Int. J. Dev. Biol. 58: 881-888 (2014)
doi: 10.1387/ijdb.150060at

THE INTERNATIONAL JOURNAL OF
DEVELOPMENTAL
BIOLOGY
www.intjdevbiol.com

The genome sequence of the corn snake (*Pantherophis guttatus*), a valuable resource for EvoDevo studies in squamates

ASIER ULLATE-AGOTE[1,2,3], MICHEL C. MILINKOVITCH[1,2,3] and ATHANASIA C. TZIKA[1,2,3,*]

[1]Laboratory of Artificial & Natural Evolution (LANE), Dept. of Genetics & Evolution, University of Geneva, [2]SIB Swiss Institute of Bioinformatics and [3]Institute of Genetics and Genomics of Geneva (iGE3), University of Geneva, Geneva, Switzerland

ABSTRACT Squamates (snakes and lizards) exhibit a striking variety of phenotypes, with little known on their generative mechanisms. Studies aiming to understand the genetic basis of this wide diversity in morphology, physiology and ecology will greatly benefit from whole genome sequencing initiatives, as they provide the foundation for comparative analyses and improve our understanding of the evolution, development and diversification of traits. Here, we present the first draft genome of the corn snake *Pantherophis guttatus*, an oviparous snake that we promote as a particularly appropriate model species for evolutionary developmental studies in squamates. We sequenced 100-base paired-end reads from multiple individuals of a single family (parents and offspring) that produced a genome assembly of 1.53 gigabases (Gb), roughly covering 75% of the expected total genome size, and 297,768 scaffolds >1 Kb. We were able to fully retrieve 86, and partially another 106, of the 248 CEGMA core genes, indicating that a high genome completeness was achieved, even though the assembly is fragmented. Using MAKER2, we annotated 10,917 genes with high confidence (Annotation Edit Distance (AED)<1) and an additional 5,263 predicted genes matched with the species' transcriptome. Numerous colour and colour pattern morphs exist in *P. guttatus*, making it an ideal model to study the genetic determinism, development, and evolution of adaptive colour traits in reptiles. Using our draft genome and a Single-Nucleotide Polymorphism (SNP) calling approach, we confirmed the interval with the causative mutation for the amelanistic phenotype, a result supported by a parallel exome-based study.

KEY WORDS: *Pantherophis guttatus, corn snake, reptile, genome, amelanistic, SNP calling*

Introduction

An increasing number of non-classical vertebrate genomes have been sequenced during the last decade, facilitating evolutionary and comparative developmental studies (Castoe *et al.*, 2013, Vonk *et al.*, 2013, Wang *et al.*, 2013). However, the order Squamata is still largely underrepresented in these initiatives, despite the wide variety of phenotypes encountered within its 10,000 species (*i.e.*, about twice as many as in mammals). In the last few years, the publications of the *Anolis carolinensis* genome (Alfoldi *et al.*, 2011), the first draft (Castoe *et al.*, 2011) and the improved version (Castoe *et al.*, 2013) of the Burmese python (*Python molurus*) genome together with draft genomes of the king cobra *Ophiophagus hannah* (Vonk *et al.*, 2013), of the rattlesnake *Crotalus mitchellii* (Gilbert *et al.*, 2014) and of the common viper *Vipera berus* demonstrate an increased interest in Squamata in general, and snakes in particular,

Abbreviations used in this paper: AED, annotation edit distance; BLAST, basic local alignment search tool; CEGMA, core eukaryotic genes mapping approach; CR1, chicken repeat 1; DCT, dopachrome tautomerase; EST, expressed sequence tag; LINE, long interspersed nuclear elements; LTR, long terminal repeat; MIR, mammalian-wide interspersed repeats; Mya, millions years ago; OCA2, oculocutaneous albinism II; RTE, retro-transposable element; SINE, short interspersed nuclear element; SNP, single-nucleotide polymorphism; *amel*, amelanistic.

***Address correspondence to:** Athanasia C.Tzika. Laboratory of Artificial & Natural Evolution (LANE), Dept. of Genetics & Evolution, University of Geneva, Sciences III, 30, Quai Ernest-Ansermet, 1211 Genève 4, Switzerland.Tel: +41(0)22 379 67 85. Fax: +41(0)22 379 67 95. E-mail: Athanasia.Tzika@unige.ch -Web: www.lanevol.org

Supplementary Material (tables, Figure, XLS files, FASTA files) for this paper is available at: http://dx.doi.org/10.1387/ijdb.150060at - See note at end of paper for a more detailed description of the files.

Accepted: 9 February 2015.

ISSN: Online 1696-3547, Print 0214-6282

as they are good models for investigating the mechanisms associated to their modified limb and body plan development (Di-Poi *et al.*, 2010, Woltering, 2012), venom evolution (Vonk *et al.*, 2013), physicochemical sensory perception (Brykczynska *et al.*, 2013), extreme fluctuations in metabolic rates (Castoe *et al.*, 2013), as well as development and patterning of scales and colours.

Over the last few years, we have been promoting the corn snake *Pantherophis guttatus*, of the Colubridae family and originating from North America, as a particularly appropriate snake model species for evolutionary developmental studies (Di-Poi *et al.*, 2010, Milinkovitch and Tzika, 2007, Tzika and Milinkovitch, 2008). These animals are easy to breed and maintain, they have a moderate size (maximum length around 1.5 meters) and a long life span (>20 years), and are harmless to humans because they are non-venomous and reluctant to bite. Of paramount importance is that corn snakes are oviparous (Fig. 1A), laying one to two clutches per year, making the species amenable to developmental studies (Fig. 1B). In addition, numerous colour and pattern morphs exist (Fig. 1C), making them ideal models to investigate the genetic determinism and underlying molecular pathways controlling adaptive colour variation in reptiles. In short, corn snakes combine a series of advantages matched by none of the other species of snakes for which a genome is available or in progress: *Thamnophis* species and rattlesnakes are ovoviviparous (hence, females must be sacrificed to access the embryos), whereas cobras, rattlesnakes and large pythons (such as Burmese pythons) are difficult or dangerous to maintain in captivity.

Here, we present the first draft genome of *P. guttatus*, produced by whole genome sequencing of multiple individuals from a single family (parents and offspring). The assembly statistics and the comparison to reference datasets show that this *P. guttatus* genome is of similar quality to other draft snake genomes (Castoe *et al.*, 2011, Gilbert *et al.*, 2014). In addition, we were able to use the assembled genomic sequence together with single-nucleotide polymorphism (SNPs) data to confirm the sequence interval containing the causal mutation of the amelanistic mutation (*amel*; lack of dark pigment from the skin and eyes, Fig. 2). Aggregation of SNPs identified a genomic area in *P. guttatus* that corresponds to a 31 Mb region of *A. carolinensis* chromosome 3 and a 17Mb region of *Gallus gallus* chromosome 1. The present analysis confirms our previous work using an exome-assembly approach (Saenko *et al.*, submitted) that identified a similar genomic interval and led to the identification of the causal *amel* mutation in corn snakes. Compared to the exome-based analysis, our genome-based approach provides longer genomic reference sequences that extend further into intronic and intergenic regions. Thus, more SNPs cosegregating with the amelanistic genotype are discovered, facilitating the retrieval of the interval of interest and increasing the number of SNPs that can be considered for posterior genotyping. Moreover, a shorter interval of *A. carolinensis* chromosome 3 was found with the genome-based (31 Mb) than with the exome-based (39 Mb) approach.

Results

The corn snake draft genome

On five Illumina lanes (100-base paired-end reads), we sequenced genomic DNA from 78 individuals of a single corn snake family: (i) the mother, heterozygous for the recessive *amelanistic* mutation (*amel*/+), and the father, homozygous for the same mutation (*amel*/*amel*) were indexed on a single lane, (ii) a pool of 20 non-indexed heterozygous *amel*/+ offspring, (iii) a pool of 20 non-indexed homozygous amelanistic offspring, and (iv and v) two pools, 18 non-indexed individuals each, of offspring grouped on the basis of another trait (unlinked to the *amel* locus). We used the filtered and trimmed Illumina reads (see Materials & Methods) of the mother to assemble 26.1 giga-bases (Gb) into contigs. Our indirect estimate of the corn snake genome size (on the basis the k-mer distribution; see Materials and Methods) is 1.74-1.79 Gb (Supp. Fig. S1), a number in agreement with genome sizes of other *Pantherophis* and *Elaphe* species (1.8-2.2 Gb, Animal Genome Size Database www.genomesize.com). Hence, the draft genome presented here has an average sequencing depth of 14.6-15.0x. During this contig building step (*i.e*, assembling overlapping contiguous reads), we obtained 4,169,571 contigs ≥ 100b with an average length of 358b and an N50 of 563b (Table 1). Afterwards, we ordered and oriented contigs into scaffolds, taking into account the paired-end reads information (*i.e.*, the distance and orientation of the two reads of each sequenced pair). For the scaffolding, we used all six genomic libraries (184.6 Gb in total, 86-106-fold average genome sequencing depth). The genome assembly (deposited in GenBank, accession number JTLQ00000000, including only contigs/scaffolds >200b) is 1.53 Gb long, *i.e.*, between 70 and 85% of the expected genome size. It comprises 1,781,284 scaffolds and singletons ≥ 100 b. The genome GC content is 39.81% and only 3.2% of the bases are unknown (N). The great difference between the scaffold average length (860b) and the N50 value (3,674)

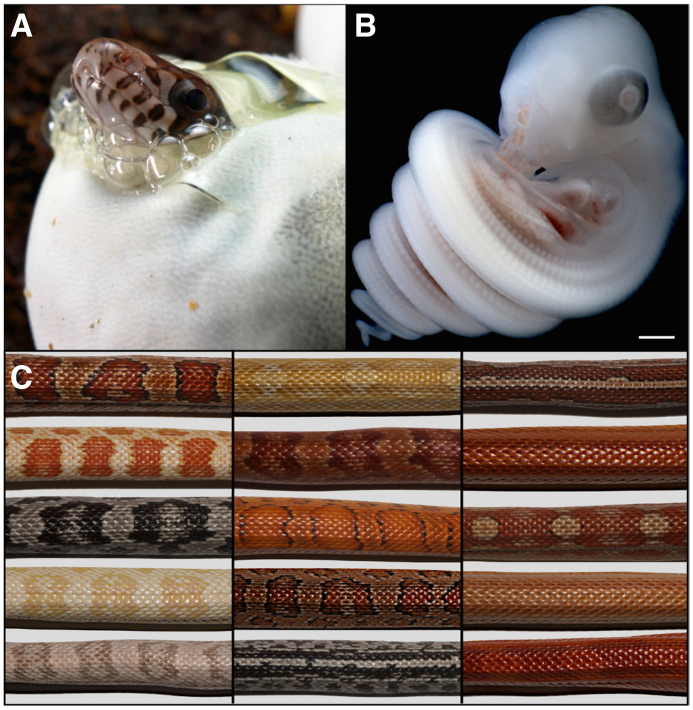


Fig. 1. The corn snake *Pantherophis guttatus* as a model species for evolutionary developmental studies. (A) *Hatching after an egg incubation of 60 days.* **(B)** *Embryonic day 10 embryo.* **(C)** *Numerous colour and colour pattern morphs exist in* Pantherophis guttatus*, making it an ideal model to study the genetic determinism of adaptive colour traits in reptiles.*

TABLE 1

CORN SNAKE GENOME ASSEMBLY STATISTICS

	Sequences ≥ 100b	**Assembly size (*b*)**	**Average length (*b*)**	**N50 (*b*)**	**Longest sequence (*b*)**	**Sequences ≥ 10Kb**
Contig assembly	4,169,571	1,493,759,741	358	563	16,051	42
Scaffolding	1,781,284	1,532,089,567 (3,2%)	860	3,674	102,731	21,405 (1.2%)

In parenthesis, the percentage of unknown (N) bases is indicated. ***b*** = number of bases

shows that many of the scaffolds are short: 297,768 scaffolds are >1Kb and half of the genome size is covered by 94,091 scaffolds (L50). These statistics are similar to those of the *P. molurus* and *C. mitchellii* draft genomes.

The Core Eukaryotic Genes Mapping Approach (CEGMA (Parra *et al.*, 2007)), defines a set of 248 highly-conserved core proteins present in a wide range of eukaryotes (from yeast to human). Our assembly includes the complete gene sequence of 86 out of the 248 CEGMA eukaryotic core genes (34.68%), a lower value compared to higher quality genomes (Castoe *et al.*, 2013, Vonk *et al.*, 2013), where more than 200 CEGMA full genes were sequenced. On the other hand, we were able to retrieve partial sequence of 192 core genes (77.42%) from our draft genome, demonstrating that genome coverage is good but fragmented.

We built a library of *P. guttatus* repeats with *RepeatModeler* and combined it with all Vertebrate repeats from RepBase (Jurka *et al.*, 2005). The use of this combined library resulted in masking 39.14% of the corn snake genomic sequence with RepeatMasker-open 4.0.5 (Smit *et al.*, 1996). We observe a highly diverse landscape of repetitive elements (Fig. 3) as in other non-avian Sauropsida (Alfoldi *et al.*, 2011, Shedlock *et al.*, 2007) and unlike in humans, where most of the non-longterminal-repeats (non-LTR) retrotransposons consist of L1 repeats. The most common long interspersed nuclear elements (LINEs) in the corn snake genome are homologous to the Chicken Repeat 1 (CR1, 2.81%), with L1, L2, RTE (retro-transposable element) and R4 also being abundant, as in *A. carolinensis* (Novick *et al.*, 2009). The identified short interspersed nuclear elements (SINEs) mainly belong to the MIR-like family (1.91%) and there are also a few LTR elements that are endogenous retroviruses. The distribution and variety of repetitive elements in the corn snake genome are similar to those of other snakes (Castoe *et al.*, 2013), except that we observe a higher proportion of DNA transposons ("cut and paste DNA", 6.45%). Finally, 16.58% of the corn snake genome corresponds to repetitive elements from unclassified groups (Fig. 3).

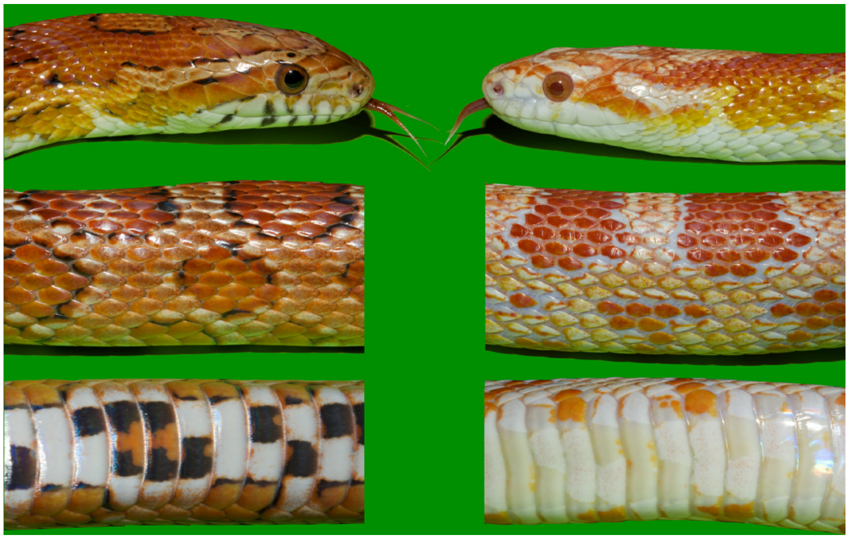


Fig. 2. Comparison of a wild-type (left) and an amelanistic (right) individual. *The lack of melanin in the amelanistic individual is obvious in the eyes, around the dorso-lateral blotches, as well as by the absence of the ventral checker pattern.*

Genome annotation

Two iterations of the MAKER2 pipeline (Holt and Yandell, 2011) yielded 24,258 predicted genes. MAKER2 provides a set of gene predictors, performs similarity searches against cDNA/ESTs or proteins databases, and combines the outputs into a set of high-quality gene predictions. Our corn snake transcriptomic data (Tzika *et al.*, 2015) together with a protein database (including the SwissProt repository and NCBI sequences of *A. carolinensis* and snakes) were used in the MAKER2 pipeline to complement the gene models computed by three *ab-initio*/ evidence-based predictors: AUGUSTUS (Stanke and Waack, 2003), SNAP (Korf, 2004) and GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008). As a quality measurement of our annotation, we used the 'annotation edit distance' (AED, (Eilbeck *et al.*, 2009)), which is based on the agreement at the nucleotide level (*i.e.*, focusing on overlapping nucleotides rather than substitutions) between the final gene annotation and the aligned evidence data (cDNA/proteins) supporting that annotation. Values closer to zero correspond to higher quality annotation (*i.e*, it is more congruent with the cDNA or protein data) and predictions with an AED=1 are interpreted as not supported by experimental data at all. An AED <1 was found for 10,917 of the 24,258 predicted genes (45%) (Supp. Files S2-S4), with the majority of them showing an AED<0.5 (8,597, 35.4%). These rather low numbers are due to the high fragmentation of the draft corn snake genome: many genes span multiple scaffolds, making their annotation impossible. Eighty percent (8,817) of the corn snake proteins with AED < 1 had a *blastp* hit (e-value <10^{-5}) against *A. carolinensis* (Ensembl v77 proteins), *P. molurus* or *O. hannah* sequences and 94% (10,237) had a *megablast* hit against our *P. guttatus* transcriptome (Tzika *et al.*, 2015). When considering proteins with the maximum AED value of 1, only 19.2% (2,565) had a hit against *A. carolinensis, P. molurus* or *O. hannah* and 39.4% (5,263) against the *P. guttatus* transcriptome. The sequences without a hit against other species but with a hit against our corn-snake transcriptome are likely to be real cDNA/proteins, coded by less-conserved or taxonomically-restricted genes.

To further assess the completeness of the *P. guttatus* draft genome, we checked for the number of *Hox* genes annotated and/or present in the assembly (Suppl. Table S1). Eleven out of 40 *Hox* genes were identified by MAKER2, with four being fully sequenced. We located homologous sequences of 27 additional *Hox* genes using BLAST searches on scaffolds <10Kbp (not considered during the annotation process). The only missing *Hox* gene in our draft genome is *Hoxd12*, a gene involved in limb development in other vertebrates. This result is consistent with the absence of *Hoxd12* in the fully-sequenced clusters of posterior *Hox* genes in *P. guttatus* (Di-Poi *et al.*, 2010) and in the *P. molurus*, *O. hannah* and other squamata genomes (Vonk *et al.*, 2013). Given the fragmented nature of the assembly, 29 of the 39 corn snake *Hox* genes were located on distinct scaffolds, besides five pairs of genes found on the same scaffold, preventing the confirmation of their cluster organisation.

Identification of the genomic region with the amelanistic *mutation*

We performed a Single-Nucleotide Polymorphism (SNP) calling approach to find the locus of the amelanistic (*amel*) mutation. The Mendelian inheritance of this phenotype indicates that it is associated to a single recessive mutation in our captive-bred population. The homozygous amelanistic corn snake individuals can be identified by their lack of dark pigmentation (melanin) in their skin and eyes; hence, they miss both the checker pattern on the ventral skin and the black contour surrounding the red blotches on their dorsal side (Fig. 2).

We used the parental genomic DNA libraries (male *amel*/*amel* and female *amel*/+) and the two libraries of their offspring pooled on the basis of their genotype (*amel*/*amel* versus *amel*/+). Each of the four libraries was aligned separately to our newly-assembled *P. guttatus* draft genome and the SNPs were extracted using FreeBayes (Garrison and Marth, 2012). We only kept biallelic SNPs, with a defined minimum sequencing depth, that were co-segregating according to the expected genotype for the amelanistic locus in each sample (see Supp. Table S2 and Materials and Methods). This filtering resulted in 19,104 SNPs distributed on 4,740 scaffolds (on average, 4 SNPs per scaffold and 1.16 SNPs/Kb). Of these scaffolds, 751 had a *megablast* hit (bitscore ≥ 100) against the *A. carolinensis* masked genome and included 5,273 SNPs (7 SNPs per sequence). Most of the hits (59%) were against *A. carolinensis* Chromosome 3, and 356 of them densely covered the 82-113Mb interval (Fig. 4A and 4B): 3,735 SNPs were present in this interval and only 131 elsewhere on the chromosome. Note that the SNPs in the 82-113Mb interval are split in two islands with only 50 hits found within the 87-97Mb sub-interval (corresponding, in *A. carolinensis,* to a gene desert, *i.e.*, a region of the genome with very few protein-coding genes). As the *P. guttatus* and *A. carolinensis* lineages separated ~166.4 million years ago (Mya), it is not surprising that similarity searches (BLAST matches) are inefficient in this, probably fast evolving, intergenic region.

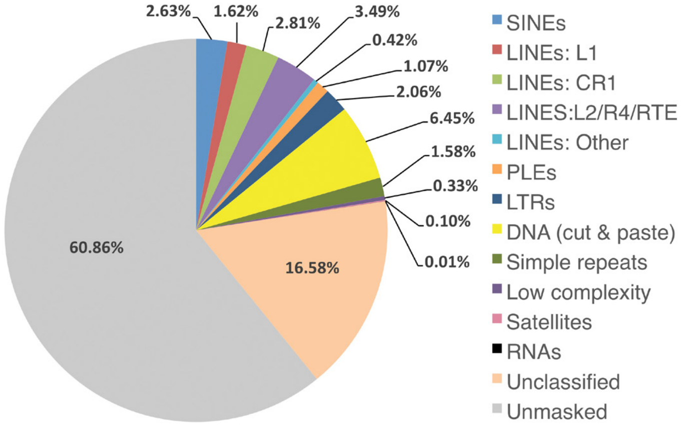


Fig. 3. Proportions of identified repetitive elements in the *Pantherophis guttatus* genome.

Another 61 *P. guttatus* scaffolds with informative SNPs aligned to a 2.2 Mb *A. carolinensis* scaffold (GL343243.1), which includes genes orthologous to those located in the 132.9-134.4 Mb interval of the *G. gallus* chromosome 1 (Fig. 5). Other parts of the *G. gallus* chromosome 1 adjacent to this interval are syntenic with *A. carolinensis* chromosome 3. Hence, our analysis indicates that the *A. carolinensis* scaffold GL343243.1 should probably be inserted in *A. carolinensis* chromosome 3 at about the position 109Mb (more specifically between the TGFBRAP1 and MGAT4A genes), *i.e.*, in the area discussed above where most SNPs informative for the amelanistic locus are localised. Our analysis also indicates that genes present in the 131.8-132.9 Mb interval in *G. gallus* chromosome 1 are homologous to *A. carolinensis* genes located on three short scaffolds: GL343940.1 (0.1 Mb), GL343456.1 (0.66 Mb) and GL343542.1 (0.45 Mb). In total, 435 out of 751 (58%) of the corn snake sequences with SNPs informative for the amelanistic locus had a hit against either one of these four *A. carolinensis* scaffolds or the 82-113 Mb interval of *A. carolinensis* chromosome 3. All these hits accounted for 4,674 SNPs (89%) of the 5,273 informative SNPs (10.7 SNPs per scaffold and 1.45 SNPs/Kb), strongly suggesting that the amelanistic locus is in this interval, as SNPs located close to the amelanistic locus tend to present the same co-segregating pattern as the causal mutation.

We also ran a *megablast* search of the 4,740 filtered scaffolds against the *G. gallus* masked genome (International Chicken Genome Sequencing, 2004). As snakes are more distantly related to birds (~274 Mya) than to lizards (~166.4 Mya), only 291 *P. guttatus* scaffolds out of 4,740 (6%) had a hit (bitscore ≥ 100). These matching scaffolds include 2,530 SNPs (13.2%, 8.7 SNPs per scaffold, and 1.09 SNPs/Kb) and 226 (77.5%) of them were against *G. gallus* chromosome 1, with 178 in the 129.5-146 Mb interval (2,312 SNPs with a density of 1.43 SNPs/Kb; Fig. 4C). These SNPs are distributed more uniformly than in *A. carolinensis* chromosome 3 (Fig. 4B), probably because this interval in the *G. gallus* chromosome 1 does not include a gene desert, contrary to the corresponding region of *A. carolinensis* chromosome 3. The synteny between *G. gallus* chromosome 1 and the *A. carolinensis* chromosome 3 and the four short scaffolds (Fig. 5) provides additional support for the localisation of the amelanistic mutation.

A greater number of *P. guttatus* scaffolds had a *megablast* hit against the *O. hannah* and *P. molurus* genomes (than against the *A. carolinensis* and *G. gallus* genomes), as expected given their closer evolutionary relationship (Supp. Table S2). Although this larger number of *P. guttatus* scaffolds also corresponds to a greater number of informative SNPs, in total 85 *O. hannah* and 120 *P. molurus* scaffolds span the interval of interest on the *A. carolinensis* chromosome 3, making the comparison with these snake species less informative in terms of synteny.

Candidate genes for the amelanistic *mutation*

Among the 205 genes in the 82-113 Mb interval of the *A. carolinensis* chromosome 3 and the four short scaffolds (Fig. 5), two genes are particularly good candidates for bearing the mutation responsible for the amelanistic phenotype in corn snakes: the oculocutaneous albinism II or P protein (OCA2) at one end of the interval and the dopachrome tautomerase (DCT) or Tyrp2 genes at the other end (Fig. 4A). Both proteins participate in the

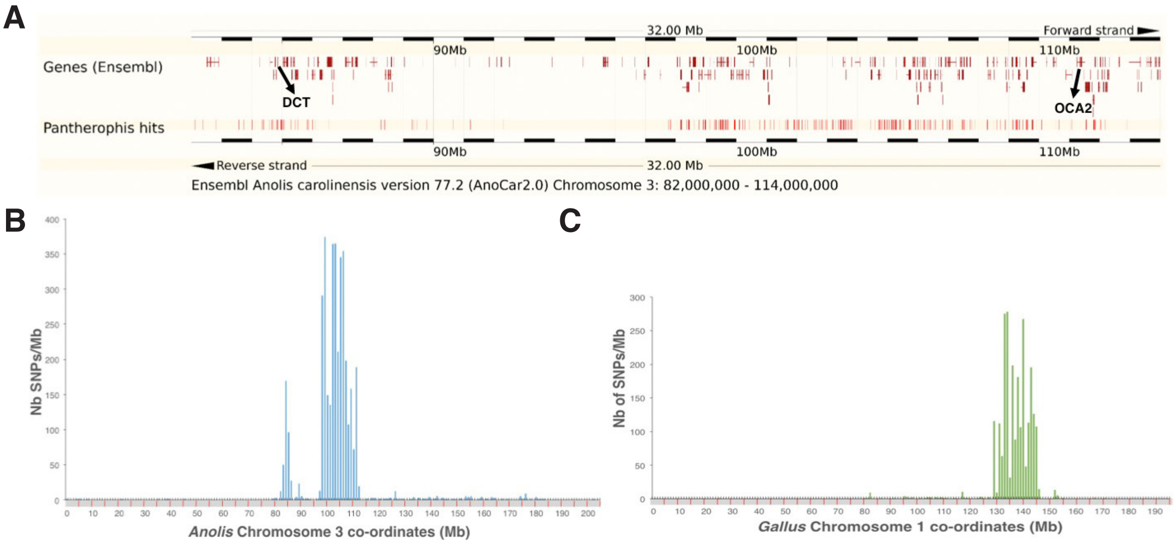


Fig. 4. Distribution of single-nucleotide polymorphisms (SNPs) and scaffolds on the interval harboring the amelanistic mutation. (A) *Schematic representation of a 32 Mb* Anolis carolinensis *chromosome 3 interval matching with* Pantherophis guttatus *sequences exhibiting SNPs co-segregating with the amelanistic mutation. Dark red bars at the top:* Anolis carolinensis *Ensembl genes. Light red bars at the bottom: location of the corn snake scaffolds BLAST hits. Black arrows: location of the two main candidate genes for the* amel *locus.* **(B,C)** *Distribution of* Pantherophis guttatus *SNPs that co-segregate with the amelanistic mutation and match (number of SNPs per Mb) with the* Anolis carolinensis *chromosome 3* **(B)** *and the* Gallus gallus *chromosome 1* **(C)**.

biosynthesis of eumelanin from tyrosine within the melanosomes of melanocytes. OCA2 is involved in the transport of the pigment out of the melanosomes, whilst DCT is an isomerase found at the organelle's membrane. Mutations in both genes are known to cause pigmentation disorders in humans and mice (Budd and Jackson, 1995, Rimoldi *et al.*, 2014), resulting in lighter coloration of the hair, the skin and the eyes. These two genes were not annotated by the MAKER2 pipeline in our draft *P. guttatus* genome because their exons are distributed in several scaffolds.

Using a SNP calling analysis based on a *P. guttatus* exome-assembly approach (Saenko *et al.*, submitted) rather than a genome-assembly (this study), we previously identified an overlapping, but longer, interval on *A. carolinensis* chromosome 3 as including the amelanistic locus. Even though the same sequencing data was used, the full-genome approach presented here allowed us to identify a higher number of co-segregating SNPs than with the exome-assembly approach. Indeed, the draft genome of the corn snake made it possible to extend the search for co-segregating SNPs further in the intronic and intergenic regions, and provided greater support for the interval containing the causal mutation. It is likely that the corn snake draft genome presented here will greatly assist the identification of mutations responsible for coloration phenotypes (Fig. 1C) or other traits in *P. guttatus* or closely-related species.

Discussion

Here, we have extended the scarce Squamata sequencing resources by assembling the first draft genome of the corn snake *P. guttatus*. We have long supported the use of this species as a model for evolutionary developmental biology in snakes (Brykczynska *et al.*, 2013, Di-Poi *et al.*, 2010, Milinkovitch and Tzika, 2007, Tzika *et al.*, 2011, Tzika and Milinkovitch, 2008, Tzika *et al.*, 2015). We are particularly interested in understanding the genetic basis of colour and colour pattern variation within and among snake species. The amelanistic mutation is the oldest mutation segregated in captive breeding populations. Since then, numerous other colour phenotypes (some of them due to single locus mutations) in corn snakes have been selected by breeders all over the world. Hence,

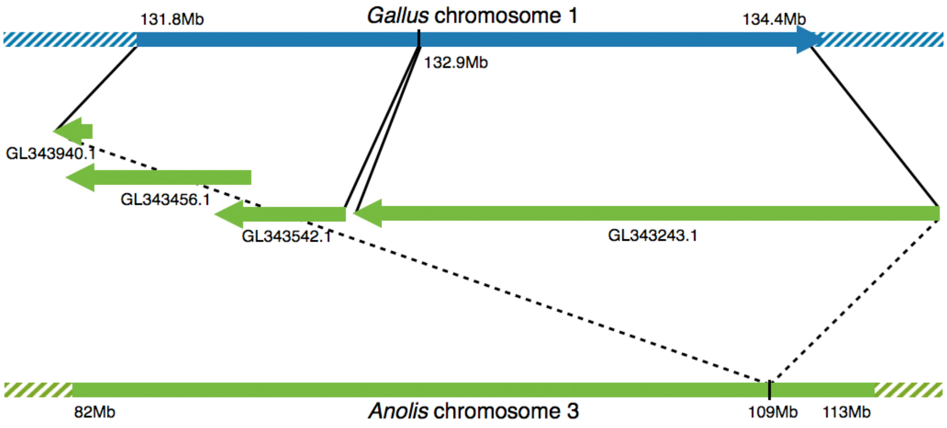


Fig. 5. Synteny diagram between the *Anolis carolinensis* (green) and the *Gallus gallus* (blue) genomic regions that match with most of the *Pantherophis guttatus* scaffolds harboring the single-nucleotide polymorphisms (SNPs) that co-segregate with the amelanistic mutation. *Boxes with diagonal lines represent synteny between the two species independently of their SNP content.*

P. guttatus is a very good model to study the pigment biosynthesis pathways and developmental mechanisms controlling skin colour patterns in squamates.

The draft corn snake genome presented here shows similar quality and sequencing depth as other draft snake genomes, but is still rather fragmented. This high fragmentation explains that only 86 out of the 248 CEGMA highly-conserved core eukaryotic genes are completely sequenced, though at least 192 were partially retrieved. Nonetheless, we did obtain a core of 10,917 genes with a good prediction support and another 5,263 with transcriptomic support. So as to reduce the fragmentation of the corn snake genome sequence, it would be necessary to sequence paired-end Illumina libraries with variable fragment size and mate-paired Illumina libraries with inserts ranging from 3 kb to 6 kb to 40 kb (Nagarajan and Pop, 2013). Furthermore, generating sequence data with additional technologies, such as Pacific Biosciences, is likely to improve the assembly as recently put forward by the Assemblathon 2 initiative (Bradnam *et al.*, 2013). Such an approach would produce longer scaffolds with less gaps and possibly resolve repetitive regions. Additional reads from a single corn snake, rather than a pool of several individuals (initially used for gene mapping; Saenko *et al.,* submitted), would also increase the sequencing depth and accuracy of the final assembly.

We demonstrated the usefulness of the corn snake genomic sequence by retrieving the region in which the locus responsible for the amelanistic mutation (*amel*) is present. To that end, we performed a SNP calling approach using our genome assembly as a reference and Illumina libraries of individuals from a single *P. guttatus* family with known genotypes for the *amel* locus. Two excellent candidate genes were located in the selected interval, corresponding to 31 Mb of the *A. carolinensis* chromosome 3: DCT and OCA2. These two genes are involved in the melanin biosynthetic pathway and their mutations are known to produce pigmentation disorders in human and mice. Using our approach, we retrieved a greater number of informative SNPs on longer scaffolds, both in coding and intergenic/intronic regions, compared to the SNPs obtained in a parallel study, based on the same raw Illumina sequence data but using an exome-assembly approach (Saenko *et al.*, submitted). In that study, we proceeded with the genotyping by PCR and Sanger sequencing of additional individuals of the same *P. guttatus* family and identified the mutation responsible for the corresponding phenotype.

Materials and Methods

Whole-genome sequencing

Genomic DNA was extracted from tissue samples of *P. guttatus* individuals using the QIAGEN DNeasy Blood and Tissue kit. All animal experiments were performed in accordance with the Swiss animal welfare regulation (permit number 1008/3421/0). We obtained DNA from two adults: (i) an amelanistic male (*i.e.*, homozygous for the recessive *amel* mutation; *amel/amel* genotype), also heterozygous for the *motley* mutation (*motley*/+ genotype) and (ii) a female heterozygous for the *amel* mutation and homozygous for the *motley* mutation (*amel*/+ and *motley/motley* genotype). *Motley* is a skin colour pattern morph that was not analysed in this study. In addition, we pooled DNA in equimolar concentrations in four offspring groups produced from these two individuals: (i) 20 heterozygous *amel*/+ individuals, (ii) 20 amelanistic (*amel/amel*) animals, (iii) 18 snakes heterozygous for the *motley* mutation (*motley*/+) and (iv) 18 motley (*motley/motley*) individuals. We constructed DNA libraries (300-400bp fragment size) for each of the parents and for each of the offspring pools using the TruSeq PE Cluster Kit v3-cBot-HS. We then sequenced the samples on an Illumina HiSeq2000 sequencer, using one lane for each of the four libraries of pooled offspring and a fifth lane for the two indexed parental libraries.

We obtained 245 to 395 millions of Illumina 100-base paired-end reads per library and performed quality trimming using sickle v1.29 (Joshi and Fass, 2011); we removed flanking bases of quality <20 and we discarded reads that had one or more unknown base(s) (N) in their sequence. We used SOAPec v2.01 (http:// soap.genomics.org.cn/soapdenovo.html) to remove the reads that were <50b and to further correct and filter the reads through the use of a k-mer distribution spectrum (for k = 23) constructed using SOAPec. Reads with low-frequency (≤3) k-mer regions are considered to contain sequencing errors and are selected by the program, which uses an algorithm to correct the bases producing the erroneous k-mers (if their quality is <30). The reads that remained uncorrected were then trimmed. Following the sequence clean-up, we obtained 21-37 Gb per library with paired-end reads (and single-end reads, when one of the reads from the pair is removed during the filtering process) showing an average length of 98b. We estimate the haploid genome size of *P. guttatus* to be 1.8-2.2 Gb, based on the c-values of closely related *Elaphe* and *Pantherophis* species (Animal Genome Size Database -http:// www.genomesize.com) and the shared number of chromosomes between *P. guttatus* and *P. obsoletus* that practically makes their karyotype indistinguishable (Baker *et al.*, 1971). Thus, we expect to have generated an average genome sequencing depth of 9.5x-20.5x per library.

We also indirectly estimated the corn-snake genome size based on the assumption that the genome sequencing depth is a function of the sequencing depth of the most frequent k-mers (Li *et al.*, 2010). To this end, we obtained the frequency distribution of k-mers of different sizes (17, 23 and 31-mers) in the raw sequencing reads of all the libraries using the program Jellyfish (Marcais and Kingsford, 2011). The peak k-mer depth was 92, 84 and 74, respectively (Supp. Fig. S1). Using the formula $M = N(L–K+1)/L$, where M is the k-mer peak depth, N the genome sequencing depth, L the average read length (98.16b) and K the selected k-mer length, we obtained a sequencing depth N of 107-110x. Thus, we estimated the genome size at 1.74-1.79Gb by dividing the total cumulated read length (191,011,161,168b) by the genome sequencing depth N. On the basis of this estimate, we expect to have generated an average genome sequencing depth of 14.6-15.0x.

Genome assembly

We assembled the *P. guttatus* genome using SOAPdenovo2-v2.04.240 (Luo *et al.*, 2012). Only the 266 million filtered reads of the (*amel*/+) mother Illumina library were considered for the contig building step (*i.e*, assembling overlapping contiguous reads). We performed multiple assemblies with a range of k-mers between 43 and 55b and selected the optimal k-mer length (45) on the basis of three assembly parameters:

(i) N50, (ii) N90 and (iii) the longest contig sequence. As k=43 and k=45 had similar statistics, we chose the longest k-mer. The average fragment size of each library was estimated using the software *bwa v0.7.5a* with default parameters by aligning a subset of its paired-end reads against the preliminary assembled contigs (Li and Durbin, 2009). For the scaffolding steps, the contigs were ordered and oriented using the paired-end reads information of all six Illumina libraries (parental and offspring). The scaffolding step was also performed in SOAPdenovo2 and all parameters were set to default, except for the F parameter that was activated to fill up gaps. Note that we performed assemblies with alternative approaches (*e.g.*, by also using multiple libraries during contig building, in addition to scaffolding) and softwares, such as *Platanus* (Kajitani *et al.*, 2014), but the quality of the final assembly did not improve.

The quality of the draft genome was assessed using CEGMA v2.5 (Parra *et al.*, 2007), optimised for vertebrates. We also used RepeatModeler-1.0.8 to identify and model the *P. guttatus* repetitive elements. To mask the final assembly with RepeatMasker-open 4.0.5 (Smit *et al.*, 1996), we considered the newly identified repeats together with all the Vertebrate repeats from RepBase Update 19.07 (Jurka *et al.*, 2005).

Genome annotation

The Core Eukaryotic Genes Mapping Approach (CEGMA (Parra *et al.*, 2007)) was used for quality control, but it additionally retrieves orthologs of the highly-conserved eukaryotic core genes in a genome, thus determining their exon-intron structure. We performed gene prediction using the automated pipeline MAKER2 (Holt and Yandell, 2011). We considered only those scaffolds that were longer than 10Kb or that were >1Kb and had a predicted annotation using the CEGMA set of highly-conserved genes in eukaryotes. For the gene annotation, we built a protein database including the SwissProt database, as well as all *A. carolinensis* and snake proteins (including the ones from *P. molurus* and *O. hannah*) from NCBI. We also used a *P. guttatus* transcriptome assembled from Illumina and 454 cDNA libraries obtained from a mix of adult organs (testis, kidneys, brain and vomeronasal organ) and three developmental stages (Brykczynska *et al.*, 2013, Tzika *et al.*, 2015).

We run the first iteration of MAKER2 combining the evidence from known mRNAs and proteins and the *ab-initio* predictions of SNAP (Korf, 2004) and GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008). For this step, the SNAP hidden Markov models (HMM) were optimised using the CEGMA output and the GeneMark-ES model parameters were obtained from self-training in genome scaffolds greater than 10Kb. We then trained the evidence-based predictor AUGUSTUS (Stanke and Waack, 2003) with the output of the first step and we run it in a second MAKER2 iteration, together with the other two gene predictors. We also modeled new SNAP HMM from the output of the previous iteration. The repetitions library including the *P. guttatus* repeats identified by RepeatModeler and the Vertebrate repeats from RepBase Update 19.07 was used to mask the genome during the annotation process.

As an additional means to verify the genome completeness, we retrieved the *Hox* proteins of *A. carolinensis* from Ensembl version 77, of *P. molurus* from NCBI, and of the *O. hannah* scaffolds that include the *Hox* clusters. We then performed BLAST searches to find their homologous sequences in the *P. guttatus* draft genome.

SNP calling

To identify the genomic interval where the *amel* mutation is located, SNP calling was performed on the genomic libraries of individuals (parents and offspring) with known genotype for the *amel* locus (excluding the libraries where offspring were segregated on the basis of the *motley* locus). First, using *bwa* v0.7.5a and default parameters, we aligned the reads of each library against the corn snake genomic scaffolds >1Kb. Second, we converted the output to BAM files and extracted all variants using *FreeBayes* v0.9.9.2 (Garrison and Marth, 2012). We then used an inhouse Python script to extract from the output VCF file only those variants that (i) were biallelic SNPs (*i.e.*, they had exactly two alleles), (ii) had a FreeBayes-estimated quality greater than 100 and (iii) had a sequencing depth between 8 and 50 for each library. At the next step, we filtered out the SNPs that deviated from the segregation expected for SNPs linked to the causal mutation. More specifically, considering only the two parental libraries, we discarded SNPs that met at least one of the following conditions: *(i)* presented both alleles in at least one amelanistic individual (*i.e.*, homozygous for the *amel* mutation) or (*ii*) showed one of the two alleles in less than 25% of the reads in the heterozygous parental library (*amel*/+ genotype) because these cannot be informative for mapping. Then, we used the four family libraries to perform the same filtering approach, but also discarding SNPs for which any of the two alleles was sequenced less than twice in the heterozygous offspring library. We compared the two filtered SNPs datasets to identify scaffolds that co-segregate with the amelanistic genotype, as they could hint to the location of the amelanistic mutation.

The identified scaffolds were compared against the *A. carolinensis* (AnoCar2.0, (Alfoldi *et al.*, 2011)), *G. gallus* (Galgal4, (International Chicken Genome Sequencing, 2004)), *O. hannah* and *P. molurus* masked genomes using *megablast* (BLAST+ release 2.2.29). We considered only hits with an e-value < 10^{-5} and a bitscore ≥ 100, keeping only the best match for each sequence.

Supplementary Files

Supplementary File S1 is a PDF including Supplementary Tables S1-S2 and Supplementary Fig S1;
Supplementary File S2 is an XLS file with MAKER2 information on the annotated *P. guttatus* proteins (AED < 1);
Supplementary Files S3 and S4 are FASTA files with sequences of predicted proteins and transcripts (AED < 1);
Supplementary File S5 is an XLS file listing the SNPs selected for mapping the *amelanistic* mutation.

Acknowledgments

We are grateful to Adrien Debry and Rosa Afonso for maintenance of the P. guttatus *colony and Suzanne Saenko for setting up the crosses and extracting the DNA. Most computations were performed at the Vital-IT Centre for high-performance computing (www.vital-it.ch) of the SIB Swiss Institute of Bioinformatics. This work was supported by grants from the University of Geneva (Switzerland), the Swiss National Science Foundation (FNSNF, grant 31003A_125060), and the SystemsX.ch initiative (project* EpiPhysX*).*

References

ALFOLDI, J., DI PALMA, F., GRABHERR, M., WILLIAMS, C., KONG, L., MAUCELI, E., RUSSELL, P., LOWE, C.B., GLOR, R.E., JAFFE, J.D. *et al.,* (2011). The genome of the green anole lizard and a comparative analysis with birds and mammals. *Nature* 477: 587-591.

BAKER, R.J., BULL, J.J. and MENGDEN, G.A. (1971). Chromosomes of Elaphe-Subocularis (Reptilla-Serpentes), with Description of an in-Vivo Technique for Preparation of Snake Chromosomes. *Experientia* 27: 1228-1229.

BRADNAM, K.R., FASS, J.N., ALEXANDROV, A., BARANAY, P., BECHNER, M., BIROL, I., BOISVERT, S., CHAPMAN, J.A., CHAPUIS, G., CHIKHI, R. *et al.,* (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience* 2: 10.

BRYKCZYNSKA, U., TZIKA, A.C., RODRIGUEZ, I. and MILINKOVITCH, M.C. (2013). Contrasted evolution of the vomeronasal receptor repertoires in mammals and squamate reptiles. *Genome Biol Evol* 5: 389-401.

BUDD, P.S. and JACKSON, I.J. (1995). Structure of the mouse tyrosinase-related protein-2/ dopachrome tautomerase (Tyrp2/Dct) gene and sequence of two novel slaty alleles. *Genomics* 29: 35-43.

CASTOE, T.A., DE KONING, A.P., HALL, K.T., CARD, D.C., SCHIELD, D.R., FUJITA, M.K., RUGGIERO, R.P., DEGNER, J.F., DAZA, J.M., GU, W. *et al.,* (2013). The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc Natl Acad Sci USA* 110: 20645-20650.

CASTOE, T.A., DE KONING, J.A., HALL, K.T., YOKOYAMA, K.D., GU, W., SMITH, E.N., FESCHOTTE, C., UETZ, P., RAY, D.A., DOBRY, J. *et al.,* (2011). Sequencing the genome of the Burmese python (Python molurus bivittatus) as a model for studying extreme adaptations in snakes. *Genome Biol* 12: 406.

DI-POI, N., MONTOYA-BURGOS, J.I., MILLER, H., POURQUIE, O., MILINKOVITCH, M.C. and DUBOULE, D. (2010). Changes in Hox genes' structure and function during the evolution of the squamate body plan. *Nature* 464: 99-103.

EILBECK, K., MOORE, B., HOLT, C. and YANDELL, M. (2009). Quantitative measures for the management and comparison of annotated genomes. *Bmc Bioinformatics* 10: 67.

GARRISON, E. and MARTH, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907.*

GILBERT, C., MEIK, J.M., DASHEVSKY, D., CARD, D.C., CASTOE, T.A. and SCHAACK, S. (2014). Endogenous hepadnaviruses, bornaviruses and circoviruses in snakes. *Proceedings. Biological sciences / The Royal Society* 281: 20141122.

HOLT, C. and YANDELL, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *Bmc Bioinformatics* 12: 491.

INTERNATIONAL CHICKEN GENOME SEQUENCING, C. (2004). Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 432: 695-716.

JOSHI, N.A. and FASS, J.N. (2011). Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at https://github.com/najoshi/sickle

JURKA, J., KAPITONOV, V.V., PAVLICEK, A., KLONOWSKI, P., KOHANY, O. and WALICHIEWICZ, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogen. Genome Res.* 110: 462-467.

KAJITANI, R., TOSHIMOTO, K., NOGUCHI, H., TOYODA, A., OGURA, Y., OKUNO, M., YABANA, M., HARADA, M., NAGAYASU, E., MARUYAMA, H. *et al.,* (2014). Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res.* 24: 1384-1395.

KORF, I. (2004). Gene finding in novel genomes. *Bmc Bioinformatics* 5: 59.

LI, H. and DURBIN, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754-1760.

LI, R.FAN, W.TIAN, G.ZHU, H.HE, L.CAI, J.HUANG, Q.CAI, Q.LI, B.BAI, Y. *et al.,* (2010). The sequence and de novo assembly of the giant panda genome. *Nature* 463: 311-317.

LUO, R., LIU, B., XIE, Y., LI, Z., HUANG, W., YUAN, J., HE, G., CHEN, Y., PAN, Q., LIU, Y. *et al.,* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1: 18.

MARCAIS, G. and KINGSFORD, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* 27: 764-770.

MILINKOVITCH, M.C. and TZIKA, A. (2007). Escaping the mouse trap: The selection of new Evo-Devo model species. In *J. Exp. Zool. Part B: Molec. Dev. Evol.*, vol. 308, pp. 337-346.

NAGARAJAN, N. and POP, M. (2013). Sequence assembly demystified. *Nat Rev Genet* 14: 157-167.

NOVICK, P.A., BASTA, H., FLOUMANHAFT, M., MCCLURE, M.A. and BOISSINOT, S. (2009). The evolutionary dynamics of autonomous non-LTR retrotransposons in the lizard anolis carolinensis shows more similarity to fish than mammals. *Molec. Biol. Evol.* 26: 1811-1822.

PARRA, G., BRADNAM, K. and KORF, I. (2007). CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23: 1061-1067.

RIMOLDI, V., STRANIERO, L., ASSELTA, R., MAURI, L., MANFREDINI, E., PENCO, S., GESU, G.P., DEL LONGO, A., PIOZZI, E., SOLDÀ, G. *et al.,* (2014). Functional characterization of two novel splicing mutations in the OCA2 gene associated with oculocutaneous albinism type II. *Gene* 537: 79-84.

SHEDLOCK, A.M., BOTKA, C.W., ZHAO, S., SHETTY, J., ZHANG, T., LIU, J.S., DESCHAVANNE, P.J. and EDWARDS, S.V. (2007). Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genome. *Proc Natl Acad Sci USA* 104: 2767-2772.

SMIT, A., HUBLEY, R. and GREEN, P. (1996). RepeatMasker Open-3.0. *RepeatMasker Open-3.0.* www.repeatmasker.org.

STANKE, M. and WAACK, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* 19 Suppl 2: ii215-ii225.

TER-HOVHANNISYAN, V., LOMSADZE, A., CHERNOFF, Y.O. and BORODOVSKY, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 18: 1979-1990.

TZIKA, A.C., HELAERS, R., SCHRAMM, G. and MILINKOVITCH, M.C. (2011). Reptiliantranscriptome v1.0, a glimpse in the brain transcriptome of five divergent Sauropsida lineages and the phylogenetic position of turtles. *Evodevo* 2: 19.

TZIKA, A.C. and MILINKOVITCH, M.C. (2008). A Pragmatic Approach for Selecting Evo-Devo Model Species in Amniotes. In *Evolving Pathways: Key Themes in Evolutionary Developmental Biology*, (ed. A, M. and G, F.). Cambridge University Press, pp.123-143.

TZIKA, A.C., ULLATE-AGOTE, A., GRBIC, D. and MILINKOVITCH, M.C. (2015). Reptilian Transcriptomes v2.0: an extensive resource for Sauropsida genomics and transcriptomics. *Genome Biol. Evol.* In press (doi:10.1093/gbe/evv106)

VONK, F.J., CASEWELL, N.R., HENKEL, C.V., HEIMBERG, A.M., JANSEN, H.J., MCCLEARY, R.J.R., KERKKAMP, H.M.E., VOS, R.A., GUERREIRO, I., CALVETE, J.J. *et al.,* (2013). The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc Natl Acad Sci USA* 110: 20651-20656.

WANG, Z., PASCUAL-ANAYA, J., ZADISSA, A., LI, W., NIIMURA, Y., HUANG, Z., LI, C., WHITE, S., XIONG, Z., FANG, D. *et al.,* (2013). The draft genomes of soft-shell turtle and green sea turtle yield insights into the development and evolution of the turtle-specific body plan. *Nat Genet* 45: 701-706.

WOLTERING, J.M. (2012). From Lizard to Snake; Behind the Evolution of an Extreme Body Plan. *Curr. Genomics* 13: 289-299.

Further Related Reading, published previously in the *Int. J. Dev. Biol.*

Sexual dimorphism of AMH, DMRT1 and RSPO1 localization in the developing gonads of six anuran species
Rafal P. Piprek, Anna Pecio, Katarzyna Laskowska-Kaszub,Jacek Z. Kubiak and Jacek M. Szymura
Int. J. Dev. Biol. (2013) 57: 891-895

Dual embryonic origin of the hyobranchial apparatus in the Mexican axolotl (Ambystoma mexicanum)
Asya Davidian and Yegor Malashichev
Int. J. Dev. Biol. (2013) 57: 821-828

Clonal analyses in the anterior pre-placodal region: implications for the early lineage bias of placodal progenitors
Sujata Bhattacharyya and Marianne E. Bronner
Int. J. Dev. Biol. (2013) 57: 753-757

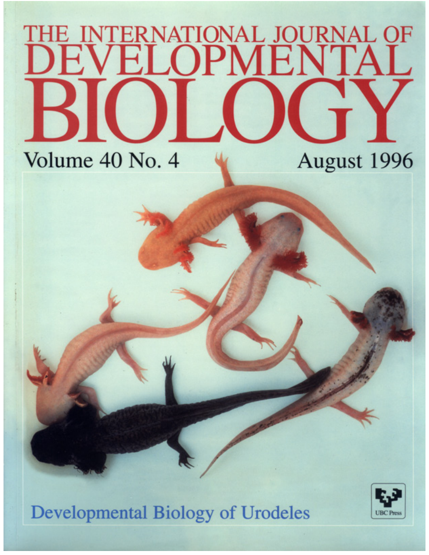
Amphibian interorder nuclear transfer embryos reveal conserved embryonic gene transcription, but deficient DNA replication or chromosome segregation
Patrick Narbonne and John B. Gurdon
Int. J. Dev. Biol. (2012) 56: 975-986

Origins of Cdx1 regulatory elements suggest roles in vertebrate evolution
Stephen J. Gaunt and Yu-Lee Paul
Int. J. Dev. Biol. (2011) 55: 93-98

Reptile scale paradigm: Evo-Devo, pattern formation and regeneration
Cheng Chang, Ping Wu, Ruth E. Baker, Philip K. Maini, Lorenzo Alibardi and Cheng-Ming Chuong
Int. J. Dev. Biol. (2009) 53: 813-826

Proteomics analysis of regenerating amphibian limbs: changes during the onset of regeneration
Michael W. King, Anton W. Neff and Anthony L. Mescher
Int. J. Dev. Biol. (2009) 53: 955-969

5 yr ISI Impact Factor (2013) = 2.879


THE INTERNATIONAL JOURNAL OF
DEVELOPMENTAL
BIOLOGY
Volume 40 No. 4
August 1996
Developmental Biology of Urodeles