Supplemental Results

Small RNAs

In model vertebrate species (ex. human, mouse, chicken, leopard frog, zebrafish) a few hundred to more than a thousand miRNAs have been identified (Kozomara and Griffiths-Jones 2014). In non-model taxa, which include the crocodilians, miRNAs are frequently identified based on sequence conservation to known miRNAs. Using this technique some conserved miRNAs in Alligator mississippiensis have been annotated by mapping small RNA reads to miRNAs from the chicken and green anole (Lyson et al. 2012) but no lineage-specific miRNAs are identifiable. Results presented here represent a first step in understanding the lineage-specific evolution of miRNAs in the crocodilians.

A total of 15 million reads from the testis library was reduced to 1.12 million unique, quality- and size-filtered reads used for miRNA prediction with miRDeep2. miRDeep2 mapped reads to 114 chicken miRNAs, confirming their presence and expression in alligator testis. Initial predictions of novel miRNAs (n = 145) were filtered using various criteria. Putative miRNAs with less than 10 reads mapping to the predicted mature miRNA (n = 15), a miRDeep score < 1 (n = 13), non-significant randFold scores (n = 11), more reads mapping to the hairpin loop than the miRNA* strand (n = 7), homology to ribosomal or transfer RNAs (n = 2), or overlapping loci (n = 2) were removed from downstream analyses. The remaining putative miRNAs were re-predicted in the alligator genome and compared to the crocodile, gharial, and chicken genomes to identify homologous miRNAs using MapMi. MapMi removed 31 putative miRNAs with homology to TEs and one putative miRNA with a low complexity sequence. Three miRDeep miRNAs failed re-prediction in MapMi, though two were identified in either the crocodile or gharial. In all, 60 putative miRNAs passed all quality filters and were predicted by both the miRDeep2 and MapMi algorithms, 25 were present in all crocodilians, 17 were alligator specific, and 11 were in the crocodilians and the chicken. Seven were present either the alligator and the gharial or the alligator and the crocodile, but not all three crocodilians. Blast results against NCBI's non-redundant nucleotide

database identified four putative miRNAs with homologs in *Anolis carolinensis* and one with *Danio rerio*. Four of the 5 miRNAs with NCBI homologs were found in all four taxa examined with MapMi (aca-mir146-a, aca-mir-34c, dar-mir-144-5, aca-mir-1388). The fifth (aca-mir-425) was in all three crocodilians, but not in the chicken. Due to the deep divergences of these taxa and strong selection on many miRNAs (Quach et al. 2009), it is likely that these putative miRNAs are functional in crocodilians. In addition, the ability to identify these conserved-functional miRNAs demonstrates the ability of the methods employed herein to identify true miRNAs that are lineage-specific. Additional work is necessary to verify and ascribe function to the putative miRNAs. Putative miRNAs were deposited in miRBase and all sequence data used for miRNA prediction was deposited in the NCBI Short read archive (PRJNA285470).

Supplemental Methods

Gene prediction

We made gene predictions using the AUGUSTUS gene prediction software version 3.0.3 (Stanke et al. 2006). AUGUSTUS predicts genes based on a hidden Markov model trained on gene structures from a related species as well as extrinsic evidence provided by the user. We provided RNA-seq alignments, repetitive element predictions, and chicken protein alignments to AUGUSTUS as extrinsic evidence. We aligned previously-published RNA-seq reads from various tissues of *Alligator mississippiensis* (Green et al. 2014) to the genome (SRA: SRP057608) using TopHat 2.0.14 (Kim et al. 2013) with default parameters. We found repetitive elements in the genome using RepeatScout (Price et al. 2005) and RepeatMasker Open-4.0 (Smit et al. 2015) with default parameters. We aligned all *Gallus gallus* (chicken) proteins from UniProt to the genome using Exonerate version 2.2.0 (Slater and Birney 2005) with the protein2genome model. Finally, we ran AUGUSTUS using these sources of extrinsic evidence and parameters trained on gene structures from *G. gallus*.

Functional annotation

We assigned protein names, gene nomenclature, and Gene Ontology (GO) terms to the predicted genes. We chose protein names based on reciprocal best hits BLAST from orthologous proteins from vertebrate species with a gene nomenclature project, specifically *G. gallus* (chicken), *A. carolinensis* (Green anole), *D. rerio* (Zebrafish), and *H. sapiens* (Human). We define orthologous proteins as those with a reciprocal best hit using default blastp parameters and an E-value cutoff of 0.00001. We assigned gene names using the same strategy, resulting in the assignment of 15,977 protein and gene names. We assigned GO terms to predicted proteins based upon a combinatorial approach. We mapped predicted proteins to InterPro identifiers and GO (assigned the GO evidence code of "IEA" or Inferred from Electronic Annotation) based on InterProScan (Jones et al. 2014). We also transferred GO using reciprocal blast from orthologous vertebrate genes experimental evidence codes (assigned the GO evidence code "ISA" or Inferred from Sequence Alignment). We merged GO annotations from these two sources, removed duplicates, and manually reviewed GO terms to eliminate those that are not species-appropriate, such as "sex chromosome" and "fin development." Following this strategy, 17,430 American alligator proteins were assigned 5,960 unique GO terms.

Small RNAs

Testis tissue was harvested from a wild-caught, reproductively mature, male alligator from Rockefeller National Wildlife (Grand Chenier, LA) and a horizontal cross section was homogenized for small RNA isolation. Small RNAs were purified using TRIzol reagent followed by an ethanol precipitation. RNA quantity and quality was measured using a Bioanalyzer, to assure that RNA Integrity Number (RIN) was greater than 7.5. The small RNA pools was prepped for Illumina sequencing using a NEBNext Small RNA Library Prep Set with converted RNA fragments ranging from 15 to 35 nt

(excluding sequencing adapters) selected via PippinHT. The resulting library was sequenced on a single MiSeq lane 1x50 nt.

Adapters and low quality base calls were removed from small RNA sequences using the FASTX-Toolkit (v0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/index.html). Specifically, reads with scores below Q20 across 50% or more of the read, after adapter trimming, were discarded. Once filtered, reads falling outside of an 18-24 nt range were culled. miRNAs were predicted from the remaining reads using the miRDeep2 pipeline (Friedländer et al. 2012). All high quality small RNA reads were mapped to known chicken (*Gallus gallus*) miRNAs (mature and hairpin) and the new alligator genome using miRDeep2's mapper.pl. Additional parameters included collapsing unique reads (-m) and limiting the maximum mapping locations to five or fewer (-r 5). Once mapped, miRNAs were predicted from reads without homology to known chicken miRNAs using the miRDeep2.pl script.

Several filters were applied to novel miRNAs predicted by miRDeep2. Any novel miRNAs that were similar to ribosomal or transfer RNAs, had fewer than 10 reads from the mature miRNA, had a miRDeep score less than 1, did not have a significant randFold score, overlapped with other predicted or known miRNAs, or contained more reads mapping to the miRNA hairpin loop than the miRNA* were removed from further analyses. Known chicken miRNAs were accepted regardless of these constraints. MapMi (Guerra-Assunção and Enright 2010) was used to identify homologous loci to the putative miRNAs predicted by miRDeep2 in the crocodile (*Crocodylus porosus*; JRXG00000000.1), gharial (*Gavialis gangeticus*; JRWT00000000.1), and chicken (CM000000.4) genomes. Initial steps in the MapMi uses Dust3 to remove low complexity sequences and then culls sequences with homology to TEs. MapMi predictions scoring less than 35 were considered low quality and removed. In addition, miRDeep2 putative miRNAs not re-predicted by MapMi in the alligator genome were removed as well.

Supplemental Figures

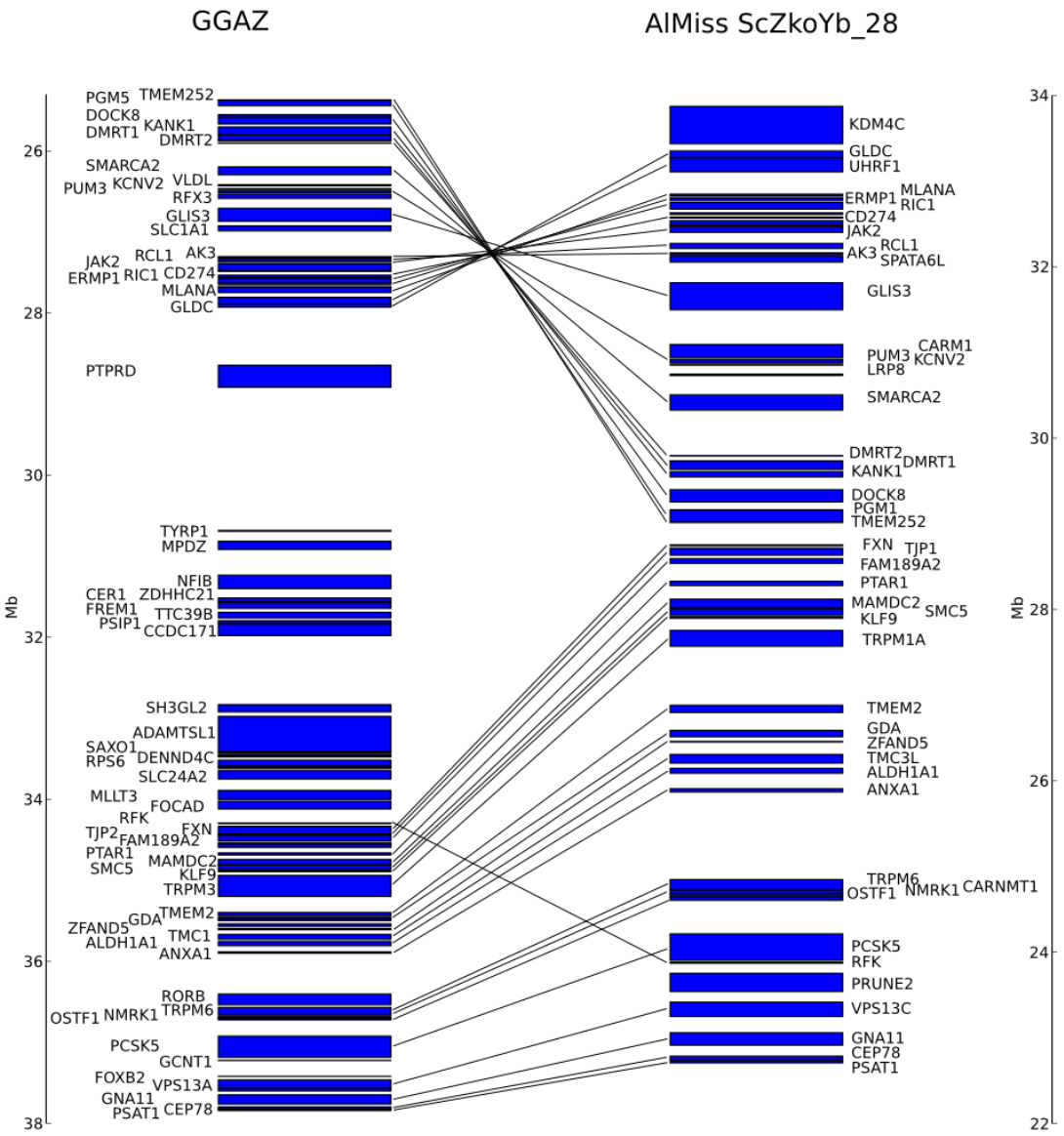


Figure S1. Synteny between the chicken Z chromosome and scaffold 28 of the alligator assembly, around the avian sex-determination gene DMRT1. Orthologous genes are connected with lines.

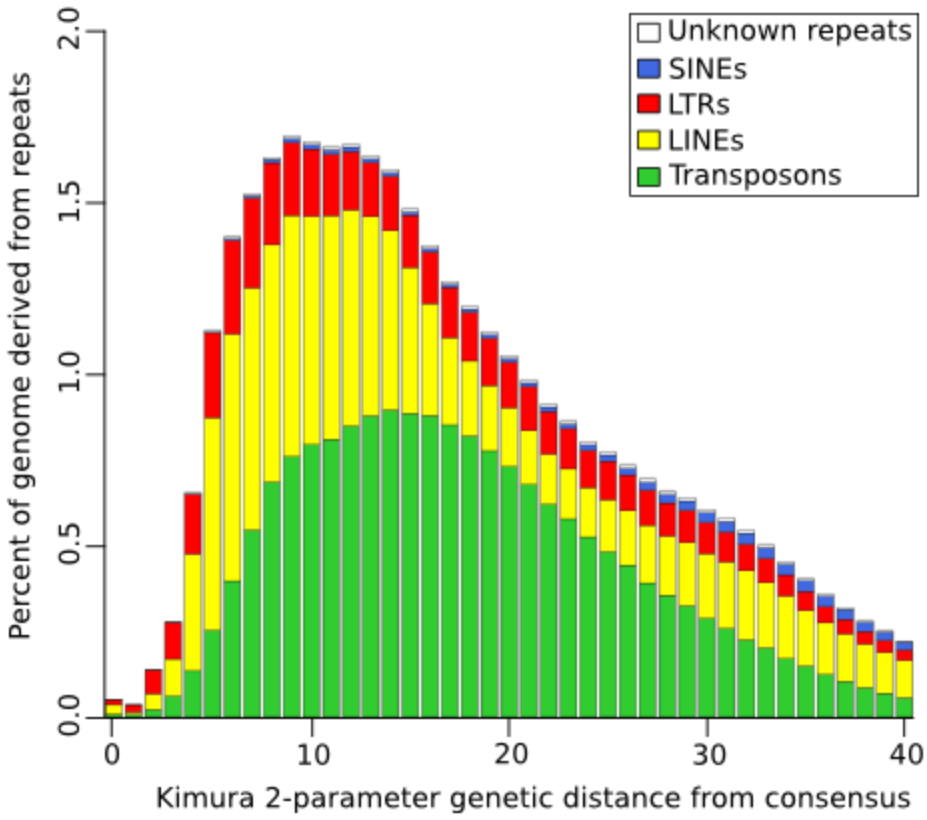
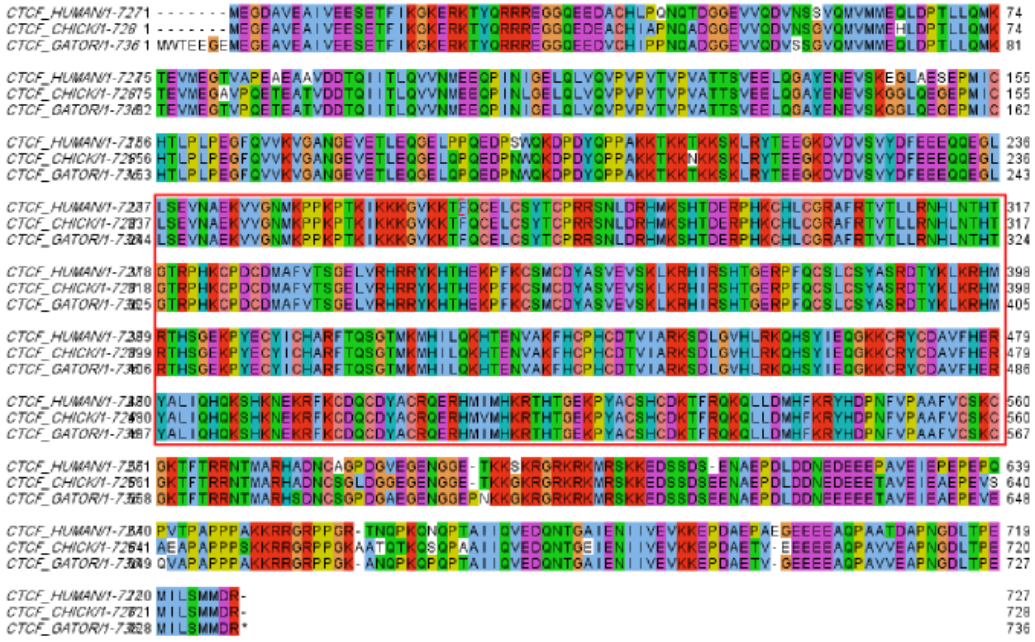


Figure S2. The Kimura 2-parameter (Kimura 1980) between individual transposable element insertions and their respective consensus sequences as a percentage of the genome. Genetic distance increases with element insertion age.

a.



b.

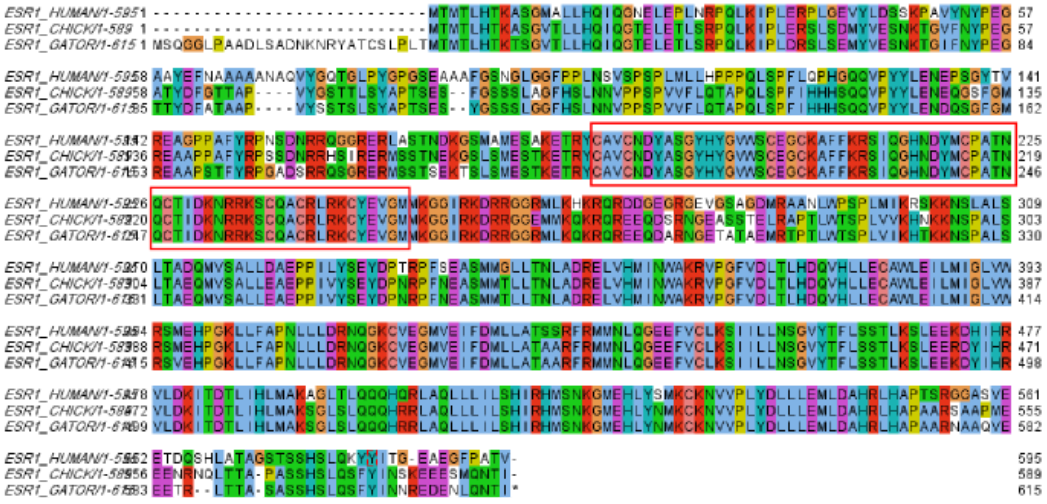


Figure S3. Alignments of the protein sequences of human, chicken, and alligator orthologs of CTCF (a) and ESR1 (b). The DNA-binding domains of each are highlighted in a red box, showing perfect conservation.

Supplemental Tables

Table S1. Scaffold joins in the saltwater crocodile and gharial genomes verified by PCR, including the primers used and results.

Table S2. Total repetitive content in new alligator assembly and percent of genome derived from all repeats as well as the three dominant TE superfamilies in crocodilians. Repeats were identified using RepeatMasker (Smit et al. 2015) and known alligator repeats present in RepBase (v21.02).

Table S3. Embryonic alligator GAM complex libraries for RNA-sequencing, along with their NCBI accessions.

Table S4. Genes determined to have sex-biased expression in alligator embryos, including expression values in FPKM, fold changes, and FDR-adjusted p-values.

Table S5. Enriched gene ontology terms for genes with male- and female-biased expression in the gonads at the 30-day time point.

Table S6. ESR1 DNA-binding domain conservation, showing perfect protein sequence conservation of the binding domain in human, mouse, chicken, alligator, and turtle orthologs of this protein.

References

Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. 2012. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* **40**: 37-52.

Green RE, Braun EL, Armstrong J, Earl D, Nguyen N, Hickey G, Vandewege MW, St John JA, Capella-Gutiérrez S, Castoe TA et al. 2014. Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**: 1254449.

Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**: 1236-1240.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol* **14**: R36.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**: 111-120.

Kozomara A, Griffiths-Jones S. 2014. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res* **42**: D68-73.

Lyson TR, Sperling EA, Heimberg AM, Gauthier JA, King BL, Peterson KJ. 2012. MicroRNAs support a turtle + lizard clade. *Biol Lett* **8**: 104-107.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21 Suppl 1**: i351-358.

Quach H, Barreiro LB, Laval G, Zidane N, Patin E, Kidd KK, Kidd JR, Bouchier C, Veuille M, Antoniewski C et al. 2009. Signatures of purifying and local positive selection in human miRNAs. *Am J Hum Genet* **84**: 316-327.

Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**: 31.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res* **34**: W435-439.