Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling

Edward S. Rice,[1] Satomi Kohno,[2] John St. John,[3] Son Pham,[4] Jonathan Howard,[5] Liana F. Lareau,[6] Brendan L. O'Connell,[1,7] Glenn Hickey,[1] Joel Armstrong,[1] Alden Deran,[1] Ian Fiddes,[1] Roy N. Platt II,[8] Cathy Gresham,[9] Fiona McCarthy,[10] Colin Kern,[11] David Haan,[1] Tan Phan,[12] Carl Schmidt,[13] Jeremy R. Sanford,[14] David A. Ray,[8] Benedict Paten,[15] Louis J. Guillette Jr.,[16,†] and Richard E. Green[1,6,7]

[1]*Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA;* [2]*Department of Biology, St. Cloud State University, St. Cloud, Minnesota 56301, USA;* [3]*Driver Group, LLC, San Francisco, California 94158, USA;* [4]*BioTuring, Incorporated, San Diego, California 92121, USA;* [5]*Department of Biochemistry, Stanford University, Stanford, California 94305, USA;* [6]*California Institute for Quantitative Biosciences, University of California, Berkeley, California 94720, USA;* [7]*Dovetail Genomics, LLC, Santa Cruz, California 95060, USA;* [8]*Department of Biological Sciences, Texas Tech University, Lubbock, Texas 79409, USA;* [9]*Institute for Genomics, Biocomputing & Biotechnology, Mississippi State University, Mississippi State, Mississippi 39762, USA;* [10]*School of Animal and Comparative Biomedical Sciences, University of Arizona, Tucson, Arizona 85721, USA;* [11]*Department of Animal Science, University of California, Davis, California 95616, USA;* [12]*HCM University of Science, Ho Chí Minh, Vietnam 748500;* [13]*Department of Animal and Food Sciences, University of Delaware, Newark, Delaware 19717, USA;* [14]*Department of Molecular, Cell, and Developmental Biology, University of California, Santa Cruz, California 95064, USA;* [15]*Center for Biomolecular Science and Engineering, University of California, Santa Cruz, California 95064, USA;* [16]*Department of Obstetrics and Gynecology, Marine Biomedicine and Environmental Science Center, Hollings Marine Laboratory, Medical University of South Carolina, Charleston, South Carolina 29412, USA*

The American alligator, *Alligator mississippiensis*, like all crocodilians, has temperature-dependent sex determination, in which the sex of an embryo is determined by the incubation temperature of the egg during a critical period of development. The lack of genetic differences between male and female alligators leaves open the question of how the genes responsible for sex determination and differentiation are regulated. Insight into this question comes from the fact that exposing an embryo incubated at male-producing temperature to estrogen causes it to develop ovaries. Because estrogen response elements are known to regulate genes over long distances, a contiguous genome assembly is crucial for predicting and understanding their impact. We present an improved assembly of the American alligator genome, scaffolded with in vitro proximity ligation (Chicago) data. We use this assembly to scaffold two other crocodilian genomes based on synteny. We perform RNA sequencing of tissues from American alligator embryos to find genes that are differentially expressed between embryos incubated at male- versus female-producing temperature. Finally, we use the improved contiguity of our assembly along with the current model of CTCF-mediated chromatin looping to predict regions of the genome likely to contain estrogen-responsive genes. We find that these regions are significantly enriched for genes with female-biased expression in developing gonads after the critical period during which sex is determined by incubation temperature. We thus conclude that estrogen signaling is a major driver of female-biased gene expression in the post-temperature sensitive period gonads.

[Supplemental material is available for this article.]

The American alligator, *Alligator mississippiensis*, like all crocodilians and many other reptiles, has temperature-dependent sex determination (TSD), in which the sex of an embryo is determined by the incubation temperature of its egg during a temperature-sensitive period (TSP) of development (Ferguson and Joanen 1982). In contrast, mammals, birds, and other animals with genetic sex determination (GSD) rely on sex chromosomes to trigger sex determination. These genetic differences induce sex differentiation during development by causing differential expression of numerous genes. Genes with sex-biased expression during development in these lineages include conserved sexual development genes such as *SOX9* and *WNT4* (De Santa Barbara et al. 1998; Hsieh et al. 2002). Such expression differences eventually cause the development of one of two sets of distinct sexual characteristics.

Corresponding authors: esrice@ucsc.edu, ed@soe.ucsc.edu
†Deceased August 6, 2015.
Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.213595.116.

27:686–696 Published by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/17; www.genome.org

However, in alligators and other species with TSD, males and females have identical genomes, leaving open the question of how differences in temperature lead to differential expression of genes between males and females during early development (Morrish and Sinclair 2002; Shoemaker-Daly et al. 2010; Kohno and Guillette 2013).

Insight into this question comes from the observation that exposing an alligator embryo to exogenous estrogen while incubated at a male-producing temperature (MPT) causes it to develop ovaries instead of testes. Estrogen, whose presence is detected and transduced via the transcription factor estrogen receptor alpha (Bull et al. 1988; Milnes et al. 2005; Kohno et al. 2015), is an early effector of sexual development genes in the American alligator, as it is in other vertebrates, including both species with GSD and TSD (Crews et al. 1989; Nakabayashi et al. 1998). In addition, *CYP19A1*, the gene coding for the enzyme aromatase, which converts androgen to estrogen, is expressed at significantly higher levels in embryos incubated at female-producing temperature (FPT) than those incubated at MPT (Gabriel et al. 2001). These two observations have led to the hypothesis that estrogen signaling is a master regulator of sex-biased gene expression in alligator embryos (Lance 2009). While it is clear that estrogen plays a critical role in inducing ovarian development at MPT, there is currently no direct evidence that the genes targeted by estrogen are actually involved in early TSD for embryos incubated at MPT.

Much work has been performed in alligators and other vertebrates with TSD to determine the initial switch that links temperature to sexual fate (Kohno et al. 2010; Schroeder et al. 2016) and the cause of increased expression of aromatase at FPT (Parrott et al. 2014; McCoy et al. 2016). One recent hypothesis for the gene acting as the initial switch in the American alligator is the thermosensitive TRP channel *TRPV4*, as it is activated at temperatures near MPT in vitro and targets gene expression of male development genes (Yatsu et al. 2015). However, less attention has been paid to the downstream effects of increased aromatase expression in these species, especially in terms of which genes are regulated by estrogen.

Estrogen signaling is best understood in humans, including the genes it targets and its role in sexual development. Whether these mechanisms and downstream effects are conserved in other vertebrates, including those with TSD, remains unknown. In humans, estrogen regulates gene expression through the transcription factors estrogen receptor alpha and beta, coded for by the genes *ESR1* and *ESR2*, respectively. The estrogen 17β-estradiol activates an estrogen receptor by binding to its ligand-binding domain, thus allowing the receptor's DNA-binding domain to bind to a well-defined enhancer sequence, the estrogen response element, promoting the expression of nearby genes (Nilsson et al. 2001; Dahlman-Wright et al. 2006). The motif to which human estrogen receptor alpha binds has been well characterized using chromatin immunoprecipitation (Gruber et al. 2004; Laganière et al. 2005). A majority of estrogen receptor alpha binding sites are distal enhancers—that is, they are far from the genes they regulate (Carroll et al. 2006; Lin et al. 2007; Welboren et al. 2009).

A majority of estrogen receptor binding events are associated with long-range intrachromosomal chromatin interactions, and these associated events are significantly enriched for RNA polymerase II recruitment (Fullwood et al. 2009). The zinc finger protein CTCF is responsible for many of these chromatin interactions (Zhang et al. 2010). Regions delineated by two CTCF binding sites that contain an estrogen receptor binding site are significantly more likely to contain estrogen-responsive genes in humans (Chan and Song 2008). It is currently unknown whether ESR1 and CTCF binding sites are predictive of estrogen-responsive regions in the genomes of other vertebrates or whether CTCF-mediated long-range chromatin interactions are involved in estrogen's inducement of female development in vertebrates with TSD. Because the estrogen response is a long-range phenomenon in humans, a contiguous genome assembly is necessary to fully explore the genome architecture of estrogen regulation in alligators.

Green et al. (2014) published the genomes of the American alligator and two other crocodilians: the saltwater crocodile *Crocodylus porosus* and the gharial *Gavialis gangeticus*, with scaffold N50s of 508 kb for the American alligator, 205 kb for the saltwater crocodile, and 127 kb for the gharial. The slow rate of molecular evolution within crocodilians (Green et al. 2014) makes this clade ideal for testing the ability to use a highly-contiguous genome assembly to scaffold the genome assemblies of related organisms based on synteny.

Results

Assembly and annotation

The updated American alligator genome assembly AllMis2 has a total length of 2.16 Gbp compared with 2.17 Gbp for the previously published assembly AllMis1, a difference within the range of variance between assembler runs. However, AllMis2 shows a 25-fold improvement in scaffold N50, a measure of contiguity, from 508 kbp to >13 Mbp.

To assess the quality and accuracy of AllMis2, we measured its concordance with previously published BAC-end pairs (Shedlock et al. 2007) that were not used in the assembly or scaffolding. By using BWA MEM (Li 2013) with default parameters, we aligned the forward and reverse reads of the 1309 BAC-end pairs to the new assembly and to the assembly prior to scaffolding using Chicago data. We found that while 142 BAC-end pairs had both ends aligning to the same scaffold of our assembly before scaffolding with Chicago, 1160 BAC-end pairs have both ends aligning to the Chicago-scaffolded assembly: 1143, or 98.5%, of these pairs aligning to the same scaffold are oriented correctly, and 1125, or 98.4%, of these correctly oriented pairs have an insert size between 70 and 180 kb. We thus conclude that AllMis2 is both accurate and an improvement over assembly not using the Chicago library.

We annotated AllMis2 for protein-coding genes using previously published RNA-seq reads (Green et al. 2014) and AUGUSTUS (Stanke et al. 2006), finding 32,052 transcripts and 24,713 genes. Moreover, we were able to assign names to 15,977 of these genes based on orthology with named genes in other vertebrate species. By use of both orthology and protein sequence analysis, we assigned 5960 unique Gene Ontology (GO) terms to 17,430 American alligator proteins.

Crocodilian versus mammalian genome synteny

While the previous assembly of the American alligator genome AllMis1 (Green et al. 2014) was sufficient to compare to other genomes at the sequence level, our new long-range assembly AllMis2 presents an opportunity to perform genome comparisons on a broader scale. We computed synteny between the alligator and chicken (Galgal4) genomes using SyMAP v4.2 (Soderlund et al. 2011). We used both AllMis2 and AllMis1 for comparison and found that the increased contiguity of AllMis2 vastly improved our ability to compute synteny between the chicken and alligator genomes, more than doubling the percentage of the

genome covered by synteny blocks from 35% to 90% and increasing the sizes of synteny blocks, with 57 of the 90 synteny blocks >10 Mb in length. Most scaffolds in the new alligator assembly correspond to a contiguous region of a chicken chromosome, although often with some intrachromosomal rearrangements (Fig. 1A–C). Some scaffolds in the alligator genome appear to

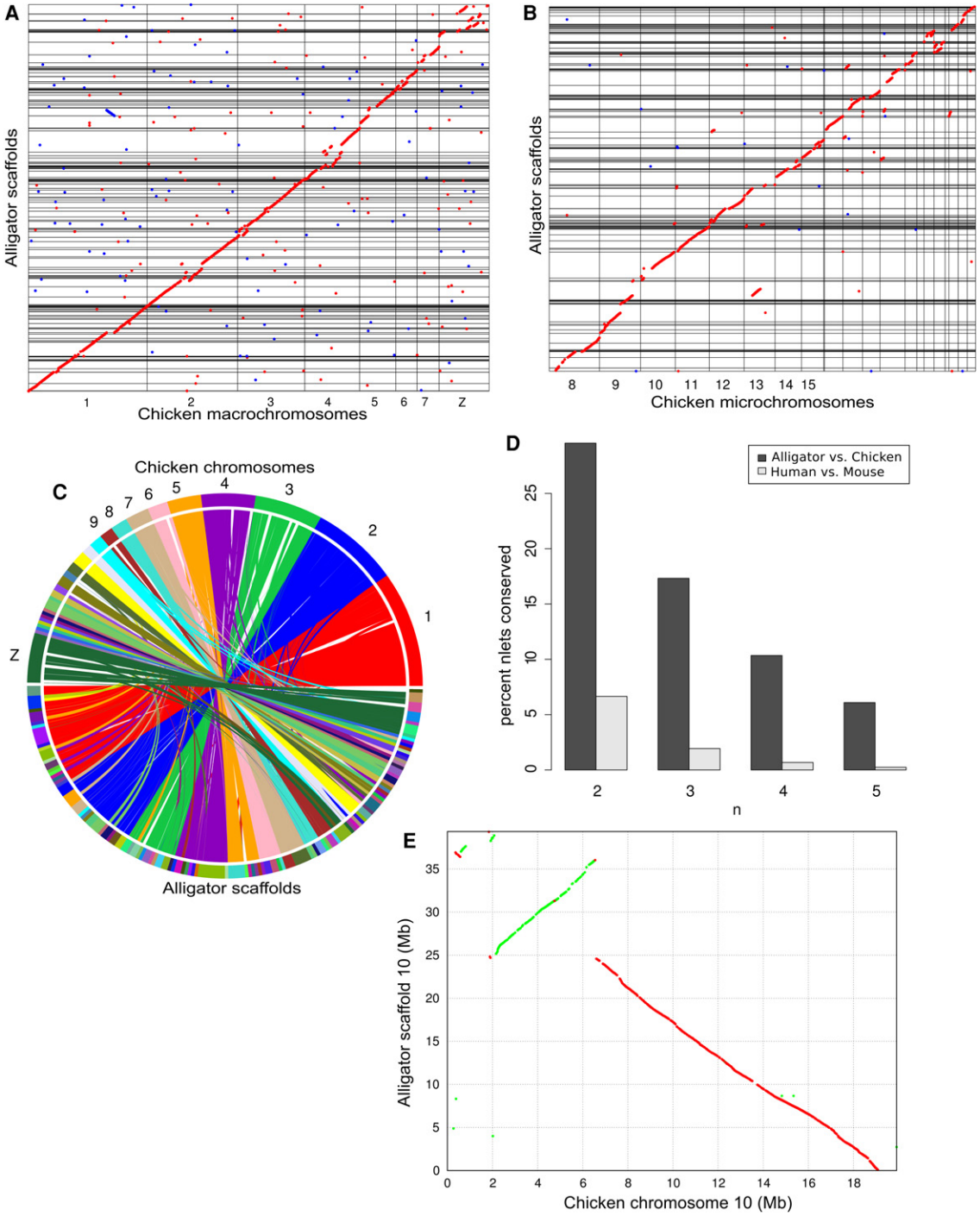

Figure 1. Our new long-range assembly of the American alligator genome allows analysis of the synteny between crocodilians and birds. (*A*,*B*) Dot plots of an anchored whole-genome alignment between the chicken and American alligator genomes show a high degree of synteny, with many long alligator scaffolds covering significant portions of chicken chromosomes, including macrochromosomes (*A*) and microchromosomes (*B*). (*C*) A circle plot of synteny between the alligator and chicken genomes made using SyMAP (Soderlund et al. 2011). (*D*) Conservation of ordered gene doublets, triplets, quadruplets, and quintuplets between alligators and chickens versus between humans and mice, showing much higher synteny between alligators and chickens than between humans and mice. (*E*) Alligator scaffold 10 covers a vast majority of the chicken microchromosome 10. However, there are several small inversions and one large inversion between the two. Green and red dots represent forward and reverse matches, respectively.

correspond to whole arms of chicken chromosomes. For example, two alligator scaffolds almost completely cover GGA7. Furthermore, the microchromosome GGA10 is almost fully covered by a single alligator scaffold, scaffold 10 (Fig. 1E), with one large inversion and numerous small local inversions.

To contrast the levels of genome rearrangement in archosaurs and mammals, we compared conservation of gene order between alligators and chickens (242 Mya TMRCA) to that between humans and mice (110 Mya TMRCA) (Crottini et al. 2012). We calculated the percentage of ordered pairs, triplets, quadruplets, and quintuplets of directly adjacent genes that occur in both alligators and chickens and both humans and mice. We found four times greater conservation of gene pair synteny between alligators and chickens than between humans and mice, nine times greater conservation of gene triplets, 15 times greater conservation of quadruplets, and 25 times greater conservation of quintuplets (Fig. 1D).

A closer look at synteny between the chicken Z Chromosome and the alligator genome reveals the expected inversion around the avian sex-determining gene *DMRT1* (Supplemental Fig. S1). This result is concordant with the *Z*-linked inversions previously predicted by examining gene synteny between the avian Z Chromosome and other reptilian outgroups such as the green anole *Anolis carolinensis*, red-tailed boa *Boa constrictor*, and Mexican musk turtle *Staurotypus triporcatus* (Kawagoshi et al. 2014; Zhou et al. 2014). While these studies show that this inversion occurred after the divergence of archosaurs from other amniotes, our result further pinpoints the time of the beginning of evolution of avian sex chromosomes by providing the first conclusive evidence that this inversion occurred in the common ancestor of birds after divergence with crocodilians.

Comparative assembly

We used the American alligator genome to scaffold the previously published genome assemblies of two other crocodilians, the saltwater crocodile *C. porosus* and the gharial *G. gangeticus*, based on synteny. These published assemblies have scaffold N50s of 205 and 127 kb, respectively. We performed comparative assembly on these genomes with Ragout (Kolmogorov et al. 2014). Through this process, we were able to increase the scaffold N50 of the saltwater crocodile genome assembly from 205 kb to 84 Mb and the gharial genome assembly from 128 kb to 96 Mb. For comparison, the mean chromosome sequence length of the saltwater crocodile and gharial genomes are 117 and 165 Mb, respectively.

To assess the accuracy of the synteny based scaffolding, we tested a random set of the scaffold joins predicted by Ragout for each species. We verified predicted scaffold joins using PCR with primers chosen such that the amplified regions would be unique in the genome assembly and would span the joins made by Ragout. We successfully amplified these gap regions for 18 out of 20 predicted joins tested in the saltwater crocodile genome and 22 out of 29 predicted joins tested in the gharial genome. Full results and primers used for join verification are in Supplemental Table S1.

Transposable elements

Repetitive sequences comprise more than one-third of the alligator genome assembly (Supplemental Table S2). Almost a quarter of the genome is derived from just three TE superfamilies: LINE CR1s (12.2%) and the DNA transposons Harbinger (7.5%) and hAT (8.2%). TEs in general appear to accumulate more slowly in crocodilians than in other vertebrate taxa (excluding Testudines), and few new TE families, or even insertions, appear in any lineage of crocodilians since their divergence (Green et al. 2014; Suh et al. 2015). Data from AllMis2 are consistent with these findings. Repeat content in general and from each of the dominant superfamilies are similar not only between alligator assemblies but also among crocodilians (Supplemental Table S2), as determined by premasked genomes (http://repeatmasker.org/genomicDatasets/RMGenomicDatasets.html; accessed March 15, 2016). Only CR1 content varies between alligator assemblies to an appreciable degree. An additional 2.6% of the AllMis2 assembly is identifiable as CR1 compared with that of AllMis1. The differences in CR1 content between assemblies may be greater than it seems when contrasted with the near uniformity in the TE annotations across existing crocodilian assemblies (Supplemental Table S2; Green et al. 2014). Highly repetitive, nearly identical sequences are difficult to assemble from short reads and are likely underrepresented in genome assemblies, so an improved assembly may be able to identify these to a greater degree. Repeats in both AllMis1 and AllMis2 are biased toward those >10% diverged from their respective consensus element (Supplemental Fig. S2). No clear "burst" of CR1 activity specific to any one divergence bin is apparent, so it is likely that the additional CR1 insertions are distributed among elements with high and low mutation loads.

Small RNAs

MicroRNAs have been identified de novo in model vertebrate species, but for nonmodel species, miRNAs are usually identified based on sequence conservation with known miRNAs in other species. We sequenced a library of small RNAs isolated from alligator testis and used the resulting reads to predict 60 putative miRNAs after filtering for quality, including one, aca-mir-425, which appears in the American alligator, saltwater crocodile, and gharial genomes, but not in the chicken genome. See Supplemental Results for more details.

Sex-biased gene expression

A crucial step toward understanding TSD in the American alligator is determining which genes are turned on or off based on temperature at various developmental stages. This necessitates the generation of a catalog of genes that show significantly different expression between eggs incubated at MPT and those incubated at FPT. To this end, we incubated a total of 168 alligator eggs at either MPT or FPT for either 0, 3, or 30 d after developmental stage 19. The TSP spans developmental stages 21 to 24 (Lang and Andrews 1994), which occur between our 3- and 30-d timepoints. We harvested the embryos after incubation, subdissected the gonad-adrenal-mesonephros (GAM) complex into its constituent parts, and performed RNA sequencing on each of these three tissues for each sample. We sequenced at least three biological replicates from different clutches for each tissue and time point combination. See Supplemental Table S3 for a list of libraries sequenced along with their NCBI accessions.

We used the resulting RNA-seq data to quantify gene expression and determine which genes are differentially expressed between developing male and female embryos at these developmental stages in these three tissues. We used Cuffdiff 2 to perform these tasks (Trapnell et al. 2013). Cuffdiff 2 generates a normalized expression value in fragments per kilobase of transcript per million mapped reads (FPKM) for each gene in each library as well as an FDR-adjusted *P*-value for determining whether gene expression is significantly different between two

sets of replicates. We considered any gene with an FDR-adjusted $P \leq 0.05$ to be differentially expressed between males and females in a given tissue at a given time point.

Due to conditions prior to egg collection, embryos can sometimes develop as a different sex than expected based on incubation temperature after collection (McCoy et al. 2015). We could not confirm sex histologically as both gonads of each embryo were used for RNA sequencing, so we confirmed the sex of each embryo by comparing gonadal expression of *CYP19A1* to *AMH* as in previous studies (Kohno et al. 2015; McCoy et al. 2015). One embryo from clutch 13 was female despite incubation at MPT (Fig. 2B), so we excluded it from differential expression analysis.

We found many genes with differential expression between males and females in each tissue at both the 3- and 30-d timepoints (Fig. 2A; Supplemental Table S4). Unsurprisingly, the gonads at the post-TSP time point displayed the most sexual dimorphism in gene expression. The genes differentially expressed between male and female embryos in these samples include many genes known to be involved in early sexual development in other vertebrates (Fig. 2B). Such male development genes include *SOX9*, which triggers testis formation, and *AMH*, which inhibits the formation of Müllerian ducts (De Santa Barbara et al. 1998). Female development genes with female-biased expression in the post-TSP gonads include *FST*, which inhibits the production of follicle-stimulating hormone (Ying et al. 1987). *CYP19A1*, which produces aromatase, the enzyme that converts androgens to estrogens (Toda and Shizuta 1993), was the gene with the largest sex-bias fold-change in either direction, with a $\log_2$ fold-change of 12.463. This is consistent with other studies of aromatase expression in embryos incubated at different temperatures (Smith et al. 1995; Gabriel et al. 2001). *ESR1*, the gene coding for estrogen receptor alpha, and *CTCF* are highly expressed in both male and female gonads at this time point, with respective average FPKM values of 24.08 and 47.02 but no significant sex bias.

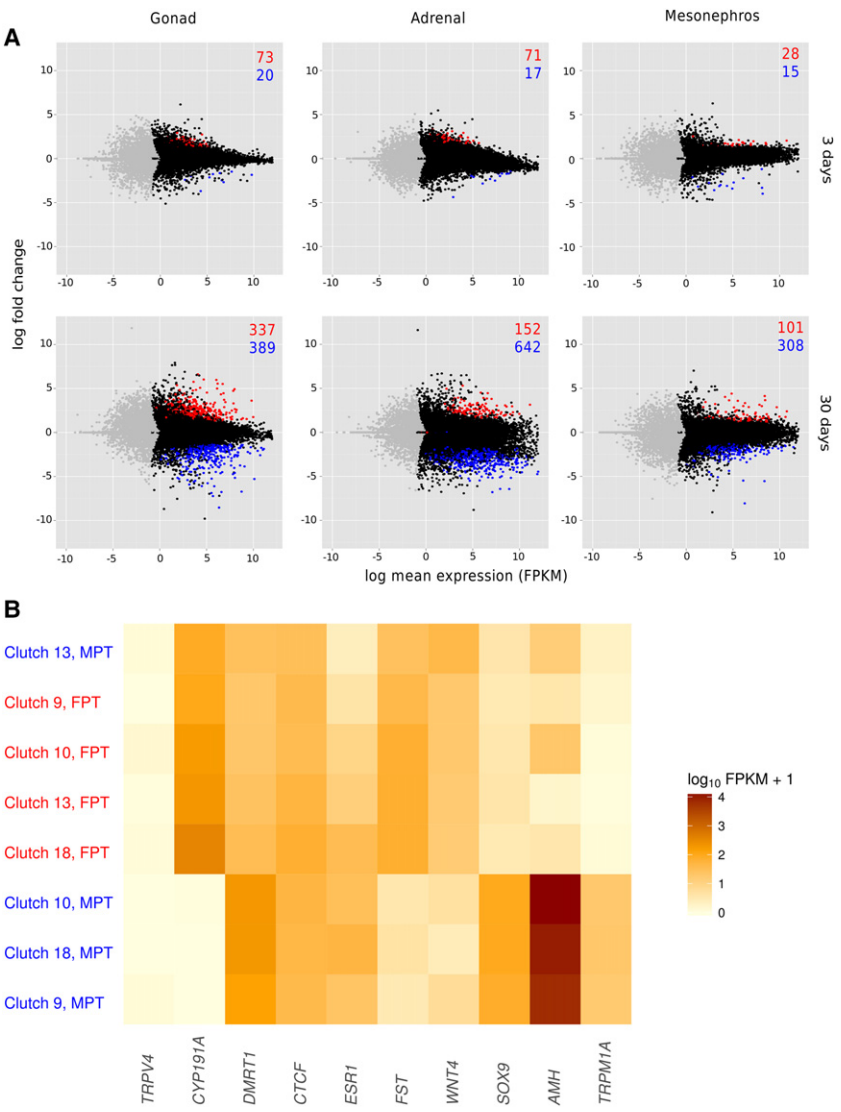

Figure 2. Sex-biased gene expression in alligator embryos. (*A*) Mean expression versus fold-change for all genes in three tissues at two developmental time points. Genes found to have female-biased expression and male-biased expression are colored in red and blue, respectively. Numbers of sex-biased genes for each tissue and time point are given in the *upper right* of each plot. (*B*) Gonadal expression of genes of interest at the 30-d time point in eight embryos. The embryo from clutch 13 incubated at MPT displays a distinctly female expression pattern despite being incubated at MPT and was thus excluded from further analyses.

We have included lists of significantly enriched GO terms among genes with male- and female-biased expression in the gonads at 30 d generated using FUNC (Prüfer et al. 2007) in Supplemental Table S5. One significantly overrepresented GO term among these male-biased genes is "detection of temperature stimulus" (GO:0016048). The only male-biased gene with this GO term is the transient receptor potential cation channel *TRPM1A*. Another transient receptor potential cation channel gene, *TRPV4*, has been suggested as one thermosensitive gene involved in TSD in the American alligator (Yatsu et al. 2015). We found no significant expression or sex-bias of *TRPV4* at any of our time points in any of the three tissues. However, Yatsu et al. (2015) found sex-biased expression of *TRPV4* only during the TSP at developmental stages 21 and 24, while we sampled only before and after the TSP.

Estrogenic regulation of gene expression

Estrogen regulation of gene expression is best understood in humans from work dissecting the molecular basis of estrogen-responsive and nonresponsive breast cancers in tissue models. That work has shown that in human estrogen-responsive tissues, estrogen promotes the expression of genes by allowing estrogen receptors to bind to enhancer DNA sequences (Dahlman-Wright et al. 2006). However, the enhancers to which estrogen receptors bind are usually distal to

the genes they regulate (Carroll et al. 2006). Due to the sex-reversing effects of estrogen exposure during crocodilian development via estrogen receptor alpha (Kohno et al. 2015) and the extreme female-biased expression of the gene coding for aromatase, we hypothesize that estrogen signaling through ESR1 binding is a major driver of female-biased gene expression during TSD in the American alligator.

We first tested this hypothesis by looking for enrichment of genes with female-biased expression in the post-TSP gonads of alligator embryos in the genomic regions surrounding computationally predicted estrogen receptor binding sites. The DNA-binding domain of ESR1 is perfectly conserved among humans, chickens, and alligators (Supplemental Fig. S3b; Supplemental Table S6), and the DNA-binding motif of ESR1 in human estrogen-responsive cells is well characterized (Gruber et al. 2004; Carroll et al. 2006; Lin et al. 2007). Therefore, we predicted ESR1 binding sites in the American alligator genome using the motif representing the human estrogen response element. We found that while 337 (2.26%) of the 14,943 genes expressed in the post-TSP gonad have female-biased expression, 62 (3.11%) of the 1991 expressed genes within 50 kb of a putative estrogen receptor binding site have female-biased expression (E = 44.9; enrichment factor = 1.38; Fisher's exact test $P = 4.79 \times 10^{-3}$). This indicates that genes are significantly more likely to have female-biased expression in the post-TSP gonad if they are near a location in the genome where ESR1 is predicted to bind.

In human tissue models of estrogen regulation of gene expression, whether a gene is likely to be estrogen responsive is based on its genomic location relative to not only estrogen receptor binding sites but also CTCF binding sites (Chan and Song 2008). Since this was established in 2008, studies of CTCF-mediated chromatin looping have shown that CTCF helps divide the genome into functional domains through a chromatin extrusion process that causes loops to form only where two adjacent CTCF binding motifs are oriented toward each other (Rao et al. 2014; Sanborn et al. 2015).

CTCF binding sites in the chicken genome have been experimentally determined (Martin et al. 2011), and the zinc finger domains of CTCF are perfectly conserved among human, chicken, and alligator orthologs (Supplemental Fig. S3a). We therefore used the CTCF binding motif in the chicken genome to predict CTCF binding sites in the American alligator genome. We used these binding site predictions and the most recent model of CTCF-mediated chromatin looping (Sanborn et al. 2015) to predict how chromatin loops form in the alligator genome (Fig. 3A). We predicted 19,482 chromatin loops based on CTCF binding sites, comparable to the 21,306 found experimentally in the human genome (Li et al. 2012); 3758 (19.3%) of these putative loops contain one or more ESR1 binding sites, and 10,074 (67.4%) of the 14,943 genes expressed in gonads after 30 d of incubation are within the boundaries of one or more predicted CTCF loops.

We found that while 337 (2.26%) of the 14,943 genes expressed in the post-TSP gonads have female-biased expression, 116 (3.09%) of the 3759 expressed genes in CTCF loops containing one or more ESR1 binding sites have female-biased expression (E = 84.8, enrichment factor = 1.37; Fisher's exact test $P = 7.76 \times 10^{-5}$). This finding shows a significant enrichment in female-biased gene expression in the regions of the genome predicted to be estrogen responsive under our model, providing support for our hypothesis that many of these genes are regulated by estrogen during sexual differentiation and development (Fig. 3B,C).

Among the female-biased genes in predicted estrogen-responsive regions is *WNT4*, a gene required for female development

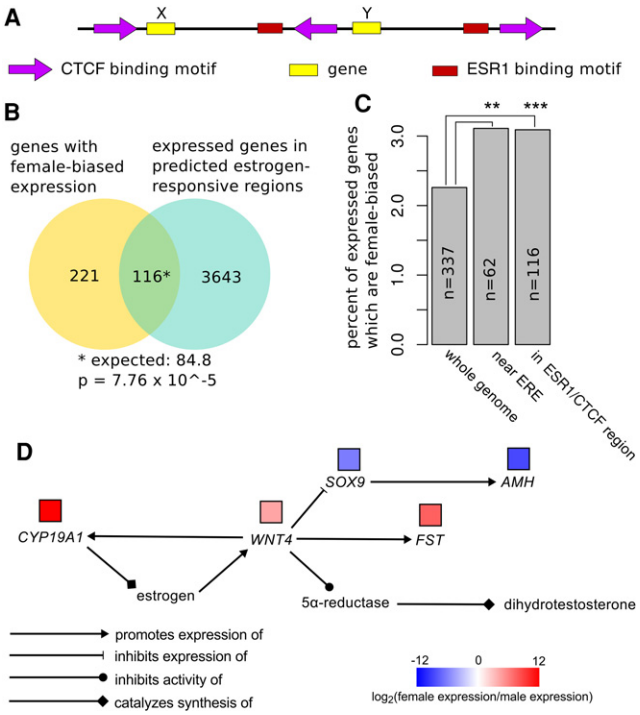

Figure 3. Genes in regions of the genome predicted to be under estrogenic regulation of gene expression are significantly more likely to be female biased in the post-TSP gonads. (*A*) Our model for predicting regions of the genome under estrogenic regulation of gene expression, based on the CTCF extrusion model (Sanborn et al. 2015) and the Chan and Song model of estrogen receptor binding site activity (Chan and Song 2008). In this example, Gene X is predicted to be estrogen responsive and Gene Y is not because Gene X is between two inward-oriented CTCF binding motifs along with an ESR1 binding site, while Gene Y is not. (*B*) Of the 14,943 genes expressed in the post-TSP gonads, 337 have female-biased expression and 3759 are in predicted estrogen-responsive genomic regions. However, 116 of these genes are both female biased and within predicted estrogen-responsive regions, a significantly higher number than the expected 84 ($P = 7.76 \times 10^{-5}$). (*C*) Percentages of expressed genes with female-biased expression in the whole genome versus near an estrogen response element and in a predicted estrogen-responsive CTCF region. Regions near an estrogen response element and predicted estrogen-responsive regions are both enriched for female-biased genes. (**) $P \leq 0.01$; (***) $P \leq 10^{-4}$. (*D*) Pathway diagram showing results of increased *CYP19A1* expression after the TSP in the gonads of embryos incubated at FPT. Sex-bias fold-changes for each gene in the pathway are shown in boxes *above* the genes.

in other vertebrates. *WNT4* suppresses *SOX9* and 5-alpha reductase activity (Fig. 3D) and promotes the formation of Müllerian ducts via frizzled receptor binding (Hsieh et al. 2002). Frizzled receptor genes *FZD2*, *FZD3*, *FZD6*, *FZD8*, and *FZD9* are all significantly expressed in the post-TSP gonads in both males and females. We therefore hypothesize that *WNT4* plays a role in sex differentiation in the American alligator similar to its role in other vertebrates, although unlike in vertebrates with GSD, its expression is determined by incubation temperature via estrogen signaling.

Discussion

We present AllMis2, an improved assembly of the American alligator (*A. mississippiensis*) genome. After demonstrating its accuracy, we used AllMis2 to examine synteny between the American

alligator and chicken (*Gallus gallus*) genomes, improve the genomes of two other crocodilian species, and predict genomic regions likely to be under estrogenic regulation of gene expression in estrogen-responsive tissues. Finally, we showed that genes in these predicted estrogen-responsive regions are significantly more likely to have female-biased expression in post-TSP gonads. We thus conclude that the genomic architecture of estrogen signaling is remarkably well conserved within vertebrates and that it is a fundamental early driver of female-biased gene expression in the post-TSP embryonic gonads of the American alligator.

Our analyses are aided by a contiguous genome, and many would not have been possible with AllMis1. Synteny blocks between the chicken genome and AllMis1 were too small and fragmented to lend significant insight to large-scale genome evolution between avians and crocodilians, while a whole-genome alignment between the chicken genome and AllMis2 shows many large synteny blocks with some inversions covering significant portions of chicken chromosomes. Synteny analysis using AllMis2 also reveals a slower rate of gene rearrangement in archosaurs than in mammals (Fig. 1C), and the first direct evidence that the initial inversion leading to the evolution of avian sex chromosomes occurred after the divergence of the crocodilian and avian lineages (Supplemental Fig. S1). Furthermore, transposable element annotation was improved by using AllMis2.

Highly repetitive, low-diversity sequences (i.e., recently active TEs) are among the most difficult to assemble, and it is likely that their presence is underestimated in genome assemblies. This could downwardly bias estimates of TE content and would particularly affect estimates of recently active TEs. It is possible that AllMis2 better represents the true TE content of the alligator genome. CR1 content increased by 2.6% between alligator assemblies (Supplemental Table S2), but sequence diversity within CR1 is similar in both assemblies (Supplemental Fig. S2). Analysis of AllMis1 suggested that TEs in general accumulate more slowly in crocodilians than in other vertebrate taxa (excluding Testudines), and few new TE families, or even insertions, have appeared in any lineage of crocodilians since their divergence (Green et al. 2014; Suh et al. 2015). Some of the variation in CR1 annotations between alligator assemblies is almost certainly due to stochasticity introduced by homology-based identification. Further, it is possible that comparable improvements to the gharial and crocodile assemblies would yield similar changes in CR1 annotation. Observations made when comparing alligator assemblies (overall increased CR1 content, few young CR1 elements) combined with our understanding of CR1 evolution in crocodilians in general (Suh et al. 2015) imply that the new alligator assembly was slightly more useful for identifying TEs.

Holleley et al. (2015) recently discovered that although the Australian bearded dragon *Pogona vitticeps* has heteromorphic sex chromosomes, it can undergo sex reversal in the wild at high temperatures. In addition, during extended hot periods, whole populations can lose their minor sex chromosomes and transition to fully TSD populations. In the context of earlier reports of thermal and hormonal overrides for GSD in several species of lizards and turtles (Barske and Capel 2008), these observations indicate that at least some components of sex determination remain sensitive to temperature even when genetic cues evolve that can override them. Here, we show that the effectors of estrogen signaling and its underlying genomic architecture are highly conserved between TSD and GSD lineages. The protein sequence of the DNA-binding domains of both ESR1 and CTCF is perfectly conserved in the alligator, human, and chicken genomes. The CTCF/ERE model for estrogen response (Chan and Song 2008) developed in estrogen-responsive human tissue culture models is predictive of female-biased gene expression in the developing alligator embryo. Aromatase and two of its downstream genes involved in sexual development in other vertebrates, *WNT4* and *SOX9*, are all differentially expressed in a temperature-dependent manner in the developing alligator embryo and in the embryos of other TSD reptiles like the red-eared slider turtle *Trachemys scripta elegans* (Ramsey and Crews 2009). We propose that some aspects of the highly conserved estrogen response may be inherently and persistently temperature sensitive. All of the 22 species of Crocodilia use TSD (Lang and Andrews 1994). Within this clade, the proposed direct link between temperature and estrogen signaling may have evolved robustness sufficient to be impervious to genetic variation. A comprehensive experimental exploration of the estrogen response in TSD versus GSD species may reveal the biochemical link between temperature and estrogen signaling.

Expression of aromatase, the enzyme that produces estrogen, has been hypothesized to be a master regulator of sex-biased gene expression in developing alligator embryos (Lance 2009) because of the ability of estrogen exposure to cause sex reversal in embryos incubated at MPT (Bull et al. 1988) and its extreme sex-biased expression in embryonic gonads after TSP (Gabriel et al. 2001). While much work is currently being performed to determine the pathway that allows aromatase expression to vary with temperature (Parrott et al. 2014; Yatsu et al. 2015; McCoy et al. 2016), less attention has been paid to the questions of which genes estrogen regulates during sexual development in American alligators or how estrogen regulates them despite its pivotal role early in embryonic sexual differentiation in alligators. Our data do not speak to the hypothesis that *TRPV4* is a component of the temperature-sensing apparatus responsible for TSD (Yatsu et al. 2015) as we find no evidence of expression of this gene in any tissue at any of our time points. Importantly, we took samples before and after the TSP. Future work measuring gene expression during the TSP may more clearly determine the roles of *TRPV4*, *TRPM1*, and perhaps other candidate thermosensitive signaling molecules.

In this article, we hypothesized that estrogen regulates gene expression in developing American alligator embryos through the same mechanism by which it is known to do so in humans and that this mechanism can explain much of the female-biased gene expression that occurs after the TSP. By using the latest model of estrogen regulation of gene expression and CTCF-mediated chromatin looping in humans, we demonstrated that the regions of the American alligator genome that are most likely to be under estrogenic regulation of gene expression are enriched for female-biased gene expression. Our results provide new evidence for Lance's hypothesis that aromatase and its production of estrogen are a major driver of sex-biased gene expression in TSD in the American alligator (Lance 2009). These results show that despite the different roles of estrogenic regulation of gene expression in sexual development between humans and alligators, much of the underlying mechanism responsible for estrogen regulation of gene expression is conserved between these two species.

Although our study does not fully elucidate the downstream effects of female-biased gene expression caused by estrogen signaling in the post-TSP gonads, *WNT4*'s female-biased expression and presence in a predicted estrogen-sensitive region provide a possible explanation for some of these effects. In mammals, *WNT4* expression prevents the formation of male-specific vasculature by preventing migration of endothelial and steroidogenic cells from mesonephros tissues to gonads (Jeays-Ward et al. 2003). It

performs this action through up-regulation of follistatin (Yao et al. 2004). *FST*, the gene coding for follistatin, is among the genes we find to have female-biased expression in the post-TSP alligator gonad, suggesting that *FST* may be among the genes indirectly regulated by estrogen signaling after the TSP. Furthermore, *WNT4* promotes expression of aromatase in mammals (Boyer et al. 2010). If the same is true in post-TSP embryonic alligator gonads, *WNT4* and aromatase may cooperate through a feed-forward mechanism in which estrogen promotes the expression of *WNT4* and *WNT4* promotes the expression of aromatase, which then creates more estrogen.

Methods

Sequencing and assembly

DNA was extracted with Qiagen blood and cell midi kits according to the manufacturer's instructions. Briefly, cells were lysed and centrifuged to isolate the nuclei. The nuclei were further digested with a combination of Proteinase K and RNase A. The DNA was bound to a Qiagen genomic column, washed, eluted and precipitated in isopropanol, and pelleted by centrifugation. After drying, the pellet was resuspended in 200 μL TE (Qiagen). We generated the Chicago library as previously described by Putnam et al. (2016). Briefly, high-molecular-weight DNA was assembled into chromatin in vitro and then chemically cross-linked before being restriction digested. The overhangs were filled in with a biotinylated nucleotide, and the chromatin was incubated in a proximity-ligation reaction. The cross-links were then reversed, and the DNA purified from the chromatin. The library was then sonicated and finished using the NEB ultra library preparation kit (NEB catalog no. E7370), according to the manufacturer's instructions, with the exception of a streptavidin bead capture step prior to indexing PCR. We sequenced the Chicago library on a single lane on the Illumina HiSeq 2500, resulting in 210 million read pairs.

The contig assembly was made with MERACULOUS (Chapman et al. 2011) and scaffolded using the Chicago library with Dovetail Genomics' HiRise scaffolder as previously described by Putnam et al. (2016).

Annotation

We made gene predictions using AUGUSTUS version 3.0.3 (Stanke et al. 2006). We provided as extrinsic evidence to AUGUSTUS RNA-seq alignments made using TopHat2 version 2.0.14 (Kim et al. 2013), repetitive region predictions made using RepeatScout (Price et al. 2005) and RepeatMasker Open-4.0 (Smit et al. 2015), and alignments of published chicken protein sequences made using Exonerate version 2.2.0 (Slater and Birney 2005). We assigned names to these predicted proteins and genes using reciprocal best hits BLAST searches between the set of predicted protein sequences and published protein sequences from related organisms. We also assigned GO terms to our predicted proteins using InterProScan (Jones et al. 2014).

To annotate the genome for microRNAs, we extracted and purified small RNAs from testis tissue of a reproductively-mature alligator caught in the Rockefeller Wildlife Refuge (Grand Chenier, LA) using TRIzol reagent followed by an ethanol precipitation. We sequenced the resulting library on a MiSeq and then, after filtering, used the miRDeep2 pipeline (Friedländer et al. 2012) and MapMi (Guerra-Assunção and Enright 2010) to align these sequences to and predict miRNAs in the alligator genome.

For more detail on our annotation process, see the Supplemental Methods.

Synteny

We created synteny maps and calculated synteny statistics using SyMAP 4.2 (Soderlund et al. 2011), considering only scaffolds of at least 100 kb and ordering the alligator scaffolds based on the chicken genome. We determined synteny for Galgal4 against both the previous version of the alligator genome (Green et al. 2014) and the updated alligator genome for comparison.

To calculate conservation of ordered gene *n*-lets between the alligator and chicken genomes, as well as the human and mouse genomes, we first found homologs in the second genome for genes in the first genome by performing a blastp search of the protein sequence of the primary isoform of each gene in the first genome against a database of all protein sequences in the second genome. We consider *n*-lets only of directly adjacent genes on the same scaffold. We then counted the number of ordered gene *n*-lets in the first genome whose homologs also appear contiguously in the same order in the second genome.

Comparative assembly

We used synteny blocks to separate large structural variants from small polymorphisms, taking a hierarchical approach, with multiple sets of synteny blocks, each defined at a different resolution, from the coarsest, karyotype level all the way down to the fine-grained base level. To create the hierarchy, we used the principles developed by Sibelia tool (Minkin et al. 2013), which can create such a hierarchy for bacterial genomes, but adapted to use a multi-size A-Bruijn graph algorithm for constructing synteny blocks from a multiple genome alignment file in HAL format (Hickey et al. 2013), produced by Progressive Cactus (Paten et al. 2011). At each level of resolution, we used Ragout (Kolmogorov et al. 2014) to decompose the input genomes into synteny blocks and join scaffolds based on this synteny.

We assessed the accuracy of joins by designing primer pairs bracketing the gaps using Primer3 (Untergasser et al. 2012). We PCR amplified saltwater crocodile or gharial DNA with these primers at annealing temperatures ranging from 58°C–62°C for 20 cycles. The joins, primers, and full results are in Supplemental Table S1.

Transposable elements

We identified transposable elements and low complexity repetitive sequences in the alligator (*A. mississippiensis*) genome using RepeatMasker Open-4.0 (Smit et al. 2015) and homology based searches with all known alligator repeats (RepBase Update v21.02). We created a repeat accumulation profile by calculating the Kimura 2-parameter (Kimura 1980) genetic distance between individual insertions and the homologous repeat in the *A. mississippiensis* library.

Egg harvesting, incubation, and dissection

All field and laboratory work were conducted under permits from the Florida Fish and Wildlife Conservation Commission and US Fish and Wildlife Service (permit no. SPGS-1 0-44). Five clutches of alligator eggs were collected from the Lake Woodruff National Wildlife Refuge, where relatively low chemical contamination of persistent organic pollutants allow American alligators to exhibit healthy reproductive activity. One egg from each clutch was dissected to identify the developmental stage of the embryo based on criteria described by Ferguson (1985). Eggs were incubated at 30°C (FPT) until they reached stage 19 based on an equation predicting their development (Kohno and Guillette 2013). At the predicted stage 19, which was before the TSP (stage 21–24) for alligator

TSD (Lang and Andrews 1994), the incubation temperature was either kept constant at FPT or increased to 33°C (MPT). The alligator embryos were dissected and the GAM complex was isolated and preserved in ice-cold RNA*later* (Ambion/Thermo Fisher Scientific) at 3 or 30 d after the stage 19. Gonadal tissues were carefully isolated from GAM under a dissection microscope after RNA stabilization in RNA*later* and stored at −80°C until RNA isolation.

RNA sequencing, expression quantification, and differential expression analysis

Total RNAs were then extracted from the GAM samples using TRIreagent LS (Sigma). Poly(A)$^+$ RNA sequencing libraries were made from each sample using the TruSeq RNA library preparation kit v1 (Illumina). A total of 60 libraries were created by PCR amplification with Illumina barcoding primers at 17 reaction cycles and quantified using a Bioanalyzer DNA 1000 kit (Agilent). Libraries were then pooled and sequenced on a HiSeq 2000 Sequencing system (Illumina).

We removed adapters from the reads using SeqPrep (https://github.com/jstjohn/SeqPrep) with default parameters and aligned them to the alligator genome using TopHat2 (Kim et al. 2013) with default parameters. We used Cuffdiff 2 to calculate normalized expression values, fold-changes, and FDR-adjusted *P*-values for each gene in each tissue at each time point (Trapnell et al. 2013). Cuffdiff 2 reports expression values normalized by transcript length and library size in FPKM for reporting the expression of individual genes in each library in values that are comparable between different genes. Expression values in FPKM are useful for generating heatmaps and reporting average expression values for a gene, but Cuffdiff 2 uses raw counts rather than FPKM for differential expression analysis. For each gonad sample at the 30 d, we compared FPKMs of two genes, *CYP19A1* and *AMH*, to verify the sex of the embryo as in previous studies (Kohno et al. 2015; McCoy et al. 2015), resulting in one sample, the embryo from clutch 13 incubated at MPT, being removed from further analysis. We used an FDR-adjusted *P*-value reported by Cuffdiff 2 for each gene for the null hypothesis that expression levels of that gene in tissues incubated at MPT and FPT are drawn from the same distribution. We considered a gene to be sex-biased if its FDR-adjusted *P*-value was ≤ 0.05.

We used FUNC to perform GO enrichment analysis (Prüfer et al. 2007). We ran the hypergeometric variant of FUNC with default options and the October 2016 release of GO tables.

Predicting estrogen-responsive regions of the alligator genome

The DNA-binding domain of estrogen receptor alpha (ESR1) and the zinc fingers of CTCF are identically conserved in protein sequence among human, chicken, and alligator (Supplemental Fig. S3), suggesting that the DNA-binding motifs of these proteins are also conserved among these species. We predicted binding locations for these proteins by searching the alligator genome for sequences matching the human ESR1-binding motif (Lin et al. 2007) and the chicken CTCF-binding motif (Martin et al. 2011) using PoSSuM-search (Beckstette et al. 2006) with *P*-value cutoffs of 4.388×10^{-6} for ESR1 and 1.214×10^{-6} for CTCF. We considered any genomic region between two inward-facing CTCF motifs within 700 kb to be possibly estrogen responsive if it contained one or more ER-binding motifs.

Data access

The sequence data from this study have been submitted to the NCBI Sequence Read Archive (SRA; http://www.ncbi.nlm.nih.gov/sra/) under accession number SRP057608 and to the BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) under accession numbers PRJNA322197, PRJNA163131, PRJNA172383, and PRJNA285470. The genome assembly AllMis2 from this study has been submitted to the NCBI Assembly database (https://www.ncbi.nlm.nih.gov/assembly/) under accession number GCA_000281125.4.

Competing interest statement

R.E.G. is a cofounder and paid consultant of Dovetail Genomics LLC. B.L.O. is a paid consultant.

Acknowledgments

We thank the Florida Fish and Wildlife Conservation Commission and the US Fish and Wildlife Service for their assistance in obtaining collection permits; Steven Weber, Darrin Schultz, Stefany Rubio (UC Santa Cruz), and Jenny Korstian (Texas Tech University) for technical assistance; and Beth Shapiro and Angela Brooks (UC Santa Cruz) for discussion regarding the project. This work was supported by the National Institutes of Health under award numbers 5U54HG007990 (National Human Genome Research Institute), GM085121, and GM109146 and by grants from the W.M. Keck Foundation and the Simons Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Barske LA, Capel B. 2008. Blurring the edges in vertebrate sex determination. *Curr Opin Genet Dev* **18:** 499–505.

Beckstette M, Homann R, Giegerich R, Kurtz S. 2006. Fast index based algorithms and software for matching position specific scoring matrices. *BMC Bioinformatics* **7:** 389.

Boyer A, Lapointe E, Zheng X, Cowan RG, Li H, Quirk SM, DeMayo FJ, Richards JS, Boerboom D. 2010. WNT4 is required for normal ovarian follicle development and female fertility. *FASEB J* **24:** 3010–3025.

Bull JJ, Gutzke WH, Crews D. 1988. Sex reversal by estradiol in three reptilian orders. *Gen Comp Endocrinol* **70:** 425–428.

Carroll JS, Meyer CA, Song J, Li W, Geistlinger TR, Eeckhoute J, Brodsky AS, Keeton EK, Fertuck KC, Hall GF, et al. 2006. Genome-wide analysis of estrogen receptor binding sites. *Nat Genet* **38:** 1289–1297.

Chan CS, Song JS. 2008. CCCTC-binding factor confines the distal action of estrogen receptor. *Cancer Res* **68:** 9041–9049.

Chapman JA, Ho I, Sunkara S, Luo S, Schroth GP, Rokhsar DS. 2011. Meraculous: de novo genome assembly with short paired-end reads. *PLoS One* **6:** e23501.

Crews D, Wibbels T, Gutzke WH. 1989. Action of sex steroid hormones on temperature-induced sex determination in the snapping turtle *(Chelydra serpentina)*. *Gen Comp Endocrinol* **76:** 159–166.

Crottini A, Madsen O, Poux C, Strauss A, Vieites DR, Vences M. 2012. Vertebrate time-tree elucidates the biogeographic pattern of a major biotic change around the K-T boundary in Madagascar. *Proc Natl Acad Sci* **109:** 5358–5363.

Dahlman-Wright K, Cavailles V, Fuqua SA, Jordan VC, Katzenellenbogen JA, Korach KS, Maggi A, Muramatsu M, Parker MG, Gustafsson JA. 2006. International Union of Pharmacology. LXIV. Estrogen receptors. *Pharmacol Rev* **58:** 773–781.

De Santa Barbara P, Bonneaud N, Boizet B, Desclozeaux M, Moniot B, Sudbeck P, Scherer G, Poulat F, Berta P. 1998. Direct interaction of SRY-related protein SOX9 and steroidogenic factor 1 regulates transcription of the human anti-Müllerian hormone gene. *Mol Cell Biol* **18:** 6653–6665.

Ferguson MWJ. 1985. Development. In *Biology of the reptilia* (ed. Gans C, et al.), Vol. 14, pp. 329–491. John Wiley and Sons, New York.

Ferguson MW, Joanen T. 1982. Temperature of egg incubation determines sex in Alligator mississippiensis. *Nature* **296:** 850–853.

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.

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH, et al. 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature* **462:** 58–64.

Gabriel WN, Blumberg B, Sutton S, Place AR, Lance VA. 2001. Alligator aromatase cDNA sequence and its expression in embryos at male and female incubation temperatures. *J Exp Zool* **290:** 439–448.

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.

Gruber CJ, Gruber DM, Gruber IM, Wieser F, Huber JC. 2004. Anatomy of the estrogen response element. *Trends Endocrinol Metab* **15:** 73–78.

Guerra-Assunção JA, Enright AJ. 2010. MapMi: automated mapping of microRNA loci. *BMC Bioinformatics* **11:** 133.

Hickey G, Paten B, Earl D, Zerbino D, Haussler D. 2013. HAL: a hierarchical format for storing and analyzing multiple genome alignments. *Bioinformatics* **29:** 1341–1342.

Holleley CE, O'Meally D, Sarre SD, Marshall Graves JA, Ezaz T, Matsubara K, Azad B, Zhang X, Georges A. 2015. Sex reversal triggers the rapid transition from genetic to temperature-dependent sex. *Nature* **523:** 79–82.

Hsieh M, Johnson MA, Greenberg NM, Richards JS. 2002. Regulated expression of Wnts and Frizzleds at specific stages of follicular development in the rodent ovary. *Endocrinology* **143:** 898–908.

Jeays-Ward K, Hoyle C, Brennan J, Dandonneau M, Alldus G, Capel B, Swain A. 2003. Endothelial and steroidogenic cell migration are regulated by WNT4 in the developing mammalian gonad. *Development* **130:** 3663–3670.

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.

Kawagoshi T, Uno Y, Nishida C, Matsuda Y. 2014. The *Staurotypus* turtles and aves share the same origin of sex chromosomes but evolved different types of heterogametic sex determination. *PLoS One* **9:** e105315.

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.

Kohno S, Guillette L Jr. 2013. Endocrine disruption and reptiles: using the unique attributes of temperature-dependent sex determination to assess impacts. In *Endocrine disrupters: hazard testing and assessment methods, (ed*. Mattheissen P), pp. 245–271. John Wiley and Sons, New York.

Kohno S, Katsu Y, Urushitani H, Ohta Y, Iguchi T, Guillette LJ Jr. 2010. Potential contributions of heat shock proteins to temperature-dependent sex determination in the American alligator. *Sex Dev* **4:** 73–87.

Kohno S, Bernhard MC, Katsu Y, Zhu J, Bryan TA, Doheny BM, Iguchi T, Guillette LJ. 2015. Estrogen receptor 1 (ESR1; ERα), not ESR2 (ERβ), modulates estrogen-induced sex reversal in the American alligator, a species with temperature-dependent sex determination. *Endocrinology* **156:** 1887–1899.

Kolmogorov M, Raney B, Paten B, Pham S. 2014. Ragout: a reference-assisted assembly tool for bacterial genomes. *Bioinformatics* **30:** i302–i309.

Laganière J, Deblois G, Lefebvre C, Bataille AR, Robert F, Giguère V. 2005. From the Cover: location analysis of estrogen receptor α target promoters reveals that *FOXA1* defines a domain of the estrogen response. *Proc Natl Acad Sci* **102:** 11651–11656.

Lance VA. 2009. Is regulation of aromatase expression in reptiles the key to understanding temperature-dependent sex determination? *J Exp Zool A Ecol Genet Physiol* **311:** 314–322.

Lang JW, Andrews HV. 1994. Temperature-dependent sex determination in crocodilians. *J Exp Zool A Ecol Genet Physiol* **270:** 28–44.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v1 [q-bio.GN].

Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, et al. 2012. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell* **148:** 84–98.

Lin CY, Vega VB, Thomsen JS, Zhang T, Kong SL, Xie M, Chiu KP, Lipovich L, Barnett DH, Stossi F, et al. 2007. Whole-genome cartography of estrogen receptor α binding sites. *PLoS Genet* **3:** e87.

Martin D, Pantoja C, Fernández Miñán A, Valdes-Quezada C, Moltó E, Matesanz F, Bogdanović O, de la Calle-Mustienes E, Domínguez O, Taher L, et al. 2011. Genome-wide CTCF distribution in vertebrates defines equivalent sites that aid the identification of disease-associated genes. *Nat Struct Mol Biol* **18:** 708–714.

McCoy JA, Parrott BB, Rainwater TR, Wilkinson PM, Guillette LJ Jr. 2015. Incubation history prior to the canonical thermosensitive period determines sex in the American alligator. *Reproduction* **150:** 279–287.

McCoy JA, Hamlin HJ, Thayer L, Guillette LJ, Parrott BB. 2016. The influence of thermal signals during embryonic development on intrasexual and sexually dimorphic gene expression and circulating steroid hormones in American alligator hatchlings (*Alligator mississippiensis*). *Gen Comp Endocrinol* **238:** 47–54.

Milnes MR, Bryan TA, Medina JG, Gunderson MP, Guillette LJ. 2005. Developmental alterations as a result of in ovo exposure to the pesticide metabolite p,p′-DDE in Alligator mississippiensis. *Gen Comp Endocrinol* **144:** 257–263.

Minkin I, Patel A, Kolmogorov M, Vyahhi N, Pham S. 2013. Sibelia: a scalable and comprehensive synteny block generation tool for closely related microbial genomes. In *Algorithms in bioinformatics* (ed. Darling A, Stoye J), pp. 215–229. Springer-Verlag, Berlin.

Morrish BC, Sinclair AH. 2002. Vertebrate sex determination: many means to an end. *Reproduction* **124:** 447–457.

Nakabayashi O, Kikuchi H, Kikuchi T, Mizuno S. 1998. Differential expression of genes for aromatase and estrogen receptor during the gonadal development in chicken embryos. *J Mol Endocrinol* **20:** 193–202.

Nilsson S, Mäkelä S, Treuter E, Tujague M, Thomsen J, Andersson G, Enmark E, Pettersson K, Warner M, Gustafsson JA. 2001. Mechanisms of estrogen action. *Physiol Rev* **81:** 1535–1565.

Parrott BB, Kohno S, Cloy-McCoy JA, Guillette LJ. 2014. Differential incubation temperatures result in dimorphic DNA methylation patterning of the SOX9 and aromatase promoters in gonads of alligator (*Alligator mississippiensis*) embryos. *Biol Reprod* **90:** 2.

Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. 2011. Cactus: algorithms for genome multiple sequence alignment. *Genome Res* **21:** 1512–1528.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics* **21**(Suppl 1): i351–i358.

Prüfer K, Muetzel B, Do H, Weiss G, Khaitovich P, Rahm E, Pääbo S, Lachmann M, Enard W. 2007. FUNC: a package for detecting significant associations between gene sets and ontological associations. *BMC Bioinformatics* **8:** 41.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW, et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res* **26:** 342–350.

Ramsey M, Crews D. 2009. Steroid signaling and temperature-dependent sex determination: reviewing the evidence for early action of estrogen during ovarian determination in turtles. *Semin Cell Dev Biol* **20:** 283–292.

Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, et al. 2014. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159:** 1665–1680.

Sanborn AL, Rao SS, Huang SC, Durand NC, Huntley MH, Jewett AI, Bochkov ID, Chinnappan D, Cutkosky A, Li J, et al. 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci* **112:** E6456–E6465.

Schroeder AL, Metzger KJ, Miller A, Rhen T. 2016. A novel candidate gene for temperature-dependent sex determination in the common snapping turtle. *Genetics* **203:** 557–571.

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

Shoemaker-Daly CM, Jackson K, Yatsu R, Matsumoto Y, Crews D. 2010. Genetic network underlying temperature-dependent sex determination is endogenously regulated by temperature in isolated cultured *Trachemys scripta* gonads. *Dev Dyn* **239:** 1061–1075.

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*. http://www.repeatmasker.org/.

Smith CA, Elf PK, Lang JW, Joss JMP. 1995. Aromatase enzyme activity during gonadal sex differentiation in alligator embryos. *Differentiation* **58:** 281–290.

Soderlund C, Bomhoff M, Nelson WM. 2011. SyMAP v3.4: a turnkey synteny system with application to plant genomes. *Nucleic Acids Res* **39:** e68.

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–W439.

Suh A, Churakov G, Ramakodi MP, Platt RN, Jurka J, Kojima KK, Caballero J, Smit AF, Vliet KA, Hoffmann FG, et al. 2015. Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biol Evol* **7:** 205–217.

Toda K, Shizuta Y. 1993. Molecular cloning of a cDNA showing alternative splicing of the 5′-untranslated sequence of mRNA for human aromatase P-450. *Eur J Biochem* **213:** 383–389.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. 2013. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat Biotechnol* **31:** 46–53.

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3: new capabilities and interfaces. *Nucleic Acids Res* **40:** e115.

Welboren WJ, van Driel MA, Janssen-Megens EM, van Heeringen SJ, Sweep FC, Span PN, Stunnenberg HG. 2009. ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *EMBO J* **28:** 1418–1428.

Yao HH, Matzuk MM, Jorgez CJ, Menke DB, Page DC, Swain A, Capel B. 2004. Follistatin operates downstream of Wnt4 in mammalian ovary organogenesis. *Dev Dyn* **230:** 210–215.

Yatsu R, Miyagawa S, Kohno S, Saito S, Lowers RH, Ogino Y, Fukuta N, Katsu Y, Ohta Y, Tominaga M, et al. 2015. TRPV4 associates environmental temperature and sex determination in the American alligator. *Sci Rep* **5:** 18581.

Ying SY, Becker A, Swanson G, Tan P, Ling N, Esch F, Ueno N, Shimasaki S, Guillemin R. 1987. Follistatin specifically inhibits pituitary follicle stimulating hormone release in vitro. *Biochem Biophys Res Commun* **149:** 133–139.

Zhang Y, Liang J, Li Y, Xuan C, Wang F, Wang D, Shi L, Zhang D, Shang Y. 2010. CCCTC-binding factor acts upstream of FOXA1 and demarcates the genomic response to estrogen. *J Biol Chem* **285:** 28604–28613.

Zhou Q, Zhang J, Bachtrog D, An N, Huang Q, Jarvis ED, Gilbert MT, Zhang G. 2014. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346:** 1246338.

Received August 1, 2016; accepted in revised form December 13, 2016.

Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling

Edward S. Rice, Satomi Kohno, John St. John, et al.

Genome Res. 2017 27: 686-696 originally published online January 30, 2017
Access the most recent version at doi:10.1101/gr.213595.116