The draft genomes of soft–shell turtle and green sea turtle yield insights into the development and evolution of the turtle–specific body plan

Zhuo Wang[#1], **Juan Pascual-Anaya**[#2], **Amonida Zadissa**[#3], **Wenqi Li**[#4], **Yoshihito Niimura**[5], **Zhiyong Huang**[1], **Chunyi Li**[4], **Simon White**[3], **Zhiqiang Xiong**[1], **Dongming Fang**[1], **Bo Wang**[1], **Yao Ming**[1], **Yan Chen**[1], **Yuan Zheng**[1], **Shigehiro Kuraku**[2], **Miguel Pignatelli**[6], **Javier Herrero**[6], **Kathryn Beal**[6], **Masafumi Nozawa**[7], **Qiye Li**[1], **Juan Wang**[1], **Hongyan Zhang**[4], **Lili Yu**[1], **Shuji Shigenobu**[7], **Junyi Wang**[1], **Jiannan Liu**[4], **Paul Flicek**[6], **Steve Searle**[3], **Jun Wang**[1,8,9], **Shigeru Kuratani**[2], **Ye Yin**[4], **Bronwen Aken**[3], **Guojie Zhang**[1,10,11], and **Naoki Irie**[2]

[1]BGI-Shenzhen, Shenzhen, China.

[2]RIKEN Center for Developmental Biology, Kobe, Japan.

[3]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

[4]BGI-Japan, Kobe, Japan.

[5]Medical Research Institute, Tokyo Medical and Dental University, Tokyo, Japan.

[6]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.

[7]National Institute for Basic Biology (NIBB) Core Research Facilities, NIBB, Okazaki, Japan.

[8]Department of Biology, University of Copenhagen, Copenhagen, Denmark.

[9]King Abdulaziz University, Jeddah, Saudi Arabia.

[10]China National GeneBank, BGI-Shenzhen, Shenzhen, China.

[11]Center for Social Evolution, Department of Biology, University of Copenhagen, Copenhagen, Denmark.

# These authors contributed equally to this work.

Abstract

The unique anatomical features of turtles have raised unanswered questions about the origin of their unique body plan. We generated and analyzed draft genomes of the soft-shell turtle (*Pelodiscus sinensis*) and the green sea turtle (*Chelonia mydas*); our results indicated the close relationship of the turtles to the bird-crocodilian lineage, from which they split ~267.9–248.3 million years ago (Upper Permian to Triassic). We also found extensive expansion of olfactory receptor genes in these turtles. Embryonic gene expression analysis identified an hourglass-like divergence of turtle and chicken embryogenesis, with maximal conservation around the vertebrate phylotypic period, rather than at later stages that show the amniote-common pattern. *Wnt5a* expression was found in the growth zone of the dorsal shell, supporting the possible co-option of limb-associated Wnt signaling in the acquisition of this turtle-specific novelty. Our results suggest that turtle evolution was accompanied by an unexpectedly conservative vertebrate phylotypic period, followed by turtle-specific repatterning of development to yield the novel structure of the shell.

The unique anatomy of turtles has raised questions about their evolution[1]. Their armor, even compared to other armored tetrapods (for example, the armadillo and Indian rhinoceros), is distinct in that the dorsal part of the shell (carapace) represents transformed vertebrae and ribs. In addition, their shoulder blades or scapulae[1] display an inside–out topology against the rib cage (Supplementary Fig. 1 and Supplementary Note), and the lack of a temporal fenestra further complicates the reconstruction of their phylogenetic position[1,2].

Three major hypotheses have been proposed for the evolutionary origin of turtles, including that they (i) constitute early–diverged reptiles, called anapsids[3], (ii) are a sister group of the lizard–snake–tuatara (Lepidosauria) clade[4] or (iii) are closely related to a lineage that includes crocodilians and birds (Archosauria)[5–8]. Even using molecular approaches, inconsistency still remains[6–9]. To clarify the evolution of the turtle–specific body plan, we first addressed the question of evolutionary origin of the turtle by performing the first genome–wide phylogenetic analysis with two turtle genomes sequenced in this project (the green sea turtle, *C. mydas*, and the Chinese soft–shell turtle, *P. sinensis*; Fig. 1a). In brief, the fragmented genomic DNA libraries of the two turtles were independently shotgun sequenced using the HiSeq 2000 sequencer and assembled using the SOAPdenovo assembler (Online Methods). The generated turtle genomes were both around 2.2 Gb in size, with the N50 lengths of scaffolds longer than 3.3 Mb (Table 1, Supplementary Figs. 2–5 and Supplementary Tables 1–9).

On the basis of the largest turtle data set so far, our phylogenetic analysis, with an orthologous set of 1,113 single–copy coding genes, robustly indicated that turtles are likely to be a sister group of crocodilians and birds (Fig. 1b, Supplementary Figs. 6 and 7 and Supplementary Tables 10–13), implying that the temporal fenestrae in the turtle skull were most likely secondarily lost in the turtle line–age[1]. A molecular evolutionary clock analysis with time constraints based on the fossil records estimated that turtles diverged from archosaurians approximately 257.4 million years ago, with a 95% credibility interval between 267.9 and 248.3 million years ago (Fig. 1b, Supplementary Fig. 7 and Supplementary Table 12). These results are consistent with the oldest turtle fossil (from 220 million years ago), named *Odontochelys*[10]. The estimated time range corresponds to the Upper Permian to Triassic period (Fig. 1b), overlapping or following shortly after the Permian extinction event[11]; this raises the question of whether the emergence of the turtle group was related to this severe extinction event, which especially involved the extinction of marine species.

Taking into consideration the phylogenetic position of turtles, we next searched for genes that could potentially explain turtle–specific characteristics (Supplementary Fig. 8 and Supplementary Tables 14–23). Unexpectedly, we found that the olfactory receptor family was highly expanded in both turtle species (Fig. 2 and Supplementary Tables 14–20). In particular, the soft–shell turtle contained 1,137 intact, possibly functional olfactory receptor genes, a number comparable to or even greater than the number of olfactory receptor genes found in most mammals[12]. Olfactory receptor gene expansion was observed mainly in the α subtype of the class I olfactory receptor genes, suggesting that turtles have superior olfaction ability against a wide variety of hydrophilic substances[12] (Fig. 2a and Supplementary Tables 19 and 20). Detailed analyses with genomic sequences further clarified that the majority of the expansion occurred after the split of the two turtle species (Fig. 2b) and that the expansion was most likely facilitated by a gene duplication process, as inferred by the clustered distribution of the olfactory receptor genes in the genome (Fig. 2c,d). These results call into question the general proposition based on mammalian studies[13,14] that vertebrates that expand their niche back into aquatic environments tend to reduce the number of olfactory receptor genes. Other than olfactory receptor gene expansion, we found that many genes involved in taste perception (Supplementary Tables 21–23) were lost in the two turtle species. Furthermore, we found that the gene for the hunger–stimulating and energy homeostasis–regulating hormone ghrelin was also lost specifically in the two turtle species (Supplementary Table 23), which could be related to their low–metabolic strategies. Further investigation of the lost genes in the two turtles identified the loss of many orthologs that are known to be important for normal development in different species, including the genes encoding UNC homeobox, FGF–binding protein 3, CXCL10 and Agouti signaling protein (Supplementary Table 23). These results, together with the identification of many other genes that show accelerated evolutionary rate in turtles (Supplementary Table 24; for example, *Bmp receptor 1b*, *Kit*, *Jak1* and *Eya4*), suggest that turtle evolution has included many alterations of the signaling cascade that is presumably involved in morphogenesis. Finally, a possible connection to longevity in turtles was also found (Supplementary Table 24); the most accelerated gene in turtles, showing evidence of positive selection (with the rate of nonsynonymous nucleotide substitutions exceeding the rate of neutral mutations,

dN/dS ratio > 1), was microsomal glutathione S–transferase 3 (*Mgst3*; dN/dS = 5.68), which is reported to function in antioxidative stress, and disrupting the homolog *Mgst3*–like in *Drosophila melanogaster* reduces lifespan[15].

In addition to changes in genomic sequences, we also investigated alterations in embryonic gene regulation that occurred after the split from the bird–crocodilian lineage. According to the recently supported developmental hourglass model[16–21], the evolutionary changes underlying major adult morphological evolution occurred primarily in the developmental stages after the period of the vertebrate common plan or the period that serves as the source of the vertebrate basic body plan, namely, the vertebrate phylotypic period[22]. However, the hourglass model has not been tested in non–model organisms, particularly in those with the atypical anatomical features of turtles; therefore, we tested whether the model held true in turtle–chicken comparison. Taking advantage of RNA sequencing (RNA–seq) technology and our previously established method[21] based on hierarchical Bayes statistics, our cross–species approach comparing whole–embryo gene expression profiles (GXPs) clearly demonstrated an hourglass–like GXP divergence in the embryogenesis of the soft–shell turtle and the chicken (Fig. 3a and Supplementary Figs. 9–12). However, the result was not robust enough to suggest that the most conserved developmental stage in turtles and birds corresponds to the vertebrate phylotype. The conserved stage could be one occurring later than the vertebrate phylotype, as a previous developmental study[23] demonstrated that turtles have a typical amniote–common plan during embryogenesis and develop turtle–specific characteristics thereafter (for example, the scapula primordium first arises outside the rib cage and only later comes to lie inside the rib cage), as if the embryo is recapitulating its own evolutionary history[23]. If the most conserved developmental stage between the two species indeed corresponds to the stage of the amniote–common plan (approximately Tokita–Kuratani[24] stage (TK) 13–14), this would indicate that the conserved stage may change depending on how distantly related the species are that are being compared, similar to the idea from the nested hourglasses model[18] (Fig. 3b), justifying, in part, the hierarchical relationship between ontogeny and phylogeny once proposed by Karl von Baer[25]. Further investigation using a statistically robust cross–species comparative analysis indicated that the soft–shell turtle TK11 and the chicken HH16 developmental stages showed the most similar GXPs (Fig. 3c, Supplementary Figs. 13 and 14 and Supplementary Table 25). Considering that the chicken stage corresponds to the previously identified phylotypic period[21], turtle stage TK11 would be an attractive candidate for the vertebrate phylotypic period. In addition to the conservation between turtle and chicken at the level of gene regulation, the identified stages showed notable similarity in morphology (Fig. 3d and Supplementary Table 26), despite the large differences in their final anatomy, the size of their eggs and the actual time scale required for embryogenesis as well as the geological time scale passed since their split (approximately 230 million years ago; Fig. 1), indicating that the morphological and molecular patterns are not uncoupled, in contrast to recent implications from plant development[26]. Taken together, these results suggest that turtles indeed conform to the developmental hourglass model (Supplementary Fig. 15) by first establishing an ancient vertebrate body plan and by developing turtle–specific characteristics thereafter.

The above results suggest that turtle–specific global repatterning of gene regulation begins after TK11 or the phylotypic period. Although turtle and chicken express many shared developmental genes in the embryo during the putative phylotypic period (Fig. 4a and Supplementary Tables 27 and 28) and have the fewest expanded or contracted gene family members expressed (Supplementary Fig. 16) at this stage, later stages showed increasing differences in their molecular patterns. We found 233 genes that showed turtle–specific increasing expression patterns after the phylotype (Fig. 4b). Considering that the chicken orthologs did not show this type of increasing expression (Supplementary Figs. 17 and 18), these 233 genes represent attractive candidates for clarifying the genomic nature of turtle–specific morphological oddities. Furthermore, our Gene Ontology (GO)–based statistical analysis identified many genes that are potentially involved in ossification and extracellular matrix regulation (Fig. 4c), suggesting the involvement of morphological characteristics appearing in turtle embryogenesis, such as extensive ossification in the shell and folding of the body wall[23,24]. The morphological specifications of turtle embryogenesis after the identified phylotypic period include the formation of the novel turtle structure called the carapacial ridge[27], which is considered to be responsible for the flabellate expansion of the turtle ribs in late development[27]. Previous molecular studies[27,28] have identified many carapacial ridge–specific coding genes, whereas no study so far has investigated carapacial ridge–specific microRNA (miRNA) expression, despite the increasing number of reports claiming the crucial roles of miRNA in various developmental processes. We therefore performed a small RNA–seq analysis of three tissues from soft–shell turtle embryos—limb, body wall and carapacial ridge (Fig. 4d and Supplementary Figs. 19–21)—and further predicted possible miRNAs by referring to the genome sequence (Supplementary Table 29). Unexpectedly, we found expression of a large number of specific miRNAs in all of the tissues (Fig. 4d and Supplementary Table 30), including the carapacial ridge (212 miRNAs). Although no definitive conclusion can be made regarding the functions of these miRNAs, our preliminary prediction–based analysis implied the possible involvement of Wnt signaling (Supplementary Fig. 21 and Supplementary Tables 31–33).

Ann Burke[29] was the first to point out the similarity of the apical ectodermal ridge of limbs and the carapacial ridge of the turtle shell. Later, increasing molecular evidence supported this hypothesis. Previous studies[27,28] have shown the carapacial ridge–specific activation of Wnt downstream genes (for example, *Lef1* expression and nuclear localization of β–catenin) and the essential role of LEF1 in carapacial ridge formation[27]; however, no Wnt ligand expression has been identified. We therefore annotated all the Wnt genes in the soft–shell turtle and green sea turtle genomes, finding a total of 20 (Supplementary Table 31), and studied their expression patterns in soft–shell turtle embryos at stage TK14, the stage when the carapacial ridge begins to be apparent (Fig. 5a). Notably, we found that *Wnt5a* was the only Wnt gene expressed in the turtle carapacial ridge region (Fig. 5b,c and Supplementary Fig. 22). With respect to the evolutionary scenario of the carapacial ridge, *Wnt5a* expression was also found in both the forelimbs and the hindlimbs, as in other amniotes, implying that part of the gene regulatory network involved in carapacial ridge development has been co–opted, most likely from the limb buds[28,29]. However, this hypothesis has to be considered with caution, particularly because we still lack functional evidence of *Wnt5a* involvement in carapacial ridge formation. Taking the findings together, the exact roles of the carapacial

ridge–expressed Wnt and miRNAs remain to be elucidated; however, our series of genome–scale results indicate the co–option of the Wnt signaling pathway in turtles and provide a basis for understanding shell evolution.

In summary, our study both highlights the evolution of the turtle body plan and offers a model to explain, at the genomic level, how the vertebrate developmental program can change to produce major evolutionary novelties in morphological phenotypes.

ONLINE METHODS

Source and sequencing of genomic DNA and error correction

The soft–shell turtle was purchased from a local farmer in Japan, and the green sea turtle was provided by the Genome 10K Project (originally collected in Ocean Park, Hong Kong). Genomic DNA was extracted from the whole blood of a female individual in each species, and we constructed a total of 18 (for the soft–shell turtle) and 17 (for the green sea turtle) libraries consisting of short–insert (170–bp, 500–bp and 800–bp) and long–insert (2–kb, 5–kb, 10–kb, 20–kb and 40–kb) libraries. Sequencing was performed using the Illumina HiSeq 2000 system, and read error correction was performed for the short–insert libraries (on the basis of the *K*–mer frequency distribution curve; Supplementary Note). Data accession numbers are given in Supplementary Table 34.

Genome assembly

Filtered and corrected data were assembled using SOAPdenovo[31,32]. We first generated contigs by constructing a de Bruijn graph with the reads from the *K*–mer–split short–insert library data. The graph was then simplified to generate the contigs by removing tips, merging bubbles and solving repeats. All sequenced reads were then realigned onto the contig sequences, and scaffolds were constructed by weighting the rates of consistent and conflicting paired–end relationships. Finally, we retrieved the read pairs with one end that uniquely mapped to the contig and the other end located in the gap region, and performed a local assembly for these collected reads to fill the gaps.

Repeat annotation and whole-genome alignment

Repeat detection was performed using the program RepeatMasker and the Genetic Information Research Institute (GIRI) repeat library. For homology–based prediction of repeats, we used the library of known repeats in the Repbase[33] database (v2008–08–01, Repbase–16.02) with RepeatMasker (v3.2.6) and RepeatProteinMask to identify transposable elements at the DNA and protein levels, respectively. The *de novo* prediction of repeats involved building a *de novo* repeat library with RepeatModeler[34] and subsequently employing RepeatMasker. Tandem repeats were searched with the Tandem Repeats Finder (TRF)[35]. Whole–genome pairwise alignments were generated by LASTZ[36].

Gene prediction for the two turtles and crocodilians

Gene prediction for the two turtle genomes employed both the *ab initio* approach (GENSCAN[37] (v2.5.5) and AUGUSTUS[38] (v1.0)) and a homolog–based approach against the repeat–masked genome, and gene sets predicted by these two approaches were further

consolidated with the GLEAN[39] program. For the soft–shell turtle, an additional 146.7 Gb of RNA–seq data was used. The proteins of other vertebrate species were mapped to the genome using TBLASTN (Legacy Blast[40] v2.2.23). Aligned sequences were then filtered and passed to GeneWise[41] (v2.2.0) along with the query sequences. The resulting data sets were integrated by GLEAN[39] into a consensus gene set. The best BLASTP match to the SwissProt and TrEMBL databases was used to assign function. The motifs and domains of the gene products were annotated with InterProScan[42] against the protein databases ProDom, PRINTS, Pfam, SMART, PANTHER and PROSITE. Gene Ontology[43] IDs for each gene were obtained from the corresponding InterPro entries. The above prediction pipeline was applied to the saltwater crocodile and American alligator genomes (from the Crocodile Genome Consortium), except for the integration step in the latter case. Gene family identification was performed using TreeFam[32].

Gene prediction for the soft-shell turtle by the Ensembl prediction pipeline

For gene expression comparison analyses between soft–shell turtle and chicken embryos, we generated and used another soft–shell turtle gene set that was created by the same Ensembl pipeline as the chicken gene set (see URLs).

GO analysis

Over–represented GO terms were investigated by testing (Fisher's exact test) the bias in frequency toward other GO terms among certain gene sets, using the total set of defined GO terms as a control distribution. Developmental genes (5,659 in total) were defined as genes with developmental GO terms, and developmental GO terms were defined as those with GO:0032502 (developmental process) as an ancestor.

Animal care and use

Experimental procedures and animal care were conducted in strict accordance with guidelines approved by the RIKEN Animal Experiments Committee (Approval IDs H14–23 and H16–10).

Phylogenetic tree reconstruction and divergence time estimation

The coding sequences of single–copy gene families conserved among the soft–shell turtle, green sea turtle, anole lizard, saltwater crocodile, chicken, zebra finch, dog, human, platypus and *Xenopus tropicalis* were extracted and aligned with guidance from amino–acid alignments created by the MUSCLE program[44]. Sequences were then concatenated to one supergene sequence for each species. PhyML[45,46] was applied to construct the phylogenetic tree under an HKY85+gamma or GTR+gamma model for nucleotide sequences and the JTT +gamma model for protein sequences. aLRT values were taken to assess the branch reliability in PhyML. RAxML[47] was also applied for the same set of sequences to build a phylogenetic tree under a GTR+gamma or JTT+gamma model for nucleotide and protein sequences, respectively, with 1,000 rapid bootstraps employed to assess the branch reliability in RAxML. The same set of codon sequences at positions 1 and 2 was used for phylogenetic tree construction and estimation of the divergence time. The PAML *mcmctree* program (PAML version 4.5)[48–50] was used to determine divergence times with the

approximate likelihood calculation method and the 'correlated molecular clock' and 'REV' substitution model. Two independent runs were performed to confirm convergence.

Gene loss analysis and gene family expansion and contraction analysis

Protein sequences of the two turtles and related species (chicken, anole lizard, *X. tropicalis* and zebra finch) were used in BLAST searches against human protein sequences (Ensembl Gene v.68), identifying homologs. Subsequently, human proteins that lacked homologs in both the turtle species but had homologs in the related species were identified as lost genes in turtle. For the statistical analysis of gene family expansion and contractions, we generated pairwise whole–genome alignments for anole lizard and soft–shell turtle and for anole lizard and green sea turtle using LASTZ[51] and created three–way alignments using MULTIZ[52]. When an anole lizard gene fell in an area of conserved sequence and there was no homologous gene in the corresponding aligned sequences of the two turtle species, we hypothesized that gene loss potentially occurred at that locus in turtle (Supplementary Note). Frameshift mutations and those introducing premature stop codons in the coding sequences were also considered to represent gene loss.

Prediction of olfactory receptor genes

Olfactory receptor genes were identified by previously described methods[53], with the exception of a first–round TBLASTN[54] search, in which 119 functional olfactory receptor genes from human, mouse and zebrafish were used as queries (Supplementary Note). To construct phylogenetic trees, the amino–acid sequences encoded by olfactory receptor genes were first aligned using the program E–INS–i in MAFFT[55]. We then constructed a phylogenetic tree using the neighbor–joining method[56] with Poisson correction distances using the program LINTREE[57]. The numbers of olfactory receptor genes in ancestral species and those of gene gains or losses in evolution were calculated by the reconciled tree method[53] with 70% bootstrap value cutoff.

Genes with accelerated evolutionary rate in the turtle lineage

Homologous genes in soft–shell turtle, green sea turtle and other vertebrate species (chicken, zebra finch, anole lizard, *X. tropicalis* and platypus) were first identified with the all–against–all BLASTP program. Orthologs were defined by reciprocal best BLAST hits (RBBHs) in humans and the other species. The full orthologous gene sets were aligned using the program MUSCLE. We then compared a series of evolutionary models within the likelihood framework using the phylogenetic tree obtained by our analysis. A branch model[50] was used to detect the average length (ω) across the tree (ω_0), the ω value of the ancestor of all soft–shell turtles, the ω value for the green sea turtle branch (ω_2) and the ω value for all of the other branches (ω_1).

Embryo sampling and mRNA extraction

Fertilized soft–shell turtle and chicken eggs purchased from local farms in Japan were incubated and staged according to previous descriptions[24,58]. Amniotic membranes were removed before mRNA extraction, and more than two individual embryos were pooled for

each sample. The RNeasy Lipid Tissue kit (Qiagen) and the Ambion MicroPoly(A) Purist kit (Life Technologies) were used for mRNA extraction.

RNA-seq for transcriptome identification

Three different types of sequencing were performed for transcriptome identification in the soft–shell turtle: (i) Titanium sequencing (about 2 Gb of clean sequence data), (ii) HiSeq strand–specific paired–end RNA–seq (two libraries were prepared by methods that retain strand–specific information, including a dUTP–based method[59,60] (19 Gb of clean data) that was modified to comply with the Illumina TruSeq RNA sample prep kit and an original method developed at BGI Sequencing that was performed with Illumina HiSeq 2000 (26 Gb of clean data) and (iii) HiSeq non–stranded RNA–seq (deep–sequencing data for gene expression analysis was also used for transcriptome identification). Further details are given in the Supplementary Note.

RNA-seq for gene expression analysis and expression comparison

Biological replicates for each developmental stage were created from an independent sample pool. Extracted mRNA samples were then sequenced with an Illumina HiSeq 2000 instrument. We identified 11,602 one–to–one orthologous genes in the soft–shell turtle and chicken using RBBH information from BLAST+ (v2.2.25)[61]. Gene expression scores were obtained from RNA–seq data by mapping clean reads to the genome using Burrows–Wheeler Aligner (BWA)[62] software (v0.5.9–r16). SAMtools[63], BEDtools[64] and the DEGseq package[65] for R (v2.14.2) were used to calculate the tag count data that were mapped to the coding regions. Normalization of the orthologous gene expression scores was performed with all samples at once by either RPKM or TMM normalization[66]. Pearson's correlation coefficients, Spearman correlation coefficients, total Euclidean distances (t–Euclidean) or total Manhattan distances (t–Manhattan) were used to estimate similarities in the gene expression profiles of the two samples being compared. Two independent random selections from all reads were performed to make the mapped–10M reads (sequencing depth–controlled data set based on randomly selected 10M tags mapped to the genome) data. The Welch two–sample t test or the Wilcoxon signed–rank test was used to detect the most conserved stages. The Holm–corrected a level was applied for these multiple comparisons. Only results reproduced by the data set from two different normalizations (RPKM[67] and TMM[66]) were considered to be significant.

Genes with a significant increase in expression levels after the phylotypic period

Turtle IAP genes were selected using the following criteria: (i) the mean expression level after the phylotypic period (TK15–TK23) was more than five times higher (Wilcoxon test) than during earlier stages (gastrula, neurula, TK7 and TK9) and (ii) the chicken orthologs of the turtle IAP genes (if any) did not show such increases in chicken (the average expression levels in HH28 and HH38 did not show more than five times higher expression than in the Prim–HH14 stages).

Wnt gene identification and cloning and whole-mount ISH

In addition to constructing the predicted gene sets, we manually searched for Wnt genes using TBLASTN. Cloning of the probes and whole–mount ISH were performed using standard methods[28] (Supplementary Note).

miRNA extraction, prediction and expression analysis

Small RNA was extracted from dissected tissues using the mirVana microRNA Isolation kit (Life Technologies). Small RNA libraries were prepared and sequenced using an Illumina HiSeq 2000 (>24 million reads per sample). These small RNA reads, together with the miRNA sequences from chicken, zebra finch and *Anolis carolinensis* from miRBase (v.18), were used to predict miRNA sequences in the genome. The program miRDeep2 (v2.0.0.3)[68] was used to predict miRNAs for this prediction. Only miRNA predictions that had *P* value lower than a significant Randfold α level ($P < 0.05$ mononucleotide shuffling and 999 permutations; see ref. 68 for details) were taken into account for subsequent comparisons. miRNA target prediction was performed with miRanda[69] (v3.3a) using the annotated 3′ UTRs of soft–shell turtle genes.

Statistical tests

To avoid an inflated type I error rate, an α level of 0.01 (further Bonferronni correction in case of multiple comparisons) was accepted for statistical significance throughout the analyses unless otherwise specified. Statistical methods were carefully chosen to properly reflect the population of interest. The Welch two–sample *t* test was used for two–sample comparisons when the data passed the Kolmogorov–Smirnov test for normal distribution; otherwise, the Wilcoxon signed–rank test was used.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Kuratani S, Kuraku S, Nagashima H. Evolutionary developmental perspective for the origin of turtles: the folding theory for the shell based on the developmental nature of the carapacial ridge. Evol. Dev. 2011; 13:1–14. [PubMed: 21210938]

2. Rieppel O. Turtles as hopeful monsters. Bioessays. 2001; 23:987–991. [PubMed: 11746214]

3. Romer, AS. Vertebrate Paleontology. 3rd edn. University of Chicago Press; Chicago: 1966.

4. Rieppel O, deBraga M. Turtles as diapsid reptiles. Nature. 1996; 384:453–455.

5. Hedges SB, Moberg KD, Maxson LR. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 1990; 7:607–633. [PubMed: 2283953]

6. Crawford NG, et al. More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. Biol. Lett. 2012; 8:783–786. [PubMed: 22593086]

7. Chiari Y, Cahais V, Galtier N, Delsuc F. Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). BMC Biol. 2012; 10:65. [PubMed: 22839781]

8. 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. [PubMed: 21943375]

9. Lyson TR, et al. MicroRNAs support a turtle + lizard clade. Biol. Lett. 2012; 8:104–107. [PubMed: 21775315]

10. Li C, Wu XC, Rieppel O, Wang LT, Zhao LJ. An ancestral turtle from the Late Triassic of southwestern China. Nature. 2008; 456:497–501. [PubMed: 19037315]

11. Zhong-Qiang C, Benton MJ. The timing and pattern of biotic recovery following the end-Permian mass extinction. Nat. Geosci. 2012; 5:375–383.

12. Niimura Y. Olfactory receptor multigene family in vertebrates: from the viewpoint of evolutionary genomics. Curr. Genomics. 2012; 13:103–114. [PubMed: 23024602]

13. Hayden S, et al. Ecological adaptation determines functional mammalian olfactory subgenomes. Genome Res. 2010; 20:1–9. [PubMed: 19952139]

14. Kishida T, Kubota S, Shirayama Y, Fukami H. The olfactory receptor gene repertoires in secondary-adapted marine vertebrates: evidence for reduction of the functional proportions in cetaceans. Biol. Lett. 2007; 3:428–430. [PubMed: 17535789]

15. Toba G, Aigaki T. Disruption of the microsomal glutathione S-transferase–like gene reduces life span of *Drosophila melanogaster*. Gene. 2000; 253:179–187. [PubMed: 10940555]

16. Duboule D. Temporal colinearity and the phylotypic progression: a basis for the stability of a vertebrate bauplan and the evolution of morphologies through heterochrony. Dev. Suppl. 1994:135–142. [PubMed: 7579514]

17. Raff, A. The Shape of Life: Genes, Development, and The Evolution of Animal Form. University of Chicago Press; Chicago: 1996.

18. Irie N, Sehara-Fujisawa A. The vertebrate phylotypic stage and an early bilaterian-related stage in mouse embryogenesis defined by genomic information. BMC Biol. 2007; 5:1. [PubMed: 17222327]

19. Kalinka AT, et al. Gene expression divergence recapitulates the developmental hourglass model. Nature. 2010; 468:811–814. [PubMed: 21150996]

20. Domazet-Lošo T, Tautz D. A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature. 2010; 468:815–818. [PubMed: 21150997]

21. Irie N, Kuratani S. Comparative transcriptome analysis reveals vertebrate phylotypic period during organogenesis. Nature Commun. 2011; 2:248. [PubMed: 21427719]

22. Richardson MK, Minelli A, Coates M, Hanken J. Phylotypic stage theory. Trends Ecol. Evol. 1998; 13:158. [PubMed: 21238240]

23. Nagashima H, et al. Evolution of the turtle body plan by the folding and creation of new muscle connections. Science. 2009; 325:193–196. [PubMed: 19590000]

24. Tokita M, Kuratani S. Normal embryonic stages of the Chinese softshelled turtle *Pelodiscus sinensis* (Trionychidae). Zoolog. Sci. 2001; 18:705–715.

25. von Baer, KE. Uber Entwickelungsgeschichte der Thiere: Beobachtung und Reflektion. Gebrüder Borntraeger; Koenigsberg: 1828.

26. Quint M, et al. A transcriptomic hourglass in plant embryogenesis. Nature. 2012; 490:98–101. [PubMed: 22951968]

27. Nagashima H, et al. On the carapacial ridge in turtle embryos: its developmental origin, function and the chelonian body plan. Development. 2007; 134:2219–2226. [PubMed: 17507399]

28. Kuraku S, Usuda R, Kuratani S. Comprehensive survey of carapacial ridge–specific genes in turtle implies co-option of some regulatory genes in carapace evolution. Evol. Dev. 2005; 7:3–17. [PubMed: 15642085]

29. Burke A. Development of the turtle carapace: implications for the evolution of a novel bauplan. J. Morphol. 1989; 199:363–378.

30. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006; 22:2971–2972. [PubMed: 17021158]

31. Li R, et al. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. 2010; 20:265–272. [PubMed: 20019144]

32. Li R, et al. The sequence and *de novo* assembly of the giant panda genome. Nature. 2010; 463:311–317. [PubMed: 20010809]

33. Jurka J, et al. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet. Genome Res. 2005; 110:462–467. [PubMed: 16093699]

34. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinformatics. 2005; 21(suppl. 1):i351–i358. [PubMed: 15961478]

35. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27:573–580. [PubMed: 9862982]

36. Brudno M, et al. Automated whole-genome multiple alignment of rat, mouse, and human. Genome Res. 2004; 14:685–692. [PubMed: 15060011]

37. Burge CB, Karlin S. Finding the genes in genomic DNA. Curr. Opin. Struct. Biol. 1998; 8:346–354. [PubMed: 9666331]

38. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. 2011; 27:757–763. [PubMed: 21216780]

39. Elsik CG, et al. Creating a honey bee consensus gene set. Genome Biol. 2007; 8:R13. [PubMed: 17241472]

40. Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002; 12:656–664. [PubMed: 11932250]

41. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004; 14:988–995. [PubMed: 15123596]

42. Zdobnov EM, Apweiler R. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics. 2001; 17:847–848. [PubMed: 11590104]

43. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet. 2000; 25:25–29. [PubMed: 10802651]

44. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32:1792–1797. [PubMed: 15034147]

45. Guindon S, et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 2010; 59:307–321. [PubMed: 20525638]

46. Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 2003; 52:696–704. [PubMed: 14530136]

47. Endres CS, Putman NF, Lohmann KJ. Perception of airborne odors by loggerhead sea turtles. J. Exp. Biol. 2009; 212:3823–3827. [PubMed: 19915124]

48. Rannala B, Yang Z. Inferring speciation times under an episodic molecular clock. Syst. Biol. 2007; 56:453–466. [PubMed: 17558967]

49. Yang Z, Rannala B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 2006; 23:212–226. [PubMed: 16177230]

50. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 2007; 24:1586–1591. [PubMed: 17483113]

51. Harris, R. *Improved Pairwise Alignment of Genomic DNA* PhD Thesis. Penn. State Univ.; 2007.

52. Blanchette M, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004; 14:708–715. [PubMed: 15060014]

53. Niimura Y, Nei M. Extensive gains and losses of olfactory receptor genes in mammalian evolution. PLoS ONE. 2007; 2:e708. [PubMed: 17684554]

54. Altschul SF, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997; 25:3389–3402. [PubMed: 9254694]

55. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005; 33:511–518. [PubMed: 15661851]

56. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 1987; 4:406–425. [PubMed: 3447015]

57. Takezaki N, Rzhetsky A, Nei M. Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. 1995; 12:823–833. [PubMed: 7476128]

58. Hamburger V, Hamilton HL. A series of normal stages in the development of the chick embryo. 1951. Dev. Dyn. 1992; 195:231–272. [PubMed: 1304821]

59. Levin JZ, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat. Methods. 2010; 7:709–715. [PubMed: 20711195]

60. Parkhomchuk D, et al. Transcriptome analysis by strand-specific sequencing of complementary DNA. Nucleic Acids Res. 2009; 37:e123. [PubMed: 19620212]

61. Camacho C, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009; 10:421. [PubMed: 20003500]

62. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–1760. [PubMed: 19451168]

63. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–2079. [PubMed: 19505943]

64. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26:841–842. [PubMed: 20110278]

65. Wang L, Feng Z, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2010; 26:136–138. [PubMed: 19855105]

66. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010; 11:R25. [PubMed: 20196867]

67. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat. Methods. 2008; 5:621–628. [PubMed: 18516045]

68. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. Nucleic Acids Res. 2012; 40:37–52. [PubMed: 21911355]

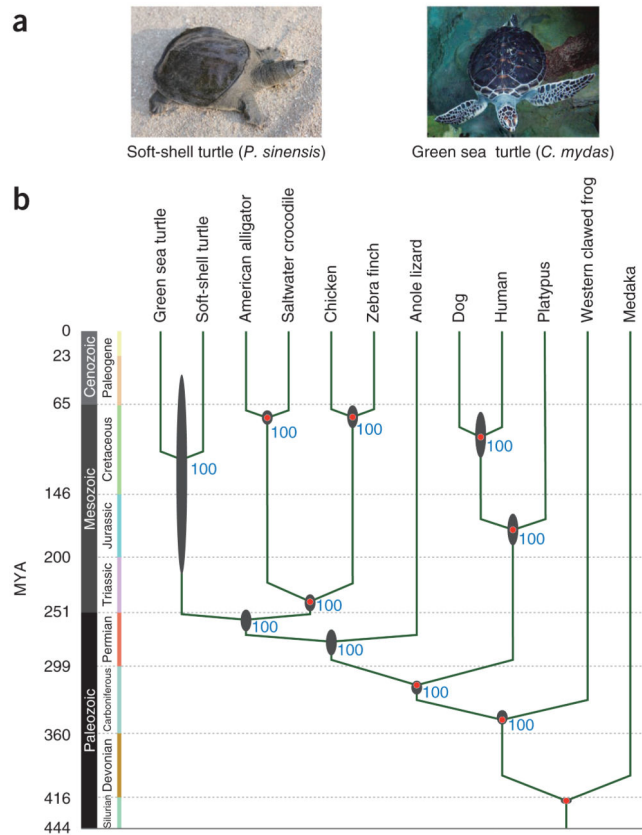69. Enright AJ, et al. MicroRNA targets in *Drosophila*. Genome Biol. 2003; 5:R1. [PubMed: 14709173]

Figure 1.

Turtle phylogeny and divergence time estimation by molecular clock analysis. (**a**) Two genome-sequenced turtles, the soft-shell turtle (*P. sinensis*) and the green sea turtle (*C. mydas*). (**b**) Estimated divergence times of 12 vertebrate species calculated using the first and second codon positions of 1,113 single-copy coding genes (Supplementary Tables 9 and 10). Tree topology is supported by 100% bootstrap values and further statistical assessment (Supplementary Fig. 5 and Supplementary Tables 11–13). The black ellipses on the nodes indicate the 95% credibility intervals of the estimated posterior distributions of the divergence times. The red circles indicate the fossil calibration times used for setting the upper and lower bounds of the estimates. MYA, million years ago.
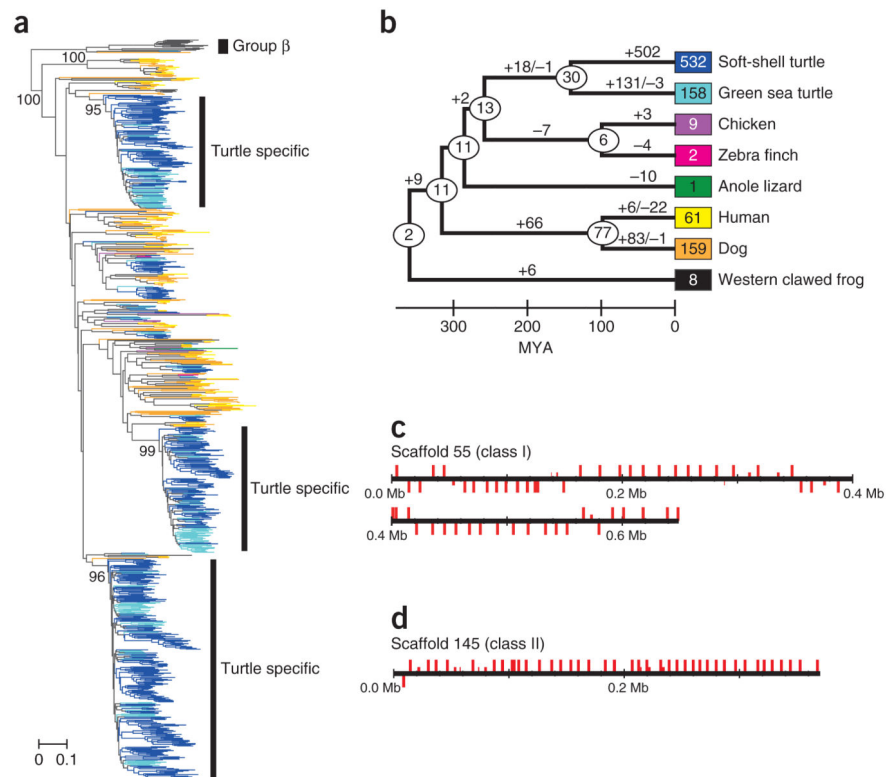
Figure 2.

Extensive expansion of olfactory receptor genes in turtles. (**a**) A neighbor-joining tree constructed with all the intact group α olfactory receptors from eight vertebrate species (soft-shell turtle, green sea turtle, chicken, zebra finch, anole lizard, human, dog and Western clawed frog), with the group β olfactory receptors as the outgroup. Bootstrap values (from 500 resamplings) are shown on the branches. The scale bar represents the number of amino-acid substitutions per site. (**b**) Expansion of group α olfactory receptor genes in the evolution of tetrapods. Numbers in boxes indicate the current number of intact group α olfactory receptor genes in each species. The number of group α olfactory receptor genes in an ancestral species is shown in an ellipse at each node, and the numbers of gene gains and losses are shown on each branch with plus and minus signs, respectively. For divergence times, we used the median values obtained from TimeTree[30]. Note that the majority of the expansion of the group α olfactory receptor genes occurred independently in each turtle lineage. The same color code for species is used in **a,b**. (**c,d**) Genomic clusters of olfactory receptor genes in scaffolds 55 and 145 of the soft-shell turtle genome. Vertical red bars represent class I (**c**) and class II (**d**) olfactory receptor genes. Bars above and below the horizontal line indicate opposite directions of transcription. Long bars depict intact olfactory receptor genes, whereas short bars depict olfactory receptor pseudogenes or gene fragments.
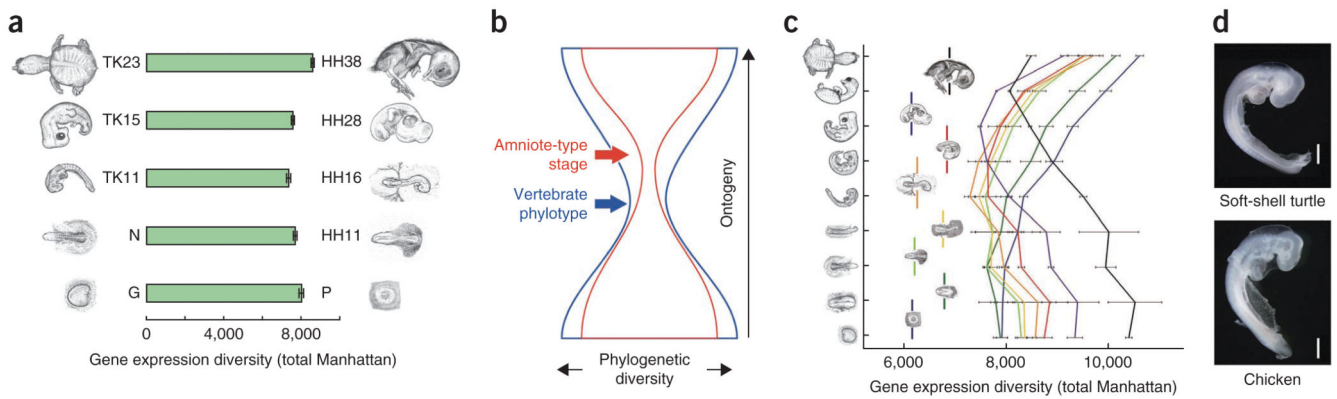
Figure 3.

The molecular divergence of turtle and chicken embryos follows the hourglass model with a maximally conserved vertebrate phylotypic period. (**a**) Distances of whole-embryo GXPs (from depth-controlled, TMM (trimmed mean of M values)-normalized data) for 11,602 orthologs in selected developmental stages of the soft-shell turtle and chicken (Supplementary Fig. 16). Error bars, s.d. ANOVA P value under heteroscedasticity $= 7 \times 10^{-7}$. (**b**) The hypothetical model (nested hourglass)[18] in which both an hourglass-like divergence and a recapitulation-like relationship between ontogeny and phylogeny can be justified. The model infers that the most conserved developmental stage changes depending on how distantly related the species are that are being compared. Comparisons within vertebrate embryos gives a vertebrate phylotype (blue arrow), and comparisons within amniotes gives an amniote-type stage (red arrow) that emerges later than the vertebrate phylotype stage. (**c**) An all-to-all comparison of turtle and chicken GXP distances (total Manhattan) indicates that the highest similarity occurs between turtle stage TK11 and chicken stage HH16 embryos (see Supplementary Fig. 18 for statistical assessment). HH16 is the stage previously identified as the vertebrate phylotypic period[21], which does not coincide with the model in **b**. Error bars, s.d. (**d**) Morphological appearance of the soft-shell turtle (stage TK11) and chicken (stage HH16) embryos that showed the highest GXP similarity (Supplementary Table 25). Scale bars, 1 mm.
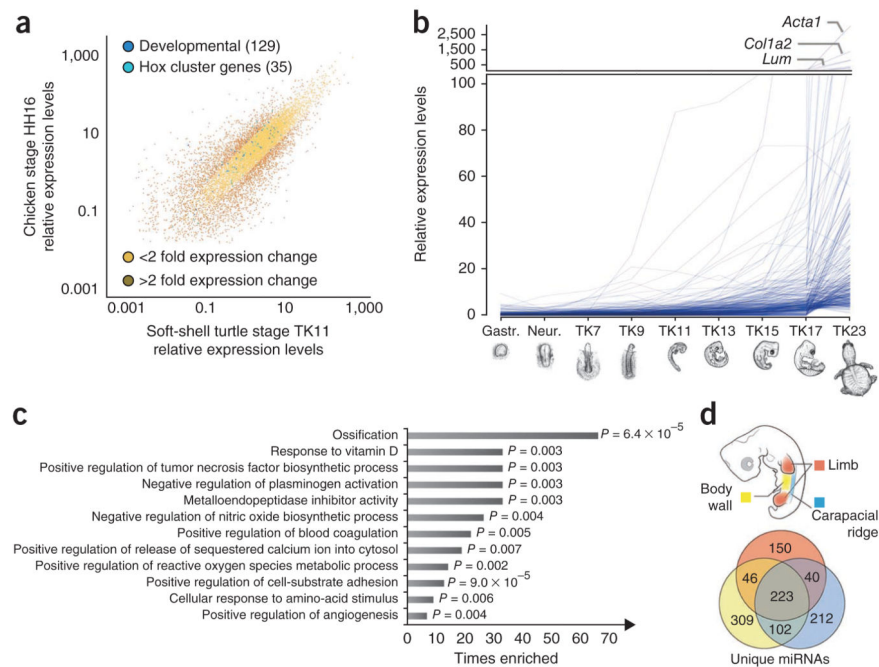
Figure 4.

Molecular characteristics of turtle embryogenesis during and after the phylotypic period. (**a**) Shared expression of developmental genes (Supplementary Table 27) in the phylotypic stages of turtle and chicken embryos. The $\log_{10}$-transformed relative expression levels of 11,602 orthologous genes from mapped-10M reads (data set based on randomly selected 10M tags mapped to the genome; Online Methods), with TMM-normalized data, were graphed on a scatterplot. Essentially the same results were obtained from other data sets (all-read-data, RPKM (reads per kilobase per million mapped reads) and TMM normalizations). The results of a statistical test to determine the groups of genes that have more similar expression can be found in the Supplementary Note. (**b**) Genes that showed a statistically significant increase in their expression level after the phylotypic period (IAP; Online Methods). Each line represents the mean expression level of each IAP (increased expression after the phylotype) gene calculated, with two biological replications, for each stage. The names of the genes with the top three highest expression levels in TK23 are shown. Consequently, 233 turtle IAP genes were found. See Supplementary Figure 18 for the expression pattern of the chicken orthologs of the turtle IAP genes. (**c**) Over-represented GO annotations for 233 turtle IAP genes with read depth–controlled, TMM-normalized data. Only the results corroborated by all of the data sets (mapped-10M reads (Online Methods), all reads, RPKM normalization and TMM normalization) are shown. Shown are *P* values calculated by Fisher's exact test. (**d**) High numbers of tissue-specific miRNAs were identified (also in the carapacial ridge) in the embryo after the phylotypic period.
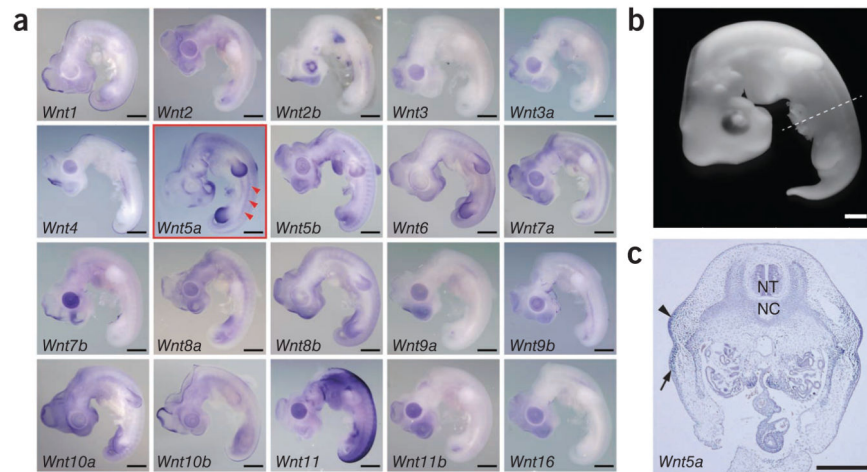
Figure 5.
Expression profiling of all 20 soft-shell turtle Wnt genes shows *Wnt5a* expression in the carapacial ridge. (**a**) Whole-mount *in situ* hybridization (ISH) was performed for all Wnt genes in the genome. *Wnt5a* (red outline) is specifically expressed in the carapacial ridge (red arrowheads), whereas most of the other genes show similar expression patterns to their known mouse and chicken counterparts. Scale bars, 0.5 mm. (**b**) Soft-shell turtle embryo at stage TK14. (**c**) Carapacial ridge expression of *Wnt5a* confirmed by ISH on a 6-μm paraffin transverse section (the sectioned level for ISH is indicated by the dashed line in **b**). The arrowhead indicates the carapacial ridge; the arrow indicates the body wall. NT, neural tube; NC, notochord. Scale bars in **b** and **c**, 0.5 mm.

Table 1

Basic statistics of two turtle genomes

	Soft-shell turtle	Green sea turtle
Estimated genome size	2.21 Gb	2.24 Gb
Sequencing depth	105.6	82.3
N50 scaffold	3.33 Mb	3.78 Mb
GC content	44.4%	43.5%
Number of coding genes	19,327	19,633