Library type	Illumina read	Physical coverage
167 paired end	2x100	21.6 x
370 paired end	2x100	54.6 x
1.8 kb mate pair	2x50	90.2 x

Table S3. Gharial sequencing libraries

Library type	Illumina read	Physical coverage
187 paired end	2x100	50 x
401 paired end	2x100	50 x
700 paired end	2x100	80 x
1.5 kb mate pair	2x50	170 x

Table S4. Genome assembly statistics

Species	Assembled genome size (Mb)	Scaffold N50 (kb)
American alligator	2,174.3	509.0
Saltwater crocodile	2,123.5	205.0
Indian gharial	2,882.7	127.6

Table S5. The number of functionally annotated proteins for each of the three crocodilian species. Proteins were annotated to GO Consortium guidelines using InterPro ("Inferred from Electronic" or IEA), Blast ("Inferred from Sequence Alignment" or ISA) and orthology to annotated proteins ("Inferred from Sequence Orthology" or ISO) to assign GO. The results were manually checked and merged to obtain a comprehensive, first pass functional annotation of crocodilian proteins.

Species	Proteins with GO annotations (number annotated)			
	IEA	ISA	ISO	Total
Alligator	20,045 (68,530)	6,539 (36,488)	7,664 (78,020)	22,720 (183,038)
Crocodile	8,692 (32,659)	1,445 (6,279)	5,543 (59,087)	10,248 (98,025)
Gharial	8,797 (31,263)	1,414 (6,232)	5,519 (58,883)	10,468 (96,378)

Table S6. Tissue list, sources, number of reads and number of assembled contigs from RNASeq.

Tissue	Source	# paired reads	# assembled contigs
Belly skin	*A. mississippiensis*	4,312,359	52650
Cerebellum	*A. mississippiensis*	2,726,389	69651
Cerebrum	*A. mississippiensis*	2,307,767	67258
Chin glands	*A. mississippiensis*	1,320,056	37478
Eye	*A. mississippiensis*	1,528,086	52616
Heart	*A. mississippiensis*	3,041,067	45869
Kidney	*A. mississippiensis*	2,401,666	62331
Liver	*A. mississippiensis*	2,889,723	47998
Midbrain	*A. mississippiensis*	2,356,254	65238
Olfactory Bulb	*A. mississippiensis*	3,279,487	77620
Ovary	*A. mississippiensis*	3,772,951	60126
Spinal Cord	*A. mississippiensis*	2,896,984	69157
Spleen	*A. mississippiensis*	1,624,639	45414
Stomach	*A. mississippiensis*	495,156	10596
Testis	*A. mississippiensis*	3,139,697	73009
Thalamus	*A. mississippiensis*	2,876,750	70007
Throat Scent Gland	*A. mississippiensis*	2,346,716	60830
Tongue	*A. mississippiensis*	1,800,951	53386
Tooth	*A. mississippiensis*	1,040,198	25683
White Matter	*A. mississippiensis*	569,596	1811
Whole blood	*A. mississippiensis*	1,867,017	284
Pool A[a]	*A. mississippiensis*	4,408,460	83502
Pool B[b]	*A. mississippiensis*	703,642	6807
Combined alligator data	*A. mississippiensis*	53,705,611	345161
Multiple tissues[c]	*C. porosus*	12,643,511	1074
Whole blood	*G. gangeticus*	1,608,049	9580

[a] Intraceolomic fat body, scute muscle, trachea, cloacal gland and gastralia.

[b] Foot scale pad, tail muscle, lung, clitoris, mendela and intestine.

[c] brain, kidney, liver, lung, spleen, skin and tongue.

Table S7. Repeat content of each genome draft based on our custom library and RepeatMasker analysis.

Class		*A. mississippiensis*			*C. porosus*			*G. gangeticus*		
		Count	bp Masked	% masked	Count	bp Masked	% masked	Count	bp Masked	% masked
ARTEFACT		19	1114	0.00%	19	1279	0.00%	18	987	0.00%
DNA	hAT	612687	170427172	8.12%	600517	163448401	7.82%	648741	173374745	8.11%
	PIF-Harbinger	591679	155470000	7.41%	578197	148664324	7.12%	610968	156177093	7.30%
	Tc/Mariner	82708	22475746	1.07%	82171	22021751	1.05%	85410	22721954	1.06%
	DNA	86374	19033671	0.91%	77681	16339264	0.78%	81328	17106337	0.80%
	Maverick	1549	1756366	0.08%	1657	2134896	0.10%	2101	2085583	0.10%
	Kolobok	6015	1036790	0.05%	5628	962709	0.05%	5945	1020643	0.05%
	PiggyBac	646	685013	0.03%	646	659867	0.03%	665	690917	0.03%
	Academ	463	384988	0.02%	432	367187	0.02%	483	387399	0.02%
Rolling Circle	Helitron	26774	5369171	0.26%	25307	4958807	0.24%	27543	5350515	0.25%
LINE	CR1	348975	200624641	9.56%	388212	218469638	10.46%	413536	215968244	10.10%
	Penelope	92419	18598004	0.89%	80450	16655520	0.80%	85060	17242900	0.81%
	RTE	25567	18354640	0.87%	25317	17753691	0.85%	27070	18522294	0.86%
	L1	13600	14934612	0.71%	13267	13487570	0.64%	14545	14445660	0.67%
	L2	29510	6824112	0.33%	28270	6523987	0.31%	29765	6823937	0.32%
	Jockey	7704	4108792	0.20%	7499	3915520	0.19%	7976	4105166	0.19%
	R2	1564	921006	0.04%	2057	1133732	0.05%	2162	1073061	0.05%
LTR	Gypsy	101246	64698663	3.08%	96154	67672971	3.24%	102504	69140401	2.40%
	ERV1	37295	22848551	1.09%	37418	23800098	1.14%	35253	21727225	0.75%
	ERV4	17507	11657192	0.56%	8640	5631835	0.27%	21239	8420678	0.29%
	LTR	22619	8572038	0.41%	21297	8030395	0.38%	7577	5271747	0.18%
	ERVL	6984	4912555	0.23%	6969	4629793	0.22%	7651	5175107	0.18%
	DIRS	1043	246836	0.01%	1002	228547	0.01%	884	218641	0.01%
Retroposon		18000	1893426	0.09%	17067	1799095	0.09%	17915	1887452	0.09%
SINE	MIR	46451	5521407	0.26%	44152	5218825	0.25%	46459	5497975	0.26%
	tRNA	34147	5361667	0.26%	32671	5095747	0.25%	34387	5370502	0.25%
	5S-Deu-L2	12643	2187758	0.10%	12030	2059027	0.10%	12708	2176956	0.10%

	U	8400	2181070	0.10%	9400	2249310	0.11%	8841	2232505	0.10%
	L2	1967	386034	0.02%	1902	368500	0.02%	2017	389724	0.02%
Unknown		27476	4171890	0.20%	27835	4748325	0.23%	25217	4132473	0.14%
Total Interspersed		**2264031**	**775644925**	**36.96%**	**2233864**	**769030611**	**36.83%**	**2172073**	**782735266**	**27.17%**
Other	Simple_repeat	307235	12241410	0.58%	264385	10416581	0.50%	301197	11996583	0.56%
	Low_complexity	45971	2061339	0.10%	41366	1838883	0.09%	43064	1931791	0.09%
	Satellite	670	415436	0.02%	653	423428	0.02%	792	480396	0.02%
	tRNA	781	114897	0.01%	752	123745	0.01%	2302	152560	0.01%
	rRNA	2270	151042	0.01%	2085	137975	0.01%	713	120430	0.01%
	snRNA	11	758	0.00%	8	545	0.00%	764	43243	0.00%
	scRNA	926	54791	0.00%	723	41094	0.00%	10	650	0.00%
Total		**2621895**	**790684598**	**37.68%**	**2543836**	**782012862**	**37.45%**	**2722510**	**803446163**	**37.57%**

Table S8. Species, data set inclusion, genome build, source, and build name or accession information for genome assemblies from which we collected UCE data using *in silico* methods.

Species	Taxon Set 1	Taxon Set 2	Genome Build	Source	Accession
Latimeria chalumnae <coelacanth>	X	X	LATCHA1	UCSC	latCha1.2bit
Xenopus (Silurana) tropicalis <African clawed frog>	X		XENTRO3	UCSC	xenTro3.2bit
Ornithorhynchus anatinus <platypus>	X	X	ORNANA1	UCSC	ornAna1.fa.gz
Monodelphis domesticus <common opossum>	X	X	MONDOM5	UCSC	monDom5.2bit
Dasypus novemcinctus <armadillo>	X	X	DASNOV2	NCBI	GCA_000208655.1
Loxodonta africana <African elephant>	X	X	LOXAFR3	UCSC	loxAfr3.2bit
Canis familiaris <dog>	X	X	CANFAM3	UCSC	canFam3.2bit
Mus musculus <mouse>	X	X	MM10	UCSC	mm10.2bit
Homo sapiens <human>	X	X	HG19	UCSC	hg19.2bit
Python molurus <python>	X	X	PYTMOL1	NCBI	GCA_000186305.1
Anolis carolinensis <Carolina anole>	X	X	ANOCAR2	UCSC	anoCar2.2bit
Columba livia <Rock Dove (pigeon)>	X	X	COLLIV1	GigaDB	doi: 10.5524/100007
Taeniopygia guttata <Zebra Finch>	X	X	TAEGUT1	UCSC	chromFa.tar.gz
Aptenodytes forsteri <Emperor Penguin>	X	X	APTFOR1	GigaDB	doi: 10.5524/100006
Gallus gallus <Red Junglefowl (chicken)>	X	X	GALGAL4	UCSC	galGal4.2bit
Struthio camelus <Ostrich>	X	X	STRCAM1	BGI	Struthio_camelus.scaf.noBacterial.fa.gz
Alligator mississippiensis <alligator>	X	X	ALLMIS1	**This study**	
Crocodylus porosus <crocodile>	X	X	CROPOR1	**This study**	
Gavialis gangeticus <gharial>	X	X	GAVGAN1	**This study**	
Chrysemys picta <painted turtle>	X	X	CHRPIC1	UCSC	chrPic1.2bit
Pelodiscus sinensis <Chinese softshell turtle>	X	X	PELSIN1	NCBI	GCA_000230535.1
Chelonia mydas <green sea turtle>	X	X	CHEMYD1	NCBI	GCA_000344595.1

Table S9. Calibration sets used to calculate evolutionary rates. The sets are named based upon the assumed age for crocodilians (the first 'young' or 'old') and birds (the second 'young' or 'old'). Dates from the 'universal' column are used whenever 'univ' appears in a column. The partially unconstrained analyses (simply called 'unconstrained' in this Table) did not use a date when there are dashes in the column. Dates that are centered below the three calibration sets to the left (those marked with 'young' or 'old' designations) were used in all three of those analyses, but not in the unconstrained analysis. The supplementary reference number for each date is presented after the upper and lower limits in millions of years ago. References to 'SM 4' are to that section of this supporting material document.

	Calibration Sets				
Node	'young-young'	'old-young'	'old-old'	unconstrained	universal
A	univ	univ	univ	univ	408-419 (*98*)
B	univ	univ	univ	univ	330.4-350.1 (*185*)
C	univ	univ	univ	univ	312.3-330.4 (*185*)
D	univ	univ	univ	univ	162.5-191.1 (*185*)
E	univ	univ	univ	univ	124.6-138.4 (*185*)
F	univ	univ	univ	univ	95.3-113 (*185*)
G	univ	univ	univ	---	85.9-89.1 (*186*)
H	univ	univ	univ	---	81.1-83.8 (*186*)
I	univ	univ	univ	---	74.6-77 (*186*)
J	univ	univ	univ	---	260-289 (*185, 187*)
K	univ	univ	univ	---	148-162 (*188*)
L	univ	univ	univ	univ	260-265 (*144*)
M	univ	univ	univ	---	110-115 (*189*)
N	univ	univ	univ	univ	234-250 (*185*)
O	77.8-83.6 (SM 4)	93.3-100 (*9*)	93.3-100 (*9*)	77.9-83.6 (SM 4)	---
P	69-72 (SM 4)	50.7-69.9 (*9*)	50.7-69.9 (*9*)	---	---
Q	97.6-107.8 (*76*)	97.6-107.8 (*76*)	91.7-141.9 (*155*)	---	---
R	83.6-94 (*76*)	83.6-94 (*76*)	81.6-126.3 (*155*)	66-86.5 (*185*)	---
S	65.8-72.4 (*76*)	65.8-72.4 (*76*)	68.6-106.1 (*76*)	---	---

Node labels are shown on Figure S8. Especially important nodes are: node C [mammal-reptile divergence (crown amniotes)], node L (turtle-archosaur divergence), node N [crocodilian-bird divergences (crown archosaurs)], node O [alligator-crocodile divergence (crown crocodilian)], node P [crocodile-gharial divergence (Longirostres)], and node Q [divergence of ostrich from other birds (Neornithes or crown birds)].

Table S10. List of the 22 species used in the comparative genomics analyses of the three crocodilian species. Columns indicate i) the scientific name of each species, ii) the number of unique longest transcripts, iii) the gene number, iv) the data source and v) when data was downloaded. Note that the number of unique longest transcripts can be smaller than the number of genes, since two longest transcripts with identical sequence are collapsed on a single entry.

Species Name	Unique sequences	Genes	Source	As of
Alligator mississippiensis	**23,309**	**23,323**	**This study**	**11/1/2012**
Crocodylus porosus	**13,314**	**13,321**	**This study**	**11/1/2012**
Gavialis gangeticus	**14,024**	**14,043**	**This study**	**11/1/2012**
Aptenodytes forsteri	15,993	16,070	PhyloBirds Project	11/1/2012
Columba livia	16,456	16,652	PhyloBirds Project	11/1/2012
Gallus gallus	15,166	15,245	Quest for Orthologs release 2012/05	11/1/2012
Struthio camelus	15,465	16,178	PhyloBirds Project	11/1/2012
Taeniopygia guttata	17,401	17,488	Ensembl release 69	11/1/2012
Chelonia mydas	15,936	18,971	PhyloBirds Project	11/1/2012
Pelodiscus sinensis	18,096	18,188	Ensembl release 69	11/1/2012
Anolis carolinensis	17,734	17,805	Ensembl release 69	11/1/2012
Python bivittatus	25,367	25,385	Snake Genomics	2/1/2013
Canis familiaris	18,986	19,025	Quest for Orthologs release 2012/05	11/1/2012
Dasypus novemcinctus	14,659	14,803	Ensembl release 69	11/1/2012
Homo sapiens	19,997	21,206	Quest for Orthologs release 2012/05	7/1/2012
Loxodonta africana	20,001	20,033	Ensembl release 69	11/1/2012
Mus musculus	21,971	22,400	Quest for Orthologs release 2012/05	11/1/2012
Danio rerio	25,037	25,037	Quest for Orthologs release 2012/05	11/1/2012
Latimeria chalumnae	19,560	19,567	Ensembl release 69	11/1/2012
Monodelphis domestica	18,613	18,662	Quest for Orthologs release 2012/05	11/1/2012
Ornithorhynchus anatinus	17,899	17,951	Quest for Orthologs release 2011/04	10/1/2011
Xenopus tropicalis	18,303	18,417	Quest for Orthologs release 2012/05	11/1/2012

Table S11. Phylome support obtained for the internal organization of the main reptile groups. Only single gene trees containing the 4 reptile, as monophyletic, groups were considered. Among the possible topologies only the three likely ones were considered on the analysis.

Categories	*A. mississippiensis*	*C. porosus*	*G. gangeticus*
Single gene trees	19,226	11,856	12,644
Trees containing the 4 reptile groups	14,521	8,644	9,171
Trees with the 4 reptile groups being monophyletic	6,585	4,037	4,055
Topology 1:(Squamata, (Turtles, (Crocodiles, Birds)))	3,013	1,959	1,908
Topology 2: (Squamata, ((Crocodiles, Turtles), Birds))	760	446	452
Topology 3: (Squamata, ((Birds, Turtles), Crocodiles))	461	275	300
Other topologies	2,351	1,357	1,395

Table S12. Genome assemblies used to construct the WGA.

UCSC genome ID	Common name	Species name
falChe1	Saker falcon	*Falco cherrug*
falPer1	Peregrine falcon	*Falco peregrinus*
ficAlb2	Collared flycatcher	*Ficedula albicollis*
zonAlb1	White-throated sparrow	*Zonotrichia albicollis*
geoFor1	Medium ground finch	*Geospiza fortis*
taeGut2	Zebra finch	*Taeniopygia guttata*
pseHum1	Tibetan ground jay	*Pseudopodoces humilis*
melUnd1	Budgerigar	*Melopsittacus undulatus*
amaVit1	Puerto Rican Parrot	*Amazona vittata*
araMac1	Scarlet macaw	*Ara macao*
colLiv1	Rock pigeon	*Columbia livia*
anaPla1	Mallard duck	*Anas platyrynchos*
galGal4	Chicken	*Gallus gallus*
melGal1	Turkey	*Meleagris gallopavo*
strCam0	Ostrich	*Struthio camelus*
allMis2	American alligator	*Alligator mississippiensis*
croPor2	Crocodile	*Crocodylus porosus*
ghaGan1	Gharial	*Gavialis gangeticus*
cheMyd1	Green seaturtle	*Chelonia mydas*
chrPic1	Painted turtle	*Chrysemys picta bellii*
pelSin1	Soft-shell turtle	*Pelodiscus sinensis*
apaSpi1	Spiny soft-shell turtle	*Apalone spinifera*
anoCar2	Carolina Anole (Lizard)	*Anolis carolinensis*

Table S13. Percentage identity for each pair of crocodilian genomes.

Genome pair	Percent ID
Alligator, crocodile	92.9%
Crocodile, gharial	95.7%
Alligator, gharial	93.4%

Table S14. 4D site substitution rates for the branches directly above each crocodilian genome.

Genome	4D substitution rate	
	Filtering strategy 1	Filtering strategy 2
Alligator	0.0263	0.0254
Crocodile	0.0221	0.0211
Gharial	0.0172	0.0167
Crocodile-gharial common ancestor	0.0147	0.0144

Table S15. TE substitution rates for the branches directly above each crocodilian genome. Note that the tree can be arbitrarily rooted between the alligator and the crocodile-gharial common ancestor.

Genome	TE substitution rate
Alligator	0.0260
Crocodile	0.0246
Gharial	0.0188
Crocodile-gharial common ancestor	0.0260

Table S16. Clean micro-insertion and -deletion rates for leaf genomes in the WGA. Note that the Carolina anole does not appear in this Table, since it has no three-leaf subtree.

Leaf genome	Clean micro-insertion rate	Clean micro-deletion rate
Parrot	0.000436	0.001174
Mallard duck	0.001804	0.002931
American alligator	0.000617	0.00046
Spiny soft-shell turtle	0.000533	0.001075
Scarlet macaw	0.000448	0.000999
Green sea turtle	0.000969	0.001583
Painted turtle	0.000944	0.001669
Rock pigeon	0.001239	0.002664
Crocodile	0.000567	0.001061
Saker falcon	0.000045	0.000059
Peregrine falcon	0.000057	0.000086
Collared flycatcher	0.001203	0.003072
Chicken	0.000866	0.001426
Medium ground finch	0.000616	0.0013
Gharial	0.000381	0.000692
Turkey	0.001042	0.002382
Budgerigar	0.000666	0.001691
Soft-shell turtle	0.000512	0.001147
Tibetan ground jay	0.000978	0.002045
Ostrich	0.002045	0.001801
Zebra finch	0.001048	0.002381
White-throated sparrow	0.000794	0.001956

Table S17. Statistics about the number of duplication events detected in single gene trees in each phylome according to their relative ages.

Species Name	Ages	Events	Trees w/ events	Ratio
A. mississippiensis	**1**: Species-specific	4,748	1,710	0.247
(trees with duplication events: 19,226)	**2**: Crocodiles	11,613	2,411	0.604
	3: Archosauria	64	55	0.0033
	4: Crocodiles, Birds & Turtles	480	339	0.025
	5: Sauria	611	403	0.0318
C. porosus	**1**: Species-specific	1,067	551	0.09
(total trees with duplication events: 11,856)	**2**: Crocodiles	7,972	1,505	0.6724
	3: Archosauria	40	35	0.0034
	4: Crocodiles, Birds & Turtles	198	153	0.0167
	5: Sauria	230	167	0.0194
G. gangeticus	**1**: Species-specific	1,476	554	0.1167
(total trees with duplication events: 12,644)	**2**: Crocodiles	8,894	1,673	0.7034
	3: Archosauria	54	38	0.0043
	4: Crocodiles, Birds & Turtles	221	166	0.0175
	5: Sauria	288	235	0.0228

Table S18. Functional enrichments for proteins duplicated at different points across different evolutionary periods.

Species	Ages	Ontology	GO Term	Name
A. mississippiensis	1	Biological Process	GO:0006278	RNA-dependent DNA replication
	1	Biological Process	GO:0006869	lipid transport
	1	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	1	Biological Process	GO:0019882	antigen processing and presentation
	1	Biological Process	GO:0042157	lipoprotein metabolic process
	1	Cellular Component	GO:0042613	MHC class II protein complex
	1	Molecular Function	GO:0003723	RNA binding

	1	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	1	Molecular Function	GO:0004984	olfactory receptor activity
	1	Molecular Function	GO:0005153	interleukin-8 receptor binding
	1	Molecular Function	GO:0030246	carbohydrate binding
C. porosus	1	Biological Process	GO:0006278	RNA-dependent DNA replication
	1	Biological Process	GO:0006869	lipid transport
	1	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	1	Biological Process	GO:0019882	antigen processing and presentation
	1	Biological Process	GO:0042157	lipoprotein metabolic process
	1	Cellular Component	GO:0042613	MHC class II protein complex
	1	Molecular Function	GO:0003723	RNA binding
	1	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	1	Molecular Function	GO:0004984	olfactory receptor activity
	1	Molecular Function	GO:0005153	interleukin-8 receptor binding
	1	Molecular Function	GO:0030246	carbohydrate binding
G. gangeticus	1	Biological Process	GO:0006278	RNA-dependent DNA replication
	1	Biological Process	GO:0006869	lipid transport
	1	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	1	Biological Process	GO:0019882	antigen processing and presentation
	1	Biological Process	GO:0042157	lipoprotein metabolic process
	1	Cellular Component	GO:0042613	MHC class II protein complex
	1	Molecular Function	GO:0003723	RNA binding
	1	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	1	Molecular Function	GO:0004984	olfactory receptor activity
	1	Molecular Function	GO:0005153	interleukin-8 receptor binding

	1	Molecular Function	GO:0030246	carbohydrate binding
A. mississippiensis	2	Biological Process	GO:0006278	RNA-dependent DNA replication
	2	Biological Process	GO:0006869	lipid transport
	2	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	2	Biological Process	GO:0019882	antigen processing and presentation
	2	Biological Process	GO:0042157	lipoprotein metabolic process
	2	Cellular Component	GO:0042613	MHC class II protein complex
	2	Molecular Function	GO:0003723	RNA binding
	2	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	2	Molecular Function	GO:0004984	olfactory receptor activity
	2	Molecular Function	GO:0005153	interleukin-8 receptor binding
	2	Molecular Function	GO:0030246	carbohydrate binding
C. porosus	2	Biological Process	GO:0006278	RNA-dependent DNA replication
	2	Biological Process	GO:0006869	lipid transport
	2	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	2	Biological Process	GO:0019882	antigen processing and presentation
	2	Biological Process	GO:0042157	lipoprotein metabolic process
	2	Cellular Component	GO:0042613	MHC class II protein complex
	2	Molecular Function	GO:0003723	RNA binding
	2	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	2	Molecular Function	GO:0004984	olfactory receptor activity
	2	Molecular Function	GO:0005153	interleukin-8 receptor binding
	2	Molecular Function	GO:0030246	carbohydrate binding
G. gangeticus	2	Biological Process	GO:0006278	RNA-dependent DNA replication
	2	Biological Process	GO:0006869	lipid transport

	2	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	2	Biological Process	GO:0019882	antigen processing and presentation
	2	Biological Process	GO:0042157	lipoprotein metabolic process
	2	Cellular Component	GO:0042613	MHC class II protein complex
	2	Molecular Function	GO:0003723	RNA binding
	2	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	2	Molecular Function	GO:0004984	olfactory receptor activity
	2	Molecular Function	GO:0005153	interleukin-8 receptor binding
	2	Molecular Function	GO:0030246	carbohydrate binding
A. mississippiensis	3	Biological Process	GO:0006278	RNA-dependent DNA replication
	3	Biological Process	GO:0006869	lipid transport
	3	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	3	Biological Process	GO:0019882	antigen processing and presentation
	3	Biological Process	GO:0042157	lipoprotein metabolic process
	3	Cellular Component	GO:0042613	MHC class II protein complex
	3	Molecular Function	GO:0003723	RNA binding
	3	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	3	Molecular Function	GO:0004984	olfactory receptor activity
	3	Molecular Function	GO:0005153	interleukin-8 receptor binding
	3	Molecular Function	GO:0030246	carbohydrate binding
C. porosus	3	Biological Process	GO:0006278	RNA-dependent DNA replication
	3	Biological Process	GO:0006869	lipid transport
	3	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	3	Biological Process	GO:0019882	antigen processing and presentation
	3	Biological Process	GO:0042157	lipoprotein metabolic process

	3	Cellular Component	GO:0042613	MHC class II protein complex
	3	Molecular Function	GO:0003723	RNA binding
	3	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	3	Molecular Function	GO:0004984	olfactory receptor activity
	3	Molecular Function	GO:0005153	interleukin-8 receptor binding
	3	Molecular Function	GO:0030246	carbohydrate binding
G. gangeticus	3	Biological Process	GO:0006278	RNA-dependent DNA replication
	3	Biological Process	GO:0006869	lipid transport
	3	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	3	Biological Process	GO:0019882	antigen processing and presentation
	3	Biological Process	GO:0042157	lipoprotein metabolic process
	3	Cellular Component	GO:0042613	MHC class II protein complex
	3	Molecular Function	GO:0003723	RNA binding
	3	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	3	Molecular Function	GO:0004984	olfactory receptor activity
	3	Molecular Function	GO:0005153	interleukin-8 receptor binding
	3	Molecular Function	GO:0030246	carbohydrate binding
A. mississippiensis	4	Biological Process	GO:0006278	RNA-dependent DNA replication
	4	Biological Process	GO:0006869	lipid transport
	4	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	4	Biological Process	GO:0019882	antigen processing and presentation
	4	Biological Process	GO:0042157	lipoprotein metabolic process
	4	Cellular Component	GO:0042613	MHC class II protein complex
	4	Molecular Function	GO:0003723	RNA binding
	4	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity

	4	Molecular Function	GO:0004984	olfactory receptor activity
	4	Molecular Function	GO:0005153	interleukin-8 receptor binding
	4	Molecular Function	GO:0030246	carbohydrate binding
C. porosus	4	Biological Process	GO:0006278	RNA-dependent DNA replication
	4	Biological Process	GO:0006869	lipid transport
	4	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	4	Biological Process	GO:0019882	antigen processing and presentation
	4	Biological Process	GO:0042157	lipoprotein metabolic process
	4	Cellular Component	GO:0042613	MHC class II protein complex
	4	Molecular Function	GO:0003723	RNA binding
	4	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	4	Molecular Function	GO:0004984	olfactory receptor activity
	4	Molecular Function	GO:0005153	interleukin-8 receptor binding
	4	Molecular Function	GO:0030246	carbohydrate binding
G. gangeticus	4	Biological Process	GO:0006278	RNA-dependent DNA replication
	4	Biological Process	GO:0006869	lipid transport
	4	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	4	Biological Process	GO:0019882	antigen processing and presentation
	4	Biological Process	GO:0042157	lipoprotein metabolic process
	4	Cellular Component	GO:0042613	MHC class II protein complex
	4	Molecular Function	GO:0003723	RNA binding
	4	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	4	Molecular Function	GO:0004984	olfactory receptor activity
	4	Molecular Function	GO:0005153	interleukin-8 receptor binding
	4	Molecular Function	GO:0030246	carbohydrate binding

A. mississippiensis	5	Biological Process	GO:0006278	RNA-dependent DNA replication
	5	Biological Process	GO:0006869	lipid transport
	5	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	5	Biological Process	GO:0019882	antigen processing and presentation
	5	Biological Process	GO:0042157	lipoprotein metabolic process
	5	Cellular Component	GO:0042613	MHC class II protein complex
	5	Molecular Function	GO:0003723	RNA binding
	5	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	5	Molecular Function	GO:0004984	olfactory receptor activity
	5	Molecular Function	GO:0005153	interleukin-8 receptor binding
	5	Molecular Function	GO:0030246	carbohydrate binding
C. porosus	5	Biological Process	GO:0006278	RNA-dependent DNA replication
	5	Biological Process	GO:0006869	lipid transport
	5	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
	5	Biological Process	GO:0019882	antigen processing and presentation
	5	Biological Process	GO:0042157	lipoprotein metabolic process
	5	Cellular Component	GO:0042613	MHC class II protein complex
	5	Molecular Function	GO:0003723	RNA binding
	5	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
	5	Molecular Function	GO:0004984	olfactory receptor activity
	5	Molecular Function	GO:0005153	interleukin-8 receptor binding
	5	Molecular Function	GO:0030246	carbohydrate binding
G. gangeticus	5	Biological Process	GO:0006278	RNA-dependent DNA replication
	5	Biological Process	GO:0006869	lipid transport
	5	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway

5	Biological Process	GO:0019882	antigen processing and presentation
5	Biological Process	GO:0042157	lipoprotein metabolic process
5	Cellular Component	GO:0042613	MHC class II protein complex
5	Molecular Function	GO:0003723	RNA binding
5	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
5	Molecular Function	GO:0004984	olfactory receptor activity
5	Molecular Function	GO:0005153	interleukin-8 receptor binding
5	Molecular Function	GO:0030246	carbohydrate binding

Table S19. Functional enrichment analysis results after removing redundancy for clusters with five or more members of specifically duplicated proteins in each species.

Species Name	Cluster ID	Size	Ontology	GO Term	Name
A. mississippiensis	1210	41	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	1210	41	Cellular Component	GO:0005634	nucleus
	1210	41	Biological Process	GO:0016032	viral reproduction
	1210	41	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	1189	29	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	1189	29	Cellular Component	GO:0005634	nucleus
	1189	29	Biological Process	GO:0016032	viral reproduction
	1189	29	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	1175	22	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	1175	22	Cellular Component	GO:0005634	nucleus
	1175	22	Biological Process	GO:0016032	viral reproduction
	1175	22	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	1160	19	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	1160	19	Cellular Component	GO:0005634	nucleus
	1160	19	Biological Process	GO:0016032	viral reproduction
	1160	19	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	1159	18	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	1159	18	Cellular Component	GO:0005634	nucleus
	1159	18	Biological Process	GO:0016032	viral reproduction
	1159	18	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	1145	17	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity

1145	17	Cellular Component	GO:0005634	nucleus
1145	17	Biological Process	GO:0016032	viral reproduction
1145	17	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1176	16	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1176	16	Cellular Component	GO:0005634	nucleus
1176	16	Biological Process	GO:0016032	viral reproduction
1176	16	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1183	15	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1183	15	Biological Process	GO:0016032	viral reproduction
1170	15	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
1170	15	Molecular Function	GO:0003723	RNA binding
1170	15	Biological Process	GO:0006278	RNA-dependent DNA replication
1181	14	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1181	14	Cellular Component	GO:0005634	nucleus
1181	14	Biological Process	GO:0016032	viral reproduction
1181	14	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1144	13	Molecular Function	GO:0016787	hydrolase activity
1161	12	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1161	12	Biological Process	GO:0016032	viral reproduction
1161	12	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1126	12	Biological Process	GO:0016032	viral reproduction
1178	11	Biological Process	GO:0016032	viral reproduction
1146	11	Biological Process	GO:0016032	viral reproduction
1135	11	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity

1135	11	Biological Process	GO:0016032	viral reproduction
1135	11	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1140	10	Biological Process	GO:0016032	viral reproduction
1130	10	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1130	10	Biological Process	GO:0016032	viral reproduction
1130	10	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1094	10	Molecular Function	GO:0030246	carbohydrate binding
1093	10	Molecular Function	GO:0030246	carbohydrate binding
1153	9	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1153	9	Cellular Component	GO:0005634	nucleus
1153	9	Biological Process	GO:0016032	viral reproduction
1153	9	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1152	9	Biological Process	GO:0016032	viral reproduction
1098	9	Biological Process	GO:0016032	viral reproduction
1081	8	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1081	8	Biological Process	GO:0016032	viral reproduction
1081	8	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1072	8	Molecular Function	GO:0016787	hydrolase activity
1061	8	Biological Process	GO:0016032	viral reproduction
1056	8	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1056	8	Biological Process	GO:0016032	viral reproduction
1056	8	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1171	7	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
1171	7	Molecular Function	GO:0003723	RNA binding

1171	7	Biological Process	GO:0006278	RNA-dependent DNA replication
1165	7	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
1165	7	Molecular Function	GO:0003723	RNA binding
1165	7	Biological Process	GO:0006278	RNA-dependent DNA replication
1096	7	Molecular Function	GO:0016787	hydrolase activity
1059	7	Biological Process	GO:0016032	viral reproduction
1055	7	Biological Process	GO:0016032	viral reproduction
1053	7	Molecular Function	GO:0016787	hydrolase activity
1037	7	Molecular Function	GO:0016787	hydrolase activity
1027	7	Molecular Function	GO:0030246	carbohydrate binding
996	6	Molecular Function	GO:0004984	olfactory receptor activity
996	6	Cellular Component	GO:0016021	integral to membrane
996	6	Biological Process	GO:0007186	G-protein coupled receptor signaling pathway
993	6	Molecular Function	GO:0016787	hydrolase activity
992	6	Molecular Function	GO:0016787	hydrolase activity
990	6	Molecular Function	GO:0016787	hydrolase activity
985	6	Molecular Function	GO:0004523	ribonuclease H activity
985	6	Molecular Function	GO:0003676	nucleic acid binding
985	6	Biological Process	GO:0015074	DNA integration
980	6	Molecular Function	GO:0030246	carbohydrate binding
968	6	Molecular Function	GO:0005200	structural constituent of cytoskeleton
968	6	Cellular Component	GO:0005882	intermediate filament
967	6	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
967	6	Biological Process	GO:0016032	viral reproduction

967	6	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
965	6	Molecular Function	GO:0016746	transferase activity, transferring acyl groups
965	6	Biological Process	GO:0008152	metabolic process
1205	6	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
1205	6	Molecular Function	GO:0003723	RNA binding
1205	6	Biological Process	GO:0006278	RNA-dependent DNA replication
1137	6	Molecular Function	GO:0003964	RNA-directed DNA polymerase activity
1137	6	Molecular Function	GO:0003723	RNA binding
1137	6	Biological Process	GO:0006278	RNA-dependent DNA replication
1134	6	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1134	6	Biological Process	GO:0016032	viral reproduction
1091	6	Molecular Function	GO:0016787	hydrolase activity
1080	6	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1080	6	Biological Process	GO:0016032	viral reproduction
1080	6	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1077	6	Molecular Function	GO:0016787	hydrolase activity
1040	6	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
1040	6	Biological Process	GO:0016032	viral reproduction
1040	6	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
1034	6	Molecular Function	GO:0016787	hydrolase activity
930	5	Biological Process	GO:0016032	viral reproduction
920	5	Molecular Function	GO:0016787	hydrolase activity
907	5	Molecular Function	GO:0030246	carbohydrate binding
906	5	Molecular Function	GO:0030246	carbohydrate binding

	903	5	Cellular Component	GO:0042612	MHC class I protein complex
	903	5	Biological Process	GO:0019882	antigen processing and presentation
	902	5	Cellular Component	GO:0042612	MHC class I protein complex
	902	5	Biological Process	GO:0019882	antigen processing and presentation
	895	5	Molecular Function	GO:0016787	hydrolase activity
	891	5	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	891	5	Biological Process	GO:0016032	viral reproduction
	891	5	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	890	5	Molecular Function	GO:0004859	phospholipase inhibitor activity
	889	5	Molecular Function	GO:0030246	carbohydrate binding
	1119	5	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	1119	5	Biological Process	GO:0016032	viral reproduction
	1119	5	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	1050	5	Molecular Function	GO:0004523	ribonuclease H activity
	1050	5	Biological Process	GO:0015074	DNA integration
	1021	5	Molecular Function	GO:0016787	hydrolase activity
	1003	5	Biological Process	GO:0015074	DNA integration
C. porosus	463	16	Molecular Function	GO:0003676	nucleic acid binding
	456	14	Molecular Function	GO:0016787	hydrolase activity
	460	13	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	460	13	Cellular Component	GO:0005634	nucleus
	460	13	Biological Process	GO:0016032	viral reproduction
	460	13	Biological Process	GO:0015074	DNA integration
	460	13	Biological Process	GO:0006355	regulation of transcription, DNA-dependent

	441	10	Molecular Function	GO:0003676	nucleic acid binding
	449	9	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	449	9	Biological Process	GO:0016032	viral reproduction
	449	9	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	430	8	Molecular Function	GO:0016787	hydrolase activity
	422	6	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	422	6	Biological Process	GO:0016032	viral reproduction
	422	6	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	415	6	Molecular Function	GO:0003676	nucleic acid binding
	411	6	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	411	6	Biological Process	GO:0016032	viral reproduction
	411	6	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	398	6	Molecular Function	GO:0003676	nucleic acid binding
	396	6	Molecular Function	GO:0004190	aspartic-type endopeptidase activity
	396	6	Biological Process	GO:0006508	proteolysis
	388	5	Molecular Function	GO:0016787	hydrolase activity
G. gangeticus	219	7	Molecular Function	GO:0003700	sequence-specific DNA binding transcription factor activity
	219	7	Biological Process	GO:0016032	viral reproduction
	219	7	Biological Process	GO:0006355	regulation of transcription, DNA-dependent
	211	6	Molecular Function	GO:0016787	hydrolase activity
	199	5	Molecular Function	GO:0016772	transferase activity, transferring phosphorus-containing groups
	199	5	Molecular Function	GO:0004672	protein kinase activity
	199	5	Biological Process	GO:0006468	protein phosphorylation

Table S20. Alligator elements that did not map to archosaur in the WGA were aligned against multiple leaf genomes using LASTZ. Elements were restricted to >30 bp to prevent spurious alignment, and repeat-masking was ignored in both the element and the target genomes.

Genome	Alligator fragments missing in archosaur with >= 1 blast hit	Alligator fragments missing in archosaur with >= 10 blast hits	Alligator coding exons missing in archosaur with >= 1 blast hit	Alligator coding exons missing in archosaur with >= 10 blast hits
Chicken	440/5000 (8.8%)	328/5000 (6.6%)	3484/22666 (15.3%)	1556/22666 (6.9%)
Zebra finch	423/5000 (8.5%)	296/5000 (6.0%)	3108/22666 (13.7%)	1252/22666 (5.5%)
Soft-shell turtle	903/5000 (18%)	619/5000 (12.3%)	6354/22666 (28.0%)	2766/22666 (12.2%)
Crocodile	4351/5000 (87%)	2300/5000 (46%)	19314/22666 (85.2%)	7458/22666 (32.9%)

Table S21. Size and diversity of the olfactory receptor repertoires of the three crocodilians in our study compared to previously published data from the Chinese alligator (*15*); chicken and zebra finch (*74*); soft shell and green sea turtles (*16*); Carolina anole, cow, dog, mouse, human, platypus and clawed frog (*190, 191*).

	Total	Intact Genes	Partial Genes	Pseudo-genes	Intact Genes	
					α	γ
Crocodile	989	593	66	330	43	549
Gharial	1139	597	153	389	36	560
American alligator	1077	465	74	538	45	419
Chinese alligator	1112	434	-	678*	37	396
Chicken	478	213	154	111	9	204
Zebra finch	552	133	198	221	2	131
Soft shell turtle	1744	1137	-	595*	532	604
Green sea turtle	849	254	-	607*	158	95
Carolina anole	146	112	4	30	1	111
Cow	2129	970	182	977	140	828
Dog	1100	811	11	278	159	651
Mouse	1391	1035	28	328	110	922
Human	802	387	0	415	58	329
Platypus	718	365	83	370	31	234
Western clawed frog	1638	824	200	614	8	752

* Partial gene and pseudogene counts were summed.

 Summed counts from the α and γ groups may not equal the total number of intact genes found because the total includes OR genes from other groups.

Table S22. The size of chemosensory receptor genes repertoires in the three crocodilians, compared to squamate reptiles and mammals. Truncated genes and putative pseudogenes are shown in parentheses next to the total number of intact genes. Partial genes that were apparently truncated owing to incomplete sequence assembly, were counted as functional intact genes as long as they display an entire seven transmembrane helices domain, other partial genes were counted as truncated genes.

	V2R[†]	T1R[†]	V1R	T2R	TAAR
Crocodile	0 (0/0)	3 (0/0)	1 (0/1)	5 (0/0)	8 (0/1)
Gharial	0 (0/0)	3 (0/0)	1 (0/1)	7 (0/2)	9 (0/2)
American alligator	0 (0/1)	2 (1/0)	1 (0/1)	6 (1/1)	7 (0/2)
Chicken	0 (0/0)	2 (0/1)	0 (0/0)	3 (0/0)	3 (0/0)
Zebra finch	0 (0/0)	1 (0/1)	0 (0/0)	5 (0/3)	1 (0/0)
Soft shell turtle	2 (0/0)	3 (0/0)	1 (0/0)	10 (0/3)	8 (0/3)
Green sea turtle	0 (0/0)	3 (0/0)	1 (0/1)	1 (1/9)	11 (0/2)
Carolina anole	62 (4/2)	3 (0/2)	1 (0/0)	41 (8/9)	3 (0/0)
Python	203 (115/97)	2 (0/0)	1 (0/1)	1 (0/1)	3 (0/1)
Monodelphis	86 (-/79)	3 (-/0)	98 (-/30)	29 (-/5)	22 (-/0)
Human	0 (-/7)	3 (-/0)	5 (-/115)	25 (-/11)	6 (-/3)
Platypus	15 (-/112)	0 (1/0)	270 (-/579)	5 (0/3)	4 (-/1)
Western clawed frog	249 (-/448)	0 (-/0)	21 (-/2)	52 (-/12)	6 (-/1)

Data from Nei et al. (*192*), except for T1Rs and T2Rs in platypus.

[†]The number of multiexon genes, V2R and T1R, was estimated based on the presence of the complete highly conserved seven transmembrane helices domain coded in the last exon.

Table S23. Reproductive characteristics and population sizes (captive, wild, and skin exports) of the three species of interest.

Species	Sex	Size at Sexual Maturity	Age of Sexual Maturity	Average Clutch Size (Range)	Expected Reproductive Lifetime	Estimated Generation Time	Wild Population Size	Captive Population Size	Number of skins exported
A. mississippiensis	M	1.8m[193, 194]	10-15[194, 195]	N/A			≈2-3 million[42]	>650,000[42]	>360,000[196]
	F	1.5-1.8m[193, 197, 198]	10-18[194, 197, 199]	39 (2-58)[193, 198]	12 (6-18)[199]	≈18[199]			
C. porosus	M	3.2m[194, 200]	15-16[194, 200]	N/A			≈400,000 non-hatchlings worldwide[201]	>100,000	>53,000[196]
	F	2.2m[200, 202]	10-12[200, 203]	50 (40-62)[202]	≈20	≈20			
G. gangeticus	M	3.5-3.7[204, 205]	12-13[204, 205]	N/A			<200 breeding adults[41]	≈500	N/A
	F	>2.7m[194, 204, 206]	10-13[204, 206]	40 (16-60)[194, 207]	20	≈20[41]			

Table S24 – Substitution spectra in crocodile sequence data. The number of observations for the reference (columns) and alternate (rows) bases are shown in each substitution class. Transversions are in blue, transitions in red.

13 alternate alleles

	A	C	G	T
A	0	4	12	4
C	5	0	3	25
G	11	3	0	3
T	2	9	4	0

12 alternate alleles

	A	C	G	T
A	0	5	31	4
C	4	0	9	35
G	22	7	0	4
T	4	27	9	0

7 alternate alleles

	A	C	G	T
A	0	478	2,278	404
C	335	0	505	2,238
G	1,931	430	0	482
T	340	1,950	382	0

2 alternate alleles

	A	C	G	T
A	0	123	407	79
C	215	0	89	433
G	406	102	0	252
T	83	382	100	0

1 alternate allele

	A	C	G	T
A	0	39,740	36,889	12,981
C	72,375	0	12,596	22,183
G	22,189	12,563	0	76,004
T	12,895	36,499	39,772	0

References

1. A. Janke, U. Arnason, The complete mitochondrial genome of *Alligator mississippiensis* and the separation between recent archosauria (birds and crocodiles). *Mol. Biol. Evol.* **14**, 1266–1272 (1997). Medline doi:10.1093/oxfordjournals.molbev.a025736

2. A. G. Sennikov, The first ctenosauriscid (Reptilia: Archosauromorpha) from the Lower Triassic of eastern Europe. *Paleontol. J.* **46**, 499–511 (2012). doi:10.1134/S0031030112050097

3. C. A. Brochu, Phylogenetic approaches toward crocodylian history. *Annu. Rev. Earth Planet. Sci.* **31**, 357–397 (2003). doi:10.1146/annurev.earth.31.100901.141308

4. G. Grigg, F. Seebacher, C. E. Franklin, Eds., *Crocodilian Biology and Evolution* (Surrey Beatty, Chipping Norton, NSW, Australia, 2001).

5. M. M. Cohen, C. Gans, The chromosomes of the order Crocodilia. *Cytogenetics* **9**, 81–105 (1970). Medline doi:10.1159/000130080

6. J. W. Lang, H. V. Andrews, Temperature-dependent sex determination in crocodilians. *J. Exp. Zool.* **270**, 28–44 (1994). doi:10.1002/jez.1402700105

7. E. Olmo, Evolution of genome size and DNA base composition in reptiles. *Genetica* **57**, 39–50 (1981). doi:10.1007/BF00057541

8. J. Harshman, C. J. Huddleston, J. P. Bollback, T. J. Parsons, M. J. Braun, True and false gharials: A nuclear gene phylogeny of crocodylia. *Syst. Biol.* **52**, 386–402 (2003). Medline doi:10.1080/10635150390197028

9. J. R. Oaks, A time-calibrated species tree of Crocodylia reveals a recent radiation of the true crocodiles. *Evolution* **65**, 3285–3297 (2011). Medline doi:10.1111/j.1558-5646.2011.01373.x

10. A. Janke, A. Gullberg, S. Hughes, R. K. Aggarwal, U. Arnason, Mitogenomic analyses place the gharial (*Gavialis gangeticus*) on the crocodile tree and provide pre-K/T divergence times for most crocodilians. *J. Mol. Evol.* **61**, 620–626 (2005). Medline doi:10.1007/s00239-004-0336-9

11. J. Alföldi, F. Di Palma, M. Grabherr, C. Williams, L. Kong, E. Mauceli, P. Russell, C. B. Lowe, R. E. Glor, J. D. Jaffe, D. A. Ray, S. Boissinot, A. M. Shedlock, C. Botka, T. A. Castoe, J. K. Colbourne, M. K. Fujita, R. G. Moreno, B. F. ten Hallers, D. Haussler, A. Heger, D. Heiman, D. E. Janes, J. Johnson, P. J. de Jong, M. Y. Koriabine, M. Lara, P. A. Novick, C. L. Organ, S. E. Peach, S. Poe, D. D. Pollock, K. de Queiroz, T. Sanger, S. Searle, J. D. Smith, Z. Smith, R. Swofford, J. Turner-Maier, J. Wade, S. Young, A. Zadissa, S. V. Edwards, T. C. Glenn, C. J. Schneider, J. B. Losos, E. S. Lander, M. Breen, C. P. Ponting, K. Lindblad-Toh, The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587–591 (2011). Medline doi:10.1038/nature10390

12. T. A. Castoe, A. P. de Koning, K. T. Hall, D. C. Card, D. R. Schield, M. K. Fujita, R. P. Ruggiero, J. F. Degner, J. M. Daza, W. Gu, J. Reyes-Velasco, K. J. Shaney, J. M. Castoe, S. E. Fox, A. W. Poole, D. Polanco, J. Dobry, M. W. Vandewege, Q. Li, R. K. Schott, A. Kapusta, P. Minx, C. Feschotte, P. Uetz, D. A. Ray, F. G. Hoffmann, R. Bogden, E. N. Smith, B. S. Chang, F. J. Vonk, N. R. Casewell, C. V. Henkel, M. K. Richardson, S. P. Mackessy, A. M. Bronikowski, M. Yandell, W. C. Warren, S. M. Secor, D. D. Pollock, The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20645–20650 (2013). Medline

13. H. B. Shaffer, P. Minx, D. E. Warren, A. M. Shedlock, R. C. Thomson, N. Valenzuela, J. Abramyan, C. T. Amemiya, D. Badenhorst, K. K. Biggar, G. M. Borchert, C. W. Botka, R. M. Bowden, E. L. Braun, A. M. Bronikowski, B. G. Bruneau, L. T. Buck, B. Capel, T. A. Castoe, M. Czerwinski, K. D. Delehaunty, S. V. Edwards, C. C. Fronick, M. K. Fujita, L. Fulton, T. A. Graves, R. E. Green, W. Haerty, R. Hariharan, O. Hernandez, L. W. Hillier, A. K. Holloway, D. Janes, F. J. Janzen, C. Kandoth, L. Kong, A. P. de Koning, Y. Li, R. Literman, S. E. McGaugh, L. Mork, M. O'Laughlin, R. T. Paitz, D. D. Pollock, C. P. Ponting, S. Radhakrishnan, B. J. Raney, J. M. Richman, J. St John, T. Schwartz, A. Sethuraman, P. Q. Spinks, K. B. Storey, N. Thane, T. Vinar, L. M. Zimmerman, W. C. Warren, E. R. Mardis, R. K. Wilson, The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol.* **14**, R28 (2013). Medline doi:10.1186/gb-2013-14-3-r28

14. F. J. Vonk, N. R. Casewell, C. V. Henkel, A. M. Heimberg, H. J. Jansen, R. J. McCleary, H. M. Kerkkamp, R. A. Vos, I. Guerreiro, J. J. Calvete, W. Wüster, A. E. Woods, J. M. Logan, R. A. Harrison, T. A. Castoe, A. P. de Koning, D. D. Pollock, M. Yandell, D. Calderon, C. Renjifo, R. B. Currier, D. Salgado, D. Pla, L. Sanz, A. S. Hyder, J. M. Ribeiro, J. W. Arntzen, G. E. van den Thillart, M. Boetzer, W. Pirovano, R. P. Dirks, H. P. Spaink, D. Duboule, E. McGlinn, R. M. Kini, M. K. Richardson, The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 20651–20656 (2013). Medline doi:10.1073/pnas.1314702110

15. Q. H. Wan, S. K. Pan, L. Hu, Y. Zhu, P. W. Xu, J. Q. Xia, H. Chen, G. Y. He, J. He, X. W. Ni, H. L. Hou, S. G. Liao, H. Q. Yang, Y. Chen, S. K. Gao, Y. F. Ge, C. C. Cao, P. F. Li, L. M. Fang, L. Liao, S. Zhang, M. Z. Wang, W. Dong, S. G. Fang, Genome analysis and signature discovery for diving and sensory properties of the endangered Chinese alligator. *Cell Res.* **23**, 1091–1105 (2013). Medline doi:10.1038/cr.2013.104

16. Z. Wang, J. Pascual-Anaya, A. Zadissa, W. Li, Y. Niimura, Z. Huang, C. Li, S. White, Z. Xiong, D. Fang, B. Wang, Y. Ming, Y. Chen, Y. Zheng, S. Kuraku, M. Pignatelli, J. Herrero, K. Beal, M. Nozawa, Q. Li, J. Wang, H. Zhang, L. Yu, S. Shigenobu, J. Wang, J. Liu, P. Flicek, S. Searle, J. Wang, S. Kuratani, Y. Yin, B. Aken, G. Zhang, N. Irie, The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat. Genet.* **45**, 701–706 (2013). Medline doi:10.1038/ng.2615

17. See supplementary materials on *Science* Online.

18. L. W. Hillier, W. Miller, E. Birney, W. Warren, R. C. Hardison, C. P. Ponting, P. Bork, D. W. Burt, M. A. M. Groenen, M. E. Delany, J. B. Dodgson, A. T. Chinwalla, P. F. Cliften, S. W. Clifton, K. D. Delehaunty, C. Fronick, R. S. Fulton, T. A. Graves, C. Kremitzki, D. Layman, V. Magrini, J. D. McPherson, T. L. Miner, P. Minx, W. E. Nash, M. N. Nhan, J. O. Nelson, L. G. Oddy, C. S. Pohl, J. Randall-Maher, S. M. Smith, J. W. Wallis, S.-P. Yang, M. N. Romanov, C. M. Rondelli, B. Paton, J. Smith, D. Morrice, L. Daniels, H. G. Tempest, L. Robertson, J. S. Masabanda, D. K. Griffin, A. Vignal, V. Fillon, L. Jacobbson, S. Kerje, L. Andersson, R. P. M. Crooijmans, J. Aerts, J. J. van der Poel, H. Ellegren, R. B. Caldwell, S. J. Hubbard, D. V. Grafham, A. M. Kierzek, S. R. McLaren,

I. M. Overton, H. Arakawa, K. J. Beattie, Y. Bezzubov, P. E. Boardman, J. K. Bonfield, M. D. R. Croning, R. M. Davies, M. D. Francis, S. J. Humphray, C. E. Scott, R. G. Taylor, C. Tickle, W. R. A. Brown, J. Rogers, J.-M. Buerstedde, S. A. Wilson, L. Stubbs, I. Ovcharenko, L. Gordon, S. Lucas, M. M. Miller, H. Inoko, T. Shiina, J. Kaufman, J. Salomonsen, K. Skjoedt, G. K.-S. Wong, J. Wang, B. Liu, J. Wang, J. Yu, H. Yang, M. Nefedov, M. Koriabine, P. J. deJong, L. Goodstadt, C. Webber, N. J. Dickens, I. Letunic, M. Suyama, D. Torrents, C. von Mering, E. M. Zdobnov, K. Makova, A. Nekrutenko, L. Elnitski, P. Eswara, D. C. King, S. Yang, S. Tyekucheva, A. Radakrishnan, R. S. Harris, F. Chiaromonte, J. Taylor, J. He, M. Rijnkels, S. Griffiths-Jones, A. Ureta-Vidal, M. M. Hoffman, J. Severin, S. M. J. Searle, A. S. Law, D. Speed, D. Waddington, Z. Cheng, E. Tuzun, E. Eichler, Z. Bao, P. Flicek, D. D. Shteynberg, M. R. Brent, J. M. Bye, E. J. Huckle, S. Chatterji, C. Dewey, L. Pachter, A. Kouranov, Z. Mourelatos, A. G. Hatzigeorgiou, A. H. Paterson, R. Ivarie, M. Brandstrom, E. Axelsson, N. Backstrom, S. Berlin, M. T. Webster, O. Pourquie, A. Reymond, C. Ucla, S. E. Antonarakis, M. Long, J. J. Emerson, E. Betrán, I. Dupanloup, H. Kaessmann, A. S. Hinrichs, G. Bejerano, T. S. Furey, R. A. Harte, B. Raney, A. Siepel, W. J. Kent, D. Haussler, E. Eyras, R. Castelo, J. F. Abril, S. Castellano, F. Camara, G. Parra, R. Guigo, G. Bourque, G. Tesler, P. A. Pevzner, A. Smit, L. A. Fulton, E. R. Mardis, R. K. Wilson, Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004). Medline doi:10.1038/nature03154

19. J. Huerta-Cepas, S. Capella-Gutiérrez, L. P. Pryszcz, M. Marcet-Houben, T. Gabaldón, PhylomeDB v4: Zooming into the plurality of evolutionary histories of a genome. *Nucleic Acids Res.* **42**, D897–D902 (2014). Medline [doi:10.1093/nar/gkt1177](doi:10.1093/nar/gkt1177)

20. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L.

Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001). Medline doi:10.1038/35057062

21. W. C. Warren, D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, A. Künstner, S. Searle, S. White, A. J. Vilella, S. Fairley, A. Heger, L. Kong, C. P. Ponting, E. D. Jarvis, C. V. Mello, P. Minx, P. Lovell, T. A. Velho, M. Ferris, C. N. Balakrishnan, S. Sinha, C. Blatti, S. E. London, Y. Li, Y. C. Lin, J. George, J. Sweedler, B. Southey, P. Gunaratne, M. Watson, K. Nam, N. Backström, L. Smeds, B. Nabholz, Y. Itoh, O. Whitney, A. R. Pfenning, J. Howard, M. Völker, B. M. Skinner, D. K. Griffin, L. Ye, W. M. McLaren, P. Flicek, V. Quesada, G. Velasco, C. Lopez-Otin, X. S. Puente, T. Olender, D. Lancet, A. F. Smit, R. Hubley, M. K. Konkel, J. A. Walker, M. A. Batzer, W. Gu, D. D. Pollock, L. Chen, Z. Cheng, E. E. Eichler, J. Stapley, J. Slate, R. Ekblom, T. Birkhead, T. Burke, D. Burt, C. Scharff, I. Adam, H. Richard, M. Sultan, A. Soldatov, H. Lehrach, S. V. Edwards, S. P. Yang, X. Li, T. Graves, L. Fulton, J. Nelson, A. Chinwalla, S. Hou, E. R. Mardis, R. K. Wilson, The genome of a songbird. *Nature* **464**, 757–762 (2010). [Medline](#) [doi:10.1038/nature08819](#)

22. W. C. Warren, L. W. Hillier, J. A. Marshall Graves, E. Birney, C. P. Ponting, F. Grützner, K. Belov, W. Miller, L. Clarke, A. T. Chinwalla, S. P. Yang, A. Heger, D. P. Locke, P. Miethke, P. D. Waters, F. Veyrunes, L. Fulton, B. Fulton, T. Graves, J. Wallis, X. S. Puente, C. López-Otín, G. R. Ordóñez, E. E. Eichler, L. Chen, Z. Cheng, J. E. Deakin, A. Alsop, K. Thompson, P. Kirby, A. T. Papenfuss, M. J. Wakefield, T. Olender, D. Lancet, G. A. Huttley, A. F. Smit, A. Pask, P. Temple-Smith, M. A. Batzer, J. A. Walker, M. K. Konkel, R. S. Harris, C. M. Whittington, E. S. Wong, N. J. Gemmell, E. Buschiazzo, I. M. Vargas Jentzsch, A. Merkel, J. Schmitz, A. Zemann, G. Churakov, J. O. Kriegs, J. Brosius, E. P. Murchison, R. Sachidanandam, C. Smith, G. J. Hannon, E. Tsend-Ayush, D. McMillan, R. Attenborough, W. Rens, M. Ferguson-Smith, C. M. Lefèvre, J. A. Sharp, K. R. Nicholas, D. A. Ray, M. Kube, R. Reinhardt, T. H. Pringle, J. Taylor, R. C. Jones, B. Nixon, J. L. Dacheux, H. Niwa, Y. Sekita, X. Huang, A. Stark, P. Kheradpour, M. Kellis, P. Flicek, Y. Chen, C. Webber, R. Hardison, J. Nelson, K. Hallsworth-Pepin, K. Delehaunty, C. Markovic, P. Minx, Y. Feng, C. Kremitzki, M. Mitreva, J. Glasscock, T. Wylie, P. Wohldmann, P. Thiru, M. N. Nhan, C. S. Pohl, S. M. Smith, S. Hou, M. Nefedov, P. J. de Jong, M. B. Renfree, E. R. Mardis, R. K. Wilson, Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008). [Medline](#) [doi:10.1038/nature06936](#)

23. A. T. Chinwalla, L. L. Cook, K. D. Delehaunty, G. A. Fewell, L. A. Fulton, R. S. Fulton, T. A. Graves, L. D. W. Hillier, E. R. Mardis, J. D. McPherson, T. L. Miner, W. E. Nash, J. O. Nelson, M. N. Nhan, K. H. Pepin, C. S. Pohl, T. C. Ponce, B. Schultz, J. Thompson, E. Trevaskis, R. H. Waterston, M. C. Wendl, R. K. Wilson, S.-P. Yang, P. An, E. Berry, B. Birren, T. Bloom, D. G. Brown, J. Butler, M. Daly, R. David, J. Deri, S. Dodge, K. Foley, D. Gage, S. Gnerre, T. Holzer, D. B. Jaffe, M. Kamal, E. K. Karlsson, C. Kells, A. Kirby, E. J. Kulbokas, E. S. Lander, T. Landers, J. P. Leger, R. Levine, K. Lindblad-Toh, E. Mauceli, J. H. Mayer, M. McCarthy, J. Meldrim, J. Meldrim, J. P. Mesirov, R. Nicol, C. Nusbaum, S. Seaman, T. Sharpe, A. Sheridan, J. B. Singer, R. Santos, B. Spencer, N. Stange-Thomann, J. P. Vinson, C. M. Wade, J. Wierzbowski, D. Wyman, M. C. Zody, E. Birney, N. Goldman, A. Kasprzyk, E. Mongin, A. G. Rust, G. Slater, A. Stabenau, A. Ureta-Vidal, S. Whelan, R. Ainscough, J. Attwood, J. Bailey, K. Barlow, S. Beck, J. Burton, M. Clamp, C. Clee, A. Coulson, J. Cuff, V. Curwen, T. Cutts, J. Davies, E. Eyras, D. Grafham, S. Gregory, T. Hubbard, A. Hunt, M. Jones, A. Joy, S. Leonard, C. Lloyd, L. Matthews, S. McLaren, K. McLay, B. Meredith, J. C. Mullikin, Z. Ning, K. Oliver, E. Overton-Larty, R. Plumb, S. Potter, M. Quail, J. Rogers, C. Scott, S. Searle, R. Shownkeen, S. Sims, M. Wall, A. P. West, D. Willey, S. Williams, J. F. Abril, R. Guigó, G. Parra, P. Agarwal, R. Agarwala, D. M. Church, W. Hlavina, D. R. Maglott, V. Sapojnikov, M. Alexandersson, L. Pachter, S. E. Antonarakis, E. T. Dermitzakis, A. Reymond, C. Ucla, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, F. Hsu, D. Karolchik, W. J. Kent, K. M. Roskin, M. S. Schwartz, C. Sugnet, R. J. Weber, P. Bork, I. Letunic, M. Suyama, D. Torrents, E. M. Zdobnov, M. Botcherby, S. D. Brown, R. D. Campbell, I. Jackson, N. Bray, O. Couronne, I. Dubchak, A. Poliakov, E. M. Rubin, M. R. Brent, P. Flicek, E. Keibler, I. Korf, S. Batalov, C. Bult, W. N. Frankel, P. Carninci, Y. Hayashizaki, J. Kawai, Y. Okazaki, S. Cawley, D. Kulp, R. Wheeler, F. Chiaromonte, F. S. Collins, A. Felsenfeld, M. Guyer, J. Peterson, K. Wetterstrand, R. R. Copley, R. Mott, C. Dewey, N. J. Dickens, R. D. Emes, L. Goodstadt, C. P. Ponting, E. Winter, D. M. Dunn, A. C. von Niederhausern, R. B. Weiss, S. R. Eddy, L. S. Johnson, T. A. Jones, L. Elnitski, D. L. Kolbe, P. Eswara, W. Miller, M. J. O'Connor, S. Schwartz, R. A. Gibbs, D. M. Muzny, G. Glusman, A. Smit, E. D. Green, R. C. Hardison, S. Yang, D. Haussler, A. Hua, B. A. Roe, R. S. Kucherlapati, K. T. Montgomery, J. Li, M. Li, S.

Lucas, B. Ma, W. R. McCombie, M. Morgan, P. Pevzner, G. Tesler, J. Schultz, D. R. Smith, J. Tromp, K. C. Worley, E. S. Lander, J. F. Abril, P. Agarwal, M. Alexandersson, S. E. Antonarakis, R. Baertsch, E. Berry, E. Birney, P. Bork, N. Bray, M. R. Brent, D. G. Brown, J. Butler, C. Bult, F. Chiaromonte, A. T. Chinwalla, D. M. Church, M. Clamp, F. S. Collins, R. R. Copley, O. Couronne, S. Cawley, J. Cuff, V. Curwen, T. Cutts, M. Daly, E. T. Dermitzakis, C. Dewey, N. J. Dickens, M. Diekhans, I. Dubchak, S. R. Eddy, L. Elnitski, R. D. Emes, P. Eswara, E. Eyras, A. Felsenfeld, P. Flicek, W. N. Frankel, L. A. Fulton, T. S. Furey, S. Gnerre, G. Glusman, N. Goldman, L. Goodstadt, E. D. Green, S. Gregory, R. Guigó, R. C. Hardison, D. Haussler, L. D. W. Hillier, A. Hinrichs, W. Hlavina, F. Hsu, T. Hubbard, D. B. Jaffe, M. Kamal, D. Karolchik, E. K. Karlsson, A. Kasprzyk, E. Keibler, W. J. Kent, A. Kirby, D. L. Kolbe, I. Korf, E. J. Kulbokas, D. Kulp, E. S. Lander, I. Letunic, M. Li, K. Lindblad-Toh, B. Ma, D. R. Maglott, E. Mauceli, J. P. Mesirov, W. Miller, R. Mott, J. C. Mullikin, Z. Ning, L. Pachter, G. Parra, P. Pevzner, A. Poliakov, C. P. Ponting, S. Potter, A. Reymond, K. M. Roskin, V. Sapojnikov, J. Schultz, M. S. Schwartz, S. Schwartz, S. Searle, J. B. Singer, G. Slater, A. Smit, A. Stabenau, C. Sugnet, M. Suyama, G. Tesler, D. Torrents, J. Tromp, C. Ucla, J. P. Vinson, C. M. Wade, R. J. Weber, R. Wheeler, E. Winter, S.-P. Yang, E. M. Zdobnov, R. H. Waterston, S. Whelan, K. C. Worley, M. C. Zody, Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002). Medline doi:10.1038/nature01262

24. G. Bejerano, M. Pheasant, I. Makunin, S. Stephen, W. J. Kent, J. S. Mattick, D. Haussler, Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004). Medline doi:10.1126/science.1098119

25. B. C. Faircloth, J. E. McCormack, N. G. Crawford, M. G. Harvey, R. T. Brumfield, T. C. Glenn, Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. *Syst. Biol.* **61**, 717–726 (2012). Medline doi:10.1093/sysbio/sys004

26. J. E. McCormack, B. C. Faircloth, N. G. Crawford, P. A. Gowaty, R. T. Brumfield, T. C. Glenn, Ultraconserved elements are novel phylogenomic markers that resolve placental

mammal phylogeny when combined with species-tree analysis. *Genome Res.* **22**, 746–754 (2012). Medline doi:10.1101/gr.125864.111

27. A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, D. Haussler, Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005). Medline doi:10.1101/gr.3715005

28. Y. Chiari, V. Cahais, N. Galtier, F. Delsuc, Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biol.* **10**, 65 (2012). Medline doi:10.1186/1741-7007-10-65

29. N. G. Crawford, B. C. Faircloth, J. E. McCormack, R. T. Brumfield, K. Winker, T. C. Glenn, More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biol. Lett.* **8**, 783–786 (2012). Medline doi:10.1098/rsbl.2012.0331

30. A. C. Tzika, R. Helaers, G. Schramm, M. C. Milinkovitch, Reptilian-transcriptome v1.0, a glimpse in the brain transcriptome of five divergent Sauropsida lineages and the phylogenetic position of turtles. *EvoDevo* **2**, 19 (2011). Medline doi:10.1186/2041-9139-2-19

31. J. V. Chamary, J. L. Parmley, L. D. Hurst, Hearing silence: Non-neutral evolution at synonymous sites in mammals. *Nat. Rev. Genet.* **7**, 98–108 (2006). Medline doi:10.1038/nrg1770

32. A. Künstner, B. Nabholz, H. Ellegren, Significant selective constraint at 4-fold degenerate sites in the avian genome and its consequence for detection of positive selection. *Genome Biol. Evol.* **3**, 1381–1389 (2011). Medline doi:10.1093/gbe/evr112

33. C. A. Brochu, Morphology, fossils, divergence timing, and the phylogenetic relationships of Gavialis. *Syst. Biol.* **46**, 479–522 (1997). Medline doi:10.1093/sysbio/46.3.479

34. G. V. Glazko, M. Nei, Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20**, 424–434 (2003). Medline doi:10.1093/molbev/msg050

35. R. Cordaux, J. Lee, L. Dinoso, M. A. Batzer, Recently integrated Alu retrotransposons are essentially neutral residents of the human genome. *Gene* **373**, 138–144 (2006). Medline doi:10.1016/j.gene.2006.01.020

36. B. Paten, J. Herrero, S. Fitzgerald, K. Beal, P. Flicek, I. Holmes, E. Birney, Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res.* **18**, 1829–1843 (2008). Medline doi:10.1101/gr.076521.108

37. M. Blanchette, E. D. Green, W. Miller, D. Haussler, Reconstructing large regions of an ancestral mammalian genome in silico. *Genome Res.* **14**, 2412–2423 (2004). Medline doi:10.1101/gr.2800104

38. Y. Niimura, Olfactory receptor multigene family in vertebrates: From the viewpoint of evolutionary genomics. *Curr. Genomics* **13**, 103–114 (2012). Medline doi:10.2174/138920212799860706

39. J. Huerta-Cepas, H. Dopazo, J. Dopazo, T. Gabaldón, The human phylome. *Genome Biol.* **8**, R109 (2007). Medline

40. Y. Niimura, M. Nei, Extensive gains and losses of olfactory receptor genes in mammalian evolution. *PLOS ONE* **2**, e708 (2007). Medline

41. C. Stevenson, R. Whitaker, in *Crocodiles. Status Survey and Conservation Action Plan,* S. C. Manolis, C. Stevenson, Eds. (Crocodile Specialist Group, Darwin, Australia, 2010), pp. 139–143.

42. R. M. Elsey, A. R. Woodward, in *Crocodiles. Status Survey and Conservation Action Plan*, S. C. Manolis, C. Stevenson, Eds. (Crocodile Specialist Group, Darwin, Australia, 2010), pp. 1–4.

43. H. Li, R. Durbin, Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011). Medline doi:10.1038/nature10231

44. R. S. W. van de Wal, B. de Boer, L. J. Lourens, P. Kohler, R. Bintanja, Reconstruction of a continuous high-resolution CO_2 record over the past 20 million years. *Clim. Past* **7**, 1459–1469 (2011). doi:10.5194/cp-7-1459-2011

45. G. Zhang, C. Li, Q. Li, B. Li, D. M. Larkin, C. Lee, J. F. Storz, A. Antunes, M. J. Greenwold, R. W. Meredith, A. Ödeen, J. Cui, Q. Zhou, L. Xu, H. Pan, Z. Wang, L. Jin, P. Zhang, H. Hu, W. Yang, J. Hu, J. Xiao, Z. Yang, Y. Liu, Q. Xie, H. Yu, J. Lian, P. Wen, F. Zhang, H. Li, Y. Zeng, Z. Xiong, S. Liu, L. Zhou, Z. Huang, N. An, J. Wang, Q. Zheng, Y. Xiong, G. Wang, B. Wang, J. Wang, Y. Fan, R. R. da Fonseca, A. Alfaro-Núñez, M. Schubert, L. Orlando, T. Mourier, J. T. Howard, G. Ganapathy, A. Pfenning, O. Whitney, M. V. Rivas, E. Hara, J. Smith, M. Farré, J. Narayan, G. Slavov, M. N. Romanov, R. Borges, J. P. Machado, I. Khan, M. S. Springer, J. Gatesy, F. G. Hoffmann, J. C. Opazo, O. Håstad, R. H. Sawyer, H. Kim, K.-W. Kim, H. J. Kim, S. Cho, N. Li, Y. Huang, M. W. Bruford, X. Zhan, A. Dixon, M. F. Bertelsen, E. Derryberry, W. Warren, R. K. Wilson, S. Li, D. A. Ray, R. E. Green, S. J. O'Brien, D. Griffin, W. E. Johnson, D. Haussler, O. A. Ryder, E. Willerslev, G. R. Graves, P. Alström, J. Fjeldså, D. P. Mindell, S. V. Edwards, E. L. Braun, C. Rahbek, D. W. Burt, P. Houde, Y. Zhang, H. Yang, J. Wang, Avian Genome Consortium, E. D. Jarvis, M. T. P. Gilbert, J. Wang, Comparative genomics reveals insights into avian genome evolution and adaptation. *Science* **346**, 1311–1320 (2014).

46. C. I. Wu, W. H. Li, Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. U.S.A.* **82**, 1741–1745 (1985). Medline doi:10.1073/pnas.82.6.1741

47. Y. H. Lin, P. J. Waddell, D. Penny, Pika and vole mitochondrial genomes increase support for both rodent monophyly and glires. *Gene* **294**, 119–129 (2002). Medline doi:10.1016/S0378-1119(02)00695-9

48. A. P. Martin, S. R. Palumbi, Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4087–4091 (1993). Medline doi:10.1073/pnas.90.9.4087

49. J. F. Gillooly, A. P. Allen, G. B. West, J. H. Brown, The rate of DNA evolution: Effects of body size and temperature on the molecular clock. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 140–145 (2005). Medline doi:10.1073/pnas.0407735101

50. L. Blueweiss, H. Fox, V. Kudzma, D. Nakashima, R. Peters, S. Sams, Relationships between body size and some life history parameters. *Oecologia* **37**, 257–272 (1978). doi:10.1007/BF00344996

51. V. M. Savage, J. F. Gillooly, J. H. Brown, G. B. West, E. L. Charnov, Effects of body size and temperature on population growth. *Am. Nat.* **163**, 429–441 (2004). Medline doi:10.1086/381872

52. M. S. Lee, A. Cau, D. Naish, G. J. Dyke, Sustained miniaturization and anatomical innovation in the dinosaurian ancestors of birds. *Science* **345**, 562–566 (2014). Medline doi:10.1126/science.1252243

53. S. Gnerre, I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, D. B. Jaffe, High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 1513–1518 (2011). Medline doi:10.1073/pnas.1017351108

54. A. M. Shedlock, C. W. Botka, S. Zhao, J. Shetty, T. Zhang, J. S. Liu, P. J. Deschavanne, S. V. Edwards, Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 2767–2772 (2007). Medline doi:10.1073/pnas.0606204104

55. R. Li, H. Zhu, J. Ruan, W. Qian, X. Fang, Z. Shi, Y. Li, S. Li, G. Shan, K. Kristiansen, S. Li, H. Yang, J. Wang, J. Wang, De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.* **20**, 265–272 (2010). Medline doi:10.1101/gr.097261.109

56. M. Stanke, M. Diekhans, R. Baertsch, D. Haussler, Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008). Medline doi:10.1093/bioinformatics/btn013

57. D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S. L. Salzberg, TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013). Medline doi:10.1186/gb-2013-14-4-r36

58. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009). Medline doi:10.1186/gb-2009-10-3-r25

59. A. F. A. Smit, R. Hubley, RepeatModeler Open-1.0; www.repeatmasker.org.

60. R. C. Edgar, E. W. Myers, PILER: Identification and classification of genomic repeats. *Bioinformatics* **21** (suppl. 1), i152–i158 (2005). Medline doi:10.1093/bioinformatics/bti1003

61. A. L. Price, N. C. Jones, P. A. Pevzner, De novo identification of repeat families in large genomes. *Bioinformatics* **21** (suppl. 1), i351–i358 (2005). Medline doi:10.1093/bioinformatics/bti1018

62. D. Ellinghaus, S. Kurtz, U. Willhoeft, LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008). Medline doi:10.1186/1471-2105-9-18

63. S. Stephen, M. Pheasant, I. V. Makunin, J. S. Mattick, Large-scale appearance of ultraconserved elements in tetrapod genomes and slowdown of the molecular clock. *Mol. Biol. Evol.* **25**, 402–408 (2008). Medline doi:10.1093/molbev/msm268

64. A. Stamatakis, RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**, 2688–2690 (2006). Medline doi:10.1093/bioinformatics/btl446

65. J. Huerta-Cepas, A. Bueno, J. Dopazo, T. Gabaldón, PhylomeDB: A database for genome-wide collections of gene phylogenies. *Nucleic Acids Res.* **36**, D491–D496 (2008). Medline doi:10.1093/nar/gkm899

66. M. R. Wallace, L. B. Andersen, A. M. Saulino, P. E. Gregory, T. W. Glover, F. S. Collins, A de novo Alu insertion results in neurofibromatosis type 1. *Nature* **353**, 864–866 (1991). Medline doi:10.1038/353864a0

67. S. Capella-Gutiérrez, J. M. Silla-Martínez, T. Gabaldón, trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009). Medline doi:10.1093/bioinformatics/btp348

68. S. Guindon, J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010). Medline doi:10.1093/sysbio/syq010

69. T. Gabaldón, Large-scale assignment of orthology: Back to phylogenetics? *Genome Biol.* **9**, 235 (2008). Medline doi:10.1186/gb-2008-9-10-235

70. J. Huerta-Cepas, J. Dopazo, T. Gabaldón, ETE: A python environment for tree exploration. *BMC Bioinformatics* **11**, 24 (2010). Medline doi:10.1186/1471-2105-11-24

71. J. Huerta-Cepas, T. Gabaldón, Assigning duplication events to relative temporal scales in genome-wide studies. *Bioinformatics* **27**, 38–45 (2011). Medline doi:10.1093/bioinformatics/btq609

72. A. Wehe, M. S. Bansal, J. G. Burleigh, O. Eulenstein, DupTree: A program for large-scale phylogenetic analyses using gene tree parsimony. *Bioinformatics* **24**, 1540–1541 (2008). Medline doi:10.1093/bioinformatics/btn230

73. K. Tamura, D. Peterson, N. Peterson, G. Stecher, M. Nei, S. Kumar, MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28**, 2731–2739 (2011). Medline doi:10.1093/molbev/msr121

74. S. S. Steiger, V. Y. Kuryshev, M. C. Stensmyr, B. Kempenaers, J. C. Mueller, A comparison of reptilian and avian olfactory receptor gene repertoires: Species-specific expansion of group gamma genes in birds. *BMC Genomics* **10**, 446 (2009). Medline doi:10.1186/1471-2164-10-446

75. B. Paten, D. Earl, N. Nguyen, M. Diekhans, D. Zerbino, D. Haussler, Cactus: Algorithms for genome multiple sequence alignment. *Genome Res.* **21**, 1512–1528 (2011). Medline doi:10.1101/gr.123356.111

76. E. D. Jarvis, S. Mirarab, A. J. Aberer, B. Li, P. Houde, C. Li, S. Y. W. Ho, B. C. Faircloth, B. Nabholz, J. T. Howard, A. Suh, C. C. Weber, R. R. da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, Md. S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K.

Munch, M. Schierup, B. Lindow, W. C. Warren, D. Ray, R. E. Green, M. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E. P. Derryberry, M. F. Bertelsen, F.H. Sheldon, R. T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M. P. Cruz Schneider, F. Prosdocimi, J. A. Samaniego, A. M. Vargas Velazquez, A. Alfaro-Núñez, P.F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D. M. Lambert, Q. Zhou, P. Perelman, A. C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B. Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F. E. Rheindt, M. Braun, J. Fjeldsa, L. Orlando, K. Barker, K. A. Jønsson, W. Johnson, K.-P. Koepfli, S. O'Brien, D. Haussler, O. A. Ryder, C. Rahbek, E. Willerslev, G. R. Graves, T. C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D. P. Mindell, J. Cracraft, E. L. Braun, T. Warnow, W. Jun, M. T. P. Gilbert, G. Zhang, Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**, 1320–1331 (2014).

77. F. K. Barker, A. Cibois, P. Schikler, J. Feinstein, J. Cracraft, Phylogeny and diversification of the largest avian radiation. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 11040–11045 (2004). Medline doi:10.1073/pnas.0401892101

78. T. F. Wright, E. E. Schirtzinger, T. Matsumoto, J. R. Eberhard, G. R. Graves, J. J. Sanchez, S. Capelli, H. Müller, J. Scharpegge, G. K. Chambers, R. C. Fleischer, A multilocus molecular phylogeny of the parrots (Psittaciformes): Support for a Gondwanan origin during the Cretaceous. *Mol. Biol. Evol.* **25**, 2141 (2008). doi:10.1093/molbev/msn160

79. R. C. Thomson, H. B. Shaffer, Sparse supermatrices for phylogenetic inference: Taxonomy, alignment, rogue taxa, and the phylogeny of living turtles. *Syst. Biol.* **59**, 42–58 (2010). Medline doi:10.1093/sysbio/syp075

80. N. Nguyen, G. Hickey, D. R. Zerbino, B. Raney, D. Earl, J. Armstrong, D. Haussler, B. Paten, Building a pangenome reference for a population. *Res. Comput. Mol. Biol.* **8394**, 207–221 (2014). doi:10.1007/978-3-319-05269-4_17

81. G. Hickey, B. Paten, D. Earl, D. Zerbino, D. Haussler, HAL: A hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29**, 1341–1342 (2013). Medline doi:10.1093/bioinformatics/btt128

82. M. J. Hubisz, K. S. Pollard, A. Siepel, PHAST and RPHAST: Phylogenetic analysis with space/time models. *Brief. Bioinform.* **12**, 41–51 (2011). Medline doi:10.1093/bib/bbq072

83. H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009). Medline doi:10.1093/bioinformatics/btp324

84. W. Jaratlerdsiri *et al*., Comparative analyses reveal adaptive MHC structure in the saltwater crocodile (*Crocodylus porosus*). *PLOS ONE* 10.1371/journal.pone.0114631 (2014).

85. M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, A. Regev, Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011). Medline doi:10.1038/nbt.1883

86. Z. Bao, S. R. Eddy, Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002). Medline doi:10.1101/gr.88502

87. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997). doi:10.1093/nar/25.17.3389

88. O. Kohany, A. J. Gentles, L. Hankus, J. Jurka, Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics* **7**, 474 (2006). Medline doi:10.1186/1471-2105-7-474

89. A. F. A. Smit, R. Hubley, P. Green, RepeatMasker Open-3.0; www.repeatmasker.org.

90. C. Feschotte, U. Keswani, N. Ranganathan, M. L. Guibotsy, D. Levine, Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol. Evol.* **1**, 205–220 (2009). Medline doi:10.1093/gbe/evp023

91. R. S. Harris, thesis, Pennsylvania State University (2007); www.bx.psu.edu/~rsharris/lastz.

92. K. Katoh, K. Kuma, H. Toh, T. Miyata, MAFFT version 5: Improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* **33**, 511–518 (2005). Medline doi:10.1093/nar/gki198

93. R. Lanfear, B. Calcott, S. Y. Ho, S. Guindon, Partitionfinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Mol. Biol. Evol.* **29**, 1695–1701 (2012). Medline doi:10.1093/molbev/mss020

94. R. Lanfear, B. Calcott, D. Kainer, C. Mayer, A. Stamatakis, Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evol. Biol.* **14**, 82 (2014). Medline doi:10.1186/1471-2148-14-82

95. M. J. Sanderson, r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003). Medline doi:10.1093/bioinformatics/19.2.301

96. M. J. Benton, P. C. J. Donoghue, Paleontological evidence to date the tree of life. *Mol. Biol. Evol.* **24**, 26–53 (2007). Medline doi:10.1093/molbev/msl150

97. D. T. Ksepka, C. A. Boyd, Quantifying historical trends in the completeness of the fossil record and the contributing factors: An example using Aves. *Paleobiology* **38**, 112–125 (2012). doi:10.1666/10059.1

98. J. Müller, R. R. Reisz, Four well-constrained calibration points from the vertebrate fossil record for molecular clock estimates. *BioEssays* **27**, 1069–1075 (2005). Medline doi:10.1002/bies.20286

99. J. F. Parham, P. C. Donoghue, C. J. Bell, T. D. Calway, J. J. Head, P. A. Holroyd, J. G. Inoue, R. B. Irmis, W. G. Joyce, D. T. Ksepka, J. S. Patané, N. D. Smith, J. E. Tarver, M. van Tuinen, Z. Yang, K. D. Angielczyk, J. M. Greenwood, C. A. Hipsley, L. Jacobs, P. J. Makovicky, J. Müller, K. T. Smith, J. M. Theodor, R. C. Warnock, M. J. Benton, Best practices for justifying fossil calibrations. *Syst. Biol.* **61**, 346–359 (2012). Medline doi:10.1093/sysbio/syr107

100. D. T. Ksepka, M. J. Benton, M. T. Carrano, M. A. Gandolfo, J. J. Head, E. J. Hermsen, W. G. Joyce, K. S. Lamm, J. S. Patané, M. J. Phillips, P. D. Polly, M. Van Tuinen, J. L. Ware, R. C. Warnock, J. F. Parham, Synthesizing and databasing fossil calibrations: Divergence dating and beyond. *Biol. Lett.* **7**, 801–803 (2011). Medline doi:10.1098/rsbl.2011.0356

101. C. A. Brochu, A new Late Cretaceous gavialoid crocodylian from eastern North America and the phylogenetic relationships of thoracosaurs. *J. Vertebr. Paleontol.* **24**, 610–633 (2004). [doi:10.1671/0272-4634(2004)024[0610:ANLCGC]2.0.CO;2](doi:10.1671/0272-4634(2004)024[0610:ANLCGC]2.0.CO;2)

102. J. Gatesy, G. Amato, M. Norell, R. DeSalle, C. Hayashi, Combined support for wholesale taxic atavism in gavialine crocodylians. *Syst. Biol.* **52**, 403–422 (2003). Medline [doi:10.1080/10635150390197037](doi:10.1080/10635150390197037)

103. S. W. Salisbury, R. E. Molnar, E. Frey, P. M. A. Willis, The origin of modern crocodyliforms: New evidence from the Cretaceous of Australia. *Proc. R. Soc. B* **273**, 2439 (2006). doi:10.1098/rspb.2006.3613

104. S. W. Salisbury, E. Frey, D. M. Martill, M. C. Buchy, A new crocodilian from the Lower Cretaceous Crato Formation of north-eastern Brazil. *Palaeontogr. Abt. A* **270**, 3 (2003).

105. A. Ősi, J. M. Clark, D. B. Weishampel, First report on a new basal eusuchian crocodyliform with multicusped teeth from the Upper Cretaceous (Santonian) of Hungary. *Neues Jahrb. Geol. Palaontol. Abh.* **243**, 169–177 (2007). [doi:10.1127/0077-7749/2007/0243-0169](doi:10.1127/0077-7749/2007/0243-0169)

106. D. Pol, A. H. Turner, M. A. Norell, Morphology of the Late Cretaceous crocodylomorph *Shamosuchus djadochtaensis* and a discussion of Neosuchian phylogeny as related to the origin of Eusuchia. *Bull. Am. Mus. Nat. Hist. 324* (2009).

107. D. C. Fortier, C. L. Schultz, A new Neosuchian crocodylomorph (Crocodyliformes, Mesoeucrocodylia) from the Early Cretaceous of north-east Brazil. *Palaeontology* **52**, 991–1007 (2009). [doi:10.1111/j.1475-4983.2009.00894.x](doi:10.1111/j.1475-4983.2009.00894.x)

108. A. D. Buscalioni, P. Piras, R. Vullo, M. Signore, C. Barbera, Early eusuchia crocodylomorpha from the vertebrate-rich Plattenkalk of Pietraroia (Lower Albian, southern Apennines, Italy). *Zool. J. Linn. Soc.* **163**, S199–S227 (2011). [doi:10.1111/j.1096-3642.2011.00718.x](doi:10.1111/j.1096-3642.2011.00718.x)

109. C. M. Holliday, N. M. Gardner, A new eusuchian crocodyliform with novel cranial integument and its significance for the origin and evolution of Crocodylia. *PLOS ONE* **7**, e30471 (2012). Medline doi:10.1371/journal.pone.0030471

110. F. C. Montefeltro, H. C. E. Larsson, M. A. G. de França, M. C. Langer, A new neosuchian with Asian affinities from the Jurassic of northeastern Brazil. *Naturwissenschaften* **100**, 835–841 (2013). Medline doi:10.1007/s00114-013-1083-9

111. C. A. Brochu, A new Late Cretaceous gavialoid crocodylian from eastern North America and the phylogenetic relationships of thoracosaurs. *J. Vertebr. Paleontol.* **24**, 610–633 (2004). doi:10.1671/0272-4634(2004)024[0610:ANLCGC]2.0.CO;2

112. X. C. Wu, A. P. Russell, D. B. Brinkman, A review of *Leidyosuchus canadensis* Lambe, 1907 (Archosauria: Crocodylia) and an assessment of cranial variation based upon new material. *Can. J. Earth Sci.* **38**, 1665–1687 (2001). doi:10.1139/e01-059

113. X. C. Wu, in *Dinosaur Provincial Park: A Spectacular Ancient Ecosystem Revealed*, P. J. Currie, E. B. Koppelhus, Eds. (Indiana Univ. Press, Bloomington, 2005), pp. 277–290.

114. R. B. Irmis, J. H. Hutchison, J. J. W. Sertich, A. L. Titus, in *At the Top of the Grand Staircase: The Late Cretaceous of Southern Utah*, A. L. Titus, M. A. Loewen, Eds. (Indiana Univ. Press, Bloomington, 2013), pp. 424–444.

115. J. A. Lillegraven, A new genus of therian mammal from the Late Cretaceous "El Gallo Formation," Baja California, Mexico. *J. Paleontol.* **50**, 437 (1976).

116. C. A. Brochu, W. Langston, T. Rowe, A new, phylogenetically significant alligatoroid from the Late Cretaceous (Campanian) of Mexico. *J. Vertebr. Paleontol.* **33**, 94A (2013).

117. W. J. Morris, in *A Guidebook to the Geology of Peninsular California*, G. Gastil, J. A. Lillegraven, Eds. (American Association of Petroleum Geologists/Society of Economic Geologists, Pacific Section, Los Angeles, 1974), pp. 60–66.

118. P. R. Renne, M. M. Fulford, C. Busbyspera, High-resolution ^{40}Ar/^{39}Ar chronostratigraphy of the Late Cretaceous El Gallo formation, Baja-California Del Norte, Mexico. *Geophys. Res. Lett.* **18**, 459–462 (1991). doi:10.1029/91GL00464

119. E. M. Roberts, A. L. Deino, M. A. Chan, ^{40}Ar/^{39}Ar age of the Kaipirowits Formation, southern Utah, and correlation of contemporaenous Campanian strata and vertebrate faunas along the margin of the Western Interior Basin. *Cretac. Res.* **26**, 307–318 (2005). doi:10.1016/j.cretres.2005.01.002

120. D. A. Eberth, in *Dinosaur Provincial Park: A Spectacular Ancient Ecosystem Revealed*, P. J. Currie, E. B. Koppelhus, Eds. (Indiana Univ. Press, Bloomington, 2005), pp. 54–82.

121. E. H. Colbert, R. T. Bird, A gigantic crocodile from the Upper Cretaceous beds of Texas. *Am. Mus. Novit.* **1688**, 1 (1954).

122. D. R. Schwimmer, *King of the Crocodylians: The Paleobiology of Deinosuchus* (Indiana Univ. Press, Bloomington, 2002).

123. W. Wahl, J. Hogbin, Deinosuchus material from the Mesaerde formation of Wyoming: Filling in a gap. *J. Vertebr. Paleontol.* **23**, 107A (2003).

124. S. G. Lucas, J. A. Spielmann, R. M. Sullivan, C. Lewis, Late Cretaceous crocodylians from the San Juan Basin, New Mexico. *New Mexico Mus. Nat. Hist. Sci. Bull.* **35**, 249 (2006).

125. H. E. Rivera-Sylva, E. Frey, The first mandible fragment of Deinosuchus (Eusuchia: Alligatoroidea) discovered in Coahuila, Mexico. *Bol. Soc. Geol. Mex.* **63**, 459 (2011).

126. A. D. Buscalioni, J. L. Sanz, M. L. Casanovas, A new species of the eusuchian crocodile Diplocynodon from the Eocene of Spain. *Neues Jahrb. Geol. Palaontol. Abh.* **187**, 1 (1992).

127. L. Ginsburg, C. Bulot, Les Diplocynodon (Reptilia, Crocodylia) de l'Orléanien (Miocène inférieur à moyen) de France. *Geodiversitas* **19**, 107 (1997).

128. J. Martin, M. Gross, Taxonomic clarification of Diplocynodon Pomel, 1847 (Crocodilia) from the Miocene of Styria, Austria. *Neues Jahrb. Geol. Palaontol. Abh.* **261**, 177–193 (2011). [doi:10.1127/0077-7749/2011/0159](doi:10.1127/0077-7749/2011/0159)

129. M. Delfino, T. Smith, Reappraisal of the morphology and phylogenetic relationships of the Middle Eocene alligatoroid Diplocynodon Deponiae (Frey, Laemmert, and Riess, 1987) based on a three-dimensional specimen. *J. Vertebr. Paleontol.* **32**, 1358–1369 (2012). [doi:10.1080/02724634.2012.699484](doi:10.1080/02724634.2012.699484)

130. C. C. Mook, A new crocodilian from the Lance Formation. *Am. Mus. Novit.* **1128**, 1 (1941).

131. A. D. Buscalioni, F. Ortega, D. Vasse, New crocodiles (Eusuchia: Alligatoroidea) from the Upper Cretaceous of southern Europe. *C. R. Acad. Sci. Ser. II A* **325**, 525 (1997). [doi:10.1016/S1251-8050(97)89872-2](doi:10.1016/S1251-8050(97)89872-2)

132. J. E. Martin, New material of the Late Cretaceous globidontan Acynodon iberoccitanus (Crocodylia) from southern France. *J. Vertebr. Paleontol.* **27**, 362–372 (2007). doi:10.1671/0272-4634(2007)27[362:NMOTLC]2.0.CO;2

133. J. E. Martin, Allodaposuchus Nopsca, 1928 (Crocodylia, Eusuchia), from the Late Cretaceous of southern France and its relationships to Alligatoroidea. *J. Vertebr. Paleontol.* **30**, 756–767 (2010). doi:10.1080/02724631003758318

134. M. Delfino, J. E. Martin, E. Buffetaut, A new species of Acynodon (Crocodylia) from the Upper Cretaceous (Santonian-Campanian) of Villaggio del Pescatore, Italy. *Palaeontology* **51**, 1091–1106 (2008). doi:10.1111/j.1475-4983.2008.00800.x

135. J. E. Martin, E. Buffetaut, Crocodilus affuvelensis Matheron, 1869 from the Late Cretaceous of southern France: A reassessment. *Zool. J. Linn. Soc.* **152**, 567–580 (2008). doi:10.1111/j.1096-3642.2007.00358.x

136. E. Puértolas, J. I. Canudo, P. Cruzado-Caballero, A new crocodylian from the late Maastrichtian of Spain: Implications for the initial radiation of crocodyloids. *PLOS ONE* **6**, e20011 (2011). Medline doi:10.1371/journal.pone.0020011

137. A. D. Buscalioni, F. Ortega, D. B. Weishampel, C. M. Jianu, A revision of the crocodyliform *Allodaposuchus precedens* from the Upper Cretaceous of the Hateg Basin, Romania. Its relevance in the phylogeny of Eusuchia. *J. Vertebr. Paleontol.* **21**, 74–86 (2001). doi:10.1671/0272-4634(2001)021[0074:AROTCA]2.0.CO;2

138. M. Delfino, V. Codrea, A. Folie, P. Dica, P. Godefroit, T. Smith, A complete skull of *Allodaposuchus precedens* Nopcsa, 1928 (Eusuchia) and a reassessment of the morphology of the taxon based on the Romanian remains. *J. Vertebr. Paleontol.* **28**, 111–122 (2008). doi:10.1671/0272-4634(2008)28[111:ACSOAP]2.0.CO;2

139. C. A. Brochu, D. C. Parris, B. S. Grandstaff, R. K. Denton Jr., W. B. Gallagher, A new species of Borealosuchus (Crocodyliformes, Eusuchia) from the Late Cretaceous-Early Paleogene of New Jersey. *J. Vertebr. Paleontol.* **32**, 105–116 (2012). doi:10.1080/02724634.2012.633585

140. R. J. Butler, S. L. Brusatte, M. Reich, S. J. Nesbitt, R. R. Schoch, J. J. Hornung, The sail-backed reptile Ctenosauriscus from the latest Early Triassic of Germany and the timing

and biogeography of the early archosaur radiation. *PLOS ONE* **6**, e25693 (2011). Medline doi:10.1371/journal.pone.0025693

141. S. J. Nesbitt, The early evolution of archosaurs: Relationships and the origin of major clades. *Bull. Am. Mus. Nat. Hist*. *352* (2011).

142. S. J. Nesbitt, J. Liu, C. Li, A sail-backed suchian from the Heshanggou Formation (Early Triassic: Olenekian) of China. *Earth Environ. Sci. Trans. R. Soc. Edinburgh* **101**, 271 (2010). doi:10.1017/S1755691011020044

143. S. L. Brusatte, G. Niedźwiedzki, R. J. Butler, Footprints pull origin and diversification of dinosaur stem lineage deep into Early Triassic. *Proc. R. Soc. B* **278**, 1107 (2011). doi:10.1098/rspb.2010.1746

144. T. R. Lyson, G. S. Bever, B. A. S. Bhullar, W. G. Joyce, J. A. Gauthier, Transitional fossils and the origin of turtles. *Biol. Lett.* **6**, 830–833 (2010). Medline doi:10.1098/rsbl.2010.0371

145. T. R. Lyson, G. S. Bever, T. M. Scheyer, A. Y. Hsiang, J. A. Gauthier, Evolutionary origin of the turtle shell. *Curr. Biol.* **23**, 1113–1119 (2013). Medline doi:10.1016/j.cub.2013.05.003

146. J. Gauthier, A. G. Kluge, T. Rowe, Amniote phylogeny and the importance of fossils. *Cladistics* **4**, 105–209 (1988). doi:10.1111/j.1096-0031.1988.tb00514.x

147. M. S. Y. Lee, Pareiasaur phylogeny and the origin of turtles. *Zool. J. Linn. Soc.* **120**, 197–280 (1997). doi:10.1111/j.1096-3642.1997.tb01279.x

148. O. Rieppel, R. R. Reisz, The origin and early evolution of turtles. *Annu. Rev. Ecol. Syst.* **30**, 1–22 (1999). doi:10.1146/annurev.ecolsys.30.1.1

149. R. L. Carroll, in *Morphology and Evolution of Turtles*, D. B. Brinkman, P. A. Holroyd, J. D. Gardner, Eds. (Springer, Dordrecht, Netherlands, 2013), pp. 19–38.

150. Y. Katsu, E. L. Braun, L. J. Guillette Jr., T. Iguchi, From reptilian phylogenomics to reptilian genomes: Analyses of c-Jun and DJ-1 proto-oncogenes. *Cytogenet. Genome Res.* **127**, 79–93 (2009). Medline doi:10.1159/000297715

151. D. J. Field, J. A. Gauthier, B. L. King, D. Pisani, T. R. Lyson, K. J. Peterson, Toward consilience in reptile phylogeny: miRNAs support an archosaur, not lepidosaur, affinity for turtles. *Evol. Dev.* **16**, 189–196 (2014). Medline doi:10.1111/ede.12081

152. J. A. Clarke, C. P. Tambussi, J. I. Noriega, G. M. Erickson, R. A. Ketcham, Definitive fossil evidence for the extant avian radiation in the Cretaceous. *Nature* **433**, 305–308 (2005). Medline doi:10.1038/nature03150

153. N. R. Longrich, T. Tokaryk, D. J. Field, Mass extinction of birds at the Cretaceous-Paleogene (K-Pg) boundary. *Proc. Natl. Acad. Sci. U.S.A.* **108**, 15253–15257 (2011). Medline doi:10.1073/pnas.1110395108

154. A. Feduccia, Avian extinction at the end of the Cretaceous: Assessing the magnitude and subsequent explosive radiation. *Cretac. Res.* **50**, 1–15 (2014). doi:10.1016/j.cretres.2014.03.009

155. M. S. Lee, A. Cau, D. Naish, G. J. Dyke, Morphological clocks in paleontology, and a mid-Cretaceous origin of crown Aves. *Syst. Biol.* **63**, 442–449 (2014). Medline doi:10.1093/sysbio/syt110

156. T. Sicheritz-Pontén, S. G. Andersson, A phylogenomic approach to microbial evolution. *Nucleic Acids Res.* **29**, 545–552 (2001). Medline doi:10.1093/nar/29.2.545

157. T. F. Smith, M. S. Waterman, Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981). Medline doi:10.1016/0022-2836(81)90087-5

158. R. C. Edgar, MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113 (2004). Medline doi:10.1186/1471-2105-5-113

159. K. Katoh, H. Toh, Recent developments in the MAFFT multiple sequence alignment program. *Brief. Bioinform.* **9**, 286–298 (2008). Medline doi:10.1093/bib/bbn013

160. A. R. Subramanian, M. Kaufmann, B. Morgenstern, DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment. *Algorithms Mol. Biol.* **3**, 6 (2008). Medline doi:10.1186/1748-7188-3-6

161. G. Landan, D. Graur, Heads or tails: A simple reliability check for multiple sequence alignments. *Mol. Biol. Evol.* **24**, 1380–1383 (2007). [Medline](#) [doi:10.1093/molbev/msm060](#)

162. I. M. Wallace, O. O'Sullivan, D. G. Higgins, C. Notredame, M-Coffee: Combining multiple sequence alignment methods with T-Coffee. *Nucleic Acids Res.* **34**, 1692–1699 (2006). [Medline](#) [doi:10.1093/nar/gkl091](#)

163. O. Gascuel, BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. *Mol. Biol. Evol.* **14**, 685–695 (1997). [Medline](#) [doi:10.1093/oxfordjournals.molbev.a025808](#)

164. H. Akaike, A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19**, 716–723 (1974). [doi:10.1109/TAC.1974.1100705](#)

165. J. Huerta-Cepas, S. Capella-Gutierrez, L. P. Pryszcz, I. Denisov, D. Kormes, M. Marcet-Houben, T. Gabaldón, PhylomeDB v3.0: An expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. *Nucleic Acids Res.* **39**, D556–D560 (2011). [Medline](#) [doi:10.1093/nar/gkq1109](#)

166. H. Shimodaira, M. Hasegawa, CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* **17**, 1246–1247 (2001). [Medline](#) [doi:10.1093/bioinformatics/17.12.1246](#)

167. S. R. Eddy, Accelerated profile HMM searches. *PLOS Comput. Biol.* **7**, e1002195 (2011). [Medline](#) [doi:10.1371/journal.pcbi.1002195](#)

168. I. Medina, J. Carbonell, L. Pulido, S. C. Madeira, S. Goetz, A. Conesa, J. Tárraga, A. Pascual-Montano, R. Nogales-Cadenas, J. Santoyo, F. García, M. Marbà, D. Montaner, J. Dopazo, Babelomics: An integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Res.* **38**, W210–W213 (2010). [Medline](#) [doi:10.1093/nar/gkq388](#)

169. F. Supek, M. Bošnjak, N. Škunca, T. Šmuc, REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE* **6**, e21800 (2011). [Medline](#) [doi:10.1371/journal.pone.0021800](#)

170. S. Song, L. Liu, S. V. Edwards, S. Wu, Resolving conflict in eutherian mammal phylogeny using phylogenomics and the multispecies coalescent model. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 14942–14947 (2012). Medline doi:10.1073/pnas.1211733109

171. E. C. Teeling, S. B. Hedges, Making the impossible possible: Rooting the tree of placental mammals. *Mol. Biol. Evol.* **30**, 1999–2000 (2013). Medline doi:10.1093/molbev/mst118

172. H. Nishihara, S. Maruyama, N. Okada, Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 5235–5240 (2009). Medline doi:10.1073/pnas.0809297106

173. S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, S. L. Salzberg, Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004). Medline doi:10.1186/gb-2004-5-2-r12

174. B. J. Raney, T. R. Dreszer, G. P. Barber, H. Clawson, P. A. Fujita, T. Wang, N. Nguyen, B. Paten, A. S. Zweig, D. Karolchik, W. J. Kent, Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser. *Bioinformatics* **30**, 1003–1005 (2014). Medline

175. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010). Medline doi:10.1093/bioinformatics/btq033

176. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013). Medline doi:10.1093/molbev/mst010

177. E. L. Sonnhammer, G. von Heijne, A. Krogh, A hidden Markov model for predicting transmembrane helices in protein sequences. In *ISMB '98 Proceedings of the 6th International Conference on Intelligent Systems for Molecular Biology* (1998), pp. 175–182.

178. M. K. Fujita, S. V. Edwards, C. P. Ponting, The Anolis lizard genome: An amniote genome without isochores. *Genome Biol. Evol.* **3**, 974–984 (2011). Medline doi:10.1093/gbe/evr072

179. S. M. Fullerton, A. Bernardo Carvalho, A. G. Clark, Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* **18**, 1139–1142 (2001). Medline doi:10.1093/oxfordjournals.molbev.a003886

180. R. C. Hardison, K. M. Roskin, S. Yang, M. Diekhans, W. J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe, S. Schwartz, T. S. Furey, S. Whelan, N. Goldman, A. Smit, W. Miller, F. Chiaromonte, D. Haussler, Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* **13**, 13–26 (2003). Medline doi:10.1101/gr.844103

181. D. Mouchiroud, G. D'Onofrio, B. Aïssani, G. Macaya, C. Gautier, G. Bernardi, The distribution of genes in the human genome. *Gene* **100**, 181–187 (1991). Medline doi:10.1016/0378-1119(91)90364-H

182. B. Boussau, M. Gouy, Efficient likelihood computations with nonreversible models of evolution. *Syst. Biol.* **55**, 756–768 (2006). Medline doi:10.1080/10635150600975218

183. J. Romiguier, V. Ranwez, E. J. Douzery, N. Galtier, Contrasting GC-content dynamics across 33 mammalian genomes: Relationship with life-history traits and chromosome sizes. *Genome Res.* **20**, 1001–1009 (2010). Medline doi:10.1101/gr.104372.109

184. L. Duret, P. F. Arndt, The impact of recombination on nucleotide substitutions in the human genome. *PLOS Genet.* **4**, e1000071 (2008). Medline doi:10.1371/journal.pgen.1000071

185. M. dos Reis, J. Inoue, M. Hasegawa, R. J. Asher, P. C. Donoghue, Z. Yang, Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proc. R. Soc. B* **279**, 3491–3500 (2012). Medline doi:10.1098/rspb.2012.0683

186. R. R. Reisz, S. P. Modesto, D. M. Scott, A new Early Permian reptile and its significance in early diapsid evolution. *Proc. R. Soc. B* **278**, 3731–3737 (2011). Medline doi:10.1098/rspb.2011.0439

187. M. E. Jones, C. L. Anderson, C. A. Hipsley, J. Müller, S. E. Evans, R. R. Schoch, Integration of molecules and new fossils supports a Triassic origin for Lepidosauria (lizards, snakes, and tuatara). *BMC Evol. Biol.* **13**, 208 (2013). Medline doi:10.1186/1471-2148-13-208

188. R. Hirayama, Oldest known sea turtle. *Nature* **392**, 705–708 (1998). doi:10.1038/33669

189. Y. Niimura, M. Nei, Evolutionary dynamics of olfactory receptor genes in fishes and tetrapods. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 6039–6044 (2005). Medline doi:10.1073/pnas.0501922102

190. Y. Niimura, On the origin and evolution of vertebrate olfactory receptor genes: Comparative genome analysis among 23 chordate species. *Genome Biol. Evol.* **1**, 34–44 (2009). Medline doi:10.1093/gbe/evp003

191. M. Nei, Y. Niimura, M. Nozawa, The evolution of animal chemosensory receptor gene repertoires: Roles of chance and necessity. *Nat. Rev. Genet.* **9**, 951–963 (2008). Medline doi:10.1038/nrg2480

192. T. Joanen, L. L. Mcnease, Ecology and physiology of nesting and early development of the American alligator. *Am. Zool.* **29**, 987 (1989).

193. M. W. J. Ferguson, in *Biology of the Reptilia*, C. Gans, F. Billett, P. F. A. Maderson, Eds. (Wiley, New York, 1985), vol. 14, pp. 329–492.

194. S. E. Klause, thesis, North Carolina State University (1984).

195. J. Caldwell, "World trade in crocodilian skins 2008–2010" (UNEP World Conservation Monitoring Centre, Cambridge, 2012).

196. G. H. Dalrymple, Growth of American alligators in the Shark Valley region of Everglades National Park. *Copeia* **1996**, 212 (1996). doi:10.2307/1446962

197. T. Joanen, Nesting ecology of alligators in Louisiana. *Proceedings of the Annual Conference of Southeastern Association of Game and Fish Commissioners* **23**, 141 (1969).

198. C. L. Abercrombie, in *Crocodiles: Their Ecology, Management, and Conservation*, P. Hall, Ed. (International Union for the Conservation of Nature and Natural Resources, Gland, Switzerland, 1989), pp. 1–16.

199. G. J. W. Webb, H. Messel, J. Crawford, M. J. Yerbury, Growth rates of *Crocodylus porosus* (Reptilia: Crocodilia) from Arnhem Land, Northern Australia. *Wildl. Res.* **5**, 385 (1978). doi:10.1071/WR9780385

200. G. J. W. Webb, S. C. Manolis, M. L. Brien, in *Crocodiles. Status Survey and Conservation Action Plan*, S. C. Manolis, C. Stevenson, Eds. (Crocodile Specialist Group, Darwin, Australia, 2010), pp. 99–113.

201. G. J. W. Webb, H. Messel, W. Magnusson, The nesting of *Crocodylus porosu*s in Arnhem Land, Northern Territory. *Copeia* **1977**, 238 (1977). doi:10.2307/1443905

202. S. C. Stirrat, D. Lawson, W. J. Freeland, R. Morton, Monitoring *Crocodylus porosus* populations in the Northern Territory of Australia: A retrospective power analysis. *Wildl. Res.* **28**, 547 (2001). doi:10.1071/WR00079

203. H. R. Bustard, S. Maharana, Size at first breeding in the gharial (*Gavialis gangeticus* (Gmelin)) (Reptilia: Crocodilia) in captivity. *J. Bombay Nat. Hist. Soc.* **79**, 206 (1982).

204. A. Das, A. K. Das, P. K. Sarma, A. K. Singh, "An assessment of assisted recovery of gharial (*Gavialis gangeticus*) in river systems of Northeast India. Final technical report" (Aaranyak, Assam, India, 2011).

205. R. Whitaker, in *Wildlife Management: Crocodiles and Alligators*, G. J. W. Webb, S. C. Manolis, P. J. Whitehead, Eds. (Surrey Beatty, Chipping Norton, NSW, Australia, 1987), pp. 63–72.

206. S. A. Hussain, Reproductive success, hatchling survival and rate of increase of gharial *Gavialis gangeticus* in National Chambal Sanctuary, India. *Biol. Conserv.* **87**, 261–268 (1999). doi:10.1016/S0006-3207(98)00065-2