

# The tuatara genome: insights into vertebrate evolution from the sole survivor of an ancient reptilian order

Neil J. Gemmell<sup>1\*</sup>, Kim Rutherford<sup>1</sup>, Stefan Probst<sup>2,3</sup>, Marc Tollis<sup>4,5</sup>, David Winter<sup>6</sup>, J. Robert Macey<sup>7</sup>, David L. Adelson<sup>8</sup>, Alexander Suh<sup>9</sup>, Terry Bertozzi<sup>8,10</sup>, José H. Grau<sup>11,12</sup>, Chris Organ<sup>13</sup>, Paul P. Gardner<sup>14</sup>, Matthieu Muffato<sup>15</sup>, Mateus Patricio<sup>15</sup>, Konstantinos Billis<sup>15</sup>, Fergal J Martin<sup>15</sup>, Paul Flicek<sup>15</sup>, Bent Petersen<sup>16</sup>, Lin Kang<sup>17</sup>, Pawel Michalak<sup>17-19</sup>, Thomas R. Buckley<sup>20,21</sup>, Melissa Wilson<sup>4</sup>, Yuanyuan Cheng<sup>22</sup>, Hilary Miller<sup>23</sup>, Ryan K. Schott<sup>24</sup>, Melissa Jordan<sup>25</sup>, Richard Newcomb<sup>25</sup>, José Ignacio Arroyo<sup>26</sup>, Nicole Valenzuela<sup>27</sup>, Tim A. Hore<sup>1</sup>, Jaime Renart<sup>28</sup>, Valentina Peona<sup>9</sup>, Claire R. Peart<sup>9,29</sup>, Vera M. Warmuth<sup>9,29</sup>, Lu Zeng<sup>8</sup>, R. Daniel Kortschak<sup>8</sup>, Joy M. Raison<sup>8</sup>, Valeria Velásquez Zapata<sup>27</sup>, Zhiqiang Wu<sup>27</sup>, Didac Santesmasses<sup>30</sup>, Marco Mariotti<sup>30</sup>, Roderic Guigó<sup>30</sup>, Shawn M. Rupp<sup>4</sup>, Victoria G. Twort<sup>20,21</sup>, Nicolas Dussex<sup>1</sup>, Helen Taylor<sup>1</sup>, Hideaki Abe<sup>1</sup>, James M. Paterson<sup>31</sup>, Daniel G. Mulcahy<sup>32</sup>, Vanessa L. Gonzalez<sup>32</sup>, Charles G. Barbieri<sup>7</sup>, Dustin P. DeMeo<sup>7</sup>, Stephan Pabinger<sup>33</sup>, Oliver Ryder<sup>34</sup>, Scott V. Edwards<sup>35</sup>, Steven L. Salzberg<sup>36</sup>, Lindsay Mickelson<sup>37</sup>, Nicola Nelson<sup>37</sup>, Clive Stone<sup>38</sup>, Ngatiwai Trust Board<sup>38</sup>

1. Allan Wilson Centre, Department of Anatomy, University of Otago, PO Box 56, Dunedin 9054, New Zealand
2. LOEWE-Center for Translational Biodiversity Genomics, Senckenberg Museum, 60325 Frankfurt, Germany
3. South African National Biodiversity Institute, National Zoological Garden, Pretoria, 0184, South Africa
4. School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ 85281 USA
5. School of Informatics, Computing, and Cyber Systems, Northern Arizona University, 1295 S Knoles Drive, Flagstaff, AZ 86011 USA
6. School of Fundamental Sciences, Massey University, Private Bag 11 222, Palmerston North, 4442, New Zealand
7. Peralta Genomics Institute, Chancellor's Office, Peralta Community College District, 333 East 8th Street, Oakland, CA 94606
8. School of Biological Sciences, The University of Adelaide, Adelaide, SA 5005, Australia
9. Department of Ecology and Genetics - Evolutionary Biology, Evolutionary Biology Centre (EBC), Uppsala University, Norbyvägen 18D, 75236 Uppsala, Sweden.
10. Evolutionary Biology Unit, South Australian Museum, Adelaide, SA 5000, Australia
11. Amedes Genetics, Amedes Medizinische Dienstleistungen GmbH, Jägerstr. 61, 10117 Berlin, Germany
12. Museum für Naturkunde Berlin, Leibniz-Institut für Evolutions- und Biodiversitätsforschung an der Humboldt-Universität zu Berlin, Invalidenstraße 43, 10115. Berlin, Germany
13. Department of Earth Sciences, Montana State University, Bozeman, MT 59717
14. Department of Biochemistry, University of Otago, PO Box 56, Dunedin 9054, New Zealand
15. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
16. Section for Evolutionary Genomics, The GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Øster Farimagsgade 5, 1353 København K, Denmark
17. Edward Via College of Osteopathic Medicine, Blacksburg, VA 24060, USA
18. Center for One Health Research, Virginia-Maryland College of Veterinary Medicine, Blacksburg, VA 24060, USA
19. Institute of Evolution, University of Haifa, Haifa, 3498838, Israel
20. Manaaki Whenua - Landcare Research, Private Bag 92170, Auckland, New Zealand
21. School of Biological Sciences, The University of Auckland, Private Bag 92019, Auckland, New Zealand
22. School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia
23. Biomatters Ltd, Level 2, 18 Shortland Street, Auckland 1010 New Zealand
24. Department of Vertebrate Zoology, National Museum of Natural History, Smithsonian Institution, Washington DC 20560, USA

25. The New Zealand Institute for Plant & Food Research, Private Bag 92169, Auckland, New Zealand
26. Departamento de Ecología, Facultad de Ciencias Biológicas, Pontificia Universidad Católica de Chile, Avenida Libertador Bernardo O'Higgins 340, Santiago, Chile.
27. Iowa State University, Department of Ecology, Evolution, and Organismal Biology, 251 Bessey Hall, Ames IA, 50111 USA.
28. Instituto de Investigaciones Biomédicas CSIC-UAM, 28029-Madrid, Spain.
29. Division of Evolutionary Biology, Faculty of Biology, Ludwig-Maximilian University of Munich, D-82152 Planegg-Martinsried, Germany
30. Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Catalonia, Spain, Universitat Pompeu Fabra (UPF), Barcelona, Catalonia, Spain
31. School of Biological Sciences, University of Canterbury, New Zealand
32. Global Genome Initiative, National Museum of Natural History, Smithsonian Institution, 1000 Constitution Ave., Washington, DC 20560, USA
33. AIT Austrian Institute of Technology, Center for Health and Bioresources, Molecular Diagnostics, Muthgasse 11, 1190 Vienna, Austria
34. San Diego Zoo Institute for Conservation Research, Escondido, CA 92027 USA
35. Department of Organismic and Evolutionary Biology and the Museum of Comparative Zoology, Harvard University, Cambridge, MA 02138 USA
36. Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205 USA
37. School of Biological Sciences, Victoria University of Wellington, PO Box 600, Wellington 6140, New Zealand
38. Ngatiwai Trust Board, 129 Port Road, PO Box 1332, Whangarei, New Zealand

\*Correspondence to: neil.gemmell@otago.ac.nz

**The tuatara (*Sphenodon punctatus*), the only living member of the archaic reptilian order Rhynchocephalia (Sphenodontia) once widespread across Gondwana, is an iconic and enigmatic terrestrial vertebrate endemic to New Zealand. A key link to the now extinct stem reptiles from which dinosaurs, modern reptiles, birds and mammals evolved, the tuatara provides exclusive insights into the ancestral amniotes. The tuatara genome, at ~5 Gbp, is among the largest vertebrate genomes assembled. Analysis of this genome and comparisons to other vertebrates reinforces the uniqueness of the tuatara. Phylogenetic analyses indicate tuatara diverged from the snakes and lizards ~250 MYA. This lineage also shows moderate rates of molecular evolution, with instances of punctuated evolution. Genome sequence analysis identifies expansions of protein, non-protein-coding RNA families, and repeat elements, the latter of which show an extraordinary amalgam of reptilian and mammalian features. Sequencing of this genome provides a valuable resource for deep comparative analyses of tetrapods, as well as for tuatara biology and conservation. It also provides important insights into both the technical challenges and the cultural obligations associated with genome sequencing.**

The tuatara (*Sphenodon punctatus*) is an iconic and enigmatic terrestrial vertebrate, unique to New Zealand<sup>1</sup>. The only living member of the archaic reptilian order Rhynchocephalia (Sphenodontia) which last shared a common ancestor with other reptiles some 250 million years ago (Figure 1), the tuatara represents an important link to the now extinct stem reptiles from which dinosaurs, modern reptiles, birds and mammals evolved and is thus important for our understanding of amniote evolution<sup>2</sup>.

It is also a species of significance in other contexts. First, the tuatara is a *taonga*, or special treasure, for Māori, who hold that tuatara are the guardians of special places<sup>1</sup>. Second, the tuatara is internationally recognised as a critically important species vulnerable to extinction due to habitat loss, predation, disease, global warming and other factors<sup>1</sup>. Third, the tuatara displays a variety of morphological and physiological innovations that have puzzled scientists since the first description of the species in 1867<sup>1,3</sup>. These include a unique combination of features shared variously with lizards, turtles, and birds, which left its taxonomic position in doubt for many decades<sup>1</sup>. That taxonomic conundrum has been largely addressed using molecular approaches<sup>4</sup>, but the timing of the split of tuatara from the lineage that forms the modern squamata (lizards and snakes), the rate of evolution of tuatara, and the number of species of tuatara remains contentious<sup>1</sup>. Last, there are aspects of tuatara biology that are unique, or atypical of reptiles. These include the possession of a parietal eye whose function remains in dispute, a unique form of temperature-dependent sex determination (TSD) that sees females produced below, and males above, 22°C, extremely low basal metabolic rates, and significant longevity<sup>1</sup>.

To shed new insight into the biology of this extraordinary species we have sequenced the tuatara genome in partnership with Ngatiwai, the Māori iwi (tribe) who hold kaitiakitanga (guardianship) over the tuatara populations located on islands in the far North of New Zealand. This partnership, unique among the genome projects undertaken to date, had a strong practical focus on developing resources and information that will improve our understanding of tuatara and aid future conservation efforts. It is an exemplar for future genome initiatives that aspire to meet access and benefit-sharing obligations to indigenous communities required under the Convention for Biological Diversity<sup>5</sup>.

We find that the tuatara genome, as well as the animal, is an amalgam of ancestral and derived characteristics. Tuatara have  $2n=36$  chromosomes in both sexes, consisting of 14 pairs of macrochromosomes and four pairs of microchromosomes<sup>6</sup>. The genome size, estimated at 5 Gbp, is among the largest vertebrate genomes sequenced to date and this is predominantly explained by an extraordinary diversity of repeat elements, many of which are unique to tuatara. Given its size and complexity this genome has posed a variety of technical challenges, so we also outline how we have innovated and adapted a variety of technologies and assembly approaches to produce the *de novo* assembly of one of the largest and most complex high-quality vertebrate genomes undertaken to date.

## Sequencing, assembly, synteny and annotation

The tuatara genome assembly is 4.3 Gbps and consists of 16,536 scaffolds with an N50 scaffold length of 3 Mbp (Table S1). Genome assessment using BUSCO<sup>7</sup> indicates 85.6% of sequences are present and complete based on the eukaryota gene set. Subsequent annotation identified 17,448 genes, of which 16,185 were identified as one-to-one orthologs (Supplementary Materials 2). Local gene-order conservation is high; 75% or more of the tuatara genes showed conservation with birds, turtles and

crocodilians (Supplementary Figures 1 and 2). We also find that components of the genome, 15 Mbp and greater, are syntenic with other vertebrates, with the protein-coding gene order and orientation maintained between tuatara, turtle, chicken, and human (Supplementary Figures 1–3).

## Genomic architecture

### Repetitive elements

In-depth repeat analyses revealed that at least (64%) of the tuatara genome assembly is composed of repetitive sequences, made up of transposable elements (31%) and low copy number segmental duplications (33%). While the total transposable element content is similar to other reptiles<sup>8–11</sup>, there are notable differences in the types of repeats found which make this genome appear more mammal-like than reptile-like. Furthermore, a number of repeat families show evidence of recent activity and greater expansion and diversity than seen in other vertebrates (Figure 2).

Among long interspersed elements (LINEs), L2 elements account for most of the LINE sequences in the tuatara genome (10% of the genome) and some may still be active, a remarkable observation given that CR1 elements are the dominant LINE in other sauropsid genomes<sup>8–10,12–14</sup> but only account for a relatively small fraction (~4%) of the tuatara genome (Figure 2 panel a and b; Table S2; Supplementary Materials 4). There are also potentially active CR1 elements in tuatara (Supplementary Materials 4). L1 elements, which are prevalent in placental mammals account for only a tiny fraction of the genome (<1%) (Table S2). Phylogenetic analysis of L2 elements identifies an L2 subfamily present in tuatara, but absent from other lepidosaurs, that is also common in monotremes (Supplementary Materials 4, Figures 4.3 and 4.4) where these repeats account for 18–23% of the genome<sup>15</sup>. These data suggest that the LINE composition of the tuatara genome is more monotreme-like than reptile-like. Whether this implies an ancestral presence of these repeats or an ancient horizontal transfer event remains to be determined, but it appears that stem sauropsid ancestors had quite a different repeat composition than has been inferred from prior comparisons using birds and lizards.

Many short interspersed elements (SINEs), TEs which depend on the enzymatic machinery of LINEs, belong to copies from ancient CORE-SINEs present in all amniotes<sup>16</sup>. We also identified at least 16 SINE subfamilies which were recently active in the tuatara genome (Figure 2 panel c; Supplementary Materials 5). Most of these SINEs are from MIR subfamilies. While consistent with the pattern described above for L2 elements, which are required for mobilisation of these SINEs, this pattern is unusual for sauropsids but reminiscent of the repeat landscape of monotreme and marsupial mammals<sup>17,18</sup>. Notably, the diversity of tuatara MIR subfamilies is higher than in marsupials<sup>17</sup>, the highest previously known in amniotes. In the human genome, hundreds of fossil MIR elements act as chromatin and regulatory domains<sup>19</sup>, which suggests that the recent activity of diverse tuatara MIR subfamilies may have influenced regulatory rewiring on rather recent evolutionary timescales.

We detected 24 novel DNA transposon families with no sequence similarity to known families or each other, suggesting frequent tuatara germline infiltration by novel DNA transposons through horizontal transfer<sup>20</sup>. No fewer than 30 DNA transposon subfamilies were recently active, spanning a diverse range of cut-and-paste transposons and polintons (Figure 2 panel d; Supplementary Materials 5). This diversity is higher than in anole lizards<sup>21</sup> and vespertilionid bats<sup>22–24</sup>, which are among the few amniotes identified with recently active DNA transposons. Strikingly, we found thousands of identical DNA transposon copies, suggesting very recent, and/or ongoing, activity. It is thus possible that cut-and-paste transposition is shaping the tuatara genome in similar ways as in vespertilionid bats<sup>22–24</sup>.

We identified approximately 7,500 full-length long terminal repeat (LTR) retroelements including endogenous retroviruses (ERVs), which we classified into at least 12 groups using phylogenetic reconstructions of their retrotranscriptase domains (Figure 2 panel e and f; Supplementary Materials 6). We identified nearly 450 ERVs from five major retroviral clades and some 5,500 Ty3/Gypsy elements representing five major clades. Notably, a Ty1/Copia element (Mtanga-like) is abundant in the tuatara genome with more than 1,140 full-length elements. The general spectrum of tuatara LTR retroelements is comparable to that of other sauropsids, but lacks Bel/Pao LTR retroelements which are present in anole<sup>25</sup>. An unexpected result is the presence of at least 37 complete spumaretroviruses in the tuatara genome (Figure 2 panel f; Supplementary Materials 6). Spumaretroviruses are considered among the most ancient ERVs<sup>26,27</sup> and were previously found only in coelacanth<sup>28</sup>, sloth<sup>29</sup>, aye-aye<sup>30</sup> and Cape golden mole<sup>31</sup>. Thus, the occurrence of several full-length endogenous spumaretroviruses in tuatara, some encoding a complete envelope gene, is the first in a sauropsid genome.

The tuatara genome further contains more than 8,000 non-coding RNA (ncRNA)-related elements. The bulk of these elements (~6,900) are derived from recently active transposable elements. Most overlap with a novel CR1-mobilized SINE (Figure 2 panel c; Supplementary Materials 7), which contains nearly 3,900 elements with strong matches to tRNA covariance models (over 10,000 elements in total). The remaining high copy number elements include sequences closely related to ribosomal RNAs, spliceosomal RNAs and signal recognition particle RNAs.

Finally, an extraordinarily high proportion (33%) of the tuatara genome originates from low copy number segmental duplications, with 6.7% of these of recent origin based on their high level of sequence identity (>94% identity), which is more than seen in other vertebrates<sup>15</sup>. The tuatara genome is 2.4x larger than the anole genome and this difference appears thus to be driven disproportionately by segmental duplications.

The overall repeat architecture of tuatara is unlike anything previously reported, showing a unique amalgam of features previously viewed as characteristic of either reptilian or mammalian lineages. This truly reflects the phylogenetic position of this astonishing evolutionary relic – a combination of ancient amniote features as well as a surprisingly dynamic and diverse repertoire of lineage-specific TEs.



## Epigenetic features

Low-coverage sequencing using post-bisulfite adaptor tagging (PBAT)<sup>32</sup> shows ~81% of CpG sites are methylated in tuatara (Figure 3a). When PBAT is used to compare against other vertebrate species, DNA methylation levels are the highest reported for an amniote — markedly different methylation to those observed in mouse, human (~70%) and chicken (~50%) and more similar to that seen in *Xenopus* (82%) and zebrafish (78%). One possible explanation for this high rate of DNA methylation is the extraordinary level of repetitive elements found in this genome, many of which appear recently active and might be regulated via DNA methylation (see above).

Examination of the CpG distribution in tuatara, sheds new light on the commonalities and diversity of nCpG distribution across vertebrates, a trait that is linked to DNA methylation and may consequently impact gene expression (Supplementary Materials 8). Tuatara display a bimodal nCpG distribution in all genomic regions examined (gene promoters, exons, introns, and intergenic sequences), similar to recent reports in painted turtles<sup>33</sup> and alligator, both of which have temperature dependent sex determination (TSD, Supplementary Figure 4). The bimodal nCpG pattern seen in introns of these TSD reptiles may mark a divide between TSD taxa and the genotypic sex determination (GSD) vertebrates previously reported to display a unimodal nCpG distribution in introns<sup>33</sup>. However, testing explicitly for the existence of bimodality in all genomic regions examined using mixed models, we find a bimodal pattern of nCpG that appears to be ancestral and conserved across all vertebrates irrespective of their sex-determining mechanism. This finding agrees with the recent detection of bimodal DNA methylation in gene bodies in human, chicken, fish, and tunicates<sup>34</sup>.

## The mitochondrial genome

The mitochondrial (mt) genome in the tuatara reference animal is 18,078 bp, containing 13 protein coding, 2 rRNA, and 22 tRNA genes, standard among animals (Supplementary Figure 5). This contradicts previous reports<sup>35,36</sup> that the tuatara mt-genome lacks 3 genes: ND5, tRNA<sup>Thr</sup>, tRNA<sup>His</sup>. These three genes are found, with an additional copy of tRNA<sup>Leu(CUN)</sup> and an additional non-coding block (NC2) in a single segment of the mt-genome. This Illumina assembly result is confirmed with 40 Oxford Nanopore mitochondrial DNA reads, of which one covers the entire molecule. Three non-coding areas with Control Region (heavy strand replication origin, O<sub>H</sub>) features (NC1–3), and two copies of tRNA<sup>Leu(CUN)</sup> adjacent to NC1 and NC2 possess identical or near identical sequence, possibly as a result of concerted evolution. A stem-and-loop structure is observed in regions encoding tRNA<sup>Asn</sup> and tRNA<sup>Cys</sup>, which may supplement for the missing structure normally observed for the origin of light strand replication (O<sub>L</sub>) in this location<sup>37</sup>. The tRNA<sup>Lys</sup> gene is duplicated with the first copy possibly a pseudo-gene<sup>35</sup>. The tRNA<sup>Cys</sup> gene encodes a tRNA with a D-arm replacement loop<sup>38</sup>.

## **The unique biology of the tuatara is reflected in its genome**

Befitting its taxonomic distinctiveness, we find that the tuatara genome displays multiple innovations in genes associated with immunity, odour reception, thermal regulation and selenium metabolism.

### **MHC**

Genes of the major histocompatibility complex (MHC) play an important role in disease resistance<sup>39</sup>, mate choice and kin recognition<sup>40</sup> and are among the most polymorphic in the vertebrate genome. Prior work<sup>41</sup> mapped the core MHC genes to tuatara chromosome 13q, although a small number of class I genes map to chromosome 4p. Here, annotation of tuatara MHC regions and comparisons to the gene organization to six other representative species (Supplementary Materials 9, Supplementary Figure 6), identified 56 genes spanning 13 scaffolds, including six class I genes, six class II genes, 19 class III genes, 18 framework genes, and seven extended class II genes. Four class I and four class II genes were not linked to other MHC related genes and were excluded from the comparative analysis as they may be located outside the core MHC region on chromosome 13<sup>41</sup>.

The genomic organization of tuatara MHC is most similar to that of the green anole, which we interpret as being typical of the Lepidosauria (Supplementary Figure 6). Unlike the chicken MHC, which was proposed to represent the minimal essential MHC gene cluster<sup>42</sup>, tuatara, anole, alligator, and turtle show a high gene content and complexity that is more similar to the MHC regions of amphibians and mammals than birds. While the majority of genes annotated in the tuatara MHC are well conserved as one-to-one orthologs, extensive genomic rearrangements were observed among these distant lineages (Supplementary Figure 6). The most noteworthy rearrangement lies in the relative locations of genes in the class I and class III regions (as defined by their locations in the human MHC), with these two regions inverted between the mammalian and frog MHC, and the genes interspersed in the tuatara and anole genomes. This finding raises questions over the traditional categorization of these genes as class I or class III region genes, as evolutionarily there is no explicable difference between the two sets of genes in terms of gene clustering, organization, or function.

### **Vision and smell**

The tuatara is a highly visual predator that is able to capture prey under extremely low light conditions<sup>1</sup>. Despite this extreme nocturnal visual adaptation, the tuatara shows strong morphological evidence of an ancestrally diurnal visual system including a fovea<sup>43</sup> and photoreceptors that while superficially rod-like appear to be evolutionary derived from cones<sup>44</sup>, similar to the situation in snakes and geckos<sup>45</sup>. The tuatara genome provides an opportunity to uncover the molecular basis of these morphological adaptations. A search for visual genes identified all five vertebrate opsin genes in the tuatara genome, and a comparative analysis revealed one of the lowest rates of visual

gene loss for any amniote (Supplementary Materials 10). This low gene loss contrasted sharply with high rates of gene loss in ancestrally nocturnal lineages (Supplementary Figure 7).

Analysis of the selective constraint operating upon a subset of visual genes (those involved in phototransduction) showed strong negative selection and no concerted evidence of long-term shifts in selective pressures found in other groups that have evolutionary modified photoreceptors i.e., rod-like cones and cone-like rods<sup>46</sup> (Supplementary Materials 10). Collectively, these results suggest a unique path to nocturnal adaptation in tuatara from a diurnal ancestor. Furthermore, the retention of five visual opsins and conserved nature of the visual system strongly suggests tuatara possess robust colour vision, potentially at low light levels where most animals are colour-blind. This surprisingly broad visual repertoire may be explained by the dichotomy in tuatara life-history: juvenile tuatara, often take up a diurnal and arboreal lifestyle to avoid the terrestrial, nocturnal, adults that may predate them<sup>1</sup>.

Odorant receptors (ORs), expressed in the dendritic membranes of olfactory receptor neurons, are responsible for the detection of odours. Species that depend strongly on their sense of smell to interact with their environment, find prey, identify kin and avoid predators may be expected to have a large number of ORs. Four hundred and seventy two genes encoding predicted ORs were identified from the genome of tuatara (Supplementary Materials 11). Of these, 341 sequences were predicted to be intact ORs, encoding proteins 303–345 amino acids in length ([10.5281/zenodo.2592798](https://doi.org/10.5281/zenodo.2592798)). The remainder consisted of 9 full length sequences missing the initial start codon, 60 partial genes, 22 genes containing frameshifts, and 40 presumed pseudogenes. Many ORs were found as tandem arrays, with up to 26 found on a single scaffold. Tuatara OR genes do not contain introns, following the conventional vertebrate OR gene structure<sup>47,48</sup>.

Odorant receptor gene number and diversity varies greatly amongst the Sauropsida. Birds possess from 182–688 genes, the green anole lizard has 156 genes, while crocodilian and testudine repertoires are much larger containing 1000–2000 genes<sup>47–51</sup>. Tuatara has a similar number of ORs to birds, but many more than the genome of the single lizard species published to date. Interestingly the tuatara genome contains a high percentage of intact OR genes, 85%, compared to other published OR sets from sauropsid genomes. This may reflect a strong reliance on olfaction by tuatara and therefore pressure to maintain a substantial repertoire of ORs (Supplementary Figure 8). There is some evidence that olfaction may play a role in identifying prey<sup>52,53</sup> and that cloacal secretions may potentially act as pheromones<sup>54,55</sup>.

## Thermoregulation

The tuatara, a behavioural thermoregulator, is notable for having the lowest optimal body temperature of any reptile (16–21 °C). They can remain active at temperatures as low as 5 °C, while temperatures over 28 °C are generally fatal<sup>1</sup>. Transient Receptor Potential (TRP) ion channel genes play an important role in thermoregulation,



participating in thermosensation and cardiovascular physiology<sup>56 57</sup>. Inhibition of cold- and heat-sensitive TRP channels in the saltwater crocodile produces alterations in thermoregulation<sup>58</sup> leading us to hypothesise that TRP genes may be linked to the extraordinary thermal tolerance of tuatara. We therefore undertook comparative genomic analysis of TRP genes in tuatara and six other amniotes (lizard, snake, alligator, chicken, turtle, and human) to look for evidence of genomic adaptations related to thermoregulation in tuatara (Supplementary Materials 12). We identified 37 TRP-like sequences, spanning all seven described subfamilies of TRPs in the tuatara genome (Supplementary Figure 9) — an unusually large repertoire of TRP genes.

Among this suite of genes we identified new thermo- and non-thermosensitive TRP genes that appear to result from gene duplication, and have been differentially retained in tuatara. For example, tuatara is unusual in possessing an additional copy of a thermosensitive TRPV-like gene (TRPV1/2/3) classically linked to the detection of moderate to extreme heat<sup>56</sup> — a feature it shares with turtles (Supplementary Figure 9). A strong signature of positive selection among heat-sensitive TRP genes (TRPA1, TRPM and TRPV) was also observed.

In general these results show a high rate of differential retention and positive selection in genes for which function in heat sensation is well established<sup>56 57</sup>. Given the role of these genes, it seems probable that these genomic changes in TRP genes are associated with the evolution of thermoregulation in this species.

### Selenoproteins and tRNA<sup>Sec</sup> genes

Tuatara are, barring tortoises, the longest-lived reptiles, likely exceeding 100 years<sup>1</sup>. This enhanced life-span may be linked to genes that afford protection against reactive oxygen species either directly as antioxidants or by regulating redox pathways<sup>59</sup>. One class of genes known to afford such protection are the selenoproteins, proteins that incorporate selenium in the form of selenocysteine (Sec), the 21st amino acid. The human genome encodes 25 selenoproteins whose roles include antioxidation, redox regulation, thyroid hormone synthesis, calcium signal transduction and others<sup>60</sup>.

We identified 26 selenoprotein genes in the tuatara genome, along with genes encoding the machinery for selenoprotein synthesis (Supplementary Materials 13). The number of selenoproteins, and protein families to which they belong, are similar to other vertebrate genomes<sup>61</sup>, but for the presence of two selenoproteins from the SELENOW (previously SelW) family. SELENOW is a short protein with unknown function that has a thioredoxin domain (Rdx family, PF10262). Two SELENOW selenoproteins, SELENOW1 and SELENOW2, were predicted to be present in the last common ancestor of vertebrates, but most genomes have only one of the two genes<sup>61</sup>. SELENOW2 was predicted to have been lost in the amniote stem (sauropsids and mammals). The presence of SELENOW2 in tuatara suggests that this loss must have occurred later in evolution, after the split of sauria and mammals.

The selenocysteine-specific tRNA (tRNA-Sec) plays a central role in both the biosynthesis and incorporation of Sec<sup>60</sup> and is usually present as a single functional copy, as is observed in the squamates. In tuatara we identified four eukaryotic tRNA-Sec genes in the same scaffold within a ~5 kbp region of the genome, each of which appears to be functional (Supplementary Materials 13).

While further work is needed, the unusual selenoprotein gene and tRNA composition of tuatara may be linked to the extraordinary longevity of tuatara, or might arise as a response to the low levels of selenium and other trace elements in New Zealand's terrestrial systems.

### Sex determination genes and markers

Tuatara have a unique mode of temperature dependent sex determination (TSD) where higher egg incubation temperatures result in males<sup>1</sup>. Thus we were interested to determine which genes previously implicated in vertebrate sex determination were conserved in tuatara (Table S3). We found orthologs for many genes known to act antagonistically in masculinising (e.g., *sfl*, *sox9*) and feminising (e.g., *rspo1*, *wnt4*) gene networks to, respectively, promote testicular or ovarian development<sup>62</sup>. We also found orthologs of several genes recently implicated in TSD, including *cirbp*<sup>62</sup>. Orthologs for some notable vertebrate sex determining genes (e.g., *dmrt1* and *amhl*) were absent in our multiple species alignments. We suspect these are not genuinely absent in tuatara and result from alignment failure due to high sequence divergence, because of the prior identification of *dmrt1* in our genome data<sup>63</sup> and its earlier localisation in the tuatara genome by gene mapping<sup>6</sup>.

Tuatara have no obviously differentiable sex chromosomes<sup>6</sup>, but we hypothesized that sex specific markers might be identifiable through comparison of known sex males and females at a genomic and epigenetic level. We found no significant sex specific differences in global CG methylation between males and females (Figure 3a). Likewise, a search for sex specific SNPs<sup>64</sup> using population genomic resources (detailed below) identified no sex specific SNPs (Figure 3b).

### Phylogeny, evolutionary rates and evidence for punctuated evolution

Albert Günther discovered that the “remarkable saurian” tuatara, described in the 18th century as “a monstrous animal of the lizard kind” is not a lizard but rather a unique reptile classified in a separate taxonomic order, Rhynchocephalia<sup>3</sup>. We placed the tuatara within the context of vertebrate evolution by carrying out multiple phylogenomic analyses that incorporated both whole genome alignments and clusters of single-copy orthologs (Supplementary Materials 14 and 15), recapitulating many patterns observed from the fossil record corroborated during the genomic era (Figure 1). After their appearance ~312 million years ago<sup>65</sup>, amniote vertebrates diversified into two groups: the synapsids, which include all mammals, and the sauropsids, which include all reptiles and birds. By analysing 818,968 fourfold degenerate sites obtained

from 27 vertebrates in a maximum likelihood framework, we obtained full phylogenomic support for a monophyletic Lepidosauria, marked by the divergence of the tuatara lineage from all squamates (lizards and snakes) during the early part of the Triassic Period ~246 million years ago as estimated using a penalized likelihood method (Supplemental Materials 14). The phylogenetic placement and age of divergence for the tuatara was similarly recovered by analysing 245 single-copy gene orthologs using concatenated, species tree, and molecular clock methods (Figure 1; Supplementary Materials 15 and 16).

The rate of molecular evolution in the tuatara has been suggested to be paradoxically high in contrast to the apparently slow rate of morphological evolution<sup>66,67</sup>. However, we find that the actual divergence in terms of DNA substitutions per site per million years at fourfold degenerate sites has been relatively low, particularly with respect to lizards and snakes, making it the slowest-evolving lepidosaurian reptile yet analysed (Supplementary Figure 10). We also find that amniote evolution in general can be described by a model of punctuated evolution, where the amount of genomic change is related to the degree of species diversification within clades<sup>68,69</sup>. The tuatara falls well below this trend, suggesting this “living fossil” has accumulated substitutions at a rate expected given the lack of rhynchocephalian diversity (Supplementary Figure 11; Supplementary Materials 16). This suggests that rates of phenotypic and molecular evolution were not decoupled throughout amniote evolution<sup>70</sup>.

## Patterns of selection

In two different sets of analyses, we find that most genes exhibit a pattern of molecular evolution that suggests the tuatara branch evolves at a different rate than the rest of the tree (Supplementary Materials 17, Table S4). Approximately 659 of 4,284 orthologs tested had significantly different  $\omega$  values (ratios of non-synonymous to synonymous substitutions or dN/dS) on the tuatara branch relative to other birds and reptiles tested. Although none of these orthologs had  $\omega$  values suggestive of strong positive selection (i.e.  $> 1$ ), the results do indicate that shifts in patterns of selection are affecting many genes and functional categories of genes across the tuatara genome, including genes involved in RNA regulation, metabolic pathways and general metabolism.

For genes involved in sex determination, we find that the entire length of the tree is shorter and that dN and dN/dS values are significantly lower than for typical genes in the genome (Supplementary Materials 17). Thus, while we find that most genes available for these analyses show strong levels of purifying selection, genes involved in sex determination are even more conserved, highlighting their critical function for reproduction and survival<sup>62</sup>.

## Population genomics

Once widespread across the supercontinent of Gondwana, Rhynchocephalia are now represented by a single species, the tuatara, found on a few offshore islands in New Zealand (Figure 1c). Tuatara have declined because of introduced pests and habitat

loss<sup>1</sup>. They remain at risk of extinction due to their highly restricted distribution and the threats imposed by disease and climate-change-induced changes in sex ratios, which could profoundly impact their survival<sup>71</sup>. Earlier population genetic studies have shown populations in northern New Zealand are genetically distinct from those in the Cook Strait, and that the population on North Brother Island in the Cook Strait might be a distinct species<sup>72</sup>. Although subsequent studies have not found evidence for species-status<sup>73</sup>, this population is still managed as a separate conservation unit.

We took advantage of the tuatara reference genome to perform the first ancestral demographic and population genomic analyses of this species. First, we investigated genome wide signals for demographic change using PSMC (Supplementary Materials 18). The reconstructed demography (Figure 3c) reveals an increase in effective population size ( $N_e$ ) detectable around 10 Mya, a drastic decrease in  $N_e$  about 1–3 Mya, and a rapid increase in  $N_e$  between 500 kya and 1 Mya. These events correlate well with the known geological history of New Zealand<sup>74</sup>, reflecting an increase in available landmass following the Oligocene drowning, a period of significant climatic cooling that likely reduced tuatara habitat, and the formation of land bridges that facilitated population expansion.

Next, we obtained population samples spanning the three axes of genetic diversity in tuatara<sup>73,75</sup>: northern New Zealand (Little Barrier Is./ Hauturu), and two islands in Cook Strait (Stephens Island / Takapourewa; North Brother Island) (Figure 3d), sampling 10 individuals each. We used a genotype-by-sequencing approach to generate 22 Gb of data from these individuals (Supplementary Materials 19). After applying filters and thinning the data to analyse only unlinked loci, we obtained 52,171 single nucleotide variants (SNVs) for population genomic analysis.

Analysis of the SNV dataset reveals that modern tuatara populations are highly differentiated (Figure 3d). Our genome-wide estimate of the fixation index ( $F_{ST}$ ) is 0.54, and more than two thirds of variable sites have an allele which is restricted to a single island. All of the island populations have relatively low genetic diversity (nucleotide diversity ranges from  $8 \times 10^{-4}$  for North Brother to  $1.1 \times 10^{-3}$  for Little Barrier). The low within-population diversity and marked population structure we observe in tuatara suggests modern island populations were isolated from each other sometime during the last glacial maximum at ~18 kya.

Our results also support the distinctiveness of the Brothers Island tuatara, which has variously been described as *Sphenodon punctatus* or *Sphenodon guntheri*<sup>72</sup>. Recently the two species have been again collapsed to one<sup>73</sup>, but we find that the Brothers Island population forms a distinct group in all clustering analyses (Figure 3d). This population is highly inbred and shows evidence of a severe bottleneck, most likely reflecting a founder event around the last glaciation<sup>75</sup>. It is not clear if the distinctiveness we observe is due to allele frequency changes brought about by this bottleneck, or is reflective of a deeper split in the population history of tuatara. Regardless of the historical reasons for the North Brother population's genetic distinctiveness, it is an important source of

genetic diversity in tuatara, containing 8,480 private alleles in our dataset, and warrants ongoing conservation as an independent unit.

## **A cultural dimension**

The tuatara is a taonga, or special treasure, for many Māori — notably Ngātiwai and Ngāti Koata who are the kaitiaki (guardians) of tuatara. The tuatara genome project was undertaken in partnership with Ngātiwai iwi; our common goal being to increase knowledge and understanding of tuatara to aid the long term conservation of this species. Ngātiwai were involved in all decision-making regarding the use of the genome data by potential project partners; for each new project proposed we discussed the benefits that might accrue from this work and how these could be shared. The need to engage with, and protect the rights of, indigenous communities in such a transparent way has seldom been considered in the genomic projects published to date, even those related to human genomics, but is now a mandated consideration under the Nagoya Protocol<sup>76</sup>. Our partnership, which required significant time and investment from both parties since its initiation in 2011, is a first step towards an inclusive model of genomic science that we hope others will adopt and improve upon. While each partnership is unique, we provide a template genome research agreement (Supplementary Materials 20) that we hope will be useful to others.

## **Discussion**

The tuatara genome is among the highest quality genomes assembled for any reptile, and one of the largest amniote genomes thus far assembled. The tuatara has a genomic architecture unlike anything previously reported. Tuatara possess a unique amalgam of features that were previously viewed as characteristic of either the mammalian or reptilian lineages. Notable among these features are unusually high levels of repetitive sequences classically considered mammalian, many of which appear to have been recently active, and a level of genome methylation that is the highest of any amniote thus far investigated. We also found a mitochondrial genome (mtDNA) gene content at odds with published reports for this species that mistakenly omitted the ND5 gene<sup>35</sup>; this gene is indeed present and nested within a repeat rich region of the mtDNA.

Phylogenetic studies reveal new insights into the timing and speed of amniote evolution, including new evidence of punctuated genome evolution across this phylogeny. Contrary to past claims<sup>66</sup> we find that the evolutionary rate for tuatara is not exceptionally fast, rather it is the slowest-evolving lepidosaurian reptile yet analysed.

Investigations of genomic innovations identified genetic candidates that may explain tuatara's ultra-low active body temperature, extraordinary longevity, and apparent resistance to infectious disease. Further functional exploration will refine our understanding of these unusual facets of tuatara biology, while the tuatara genome itself will enable many future studies to explore the evolution of complex systems across the vertebrates in a more complete way than has previously been possible.



Our population genomic work supports the taxonomic distinctiveness of the Brothers Island tuatara, which has variously been described as *Sphenodon punctatus* or *Sphenodon guntheri*<sup>72</sup> over the past three decades, but has recently been resynonymised<sup>73</sup>. Our findings show significant genetic differences among the populations, leading us to suggest that a separate conservation plan be re-established for each population.

Finally, this landmark genome will greatly aid future work on population differentiation, inbreeding, and local adaptation in this global icon, the last remaining species of the once globally dominant reptilian order Rhynchocephalia (Sphenodontia).

## Methods

### General methods

A full description of the Methods can be found in the Supplementary Materials.

### Sampling and sequencing

A blood sample was obtained from a large male tuatara from Lady Alice Island (35°53'24.4"S 174°43'38.2"E), New Zealand, with appropriate ethical and iwi permissions. Total genomic DNA and RNA were extracted and sequenced using the Illumina HiSeq 2000 and MiSeq sequencing platforms (Illumina, San Diego, CA, USA) supported by New Zealand Genomics Ltd (Supplementary Materials 1).

### Genome, transcriptome and epigenome

#### *Genome assembly and scaffolding*

Raw reads were *de novo* assembled using Allpaths-LG (version 49856)<sup>77</sup>. With a total input data of 5,741,034,516 reads for the paired-end libraries and 2,320,886,248 reads of the mate pair libraries, several setups were tested for the optimal proportion of input data. Based on assembly statistics an optimal setup was found using 85% of the fragment libraries and 100% of the jumping libraries. Gaps were closed using GapCloser<sup>78</sup> and the assembly scaffolded with the transcriptome (see below) using L\_RNA\_Scaffolder<sup>79</sup>. We further scaffolded the assembly using Chicago libraries and HiRise<sup>80</sup>.

We assembled a *de novo* transcriptome as a reference for read mapping using total RNA derived from the blood sample of our reference male using a standard RNA-Seq pipeline<sup>81</sup>, and a collection of prior transcriptomic data collected from early stage embryos<sup>82</sup>. In total we had 131,580,633 new 100bp read pairs and 60,637,100 prior 50 bp read pairs<sup>82</sup>. These were assembled using Trinity v2.2.0<sup>83,84</sup> running default parameters. To reduce redundancy in the assembled transcripts, we used CD-HIT-EST v4.6.6<sup>85</sup> and then assessed assembly quality and completeness before and after redundancy removal using BUSCO v1.22<sup>7</sup> and Transrate v1.0.3<sup>86</sup> (Supplementary Materials, section 1.4).

Low-coverage BS-seq was undertaken using a modified post-bisulfite adapter tagging (PBAT) method<sup>87</sup> to explore global patterns of methylation in the tuatara genome for representatives of both sexes (12 males, 13 females) (Figure 1c, 3d). Briefly, purified DNA was subjected to bisulfite conversion using the EZ Methylation Direct Mag Prep kit (Zymo, D5044). Converted DNA was then subjected to single-ended sequencing (1x 150 bp) on an Illumina MiSeq. Raw reads were trimmed in two steps using TrimGalore v0.4.0 and trimmed reads were mapped using Bismark v0.14.3 with the --pbat option<sup>87</sup>. Global methylation levels for each species were calculated from Bismark reports by dividing methylated cytosine calls by total cytosine calls in both CG and non-CG contexts.

## Repeat and gene annotation

We used a combination of *ab initio* repeat identification in CARP/RepeatModeler/LTRharvest<sup>15,88,89</sup>, manual curation of specific newly identified repeats, and homology to repeat databases to investigate the repeat content of the tuatara genome (Supplementary Materials 1.6). From these three complementary repeat identification approaches, the CARP results were in-depth annotated for LINEs and segmental duplications (Supplementary Materials 4), the RepeatModeler results were in-depth annotated for SINEs and DNA transposons (Supplementary Materials 5), and the LTRharvest results were in-depth annotated for LTR retrotransposons (Supplementary Materials 6).

For the gene annotation we used RepeatMasker (v4.0.3) along with our partially curated RepeatModeler library plus Repbase sauropsid repeat database to mask transposable elements in the genome sequence prior to the gene annotation. We did not mask simple repeats at this point to allow for more efficient mapping during the homology based step in the annotation process. Simple repeats were later soft-masked and protein-coding genes predicted using MAKER2<sup>90</sup>. We used anole lizard (*A. carolinensis*, version AnoCar2.0), python (*P. bivittatus*, version bivittatus-5.0.2,) and RefSeq ([www.ncbi.nlm.nih.gov/refseq](http://www.ncbi.nlm.nih.gov/refseq)) as protein homology evidence, which we integrated with *ab initio* gene prediction methods including BLASTX, SNAP<sup>91</sup> and Augustus<sup>92</sup>. Non-coding RNAs were annotated using Rfam covariance models (v13.0)<sup>93</sup> (Supplementary Materials 7).

## Ortholog calling

We performed a phylogenetic analysis to infer orthology relationships between the tuatara and 25 other species (Supplementary Materials Tables 2.1 and 2.2) using the Ensembl GeneTree method<sup>94</sup>. Multiple-sequence alignments, phylogenetic trees and homology relationships were extracted in various formats ([10.5281/zenodo.2542570](https://doi.org/10.5281/zenodo.2542570)). We also calculated the Gene Order Conservation (GOC) score<sup>95</sup>, which uses local synteny information around a pair of orthologous genes to compute how much the gene order is conserved. For each of these species, we chose the paralogue with the best GOC score and sequence similarity, which resulted in a total set of 3,168 clusters of orthologs (Supplementary Materials Table 2.3).

## Gene tree reconstructions and substitution rate estimation

We constructed phylogenies using just fourfold degenerate site data derived from whole genome alignments for 27 tetrapods, analysed as a single partition in RAxML v8.2.3<sup>96</sup>. Using the topology and branch lengths obtained from the best maximum likelihood phylogeny, we estimated absolute rates of molecular evolution in terms of substitution per site per million years and estimated the divergence times of amniotes via the semiparametric penalized likelihood (PL) method<sup>97</sup> with the program r8s v1.8<sup>98</sup> (Supplementary Materials 14.5).

We also generated gene trees based on 245 single-copy orthologs found between all species using a maximum likelihood based multi-gene approach (Supplementary Materials 15). Sequences were aligned using the codon-based aligner Prank<sup>99</sup>. The FASTA format alignments were then converted to Phylip using the catfasta2phyml.pl script (<https://github.com/nylander/catfasta2phyml>). Next, we used the individual exon Phylip files for gene tree reconstruction using RaxML<sup>96</sup> using a GTR + G model<sup>96</sup>. Subsequently, we binned all gene trees to reconstruct a species tree and carried out bootstrapping using Astral<sup>100</sup>. Astral applies binning of gene trees with similar topologies, based on an incompatibility graph between gene trees. Subsequently, it chooses the most likely species tree under the multi-species coalescent model (Supplementary Materials Figure 15.1).

## Divergence times and tests of punctuated evolution

We inferred time-calibrated phylogenies with BEAST v2.4.8<sup>101</sup> using the CIPRES Science Gateway<sup>102</sup> to explore divergence times (Supplementary Materials 16.1). Subsequently, we used Bayesian phylogenetic generalized least squares to regress the total phylogenetic path length (of four-fold degenerate sites) on the net number of speciation events (nodes in a phylogenetic tree) as a test for punctuated evolution<sup>103</sup>. (Supplementary Materials 16.2).

## Analysis of genomic innovations

We explored the genomic organisation and sequence evolution of genes associated with immunity, vision, smell, thermoregulation, longevity and sex determination (Supplementary Materials 8–13). Tests of selection were undertaken across multiple genes, including those linked to metabolism, vision and sex determination (Supplementary Materials 17) using multispecies alignments and PAML<sup>104</sup>.

## Population genomics

Demographic history was inferred from the diploid sequence of our tuatara genome using a pairwise sequential Markovian coalescent (PSMC) method<sup>105</sup> as detailed in Supplementary Materials 18. We also sampled 10 animals from each of three populations that span the main axes of genetic diversity in tuatara (Supplementary Materials Table 19.1) and used a modified genotype-by-sequencing approach<sup>106</sup> to obtain 52,171 single nucleotide variants (SNVs) that we used for population genomic

analysis, investigations of loci associated with sexual phenotype and estimates of genetic load (Supplementary Materials 19).

## Reporting summary

Further information on experimental design is available in the Nature Research Reporting Summary linked to this article.

## Permits and ethics

Samples were collected under Victoria University of Wellington Animal Ethics approvals 2006R12; 2009R12; 2012R33; 22347 and held and used under permits 45462-DOA and 32037-RES 32037-RES issued by the New Zealand Department of Conservation.

## Data availability

The Tuatara Genome Consortium Project Whole Genome Shotgun and genome assembly are registered under the umbrella BioProject PRJNA418887 and BioSample SAMN08038466. Transcriptome read data are submitted under SRR7084910 (whole blood) together with prior data SRR485948. The transcriptome assembly is submitted to GenBank with ID GGNQ000000000.1. Illumina short-read and nanopore long read sequence are in SRAs associated with PRJNA445603. The assembly (GCA\_003113815.1) described in this paper is version QEPC000000000.1 and consists of sequences QEPC01000001-QEPC01016536. Maker gene predictions are available from Zenodo, DOI: [10.5281/zenodo.1489353](https://doi.org/10.5281/zenodo.1489353). The repeat library database developed for tuatara is available from Zenodo, DOI: [10.5281/zenodo.2585367](https://doi.org/10.5281/zenodo.2585367).

## References

1. Cree, A. *Tuatara: Biology and Conservation of a Venerable Survivor*. (Canterbury University Press, 2014).
2. Jones, M. E. H. & Cree, A. Tuatara. *Curr. Biol.* **22**, R986–7 (2012).
3. Gunther, A. XIX. Contribution to the anatomy of hatteria (rhynchocephalus, Owen). *Philosophical Transactions of the Royal Society of London* **157**, 595–629 (1867).
4. Jones, M. E. H. *et al.* Integration of molecules and new fossils supports a Triassic

- origin for Lepidosauria (lizards, snakes, and tuatara). *BMC Evol. Biol.* **13**, 208 (2013).
5. Wynberg, R. & Laird, S. A. Fast science and sluggish policy: The Herculean task of regulating biodiscovery. *Trends Biotechnol.* **36**, 1–3 (2018).
6. O’Meally, D., Miller, H., Patel, H. R., Graves, J. A. M. & Ezaz, T. The first cytogenetic map of the tuatara, *Sphenodon punctatus*. *Cytogenet. Genome Res.* **127**, 213–223 (2009).
7. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
8. Alföldi, J. *et al.* The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* **477**, 587–591 (2011).
9. Shaffer, H. B. *et al.* The western painted turtle genome, a model for the evolution of extreme physiological adaptations in a slowly evolving lineage. *Genome Biol.* **14**, R28 (2013).
10. Green, R. E. *et al.* Three crocodilian genomes reveal ancestral patterns of evolution among archosaurs. *Science* **346**, 1254449 (2014).
11. Pasquesi, G. I. M. *et al.* Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.* **9**, 2774 (2018).
12. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).
13. Warren, W. C. *et al.* The genome of a songbird. *Nature* **464**, 757 (2010).



14. Suh, A. *et al.* Multiple lineages of ancient CR1 retroposons shaped the early genome evolution of amniotes. *Genome Biol. Evol.* **7**, 205–217 (2014).
15. Zeng, L., Kortschak, R. D., Raison, J. M., Bertozzi, T. & Adelson, D. L. Superior ab initio identification, annotation and characterisation of TEs and segmental duplications from genome assemblies. *PLOS One* **13**, e0193588 (2018).
16. Gilbert, N. & Labuda, D. CORE-SINEs: Eukaryotic short interspersed retroposing elements with common sequence motifs. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 2869–2874 (1999).
17. Mikkelsen, T. S. *et al.* Genome of the marsupial *Monodelphis domestica* reveals innovation in non-coding sequences. *Nature* **447**, 167–177 (2007).
18. Warren, W. C. *et al.* Genome analysis of the platypus reveals unique signatures of evolution. *Nature* **453**, 175–183 (2008).
19. Wang, J. *et al.* MIR retrotransposon sequences provide insulators to the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E4428–37 (2015).
20. Schaack, S., Gilbert, C. & Feschotte, C. Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. *Trends Ecol. Evol.* **25**, 537–546 (2010).
21. Novick, P. A., Smith, J. D., Floumanhaft, M., Ray, D. A. & Boissinot, S. The evolution and diversity of DNA transposons in the genome of the lizard *Anolis carolinensis*. *Genome Biol. Evol.* **3**, 1–14 (2011).
22. Ray, D. A. *et al.* Multiple waves of recent DNA transposon activity in the bat, *Myotis lucifugus*. *Genome Res.* **18**, 717–728 (2008).
23. Pagán, H. J. T. *et al.* Survey sequencing reveals elevated DNA transposon activity,

- novel elements, and variation in repetitive landscapes among vesper bats. *Genome Biol. Evol.* **4**, 575–585 (2012).
24. Mitra, R. *et al.* Functional characterization of piggyBat from the bat *Myotis lucifugus* unveils an active mammalian DNA transposon. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 234–239 (2013).
  25. Tollis, M. & Boissinot, S. The transposable element profile of the anolis genome: How a lizard can provide insights into the evolution of vertebrate genome size and structure. *Mob. Genet. Elements* **1**, 107–111 (2011).
  26. Aiewsakun, P. & Katzourakis, A. Marine origin of retroviruses in the early Palaeozoic Era. *Nat. Commun.* **8**, 13954 (2017).
  27. Hayward, A., Cornwallis, C. K. & Jern, P. Pan-vertebrate comparative genomics unmasks retrovirus macroevolution. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 464–469 (2015).
  28. Han, G.-Z. & Worobey, M. An endogenous foamy-like viral element in the coelacanth genome. *PLOS Pathog.* **8**, e1002790 (2012).
  29. Katzourakis, A., Gifford, R. J., Tristem, M., Gilbert, M. T. P. & Pybus, O. G. Macroevolution of complex retroviruses. *Science* **325**, 1512 (2009).
  30. Han, G.-Z. & Worobey, M. An endogenous foamy virus in the aye-aye (*Daubentonia madagascariensis*). *J. Virol.* **86**, 7696–7698 (2012).
  31. Lee, A., Nolan, A., Watson, J. & Tristem, M. Identification of an ancient endogenous retrovirus, predating the divergence of the placental mammals. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **368**, 20120503 (2013).
  32. Peat, J. R., Ortega-Recalde, O., Kardailsky, O. & Hore, T. A. The elephant shark

methyloome reveals conservation of epigenetic regulation across jawed vertebrates.

*F1000Res.* **6**, 526 (2017).

33. Radhakrishnan, S., Literman, R., Mizoguchi, B. A. & Valenzuela, N. MeDIPseq and nCpG analyses illuminate sexually dimorphic methylation of gonadal development genes with high historic methylation in turtle hatchlings with temperature-dependent sex determination. *Epigenetics Chromatin* **10**, 1–16 (2017).
34. Keller, T. E., Han, P. & Yi, S. V. Evolutionary transition of promoter and gene body DNA methylation across invertebrate–vertebrate boundary. *Mol. Biol. Evol.* **33**, 1019–1028 (2016).
35. Rest, J. S. *et al.* Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Mol. Phylogenet. Evol.* **29**, 289–297 (2003).
36. Subramanian, S., Mohandesan, E., Millar, C. D. & Lambert, D. M. Distance-dependent patterns of molecular divergences in tuatara mitogenomes. *Sci. Rep.* **5**, 8703 (2015).
37. Macey, J. R., Larson, A., Ananjeva, N. B., Fang, Z. & Papenfuss, T. J. Two novel gene orders and the role of light-strand replication in rearrangement of the vertebrate mitochondrial genome. *Mol. Biol. Evol.* **14**, 91–104 (1997).
38. Macey, J. R., Larson, A., Ananjeva, N. B. & Papenfuss, T. J. Replication slippage may cause parallel evolution in the secondary structures of mitochondrial transfer RNAs. *Mol. Biol. Evol.* **14**, 30–39 (1997).
39. Hedrick, P. W. & Kim, T. J. Genetics of complex polymorphisms: Parasites and maintenance of the major histocompatibility complex variation. in *Evolutionary Genetics: From Molecules to Morphology* (eds. Singh, R. S. & Krimbas, C. B.)

204–234 (Cambridge University Press, 2000).

40. Milinski, M. The major histocompatibility complex, sexual selection, and mate choice. *Annu. Rev. Ecol. Evol. Syst.* **37**, 159–186 (2006).
41. Miller, H. C. *et al.* Major histocompatibility complex genes map to two chromosomes in an evolutionarily ancient reptile, the tuatara *Sphenodon punctatus*. *G3* **5**, 1439–1451 (2015).
42. Kaufman, J. Co-evolving genes in MHC haplotypes: The ‘rule’ for nonmammalian vertebrates? *Immunogenetics* **50**, 228–236 (1999).
43. Walls, G. L. *The vertebrate eye and its adaptive radiation*. (Bloomfield Hills, Cranbrook Institute of Science, 1942).
44. Meyer-Rochow, V. B., Wohlfahrt, S. & Ahnelt, P. K. Photoreceptor cell types in the retina of the tuatara (*Sphenodon punctatus*) have cone characteristics. *Micron* **36**, 423–428 (2005).
45. Schott, R. K. *et al.* Evolutionary transformation of rod photoreceptors in the all-cone retina of a diurnal garter snake. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 356–361 (2016).
46. Schott, R. K., Bhattacharyya, N. & Chang, B. S. Evolutionary signatures of photoreceptor transmutation in geckos reveal potential adaptation and convergence with snakes. *Evolution* (2019). doi:10.1111/evo.13810
47. Khan, I. *et al.* Olfactory receptor subgenomes linked with broad ecological adaptations in Sauropsida. *Mol. Biol. Evol.* **32**, 2832–2843 (2015).
48. Vandewege, M. W. *et al.* Contrasting patterns of evolutionary diversification in the olfactory repertoires of reptile and bird genomes. *Genome Biol. Evol.* **8**, 470–480

- (2016).
49. Steiger, S. S., Fidler, A. E. & Kempnaers, B. Evidence for increased olfactory receptor gene repertoire size in two nocturnal bird species with well-developed olfactory ability. *BMC Evol. Biol.* **9**, 117 (2009).
  50. Kishida, T. & Hikida, T. Degeneration patterns of the olfactory receptor genes in sea snakes. *J. Evol. Biol.* **23**, 302–310 (2010).
  51. Liu, Y. *et al.* *Gekko japonicus* genome reveals evolution of adhesive toe pads and tail regeneration. *Nat. Commun.* **6**, 10033 (2015).
  52. Besson, A. A. *et al.* Evidence of food chemical discrimination in tuatara (O. Rhynchocephalia): Comparison with a gekkotan lizard (O. Squamata). *J. Herpetol.* **43**, 124–131 (2009).
  53. Cooper, W. E., Ferguson, G. W. & Habegger, J. J. Responses to animal and plant chemicals by several iguanian insectivores and the tuatara, *Sphenodon punctatus*. *J. Herpetol.* **35**, 255–263 (2001).
  54. Wu, N. C. & Waas, J. R. No evidence for across-population scent discrimination of cloacal gland secretions in tuatara (*Sphenodon punctatus*). *J. Herpetol.* **51**, 178–185 (2017).
  55. Flachsbarth, B., Fritzsche, M., Weldon, P. J. & Schulz, S. Composition of the cloacal gland secretion of tuatara, *Sphenodon punctatus*. *Chem. Biodivers.* **6**, 1–37 (2009).
  56. Nilius, B. & Owsianik, G. The transient receptor potential family of ion channels. *Genome Biol.* **12**, 218 (2011).
  57. Tan, C.-H. & McNaughton, P. A. The TRPM2 ion channel is required for



- sensitivity to warmth. *Nature* **536**, 460–463 (2016).
58. Seebacher, F. & Murray, S. A. Transient receptor potential ion channels control thermoregulatory behaviour in reptiles. *PLOS One* **2**, e281 (2007).
  59. Halliwell, B. & Gutteridge, J. M. C. *Free Radicals in Biology and Medicine*. (Oxford University Press, 2015).
  60. Labunskyy, V. M., Hatfield, D. L. & Gladyshev, V. N. Selenoproteins: Molecular pathways and physiological roles. *Physiol. Rev.* **94**, 739–777 (2014).
  61. Mariotti, M. *et al.* Composition and evolution of the vertebrate and mammalian selenoproteomes. *PLOS One* **7**, e33066 (2012).
  62. Capel, B. Vertebrate sex determination: Evolutionary plasticity of a fundamental switch. *Nat. Rev. Genet.* **18**, 675–689 (2017).
  63. Janes, D. E. *et al.* Molecular evolution of Dmrt1 accompanies change of sex-determining mechanisms in reptilia. *Biol. Lett.* **10**, 20140809 (2014).
  64. Stovall, W. R. *et al.* Genetic sex assignment in wild populations using genotyping-by-sequencing data: A statistical threshold approach. *Mol. Ecol. Resour.* **18**, 179–190 (2018).
  65. Blair Hedges, S. & Kumar, S. *The Timetree of Life*. (Oxford University Press, 2009).
  66. Hay, J. M., Subramanian, S., Millar, C. D., Mohandesan, E. & Lambert, D. M. Rapid molecular evolution in a living fossil. *Trends Genet.* **24**, 106–109 (2008).
  67. Miller, H. C., Moore, J. A., Allendorf, F. W. & Daugherty, C. H. The evolutionary rate of tuatara revisited. *Trends Genet.* **25**, 13–15 (2009).
  68. Landis, M. & Schraiber, J. G. Pulsed evolution shaped modern vertebrate body

- sizes. *Proc Natl Acad Sci USA* **114**, 13224–13229 (2017).
69. Webster, A. J., Payne, R. J. H. & Pagel, M. Molecular phylogenies link rates of evolution and speciation. *Science* **301**, 478 (2003).
  70. Subramanian, S., Hay, J. M., Mohandesan, E., Millar, C. D. & Lambert, D. M. Molecular and morphological evolution in tuatara are decoupled. *Trends Genet.* **25**, 16–18 (2009).
  71. Mitchell, N. J., Kearney, M. R., Nelson, N. J. & Porter, W. P. Predicting the fate of a living fossil: How will global warming affect sex determination and hatching phenology in tuatara? *Proc. Biol. Sci.* **275**, 2185–2193 (2008).
  72. Daugherty, C. H., Cree, A., Hay, J. M. & Thompson, M. B. Neglected taxonomy and continuing extinctions of tuatara (*Sphenodon*). *Nature* **347**, 177 (1990).
  73. Hay, J. M., Sarre, S. D., Lambert, D. M., Allendorf, F. W. & Daugherty, C. H. Genetic diversity and taxonomy: A reassessment of species designation in tuatara (*Sphenodon*: Reptilia). *Conserv. Genet.* **11**, 1063–1081 (2010).
  74. Cooper, A. & Cooper, R. A. The Oligocene bottleneck and New Zealand biota: Genetic record of a past environmental crisis. *Proc. Biol. Sci.* **261**, 293–302 (1995).
  75. MacAvoy, E. S. *et al.* Genetic variation in island populations of tuatara (*Sphenodon* spp) inferred from microsatellite markers. *Conserv. Genet.* **8**, 305–318 (2007).
  76. Buck, M. & Hamilton, C. The Nagoya Protocol on access to genetic resources and the fair and equitable sharing of benefits arising from their utilization to the Convention on Biological Diversity. *Rev. Eur. Community Int. Environ. Law* **20**, 47–61 (2011).

77. Gnerre, S. *et al.* High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 1513–1518 (2011).
78. Luo, R. *et al.* SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. *Gigascience* **1**, 18 (2012).
79. Xue, W. *et al.* L\_RNA\_scaffolder: Scaffolding genomes with transcripts. *BMC Genomics* **14**, 604 (2013).
80. Putnam, N. H. *et al.* Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
81. Todd, E. V., Black, M. A. & Gemmell, N. J. The power and promise of RNA-seq in ecology and evolution. *Mol. Ecol.* **25**, 1224–1241 (2016).
82. Miller, H. C., Biggs, P. J., Voelckel, C. & Nelson, N. J. De novo sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (*Sphenodon punctatus*). *BMC Genomics* **13**, 439 (2012).
83. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
84. Haas, B. J. *et al.* De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Protoc.* **8**, 1494–1512 (2013).
85. Li, W. & Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* **22**, 1658–1659 (2006).
86. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of de novo transcriptome assemblies. *Genome*

- Res.* **26**, 1134–1144 (2016).
87. Peat, J. R. *et al.* Genome-wide bisulfite sequencing in zygotes identifies demethylation targets and maps the contribution of TET3 oxidation. *Cell Rep.* **9**, 1990–2000 (2014).
  88. Smit, A., Hubley, R. & Green, P. *RepeatModeler Open-1.0. 2008–2015*.
  89. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
  90. Holt, C. & Yandell, M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**, 491 (2011).
  91. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5**, 59 (2004).
  92. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19 Suppl 2**, ii215–25 (2003).
  93. Kalvari, I. *et al.* Rfam 13.0: Shifting to a genome-centric resource for non-coding RNA families. *Nucleic Acids Res* **46**, D335–342 (2018).
  94. Herrero, J. *et al.* Ensembl comparative genomics resources. *Database* **2016**, (2016).
  95. Aken, B. L. *et al.* Ensembl 2017. *Nucleic Acids Res.* **45**, D635–D642 (2017).
  96. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
  97. Sanderson, M. J. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Mol. Biol. Evol.* **19**, 101–109 (2002).

98. Sanderson, M. J. r8s: Inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301–302 (2003).
99. Löytynoja, A. & Goldman, N. An algorithm for progressive multiple alignment of sequences with insertions. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 10557–10562 (2005).
100. Mirarab, S. *et al.* ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics* **30**, i541–8 (2014).
101. Suchard, M. A. & Rambaut, A. Many-core algorithms for statistical phylogenetics. *Bioinformatics* **25**, 1370–1376 (2009).
102. Miller, M. A., Pfeiffer, W. & Schwartz, T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. in *2010 Gateway Computing Environments Workshop (GCE)* (2010). doi:10.1109/gce.2010.5676129
103. Pagel, M., Venditti, C. & Meade, A. Large punctuational contribution of speciation to evolutionary divergence at the molecular level. *Science* **314**, 119–121 (2006).
104. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
105. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
106. Dodds, K. G. *et al.* Construction of relatedness matrices using genotyping-by-sequencing data. *BMC Genomics* **16**, 1047 (2015).



## **Supplementary Information.**

Supplementary Figures 1–11

Supplementary Tables 1–4

Supplementary Materials 1–20.

## **Acknowledgements.**

N.J.G. acknowledges the support of Ngatiwai iwi, Allan Wilson Centre, University of Otago, New Zealand Department of Conservation, New Zealand Genomics Limited, Illumina. J.I.A. was supported by CONICYT National Doctoral Scholarship N°21130515. Ensembl annotation was supported by the Wellcome Trust (WT108749/Z/15/Z) and the European Molecular Biology Laboratory.

We also acknowledge Ngāti Kaota, Te Ātiawa and Ngāti Manuhiri for granting permission to reuse tuatara samples obtained from Stephens Island (Takapourewa), Little Barrier Island (Hauturu) and North Brother Island, respectively. We gratefully acknowledge all the people involved in obtaining and curating the samples held in the Victoria University tuatara collection.

Finally we thank Aleksey Zimin, Daniela Puiu, Guillaume Marcais, James Yorke and Ross Crowhurst for help and discussions about genome assembly, Ian Fiddes, Joel Armstrong and Benedict Paten for help with comparative genome alignments and annotation, the National eScience Infrastructure (NeSI) and Swedish National Infrastructure for Computing (SNIC) through the Uppsala Multidisciplinary Center for Advanced Computational Science (UPPMAX) for computational support, Robbie McPhee for help with figures, and Tamsin Braisher for manuscript coordination and editing.

## **Contributions**

N.J.G. designed the original concept and scientific objectives and oversaw the project and analyses. N.J.G., L.M., N.N., H.T., O.R., S.V.E., C.S. contributed samples or assisted in sample preparation and permitting. N.J.G., K.M.R., S.P., M.T., D.W., R.M., D.L.A., A.S., T.B., J.H.G., C.O., P.P.G., M.M. M.P., K.B., F.J.M., P.F., B.P., L.K., P.M., T.R.B., M.W., Y.C., H.M., R.K.S., M.J., R.N., J.I.A., N.V., T.A.H., J.R., V.P., C.R.P., V.M.W., L.Z., R.D.K., J.M.R., V.V.Z., Z.W., D.S., M.M., R.G., S.M.R., V.G.T., N.D., H.A., J.M.P., D.G.M., V.L.G., C.G.B., D.P.D., S.P., S.L.S. planned and carried out experiments or analyses. N.J.G., K.M.R., S.P., M.T., D.W., R.M., D.L.A., A.S., T.B., J.H.G., C.O., P.P.G., M.M. M.P., K.B., F.J.M., B.P., L.K., P.M., T.R.B., M.W., Y.C., H.M., R.K.S., M.J., R.N., J.I.A., N.V., T.A.H., J.R., V.P., C.R.P., V.M.W., L.Z., R.D.K., J.M.R., V.V.Z., Z.W., D.S., M.M., R.G., S.M.R., V.G.T., N.D., H.T., H.A., J.M.P., D.G.M., V.L.G., C.G.B., D.P.D., S.P., S.L.S. contributed to the interpretation and presentation of results in the main manuscript and supplementary documents. N.J.G. wrote the first draft of the manuscript with input from all other authors.

## **Correspondence**

Correspondence and requests for materials should be addressed to Neil Gemmell (e-mail: [neil.gemmell@otago.ac.nz](mailto:neil.gemmell@otago.ac.nz)).

## **Competing Interests**

The authors declare no competing interests.

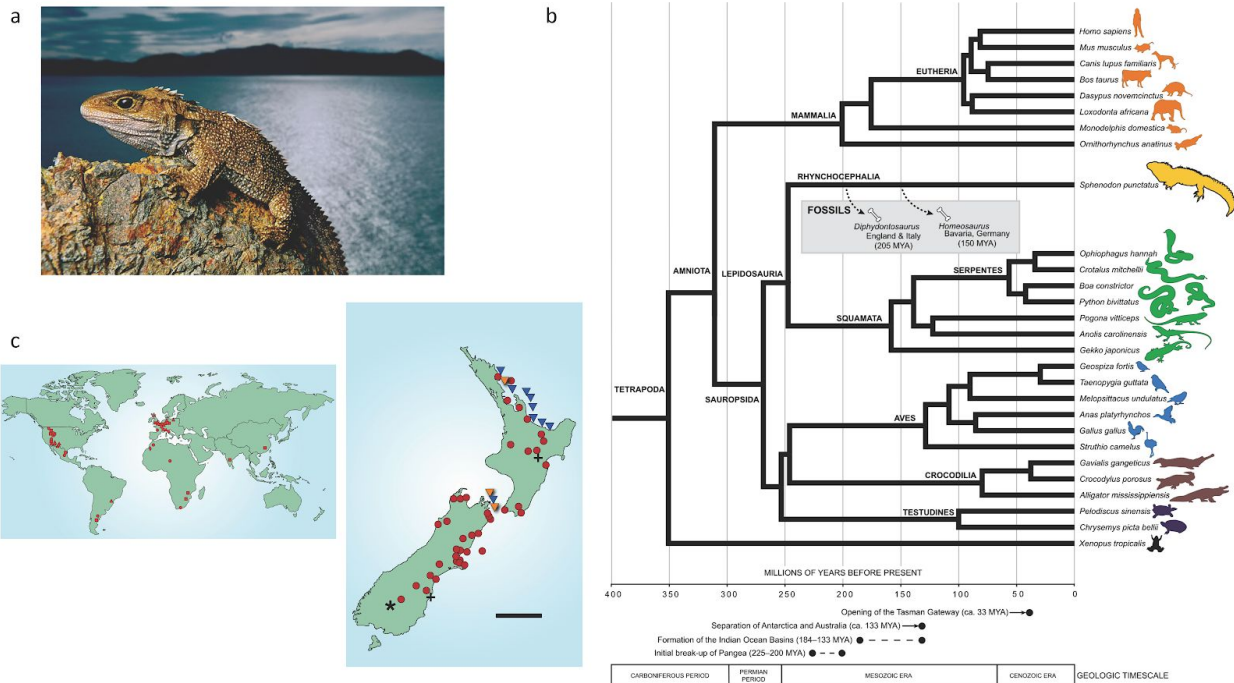


Figure 1. a) The tuatara, *Sphenodon punctatus*, is the sole survivor of the Order Rhynchocephalia. b) The rhynchocephalians appear to have originated in the Early Mesozoic (~250-240 mya) and were common and globally distributed for much of that era (c). Their geographical range progressively contracted after the Early Jurassic (~200-175 mya); the most recent fossil record outside of New Zealand is from Argentina in the Late Cretaceous (~70 mya). c) The last bastion of the rhynchocephalians are 32 islands off the coast of New Zealand, which have recently been augmented by the establishment of ~10 new island or mainland sanctuary populations using translocations. The current global population is estimated at ~100,000 individuals. Rhynchocephalian fossil localities are redrawn from Jones et al. (2009). Legend: triangles, Triassic; squares, Jurassic; filled circles, Cretaceous; diamonds, Palaeocene; asterisk, Miocene; pluses, Pleistocene; open circles, Holocene; down triangles, extant populations; orange down triangles, populations investigated in this study. Scale bar represents 200 Km. Photo credit: Frans Lanting.

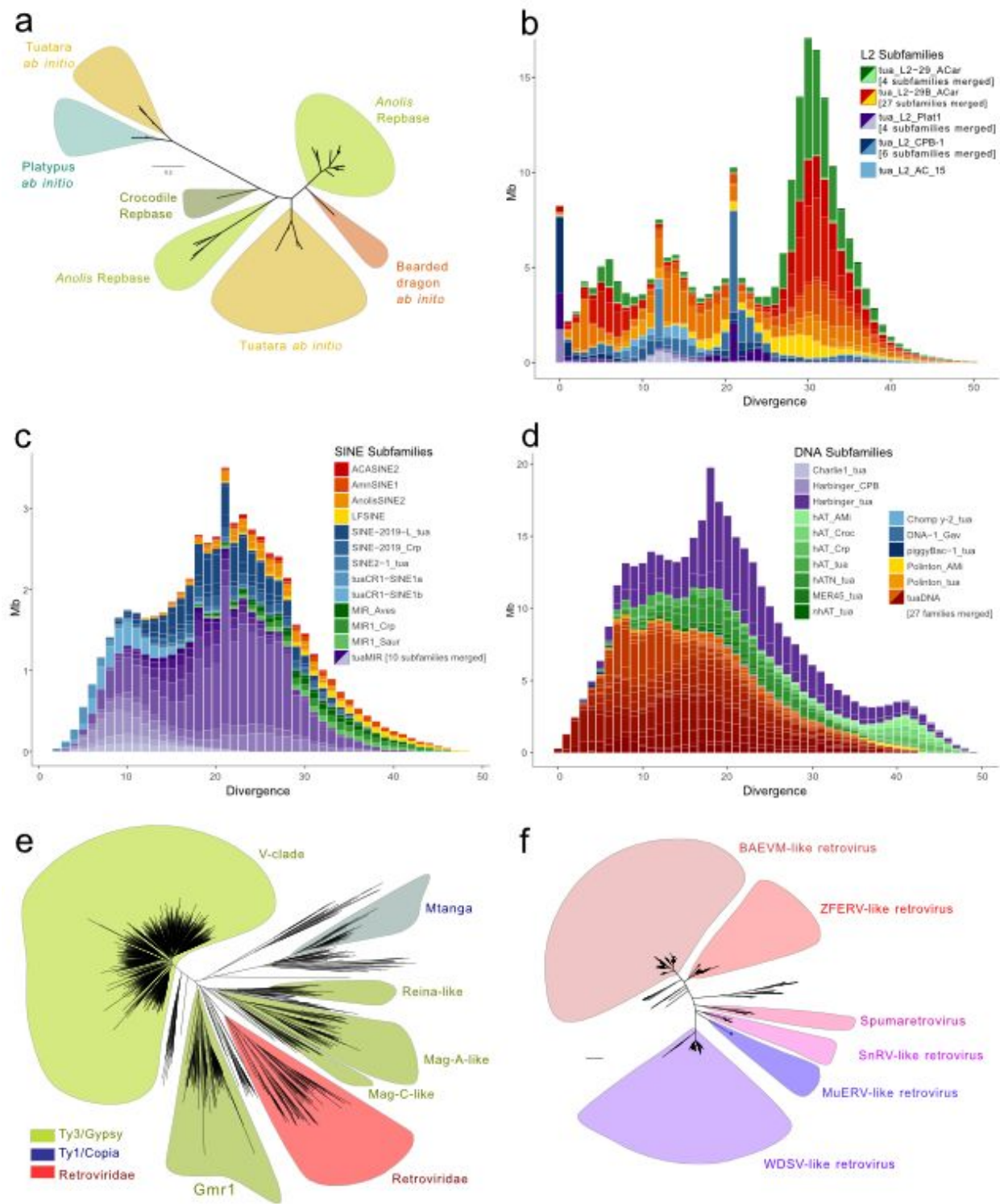


Figure 2. Analysis of the repeat landscape in the tuatara genome identifies unique repeat families, evidence of recent activity and a greater expansion and diversity of

repeats than any other vertebrate. a) Phylogenetic analysis based on the RT domain of L2 repeats identifies two L2 subfamilies; one typical of other lepidosaurs and one that is similar to platypus L2. Phylogeny based on L2 elements >1.5 kbp long with RT domain >200 aa. b) Landscape plot of L2 retrotransposons, the dominant LINEs observed in the tuatara genome, accounting for 10% of the genome. Only LINE subfamilies that occupy more than 1,000 bp are shown. c) Landscape plot of SINE retrotransposons suggests the tuatara genome is dominated by MIR sequences most typically associated with mammals, with tuatara now the amniote genome in which the greatest MIR diversity has been observed. Only SINE subfamilies that occupy more than 1,000 bp are shown. d) Landscape plot of DNA transposons including 24 novel DNA transposon families from tuatara. e) The tuatara genome contains ~7500 full-length long terminal repeat (LTR) retroelements including nearly 450 endogenous retroviruses (ERVs). The general spectrum of tuatara LTR retroelements is comparable to that of other sauropsids, but lacks Bel-Pao LTR retroelements which are present in anole, and has an abundance of Ty1/Copia retroelements. f) Tuatara ERVs span five major retroviral clades and include at least 37 complete spumaretroviruses previously found only in coelacanth, sloth, aye-aye and Cape golden mole. This is the first report of this ancient group of ERVs in a sauropsid.

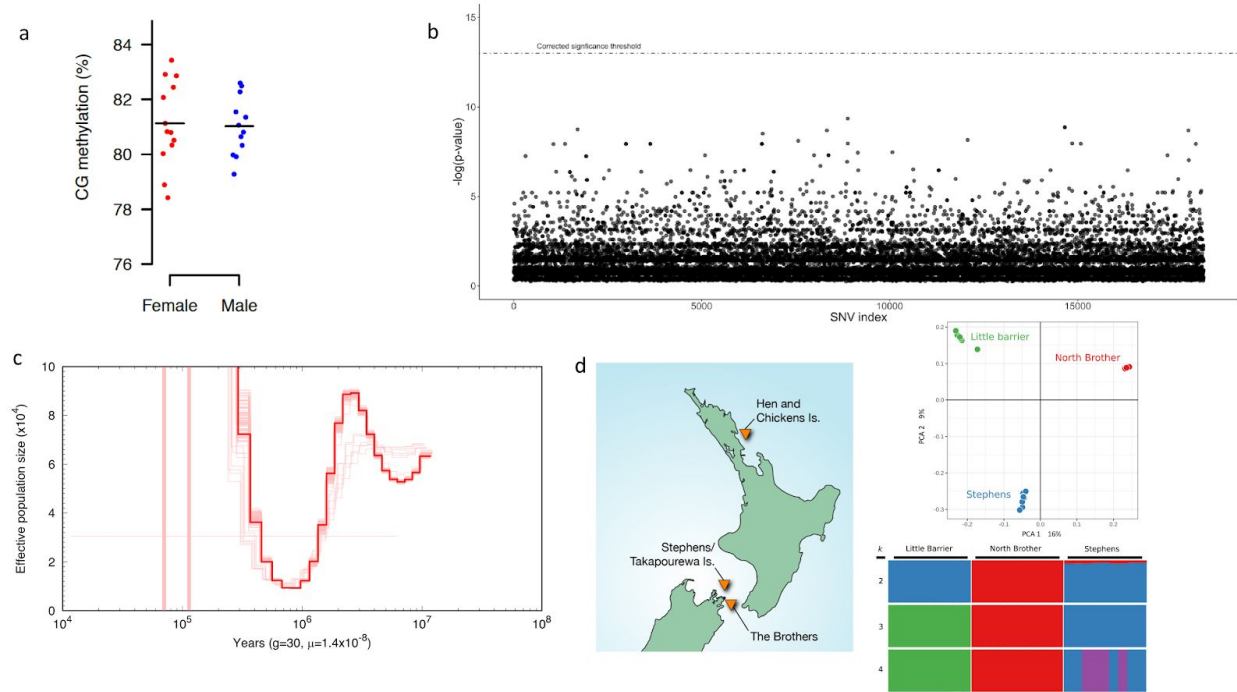
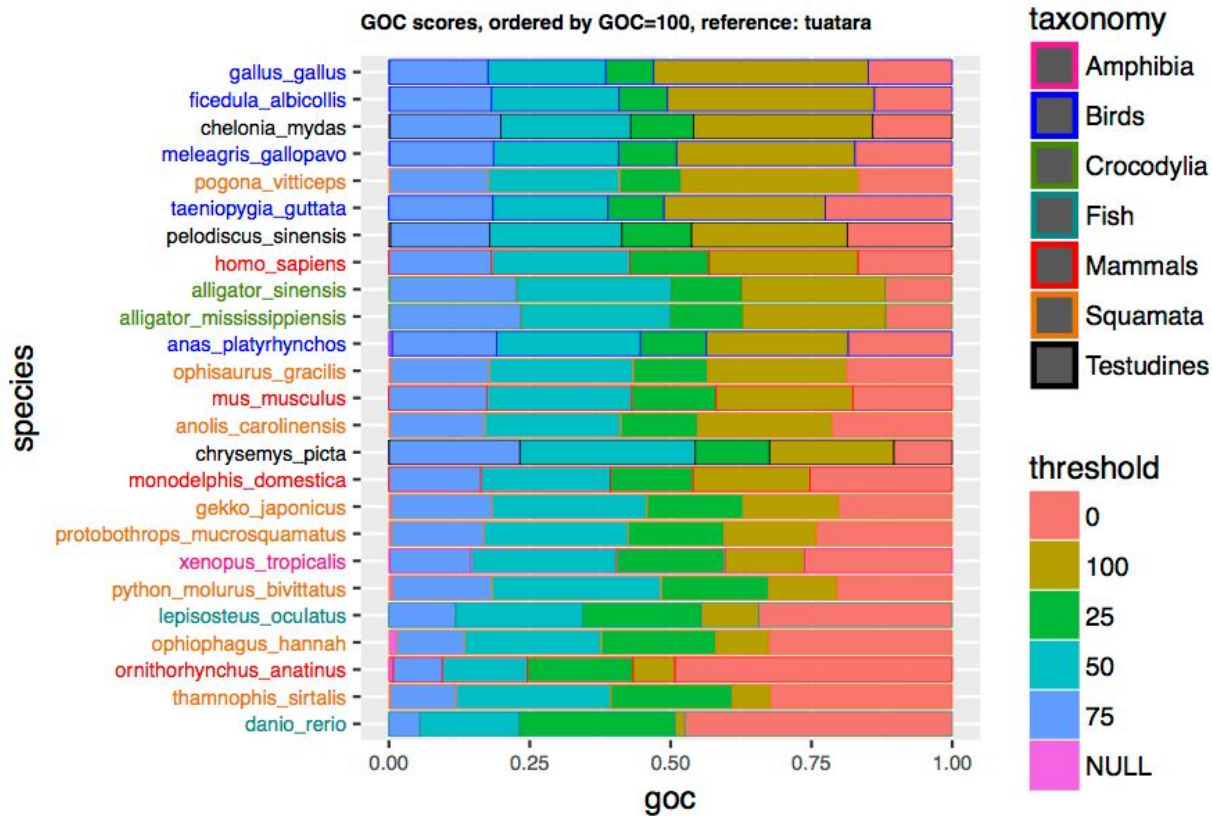


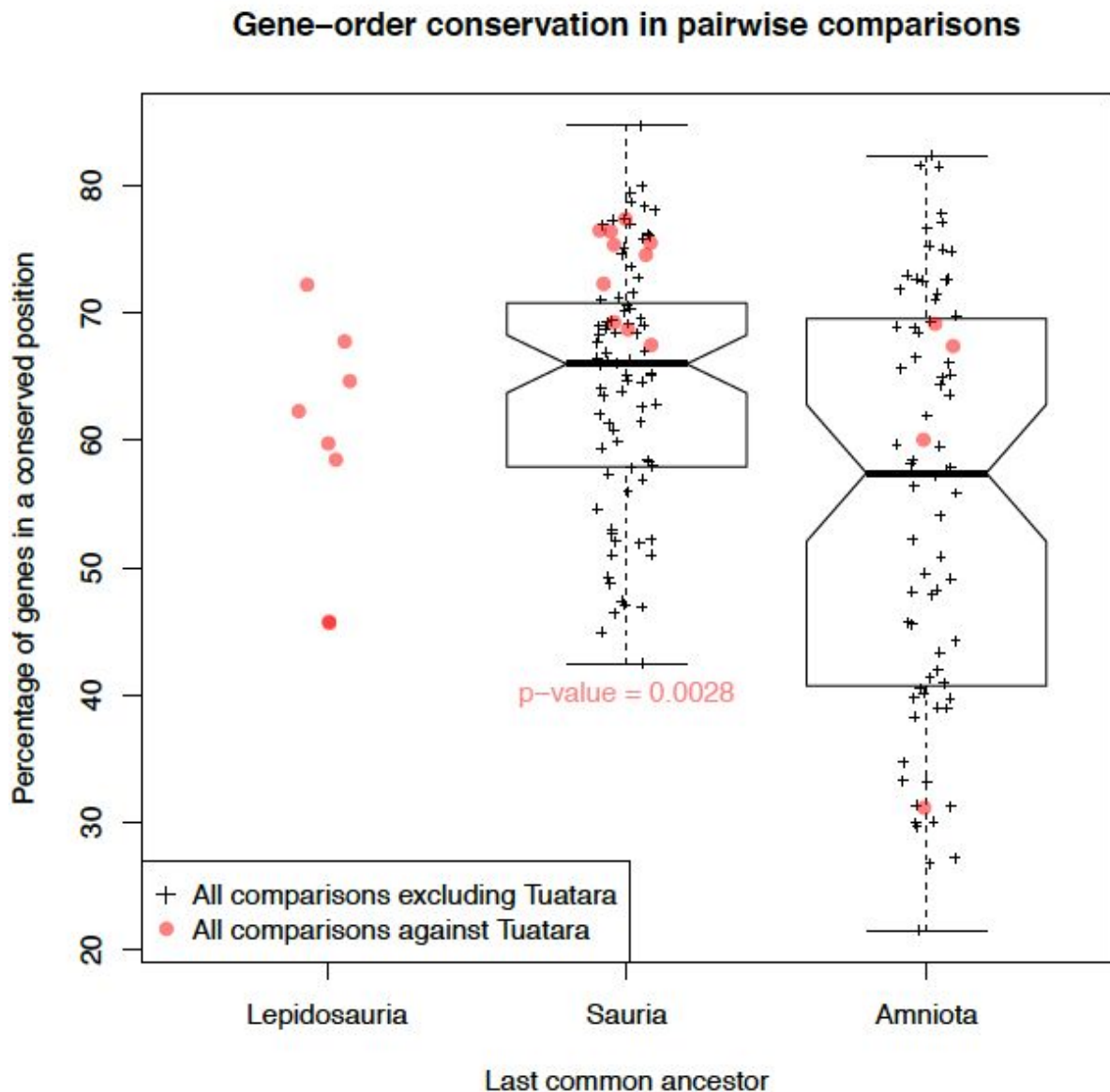
Figure 3. a) Methylation levels in the tuatara genome are high, mean 81%, but show no significant differences among the sexes. b) No single nucleotide variant (SNV) is significantly differentiated with respect to sex in the tuatara genome. The dashed line represents the threshold for statistical significance after accounting for multiple testing. c) PSMC plots of the demographic history of tuatara using a mutation rate of  $1.4 \times 10^{-8}$  substitutions per site per generation and a generation time of 30 years. d) We examined the three known axes of genetic diversity in tuatara: northern New Zealand (Little Barrier Is./ Hauturu), and two islands in Cook Strait (Stephens Is. / Takapourewa; North Brother Island) using genotype by sequencing methods. Principal component analysis and Structure plots demonstrate significant structure among tuatara populations and strongly support prior suggestions that tuatara on the Brothers Island are genetically distinct and warrant separate management.

**Supplementary Figure 1.** Gene Order Conservation score distribution using the tuatara as a reference. Species are ordered by the proportion of top-scoring orthologs.

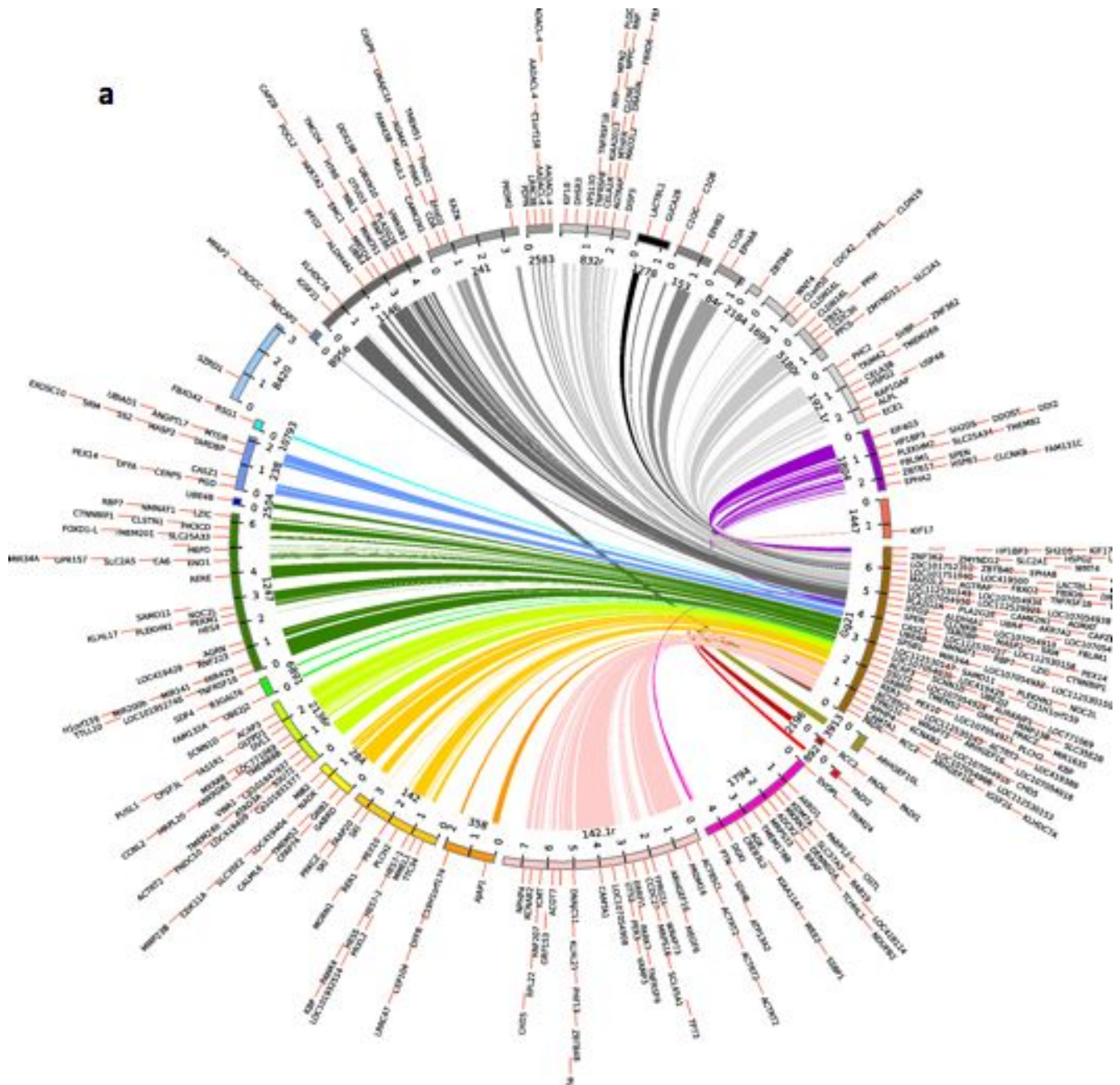




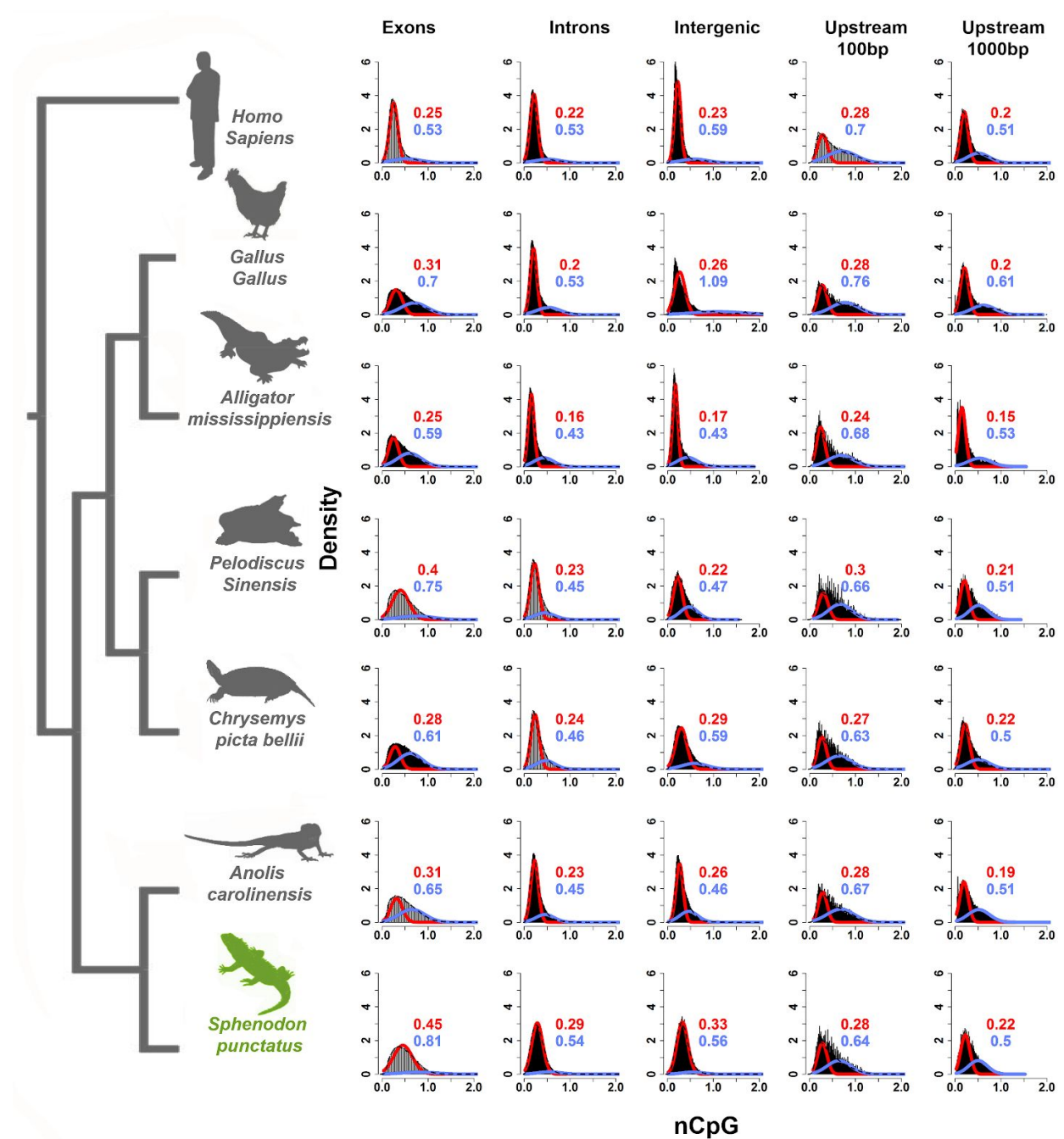
**Supplementary Figure 2.** Gene-order conservation versus divergence time. For each taxonomic grouping, we analyzed the percentage of genes that are found in a conserved position across all pairs of genomes. Pairwise comparisons involving Tuatara are shown in plain red circles, whilst comparisons that do not involve Tuatara are black (boxplot and "+" signs). The conservation of gene-order between Tuatara, birds and turtles is significantly higher ( $p\text{-value} = 2.8\text{e-}3$ ) than that observed between squamates birds and turtles. As Tuatara is the only remaining rhynchocephalian, there is no control distribution for the Lepidosauria ancestor.



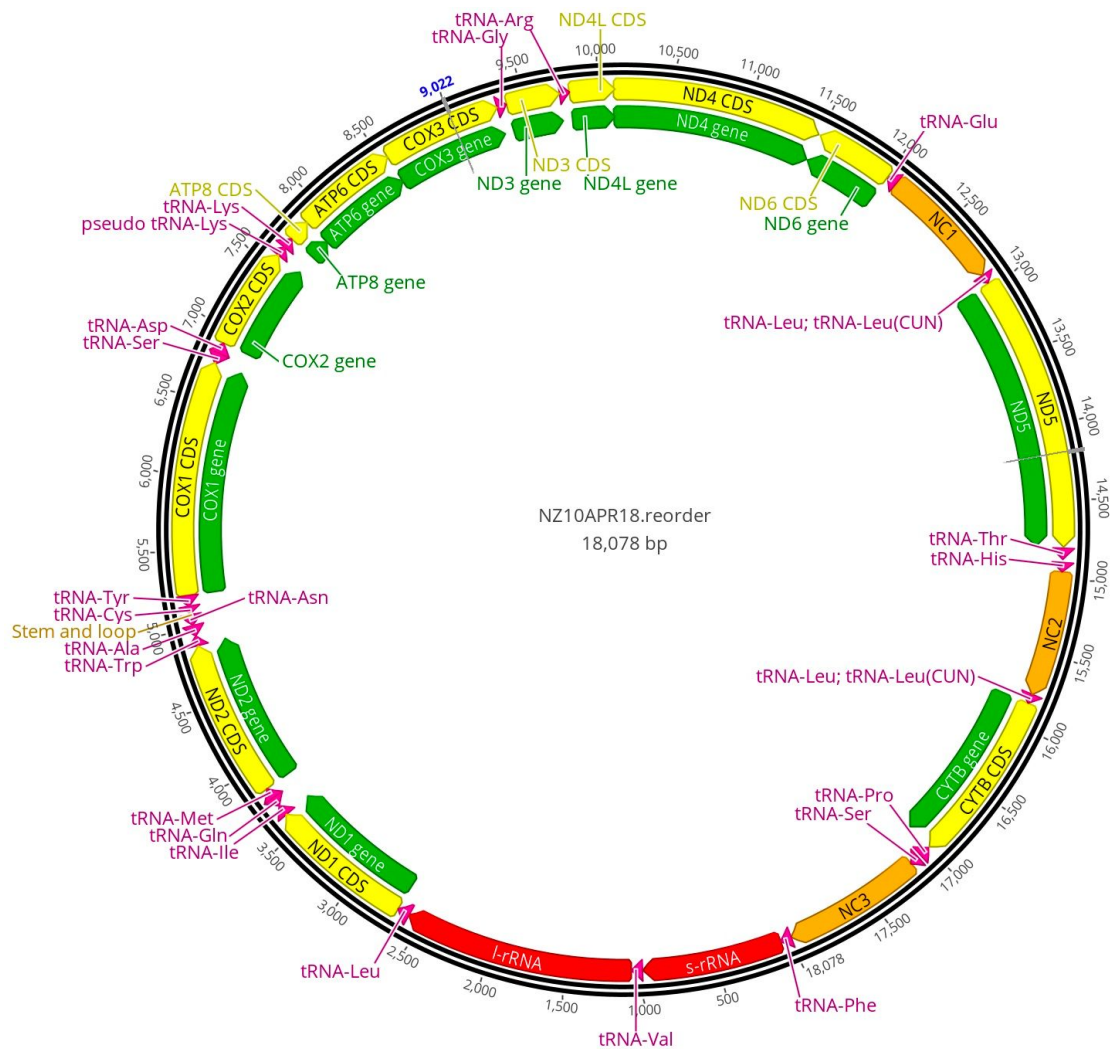
**Supplementary Figure 3. Synteny between *G. gallus* chromosome 21 and tuatara contigs.** Circos plot highlighting the synteny between chicken chromosome 21 (assembly GRCg6a) and multiple tuatara contigs. Gene names shown derive from the chicken assembly.



**Supplementary Figure 4.** Normalized CpG distributions (nCpG) for tuatara and other vertebrates.

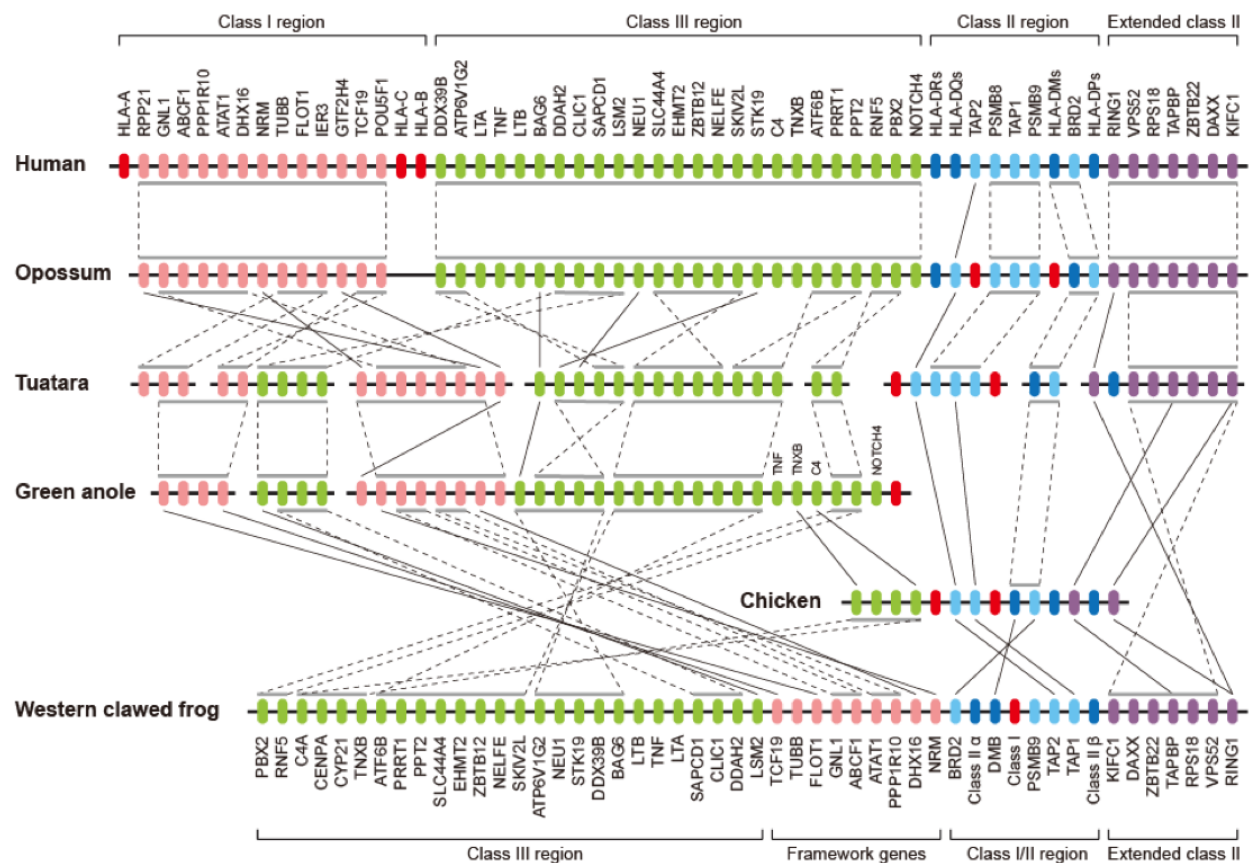


**Supplementary Figure 5.** The mitochondrial (mt) genome in the Lady Alice Island reference animal is 18,078 bp, containing 13 protein coding, 2 rRNA, and 22 tRNA genes, standard among animals contradicting prior reports that 3 genes: ND5, tRNA<sup>Thr</sup>, tRNA<sup>His</sup> were absent. Three non-coding areas with Control Region (heavy strand replication origin, O<sub>H</sub>) features (NC1–3), and two copies of tRNA<sup>Leu</sup>(CUN) adjacent to NC1 and NC2 possess identical or near identical sequence, possibly as a result of concerted evolution. The gene and structure order is: tRNA<sup>Phe</sup>, 12S rRNA, tRNA<sup>Val</sup>, 16S rRNA, tRNA<sup>Leu</sup>(UUR), ND1, tRNA<sup>Ile</sup>, tRNA<sup>Gln</sup>, tRNA<sup>Met</sup>, ND2, tRNA<sup>Trp</sup>, tRNA<sup>Ala</sup>, tRNA<sup>Asn</sup>, O<sub>L</sub>-like structure, tRNA<sup>Cys</sup>, tRNA<sup>Tyr</sup>, COI, tRNA<sup>Ser</sup>(UCN), tRNA<sup>Asp</sup>, COII, pseudo-tRNA<sup>Lys</sup>, tRNA<sup>Lys</sup>, ATP8, ATP6, COIII, tRNA<sup>Gly</sup>, ND3, tRNA<sup>Arg</sup>, ND4L, ND4, ND6, tRNA<sup>Glu</sup>, NC1, tRNA<sup>Leu</sup>(CUN) copy one, ND5, tRNA<sup>Thr</sup>, tRNA<sup>His</sup>, NC2, tRNA<sup>Leu</sup>(CUN) copy two, Cytb, tRNA<sup>Pro</sup>, tRNA<sup>Ser</sup>(AGY), and NC3.

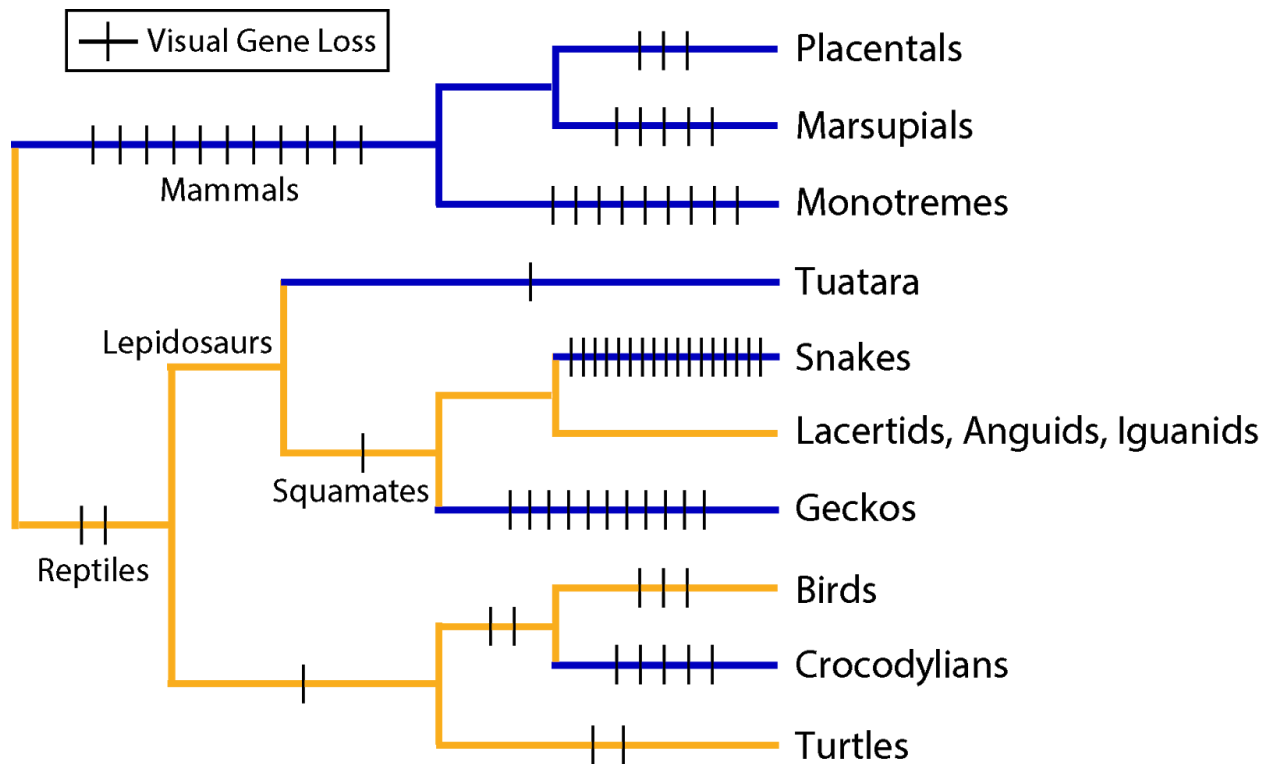




**Supplementary Figure 6.** Comparative analysis of the MHC core region. Only genes that were annotated in the tuatara genome were included in the analysis. Orthologs between species are connected by a solid line; the grey bars above/below genes indicate syntenic blocks and are linked by dashed lines between species. Anolis class I/II and extended class II regions are not shown due to the high degree of genome assembly fragmentation in these regions. Color legend: red – class I genes, pink – class I region framework genes, green – class III genes, dark blue – class II genes, light blue – class II region framework genes, purple – extended class II genes.

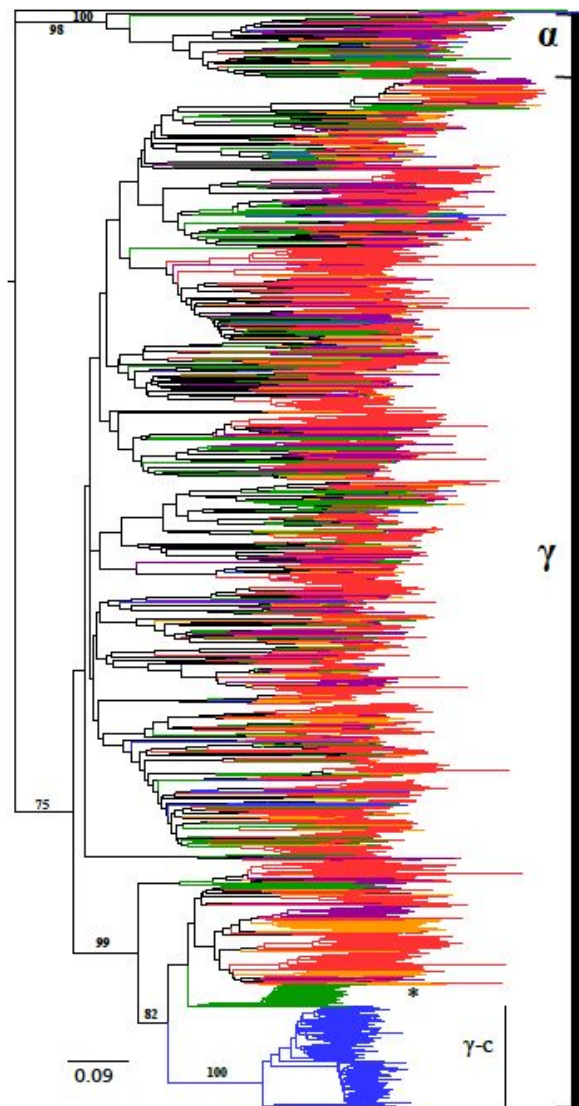


**Supplementary Figure 7.** Phylogenetic tree of reptiles depicting inferred visual gene losses. Lineages are coloured based on a rough approximation of their ancestral activity pattern (blue, nocturnal; yellow, diurnal). Note that that tuatara lineage has experienced some of the lowest rates of gene loss despite a nocturnal ancestry, which in other lineages is associated with increased gene loss.

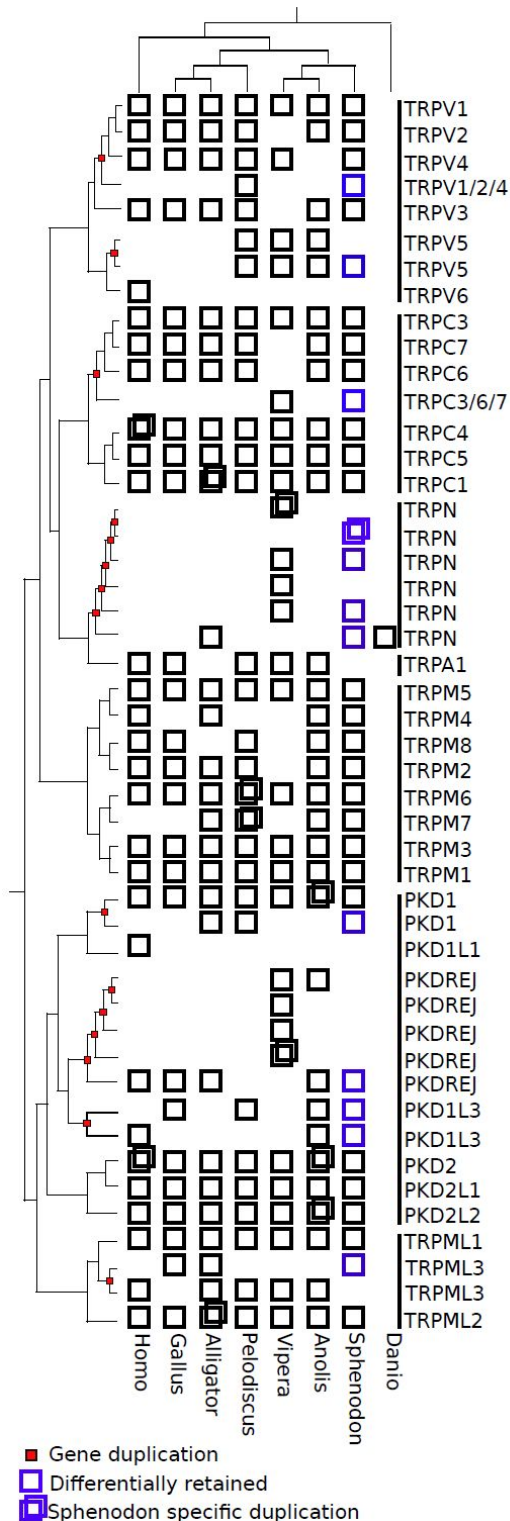




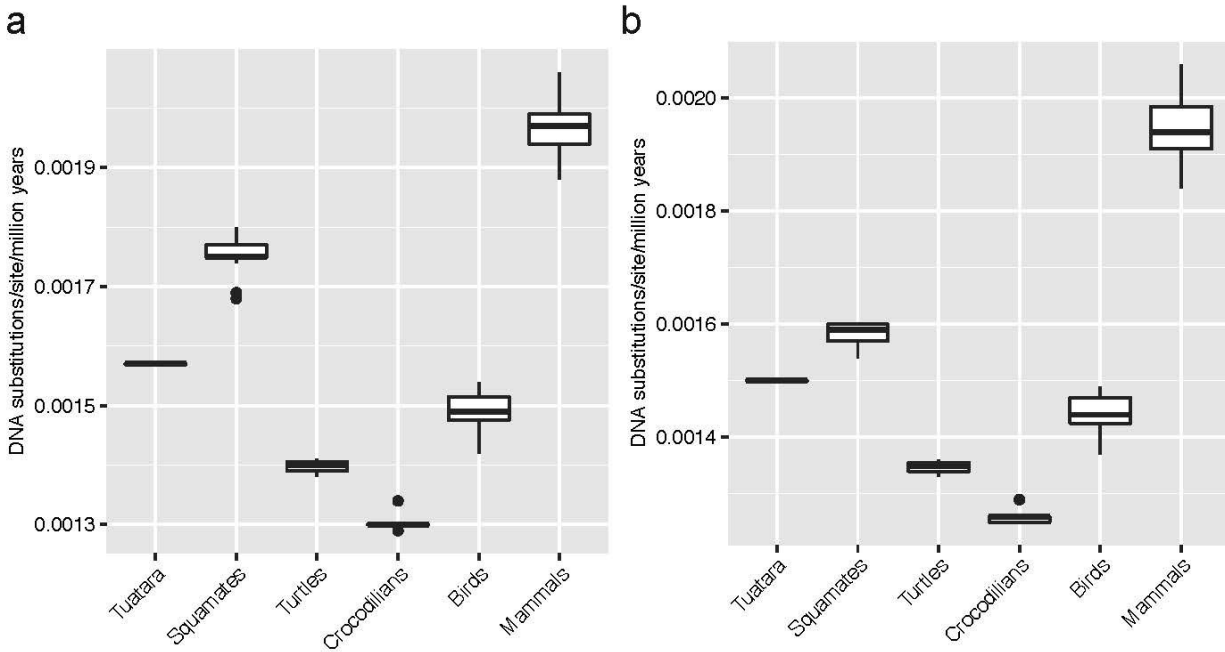
**Supplementary Figure 8.** The evolutionary history of terrestrial Sauropsid ORs was inferred using the Neighbor-joining method. The unrooted tree contains 3213 amino acid sequences. Branches are coloured according to the following categories: Green – tuatara, Blue – birds (*Gallus gallus*, *Taeniopygia guttata*), Red – snakes (*Notechis scutatus*, *Ophiophagus Hannah*, *Protobothrops mucrosquamatus*, *Pseudonaja textilis*, *Python bivittatus* and *Thamnophis sirtalis*), Orange – lizards (*Anolis carolinesis*, *Pogona vitticeps*) and Purple – gecko (*Gekko japonicas*). Bootstrap support values above 75% (1000 replicates) are indicated for major branch splits relating to the different OR groups and branches leading to the species-specific OR expansions in birds (group  $\gamma$ -c) and tuatara (\*).



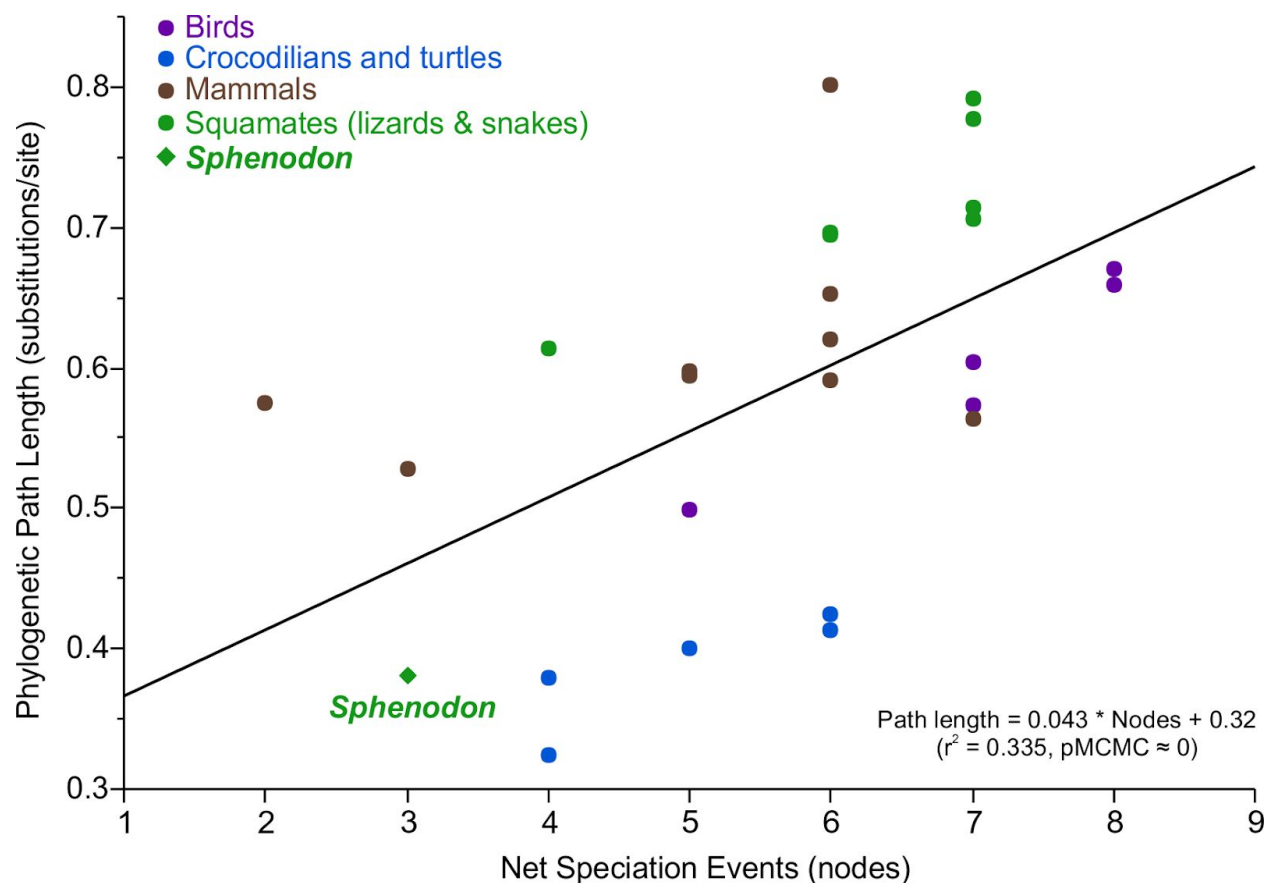
**Supplementary Figure 9.** The repertoire of TRP genes identified in tuatara (*Sphenodon punctatus*) and other six vertebrate species: lizard (*Anolis carolinensis*), viper (*Vipera berus*), turtle (*Pelodiscus sinensis*), alligator (*Alligator mississippiensis*), chicken (*Gallus gallus*), and human (*Homo sapiens*). Small red squares on nodes indicate gene duplications, blue boxes indicate gene differential retentions and duplicated boxes indicate species-specific recent duplicates. Empty spaces are differential losses.



**Supplementary Figure 10. Estimated DNA substitution rates of amniote clades based on fourfold degenerate sites.** Boxplots showing the distribution of estimated substitution rates by clade using semiparametric penalized likelihood in r8s with (a) fossil constraints from Benton et al. (Benton et al. 2015) and Head (J. J. Head 2015) and (b) median TMRCA estimates from timetree.org.



**Supplementary Figure 11. Testing for evidence of punctuated evolution.** The process of punctuated genome evolution predicts that the amount of evolution in a given species' genome should correlate with the net number of speciation events. We used Bayesian phylogenetic generalized least squares to regress the total phylogenetic path length (of four-fold degenerate sites) on the net number of speciation events (nodes in a phylogenetic tree). We find strong evidence for punctuated evolution, which accounts for 33.5% ( $r^2$ ; 95% credible interval = 0.34 to 0.38) of deviation from the molecular clock at four-fold degenerate sites.



**Supplementary Table 1. Assembly statistics and quality metrics for the tuatara genome.**

	ALLPATHS-LG assembly	Final assembly
Number of scaffolds	48545	16537
Total size of scaffolds	4338416404	4272217537
Longest scaffold	6644062	29987930
N50 scaffold length	322768	3052611
Number of contigs	580919	270321
Longest contig	141552	347348
N50 contig length	10517	27421
CEGMA complete count	89	98
CEGMA complete %	35.89%	39.52%
CEGMA partial count	209	210
CEGMA partial %	84.27%	84.68%
Complete BUSCOs	2826	3381
Complete %	71.5	85.6
Complete and single-copy BUSCOs	2800	3337
Complete and duplicated BUSCOs	26	44
Fragmented BUSCOs	674	324
Missing BUSCOs	450	245
Total BUSCO groups searched	3950	3950

**Supplementary Table 2: Copy number and fraction of tuatara genome covered by interspersed repeats.**

	<i>ab initio library</i>			<i>RMD library</i>		
<b>Group</b>	<b>Number</b>	<b>Total bp</b>	<b>Percentage</b>	<b>Number</b>	<b>Total bp</b>	<b>Percentage</b>
Non-LTR retrotransposons						
LINE L2	1,479,425	428,138,325	10.022	1,149,043	491,620,677	11.507
LINE CR1	815,011	162,227,767	3.797	500,647	149,462,580	3.498
LINE RTE	138,126	31,151,317	0.729	184,481	97,841,699	2.290
LINE L1	239,829	30,179,939	0.706	49,905	29,081,314	0.681
SINE	329,988	32,931,264	0.771	184,701	27,664,250	0.647
Others	209,995	21,413,494	0.501	46,718	21,910,444	0.514
	3,212,374	706,042,106	16.526	2,115,495	817,580,964	19.137
ERV						
ERV1	142,313	17,564,350	0.411	20,968	9,453,167	0.221
Others	145,005	15,355,898	0.359	7,149	1,948,963	0.046
	287,318	32,920,248	0.770	28,117	11,402,130	0.267
DNA transposons						
hAT	439716	57197609	1.339	668,329	194,724,290	4.558
Harbinger	435048	80071368	1.874	463,162	180,461,900	4.224
Others	1,564,326	224,599,096	5.257	167,048	36,855,177	0.863
	2439090	361868073	8.470	1,298,539	412,041,367	9.645
LTRs						
DIRS	256,478	58,538,598	1.370	130,860	78,246,082	1.831
Gypsy	570,277	83,980,917	1.966	98,683	62,432,545	1.461
LTR others	221,043	26,438,648	0.619	259,570	115,859,366	2.713
	1,047,798	168,958,163	3.955	489,113	256,537,993	6.005
Others	599,649	67,376,623	1.578	316,948	58,309,055	1.364
Well-annotated	7,586,229	1,337,165,213	31.299	4,248,212	1,555,871,509	36.418
Unknown	7875491	1424963821	33.354	3,203,800	637,030,810	14.911
<b>Total</b>	<b>15,461,720</b>	<b>2,762,129,034</b>	<b>64.653</b>	<b>7,452,012</b>	<b>2,192,902,319</b>	<b>51.329</b>



**Supplementary Table 3. Vertebrate sex-determining genes found in the tuatara genome and their substitution rates.** We computed substitution rates,  $d_N$ ,  $d_S$ , and  $d_N/d_S$  for genes previously identified as playing a primary role in sex determination (Supplementary materials 16). For eight sex-determining genes we did not identify homologous sequence in the tuatara in the multiple species alignment (gray). This does not mean that they don't exist in the tuatara genome, but that they were not found in the current version of the multiple alignments, perhaps because of high divergence.

TranscriptID	GeneSymbol	dN	dS	dNdS	TreeLength	Branch length in tuatara
ENSACAT00000009605	AR (DHTR)	0.0900	2.0497	0.0439	1.9847	0.0571163
ENSACAT00000008263	ARX	0.0749	1.4746	0.0508	1.3473	0.0348331
ENSACAT00000000495	CIRBP	0.0461	2.4627	0.0187	Tr2.1876	0.0636031
ENSACAT00000004647	CSDC2	0.0933	1.4834	0.0629	1.5226	0.106651
ENSACAT00000003256	CSDE1	0.0464	0.8424	0.0551	0.8258	0.0575427
ENSACAT00000015874	DHCR7	0.3298	1.6549	0.1993	2.1308	0.134717
ENSACAT00000010360	EMX2	0.1142	1.8239	0.0626	1.7465	0.0348258
ENSACAT00000006243	ESR1	0.1894	1.4322	0.1322	1.6464	0.113636
ENSACAT00000001292	FEM1B	0.0809	2.3919	0.0338	2.2297	0.0435242
ENSACAT00000015028	JAKMIP1	0.3526	2.9505	0.1195	3.2637	0.056045
ENSACAT00000014908	JAKMIP2	0.0494	1.5791	0.0313	1.4699	0.0361767
ENSACAT00000008030	JAKMIP3	0.0968	1.2884	0.0751	1.3204	0.095763
ENSACAT00000002517	LHX1/LIM1	0.1131	2.3564	0.0480	2.2445	0.0276626
ENSACAT00000014613	LHX9	0.0500	0.9320	0.0536	0.8877	0.0480989
ENSACAT00000030595	M33/TUSC3	0.0155	1.4471	0.0107	1.2941	0.0125199
ENSACAT00000014130	PAX2	0.0271	0.8004	0.0339	0.7732	0.0621367
ENSACAT00000030504	PDGFB	0.3974	1.2046	0.3299	1.9014	0.280557
ENSACAT00000017280	RSPO1	0.2540	1.0624	0.2391	1.4557	0.192456
ENSACAT00000016485	SF1	0.0415	2.0755	0.0200	1.8859	0.00712568
ENSACAT00000008756	SOX9	0.1188	1.8735	0.0634	1.8038	0.129618
ENSACAT00000018021	STAT3	0.0357	2.1405	0.0167	1.9370	0.0282673
ENSACAT00000005401	UBA1	0.1646	1.9622	0.0839	2.0987	0.11493
ENSACAT00000012577	WNT4	0.4560	2.9148	0.1564	3.2907	0.0161123
ENSACAT00000002516	WT1	0.0986	1.2456	0.0792	1.3415	0.110817
ENSACAT00000009898	ZFPM2/FOG2	0.2839	1.3392	0.2120	1.7684	0.227515
ENSACAT00000015597	AMH	0.5398	1.9654	0.2747	2.8312	NA
ENSACAT00000002340	DMRT1	0.0895	0.7924	0.1129	0.8935	NA
ENSACAT00000001907	GATA4	0.1677	1.5525	0.1080	1.6969	NA
ENSACAT00000007239	MSL2	0.0344	0.5761	0.0597	0.5901	NA
ENSACAT00000010492	SDC1	0.7469	1.5435	0.4839	2.9324	NA
ENSACAT00000006298	STAT1	0.1039	0.8107	0.1282	0.9141	NA
ENSACAT00000025425	STAT2	0.3812	1.1877	0.3210	1.8272	NA
ENSACAT00000009792	STAT6	0.2880	1.1722	0.2457	1.6266	NA

**Supplementary Table 4. Tuatara–squamate gene tree evolution compared between all genes and sex determining genes.** A log-likelihood test comparing the likelihood of a model where the rate of evolution was the same on all branches compared to the likelihood of a model where there was a different rate of evolution on the tuatara branch was conducted for all genes present in both the tuatara and anole (since it was the reference for the alignment) and were present in a minimum of three species.

		<b>Genes which Are Significantly Different from the Null Model</b>	<b>Total Number of Genes</b>
<b>All Genes</b>	<b>Percentage</b>	75.91%	9,586
	<b>Number of Genes</b>	7,277	
<b>Sex Determining Genes</b>	<b>Percentage</b>	84.0%	25
	<b>Number of Genes</b>	21	
<b>Non-Sex-Determining Genes</b>	<b>Percentage</b>	75.89%	9,561
	<b>Number of Genes</b>	7,256	