**Genome Biology**

Genome Biology



# The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage

*Genome Biology* 2013, **14**:R28 doi:10.1186/gb-2013-14-3-r28

John Abramyan (jabramyan@gmail.com)
Daleen Badenhorst (daleen.badenhorst@gmail.com)
Kyle K Biggar (k.k.biggar@gmail.com)
Glen M Borchert (gmborch@ilstu.edu)
Christopher W Botka (Christopher_Botka@hms.harvard.edu)
Rachel M Bowden (rmbowde@ilstu.edu)
Edward L Braun (ebraun68@ufl.edu)
Anne M Bronikowski (abroniko@iastate.edu)
Benoit G Bruneau (bbruneau@gladstone.ucsf.edu)
Leslie T Buck (les.buck@utoronto.ca)
Blanche Capel (b.capel@cellbio.duke.edu)
Todd A Castoe (todd.castoe@uta.edu)
Mike Czerwinski (michael.czerwinski@duke.edu)
Kim D Delehaunty (kimdelehaunty@gmail.com)
Scott V Edwards (sedwards@fas.harvard.edu)
Catrina C Fronick (cstrowma@watson.wustl.edu)
Matthew K Fujita (mkfujita@uta.edu)
Lucinda Fulton (lfulton@watson.wustl.edu)
Tina A Graves (tgraves@watson.wustl.edu)
Richard E Green (ed@soe.ucsc.edu)
Wilfried Haerty (wilfried.haerty@dpag.ox.ac.uk)
Ramkumar Hariharan (rmhariharan@rgcb.res.in)
LaDeana H Hillier (lhillier@watson.wustl.edu)
Alisha K Holloway (Alisha.holloway@gladstone.ucsf.edu)
Daniel Janes (DANIEL.JANES@NIH.GOV)
Fredric J Janzen (fjanzen@iastate.edu)
Cyriac Kandoth (ckandoth@wustl.edu)
Lesheng Kong (lesheng.kong@dpag.ox.ac.uk)
Jason de Koning (jason.de.koning@gmail.com)
Yang Li (yang.li@dpag.ox.ac.uk)
Robert Literman (literman@iastate.edu)
Elaine R Mardis (emardis@watson.wustl.edu)
Suzanne E McGaugh (Suzanne.mcgaugh@duke.edu)
Patrick Minx (pminx@watson.wustl.edu)
Lindsey Mork (lindsey.mork@usc.edu)
Michelle O’Laughlin (mharriso@watson.wustl.edu)
Ryan T Paitz (rpaitz@ilst.edu)
David D Pollock (David.Pollock@ucdenver.edu)
Chris P Ponting (chris.ponting@dpag.ox.ac.uk)

**Genome Biology**

Genome Biology

Srihari Radhakrishnan (srihari@iastate.edu)
Brian J Raney (braney@soe.ucsc.edu)
Joy M Richman (richman@dentistry.ubc.ca)
John St. John (jstjohn@soe.ucsc.edu)
Tonia Schwartz (schwartz@gmail.com)
Arun Sethuraman (arunseth@iastate.edu)
Bradley Shaffer (brad.shaffer@ucla.edu)
Andrew M Shedlock (shedlockam@cofc.edu)
Phillip Q Spinks (pqspinks@ucla.edu)
Kenneth b Storey (Kenneth_Storey@carleton.ca)
Nay Thane (naythane@yahoo.com)
Robert C Thomson (thomsonr@hawaii.edu)
Nicole Valenzuela (nvalenzu@iastate.edu)
Tomas Vinar (tomas.vinar@fmph.uniba.sk)
Daniel E Warren (dwarren4@slu.edu)
Wesley C Warren (wwarren@watson.wustl.edu)
Richard K Wilson (rwilson@watson.wustl.edu)
Laura m Zimmerman (lmzimme@ilstu.edu)
Omar Hernandez (omarherpad@gmail.com)
Chris T Amemiya (camemiya@benaroyaresearch.org)

**ISSN** 1465-6906

**Article type** Research

**Submission date** 18 October 2012

**Acceptance date** 19 March 2013

**Publication date** 28 March 2013

**Article URL** http://genomebiology.com/2013/14/3/R28

# The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage

H Bradley Shaffer[1,2], Patrick Minx[3], Daniel E Warren[4], Andrew M Shedlock[5,6], Robert C Thomson[7], Nicole Valenzuela[8], John Abramyan[9], Chris T Amemiya[10], Daleen Badenhorst[8], Kyle K Biggar[11], Glen M Borchert[12,13*], Christopher W Botka[14], Rachel M Bowden[12], Edward L Braun[15], Anne M Bronikowski[8], Benoit G Bruneau[16,17], Leslie T Buck[18], Blanche Capel[19], Todd A Castoe[20,21*], Mike Czerwinski[19], Kim D Delehaunty[3], Scott V Edwards[22], Catrina C Fronick[3], Matthew K Fujita[21*,23], Lucinda Fulton[3], Tina A Graves[3], Richard E Green[24], Wilfried Haerty[25], Ramkumar Hariharan[26], Omar Hernandez[27], LaDeana W Hillier[3], Alisha K Holloway[16], Daniel Janes[8], Fredric J Janzen[8], Cyriac Kandoth[3], Lesheng Kong[25], A P Jason de Koning[20], Yang Li[25], Robert Literman[8], Suzanne E McGaugh[28], Lindsey Mork[19], Michelle O'Laughlin[3], Ryan T Paitz[12], David D Pollock[20], Chris P Ponting[25], Srihari Radhakrishnan[8,29], Brian J Raney[30], Joy M Richman[9], John St. John[24], Tonia Schwartz[8,29], Arun Sethuraman[8,29], Phillip Q Spinks[1,2], Kenneth B Storey[11], Nay Thane[3], Tomas Vinar[31], Laura M Zimmerman[12], Wesley C Warren[3], Elaine R Mardis[3], and Richard K Wilson[3]

*current address

1. Department of Ecology and Evolutionary Biology, University of California, Los Angeles, Los Angeles, CA 90095-1606, USA

2. La Kretz Center for California Conservation Science, Institute of the Environment and

Sustainability, University of California, Los Angeles, Los Angeles, CA 90095-1496, USA

3. The Genome Institute, Washington University School of Medicine, Campus Box 8501, 4444 Forest Park Avenue, St Louis, MO 63108, USA

4. Department of Biology, Saint Louis University, Saint Louis, MO 63103, USA

5. College of Charleston Biology Department and Grice Marine Laboratory, Charleston, SC 29424, USA

6. Medical University of South Carolina College of Graduate Studies and Center for Marine Biomedicine and Environmental Sciences, Charleston, SC 29412, USA

7. Department of Biology, University of Hawaii at Manoa, Honolulu, HI 96822, USA

8. Department of Ecology, Evolution, and Organismal Biology, Iowa State University, Ames IA 50011, USA

9. Faculty of Dentistry, Life Sciences Institute University of British Columbia, Vancouver BC, Canada

10. Benaroya Research Institute at Virginia Mason, 1201 Ninth Avenue, Seattle, WA 98101 USA

11. Department of Biology and Institute of Biochemistry, Carleton University, 1125 Colonel By Drive, Ottawa, ON, Canada K1S 5B6, Canada

12. School of Biological Sciences, Illinois State University, Campus Box 4120, Normal, IL 61790, USA

13. Department of Biological Sciences, Life Sciences Building, Rm 124, University of South Alabama, Mobile, AL 36688-0002, USA

14. Research Computing, Harvard Medical School, 25 Shattuck St., Boston, MA 02115, USA

15. Department of Biology, University of Florida, Gainesville, FL 32611 USA

16. Gladstone Institute of Cardiovascular Disease, San Francisco, CA 94158, USA

17. Cardiovascular Research Institute and Department of Pediatrics, University of California, San Francisco, San Francisco, CA 94158, USA

18. Department of Cell and Systems Biology, University of Toronto, Toronto, ON, Canada M5S 3G5, Canada

19. Department of Cell Biology, Duke University Medical Center, Durham, NC 27710, USA

20. Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

21. Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA

22. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA

23. Museum of Comparative Zoology and Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA

24. Baskin School of Engineering University of California, Santa Cruz Santa Cruz, CA 95064, USA

25. MRC Functional Genomics Unit, Department of Physiology, Anatomy and Genetics, Henry Wellcome Building of Gene Function, University of Oxford, Parks Road, Oxford, OX13PT, UK

26. Cancer Research Program, Rajiv Gandhi Centre for Biotechnology, Poojapura, Thycaud P.O, Thiruvananthapuram, Kerala 695014, India

27. FUDECI, Fundación para el Desarrollo de las Ciencias Físicas, Matemáticas y Naturales. Av, Universidad, Bolsa a San Francisco, Palacio de Las Academias, Edf. Anexo, Piso 2, Caracas, Venezuela

28. Biology Department, Box 90338, Duke University, Durham, NC 27708, US

29. Bioinformatics and Computational Biology Laboratory, Iowa State University, Ames, IA 50011, USA

30. Center for Biomolecular Science and Engineering, School of Engineering, University of California Santa Cruz (UCSC), Santa Cruz, CA 95064, USA

31. Faculty of Mathematics, Physics and Informatics, Comenius University, Mlynska Dolina, Bratislava 84248, Slovakia

Correspondence and requests for materials should be addressed to HBS (Email: brad.shaffer@ucla.edu), or PM (Email: pminx@genome.wustl.edu).

**Abstract**

*Background:* We describe the genome of the western painted turtle, *Chrysemys picta bellii*, one of the most widespread, abundant and well-studied turtles. We place the genome into a comparative evolutionary context, and focus on genomic features associated with tooth loss, immune function, longevity, sex differentiation and determination, and the species' physiological capacities to withstand extreme anoxia and tissue freezing.

*Results:* Our phylogenetic analyses confirm that turtles are the sister group to living archosaurs, and demonstrate an extraordinarily slow rate of sequence evolution in the painted turtle. The ability of the painted turtle to withstand complete anoxia and partial freezing appears to be associated with common vertebrate gene networks, and we identify candidate genes for future functional analyses. Tooth loss shares a common pattern of pseudogenization and degradation of tooth-specific genes with birds, although the rate of accumulation of mutations is much slower in the painted turtle. Genes associated with sex differentiation generally reflect phylogeny rather than convergence in sex determination functionality. Among gene families that demonstrate exceptional expansions or show signatures of strong natural selection, immune function and musculoskeletal patterning genes are consistently overrepresented.

*Conclusions:* Our comparative genomic analyses indicate that common vertebrate regulatory networks, some of which have analogs in human diseases, are often involved in the western painted turtle's extraordinary physiological capacities. As these regulatory pathways are analyzed at the functional level, the painted turtle may offer important insights into the management of a number of human health disorders.

**Keywords**

amniote phylogeny, anoxia tolerance, chelonian, freeze tolerance, genomics, longevity, phylogenomics, physiology, turtle, evolutionary rates.

## Background

Turtles (also known as chelonians or Testudines) are an enigma. As the vertebrate

paleontologist Alfred Romer noted half a century ago, “The chelonians are the most bizarre, and yet in many respects the most conservative, of reptilian groups. Because they are still living, turtles are commonplace objects to us; were they entirely extinct, [they] would be a cause for wonder.”[1] From the Triassic to the present, turtles have been morphologically conservative, and even the earliest turtles[2] are instantly recognizable. The living crown group of turtles extends back at least 210 million years[3] and is characterized by a number of unique morphological and physiological features. Besides their distinctive shell, turtles have extremely long lifespans, are often reproductively active at very advanced ages, often determine sex by the temperature at which eggs incubate, are the most anoxia-tolerant tetrapods known, and have the capacity in some species to freeze nearly solid, thaw, and survive with negligible tissue damage. The Western Painted Turtle genome harbors a wealth of information on the genetic basis of these and other adaptations that characterize this unique vertebrate lineage.

Two of the great physiological challenges to vertebrate survival are hypoxia and cold tolerance. Particularly for temperate ectotherms like the Western Painted Turtle, the two are closely linked, because winter hibernation often occurs underwater in ice-locked ponds, and involves long periods with limited access to oxygen. The Western Painted Turtle is capable of surviving, with no loss of physiological function, four months under conditions of exceptionally low oxygen availability at 3°C[4] and at least 30 hours at 20ºC[5]. This anoxia tolerance, when combined with the ability to survive freezing of 50% body water[6], allows hatchling Painted Turtles to endure long winters in their nests across the northern part of their range in North America. It also provides an unprecedented model to study natural mechanisms that protect the heart and brain from

hypoxia-induced injury. Cardiac infarct and cerebral stroke are the first and third leading causes of death in the United States[7], and while conventional therapies continue to extend human lifespan, progress in improving outcomes from these conditions has been limited. Our genomic analyses indicate that Painted Turtles frequently achieve their extreme physiological capacities, at least in part, using conserved amniote molecular pathways; functional analyses of these pathways across turtles with varying physiological capacities thus may provide important insights for human disease prevention.

**Results and discussion**

*Reference genome*: We sequenced the nuclear genome of a single female Western Painted Turtle, *Chrysemys picta bellii*, that we field collected from southern Washington, using a combination of next-generation whole genome shotgun and Sanger-based BAC end reads (see Materials and Methods, *Sequencing and Assembly*, Additional file 1, Tables S1, S2). The assembly averages 18-fold coverage across 2.59 Gb with an N50 scaffold size of 5.2 Mb, and represents at least 93% of the genome. By all available measures, the assembled sequences have sufficient nucleotide and structural accuracy to provide a suitable template for initial analysis (see Materials and Methods, *Assembly Quality and Coverage Assessments*).

*Genome annotation:* After soft masking the *C. picta bellii* genome with RepeatMasker[8], gene annotation was performed using the homology based pipeline GPIPE[9-11] using a non-redundant protein set from human (Ensembl release 66), chicken (Ensembl release 66) and green anole (Ensembl release 66) as template. Based on the quality of the alignments with the template proteins, the conservation of exon

boundaries and the absence of frame shifts and premature stop codons, we predicted a total of 21,796 protein-coding gene models in *C. picta*, including 144,670 exons (average 6.63 exons per gene), and an average transcript size of 1,023 nucleotides (median 743 nucleotides). Using cDNAs obtained through 454 sequencing of libraries derived from brain, testes, ovaries, and trunk, we identified a total of 40,091 exons within 7,961 gene models to which cDNAs could be mapped.

*Repeat structure:* Approximately 10% of the *C. picta* assembly contains an abundance of transposable elements (TEs) that include nearly 80 distinct lineages of RNA-derived retrotransposons and DNA transposons, suggesting a long and dynamic history of clade-specific genomic diversification (see Additional file 1, Table S3, Additional file 2, Figures S1-S4). The Western Painted Turtle exhibits intermediate TE copy number relative to birds and the lizard *Anolis*, and is rich in LTR elements including endogenous retroviruses, LINEs in the CR1 and RTE families, predominantly MIR-like SINEs, and DNA transposons (see Additional file 2, Figures S1-S3). These transposons include 385 *SPIN* elements in the hAT-Charlie family not previously detected by slot blot hybridization assays for seven turtle and four crocodilian species[12]. Consistent with the close evolutionary relationship between turtles and archosaurs (birds and crocodilians, see below), these elements and the overall genome have a GC content of 43% that is more similar to birds than *Anolis*[9, 13] (see Materials and Methods, *Repeat Structure*, Additional file 1, Table S3, Additional file 2, Figures S1-S6). *Chrysemys p. bellii* also exhibits a moderate density of tandem repeats (1% genomic sequence coverage with an average density of 111 repeats per MB) with length and frequency distributions more similar to birds than to *Anolis*[13]. Overall, the repetitive landscape of *C. picta* exhibits a

substantial amount of lineage-specific evolution that distinguishes turtles from other major amniote taxa but exhibits some similarities to archosaurs, in keeping with their sister group relationship. Long generation times and a slow rate of molecular evolution may have facilitated the diversification of turtle repeats, potentially impacting both genomic stability and dynamics of transcriptome function[14-17].

*Isochore structure*: The presence of GC-rich isochores is a well-known feature of birds and mammals, but is a minor component of genomic structure in the lizard *Anolis*. The Western Painted Turtle genome has an average GC proportion of 0.43, which is consistent with other amniotes (see Additional file 2, Figure S7). At a 3-kb scale, the standard deviation of GC content is 0.059, which is also intermediate among vertebrate genomes (see Additional file 2, Figure S7). The standard deviation of GC content in the Western Painted Turtle is intermediate between those of the lizard *Anolis* and mammals/birds for sliding window sizes ranging from 5-320kb (Figure 1), suggesting that the gene-rich isochores that characterize the endothermic birds and mammals are not as prominent a feature of the Western Painted Turtle genome (see Materials and Methods, *Isochores*, Additional file 1, Table S4, Additional file 2, Figures S7-S9). For the Western Painted Turtle, we found a weak but significant correlation between the GC content of protein-coding genes and their flanking sequence, indicating a slight, but potentially important relationship between genomic environment and the nucleotide composition of genes (see Additional file 2, Figure S8). We also found a slight negative relationship between the GC content and the length of intergenic sequences in the Western Painted Turtle (not shown); thus, GC rich regions tend to be more gene dense. This is a strong relationship in mammals and birds, but is nonexistent in *Anolis*.

To examine the evolution of GC content in the context of the vertebrate phylogeny, we quantified GC content at third codon positions (GC3) using the 2,366 simple orthologs (1:1) identified from the OPTIC pipeline orthology predictions for zebra fish, pufferfish, chicken, zebra finch, Western Painted Turtle, green anole, platypus, mouse, and human (see Materials and Methods, *Identification of gene family expansion/contraction* for methods on determining gene homology). We used the program NHML (with default parameters) to estimate (1) ancestral GC content, and (2) GC3*, the equilibrium GC content, which can be interpreted as the GC content toward which a lineage is evolving. Our results are consistent with trends from previous phylogenetic analyses of GC content[18, 19], with the exception that chicken seems to be in equilibrium with regard to GC3. The Western Painted Turtle shows a striking decrease in GC3 from its current value of 46.74% to a GC3* value of 38.90%, indicating an erosion of GC content that is also seen in *Anolis* (see Additional file 2, Figure S9)[18].

One mechanism that can contribute to this erosion is homogenization of recombination. Recombination is correlated with several evolutionary processes and genomic features. For instance, regions with higher recombination activity experience more efficient selection as well as higher GC content in mammals and birds. It stands to reason that genes with higher GC3 will have a lower lineage-specific dN/dS; that is, genes with higher GC content will experience more efficient selection. To test this, we divided up the genes from human, chicken, and Western Painted Turtle into “high GC3” and “low GC3” genes based on the GC3 values of genes for each taxon. We then examined the distribution of dN/dS values between these two groups for each taxon. We expected, if recombination has a landscape similar to mammals and birds, that the “high

GC3" genes will have a lower lineage-specific dN/dS and "low GC3" genes will have greater dN/dS values. We found this to be the case in human and chicken[18], indicating a heterogeneous recombination landscape (see Additional file 1, Table S4). In the painted turtle, we found that there is an even greater disparity in dN/dS between "high GC3" and "low GC3" genes than in human and chicken, indicating that an even more heterogeneous landscape exists in turtle. This may indicate that rather than a recombination-based mechanism driving GC content in turtle (e.g. GC-biased gene conversion), mutational biases are playing an important role in the trajectory of GC3.

*Phylogeny and evolutionary rates:* The phylogenetic position of turtles has remained one of the last unresolved problems in vertebrate evolutionary history, with recent hypotheses suggesting widely disparate placements[20, 21]. Our phylogenetic analysis of 1955 sets of rigorously screened gene orthologs (see Materials and Methods, *Multiple alignments and gene orthologs*) for eight vertebrate species (human, platypus, chicken, zebrafinch, anole, turtle, python, and alligator), analyzed separately or as a concatenated dataset, concur with two recent phylogenomic analyses[20, 22] in placing turtles as the sister group to Archosauria with strong statistical support (Figure 2). Thus, based on independent, genome-scale analyses, the phylogenetic placement of turtles as well-nested within diapsid amniotes appears to be relatively secure.

We also estimated the relative rate of substitution in a smaller dataset that was designed to minimize missing data. This dataset comprised 309 orthologs that were identified in all eight species. Our analyses indicate that the turtle lineage has undergone a remarkable substitution-rate slowdown relative to other amniotes (Figure 2). Estimates of relative evolutionary rates under a relaxed molecular clock suggest that turtles have the

slowest rate of substitution among the eight representative amniote lineages analyzed. Turtle genomes evolve at about one-third the rate seen in humans, and roughly one-fifth the rate of the fastest-evolving python lineage (see Materials and Methods, *Phylogeny and substitution rate*, Additional file 1, Tables S5-S6). Given the long generation time that characterizes turtles, our comparative analysis is consistent with the negative relationship between generation time and rate of molecular evolution found in reptiles[23] and other amniotes[24], although the observed slowdown in archosaurs and turtles may also suggest a broad, lineage-specific effect.

*Extreme anoxia tolerance in the painted turtle:* Although all turtles can withstand anoxia for a few hours with no discernable tissue damage, the Painted Turtle is a candidate for the most extreme anoxia-tolerant tetrapod known. To explore the transcriptomic basis of this extreme anoxia tolerance, we assembled a gene expression profile by sequencing poly A-enriched RNA isolated from the ventricle (heart) and telencephalon (brain) of normoxic and anoxic (N=4 turtles/group, 24 hours at 19ºC) adult Western Painted Turtles (see Materials and Methods, *Anoxic gene expression*). FPKM (Fragments per kilobase of exon model per million mapped fragments) values from 13,236 Western Painted Turtle genes with human orthologs were interrogated (from a starting pre-filtering pool of 22,174 gene orthologs) and analyzed with ANOVA. Differential gene expression significantly increased in brain (19 genes) and heart (23 genes) (see Additional file 1, Tables S7, S8), mirroring previous work showing up-regulated gene expression in response to hypoxia in other vertebrate tissues, including many cancers.

The largest overall change in expression was in *APOLD1*, an apolipoprotein encoding gene whose transcript levels increased 128-fold in telencephalon and 19-fold in ventricle (see Additional file 1, Tables S7, S8; Additional file 2, Figure S10). *APOLD1* expression moderately increases during hypoxia in human microvascular endothelial cell culture, although its exact function remains unclear[25]. Other highly differentially expressed genes (>10-fold; *FOS, JUNB, ATF3, PTGS2, BTG1/2* and *EGR1*) encode proteins that, individually and in dimeric forms, have been implicated in the control of cellular proliferation, cancers, and tumor suppression[26-29]. The 30-fold increase in a gene orthologous to *SLC2A1* (see Additional file 1, Table S8, Additional file 2, Figure S11), which encodes the glucose transporter GLUT-1, is also exceptional since deficiencies in membrane glucose transport underlie diabetes in humans. An understanding of the mechanism by which membrane GLUT-1 levels increase in the turtle would be a useful contribution to human diabetes research. Decreases in gene expression were fewer and found only in ventricle (see Additional file 1, Table S9; Additional file 2, Figure S12), but included decreases in *CDO*, which is important in regulating intracellular cysteine as well as levels of the endogenous metabolic depressant hydrogen sulfide[30, 31], and genes involved in mRNA splicing (*SRSF5*)[32] and tumor proliferation (*MKNK2*)[33].

These analyses demonstrate the power of the Western Painted Turtle as a model for the evolution of anoxia tolerance by regulatory changes utilizing broadly conserved vertebrate genes, including many genes that lead to pathogenesis in humans. Clearly, further study of the processes that link these regulatory changes to anoxia tolerance are an next important step. Although this is yet to be tested, we also note that the regulatory

pathways that evolved in the Western Painted Turtle could lead to the identification of targets for therapeutic intervention in human diseases involving hypoxic injury and possibly tumorigenesis.

*A movel microRNA associated with freeze tolerance in hatchling painted turtles:* Freeze tolerance constitutes a second suite of physiological adaptations that are integral to winter survival for hatchling Painted Turtles and other species that overwinter in shallow terrestrial nests. Molecular adaptations that underlie natural freezing survival in *C. picta* include strong metabolic rate depression, use of anaerobic metabolism (see *Extreme Anoxia Tolerance in the Painted Turtle*), and selective up-regulation of genes involved in key cellular processes[34].

Entrance into hypometabolism involves regulatory changes in multiple metabolic processes coordinated by extracellular stimuli that are readily induced and reversed to allow smooth transitions to and from the frozen state. MicroRNA regulation of mRNA transcripts meets these criteria and is involved in other models of stress-induced metabolic rate depression[35]. Using the Western Painted Turtle genome, we retrieved the precursor sequence of miR-29b, a microRNA involved in DNA methylation and regulation of glucose transport[36, 37] that is often associated with freeze and anoxia tolerance (see Materials and Methods, *Freeze tolerance*). Based on this sequence, the secondary structure of Western Painted Turtle pre-miR-29b was predicted to contain a single nucleotide mutation (nuc-43) resulting in a larger terminal stem-loop compared to the less freeze tolerant turtle *Apalone spinifera* and *Homo sapiens*. Although the functional significance of this mutation is unknown, microRNAs are generally extremely conserved across vertebrates, and nucleotide structures that restrain the terminal loop

region (as predicted for human and other turtles) can decrease the efficiency of Dicer processing of precursor microRNA transcripts in the range of 50% (Figure 3A)[38]. In addition to loop flexibility, slight alterations to loop structure and nucleotide sequence can influence interactions between pre-microRNA and terminal loop binding proteins, impacting processing efficiency. Consistent with the hypothesis that enhanced microRNA processing under low temperature stress facilitates freezing survival, quantitative RT-PCR (see Materials and Methods, *Freeze tolerance*) revealed a mild but statistically significant 1.3-fold increase in processed mature miR-29b levels in liver of hatchling turtles in response to 24 h freezing; expression was maintained and possibly increased during subsequent thawing (Figure 3B).

Although these results require additional functional analyses and are clearly preliminary, they point to future work on miR-29b as a potential candidate for freeze tolerance work on turtles with this physiological capacity. With refined genomic and comparative data across freeze tolerant and intolerant turtles, future studies of turtle freeze tolerance should help confirm or refute our interpretation that mutations in miR-29b are an important component of freeze tolerance in turtles.

*Tooth loss pseudogenization:* Turtles lost the ability to form teeth ~150-200 million years ago, making them the oldest extant edentulous lineage of tetrapods (birds lost teeth ~80-100 mya)[39]. Previous studies in birds and edentulous mysticete (baleen) whales demonstrated that tooth loss is closely associated with the pseudogenization and subsequent degradation of the tooth-specific genes enamelin (*ENAM*), amelogenin (*AMEL*), ameloblastin (*AMBN*), dentin sialophosphoprotein (*DSPP*), and enamelysin (*MMP20*)[40, 41]. We identified the majority of turtle pseudo-exons in their

chromosomally syntenic regions (see Materials and Methods, *Tooth loss*) when compared to other amniotes (Figure 4), consistent with the very slow rate of genomic change seen in chelonians (e.g. Figure 2). Turtle *ENAM*, *AMEL*, and *MMP20* all contain premature stop codons (exons 5, 3 and 2 respectively) in addition to highly degenerated sequences. *AMBN*, while somewhat more conserved, has a premature stop codon in exon 7. While *DSPP* exons 1 and 2 are relatively conserved, all subsequent exons were unidentifiable. Sequence identity scores between pseudogene exons identified in turtle and chicken were not significantly different from each other compared to their functional orthologs in crocodilians (see Materials and Methods, *Tooth loss*, Additional file 1, Tables S10, S11), even though turtles lost their teeth ~50-100 million year earlier.

This extremely conservative pattern of tooth-loss pseudogenization across amniotes is consistent with a single evolutionary origin (and regulatory network) of teeth, and suggests that the deterioration of this pathway evolved independently (that is, is homoplastic) in turtles, whales and birds. This is also consistent with the fossil record, as early members of all three lineages are known to be toothed. However, concordant with their overall slow rate of molecular evolution, the tooth-specific genes in turtles have accumulated mutations at roughly half the rate of accumulation found in birds.

*The genomic basis of longevity in turtles:* One of the defining features of turtles as a lineage is their extreme longevity (many species live 100 years or more), and we used the Western Painted Turtle genome to investigate this quintessential chelonian feature. Based on previous work implicating the shelterin complex encoding genes in exceptional longevity in the naked mole rat[42], we evaluated (by BLAST searches of all available turtle sequence data including unplaced scaffolds, see Materials and Methods, *Aging and*

*longevity*, Additional file 1, Table S12) the status of the shelterin complex in the Western Painted Turtle genome. Even with this comprehensive search, we were unable to find orthologs for three of the five genes (*POT1, TERF2IP, TEP1*) in the Western Painted Turtle. Given that *TEP1* is also absent in birds, this result strongly suggests that turtles (and their sister-group, the archosaurs) do not share this longevity mechanism with the naked mole rat.

We also examined genes that have apparently been lost in the Western Painted Turtle (and were also absent in our searches of all other available turtle genomes) to investigate their relevance to aging based on their orthology to known aging-linked genes in model organisms[43]. Specifically, lowered activity of *ATP5O* in the nematode *C. elegans* increases longevity[44], while *PLCG2* is a crucial intracellular signaling modulator and seems to be negatively affected by aging[45]. Although confirming the absence of genes is difficult with incompletely assembled genomes, the Western Painted Turtle genome is at least 93% complete, and their absence in other turtle genomes is compelling (see Additional file 1, Table S12). Among these presumably missing genes, the lack of *ATP5O* (for which we found no hits in any turtle) and *PLCG2* (where we found evidence for a total of six out of 30 exons across all turtles) may be important in the extraordinary longevity of turtles.

*Temperature-dependent sex determination/differentiation (TSD) genes:* Since the first realization that many, but not all, turtles have TSD, turtles have become a model system for comparing the gene networks controlling genotypic sex determination (GSD) and TSD. Phylogenetic reconstruction indicates that the ancestral condition of sex determination in turtles and crocodilians was thermosensitive (TSD), and that GSD has

re-evolved in several turtle lineages[46]. Although it is now clear that TSD and GSD each encompass multiple mechanisms whose divergence involves regulatory and structural evolution affecting the level of plasticity and canalization of vertebrate sexual development[47, 48], it also remains the case that transitions between TSD and GSD have occurred many times, and that TSD is the ancestral condition in turtles. Genomic analyses of TSD and GSD turtles (and crocodilians) can provide important clues to help decipher the changes in genetic architecture that underlie these evolutionary transitions. Comparative analysis of genomes and transcriptomes from TSD turtles (*Chrysemys picta*, *Chelydra serpentina*, *Trachemys scripta*) and the GSD softshell turtle *Apalone mutica* (all data produced by our group) from early through late embryonic stages revealed that virtually all of the known vertebrate genes involved in sexual differentiation are present in turtle genomes and active during sexual development (see Materials and Methods, *Sex determination/differentiation*, Additional file 1, Table S13).

We took a gene-tree reconstruction approach to examine the phylogenies of the coding regions of five key genes involved in the gonadogenesis regulatory network whose transcriptional responses have been studied in the Western Painted Turtle (*WT1*, *SF1*, *SOX9*, *DMRT1*, and *AROMATASE*[48, 49], Figure 5). Although the roles of these genes in the TSD/GSD transition remains incompletely understood, they are important in sexual differentiation in a variety of vertebrates including reptiles. Our primary goal was to ask whether these individual gene trees cluster taxa based on their phylogenetic relationships (as might be expected if independent TSD/GSD transitions have evolved that do not mask phylogeny) or on their TSD/GSD phenotype. Consistent with their phylogenetic relationships, our gene tree analyses generally placed the monophyletic set

of turtle orthologs as the sister group to archosaurs, (compare the relationships of turtles and crocodilians in Figure 2 with Figure 5), although in one case (*WT1)* TSD turtles and crocodilians were sister groups (Figure 5). However, within-turtle relationships of these five gene trees often resolve the GSD softshell *Apalone spinifera* out as sister group to the remaining turtles, rather than in its generally established placement as sister to the remaining cryptodires[50]. It is well known that estimates of individual gene trees can differ from species trees for purely statistical reasons, and the interrelationships of softshells to other turtles has been notoriously difficult to determine with molecular data[50-52].

Overall, there is no compelling evidence of clustering TSD and GSD turtles, or TSD and GSD vertebrates that is contrary to their phylogenetic relationships, suggesting that strong convergence at the molecular level has not occurred in these markers. Interestingly, dn/ds analysis revealed that the molecular evolution of these elements is driven overwhelmingly by purifying selection, with only few instances of neutral evolution between some closely related species pairs such as *Trachemys scripta* (TSC) and *Chrysemys picta* (CPI) for *SF1*, *AROMATASE* and *Wt1,* TSC and *Apalone spinifera* (ASP) for *SF1,* ASP and CPI for *SF1*. Thus, these analyses indicate that the primary patterns of gene tree evolution in these loci associated with sex determination are driven by their organismal (phylogenetic) history rather than TSD/GSD functionality.

*Immune system genomics*: Given the striking preponderance of expansions of immune function genes (see below), and their potential importance in the extended life spans of turtles, we characterized a large panel of immune-function genes in the Western Painted Turtle genome. We aligned the *C. p. bellii* genome against a sequence database

of ~3000 immune-function related genes developed from a diverse set of 14 vertebrates ranging from lamprey to mammals (see Materials and Methods, *Immune system*). Blast searches of the *C. p. bellii* genome against this database resulted in the identification of 110 genes, 100 of which were confirmed with reciprocal alignments; 73 were also identified in either cDNA or predicted gene sequences (see Additional file 1, Table S14). The cDNA represented a small number of tissues/developmental stages, and 73/110 (66%) confirmation of expression is very encouraging.

The adaptive immune response of turtles is generally slower and less robust than its mammalian counterpart, and does not consistently demonstrate evidence of a memory response[53, 54]. However, we identified several major components necessary for adaptive immunity and generation of immune memory including *CD4*, *MHCII* and the immunoglobulin heavy chain locus (see Additional file 1, Table S14). Our analysis also demonstrates that the Western Painted Turtle has a unique repertoire of Toll-like Receptors (TLRs), comprised of those found in amphibians, fish, birds, and mammals. This includes a TLR15-like receptor that has previously only been defined in birds, and is known to interact with bacterial pathogens including *Salmonella*[55] (see Additional file 1, Table S15). Given the delayed adaptive response and poor generation of immune memory, combined with their diverse set of TLRs, we predict that turtles should rely more heavily on the non-specific innate immune response to effectively recognize and initiate appropriate responses to pathogens. This initial response would be followed by a more moderate adaptive response that, because of the low specificity due to lack of immune memory formation, may serve as a general mechanism to combat remaining pathogens. Given the overall low specificity of their innate and adaptive immune

responses, it seems that turtles are able to adequately balance their immune compartments to eliminate pathogens, while simultaneously avoiding damage to self-tissues as a result of an overactive immune response.

*Gene family expansions*: Gene family expansions point to candidate sets of genes of particular importance in chelonian survival and evolution. After annotating the Western Painted Turtle genome (see Materials and Methods, *Identification of gene family expansion/contraction*, Additional file 1, Table S16), we used phylogenetic reconstructions of the genomes of three mammals (human, mouse, platypus), two birds (chicken, zebrafinch), one lizard (green anole), and two fish (tetraodon, zebrafish) to identify one-to-one orthologs, as well as gene losses and gene family expansions in the Western Painted Turtle genome. We identified 3,222 one-to-one orthologs across all nine species, 4,828 genes among the seven amniote species, and 103 gene families including 957 gene predictions that show expansion in the Western Painted Turtle lineage. Among these expanded gene families, 15 of the 27 with four or more members, which jointly account for 623 of 957 gene predictions, were annotated as being involved in immune response (see Materials and Methods, *Expansion of gene families involved in the immune response*, Figure 6, Additional file 1, Table S17); an additional large expansion (106 members, 101 confirmed by manual curation) was evident among the beta-keratins (see Materials and Methods, *Beta-keratin expansions*) involved in the formation of scales, claws and scutes that encase the shell[56]. Additional analyses using beta-keratin mRNAs extracted from the precursor cells of the shell of *Pseudemys nelsoni*[56] indicates that there have been independent lineage-specific expansions of the beta-keratins in birds and

turtles associated with the formation of feathers and the shell (Li et al., unpublished results).

*Patterns of natural selection:* Genomic scans for positive selection across turtles constitute a complementary strategy to identify genes underlying chelonian adaptations. We examined a carefully screened ortholog set of 4136 genes (see Materials and Methods, *Ortholog sets*) for eight vertebrate species (human, platypus, chicken, zebrafinch, anole, turtle, python, and alligator) to detect signs of turtle lineage-specific positive selection. Using branch-site likelihood-ratio tests[57]with reduced parameterization[58] (see Materials and Methods, *Positive selection*), we identified 671 genes under positive selection (false discovery rate < 0.1) (Accessory Data File 1). Among these genes were several categories of interest to notable physiological in turtles, several of which we highlight here.

There were nine genes containing ankyrin repeat motifs (the most significant was *ANKRD32*, $P=1.1x10^{-17}$), which are typically sites of protein-protein interactions. Furthermore, some of these ankyrin-repeat-motif genes contained SOCS box (suppressor of cytokine signaling) domains as well (*ASB14*, $P=1.2x10^{-2}$ and *ASB18*, $P=8.8x10^{-3}$) and are involved in protein turnover regulation[59]. In addition, a number of chemokine receptors – *CCR4* (P=1.1x10-7), *CCR5* ($P=7.2x10^{-5}$), and *CCR10* ($P=1.0x10^{-19}$), as well as *CCRL1* ($P=4.2x10^{-9}$), showed evidence of positive selection in our analysis. These G-protein coupled receptors bind specific cytokines (chemokines), are involved in chemokine-mediated signaling, and are generally pro-inflammatory/immune responsive[60]. We found evidence for significant positive selection in *DMRT2* (P=4.2x10-4). *DMRT*s have been found to associate with sexual determination and

development (see earlier section on *Temperature-Dependent Sex Determination (TSD) genes*, also reviewed in[61]).

Related to oxidative phosphorylation and free-radical scavenging, several positively selected genes were involved directly (e.g., *ATP5S*, $P=3.7x10^{-6}$; *ATP5H*, $P=3.1x10^{-4}$; *COX15*, $P=1.4x10^{-4}$; *ATP5G3*, $P=1.4\ x10^{-3}$; *COX7A2*, $P=6.0x10^{-3}$; *ATP5B*, $P=1.9\ x10^{-2}$; *DAP3*, $P=8.3x10^{-6}$) or indirectly (e.g., *SOD1*, $P=1.5x10^{-7}$; *ACO2*, $P=1.4x10^{-2}$) in this process. Adaptations within genes in the process of ATP formation (specifically those that are subunits of ATP synthase) and anti-oxidant defenses have been proposed as mechanisms of life-history evolution in reptiles[62]. Several additional genes involved in life history traits were also under positive selection, including those involved in fertility (*FSHB*, $P=7.4x10^{-6}$), reproduction/immune functionality (*prolactin receptor*, $P=4.4x10^{-8}$), and aging (*SIRT3*, $P=2.1x10^{-3}$; *CLK1*, $P=1.1x10^{-2}$). In general, these 671 positively selected genes are involved in diverse functions that span biological processes. Although a numerically large set, our careful filtering and criteria for ortholog consideration suggests they are a robust set that is larger than would be expected when compared to naked mole rat or human[42, 63].

We detected 171 GO functional categories showing enrichment for genes under positive selection (nominal P-values < 0.05, Mann-Whitney U-test), however, none were statistically significant after multiple testing correction (see Accessory Data File 1 for an overview of genes under positive selection and GO category enrichments).

## Conclusions

The Western Painted Turtle, and chelonians generally, comprise a unique combination of extremely conservative evolutionary history interspersed with some of the most unique physiological and behavioral adaptations found in amniotes. Our analyses of the Western Painted Turtle genome indicate that common vertebrate regulatory pathways are often involved with these novel phenotypes, and additional functional experiments can now investigate the ways in which these pathways have been modified in turtles. Our extensive analyses of anoxia tolerance provides particularly strong support for the interpretation that the Western Painted Turtle utilizes common vertebrate pathways to achieve its extraordinary physiological abilities; temperature-dependent sex determination and immune system functionality also appear to utilize common suites of vertebrate genes. Genomic analyses of longevity and particularly tooth loss, both of which characterize all living chelonians, suggest that patterns of gene loss are also key elements of turtle evolutionary novelties. The Western Painted Turtle genome, enabled by both comparative genomics and functional experimentation, has provided and will continue to provide windows into the evolution of physiological novelties, perhaps including some with biomedical and cryopreservation applications.

One aspect of turtle evolution that is proceeding at a rapid and accelerating pace is human-mediated extinction. Although the lineages represented by living turtles have survived countless challenges in the last 210 million years, current estimates are that at least 50% of the 330 recognized species of living chelonians are threatened with extinction[64]. Turtles far outstrip amphibians, mammals, and birds in their proportion of at-risk species, and the survival likelihood of many species is bleak. Future comparative genomics work on turtles, including comparisons among species that vary in their

longevity, anoxia and freeze tolerances, immunocompetency, and a host of other key human challenges, requires healthy populations of the remaining diversity of turtles. The challenge, for comparative biology and conservation alike, is to preserve the remaining diversity of living turtles as we continue to unravel their secrets for success.

## Materials and methods

*Sequencing and assembly.* A single *Chrysemys picta belli* (Western Painted Turtle) was sequenced at The Genome Center, Washington University School of Medicine, St Louis, Missouri. The whole genome shotgun library primary donor-derived reads (B. Shaffer lab, female, field number: RCT428, locality: WA Grant Co, small lake 1.3 miles south of Potholes Reservoir, tissue accession number: HBS 112648) and BAC end reads (BAC library source: VMRC CHY3: J. Froula, JGI (from C. Amemiya lab) female, strain: MVZ #238119, Locality: Frenchman Hills wasteway 9.0 mi S via Dodson road of junction with Hwy I-90, Grant Co., Washington) were assembled using Roche's Newbler (version 2.6) with stringent parameters. Newbler uses all of the input single and paired end read data (including the paired BAC end data) to create contigs and then, focusing on the paired end read data along with estimates of insert size, organizes those contigs into larger scaffolds. After removing contamination, the resulting assembly was labeled as 3.0.1. All scaffolds larger than 500 bases (81,642 scaffolds with a total size of 2.59 Gb, N50 scaffold size of 3.01 Mb (N50 number is 248)) were retained for submission to the public databases.

After the assembly was complete, 15X of paired end sequencing data were generated on the Illumina platform and used only for error correction in the reference

assembly; the Illumina paired end data were not used to aid in scaffolding of existing contigs. For error correction, the Illumina data were aligned against the 3.0.1 assembly using bwa[65] and processed using samtools and bcftools[66]. Based on the paired end mapping data, all duplicate mapped reads were removed. One and two basepair indels were introduced into the reference for all cases where there were $\geq$3 and $\leq$200 reads aligned (mapping quality $\geq$40 and the indel was $\geq$10 bases from the end of the alignment), and where all reads disagreed with the reference and agreed with one another. There were a total of 27,296 indels introduced into 24,712 contigs.

The assembly data were aligned utilizing BLASTZ[67] to align and score non-repetitive turtle regions against the following repeat-masked genomes: anole (anoCar2), human (hg19), chicken (galGal3) and opossum (monDom5). Alignment chains differentiated between orthologous and paralogous alignments[68] and only "reciprocal best" alignments were retained in the alignment set. The alignments were post-filtered in the following ways: (1) only alignments that extended over at least 2000 bases where the relative expansion/contraction was less than 10X were retained, (2) alignments were then smoothed by removing any single alignments that were <10 kb and occurred as a single alignment in between a large block of separate alignments to the same chromosome. The relative scaffold ordering was then examined in the four pairwise alignments. If at least three of the different pairwise alignments with the other species all suggested a given order and orientation, that pairwise ordering was retained in a list of valid orders (and orientations). Then the consistent pairwise alignments were linked into groups. The AGP was created using those lists of ordered and oriented scaffolds. Because ordering by homology is not absolutely confident, the gaps between scaffolds were annotated as

"contig" gaps including a "no" in the final column indicating that there is no spanning clone closing the gap. There was approximately 1.2 Gb of sequence organized into 290 ordered groups leaving 80,697 individual scaffolds totaling 1.3 Gb. The N50 scaffold size rose to 5.2 Mb (N50 number is 148).

*Assembly quality and coverage assessments.* As indicated by comparisons of the submitted assembly with a set of 64 finished Western Painted Turtle BACs (BAC library source: VMRC CHY3; J. Froula JGI (from C. Amemiya, Benaroya Research Institute, Seattle, WA) Female; Strain MVZ #238119; Locality: WA: Grant Co: small lake 1.3 miles south of potholes reservoir) totaling 9.3 Mb of finished sequence, structural accuracy of the assembled sequence is sufficient for these analyses. These completed BAC sequences were not included in the assembly and thus provide an important data set for assessing assembly accuracy and coverage. Some small supercontigs (most <5 kb) were not positioned within larger supercontigs (<1 event per 500 kb). While these are not strictly errors, they do affect overall assembly statistics. There are also small, undetected overlaps (most <1 kb) between consecutive contigs (~1 event per 30 kb), occasional local mis-ordering of small contigs (~1 event per Mb), and small contigs incorrectly inserted within larger supercontigs (<1 event per 275 kb). Overall, the rate of rearrangements with respect to finished BACs was comparable to previous next generation WGS assemblies. Nucleotide-level accuracy is high by several measures. Over 99% of the consensus bases in the Western Painted Turtle sequence have quality scores[69] of at least Q40 corresponding to an error rate of $\leq 10^{-4}$. Comparison of the WGS sequence to the 9.3 Mb of finished BACs from the sequenced individual is consistent with this estimate, giving a high quality discrepancy rate of $3x10^{-3}$ substitutions and $2x10^{-4}$ indels which is no more

than expected given the heterozygosity rate. The rate of substitutions is due to the polymorphism rate. By restricting analysis to high-quality bases, the nucleotide-level accuracy of the WGS assembly is sufficient for analyses presented here. As with the chimpanzee and other whole genome shotgun-based assemblies, the most problematic regions are those containing segmental duplications (Chimpanzee Sequencing and Analysis Consortium, 2005).

We estimate that Western Painted Turtle genome sequence covers at least 93% of the full genome sequence. To obtain this estimate, we first evaluated the coverage using the results of the alignments of the assembly against the 64 finished Western Painted Turtle BACs. The overall coverage of those BACs exceeded 93%. Second, we aligned a set of Western Painted Turtle cDNAs generated by this project against the genome assembly using BLAT[70]. The cDNA libraries were constructed from several tissue sources (see Additional file 1, Table S2) and were sequenced in our lab on the 454 Life Sciences instrument using methods previously reported[71]. The reads were assembled using the Newbler software package provided by 454 Life Sciences. The coverage estimates per tissue range from 93 to 98% when asking that at least 50% of the EST align to the genome or from 91 to 96% when requiring more than 90% of the EST aligns to the genome (see Additional file 1, Table S2).

Finally, we estimated coverage by looking at the coverage of a related genome using BLAT[70]. Over 96% of the draft assembly of the 1.5 Gb *Trachemys scripta* genome (separated by approximately 10-15 My from the Western Painted Turtle) aligned with the Western Painted Turtle genome.

*Repeat structure.* TE sequence divergence in three turtle genome assemblies

reveal a distribution that contrasts with the high turnover of younger L1s in the lizard (*Anolis*), the skewed accumulation of older TEs in the alligator, and near complete lack of SINEs and active CR1s in the small, homogenous genomes of birds (see Additional file 1, Figures S1-S5)[9, 11, 72]. The average G+C content of *C. picta* mobile elements is the same as the genome-wide average of 43% and the range of values for TE content and G+C among the N50 scaffolds is more similar to those observed in chicken than in *Anolis* (see Additional file 2, Figure S6)[9], consistent with its closer phylogenetic relationships to archosaurs.

Identification and classification of repetitive elements in the *C. picta* assembly were carried out on the full original *C. picta* assembly sequence (*C. picta bellii* v3.0.1) using the RepeatMasker version 3.3.0[8], Tandem Repeat Finder version 4.0.4[73] and Phobos version 3.3.12[74] software packages. For all available genome assemblies investigated RepeatMasker was run with the BLAST engine and repeat classification was carried out using the Vertebrate library from version 20110920 of the RepBase database. We employed Phobos using default parameters. Tandem Repeat Finder was run with the default alignment parameters except for a reduced MaxPeriod value of 200 instead of the default 500, and with exclusion of HTML output. These parameter settings were directly comparable to summary statistics available through TRDB for the most recent whole-genome assemblies of amniote species. Results from RepeatMasker were analyzed using RMPipeline[75], a set of generalized programs for analyzing RepeatMasker output written using Perl. These programs can be used to process any RepeatMasker output files and are publicly available and free to use under the GPLv3 license. Graphs were created using RMPipeline results, some additional Perl scripts, and Microsoft Excel.

*Isochores.* The absolute GC content of the assembly (after removing scaffolds with >20% missing data) is 0.434. We examined whether the assembly exhibited any bias in GC content. We divided the assembly into 4 equally-sized bins of increasing scaffold size (after omitting scaffolds with >20% missing data). The absolute GC contents of each bin were (range of scaffold lengths are indicated in bp): 0.496 (501-591), 0.498 (591-735), 0.496 (735-1,039), 0.433 (1,039-26,452,378). Because it appears there is a bias for smaller scaffolds to have a larger GC proportion, we focused our analyses of genomic GC content to those > 320 kb, a subset of the genome whose absolute GC is 0.430, a value very close to the whole-genome absolute GC. To generate the distributions of GC content, we divided up the genomes of human, dog, frog, turkey, zebrafinch, chicken, and Western Painted Turtle (scaffolds >320 kb) into 3 kb windows, using the GC content of these windows as measures (see Additional file 2, Figure S7). We also examined GC variation at different spatial scales, using non-overlapping windows of 5, 20, 80, and 320 kb (Figure 1). As window size quadruples, standard deviation should decrease by 50% for a completely homogeneous genome[76]. To determine the relationship between GC3 and flanking sequence, we used 10 kb upstream of the start codon and 10 kb downstream of the stop codon as the 20 kb flanking sequence. Only those flanking sequence with 80% complete data (allowing 20% combined missing data or clipped ends due to proximity of the gene to the ends of the scaffold) were considered. To examine the relationship between gene density and GC content, we divided up intergenic sequences into 10 equal-sized bins of increasing size and calculated the GC content of each bin. For the Western Painted Turtle, we found a weak but significant correlation between the GC content of

protein-coding genes and their flanking sequence, indicating that genomic environment influences the nucleotide composition of genes (see Additional file 2, Figure S8).

*Multiple alignments and gene orthologs.* Comparative genomic analyses (including studies of phylogenetic relationships, selection, conserved elements, and accelerated regions) are prone to artifacts derived from biases introduced by differences in gene prediction methods used in draft genome annotations of individual genomes included in the study, as well as gene prediction errors. In order to avoid having such biases dominate analyses, one can chose a well-annotated reference genome (in our case, human or chicken, whichever is more appropriate for a particular analysis), and annotations are remapped from the reference to the target genomes through multiple alignment. This step is followed by extensive checks to ensure the quality of derived annotations in target genomes.

A disadvantage of this approach is that novel elements introduced in non-reference genomes are not covered by the analysis. In case of human-referenced orthologs, the analysis only includes genes preserved throughout amniote evolution (since mammals are the sister groups of the remaining amniotes), while in the case of chicken-referenced orthologs, we analyze genome elements preserved during the evolutionary diversification of turtles and archosaurs (e.g. Figure 2). Thus, reference derived ortholog sets are best used in analyses requiring conservative high-confidence gene sets, and are not suitable for estimating target genome characteristics, such as numbers of genes, exons, or novel elements.

To construct a set of high-confidence orthologs, we used a methodology developed by Kosiol and colleagues[58]. First, we created a multiple alignment of human

(hg19), platypus (ornAna1), chicken (galGal3), zebrafinch (ornAna1), anole (anoCar2), turtle, python, and alligator, using a standard UCSC genome browser pipeline[77] based on BLASTZ[67] and multiz[78]. We based ortholog predictions on the human gene catalog of 21,360 genes (including RefSeq, UCSC known genes, ENSEMBL, and VEGA genes), which were remapped to all of the above species through these multiple alignments. We observed high variability for positions of translation start sites and stop codons, thus we also evaluated incomplete gene models, where we removed 10% on each end of the gene. Altogether, our gene set contained more than 378,000 alternative gene models.

Series of filters were run to identify which of these gene models can be considered high-confidence orthologs. For a gene model to be considered clean in a particular genome, we required that (a) it was covered by a single chain within the syntenic (for platypus and chicken) or reciprocal-best (for zebrafinch, anole, turtle, python, and alligator) net created using the UCSC genome browser pipeline, (b) there were no significant gaps in the gene alignments, (c) there were no frameshifts uncorrected within a short window of sequence, and (d) all elements important for the gene structure (donor sites, acceptor sites, translations start sites, and stop codons) were preserved. For each gene, we selected a single gene model that was clean in turtle, giving preference to the models that were clean in the most species and were the longest. The gene was excluded if it did not have any gene model satisfying these conditions (see Additional file 1, Table S6, which shows the number of genes filtered out in each step.) This approach resulted in 4786 high-confidence orthologs, out of which 3318 are incomplete (shifted start codon or stop codon). Out of these genes, 312 covered 2 species

(human and turtle), 622 covered 3 species, 757 covered 4 species, 896 covered 5 species, 1048 covered 6 species, 842 covered 7 species, and 309 covered all 8 species. An additional 12 genes that were incompletely covered in the reference genome were detected in the last stages of comparison and removed in post-processing.

*Phylogeny and substitution rate.* We estimated phylogeny using the set of 1955 orthologs that we identified in at least five of the eight genomes that we examined and had the potential to be informative about the phylogenetic position of turtles. We partitioned the dataset by codon position, using an independent GTR model for each position and allowing for gamma distributed rate variation among sites. We ran 4 independent analyses for 10 million generations, sampling every 1000 generations in MrBayes v. 3.1.2[79]. We then estimated the relative rate of substitution in a smaller dataset that was designed to minimize missing data. This dataset comprised 309 orthologs that were identified in all eight species. We used a UCLN relaxed clock model implemented in BEAST v. 1.7.1[80]. We partitioned the dataset by codon position, using independent general time reversible models of DNA substitution allowing gamma distributed rate variation for each position. We set the log normal distribution describing among-branch substitution rate variation to mean 1.0 and standard deviation of 0.33 and estimated relative substitution rates on the topology shown in Figure 2. We carried out three replicate runs, ensuring convergence and adequate mixing by inspecting samples from the MCMC in Tracer[81]. Each analysis was run for 10 million generations, sampled every 1000 generations. Rates varied by a factor of approximately 5, ranging from the lowest relative rate of 0.33 (in turtle) to a high of 1.67 (in python; see Additional file 1, Table S5).

*Anoxic gene expression.* To better understand the transcriptomic changes that might underlie the profound anoxia tolerance of the Western Painted Turtle, differential gene expression was investigated in telencephalon and ventricle from Western Painted Turtles that were either normoxic or submerged in anoxic water 24 hours at 19°C (N=4 per group, 8 total; mean ± SD 238.6±23 g, range 198-274 g) using RNA-seq methodology. At the end of the submergence period, the turtles, which appeared sedated due to profound metabolic depression, were removed from the chamber and quickly euthanized. The telencephalon was removed from the braincase, stripped of any adherent meninges, and flash-frozen in freeze-clamps previously cooled in liquid nitrogen. A 2 cm x 4 cm window was quickly cut in the plastron with a bone saw, exposing the still-beating heart, which was quickly removed, bisected, blotted on gauze to remove any blood, and quickly flash-frozen. Water was considered anoxic when oxygen concentrations were undetectable with a submerged oxygen electrode (YSI D200) while bubbling the water with nitrogen gas. Frozen tissue samples (22-109 mg) were ground to a fine powder under liquid nitrogen with a mortar and pestle and transferred to a dry-ice cooled test tube with a liquid nitrogen-cooled spatula. One milliliter of room-temperature Trizol® reagent (Life Technologies) per 50-100 mg tissue was added to the tube, which was immediately vortexed. All subsequent RNA isolation steps were performed according to the Trizol manufacturer's instructions. The final RNA pellet was resuspended in DEPC-treated water and treated with DNAse I (Life Technologies) according the manufacturer's instructions in order to remove any DNA contamination. RIN values for the samples were all greater than 7.4 (Agilent 2100 Bioanalyzer). cDNA library construction and sequencing was carried out using previously described

method[11, 82].

Paired-end 2x100bp reads generated from poly(A) selected RNA-seq libraries from all 16 samples were aligned to the latest *C. picta* assembled reference sequence, using TopHat 1.4.0[83], which also splits reads to align them across known and novel splice junctions. For known splice junction loci, a GTF (Gene Transfer Format) file of OPTIC annotations was provided. To estimate transcript and gene abundances, Cufflinks 1.3.0[84] was used. This generates normalized FPKMs (Fragments per kilobase of exon model per million mapped fragments) for each annotated gene and transcript as defined in the OPTIC based annotations. The Cufflinks parameter -G was used to exclude novel isoforms, in order to exclude large outliers (regions with extraordinarily high read-depths) that causes the Cufflinks normalization method to introduce a loss of sensitivity. The per-gene FPKMs were $\log_2$ transformed and compared across treatments and tissues by ANOVA assuming a normal/Gaussian distribution[85] with FPR multiple testing correction using JMP Genomics 5.1. Genes were excluded from the analysis if the median FPKM equaled zero for three out of the four sampling groups. The results of genes showing greater than two-fold increases are shown in Additional file 1, Tables S7, S8; down-regulated genes are shown in Additional file 1, Table S9; and RNA-seq read depths for the most highly up and down-regulated genes are shown in Additional file 2, Figures S10-S12.

*Freeze tolerance.* The Mfold (v.2.3) computer program was used to predict RNA structure[86]. The program predicts secondary structure based on the energy minimization method and thermodynamic parameters. We initially searched the *C. picta* assembly (v.3.0.1) for the sequence of premiR-29b using BLAST+ (v.2.2.18)[87]. We

focus on this micro-RNA because, in conjunction with ongoing experiments (Storey, unpublished results), we find that miR-29b increases in expression levels for many models of metabolic rate depression (hibernating mammals, freeze tolerance and anoxia tolerance). This is most likely due to its proposed role in regulating the PI3K/Akt signaling pathway, a pathway that is commonly differentially regulated in response to environmental stress and has been shown to control glucose metabolism and transport, survival (apoptosis), translation processes and cell cycle arrest. This microRNA continually proves to be a utilized regulatory response to severe environmental stresses. Small RNAs, including miRNAs, were isolated using the mirVana miRNA isolation kit from Ambion Inc. (P/N: 1560) according to the manufacturer's protocol. Samples (~100 mg) were homogenized 1:10 w:v in lysis/binding buffer, left on ice for 10 min and then a mixture of acid phenol:chloroform was added in a 1:1 ratio. Samples were centrifuged for 5 min at 10,000 ×g and the supernatant was collected. Small miRNAs (<200 nt) were isolated using the enrichment protocol provided with the kit involving two sequential filtrations through glass-fiber filters at different ethanol concentrations. RNA concentration was determined by absorbance at 260 nm and the ratio of absorbance at 260/280 nm was used as an indicator of RNA purity.

To determine the expression of mature miR-29b from *C. picta*, a modified v miRNA-specific reverse transcription and qRT-PCR procedure was performed. A 5.0 µL aliquot of small RNA (0.2 ng/µl) was incubated with 1µL of 250 nM microRNA-specific stem-loop primer (5’-CTCACAGTACGTTGGTAT CCTTGTGATGTTCGATGCCATATTGTACTGTGAGAACACTGA-3’). The reaction was heated at 95°C for 5 min to denature the RNA, and then incubated for 5 min at 60°C

to anneal the stem loop primer. After cooling on ice for 1 min, the remaining reagents (4 µL of 5x first strand buffer, 2 µL of 0.1 M DTT, 1 µL of dNTP mixture containing 25 mM of each nucleotide, and 1 µL of M-MLV reverse transcriptase) were added. The reaction proceeded for 30 min at 16°C, followed by 30 min at 42°C, and 85°C for 5 min. Following reverse transcription, the RT product was stored at -20°C. Real-time PCR was performed on a BioRad MyiQ2 Detection System (P/N: 170-9790, BioRad). The 25 µL qRT-PCR reaction included 5 µL RT product, 12.5 µL SsoFast EvaGreen Supermix (P/N: 172-5201, BioRad), 0.5 µL of 12.5 µM forward primer (5'-ACACTCCAGCTGGGTAGCACCATTTGAAATC-3'), 0.5 µL of 12.5 µM reverse primer (5'-CTCACAGTACGTTGGTATCCTTGTG-3') and 6.5 µL nuclease free water. Reactions were incubated in a 96-well plate at 95°C for 3 min, followed by 40 cycles of 95°C for 15 s and 60°C for 1 min. A melting curve analysis was performed for each miRNA analyzed. All reactions were run in triplicate.

*Tooth loss.* We initially searched the *C. picta* scaffolds for individual exons of *ENAM*, *AMEL*, *AMBN*, *DSPP*, and *MMP20* using BLAST+ (v. 2.2.18)[87]. We used crocodilian sequences for *AMEL* - AF095568, *AMBN* - AY043290, and *ENAM* - GU344683.1, MMP20 – DQ885891.1 and human sequences for *DSPP* - NM_014208 and part of *MMP20*- NM_004771.3. We were able to identify from one to several conserved exons from the *C. picta* pseudogenes, thus providing us with an anchor point for further analysis. Subsequently, we utilized UniDPlot, which is a tool for the detection of poorly conserved DNA regions and was previously used to find pseudogenes in chicken[41, 88]. Finally, we utilized T-coffee to align homologous exons and manual curation to identify GT-AG exon-intron junctions. Identity scores were calculated using LALIGN[89]. Gene

positions within chromosomally syntenic regions were analyzed using lizard (*A. carolinensis* genome assembly 2.0), chicken (*Gallus gallus* genome assembly 4.0), and the UCSC Human Genome Browser.

*Aging and longevity*. We obtained the individual exon sequences for all five shelterin complex encoding genes and for the genes in HAGR, from NCBI. We then searched all *C. picta* sequence and RNA-seq data to identify orthologs of individual exons of these genes using BLAST+ (v. 2.2.18)[87]. We used either chicken or anole (lizard) sequences as the query sequence. While *TEP1*, *TERF2IP*, and *ATP5O* were completely absent from *C. picta* genome, partial fragmented forms of the other genes were found. To avoid draft assembly artifacts, we confirmed our results by carrying out similar searches for these genes in all four turtle genome available to us (See Additional file 1, Table S12).

*Sex determination/differentiation*. 454 reads (generated by us) from the transcriptomes of *C. p. bellii*, *Chelydra serpentina*, *Apalone mutica*, and *Podocnemis expansa* were combined and mapped to *C. picta* assembly 3.0.1 using GMAP[90]. The resulting SAM file was then run through Cufflinks 1.3.0[91] to obtain a GTF file containing a single list of putative genes in the Western Painted Turtle genome. This GTF file was used in CuffLinks as the reference GTF for subsequent CuffLinks runs on each 454 dataset. cDNA sequences per tissue and species were extracted using R and the bioconductor package ShortRead from this reference GTF file. Each transcriptome 454 dataset was mapped to the *C. picta* assembly 3.0.1 using GMAP.

DNA coding sequences from the genomes or transcriptomes of multiple vertebrates of 34 genes in the sex determination/differentiation network of vertebrates or

linked to sex chromosomes in chicken (see Additional file 1, Table S13) were extracted and aligned using CLUSTALW in Geneious Pro[92] and artificial frameshifts and other errors were manually corrected. Rates of molecular evolution were evaluated by calculating dN, dS, and dN/dS per gene in MEGA5[93]. Tests of neutrality, positive and purifying selection were carried out in MEGA5 using the codon-based Z-test, using the Nei-Gojobori method[94], where the variance of the difference was computed using the bootstrap method with 500 replicates. Optimal models of DNA evolution were inferred per gene and gene-specific phylogenetic trees were built by maximum likelihood with MEGA5, and topologies contrasted among genes with the species phylogenetic relationships.

*Immune system.* A unique sequence database was generated from Ensembl[95] consisting of ~3000 immune genes from human (*Homo sapiens*, GRCh37), mouse (*Mus musculus*, NCBIM37), rat (*Rattus norvegicus*, RGSC3.4), chicken (*Gallus gallus*, WASHUC2), Fugu (*Takifugu rubripes*, FUGU4), Medaka (*Oryzias latipes*, MEDAKA1), Anole (*Anolis carolinensis*, AnoCar2.0), Stickleback (*Gasterosteus aculeatus*, BROADS1), Turkey (*Meleagris gallopavo*, UMD2), Xenopus (*Xenopus tropicalis*, JGI_4.2), Tetraodon (*Tetraodon nigroviridis*, TETRAODON8), Zebrafinch (*Taeniopygia guttata*, taeGut3.2.4), Zebrafish (*Danio rerio*, Zv9) and Sea Lamprey (*Petromyzon marinus*, Pmarinus_7.0). Sequences, Ensembl Gene ID and Gene Name were obtained from Ensembl directly or using the Biomart mining utility[96] when available. Pairwise alignments were obtained using in-house BLAST (BLASTN 2.2.15)[97] comparing query immune gene sequences to the *C. picta* genome assembly and unassembled sequencing reads, gene predictions, and cDNA reads.

*Identification of gene family expansion/contraction*. To identify gene family expansions and contractions, we built phylogenetic trees for all predicted genes in *C. picta* with their orthologs in human, mouse, platypus, chicken, zebrafinch, green anole and using the pufferfish and zebrafish as outgroups.

Orthology assignments and orthologous groups were defined using the OPTIC pipeline[10, 98]. Orthology assignments are based upon the computation of pairwise orthologs using PhyOP[99] using BLASTP searches with an E-value threshold of $10^{-5}$ and a minimum size cut-off equal to 75% of the smaller sequence. The alignments were weighted according to the normalized bit score:

$s_{ij}=1-((\max[s'ij,s'ji])/\min(s'ij, sji).$

Where $s'_{ij}$ is the bit score for a BLASTP alignment between sequence i and j.

A tree based orthology method implemented within PhyOP[99] was used to define clusters of orthologous groups. For each cluster, genes were aligned using MUSCLE[100], genes with multiple transcripts were collapsed into sequences of non-redundant exons and phylogenetic trees were estimated using TreeBeST[101]. Rates of non-synonymous substitutions per non-synonymous sites ($d_N$) and rates of synonymous substitutions per synonymous sites ($d_S$) and their ratio ($d_N/d_S$) were estimated for each branch of the tree with PAML[102]. Rates were not allowed to vary between sites. To remove biases associated with poor alignments, translated sequences were masked with SEG[103] and corresponding masked codons were removed; poorly aligned columns were also removed using Gblocks[104].

A total of 20,234 orthologous groups were found, of which 12,938 have at least one gene prediction from *C. picta* and 1,176 groups contain at least two *C. picta* gene models. All orthology / paralogy predictions are available at[105]. We identified a total of 4,828 genes with one-to-one orthologous relationship between all amniotes, and 3,222 between all species when pufferfish and zebrafish are included. A total of 604 predicted gene models in *C. picta* had no predicted orthologs; these include rapidly-evolving genes as well as problematic gene models that survived our filters. We also identified 568 groups with genes in human, mouse, platypus, chicken, zebrafinch and green anole but that have no detectable orthologs in the current version of the *C. picta* genome assembly. These currently absent genes will contain genes absent from the current assembly, as well as rapidly-evolving genes.

In order to reach a conservative estimate of the number of genes within a family and to remove any residual biases associated with the assembly process, we estimated the pairwise amino acid identity between every pair of members of a family and rejected duplicated genes that are more than 97% identical. A summary of gene expansions is presented in Figure 6.

*Beta-keratin expansions*. Beta keratins have previously been described to be an important component of the corneous layers of the reptilian epidermis forming the scales, claws and beak. In birds, they are the major component of feathers[106]. We identified a total of 106 gene models (101 complete) in *C. picta* that share significant sequence similarity with avian and green anole beta-keratins. Using beta-keratin mRNAs extracted from the precursor cells of the shell of *Pseudemys nelsoni*[56], and the phylogeny built

with PhyML[107], we identified 41 and 60 putative non-shell and shell proteins in *C. picta* respectively.

*Expansion of gene families involved in the immune response*. Among the families with the largest expansions (≥ 4 members), 15 are related to the innate or adaptive immune response.

As part of the adaptive immune response, we identified 365, 131, and 94 predicted gene models in *C. picta* that cluster with the immunoglobulin heavy chain, lambda, and kappa chain variable regions respectively in mouse. The large number of genes from these two families is of prime importance in the generation of antibody diversity through V(D)J recombination. Both the immunoglobulin heavy and light chain variable regions are known to be among the most dynamic gene regions in the human genome, and immunoglobulin genes are known to show high allelic and copy number variation[108, 109]. Interestingly the imunoglobulin kappa chains have been lost in the bird genomes[110]. These authors predicted this loss to predate the divergence between Passeriformes and Galliformes (100 Mya). In agreement with this, our analysis show that the immunoglobulin kappa chains were present in the common ancestor of the birds and turtles ~260 Mya.

We also identified expansions of several gene families that form part of the innate immune system. These gene products are expressed on the surface of natural killer (NK) cells (NK cells' C-type lectin-like and NK cells' immunoglobulin-like receptors) or are secreted by these NK cells (for example, granzymes). NK receptors previously shown to belong to the LCR in human, mouse and chicken are known to have undergone lineage-specific expansion in each of these lineages[111-114]. We searched the *C. picta*

polypeptide predictions belonging to these two families for transmembrane domains[115] and found that only 6 of 27 putative NK cells' C-type lectin-like and 14 of 35 putative NK cells' immunoglobulin-like receptors possess transmembrane domains.

*Ortholog sets.* We based our study of positive selection on the set of carefully screened orthologs for eight vertebrate species (human, platypus, chicken, zebrafinch, anole, turtle, python, and alligator, see Materials and Methods, *Multiple Alignments and Gene Orthologs*). From among 4786 high-confidence ortholog sets, each covering between 2 and 8 species, we selected 4136 sets that covered human, turtle, and at least one of the outgroup genomes (chicken, alligator, zebrafinch).

*Positive selection.* We detected signs of positive selection using likelihood ratio tests[57] with reduced parameterization[58]. P-values were estimated assuming a null distribution that is a 50:50 mixture of $\chi 2$ distribution with one degree of freedom, and a point mass at zero, leading to conservative P-value estimates[116]. The branch leading to the turtle was designated as a forward branch, with some sites allowing dN/dS>1, while all other branches were background branches, disallowing positive selection. The results were corrected for multiple testing using Benjamini and Hochberg false discovery rate control (FDR). Accessory Data File 1 shows the results of likelihood-ratio tests for all genes with nominal P-values < 0.05 (890 genes), indicating genes with FDR<0.1 (671 genes).

We also examined GO functional categories for enrichment for genes under positive selection, using Mann-Whitney U-test with Holm's correction for multiple testing[117]. No functional categories were statistically significantly enriched for genes

under positive selection after multiple testing correction. Accessory Data File 1 shows 171 GO categories with nominal P-values < 0.05.

## Competing interests

There are no competing financial interests associated with this work.

## Authors' contributions

HBS, PM, AMS, RCT and NV comprise the organizing committee of the Western Painted Turtle genome sequencing project. Manuscript organization and editing: HBS, PQS, RCT, WCW, CPP, WH and PM. BAC library construction: CTA. Project management and data production: LF, KDD, CCF, MO and TAG. Assembly and analysis: PM, NT and LWH. RNA-seq and anoxia analysis: DEW, CK and LTB. Freeze tolerance: KBS and KKB. Gene model predictions, orthology prediction and analysis: CPP, WH, LK and YL. Repeat element and isochore analysis: AMS, CWB and MKF. Comparative alignments: BJR. Phylogenetic analyses: RCT, TV, TAC, DDP, APJK, REG, JSt.J and ELB. Gene analyses, sex chromosome genes: FJJ, SEM, AMB, TS, AS, NV, RL, DB, SR, DJ, SVE, BC, MC, OH and LM. Gene analyses, aging: RH. Gene analyses, immuno-response: RMB, GMB, LMZ and RTP. Gene analyses, enamel: JA and JMR. Turtle accelerated regions: AKH and BGB. Principal investigators: ERM, WCW and RKW. All authors read and approved the final manuscript.

## Author information

The *Chrysemys picta bellii* whole-genome shotgun project has been deposited in NCBI GenBank under the project accession AHGY00000000. The raw input data for *Chrysemys picta bellii* (BioProject ID: 78657) was deposited to the trace archive and the SRA under the project accession SRP012057. The *Apalone spinifera* whole genome data can be found at the NCBI SRA under the accession numbers SRX217616-7. Specimen collection for the *C. p. bellii* was authorized by the Washington Department of Fish and Wildlife under scientific collecting permit 08-086 (to R.C.T.) and complied with IACUC standards at UC Davis (H.B.S. protocol holder). Ethical (IACUC) approvals for all experiments involving living turtles were obtained at the university where the experiment or field work were conducted. Correspondence: Brad Shaffer, brad.shaffer@ucla.edu; Pat Minx, pminx@genome.wustl.edu.

## Acknowledgements

The sequencing of the Western Painted Turtle genome was funded by the National Human Genome Research Institute (NHGRI). Further research support included grants to HBS (NSF DEB 0817042) NV (NSF IOS 0743284), NV & SVE (MCB 0815354), FJJ and AS (NSF DEB 0640932), KBS and LTB **(**NSERC), TV (FP7 IRG-224885, VEGA 1/1085/12), and SVE, CTA and JR Macey (NSF IOS 0207870/0431717). Resources for exploring the sequence and annotation data are available on browser displays available at UCSC[118], Ensembl[119], NCBI[120], and MRC[105]. We thank Andrew Severin (Iowa State Genome Informatics Facility) for help with data analyses, Erik Larson (Illinois State) for contributing Glen Borchert's time to the project, The

Genome Institute members and Michael Montague for manuscript review, Louise Whitaker (UCLA) for bibliographic assistance, and Bronwen Aken (Wellcome Trust Sanger Institute) for displaying the genome annotation on ENSEMBL. During the development of this project we received useful input from Naoki Irie (RIKEN) and Guojie Zhang (BGI). We also thank the following members of The Genome Institute: Michael Montague for manuscript review, Chad Tomlinson for assembling cDNAs, Bob Fulton and Aye Wollam for finishing BACs used in assembly quality assessment.

## References

1. Romer AS: *Vertebrate paleontology 3rd Ed.*: Univ. Chicago Press; 1967.
2. Li C, Wu X-C, Rieppel O, Wang L-T, Zhao L-J: **An ancestral turtle from the Late Triassic of southwestern China.** *Nature* 2008, **456:**497-501.
3. Gaffney ES, Jenkins FA: **The cranial morphology of Kayentachelys, an Early Jurassic cryptodire, and the early history of turtles.** *Acta Zoologica* 2010, **91:**335-368.
4. Ultsch GR, Jackson DC: **Long-term submergence at 3-degrees-C of the turtle, *Chrysemys picta bellii*, in normoxic and severely hypoxic water. 1. Survival, gas-exchange and acid-base status.** *J Exp Biol* 1982, **96:**11-28.
5. Johlin JM, Moreland FB: **Studies of the blood picture of the turtle after complete anoxia.** *J Biol Chem* 1933, **103:**107-114.
6. Storey KB, Storey JM, Brooks SPJ, Churchill TA, Brooks RJ: **Hatchling turtles survive freezing during winter hibernation.** *Proceedings of the National Academy of Sciences of the United States of America* 1988, **85:**8350-8354.

7. Keenan NL, Shaw KM: **Coronary Heart Disease and Stroke Deaths - United States, 2006.** *Morb Mortal Weekly Rep* 2011, **60:**62-66.

8. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0.** 1996-2010.

9. Alföldi J, Di Palma F, Grabherr M, Williams C, Kong L, Mauceli E, Russell P, Lowe CB, Glor RE, Jaffe JD, Ray DA, Boissinot S, Shedlock AM, Botka C, Castoe TA, Colbourne JK, Fujita MK, Moreno RG, ten Hallers BF, Haussler D, Heger A, Heiman D, Janes DE, Johnson J, de Jong PJ, Koriabine MY, Lara M, Novick PA, Organ CL, Peach SE, et al: **The genome of the green anole lizard and a comparative analysis with birds and mammals.** *Nature* 2011, **477:**587-591.

10. Heger A, Ponting CP: **Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes.** *Genome Res* 2007, **17:**1837-1849.

11. Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S, Heger A, Kong LS, Ponting CP, Jarvis ED, Mello CV, Minx P, Lovell P, Velho TAF, Ferris M, Balakrishnan CN, Sinha S, Blatti C, London SE, Li Y, Lin YC, George J, Sweedler J, Southey B, Gunaratne P, Watson M, et al: **The genome of a songbird.** *Nature* 2010, **464:**757-762.

12. Gilbert C, Hernandez SS, Flores-Benabib J, Smith EN, Feschotte C: **Rampant Horizontal Transfer of SPIN Transposons in Squamate Reptiles.** *Mol Biol Evol* 2012, **29:**503-515.

13. Shedlock AM, Botka CW, Zhao SY, Shetty J, Zhang TT, Liu JS, Deschavanne PJ, Edward SV: **Phylogenomics of nonavian reptiles and the structure of the**

**ancestral amniote genorne.** *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104:**2767-2772.

14. Feschotte C: **Opinion - Transposable elements and the evolution of regulatory networks.** *Nature Reviews Genetics* 2008, **9:**397-405.

15. Janes DE, Organ CL, Fujita MK, Shedlock AM, Edwards SV: **Genome Evolution in Reptilia, the Sister Group of Mammals.** In *Annual Review of Genomics and Human Genetics, Vol 11. Volume* 11. Edited by Chakravarti A, Green E; 2010: 239-264: *Annual Review of Genomics and Human Genetics*].

16. Kazazian HH: **Mobile elements: Drivers of genome evolution.** *Science* 2004, **303:**1626-1632.

17. Shedlock AM: **Phylogenomic investigation of CR1 LINE diversity in reptiles.** *Syst Biol* 2006, **55:**902-911.

18. Fujita MK, Edwards SV, Ponting CP: **The Anolis Lizard Genome: An Amniote Genome without Isochores.** *Genome Biology and Evolution* 2011, **3:**974-984.

19. Galtier N, Gouy M: **Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis.** *Mol Biol Evol* 1998, **15:**871-879.

20. Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC: **More than 1000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs.** *Biol Lett* 2012.

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

22. 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**.

23. Bromham L: **Molecular clocks in reptiles: Life history influences rate of molecular evolution.** *Mol Biol Evol* 2002, **19:**302-309.

24. Sayres MAW, Venditti C, Pagel M, Makova KD: **Do variations in substitution rates and male mutation bias correlate with life-history traits? A study of 32 mammalian genomes.** *Evolution* 2011, **65:**2800-2815.

25. Regard JB, Scheek S, Borbiev T, Lanahan AA, Schneider A, Demetriades AM, Hiemisch H, Barnes CA, Verin AD, Worley PF: **Verge: A novel vascular early response gene.** *J Neurosci* 2004, **24:**4092-4103.

26. van Dam H, Castellazzi M: **Distinct roles of Jun : Fos and Jun : ATF dimers in oncogenesis.** *Oncogene* 2001, **20:**2453-2464.

27. Passegue E, Wagner EF: **JunB suppresses cell proliferation by transcriptional activation of p16(INK4a) expression.** *EMBO J* 2000, **19:**2969-2979.

28. Winkler GS: **The Mammalian Anti-Proliferative BTG/Tob Protein Family.** *Journal of Cellular Physiology* 2010, **222:**66-72.

29. Levin WJ, Press MF, Gaynor RB, Sukhatme VP, Boone TC, Reissmann PT, Figlin RA, Holmes EC, Souza LM, Slamon DJ: **Expression patterns of immediate-early transcription factors in human nonsmall cell lung-cancer.** *Oncogene* 1995, **11:**1261-1269.

30. Blackstone E, Morrison M, Roth MB: **H2S induces a suspended animation-like state in mice.** *Science* 2005, **308:**518-518.

31. Ueki I, Roman HB, Valli A, Fieselmann K, Lam J, Peters R, Hirschberger LL, Stipanuk MH: **Knockout of the murine cysteine dioxygenase gene results in severe impairment in ability to synthesize taurine and an increased catabolism of cysteine to hydrogen sulfide.** *American Journal of Physiology-Endocrinology and Metabolism* 2011, **301:**E668-E684.

32. Long JC, Caceres JF: **The SR protein family of splicing factors: master regulators of gene expression.** *Biochem J* 2009, **417:**15-27.

33. Wheater MJ, Johnson PW, Blaydes JP: **The role of MNK proteins and eIF4E phosphorylation in breast cancer cell proliferation and survival.** *Cancer Biology & Therapy* 2010, **10:**728-735.

34. Storey KB: **Reptile freeze tolerance: Metabolism and gene expression.** *Cryobiology* 2006, **52:**1-16.

35. Biggar KK, Storey KB: **The emerging roles of microRNAs in the molecular responses of metabolic rate depression.** *Journal of Molecular Cell Biology* 2011, **3:**167-175.

36. Garzon R, Liu SJ, Fabbri M, Liu ZF, Heaphy CEA, Callegari E, Schwind S, Pang JX, Yu JH, Muthusamy N, Havelange V, Volinia S, Blum W, Rush LJ, Perrotti D, Andreeff M, Bloomfield CD, Byrd JC, Chan K, Wu LC, Croce CM, Marcucci G: **MicroRNA-29b induces global DNA hypomethylation and tumor suppressor gene reexpression in acute myeloid leukemia by targeting directly DNMT3A and 3B and indirectly DNMT1.** *Blood* 2009, **113:**6411-6418.

37. He AB, Zhu LB, Gupta N, Chang YS, Fang F: **Overexpression of micro ribonucleic acid 29, highly up-regulated in diabetic rats, leads to insulin resistance in 3T3-L1 adipocytes.** *Mol Endocrinol* 2007, **21:**2785-2794.

38. Zhang XX, Zeng Y: **The terminal loop region controls microRNA processing by Drosha and Dicer.** *Nucleic Acids Res* 2010, **38:**7689-7697.

39. Davit-Beal T, Tucker AS, Sire JY: **Loss of teeth and enamel in tetrapods: fossil record, genetic data and morphological adaptations.** *J Anat* 2009, **214:**477-501.

40. Meredith RW, Gatesy J, Cheng J, Springer MS: **Pseudogenization of the tooth gene enamelysin (MMP20) in the common ancestor of extant baleen whales.** *Proceedings Of The Royal Society B-Biological Sciences* 2011, **278:**993-1002.

41. Sire JY, Delgado SC, Girondot M: **Hen's teeth with enamel cap: from dream to impossibility.** *BMC Evol Biol* 2008, **8**.

42. Kim EB, Fang XD, Fushan AA, Huang ZY, Lobanov AV, Han LJ, Marino SM, Sun XQ, Turanov AA, Yang PC, Yim SH, Zhao X, Kasaikina MV, Stoletzki N, Peng CF, Polak P, Xiong ZQ, Kiezun A, Zhu YB, Chen YX, Kryukov GV, Zhang Q, Peshkin L, Yang L, Bronson RT, Buffenstein R, Wang B, Han CL, Li QY, Chen L, et al: **Genome sequencing reveals insights into physiology and longevity of the naked mole rat.** *Nature* 2011, **479:**223-227.

43. de Magalhaes JP, Costa J, Toussaint O: **HAGR: the human ageing genomic resources.** *Nucleic Acids Res* 2005, **33:**D537-D543.

44. Dillin A, Hsu AL, Arantes-Oliveira NA, Lehrer-Graiwer J, Hsin H, Fraser AG, Kamath RS, Ahringer J, Kenyon C: **Rates of behavior and aging specified by mitochondrial function during development.** *Science* 2002, **298:**2398-2401.

45. Undie AS, Wang HY, Friedman E: **Decreased phospholipase C-beta immunoreactivity, phosphoinositide metabolism, and protein-kinase-C activation in senescent F344 rat-brain.** *Neurobiol Aging* 1995, **16:**19-28.

46. Janzen FJ, Krenz JG: *Phylogenetics: which was first, TSD of GSD?*; 2004.

47. Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, Rice WR, Valenzuela N: **Are all sex chromosomes created equal?** *Trends Genet* 2011, **27:**350-357.

48. Valenzuela N, Neuwald JL, Literman R: **Transcriptional evolution underlying vertebrate sexual development.** *Dev Dyn* 2012.

49. Valenzuela N: **Multivariate Expression Analysis of the Gene Network Underlying Sexual Development in Turtle Embryos with Temperature-Dependent and Genotypic Sex Determination.** *Sexual Development* 2010, **4:**39-49.

50. Barley AJ, Spinks PQ, Thomson RC, Shaffer HB: **Fourteen nuclear genes provide phylogenetic resolution for difficult nodes in the turtle tree of life.** *Mol Phylogen Evol* 2010, **55:**1189-1194.

51. Krenz JG, Naylor GJP, Shaffer HB, Janzen FJ: **Molecular phylogenetics and evolution of turtles.** *Mol Phylogen Evol* 2005, **37:**178-191.

52. Shaffer HB, Meylan P, McKnight ML: **Tests of turtle phylogeny: Molecular, morphological, and paleontological approaches.** *Syst Biol* 1997, **46:**235-268.

53. Grey HM: **Phylogeny of immune response-studies on some physical chemical serologic characteristics of antibody produced in turtle.** *J Immunol* 1963, **91:**819-&.

54. Zimmerman LM, Vogel LA, Bowden RM: **Understanding the vertebrate immune system: insights from the reptilian perspective.** *J Exp Biol* 2010, **213:**661-671.

55. Brownlie R, Allan B: **Avian toll-like receptors.** *Cell Tissue Res* 2011, **343:**121-130.

56. Dalla Valle L, Nardi A, Toni M, Emera D, Alibardi L: **Beta-keratins of turtle shell are glycine-proline-tyrosine rich proteins similar to those of crocodilians and birds.** *J Anat* 2009, **214:**284-300.

57. Yang ZH, Nielsen R: **Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages.** *Mol Biol Evol* 2002, **19:**908-917.

58. Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A: **Patterns of Positive Selection in Six Mammalian Genomes.** *PLoS Genet* 2008, **4**.

59. Kohroki J, Nishiyama T, Nakamura T, Masuho Y: **ASB proteins interact with Cullin5 and Rbx2 to form E3 ubiquitin ligase complexes.** *FEBS Lett* 2005, **579:**6796-6802.

60. Murdoch C, Finn A: **Chemokine receptors and their role in inflammation and infectious diseases.** *Blood* 2000, **95:**3032-3043.

61. Kopp A: **Dmrt genes in the development and evolution of sexual dimorphism.** *Trends Genet* 2012, **28:**175-184.

62. Schwartz TS, Bronikowski AM: **Molecular stress pathways and the evolution of reproduction and aging in reptiles** In *Molecular Mechanisms of Life History Evolution.* Edited by Flatt T, Heyland A. Oxford, UK: Oxord University Press; 2011 193-209

63. Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D: **Estimates of Positive Darwinian Selection Are Inflated by Errors in Sequencing, Annotation, and Alignment.** *Genome Biology and Evolution* 2009, **1:**114-118.

64. Turtle Taxonomy Working Group: **Turtles of the World, 2011 Update: Annotated Checklist of Taxonomy, Synonymy, Distribution, and Conservation Status.** In *Chelonian Res Monogr* (van Dijk PP, Iverson JB, Shaffer HB, Bour R, Rhodin AGJ eds.), vol. 5. pp. 78pp.; 2011:78pp.

65. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.

66. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data P: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

67. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, Miller W: **Human-mouse alignments with BLASTZ.** *Genome Res* 2003, **13:**103-107.

68. Kent WJ, Baertsch R, Hinrichs A, Miller W, Haussler D: **Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes.**

*Proceedings of the National Academy of Sciences of the United States of America* 2003, **100:**11484-11489.

69. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8:**175-185.

70. Kent WJ: **BLAT - The BLAST-like alignment tool.** *Genome Res* 2002, **12:**656-664.

71. Shin H, Hirst M, Bainbridge MN, Magrini V, Mardis E, Moerman DG, Marra MA, Baillie DL, Jones SJM: **Transcriptome analysis for Caenorhabditis elegans based on novel expressed sequence tags.** *BMC Biol* 2008, **6**.

72. Hillier LW, Miller W, Birney E, Warren W, Hardison RC, Ponting CP, Bork P, Burt DW, Groenen MAM, Delany ME, Dodgson JB, Chinwalla AT, Cliften PF, Clifton SW, Delehaunty KD, Fronick C, Fulton RS, Graves TA, Kremitzki C, Layman D, Magrini V, McPherson JD, Miner TL, Minx P, Nash WE, Nhan MN, Nelson JO, Oddy LG, Pohl CS, Randall-Maher J, et al: **Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution.** *Nature* 2004, **432:**695-716.

73. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27:**573-580.

74. Mayer C: **Phobos 3.3.11.** 2006-2010.

75. **https://github.com/hmsrc/RMPipeline.**

76. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J,

Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, et al: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921.

77. Dreszer TR, Karolchik D, Zweig AS, Hinrichs AS, Raney BJ, Kuhn RM, Meyer LR, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Pohl A, Malladi VS, Li CH, Learned K, Kirkup V, Hsu F, Harte RA, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Diekhans M, Cline MS, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: extensions and updates 2011.** *Nucleic Acids Res* 2012, **40:**D918-D923.

78. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14:**708-715.

79. Ronquist F, Huelsenbeck JP: **MrBayes 3: Bayesian phylogenetic inference under mixed models.** *Bioinformatics* 2003, **19:**1572-1574.

80. Drummond AJ, Rambaut A: **BEAST: Bayesian evolutionary analysis by sampling trees.** *BMC Evol Biol* 2007, **7:**214.

81. Rambaut A, Drummond AJ: **Tracer v1.4, Available from http://beast.bio.ed.ac.uk/Tracer.** 2007.

82. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA.** *Nucleic Acids Res* 2009, **37**.

83. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105-1111.

84. Roberts A, Pimentel H, Trapnell C, Pachter L: **Identification of novel transcripts in annotated genomes using RNA-Seq.** *Bioinformatics* 2011, **27:**2325-2329.

85. Lu C, King RD: **An investigation into the population abundance distribution of mRNAs, proteins, and metabolites in biological systems.** *Bioinformatics* 2009, **25:**2020-2027.

86. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic Acids Res* 2003, **31:**3406-3415.

87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.

88. Girondot M, Sire JY: **UniDPlot: A software to detect weak similarities between two DNA sequences.** *Journal of Bioinformatics and Sequence Analysis* 2010, **2:**69-74.

89. Huang XQ, Miller W: **A time-efficient, linear-space local similarity algorithm.** *Advances in Applied Mathematics* 1991, **12:**337-357.

90. Wu TD, Watanabe CK: **GMAP: a genomic mapping and alignment program for mRNA and EST sequences.** *Bioinformatics* 2005, **21:**1859-1875.

91. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511-U174.

92. Biomatters: **Geneious Pro, Available from http://www.geneious.com/.**

93. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods.** *Mol Biol Evol* 2011, **28:**2731-2739.

94. Nei M, Gojobori T: **Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions.** *Mol Biol Evol* 1986, **3:**418-426.

95. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, Down T, Durbin R, Fernandez-Suarez XM, Flicek P, Graf S, Hammond M, Herrero J, Howe K, Iyer V, Jekosch K, Kahari A, Kasprzyk A, Keefe D, Kokocinski F, Kulesha E, London D, Longden I, Melsopp C, Meidl P, Overduin B, et al: **Ensembl 2006.** *Nucleic Acids Res* 2006, **34:**D556-D561.

96. Haider S, Ballester B, Smedley D, Zhang JJ, Rice P, Kasprzyk A: **BioMart Central Portal-unified access to biological data.** *Nucleic Acids Res* 2009, **37:**W23-W27.

97. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.

98. Heger A, Ponting CP: **OPTIC: orthologous and paralogous transcripts in clades.** *Nucleic Acids Res* 2008, **36:**D267-D270.

99. Goodstadt L, Ponting CP: **Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human.** *PLoS Comp Biol* 2006, **2:**1134-1150.

100. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res* 2004, **32:**1792-1797.

101. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E: **EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates.** *Genome Res* 2009, **19:**327-335.

102. Yang ZH: **PAML 4: Phylogenetic analysis by maximum likelihood.** *Mol Biol Evol* 2007, **24:**1586-1591.

103. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Computer Methods for Macromolecular Sequence Analysis* 1996, **266:**554-571.

104. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540-552.

105. **http://genserv.anat.ox.ac.uk/clades/vertebrates_cpicta2/**

106. Fraser RDB, Parry DAD: **Molecular packing in the feather keratin filament.** *Journal of Structural Biology* 2008, **162:**1-13.

107. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O: **New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.** *Syst Biol* 2010, **59:**307-321.

108. Matsuda F, Ishii K, Bourvagnet P, Kuma K, Hayashida H, Miyata T, Honjo T: **The complete nucleotide sequence of the human immunoglobulin heavy chain variable region locus.** *J Exp Med* 1998, **188:**2151-2162.

109. Watson CT, Breden F: **The immunoglobulin heavy chain locus: genetic variation, missing data, and implications for human disease.** *Genes Immun* In Press.

110. Das S, Mohamedy U, Hirano M, Nei M, Nikolaidis N: **Analysis of the Immunoglobulin Light Chain Genes in Zebra Finch: Evolutionary Implications.** *Mol Biol Evol* 2010, **27:**113-120.

111. Barten R, Torkar M, Haude A, Trowsdale J, Wilson MJ: **Divergent and convergent evolution of NK-cell receptors.** *Trends Immunol* 2001, **22:**52-57.

112. Laun K, Coggill P, Palmer S, Sims S, Ning ZM, Ragoussis J, Volpi E, Wilson N, Beck S, Ziegler A, Volz A: **The leukocyte receptor complex in chicken is characterized by massive expansion and diversification of immunoglobulin-like loci.** *PLoS Genet* 2006, **2:**707-718.

113. Nikolaidis N, Makalowska I, Chalkia D, Makalowski W, Klein J, Nei M: **Origin and evolution of the chicken leukocyte receptor complex.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102:**4057-4062.

114. Viertlboeck BC, Gick CM, Schmitt R, Du Pasquier L, Gobel TW: **Complexity of expressed CHIR genes.** *Developmental and Comparative Immunology* 2010, **34:**866-873.

115. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL: **Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes.** *J Mol Biol* 2001, **305:**567-580.

116. Zhang JZ, Nielsen R, Yang ZH: **Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level.** *Mol Biol Evol* 2005, **22:**2472-2479.

117. Holm S: **A simple sequentially rejective multiple test procedure.** *Scandinavian Journal of Statistics* 1979, **6:**65-70.

118. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D: **The human genome browser at UCSC.** *Genome Res* 2002, **12:**996-1006.

119. **http://pre.ensembl.org/index.html**

120. **http://www.ncbi.nlm.nih.gov**

**Description of additional files**

We provide two Additional files, both in Microsoft word (.docx). Additional file 1 contains Tables S1-S17), and Additional file 2 contains Figures S1-S12). We also provide Additional file 3 in support of the selection scan analysis in Microsoft Excel format. In it, the sheet labeled **mwu_turtle_go** details enrichment of GO categories for positive selection on the Western Painted Turtle lineage with nominal P-values < 0.05 (Mann-Whitney U-test). No categories were significantly enriched for positive selection after application of multiple testing correction. The second sheet, labeled **lrt_turtle,** shows genes under positive selection on the Western Painted Turtle lineage. All genes with nominal P-values < 0.05 (likelihood ratio branch-site test) are shown, and the 671 genes

that were statistically significant after applying multiple testing correction (FDR < 0.1) are also noted.

**Figure legends**

**Figure 1. Standard deviation of GC content at different spatial scales.** Genomes were partitioned into non-overlapping windows (5-, 20-, 80-, and 320-kb). As window size increases, variation in GC content naturally decreases. The Western Painted Turtle exhibits a pattern consistent with high variation in nucleotide composition at smaller scales, rather than sustained isochoric variation at larger scales seen in mammals and birds. The expected pattern of decreasing standard variation assumes a compositionally homogeneous genome with a mean GC proportion of 0.41.

**Figure 2**. **A revised phylogeny of major amniote lineages and their rates of molecular evolution.** a) Bayesian phylogram depicting the relationships of the eight primary amniote lineages, and their rates of molecular evolution. The phylogeny demonstrates the sister group relationship of turtle and archosaurs (allligator plus birds). The numbers at nodes denote posterior probabilities (all are at the maximum of 1.0). b) The histogram shows the relative rate of substitution inferred for each lineage under a relaxed clock. For analysis details, see Materials and Methods, *Phylogeny and substitution rate*).

**Figure 3. Western painted turtle miR-29b and response to freezing.** a) Nucleotide sequence and predicted secondary structure of pre-miR-29b transcripts from *H. sapiens*,

*A. spinifera* and *C. picta bellii* at 25 C. Nucleotide substitution which leads to differential terminal stem-loop formation that is unique to *C. picta bellii* is circled. b) Relative expression levels of miR-29b as assessed by quantitative RT-PCR in liver samples of hatchling Western Painted Turtles under control (5°C acclimated), 24 h frozen (at -2.5°C) or 4 h thawed (at 5°C) conditions. Data are means ± s.e.m. (n = 5 different animals). Parallel analysis of 5S rRNA found no significant changes between control and experimental conditions for this reference RNA. * Significantly different from the corresponding control ($P < 0.05$).

**Figure 4. Conserved syntenic regions containing tooth-specific genes across toothed (human, anole) and edentulous (turtle, chicken) vertebrates.** *AMBN* and *ENAM* are in a reptile-specific chromosomal region, precluding the use of human as a reference sequence for these genes. Dashed outlines indicate pseudogenization.

**Figure 5**. **Maximum likelihood estimates of the phylogenetic relationships among taxa for five genes involved in gonadogenesis.** Branch lengths are proportional to the number of substitutions per site; numbers at nodes are bootstrap proportions based on 500 pseudoreplicates. Colored branches denote the taxonomic group for each taxon. Tip font colors denote sex-determining mechanisms (red = TSD, gray = GSD). For all species, the full coding region was utilized except where only partial sequences were available, in which case the tip is denoted as (P).

**Figure 6. Gene families showing expansion in the Western Painted Turtle lineage.**

The number of genes within a family is provided in front of each bar. Gene families associated with the immune response are shown in red.

**Additional file legends**

**Additional file 1. Supplementary tables.** Tables S1-S17 contain additional information in support of the painted turtle assembly (Tables S1-S2), transposable elements (Table S3), isochores (Table S4), phylogeny and evolutionary rates (Tables S5-S6), anoxia (Tables S7-S9), tooth loss (Tables S10-S11), longevity (Table S12), sex determination (Table S13), immune function (Tables S14-S15), and gene family expansions (Tables S16-S17).

**Additional file 2. Supplementary figures.** Figures S1-S12 contain additional information in support of the painted turtle repeat analyses (Figures S1-S6), isochores (Figures S7-S9), and anoxia tolerance (Figures S10-12).

**Additional file 3. Selection scan analysis**. The sheet labeled **mwu_turtle_go** provides additional information on the enrichment of GO categories for positive selection on the Western Painted Turtle, while the sheet labeled **lrt_turtle** lists all genes under positive selection on the Western Painted Turtle lineage. All genes with nominal P-values < 0.05 (likelihood ratio branch-site test) are shown, and the 671 genes that were statistically significant after applying multiple testing correction (FDR < 0.1) are also noted.

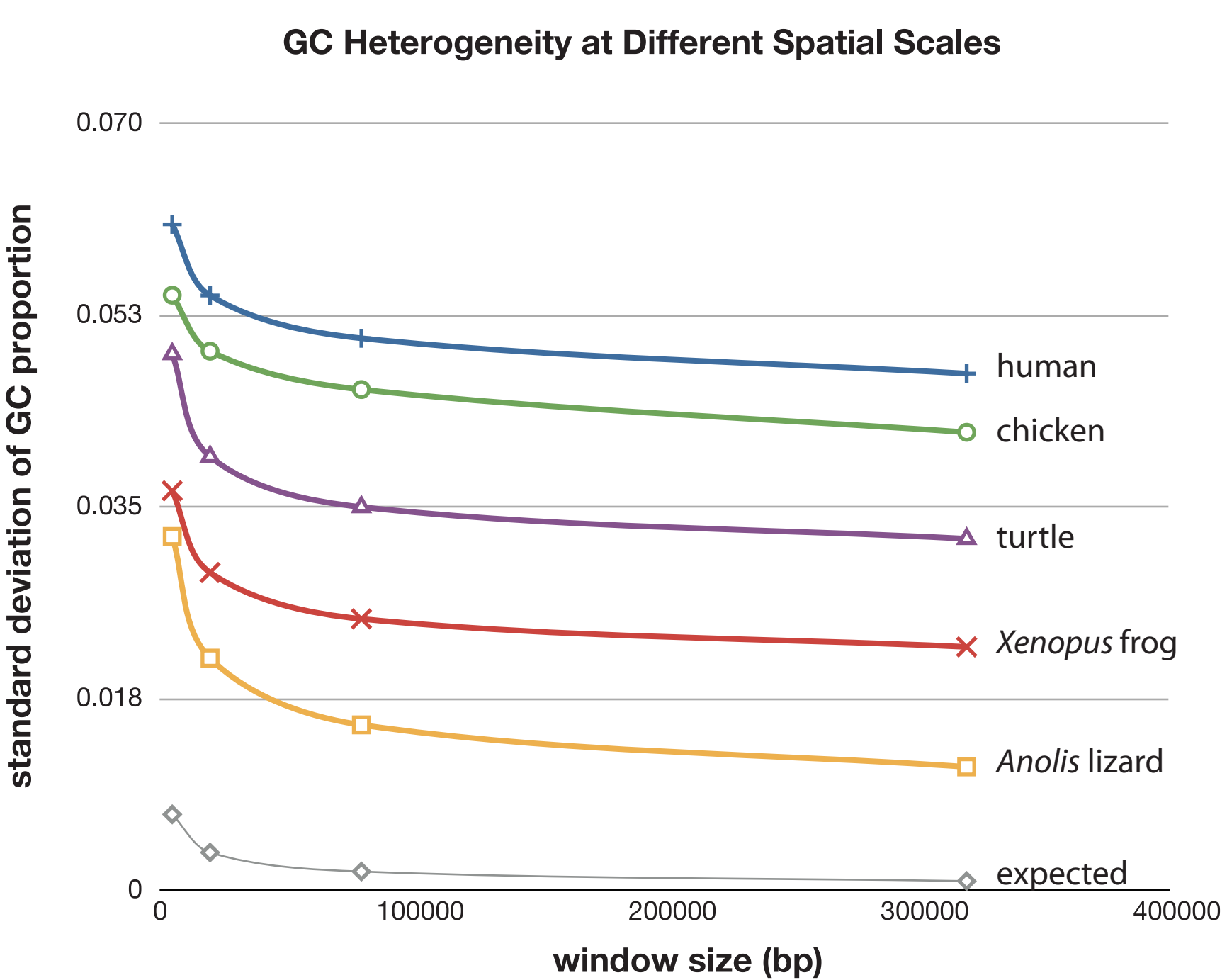

GC Heterogeneity at Different Spatial Scales
standard deviation of GC proportion
0.070
0.053
0.035
0.018
0
0
100000
200000
300000
400000
window size (bp)
human
chicken
turtle
Xenopus frog
Anolis lizard
expected

Figure 1

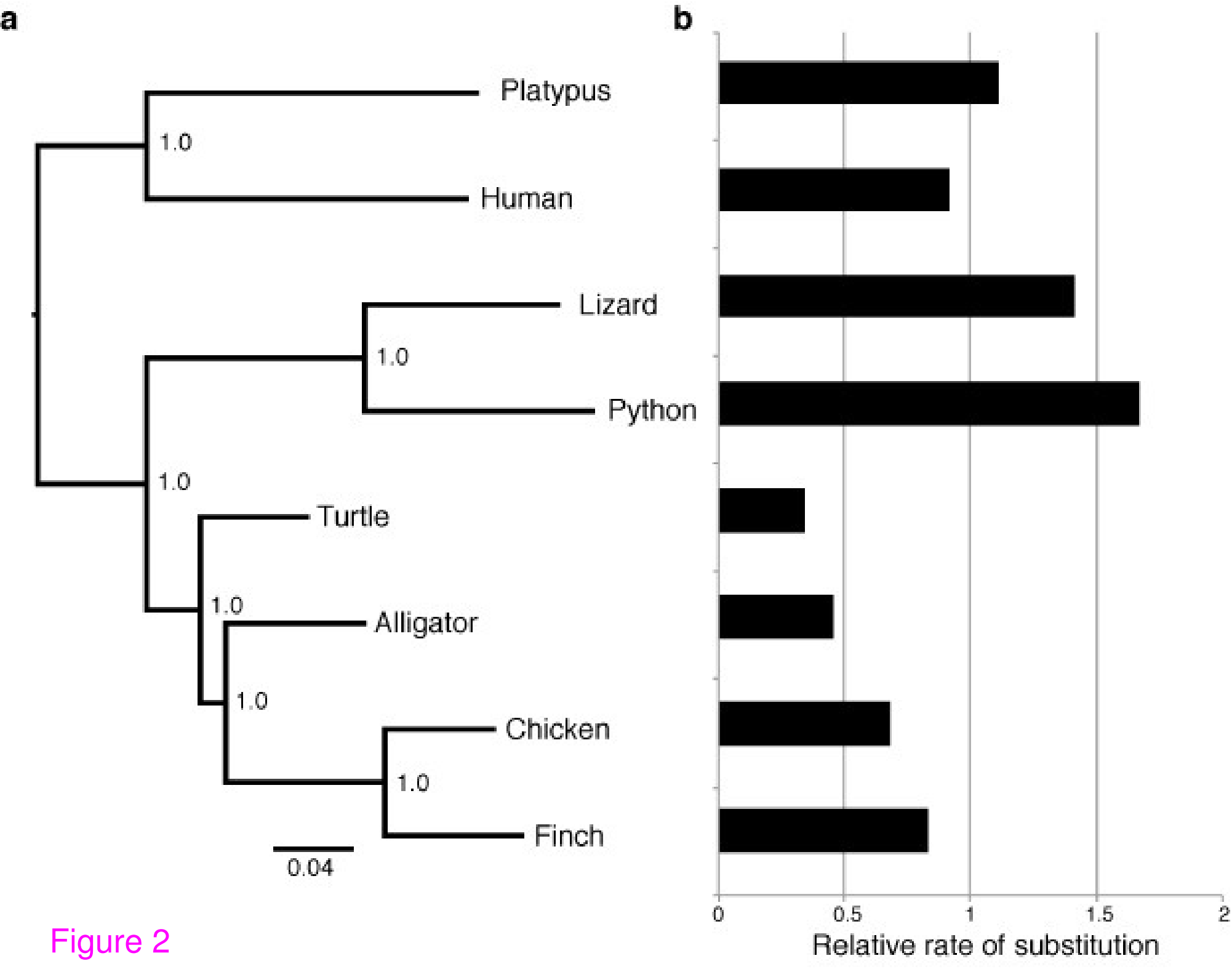

a
Platypus
1.0
Human
Lizard
1.0
Python
1.0
Turtle
1.0
Alligator
1.0
Chicken
1.0
Finch
0.04
b
0
0.5
1
1.5
2
Relative rate of substitution

Figure 2

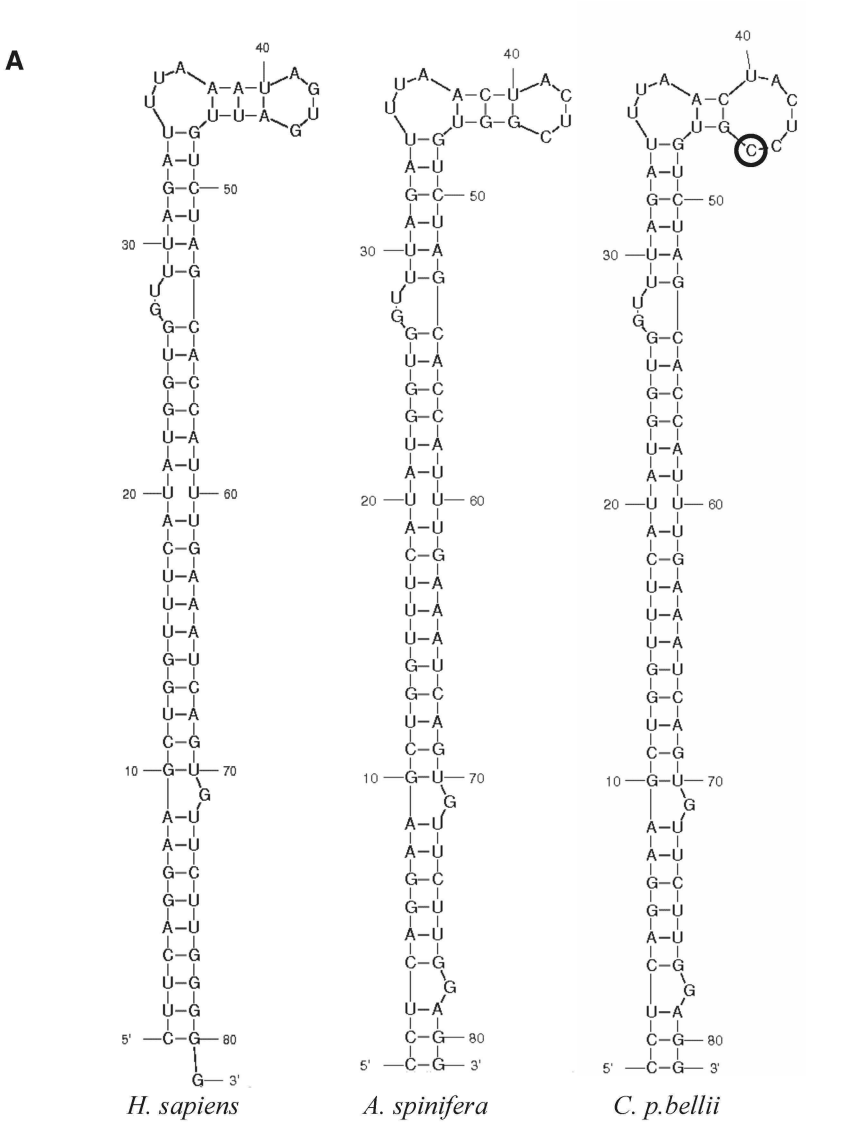
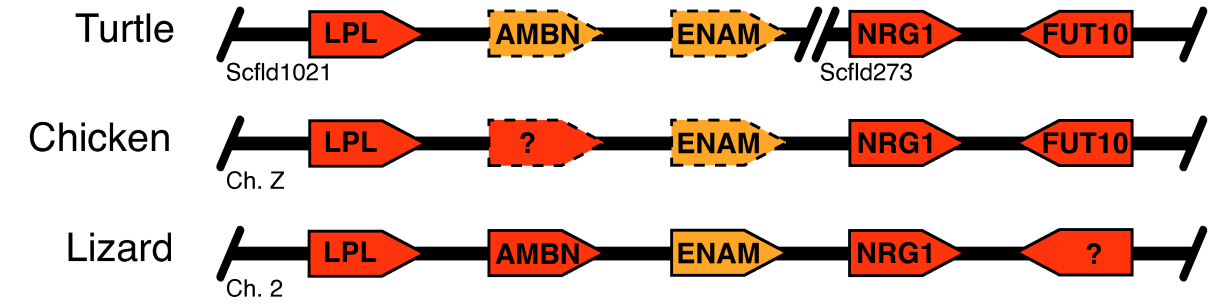
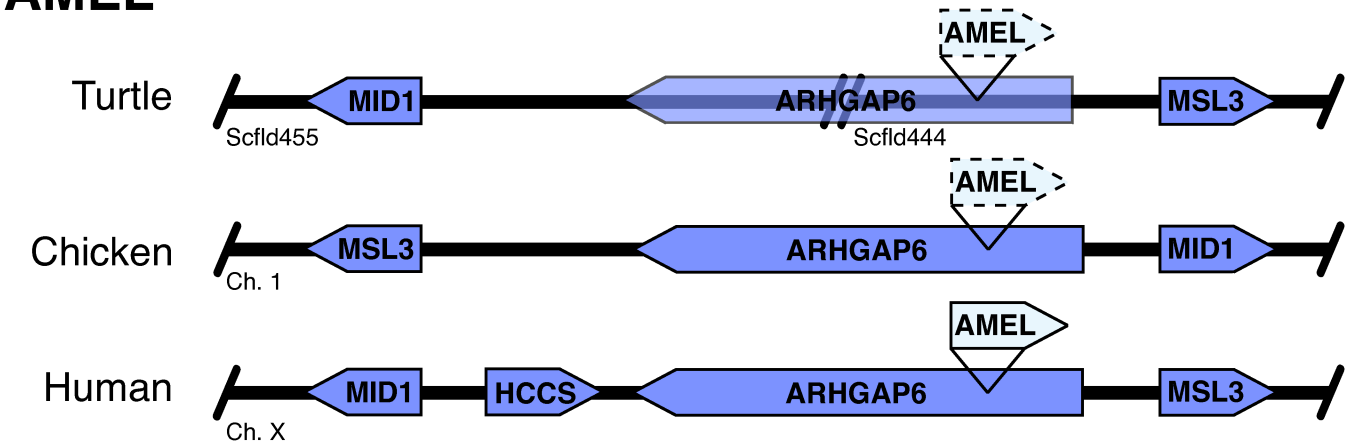

A
H. sapiens
A. spinifera
C. p.bellii
B
miR-29b Expression
Relative Fold Expression
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
*
*
Control
24 hr Frozen
4 hr Recovery

Figure 3

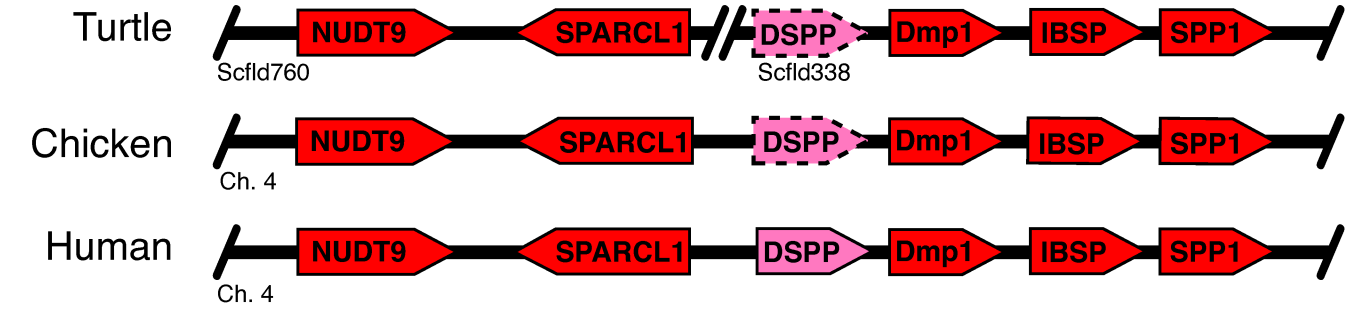
## AMBN & ENAM




## AMEL




## DSPP




## MMP20

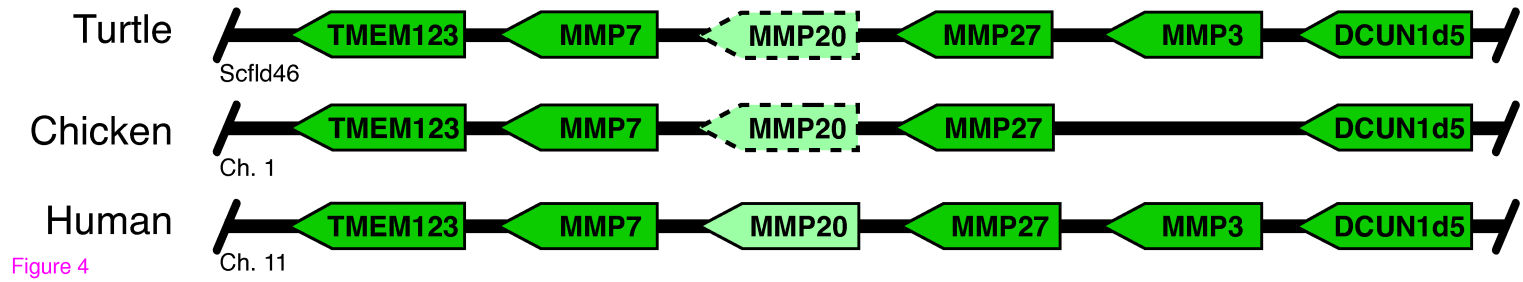
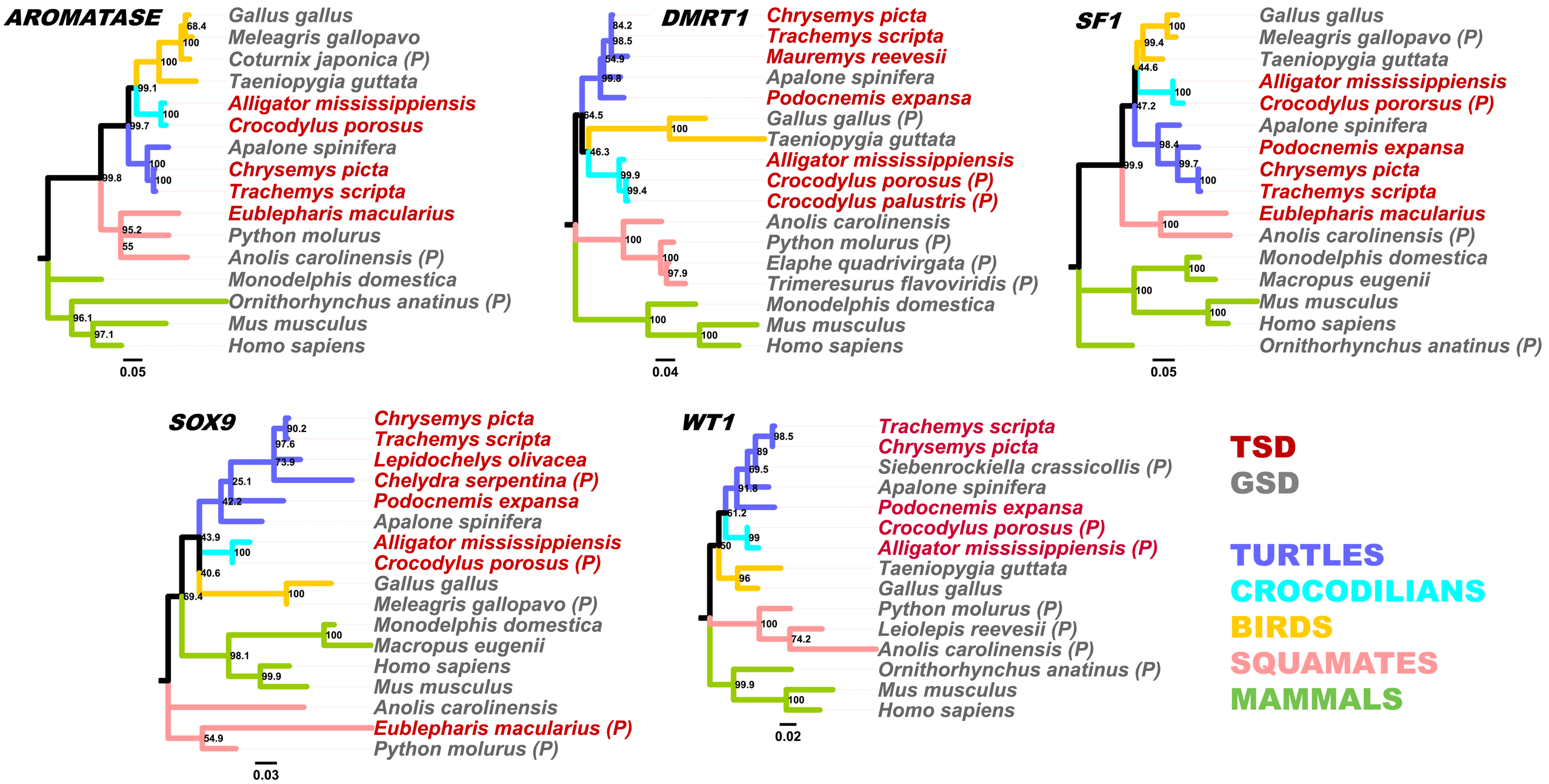



Figure 4


AROMATASE
Gallus gallus
Meleagris gallopavo
Coturnix japonica (P)
Taeniopygia guttata
Alligator mississippiensis
Crocodylus porosus
Apalone spinifera
Chrysemys picta
Trachemys scripta
Eublepharis macularius
Python molurus
Anolis carolinensis (P)
Monodelphis domestica
Ornithorhynchus anatinus (P)
Mus musculus
Homo sapiens
68.4
100
100
99.1
100
99.7
100
100
99.8
95.2
55
96.1
97.1
0.05
DMRT1
Chrysemys picta
Trachemys scripta
Mauremys reevesii
Apalone spinifera
Podocnemis expansa
Gallus gallus (P)
Taeniopygia guttata
Alligator mississippiensis
Crocodylus porosus (P)
Crocodylus palustris (P)
Anolis carolinensis
Python molurus (P)
Elaphe quadrivirgata (P)
Trimeresurus flavoviridis (P)
Monodelphis domestica
Mus musculus
Homo sapiens
84.2
98.5
54.9
99.8
64.5
100
46.3
99.9
99.4
100
100
97.9
100
100
0.04
SF1
Gallus gallus
Meleagris gallopavo (P)
Taeniopygia guttata
Alligator mississippiensis
Crocodylus porosus (P)
Apalone spinifera
Podocnemis expansa
Chrysemys picta
Trachemys scripta
Eublepharis macularius
Anolis carolinensis (P)
Monodelphis domestica
Macropus eugenii
Mus musculus
Homo sapiens
Ornithorhynchus anatinus (P)
100
99.4
44.6
100
47.2
98.4
99.9
99.7
100
100
100
100
100
0.05
SOX9
Chrysemys picta
Trachemys scripta
Lepidochelys olivacea
Chelydra serpentina (P)
Podocnemis expansa
Apalone spinifera
Alligator mississippiensis
Crocodylus porosus (P)
Gallus gallus
Meleagris gallopavo (P)
Monodelphis domestica
Macropus eugenii
Homo sapiens
Mus musculus
Anolis carolinensis
Eublepharis macularius (P)
Python molurus (P)
90.2
97.6
73.9
25.1
42.2
43.9
100
40.6
69.4
100
100
98.1
99.9
54.9
0.03
WT1
Trachemys scripta
Chrysemys picta
Siebenrockiella crassicollis (P)
Apalone spinifera
Podocnemis expansa
Crocodylus porosus (P)
Alligator mississippiensis (P)
Taeniopygia guttata
Gallus gallus
Python molurus (P)
Leiolepis reevesii (P)
Anolis carolinensis (P)
Ornithorhynchus anatinus (P)
Mus musculus
Homo sapiens
98.5
89
69.5
91.8
61.2
99
50
96
100
74.2
99.9
100
0.02
TSD
GSD
TURTLES
CROCODILIANS
BIRDS
SQUAMATES
MAMMALS

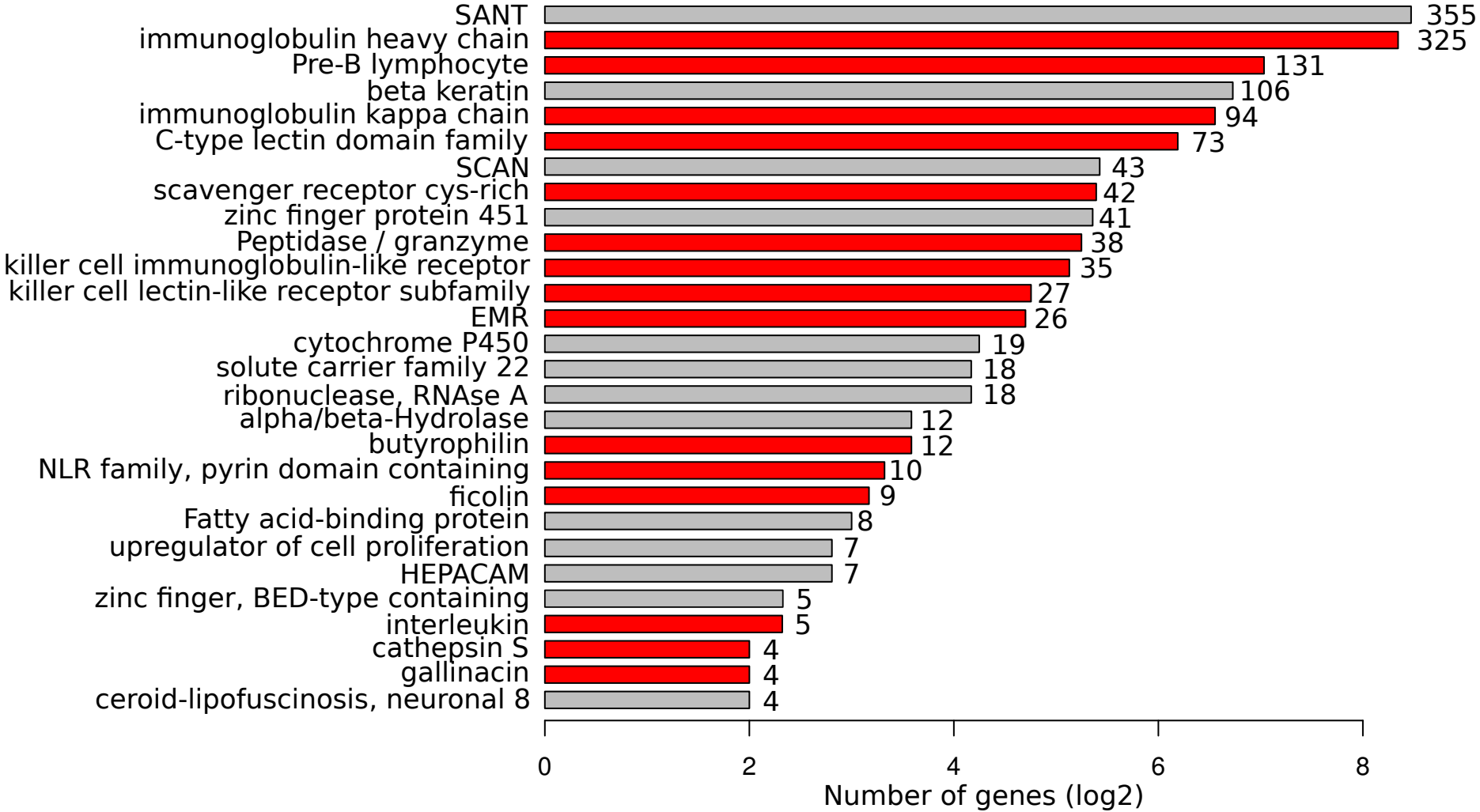

SANT
immunoglobulin heavy chain
Pre-B lymphocyte
beta keratin
immunoglobulin kappa chain
C-type lectin domain family
SCAN
scavenger receptor cys-rich
zinc finger protein 451
Peptidase / granzyme
killer cell immunoglobulin-like receptor
killer cell lectin-like receptor subfamily
EMR
cytochrome P450
solute carrier family 22
ribonuclease, RNAse A
alpha/beta-Hydrolase
butyrophilin
NLR family, pyrin domain containing
ficolin
Fatty acid-binding protein
upregulator of cell proliferation
HEPACAM
zinc finger, BED-type containing
interleukin
cathepsin S
gallinacin
ceroid-lipofuscinosis, neuronal 8
355
325
131
106
94
73
43
42
41
38
35
27
26
19
18
18
12
12
10
9
8
7
7
5
5
4
4
4
0
2
4
6
8
Number of genes (log2)

Figure 6

**Additional files provided with this submission:**

Additional file 1: Additional file 1.doc, 788K
http://genomebiology.com/imedia/5064405995182527/supp1.doc
Additional file 2: Additional file 2.docx, 4340K
http://genomebiology.com/imedia/1468159519951825/supp2.docx
Additional file 3: Additional file 3.xlsx, 82K
http://genomebiology.com/imedia/5529374829518264/supp3.xlsx