# Jurassic Genomics: Using the Tuatara Genome to re-examine the basal reptilian phylogeny

Developing a new phylogenomic-scale genetic alignment, in regard to the assessment of tuatara and the higher level amniote evolutionary tree.

Allan Wilson Center Summer Studentship Report 26/02/2013
Alexander Boast
Supervisor: Neil Gemmell
Otago University

## Introduction and Aims of this project.

The Tuatara (*Sphenodon* spp.) endemic to New Zealand are the only extant members of the Rhynchocephalia, comprising one of the five extant major groupings of Sauropsids (birds and reptiles). Rhynchocephalians, (the majority found within the suborder Sphenodontia) are known from several fossil species up to Triassic in age from both Hemispheres, and largely retained a conservative, unspecialised 'lizard-like' morphology similar to the modern forms throughout their history (Wu 1994, Evans & Borsuk-Białynicka 2009). The group disappeared from the global fossil record shortly preceding the K/T mass extinction 65 Ma, and it is unknown whether they persisted outside New Zealand during the Tertiary (e.g. Apesteguía & Roughier 2007, Jones *et al.* 2009). Rhynchocephalians are widely accepted as the sister group to the extant ~8,000 species of squamates (snakes, lizards, amphisbaenians and mosasaurs) forming the Lepidosauromorpha as supported by several morphological synapomorphies and phylogenetic studies (e.g. Hedges 1994, Hedges & Poling 1999, Rest *et al*. 2003, Crawford *et al.* 2012, Mulcahy *et al*. 2012). The two groups are likely to have diverged near to the Permo-Triassic boundary 252 Ma, close to the radiation of sauropsids (Evans 2003, Evans & Jones 2010, Rest *et al*. 2012, Mulcahy *et al.* 2012). For these reasons Tuatara are popularly known as 'living fossils', and constitute a taxon of considerable evolutionarily importance and a national conservation priority.

With the upcoming developments in genomics and next-generation sequencing, it has become increasing possible to develop large amounts of genetic material of non-model organisms, and thus use of high numbers of orthologous regions in contemporary phylogenetic studies. 'Phylogenomics' are now being used in the exploration of poorly understood and particularly deep evolutionary relationships, such as interfamily relationships of neognathous birds (Hacket *et al.* 2008, Künster et al. 2010) and mammals (McCormack *et al.* 2011, Reis *et al*. 2012); to studies at the kingdom or domain level (e.g. Hampl *et al.* 2009). The radiation of the major amniote lineages has been explored in several molecular studies including a number at a phylogenomic level (Tzika *et al*. 2011, Chiari *et al*. 2012, Crawford *et al.* 2012). The recent article by Chiari et al. (2012) as well as the brain-transcriptome assemblies of Tzika et al. (2011, http://www.reptilian-transcriptomes.org), developed large alignments from a synthesis of completed genome projects and *de novo* transcriptome assemblies, although Tuatara were not explored in either. There has since been the release of a Tuatara developmental transcriptome (Miller *et al*. 2012), as well as the recent completion of the first turtle genome (*Pelodiscus sinensis*). The relationship of the major amniote lineages still remains largely equivocal, and as a result there is a significant opportunity towards the development of a new and more complete synthesis.

The initial aims of this project have been to use the developmental transcriptome of Tuatara (Miller *et al*. 2012), and experiment in the development of an alignment using large and recently released genetic data assemblies. This project at the time of this report has now progressed to a rough completion of this alignment from a number of differing sources; although this is still yet to undergo a phylogenetic analysis.

## Preliminary Approach and project development.

This particular study was largely based off the approach used by Chiari *et al*. (2012), who used a total of 248 orthologous coding regions of 16 taxa in the exploring the phylogenetic relationships of turtles. This included several species available from ensembl, new transcriptomic assemblies from other sources (Künster *et al*. 2010, Castoe *et al.* 2011), as well as new material developed from the team (some of which was previously published in Gayral *et al*. 2011). However the pipeline used an approach whereby orthologs were constricted to whole taxonomic groupings as opposed to individual species, and for no single region are all species represented as a result. The alignment used is available on Dryad Digital Repository (http://datadryad.org/).

An initial approach was to experiment developing orthologs using a one-way tBLASTx search against *Gallus* from the alignment from Chiari *et al*. to the Tuatara transcriptome, which resulted in 87 orthologs. Experimental neighbor-joining analyses confirmed a nested position of Tuatara within the tree obtained. A second approach was using the recently released set of reptile brain-transcriptome assemblies released by *Tzika et al*. (*Crocodylus niloticus, Elaphe guttata, Pogona vitticeps and Trachemys scripta*). A reciprocal tBLASTx search from the cDNA transcript of *Gallus gallus* against these four species, with Tuatara returned a total of 158 orthologous regions for all taxa.

It was decided to construct a new alignment, utilising material from genome projects on ensembl and a number of *de novo* transcriptome assemblies. Additionally, the use of merging distantly related taxa in single sequences as explored in Tzika *et al.*, or using a gapped dataset as in Chiari *et al*. was decided not to be used. A large number of trial runs were attempted, initially incorporating the transcriptomes from Tzika *et al.*, however after consistently poor returns these assemblies were omitted from the final analysis. In addition, contact was made with the team responsible for the transcriptomes released in Gayral *et al*. (2011) and Chiari *et al*. (2012), to which we were granted access.

## Methods and Preliminary Results.

A pipeline closely based off that of Chiari *et al.* (2012) was devised to develop a new set of orthologs from a synthesis of several differing sources. The cDNA sequences from the chicken (*Gallus gallus*) genome were used to find and download orthologous coding regions as predicted by the EnsemblCompara phylogenetic assessment of orthology (Vilella *et al*. 2009) from completed genome projects using a customized perl script. The taxon set covers most amniote lineages (Ensembl release 70): *Homo sapiens*, *Monodelphis domestica*, *Ornithoryhnchus anatinus*, *Anolis carolinensis*, *Pelodiscus sinensis* and *Taeniopygia guttata*; as well as one amphibian, *Xenopus tropicalis*, and the coelacanth, *Latimeria chalumnae* to serve as outgroups; of which a total of 7,807 orthologous coding regions were found shared between all nine taxa (see Table. 1).

For the selection of additional taxa, reciprocal tBLASTx searches were performed ($1e^{-100}$ expectancy value, >50% identity, default parameters), on a number of selected transcriptome assemblies. The use of closely related taxa was avoided to cover the widest taxonomic range without compromising the final ortholog count, and these were performed sequentially and ranked according to priority and taxonomic significance (see Table 2). This was continued until a compromise between taxonomic coverage and ortholog count was reached. The final dataset as found here comprised 195 orthologs shared between a total of 14 taxa covering all major lineages, as well as incorporating representatives of each of the major groupings of mammals and turtles, and also covering a significant span of squamate and avian diversity.

## Project continuation and future aims.

It needs to be noted that although this is currently likely the final set of taxa to be used in the complete phylogeny, there are a number of unassembled cDNA libraries available on the sequence read archive (http://www.ncbi.nlm.nih.gov/sra), which require extensive analysis prior to use. Among these include large multi-tissue transcriptomes of a Crocodilian, *Alligator mississipiensis* (Künster *et al.* 2011), and two paleognathous birds, *Apteryx australis* (Subramanian *et al.* 2010) and *Dromaius novaehollandiae* (Künster *et al.* 2011). In addition the crocodile genome project (St. John *et al.* 2012, http://crocgenomes.org/) is currently developing assemblies from a representative of each of the three extant crocodilian families; *Alligator mississipiensis* (Alligatoroidea), *Crocodylys porosus* (Crocodyloidea) and *Gavialis gangeticus* (Gavialidae), and we are in negotiations with the team involved.

Our project was constrained to the necessity of a crocodilian for complete taxonomic coverage and the use of a restrictive jaw-tissue transcriptome of *Caiman crocodylus*, which appears to have significantly affected on our final ortholog count. Pending on time, and logistical limitations, we would hope to include one or more of the more complete Crocodilian cDNA libraries to our assembly, which should result in a much larger set of orthologs. Additionally our dataset does not include any paleognathous birds (one of the two avian superorders), and we have compromised by use of two neognaths; a galliform (*Gallus*) and a derived passerine (*Taeniopygia*); aiming to cover a wide span of genetic diversity due to the widely accepted basal relationship of galliformes in the neognathidae (e.g. Hackett *et al.* 2008). In addition the blast results as shown were those obtained by the author, and my colleague Kim Rutherford running a different script has obtained a lower number using stricter criteria and accounting for possible duplicates, which suggests a smaller final set will be used than indicated here.

Once the final dataset is complete, we will be utilizing the NeSI supercomputer (New Zealand eScience Infrastructure), for use of Bayesian and maximum likelihood methodologies to develop a phylogenetic tree. A molecular clock may be established pending on logistical developments. The results of this analysis will aim to be drafted for publication in a peer-reviewed journal early this year.

## References

Apestiguía S, Roughier GW (2007). A Late Campanian Sphenodontid Maxilla from Northern Patagonia. *American Museum Noviates* **3581:** 1-11

Castoe TA, Fox SE, de Koning J, Poole AW, Daza JM, Smith EN, Mockler TC, Secor SM, Pollock DD (2011). A multi-organ transcriptome resource for the Burmese Python (*Python molorus bivittatus*). *BMC Research Notes* **4:** 310

Chiari Y, Cahais V, Galtier N, Delsuc F (2012). Phylogenomic analyses support the position of turtles as the sister group of birds and crocodiles (Archosauria). *BMC Biology* **10:** 65

Crawford NG, Faircloth BC, McCormack JE, Brumfield RT, Winker K, Glenn TC (2012). More than 1,000 ultraconserved elements provide evidence that turtles are the sister group of archosaurs. *Biology Letters* doi:10.1098/rsbl.2012.0331

Evans SE (2003). At the feet of the dinosaurs: the early history and radiation of lizards. *Biological Reviews* **78:** 513-551

Evans SE, Jones MEH (2010). The Origin, Early History and Diversification of Lepidosauromorph Reptiles. in S. Bandyopadhyay (ed.), New Aspects of Mesozoic Biodiversity. Springer-Verlag Berlin Heidelberg. 2010.

Evans SE, Borsuk-Białynicka M (2009). A small lepidosauromorph reptile from the Early Triassic of Poland. *Paleonologia Polonica* **65:** 179-202

Hackett SJ, Kimball RT, Reddy S, Bowie RCK, Braun EL, Braun MJ, Chojnowski JL, Cox WA, Han K-L, Harshman J, Huddleston CJ, Marks BD, Miglia KJ, Moore WE, Sheldon FH, Steadman DW, Witt CC, Yuri T (2008). A Phylogenomic Study of Birds Reveals Their Evolutionary History. *Science* **320:** 1763-1767

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ (2009). Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings of the National Academy of Sciences of the USA* **106:** 3859-3864

Hedges SB (1994). Molecular evidence for the origin of birds. *Proceedings of the National Academy of Sciences of the USA* **91:** 2621-2624

Hedges SB, Poling LL (1999). A Molecular Phylogeny of Reptiles. *Science* **283:** 998-1001

Jones MEH, Tennyson AJD, Worthy JP, Evans SE, Worthy TH (2009). A sphenodontine (Rhynchocephalia) from the Miocene of New Zealand and palaeobiogeography of the tuatara (*Sphenodon*). *Proceedings of the Royal Society B: Biological Sciences* **276:** 1385-1390

Gayral P, Weinert L, Chiari Y, Tsagkogeorga G, Ballenghien M, Galtier N (2011). Next-generation sequencing of transcriptomes: a guide to RNA isolation in nonmodel animals. *Molecular Ecology Resources* **11:** 650-661

Künster A, Wolf JBW, Backström N, Whitney O, Balakrishnan C, Day L, Edwards SV, Janes ED, Schlinger BA, Wilson RK, Jarvis ED, Warren WC, Ellegren H (2010). Comparative genomics based on massive parallel transcriptome sequencing reveals patterns of substitution and selection across 10 bird species. *Molecular Ecology* **19:** 266-276

McCormack JD, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC (2011). Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. Genome Research **22:** 246-754

Miller HC, Biggs PJ, Voelckel C, Nelson NJ (2012). De novo sequence assembly and characterisation of a partial transcriptome for an evolutionarily distinct reptile, the tuatara (*Sphenodon punctatus*). *BMC Genomics* **13:** 439

Mulcahy DG, Noonan BP, Moss T, Townsend TM, TW Reeder, Sites JW Jr, Wiens JJ (2012). Estimating divergence dates and evaluating dating methods using phylogenomic and mitochondrial data in squamate reptiles. *Molecular Phylogenetics and Evolution* **65:** 974-991

Reis M dos, Inoe J, Hasegawa M, Asher RJ, Donoghue PCJ, Yang Z (2012). Phylogenomic datasets provide both precision and accuracy in estimating the timescale of placental mammal phylogeny. *Proceedings of the Royal Society B: Biological Sciences* **279:** 3491-3500

Rest JS, Ast JC, Austin CC, Wadell PJ, Tibbetts EA, Hay JM, Mindell DP (2003). Molecular systematics of primary reptilian lineages and the tuatara mitochondrial genome. *Molecular Phylogenetics and Evolution* **29:** 289-297

Schwartz TS, Tae H, Yang Y, Mockaitis K, Hemert JHV, Proulx SR, Choi J-H, Bronikowski AM (2010). A garter snake transcriptome: pyrosequencing, *de novo* assembly, and sex-specific differences. *BMC Genomics* **11:** 694

Subramanian S, Huynen L, Millar CD, Lambert DM (2010). Next generation sequencing and analysis of a conserved transcriptome of New Zealand's kiwi. *BMC Evolutionary Biology* **10:** 387

St. John JA, Braun EL, Isberg SR, Miles LG, Chong AY, Gongora J, Dalzell P, Moran C, Bed'Hom B, Abzhanov A, Burgess SC, Cooksey AM, Castoe TA, Crawford NG, Densmore LD, Drew JC, Edwards SV, Faircloth BC, Fujita MK, Greenwold MJ, Hoffmann FG, Howard JM, Iguchi T, Janes DE, Khan SY, Kohno S, de Koning APJ, Lance SL, McCarthy FM, McCormack JE, Merchant ME, Peterson DG, Pollock DD, Pourmand N, Raney BJ, Roessler KA, Sanford SR, Sawyer RH, Schmidt CJ, Triplett EW, Tuberville TD, Venegas-Anaya M, Jarvis ED, Guillette Jr. LJ, Glenn TC, Green RE, Ray DA (2012). Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology* **13:** 415

Tzika AC, Helaers R, Schramm G, Milinkovitch C (2012). Reptilian-transcriptome v1.0, a glimpse in the brain transcriptome of five divergent Sauropsida lineages and the phylogenetic position of turtles. *EvoDevo* **2:** 19

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

Wu X-C. 1994: Late Triassic-Early Jurassic sphenodontians from China and the phylogeny of the Sphenodontia. in Fraser NC & Sues H-D, In the Shadow of the Dinosaurs. Cambridge University Press, New York. 1994

## Figures

**Table 1.** Set of taxa used in the EnsemblCompara phylogenetic prediction of orthology to develop a base dataset. Shown are the number of cDNA *Gallus* regions where at least one ortholog was found. The cell in the bottom right corner shows the total number of *Gallus* regions found to be shared between all taxa.

| Species | Common Name | Major Clade | Subclade | Othologs to *Gallus* |
|---|---|---|---|---|
| *Latimeria chalumnae* | West Indian Coelacanth | Sarcopterygii | Coelacanthiformes | 12,622 |
| *Xenopus tropicalis* | Western Clawed Frog | Amphibia | Anura | 12,341 |
| *Ornithorhynchus anatinus* | Platypus | Mammalia | Prototheria | 11,916 |
| *Monodelphis domestica* | Gray Short-Tailed Opossum | Mammalia | Metatheria | 12,947 |
| *Homo sapiens* | Human | Mammalia | Eutheria | 13,161 |
| *Anolis carolinensis* | Carolina anole | Squamata | Iguania | 12,432 |
| *Pelodiscus sinensis* | Chinese Softshell Turtle | Testudines | Cryptodira (Trionychia) | 13,137 |
| *Taeniopygia guttata* | Zebra finch | Aves | Neognathidae (Passeriformes) | 13,604 |
| *Gallus gallus* | Chicken | Aves | Neognathidae (Galliformes) | 7,807 |

**Table 2.** Additional taxa and transcriptomal assemblies used in the development of the final ortholog set. Taxa are listed sequentially (line breaks indicate each step-wise group), and the assembly with the highest reciprocal ortholog count to the prior set of *Gallus* cDNA shared between all taxa was selected (highlighted). The addition of new taxa was terminated at 195 orthologs.

| Species | Common Name | Major Clade | Subclade | Orthologs to *Gallus* | Orthologs to Prior step | Tissue | Reference |
|---|---|---|---|---|---|---|---|
| *Sphenodon punctatus* | Tutatara | Rhynchocephalia | Sphenodontia | 2,517 | 1,740 | Embryo | Miller *et al*. 2012 |
| *Caiman crocodilus* | Spectacled Caiman | Crocodylomorpha | Alligatoridae | 1,270 | 536 | Jaw | Gayral *et al*. 2011, Chiari *et a*l. 2012 |
| *Thamnophis elegans* | Western Garter Snake | Squamata | Scleroglossa (Serpentes) | 5,264 | 363 | Multi-Tissue | Swartz *et al.* 2010 |
| *Python molorus* | Burmese Python | Squamata | Scleroglossa (Serpentes) | 1,417 | 311 | Multi-Tissue | Castoe *et al.* 2010 |
| *Phrynops hilarii* | Hilaire's Side-necked Turtle | Testudines | Pleurodira | 2,193 | 202 | Blood | Chiari *et al*. 2012 |
| *Podarcis* sp. | Wall Lizard | Squamata | Scleroglossa (Scincomorpha) | 978 | 100 | Liver | Gayral *et al*. 2011, Chiari *et al*. 2012 |
| *Caretta caretta* | Loggerhead Sea Turtle | Testudines | Cryptodira (Chelonioidea) | 2,570 | 163 | Blood | Chiari *et al*. 2012 |
| *Chelonoidis nigra* | Galápagos Giant Tortoise | Testudines | Cyrptodira (Testudinoidea) | 3,357 | 195 | Blood | Chiari *et al*. 2012 |