Genome Biology

OPEN LETTER

Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes

John A St John[1], Edward L Braun[2], Sally R Isberg[3,5], Lee G Miles[5], Amanda Y Chong[5], Jaime Gongora[5], Pauline Dalzell[4,5], Christopher Moran[5], Bertrand Bed'Hom[6], Arkhat Abzhanov[7], Shane C Burgess[8], Amanda M Cooksey[8], Todd A Castoe[9], Nicholas G Crawford[10], Llewellyn D Densmore[11], Jennifer C Drew[12], Scott V Edwards[7], Brant C Faircloth[13], Matthew K Fujita[7], Matthew J Greenwold[14], Federico G Hoffmann[8,15], Jonathan M Howard[16], Taisen Iguchi[17], Daniel E Janes[18,19], Shahid Yar Khan[1], Satomi Kohno[20], AP Jason de Koning[9], Stacey L Lance[21], Fiona M McCarthy[22], John E McCormack[23], Mark E Merchant[24], Daniel G Peterson[8,25], David D Pollock[9], Nader Pourmand[1], Brian J Raney[26], Kyria A Roessler[1], Jeremy R Sanford[16], Roger H Sawyer[14], Carl J Schmidt[27], Eric W Triplett[28], Tracey D Tuberville[21], Miryam Venegas-Anaya[11], Jason T Howard[29], Erich D Jarvis[29], Louis J Guillette Jr[20], Travis C Glenn[30], Richard E Green[1], and David A Ray*[8,15]

Abstract

The International Crocodilian Genomes Working Group (ICGWG) will sequence and assemble the American alligator (*Alligator mississippiensis*), saltwater crocodile (*Crocodylus porosus*) and Indian gharial (*Gavialis gangeticus*) genomes. The status of these projects and our planned analyses are described.

Keywords Genomics, evolution, Crocodylia, Archosauria, amniote

The importance of reptilian genomics

The study of reptilian genomes is essential if we are to understand the patterns of genomic evolution across amniotes (mammals, birds and non-avian reptiles). Non-avian reptiles differ from mammals and birds in several ways: they have diverse sex-determining systems, are exothermic ('cold blooded') and have extreme physiology. Non-avian reptiles are divided into four extant orders: Crocodylia (crocodiles and alligators; approximately 25 species), Sphenodontia (tuatara; two species), Squamata (lizards and snakes; approximately 7,900 species) and Testudines (turtles; approximately 300 species). The clade's most recent common ancestor is thought to have lived around 275 million years ago (Mya) [1], and birds (class Aves) are nested within reptiles (class Reptilia) (Figure 1). Although they are more diverse than birds and mammals, non-avian reptiles have not been a major focus of genome sequencing efforts [2,3]. The green anole (*Anolis carolinensis*) is the only non-avian reptilian genome sequence published to date [4]. There are, however, ongoing initiatives to sequence the genomes of the painted turtle (*Chrysemys picta*; (see NHGRI Genome Sequencing Proposals [5], the garter snake (*Thamnophis sirtalis* [6]), the king cobra (*Ophiophagus hannah*; M.K. Richardson, personal communication and the Burmese python (*Python molurus bivittatus* [7]). Although these projects will provide considerable insight into the evolution of both reptilian and amniote genomes, they only begin to address the diversity represented within reptiles, and do not include any crocodilians.

Order Crocodylia is a key group within Reptilia and genome drafts from crocodilians would provide insights into ancestral reptilian and amniote genomes. These genome assemblies will also enable more detailed inferences on the evolution of three additional lineages of substantial interest to vertebrate biologists: dinosaurs, pterosaurs and birds. Crocodilians and birds are the only extant members of Archosauria (a clade that also includes dinosaurs and pterosaurs along with several extinct lineages) [8]. Among archosaurs, only the genomes of chicken (*Gallus gallus* [9]), turkey (*Meleagris gallopavo* [10]) and zebra finch (*Taeniopygia guttata* [11]) have been sequenced, although several additional avian genomes, such as the Mallard duck (*Anas platyrhynchos* [12], budgerigar (*Melopsittacus undulatus*, a type of parrot) and a set of other avian taxa [13] are currently underway [14]. Crocodilians are the best extant outgroup for comparative analysis of avian genomes, and, as such,

*Correspondence: dray@bch.msstate.edu
[15]Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762, USA
Full list of author information is available at the end of the article

would substantially enhance analyses of the large set of bird genomes that are expected to be available shortly. Avian and crocodilian genomes provide the best hope for elucidating the gene and genomic properties of dinosaurs and other extinct archosaurs, about which we have learned surprising amounts (for example, genome size and limited protein sequences) considering we have no access to the DNA of these organisms [15-19]. In the broadest sense, Crocodylia represent an important vertebrate clade, and their genomes hold information that will illuminate the underlying relationships among all amniotes. In addition, crocodilians present several interesting biological questions that can be approached from a genomic perspective, many of these will be discussed below.

Background on crocodilians and project justification

The order Crocodylia, which typically refers to the clade that includes the extant crocodilians [20], is an ecologically successful group of reptiles that originated in the mid- to upper-Cretaceous period (approximately 100 Mya) [21,22]. Crocodilians are apex predators in the marine and freshwater habitats where they reside. They play a major role in warm-water ecosystems throughout the world. Extant crocodilians are members of a larger group, termed the Crocodylomorpha, that appeared in the fossil record by the upper Triassic (about 200-250 Mya) [8,1], a date coincident with molecular estimates of the avian-crocodilian divergence [2,22,23]. Crocodylia is divided into three families with extant members, Alligatoridae (alligators and caimans), Crocodylidae (crocodiles) and Gavialidae (gharials) [21,23]; the Gavialidae are traditionally thought to be the outgroup of a clade comprising Alligatoridae and Crocodylidae [21]. However, recent phylogenetic analyses of both molecular data [22,24] and combined molecular and morphological data [25] support a closer relationship between Crocodylidae and Gavialidae (Figure 1).

Crocodilians have been a part of the human narrative for centuries, appearing in modern popular culture (for example, the wildlife documentary series *The Crocodile Hunter*), scientific documentaries, as ancient mummies and in cave paintings. They are prized for their hides and meat, and some species, such as the American alligator, the Nile crocodile (*Crocodylus niloticus*) and the saltwater crocodile, are ranched (that is their eggs are brought in from the wild) and/or farmed (in which captive breeding stock produce the eggs). Globally, crocodilians are a source of trade worth more than $US500 million [26]. However, crocodilians likely have their most profound economic impact as tourist attractions [27,28]. Thoughtful ecotourism could be the best hope for saving endangered crocodilians, such as the critically endangered gharial, from extinction and their habitats from destruction.

Given their popularity, their status as the sister group of dinosaurs, and their inherent public fascination, efforts focused on crocodilian genomics are ideally suited for education and outreach focused on evolution and comparative genomics. Indeed, the preliminary data from our efforts has been used in a pilot genomics course at the University of Florida that integrates with undergraduate research. The consortium plans to make material for genomics pedagogy and public outreach available in parallel with the release of the genome assemblies.

In addition to their ecological, sociological and economic significance, crocodilians have genomes that will be useful sources of data for biological and biomedical research. Alligator serum has been shown to contain broad spectrum antibiotic peptides [29-32]. The American alligator has been used extensively as a model for examining the environmental impact of various contaminants, including endocrine disrupting xenobiotics [33-36]. Crocodilians represent important research organisms for diverse fields that include evolution and phylogenetics [25,37-39], functional morphology [37,40], osmoregulation [37], sex determination [41-45], hybridization [46-48] and population genetics [49-51]. To provide the genomic resources necessary to expand our understanding of these fascinating organisms, the ICGWG is obtaining and assembling genome sequences for the American alligator, saltwater crocodile, and gharial, one representative from each of the extant crocodilian families. For further information about the project and preliminary assemblies, see Ref. [52].

Properties of crocodilian genomes and available genomic resources

Short of whole genome sequencing, much work has been done on crocodilian genomes, especially the American alligator and Australian saltwater crocodile. The genome of the American alligator is approximately 2.5 gigabases [53] comprising 16 pairs of chromosomes [54,55]. The genome size of the saltwater crocodile is around 2.78 gigabases [56] with 17 pairs of chromosomes [54,57]. The genome size of the gharial is currently unknown, although it is likely to be approximately 2-3 gigabases, given the genome sizes of other crocodilians. Like the American alligator, the gharial has 16 chromosomes [54]. Unlike organisms with genetic sex-determination systems, crocodilians are not thought to have sex chromosomes [54]. Instead sex is determined by incubation temperature of the egg [42]. Although microchromosomes are common among other reptiles (including birds), and there is striking variation in chromosome sizes within crocodilians, the smallest crocodilian chromosomes are not generally regarded as small enough to be classified as microchromosomes [54,58,57,55].

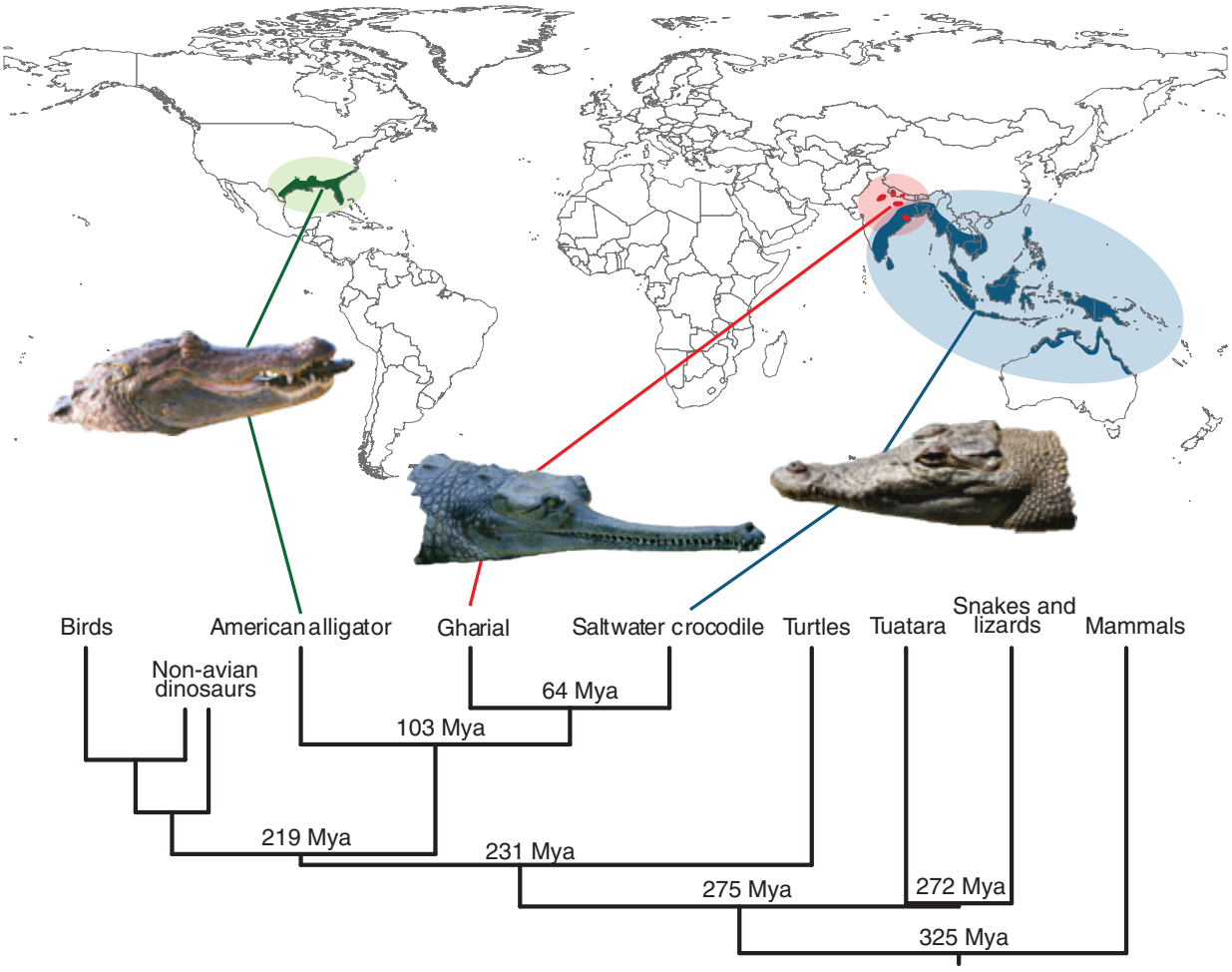


Figure 1. Amniote phylogeny emphasizing the crocodilians. The geographic ranges of the three crocodilians of interest are shown, along with approximate times of divergence of each group based upon the Timetree of Life [1]. On the basis of the fossil record, the origins of dinosaurs and birds were Triassic and upper Jurassic, respectively [86], and birds from within dinosaurs [86,87]. The phylogenetic position of turtles is unclear [2,88,89], however for simplicity we chose the consensus estimated position and divergence time presented in the Timetree of Life [1]. The photos of the American alligator (*Alligator mississippiensis*), the saltwater crocodile (*Crocodylus porosus*) were kindly provided by Louis Guillette and the photo of the Indian gharial (*Gavialis gangeticus*) was provided by Alan Wolf . Mya; million years ago.

As in birds, the most common transposable elements (TEs) in crocodilian genomes are Long INterspersed Elements (LINEs) of the chicken repeat 1 (CR1) family [59]. Earlier studies indicated that the majority of CR1 LINEs in crocodilians are fairly short (typically <2kbp [59]). Indeed, our efforts to identify novel repeats in preliminary saltwater crocodile and American alligator genome assemblies show that the most abundant repeats in the current assemblies are less than 1kbp (Figure 2). The observation that this relatively well-characterized and short class of TE insertions is the predominant family of repeats in crocodilians suggest that assembling the genomes of these organisms will be a manageable project, compared with a typical repeat-rich mammalian genome that contains a greater proportion of longer repetitive elements.

Libraries of bacterial artificial chromosomes (BACs) are available for all three species of interest and these will be used for each genome project. The American alligator BAC library currently has about 10X clone coverage [60], the saltwater crocodile library has approximately 3.7X clone coverage [56] and the gharial library has about 5.7X clone coverage, assuming it is a 2.7 gigabase genome (X. Shan, unpublished data). Several large-scale nucleotide datasets have been collected for the American alligator, including 21 assembled BAC sequences completed through the NISC Comparative Sequencing Initiative [61], and 3,276 Sanger BAC-end reads [59]. A

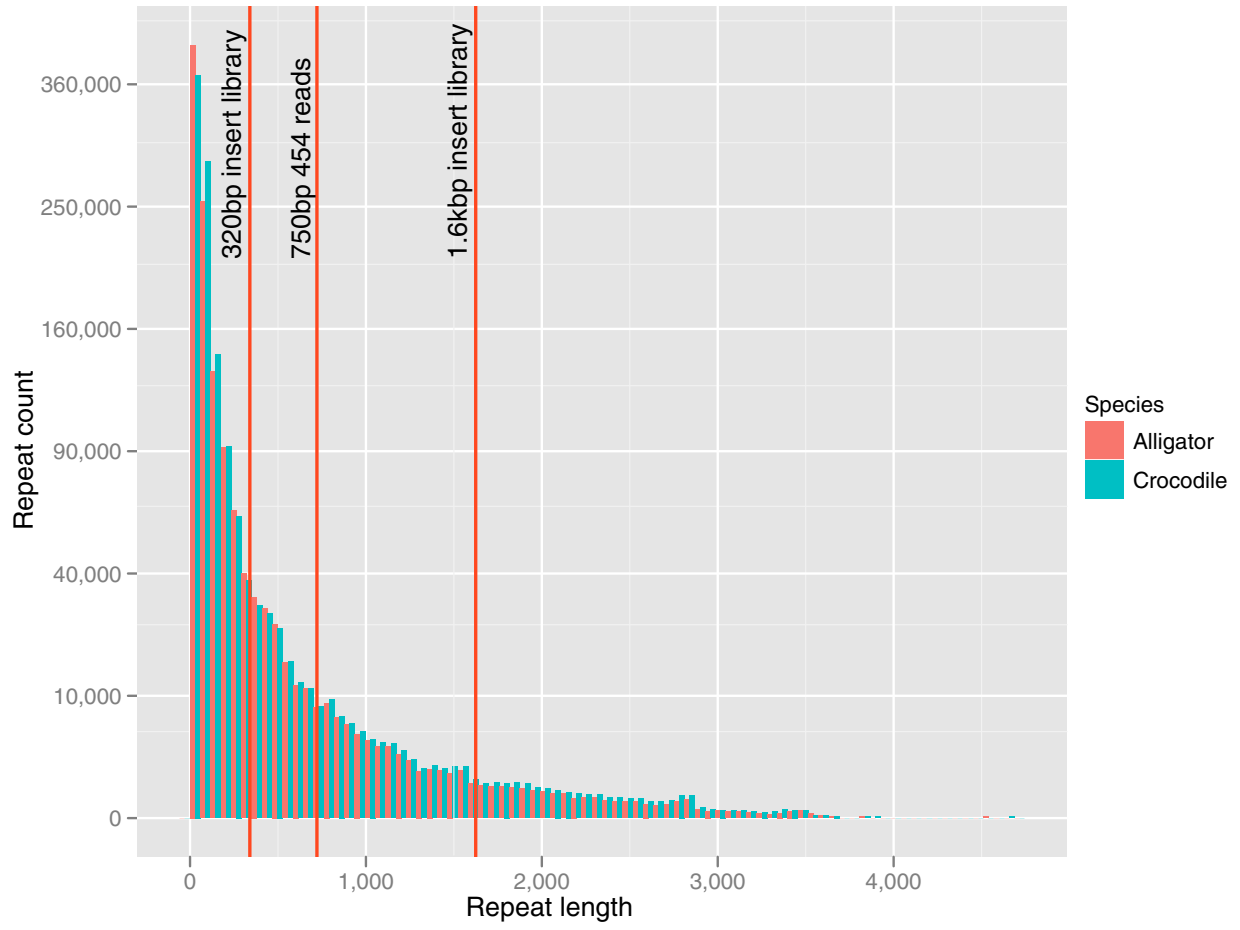


Figure 2. The distribution of repeats of different length in the alligator and crocodile assemblies. Overlaid are some of the library insert size or fragment sizes we have made for the various assemblies. Note however that the current crocodile assembly in this figure does not include the 454 data.

linkage map based on microsatellite loci [62] for the saltwater crocodile is also available. Additionally some saltwater crocodile microsatellite loci have been mapped by fluorescence *in situ* hybridization (FISH) to physical chromosomes using fosmids and BACs ([58] and P. Dalzell unpublished data), which will facilitate anchoring portions of the genome assembly to chromosomes.

In addition to genomic sequences and mapping information, both Sanger and 454 transcriptome data for the crocodile and alligator are available [63,64]. Transcriptome data will be further augmented by a diversity of tissue-specific cDNA libraries from multiple species that will be sequenced using Illumina RNA-seq to assist gene annotations. The cDNA sequences will also enable further scaffold ordering and orientation for transcripts that are split between multiple genomic fragments [65]. We will use these legacy and new data to further improve the initial *de novo* assemblies. To view the preliminary assemblies, see Ref. [52].

Sequencing strategy for the three crocodilian genomes

Owing to the availability of diverse legacy data, we are pursuing different strategies for the sequencing and assembly of each genome, as described below.

For the American alligator genome, we are following the Allpaths-LG recommended pipeline [66] of a combination of high coverage pairs of overlapping reads with a second, moderate coverage, longer insert mate-pair library. This pipeline has yielded good results with a variety of assemblies including *de novo* reassemblies of mouse and human [66], and was successfully employed in an independently evaluated genome assembly contest [67]. We have combined approximately 50x coverage from an overlapping, Illumina, short-insert library with about 20x coverage from an Illumina 2kbp mate-pair library. To investigate genetic variation and increase coverage, we will combine these reads with a set of short, non-overlapping 2x100 bp Illumina reads at approximately 50x

coverage. In addition to providing deeper coverage, these data will also provide information about genetic variation in American alligators due to single nucleotide polymorphism differences between the diploid chromosomes of an individual. We will further scaffold the assembly using low coverage BAC-end sequences, and we will carry out FISH mapping to assign scaffolds to chromosomes.

To sequence the saltwater crocodile genome, we are combining high coverage Illumina short insert sequencing with low coverage 454 libraries in a hybrid approach, similar to that used for the turkey genome [10]. We currently have about 80x coverage from a non-overlapping, short-insert library and an additional 40x from an overlapping short-insert library. We also plan to generate about 20x coverage from an Illumina 2kbp mate-pair library. To supplement the Illumina data, we have generated 1x coverage of unpaired 454 reads (about 700bp in length), and plan to generate an additional 2x coverage from 3kbp and 6kbp paired 454 reads. We will also end-sequence the crocodile BAC library using a method similar to the fosmid-based ShARC method described by Gnerre *et al*. [66]. Some of these BACs are known to contain microsatellite DNA markers used in the crocodile linkage map [62] and others have already been FISH mapped to chromosomes in the crocodile [58]. We will integrate this information for scaffolding and assigning scaffolds to chromosomes. As with the American alligator genome, we are also generating transcriptome data for the saltwater crocodile for both annotation and scaffolding purposes. We will also use the 454 brain transcriptome data that exists for the American alligator [64] and the Nile crocodile [68] in our analyses. We will use these EST and RNA-seq data, along with the other resources described above, to further order and orient scaffolds within the assembly.

Finally, we will assemble the gharial genome using a hybrid approach similar to that used for the saltwater crocodile. To do this, we have generated 40x coverage from an overlapping short-insert library. This will be combined with sequences from 400 bp and 700 bp paired-end Illumina libraries sequenced to give approximately 30x coverage, as well as 2-3x genome coverage consisting of 454 shotgun reads and 3kbp and 6kbp paired-end 454 libraries with FLX+ reads. Finally we will generate approximately 20x coverage from an Illumina 2kbp mate-pair library. The gharial is a critically endangered species, making it nearly impossible to collect a wide variety of tissues for transcriptome data. Nonetheless, we have collected blood, which will be used to generate Illumina RNA-seq data. As with the American alligator and saltwater crocodile, we will use *de novo* assembled transcripts to improve the assembly.

Project timeline and goals

The first phase of our sequencing effort, in which we generate high coverage short insert and overlapping libraries, has been completed for American alligator and saltwater crocodile and is ongoing for the Indian gharial. The data generated for alligator and crocodile were used to generate early draft assemblies for those genomes. The second phase will involve generating longer distance mate-pair libraries and BAC-end sequences to improve the assemblies. We plan to have the data gathered for this phase by mid-March 2012. The third and final phase will involve FISH mapping the BACs to assign scaffolds to chromosomes. When all three phases are completed the assemblies should be as contiguous as possible, given the combination of high coverage short distance information generated in phase one with lower coverage long distance information generated in phase two. The third phase is not critical for the most pressing questions involving crocodilian genomics; individual genes and their regulatory regions will be of primary interest, as opposed to the long-range linkage required for identifying selective sweeps. Thus we will proceed with this third phase in parallel with our other comparative genomic analyses. Once the three genomes are assembled, we will perform comparative genomic analyses both within Order Crocodylia, and among crocodilians and other members of Reptilia.

The completion of each of these phases will be publicly communicated via the website, and links to the data and assemblies will be available to researchers with restrictions as detailed below. We anticipate data collection and initial analyses to be complete by June 2012, and we plan to submit the genome paper within one year of finalizing these initial analyses. The Toronto Statement [69] suggests that there be a one-year period of initial analyses and publication, after which the broader community would be free to use this data in an unrestricted manner. Precise dates at which we complete data collection and initial analysis, and thus the beginning of the embargo period on the genome data, will be promptly posted on the website [52].

Status of the current preliminary genome assemblies

Preliminary assemblies for alligator and crocodile are available. The assembly for alligator additionally uses information from a 120x physical coverage, Illumina 1.5kbp mate-pair library. The current crocodile assembly was generated with 80x coverage from a 380bp paired-end Illumina library. The statistics for the length and contiguity of these two assemblies are shown in Table 1. These assembly statistics are on par with other early stage *de novo* assemblies using short read data [7,70].

To obtain early estimates of potential TE content, we analyzed the current assemblies using RepeatMasker and

Table 1. Overview of the current draft assemblies[a]

Genome	Estimated Length (Gbp)	Assembly Length (Gbp)	Estimated % Coverage	Contig N50 (Kbp)	Contig N90 (Kbp)	Scaffold N50 (Kbp)	Scaffold N90 (Kbp)
American alligator	2.5	2.17	86.8	28.0	6.9	106.2	22.5
Seawater crocodile	2.78	2.14	77.0	13.3	3.0	28.2	6.6
Indian gharial	2.5	N/A[b]	N/A[b]	N/A[b]	N/A[b]	N/A[b]	N/A[b]

[a]Statistics of the current draft assemblies assuming the conversion between C-value and bp is 0.987×10^9 bp/pg [90]. For this table, we calculated N50 in terms the size of our assembly rather than the estimated genome size. [b]N/A: not available as the genome sequencing and assembly is in progress.

a custom repeat library. The library consisted of all vertebrate TEs identified in RepBase [71] and a set of potential TEs identified by applying RepeatScout [72] to both raw 454 data and to the current assemblies (D. Ray, unpublished data). Consistent with earlier studies [59,73,74], much of the repetitive content of the genome comprises non-long terminal repeat (non-LTR) retrotransposons from the CR1 family (Figure 3). There is also high content of Chompy-like miniature inverted-repeat transposable elements (MITEs) [75], Penelope retrotransposons, ancient short interspersed repetitive elements (SINEs), and satellite/low complexity regions. Overall, 23.44% of the alligator and 27.22% of the crocodile genome assemblies are annotated as repetitive compared with 50.63% seen in humans. Thus, this preliminary analysis provides further evidence that these reptilian genomes might be easier to assemble than typical mammalian genomes due to their lower repeat content.

We also examined GC content across the assemblies (Figure 4). Alligators and crocodiles appear to have a higher mean GC content than many other vertebrates. Additionally their large standard deviation in GC content across contigs is similar to that of birds and mammals, suggesting that their base composition is heterogeneous and likely contains GC-rich isochores. This is unlike the situation in the lizard (*Anolis*) and frog (*Xenopus*), which lack strong isochores based upon analyses of genomic data [76], or the turtle *Trachemys scripta*, which appears to lack strong isochores based upon analyses of expressed genes [77]. However, these results are consistent with previous analyses of ESTs that suggested the existence of GC-rich isochores in the alligator genome [62,77]. Thus, these crocodilian genome data extend the results of the previous analyses and confirm the genome-wide nature of GC-content heterogeneity in crocodilian. We expect improved crocodilian genome assemblies to further illuminate the details of isochore structure in reptiles.

Quality control of intermediate assemblies and raw data

For the alligator genome, we have collected nearly 1.8 billion pairs of Illumina reads from embryos at different developmental stages that were incubated at 'male producing ' (33.5°C) and 'female producing' (30°C) temperatures. From these data, we produced a set of rigorously filtered transcript sequences that we will use to assess the completeness and contiguity of the alligator assembly. These transcripts were assembled using the OASES [78] module of velvet [79] as follows. The initial assembly of the RNA-seq paired-end reads produced 749,838 fragments. We identified the longest open reading frames from each and translated them into putative proteins. We then compared these with the set of known protein sequences in the Swiss-Prot database [80], removing proteins that were more than 10% different in length from the full length Swiss-Prot hit, this removed all but 16,972 putative transcripts. We then focused on the CDS sequence of these genes and removed sequences with less than 5x RNA-seq coverage in any 30-bp window of the sequence. This procedure yielded 2,570 high-confidence alligator CDS sequences. We used these sequences to assess the quality and completeness of the current alligator assembly with results shown in Figure 5. Overall, more than 95% of these filtered CDS sequences were full length on a single scaffold. The improvement garnered by subsequent assemblies will be assessed using these data in the same manner. We will assess the quality and completeness of crocodile and gharial genomes in a similar manner.

Because we do not yet have a set of assembled transcripts for the crocodile genome, we instead used a comparative genomics approach for quality assessment on our early assemblies. For example, we generated two pre-release draft saltwater crocodile assemblies, the second of which (here called Crocodile B) had a slightly lower N50 but a greater overall length and slightly greater mean contig size relative to the first version (here called Crocodile A). Because these statistics conflicted, we aligned the two competing versions of the saltwater crocodile genome to the chicken reference genome (UCSC galGal3) using the UCSC multiz genome alignment pipeline [81]. We then analyzed regions of the multiway alignment that overlapped chicken genes in the n-scan gene track. With these gene alignments we compared the total number of genes that could be aligned across the two assemblies and the overall level of gene fragmentation for the genes that aligned between the two assemblies (Figure 6). Based on this analysis, we

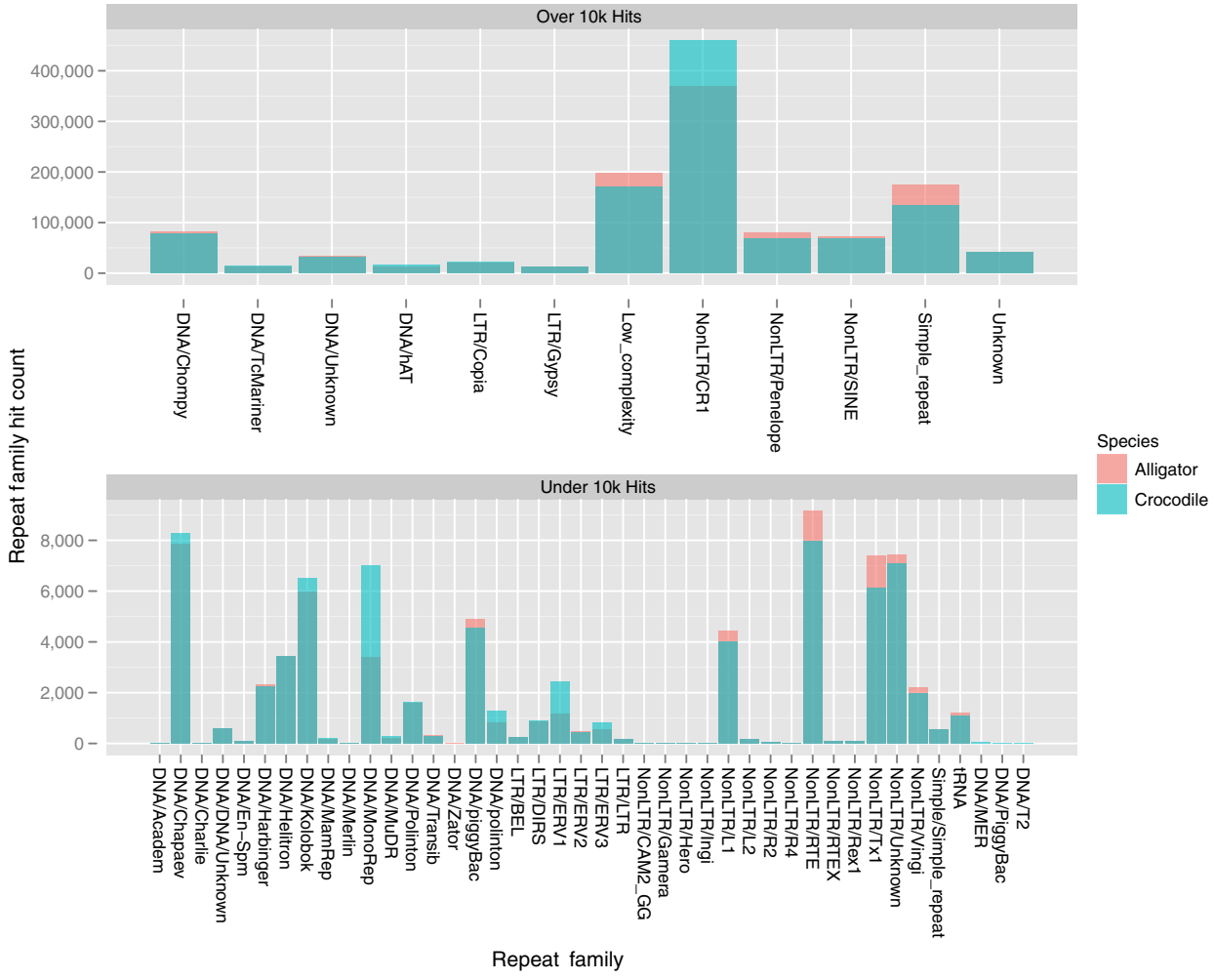


Figure 3.The size of different repeat families classified in our current alligator and crocodile assemblies. Despite more long-distance insert libraries for alligator, more repeats were found in the crocodile assembly. This strongly suggests that crocodiles have more repeats than do alligators, and perhaps the difference will become even more striking as the crocodile assembly improves.

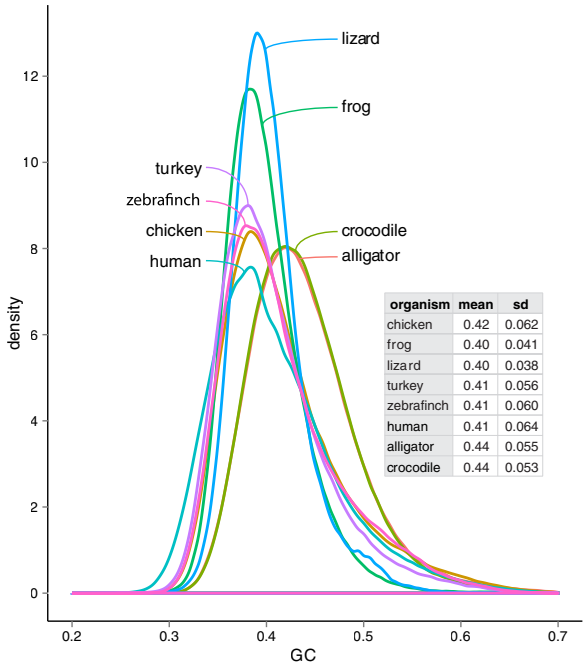
determined not only that N50 was reduced in Crocodile B but that gene contiguity was also reduced. This indicates that assembly B was not introducing false joins to achieve a higher N50, as its joins resulted in more intact gene alignments.

We will employ additional quality metrics to detect and describe the collapse of segmental duplications within our assemblies. Specifically, read-depth is a sensitive measure of this assembly artifact. Preliminary analysis suggests that such artifacts are not common in alligator or crocodile genomes (data not shown).We will employ a final form of quality control by examining the relative synteny of our three crocodilian candidate assemblies. Because alligators, crocodiles, and gharials appear to have undergone few chromosome-level rearrangements [54], we expect a high level of synteny between accurate assemblies. Once we begin scaffolding all of our assemblies with longer mate-pair and BAC data, we will assess their relative quality by measuring the effect on overall crocodilian synteny.

Planned analyses and experiments

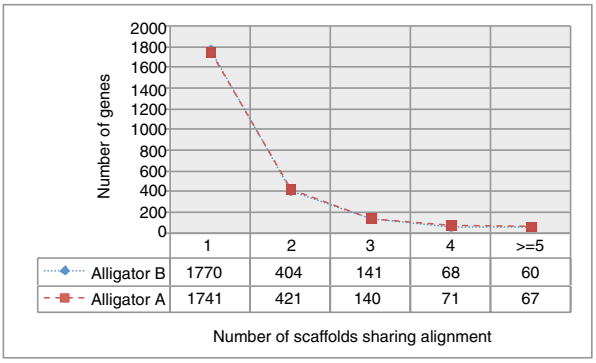
Here we outline major questions, types of analyses and analytical goals that will be included in the core publication of these completed genomes. The Toronto Statement [69] suggests these questions should be articulated to identify these topics as embargoed during preparation of the genome publication. The ICGWG will address a number of research questions at both the level of genome evolution and crocodilian biology that we describe below.

A crucial step in making genome resources useful to the scientific community is generating gene annotations. We will perform gene finding for crocodilians using the Ensembl [82] and Augustus [83] annotation pipelines and



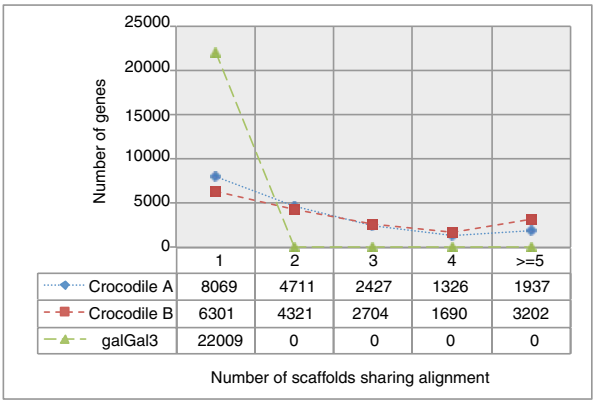
organism	mean	sd
chicken	0.42	0.062
frog	0.40	0.041
lizard	0.40	0.038
turkey	0.41	0.056
zebrafinch	0.41	0.060
human	0.41	0.064
alligator	0.44	0.055
crocodile	0.44	0.053

Figure 4. The distribution of GC proportion across several species. Note that alligators and crocodiles have a higher overall proportion of GC than many other vertebrates, as predicted by early BAC-end scans [42]. Abbreviation: SD; standard deviation.



	1	2	3	4	>=5
Alligator B	1770	404	141	68	60
Alligator A	1741	421	140	71	67

Figure 5. Using *de novo* assembled alligator transcripts, the level of gene presence and fragmentation in two alternate alligator assemblies were compared. These results suggest that the new assembly (assembly B) is an improvement over the earlier effort (assembly A).



	1	2	3	4	>=5
Crocodile A	8069	4711	2427	1326	1937
Crocodile B	6301	4321	2704	1690	3202
galGal3	22009	0	0	0	0

Figure 6. Using gene regions in a whole genome alignment to the chicken reference genome (galGal3), we compared the number of scaffolds of each assembly that the alignment of each gene is split between. Although many genomic rearrangements may exist between chicken and crocodile, assuming that breakpoints tend to happen between genes rather than within genes, this method allows us to assess the relative quality of assemblies in the same manner as Figure 5 when assembled and verified transcripts are not yet available.

combine the output. We will also partner with groups sequencing additional avian genomes and update the crocodile annotations as needed. Gene finders will initially be trained using the chicken genome and the results from the pipelines will be compared to identify accuracy at both the gene and exon level. Genes will be assigned standardized gene nomenclature based on chicken gene names where there is an unambiguous 1:1 functional ortholog, or a gene identifier in cases where this is not possible. We will also provide preliminary functional annotation for proteins and transcripts using standard Gene Ontology Consortium methods, including functional analysis of motifs and domains and manual curation of orthologs. The ICGWG will perform these analyses to complement and extend those performed by NCBI and Ensembl once the draft genomes are submitted to those organizations.

One major focus will be the large-scale structure of crocodilian genomes, focusing on the degree of syntenic conservation at different scales within these genomes. Karyotype analysis suggests a remarkable conservation of synteny among crocodilians, with the alligator and crocodile having undergone fewer than five chromosomal rearrangements visible at the microscopic level [54] despite 80 million years of evolutionary divergence. However, the level of syntenic conservation at small scales within these genomes remains unclear, and we expect our genome assemblies to illuminate this topic. Microchromosomes are absent in crocodilians [54,55,59] but present in birds, lizards and snakes, tuatara, and turtles [4,84]. This absence in crocodilians almost certainly represents a derived feature of crocodilians. We will examine the fate of these genetic units within crocodilian genomes. Do microchromosomes comprise linked components within the genomes of the only major reptilian clade without microchromosomes?

Recent work showed that the lizard, *Anolis carolinensis*, unlike other amniotes sequenced to date (with the possible exception of turtles [77]), has a homogeneous genome that lacks GC-rich isochores [76,4]. Our

preliminary analyses indicate that crocodilians have a higher GC-content and greater heterogeneity than *Anolis* (Figure 4), but these analyses are less clear regarding the scale of the observed GC-content variation. Do crocodilians have GC-rich isochores that are similar to those in mammals and birds or do the patterns of GC-content heterogeneity appear distinct?

We will also carry out a number of traditional analyses of genome content using the crocodilian genomes, focusing on repeated sequences and gene families. These analyses include the evolution of repeat families and patterns of TE proliferation. We will compare the repeat family content within crocodilian genomes and with other reptiles and amniotes. Additionally, we will conduct analyses of gene family evolution within reptiles and crocodilians to identify specific genes and other functional elements, including the identification of ultra-conserved regions and potential micro RNA sequences, with a special focus on those sequences that could have been gained or lost both within the crocodilians and in comparison to the other relevant lineages that are now available for investigation.

We will use these three crocodilian genomes to infer their ancestral genome. This, combined with existing and soon to be released bird genomes, will enable some inference of the ancestral archosaur genome. Reconstructing the ancestral archosaur genome has obvious implications for expanding our understanding of the genomes of extinct archosaurs, like the non-bird dinosaurs and pterosaurs (Figure 1).

There are also several biological questions specific to crocodilians that we will address by analyzing genomic and RNA-seq data and via experimental techniques. For example, despite having a temperature-dependent sex-determination system seemingly without sex chromosomes, the sexes of crocodilians have been shown to have very different recombination rates [62]. Identification of the genes that are differentially expressed in the male and female crocodilian gonads might provide insight into the perplexing observation.

SNP discovery arising from the genome sequencing is particularly relevant to farm-bred saltwater crocodiles. Large panels of SNP markers will enable more refined linkage maps [62], more precise mapping of quantitative trait loci (QTL) than is currently possible with microsatellite markers [62] and eventually the implementation of genomic selection in crocodile breeding programs.

Eventually members of the ICGWG hope to address additional questions beyond the scope of the initial genome paper. These might be presented in satellite publications. One of these involves the sex determination system of American alligators. Which genes are the initial temperature sensitive regulators that trigger the downstream, largely conserved [85] sex-determination system? Having the genome sequences available for these three crocodilians will enable a new wave of discoveries about the evolutionary histories of crocodilians, non-avian reptiles and birds, and amniotes generally.

How other groups can join the consortium, or publish independently with our early release data

This project is affiliated with the Genome 10K (G10K) initiative [14]. We invite other G10K affiliates and the broader scientific community to access and make use of the draft assembly and raw read data that we have produced. Any group performing non-genome-scale analyses that are sufficiently independent of the analyses described above are welcome to use these data without restriction. As a matter of courtesy and to avoid duplicated effort, we request that competing genome-scale projects or analyses that overlap with the areas stated above disclose their status to the ICGWG consortium (formal inquiries and requests to join the working group should be made to D.A.R.) and cite this and subsequent papers that provide the data. Versioned assemblies, further project description, and a complete list of current ICGWG members can be accessed on the website dedicated to this project [52].

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported by grants to D.A.R. (MCB-1052500, MCB-0841821, DEB-1020865 from the U.S. National Science Foundation) and funds from the Institute for Genomics, Biocomputing and Biotechnology at Mississippi State University. E.L.B., E.W.T., and collaborators at the University of Florida were supported by funds from the U.S. National Science Foundation (DUE-0920151). T.I. received financial support from the National Institute for Basic Biology and Grants-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, Science and Technology of Japan. S.R.I., L.G.M., J.G., P.D. and C.M. were supported by Australian Rural Industries Research and Development Corporation grants (RIRDC PRJ-000549, RIRDC PRJ-005355, RIRDC PRJ-002461). M.K.F. received financial support from a U.S. National Science Foundation Biological Informatics Postdoctoral Fellowship (DBI-0905714). R.E.G. is a Searle Scholar and a Sloan Fellow. E.D.J. was supported by the Howard Hughes Medical Institute and the National Institutes of Health. We are grateful to Kent Vliet (University of Florida) and the Alligator Farm (St. Augustine, Florida) for providing access to fresh gharial blood.

Author details
[1]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064, USA. [2]Department of Biology, University of Florida, Gainesville, FL 32611 USA. [3]Porosus Pty Ltd, PO Box 86, Palmerston, NT 0831, Australia. [4]South Eastern Area Laboratory Services, Randwick, NSW 2031, Australia. [5]Faculty of Veterinary Science, University of Sydney, NSW 2006, Australia. [6]INRA, AgroParisTech, UMR1313 Animal Genetics and Integrative Biology, Jouy-en-Josas, F78352, France. [7]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA. [8]Institute for Genomics, Biocomputing and Biotechnology, Mississippi State University, Mississippi State, MS 39762, USA. [9]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA. [10]Department of Biology, Boston University, 5 Cummington Street, Boston, MA 02215, USA. [11]Department of Biological Sciences, Texas Tech University, Lubbock, TX 79409, USA. [12]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611 USA. [13]Department of Ecology and Evolutionary Biology, University of California, 621 Charles E. Young Drive South, Los Angeles, CA 90095, USA. [14]Department of Biological Sciences,

University of South Carolina, Columbia, SC 29205, USA. [15]Department of Biochemistry, Molecular Biology, Entomology and Plant Pathology, Mississippi State University, Mississippi State, MS 39762, USA. [16]Department of Molecular, Cell & Developmental Biology, University of California, Santa Cruz, CA 95064, USA. [17]Okazaki Institute for Integrative Bioscience, National Institute for Basic Biology, National Institutes of Natural Sciences, 5-1 Higashiyama, Myodaiji, Okazaki 444-8787, Japan. [18]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Ave, Cambridge, MA 02138, USA. [19]Current address: National Institute of General Medical Sciences, National Institutes of Health, Bethesda, MD 2089, USA. [20]Hollings Marine Laboratory, Marine Biomedicine and Environmental Sciences, Department of Obstetrics and Gynecology, Medical University of South Carolina, Charleston, SC 29412, USA. [21]Savannah River Ecology Laboratory, University of Georgia, PO Drawer E, Aiken, SC 29802, USA. [22]College of Veterinary Medicine, Mississippi State University, Mississippi State, MS 39762, USA. [23]Moore Laboratory of Zoology, Occidental College, 1600 Campus Rd, Los Angeles, CA 90041, USA. [24]Department of Chemistry, McNeese State University, Lake Charles, LA 70609, USA. [25]Department of Plant and Soil Sciences, Mississippi State University, Mississippi State, MS 39762, USA. [26]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, CA 95064, USA. [27]Department of Animal & Food Sciences University of Delaware, Newark DE, 19717, USA. [28]Department of Microbiology and Cell Science, University of Florida, Gainesville, FL 32611 USA. [29]Howard Hughes Medical Institute, Duke University Medical Center, Department of Neurobiology, Box 3209, Durham, NC 27710, USA. [30]Department of Environmental Health Science and Georgia Genomics Facility, Environmental Health Science Building, University of Georgia, Athens, GA 30602, USA.

Published: 31 January 2012

References

1. Hedges SB, Kumar S: *The Timetree of Life*. Oxford University Press, USA; 2009.
2. Katsu Y, Braun EL, Guillette LJ, Iguchi T: **From reptilian phylogenomics to reptilian genomes: analyses of c-Jun and DJ-1 proto-oncogenes.** *Cytogenet Genome Res* 2009, **127**:79–93.
3. Janes DE, Organ CL, Fujita MK, Shedlock AM, Edwards SV: **Genome evolution in Reptilia, the sister group of mammals.** *Annu Rev Genomics Hum Genet* 2010, **11**:239–264.
4. Alföldi J, di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD, Ray DA, Boissinot S, Shedlock AM, Botka C, Castoe TA, Colbourne JK, Fujita MK, Moreno RG, Hallers ten BF, Haussler D, Heger A, Heiman D, Janes DE, Johnson J, de Jong PJ, Koriabine MY, Lara M, Novick PA, Organ CL, Peach SE, et al.: **The genome of the green anole lizard and a comparative analysis with birds and mammals.** *Nature* 2011, **477**:587-591.
5. NHGRI Genome Sequencing Proposals [http://www.genome.gov]
6. Castoe TA, Bronikowski AM, Brodie ED, Edwards SV, Pfrender ME, Shapiro MD, Pollock DD, Warren WC: **A proposal to sequence the genome of a garter snake (*Thamnophis sirtalis*).** *Stand Genomic Sci* 2011, **4**:257–270.
7. Castoe TA, de Koning AJ, Hall KT, Yokoyama KD, Gu W, Smith EN, Feschotte C, Uetz P, Ray DA, Dobry J, Bogden R, Mackessy SP, Bronikowski AM, Warren WC, Secor SM, Pollock DD: **Sequencing the genome of the Burmese python (*Python molurus bivittatus*) as a model for studying extreme adaptations in snakes.** *Genome Biol* 2011, **12**:406.
8. Brusatte S, Benton M, Desojo J, Langer M: **The higher-level phylogeny of Archosauria (*Tetrapoda: Diapsida*).** *J Syst Paleontol* 2010, **8**:3-47.
9. International Chicken Genome Sequencing Consortium: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432**:695–716.
10. Dalloul RA, Long JA, Zimin AV, Aslam L, Beal K, Blomberg LA, Bouffard P, Burt DW, Crasta O, Crooijmans RPMA, Cooper K, Coulombe RA, De S, Delany ME, Dodgson JB, Dong JJ, Evans C, Frederickson KM, Flicek P, Florea L, Folkerts O, Groenen MAM, Harkins TT, Herrero J, Hoffmann S, Megens H-J, Jiang A, de Jong P, Kaiser P, Kim H, et al.: **Multi-platform next-generation sequencing of the domestic turkey (*Meleagris gallopavo*): genome assembly and analysis.** *PLoS Biol* 2010, **8**: e1000475.
11. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Künstner A, Searle S, White S, Vilella AJ, Fairley S, Heger A, Kong L, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TAF, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin Y-C, George J, Sweedler J, Southey B, Gunaratne P, Watson M, et al.: **The genome of a songbird.** *Nature* 2010, **464**:757–762.
12. Mallard duck (*Anas platyrhynchos)* project [http://pre.ensembl.org/Anas_platyrhynchos/Info/Index]
13. The Avian Genomes Project [http://aviangenomes.org]
14. Genome 10K Community of Scientists: **Genome 10K: A proposal to obtain whole-genome sequence for 10 000 vertebrate species.** *J Hered* 2009, **100**:659–674.
15. Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV: **Origin of avian genome size and structure in non-avian dinosaurs.** *Nature* 2007, **446**:180–184.
16. Organ CL, Brusatte SL, Stein K: **Sauropod dinosaurs evolved moderately sized genomes unrelated to body size.** *Proc R Soc Lond B: Biol Sci* 2009, **276**:4303–4308.
17. Organ CL, Janes DE, Meade A, Pagel M: **Genotypic sex determination enabled adaptive radiations of extinct marine reptiles.** *Nature* 2009, **461**:389–392.
18. Schweitzer MH, Zheng W, Organ CL, Avci R, Suo Z, Freimark LM, Lebleu VS, Duncan MB, Vander Heiden MG, Neveu JM, Lane WS, Cottrell JS, Horner JR, Cantley LC, Kalluri R, Asara JM: **Biomolecular characterization and protein sequences of the Campanian hadrosaur *B. canadensis*.** *Science* 2009, **324**:626–631.
19. Organ CL, Shedlock AM: **Palaeogenomics of pterosaurs and the evolution of small genome size in flying vertebrates.** *Biol Lett* 2009, **5**:47–50.
20. Brochu CA, Wagner JR, Jouve S, Sumrall CD, Densmore LD: **A correction corrected: consensus over the meaning of Crocodylia and why it matters.** *Syst Biol* 2009, **58**:537–543.
21. Brochu C: **Phylogenetic approaches toward crocodylian history.** *Annual Review of Earth and Planetary Sciences* 2003, **31**:357-397
22. Roos J, Aggarwal RK, Janke A: **Extended mitogenomic phylogenetic analyses yield new insight into crocodylian evolution and their survival of the Cretaceous-Tertiary boundary.** *Molecular Phylogenetics and Evolution* 2007, **45**:663–673.
23. Densmore LD: **Biochemical and immunological systematics of the Order Crocodilia.** *Evol Biol* 1983, **16**:397–465.
24. Harshman J, Huddleston CJ, Bollback JP, Parsons TJ, Braun MJ: **True and false gharials: a nuclear gene phylogeny of crocodylia.** *Syst Biol* 2003, **52**:386–402.
25. Gatesy J, Amato G, Norell M, Desalle R, Hayashi C: **Combined support for wholesale taxic atavism in gavialine crocodylians.** *Syst Biol* 2003, **52**:403–422.
26. Ross JP: *Crocodiles. Status Survey and Conservation Action Plan*. 2nd ed. IUCN, Gland, Switzerland and Cambridge, UK: ICUN/SSC Crocodile Specialist Group; 1998.
27. Ryan C: **Saltwater crocodiles as tourist attractions.** *J Sustain Tour* 1998, **6**:314–327.
28. Ryan C, Harvey K: **Who likes saltwater crocodiles? Analysing socio-demographics of those viewing tourist wildlife attractions based on saltwater crocodiles.** *J Sustain Tour* 2000, **8**:426–433.
29. Merchant M, Thibodeaux D, Loubser K, Elsey RM: **Amoebacidal effects of serum from the American alligator (*Alligator mississippiensis*).** *J Parasitol* 2004, **90**:1480–1483.
30. Merchant ME, Roche C, Elsey RM, Prudhomme J: **Antibacterial properties of serum from the American alligator (*Alligator mississippiensis*).** *Comp Biochem Physiol B* 2003, **136**:505–513.
31. Merchant ME, Pallansch M, Paulman RL, Wells JB, Nalca A, Ptak R: **Antiviral activity of serum from the American alligator (*Alligator mississippiensis*).** *Antiviral Res* 2005, **66**:35–38.
32. Merchant ME, Leger N, Jerkins E, Mills K, Pallansch MB, Paulman RL, Ptak RG: **Broad spectrum antimicrobial activity of leukocyte extracts from the American alligator (*Alligator mississippiensis*).** *Vet Immunol Immunopath* 2006, **110**:221–228.
33. Milnes MR, Guillette LJ: **Alligator tales: New lessons about environmental contaminants from a sentinel species.** *BioScience* 2008 ,**58**:1027-1036.
34. Campbell KR: **Ecotoxicology of crocodilians.** *Appl Herpetol* 2003, **1**:45–163.
35. Guillette LJ, Gross TS, Masson GR, Matter JM, Percival HF, Woodward AR: **Developmental abnormalities of the gonad and abnormal sex hormone concentrations in juvenile alligators from contaminated and control lakes in Florida.** *Environ Health Perspect* 1994, **102**:680–688.
36. Wu TH, Cañas JE, Rainwater TR, Platt SG, McMurry ST, Anderson TA: **Organochlorine contaminants in complete clutches of Morelet's crocodile (*Crocodylus moreletii*) eggs from Belize.** *Environ Pollut* 2006, **144**:151–157.
37. Grigg GC, Seebacher F, Franklin CE: *Crocodilian Biology and Evolution*. Surrey Beatty & Sons; 2001.

38. Brochu CA: **Calibration age and quartet divergence date estimation.** *Evolution* 2004, **58**:1375–1382.
39. Brochu CA: **Morphology, fossils, divergence timing, and the phylogenetic relationships of *Gavialis*.** *Syst Biol* 1997, **46**:479–522.
40. Rayfield E: **Establishing a framework for archosaur cranial mechanics.** *Paleobiology* 2008, **34**:494-515.
41. Deeming DC, Ferguson MWJ: **The mechanism of temperature dependent sex determination in Crocodilians: A hypothesis.** *Integr Comp Biol* 1989, **29**:973–985.
42. Lang JW, Andrews HV: **Temperature-dependent sex determination in crocodilians.** *J Exp Zool* 1994, **270**:28–44.
43. Western PS, Harry JL, Marshall Graves JA, Sinclair AH: **Temperature-dependent sex determination in the American alligator: expression of SF1, WT1 and DAX1 during gonadogenesis.** *Gene* 2000, **241**:223–232.
44. Pieau C, Dorizzi M, Richard-Mercier N: **Temperature-dependent sex determination and gonadal differentiation in reptiles.** *Cell Mol Life Sci* 1999, **55**:887–900.
45. Ferguson MW, Joanen T: **Temperature of egg incubation determines sex in *Alligator mississippiensis*.** *Nature* 1982, **296**:850–853.
46. Cedeño-Vázquez JR, Rodriguez D, Calmé S, Ross JP, Densmore LD, Thorbjarnarson JB: **Hybridization between *Crocodylus acutus* and *Crocodylus moreletii* in the Yucatan Peninsula: I. Evidence from mitochondrial DNA and morphology.** *J Exp Zool A Ecol Genet Physiol* 2008, **309**:661–673.
47. Ray DA, Dever JA, Platt SG, Rainwater TR, Finger AG, McMurry ST, Batzer MA, Barr B, Stafford PJ, McKnight J, Densmore LD: **Low levels of nucleotide diversity in *Crocodylus moreletii* and evidence of hybridization with *C. acutus*.** *Conserv Genet* 2004, **5**:449–462.
48. Weaver JP, Rodriguez D, Venegas-Anaya M, Cedeño-Vázquez JR, Forstner MRJ, Densmore LD: **Genetic characterization of captive Cuban crocodiles (*Crocodylus rhombifer*) and evidence of hybridization with the American crocodile (*Crocodylus acutus*).** *J Exp Zool A Ecol Genet Physiol* 2008, **309**:649–660.
49. Davis LM, Glenn TC, Elsey RM, Dessauer HC, Sawyer RH: **Multiple paternity and mating patterns in the American alligator, *Alligator mississippiensis*.** *Mol Ecol* 2001, **10**:1011–1024.
50. Davis LM, Glenn TC, Strickland DC, Guillette LJ, Elsey RM, Rhodes WE, Dessauer HC, Sawyer RH: **Microsatellite DNA analyses support an east-west phylogeographic split of American alligator populations.** *J Exp Zool* 2002, **294**:352–372.
51. Ryberg WA, Fitzgerald LA, Honeycutt RL, Cathey JC: **Genetic relationships of American alligator populations distributed across different ecological and geographic scales.** *J Exp Zool* 2002, **294**:325–333.
52. The International Crocodilian Genomes Working Group [http://crocgenomes.org].
53. Krishan A, Dandekar P, Nathan N, Hamelik R, Miller C, Shaw J: **DNA index, genome size, and electronic nuclear volume of vertebrates from the Miami Metro Zoo.** *Cytometry A* 2005, **65**:26–34.
54. Cohen MM, Gans C: **The chromosomes of the order Crocodilia.** *Cytogenet Genome Res* 1970, **9**:81–105.
55. Valleley EM, Harrison CJ, Cook Y, Ferguson MW, Sharpe PT: **The karyotype of *Alligator mississippiensis*, and chromosomal mapping of the ZFY/X homologue, Zfc.** *Chromosoma* 1994, **103**:502–507.
56. Shan X, Ray DA, Bunge JA, Peterson DG: **A bacterial artificial chromosome library for the Australian saltwater crocodile (*Crocodylus porosus*) and its utilization in gene isolation and genome characterization.** *BMC Genomics* 2009, **10 Suppl 2**:S9.
57. King M, Honeycutt R, Contreras N: **Chromosomal repatterning in crocodiles: C, G and N-banding and the in situ hybridization of 18S and 26S rRNA cistrons.** *Genetica* 1986, **70**:191-201.
58. Dalzell P, Miles LG, Isberg SR, Glenn TC, King C, Murtagh V, Moran C: **Standardized Reference Ideogram for Physical Mapping in the Saltwater Crocodile (*Crocodylus porosus*).** *Cytogenet Genome Res* 2009, **127**:204–212.
59. Shedlock AM, Botka CW, Zhao S, Shetty J, Zhang T, Liu JS, Deschavanne PJ, Edwards SV: **Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome.** *P Natl Acad Sci Usa* 2007, **104**:2767–2772.
60. Miyake T, Amemiya CT: **BAC libraries and comparative genomics of aquatic chordate species.** *Comp Biochem Physiol C Toxicol Pharmacol* 2004, **138**:233–244.
61. NISC Comparative Sequencing Initiative [http://www.nisc.nih.gov]
62. Miles LG, Isberg SR, Glenn TC, Lance SL, Dalzell P, Thomson PC, Moran C: **A genetic linkage map for the saltwater crocodile (*Crocodylus porosus*).** *BMC Genomics* 2009, **10**:339.
63. Chojnowski JL, Franklin J, Katsu Y, Iguchi T, Guillette LJ, Kimball RT, Braun EL: **Patterns of vertebrate isochore evolution revealed by comparison of expressed mammalian, avian, and crocodilian genes.** *J Mol Evol* 2007, **65**:259–266.
64. Nabholz B, Künstner A, Wang R, Jarvis ED, Ellegren H: **Dynamic evolution of base composition: causes and consequences in avian phylogenomics.** *Mol Biol Evol* 2011, **28**:2197–2210.
65. Mortazavi A, Schwarz EM, Williams B, Schaeffer L, Antoshechkin I, Wold BJ, Sternberg PW: **Scaffolding a *Caenorhabditis* nematode genome with RNA-seq.** *Genome Res* 2010, **20**:1740–1747.
66. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB: **High-quality draft assemblies of mammalian genomes from massively parallel sequence data.** *Proc Natl Acad Sci USA* 2011, **108**:1513–1518.
67. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, Yu HOK, Buffalo V, Zerbino DR, Diekhans M, Nguyen N, Ariyaratne PN, Sung W-K, Ning Z, Haimel M, Simpson JT, Fonseca NA, Birol I, Docking TR, Ho IY, Rokhsar DS, Chikhi R, Lavenier D, Chapuis G, Naquin D, Maillet N, Schatz MC, Kelley DR, Phillippy AM, Koren S, et al.: **Assemblathon 1: A competitive assessment of de novo short read assembly methods.** *Genome Res* 2011, **21**:2224–2241.
68. Tzika AC, Helaers R, Schramm G, Milinkovitch MC: **Reptilian-transcriptome v1.0, a glimpse in the brain transcriptome of five divergent Sauropsida lineages and the phylogenetic position of turtles.** *EvoDevo* 2011, **2**:19.
69. Toronto International Data Release Workshop Authors, Birney E, Hudson TJ, Green ED, Gunter C, Eddy S, Rogers J, Harris JR, Ehrlich SD, Apweiler R, Austin CP, Berglund L, Bobrow M, Bountra C, Brookes AJ, Cambon-Thomsen A, Carter NP, Chisholm RL, Contreras JL, Cooke RM, Crosby WL, Dewar K, Durbin R, Dyke SOM, Ecker JR, Emam El K, Feuk L, Gabriel SB, Gallacher J, Gelbart WM, et al.: **Prepublication data sharing.** *Nature* 2009, **461**:168–170.
70. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC-C, Zhou Y, Cao J, Sun X, Fu Y, et al.: **The sequence and de novo assembly of the giant panda genome.** *Nature* 2010, **463**:311–317.
71. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110**:462–467.
72. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes.** *Bioinformatics* 2005, **21 Suppl 1**:i351–i358
73. Shedlock A: **Phylogenomic investigation of CR1 LINE diversity in reptiles.** *Syst Biol* 2006, **55**:902–911.
74. Kordis D: **Transposable elements in reptilian and avian (*Sauropsida*) genomes.** *Cytogenet Genome Res* 2009, **127**:94–111.
75. Ray D, Hedges D, Herke S, Fowlkes J, Barns E, LaVie D, Goodwin L, Densmore L, Batzer M: **Chompy: An infestation of MITE-like repetitive elements in the crocodilian genome.** *Gene* 2005, *362*:1-10.
76. Fujita MK, Edwards SV, Ponting CP: **The Anolis lizard genome: An amniote genome without isochores.** *Genome Biol Evol* 2011, **3**:974–984.
77. Chojnowski JL, Braun EL: **Turtle isochore structure is intermediate between amphibians and other amniotes.** *Integr Comp Biol* 2008, **48**:454–462.
78. OASES [http://www.ebi.ac.uk/~zerbino/oases]
79. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18**:821–829.
80. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28**:45–48.
81. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14**:708–715.
82. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Chen Y, Clapham P, Coates G, Fairley S, Fitzgerald S, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Larsson P, Longden I, McLaren W, Overduin B, Pritchard B, Riat HS, Rios D, Ritchie GRS, Ruffier M, Schuster M, et al.: **Ensembl 2011.** *Nucleic Acids Res* 2011, **39**:D800–806.
83. Stanke M, Diekhans M, Baertsch R, Haussler D: **Using native and syntenically mapped cDNA alignments to improve de novo gene finding.** *Bioinformatics* 2008, **24**:637–644.
84. Norris TB, Rickards GK, Daugherty CH: **Chromosomes of tuatara,**

Sphenodon, a chromosome heteromorphism and an archaic reptilian karyotype. *Cytogenet Genome Res* 2004, **105**:93–99.
85. Haag ES, Doty AV: **Sex determination across evolution: Connecting the dots.** *PLoS Biol* 2005, **3**:e21.
86. Sereno PC: **The Evolution of Dinosaurs.** *Science* 1999, **284**:2137–2147.
87. Chiappe LM: *Glorified Dinosaurs: The Origin and Early Evolution of Birds*. 1st ed. Wiley-Liss; 2007 .
88. Shen X-X, Liang D, Wen J-Z, Zhang P: **Multiple genome alignments facilitate development of NPCL markers: A case study of tetrapod phylogeny focusing on the position of turtles.** *Mol Biol Evol* 2011, **28**:3237-3252.
89. Lyson TR, Sperling EA, Heimberg AM, Gauthier JA, King BL, Peterson KJ: **MicroRNAs support a turtle + lizard clade.** *Biol Lett* 2011, **8**:104-107.
90. Dolezel J, Bartos J, Voglmayr H, Greilhuber J: **Nuclear DNA content and genome size of trout and human.** *Cytometry A* 2003, **51**:127–128.

doi:10.1186/gb-2012-13-1-415
Cite this article as: St John JA, *et al.*: Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology* 2012, **13**:415.