Multi-tissue transcriptomes of caecilian amphibians highlight incomplete knowledge of vertebrate gene families

María Torres-Sánchez, Christopher Creevey, Etienne Kornobis, David Gower, Mark Wilkinson, Diego San Mauro

DNA Research, 2019, 26(1), 13–20
doi: 10.1093/dnares/dsy034
Advance Access Publication Date: 22 October 2018
Full Paper

Full Paper

Multi-tissue transcriptomes of caecilian amphibians highlight incomplete knowledge of vertebrate gene families

María Torres-Sánchez[1†], Christopher J. Creevey[2], Etienne Kornobis[3], David J. Gower[4], Mark Wilkinson[4], and Diego San Mauro [1*]

[1]Department of Biodiversity, Ecology and Evolution, Complutense University of Madrid, 28040 Madrid, Spain, [2]Institute for Global Food Security, School of Biological Sciences, Queen's University Belfast, Belfast, BT7 1NN, UK, [3]Institut Pasteur, Bioinformatics and Biostatistics Hub, C3BI, USR 3756 IP CNRS, Paris, France, and [4]Department of Life Sciences, The Natural History Museum, London SW7 5BD, UK

*To whom correspondence should be addressed. Tel. +34 913944948. Fax. +34 913944947. Email: dsanmaur@ucm.es

[†]Present address: Department of Neuroscience, Spinal Cord and Brain Injury Research Center & Ambystoma Genetic Stock Center, University of Kentucky, Lexington, KY, 40536, USA

Edited by Dr. Yuji Kohara

Received 13 February 2018; Editorial decision 10 September 2018; Accepted 13 September 2018

Abstract

RNA sequencing (RNA-seq) has become one of the most powerful tools to unravel the genomic basis of biological adaptation and diversity. Although challenging, RNA-seq is particularly promising for research on non-model, secretive species that cannot be observed in nature easily and therefore remain comparatively understudied. Among such animals, the caecilians (order Gymnophiona) likely constitute the least known group of vertebrates, despite being an old and remarkably distinct lineage of amphibians. Here, we characterize multi-tissue transcriptomes for five species of caecilians that represent a broad level of diversity across the order. We identified vertebrate homologous elements of caecilian functional genes of varying tissue specificity that reveal a great number of unclassified gene families, especially for the skin. We annotated several protein domains for those unknown candidate gene families to investigate their function. We also conducted supertree analyses of a phylogenomic dataset of 1,955 candidate orthologous genes among five caecilian species and other major lineages of vertebrates, with the inferred tree being in agreement with current views of vertebrate evolution and systematics. Our study provides insights into the evolution of vertebrate protein-coding genes, and a basis for future research on the molecular elements underlying the particular biology and adaptations of caecilian amphibians.

Key words: gene families, Gymnophiona, phylogenomics, RNA-seq, skin-specific genes

1. Introduction

High-throughput sequencing (HTS) technologies and associated bioinformatics are transforming the study of evolutionary and comparative genetics, offering an unprecedented opportunity to characterize and understand diversity and function in both model and non-model organisms.[1–3] In this context, one recent revolution is the use of HTS technologies to analyse sets of RNA molecules, transcriptomes, on a massively parallel scale.[4,5] The transcriptome is a snapshot in time of genes transcribed in the tissue

or cells sampled. Investigation of transcriptomes can allow the identification of functional elements of genomes, reveal molecular constituents of cells and tissues, help understand organismal development and disease,[6] and has the potential to uncover the role of tissue-specific evolution in biological diversity.[7] Having entered the phylogenomics era, RNA-seq has also become a powerful complement of *de novo* genome sequencing, particularly helping with functional annotation[8] and gene expression assessment, and is sometimes the only practical approach to scan and survey gene diversity in organisms with large genomes that still lack reference genomic data.[9] A general strategy for this approach is to pool the mRNA data from a wide range of tissues (from different individuals and/or stages of development) to assemble a reference dataset of the genes of the species (i.e. a proxy of the reference genome of the species).

We have applied the pooling of tissue-specific reads from RNA-seq to the study of tissue-specific transcriptomic landscapes of five species of caecilian amphibians (order Gymnophiona) representing four of the ten currently recognized families (Caeciliidae, Rhinatrematidae, Siphonopidae and Typhlonectidae) and a range of ecologies and degrees of evolutionary divergence (including coverage of both branches of the basal evolutionary divergence within the order).[10] Caecilians are, along with frogs and salamanders, one of the three orders of extant amphibians. They are a highly specialized group with elongate, annulated, limbless bodies, reduced visual systems and with paired bilateral sensory tentacles on the snout.[11] There are 207 currently recognized extant species classified in 32 genera, with mainly tropical distributions and mainly burrowing habits.[12–14] Most are terrestrial as adults, living in soil, but several species of the Typhlonectidae (including the one sampled here) are fully aquatic. Caecilians are an old group, with at least 250 million years (myr) of separate evolution from their sister-group, the frogs and salamanders.[15–19] Due to their specialized body form, ecological distinctiveness and phylogenetic position in the vertebrate tree of life, caecilians are interesting for macro-evolutionary, life history and evolutionary developmental biology research.[11]

We provide a first large-scale characterization of caecilian genomes using multi-tissue transcriptomic landscapes generated with RNA-seq. We use two complementary approaches to investigate features of caecilian protein-coding sequences in a vertebrate comparative framework. First, we assess the degree to which homologous elements of caecilian functional genes of varying tissue specificity can be identified across 51 other vertebrates. This reveals a high number of unclassified candidate gene families that are transcribed differentially across tissue types in caecilians. Comparisons between the already known vertebrate gene families and the potentially novel gene families found in caecilians highlight the relevance of skin-specific genes and the poor characterization of the molecular elements of caecilian skin. Here, we start addressing this knowledge gap by identifying protein domains for the caecilian skin-specific genes. Second, we infer the phylogenetic relationships of the five sampled caecilian species and the same set of 51 vertebrates based on candidate orthologous genes. This study provides new information about the functional elements of the genome and phylogenomics of caecilians and highlights distinctive and singular genes for the most neglected amphibian order.

2. Materials and methods

2.1. Sample preparation and high-throughput sequencing

This study includes novel data from five caecilian species: *Rhinatrema bivittatum* (Guérrin-Méneville, 1838), *Caecilia tentaculata* Linnaeus, 1758, *Typhlonectes compressicauda* (Duméril & Bibron, 1841), *Microcaecilia unicolor* (Duméril, 1861) and *Microcaecilia dermatophaga* Wilkinson, Sherratt, Starace & Gower, 2013. Different tissues (skin, posterior skin [from the posterior end of the body], foregut, muscle, liver, kidney, lung, heart, spleen and testis) were collected from freshly sacrificed, captive (but wild caught, in French Guiana) maintained specimens anesthetized with tricaine methanesulphonate (MS222). Biopsy samples were cut into pieces thinner than 0.25 cm in any single dimension, immediately soaked in RNAlater stabilization solution (Qiagen), incubated at 4°C overnight (to allow the solution to thoroughly penetrate the tissue) and stored at −20°C. Numbers of specimens and of tissues sampled per species, voucher and sampling information are given in Table 1 and Supplementary Table S1.

RNA was isolated using the RNeasy Fibrous Tissue Mini Kit (Qiagen) using the manufacturer's instructions, following tissue disruption and homogenization with TissueRuptor (Qiagen). RNA quantity and quality was assessed with Qubit 2.0 fluorometer, NanoDrop 1000 spectrophotometer and Agilent 2100 Bioanalyzer (RNA Nano Chip). Forty RNA extractions with RNA integrity number, RIN,[20] values ranging from 7.8 to 10 were selected for RNA-seq. These 40 selected samples included RNA extractions of skin, liver and kidney for all five caecilian species, as well as a selection of other tissues (foregut, muscle, lung, heart, spleen, testis) each available for only a subset of the species (see Supplementary Table S1). Unstranded paired-end sequencing after poly-A enrichment and TruSeq library preparation was carried out on the Illumina HiSeq2000 platform at Macrogen (16 RNA extraction samples) and BGI Tech Solutions (24 RNA extraction samples) using ten dual flow cells, two lanes per sample. All RNA extractions from the same tissue were sequenced by the same company.

2.2. Raw data processing and *de novo* assembly

Paired-end RNA-seq raw reads (100 nucleotides long) of each of the 40 tissue samples were trimmed individually and filtered by PRINSEQ 0.20.3[21] after inspection of the FastQC 0.11.2[22] quality control report. In all cases, the first 15 bases from the 5′ end of the reads, optical duplicates and reads with an average Phred quality score[23] below 25 were removed. Separate *de novo* assemblies were performed for each of the five caecilian species employed in the study (species-specific transcriptome assemblies). These were carried out by pooling together all reads (filtered and trimmed) for tissue samples belonging to the same species (Supplementary Table S1). Reads were also pooled for all (both) specimens for each of the two species for which multiple specimens were sampled. A few preliminary *de novo* assembly runs of separate tissue samples (single-tissue transcriptome assemblies) were conducted on the TRUFA platform[24] to explore parameter settings and run times.

De novo species-specific assemblies were performed with Trinity r20140717[25] using 60 Gb of RAM (–max_memory 60G) and prior *in silico* normalization (with otherwise default settings[26]). TransDecoder 2.0[26] was used with default settings to identify candidate protein-coding genes from the subsets of contigs with open reading frame (ORFs) in the five caecilian species-specific transcriptomes. Reads were mapped back to each assembly with Bowtie 2.0.2,[27] post-processed with SAMtools[28] and gene expression was estimated using the counts of reads mapping to each assembly with HTSeq 0.6.1.[29] Multiple measures (N50, median contig length, average contig length, alignment percentage) were used for assessing the accuracy of each of the five caecilian species-specific assemblies.[30,31] Likewise, we used a computational method, CEGMA 2.4,[32] to estimate the percentage completeness of each caecilian

Table 1. Information on the species-specific caecilian transcriptome assemblies and their annotation

Species	N	T	Contigs	% CEGs	Protein-coding genes	veNOG annotation	KVGF annotation
Caecilia tentaculata	1	10	142,502	97.18	27,384	18,368	12,937
Microcaecilia dematophaga	1	4	106,298	97.18	22,058	17,099	11,670
Microcaecilia unicolor	2	9	146,348	97.58	26,302	18,487	12,719
Rhinatrema bivittatum	2	10	201,584	97.58	34,654	19,863	13,429
Typhlonectes compressicauda	1	7	134,394	97.58	27,603	18,302	12,293

N: number of specimens; T: number of tissues; % CEGs: percentage completeness core eukaryotic genes; veNOG annotation: number of genes with similarity match in veNOG database; KVGF annotation: number of known vertebrate gene families with caecilian genes.

transcriptome, and compared these with the completeness percentages of the genome assemblies of the frog *Xenopus tropicalis* Gray, 1864 v9.0 and v4.1.[33] Finally, we compared our species-specific transcriptomes to other transcriptomes recently generated for four species of caecilians, including for two of our sampled species (*R. bivittatum*, *T. compressicauda*, *T. natans* [Fischer, 1880] and *Geotrypetes seraphini* [Duméril, 1859]).[34] These previously published caecilian transcriptomes are not associated with tissue-expression information and they contain fewer ORFs than do our transcriptomes for the same species. Using similarity searches, we determined that the vast majority (89.83%) of the protein-coding genes from the previous transcriptomes occur also in our transcriptomes (using BLAST, blastp version 2.2.28[35] with *e*-value threshold of 1e-20: data not shown). Thus, the previously published caecilian genomic data were not used in our subsequent analyses.

2.3. Multigene family analysis

Contigs of the five new species-specific caecilian transcriptomes containing ORFs were aligned against predefined vertebrate-specific gene families (veNOGs) from the EggNOG 4.1 database[36] using blastp, applying a conservative *e*-value threshold of 1e-20 (applying less conservative 1e-10 or 1e-5 cutoffs does not result in substantially greater annotation percentages: data not shown). Contigs with expression levels below 100 total read counts were discarded and not used in subsequent analyses. We classified all caecilian annotations (from the pooled contigs of the five species) according to the gene-expression presence across the tissues sampled. For tissue expression analysis, contigs were postulated as being expressed in a particular tissue of a particular transcriptome if they had a minimum of 10 reads aligning to them. This allowed a scale of 'tissue presence' to be generated, ranging from those genes found expressed in every tissue type to those found expressed in only one tissue type. The distribution of all homologues of the caecilian protein-coding genes on the vertebrate taxonomy tree from the NCBI taxonomy database was generated and visualized using phyloT and ITOL,[37] respectively. Vertebrate taxonomy tree was built using the unique identifier, taxids, of the species that are included in the EggNOG database.

Where possible, caecilian gene families were annotated with the same function as those vertebrate gene families with the best BLAST match (smallest *e*-value and highest BIT score) in EggNOG identified above. Transcripts with no hits to the known vertebrate gene families in EggNOG were clustered using CD-HIT 4.6.4[38] with a threshold of 90% amino acid sequence identity to ensure same function of the sequences clustered. These clusters were compared against protein-coding genes from currently available amphibian genomes (*Ambystoma mexicanum* [Shaw & Nodder, 1798], *Nanorana parkeri* [Stejneger, 1927], *Rana catesbeiana* Shaw 1802 and *Xenopus laevis* [Daudin, 1802])[33,39–41] that are not included in the EggNOG database, using blastp, with an *e*-value threshold of 1e-20. Clusters that remained without similarity hits after the searches against the amphibian genomes were classified as potentially (candidate) novel caecilian gene families. Of these, we calculated the number of tissues in which any gene family was expressed (as described earlier). In addition, to characterize the different tissues with a more restrictive approach than the previously used tissue presence classification, tissue specificity was postulated when 95% of total read counts in each caecilian contig belonged to a single tissue for both unclassified and known vertebrate gene families. To test if there was a greater number of candidate novel genes specific to a particular tissue type than expected by chance, the relative abundance of known vertebrate gene families versus those of candidate novel caecilian gene families were compared using a two-tailed Fisher's exact test conducted with R 3.3.0,[42] with the null hypothesis that there was no difference in the numbers of tissue-specific candidate novel genes. Finally, our characterization of tissue specificity expression was completed with the inference of protein–protein interactions (PPIs) and functional enrichment pathways using STRING[43] with the option of auto-detect organism for the known vertebrate gene families; and the Pfam[44] annotation of the uncharacterized, candidate novel caecilian gene families using HMMER 3.0[45] with default parameters to identify protein domains.

2.4. Orthology prediction and phylogenomic analysis

To carry out a phylogenomic analysis we identified candidate orthologous genes from across vertebrates, including our caecilian samples. To do this we used OrthoFinder 0.2.8[46] and used as input all predicted protein-coding genes from the caecilian transcriptomes and all protein-coding sequences for the 51 vertebrates represented in the EggNOG database. From the results of OrthoFinder analysis we filtered out any groups (orthogroups) that had more than one gene copy in one species (co-orthologues for different species and paralogues for the same species). Multiple-sequence alignments were performed individually for each of the resulting filtered orthogroups using MAFFT 7.245[47] with default settings, and individual gene trees were inferred using approximately maximum-likelihood with FastTree 2.1.8[48] and the JTT+CAT model of amino acid substitutions.[49] We reconstructed a supertree using ASTRAL 4.10.11, which provides statistically consistent species tree inference from gene trees subject to incomplete lineage sorting,[50,51] and computed posterior probabilities and quartet support for the internal branches of the main recovered topology.

3. Results and discussion

3.1. *De novo* transcriptome assemblies

In total, RNA sequencing yielded nearly two billion reads (1,963,110,986), averaging 49 million reads per library. The five species-specific assemblies from pooled reads of all tissues of each

species resulted in transcriptomes of a mean of 146,227 contigs with N50 values of 1,263–1,884 (Supplementary Table S2). Tissue-specific RNA-seq reads and species-specific *de novo* transcriptome assemblies are available from NCBI through BioProject ID number PRJNA387587. The maximum and minimum contig lengths were 27,126 and 201 (default minimum size parameter used in the assembly program) bases, respectively. The longest contig was reconstructed from the *R. bivittatum* transcriptome and only a few very long (see Supplementary Fig. S1) contigs were present in any of the species-specific caecilian transcriptomes. In addition to transcriptome metrics, we assessed the quality of the *de novo* assemblies by the extent to which each pair of raw reads (more than 95%) could be mapped to the same contig (Supplementary Table S2). On average, 27,600 protein-coding genes were identified from the contigs with ORFs, (Table 1 and Supplementary Table S2). Our caecilian transcriptome reconstructions were supported also by the annotation. At least 241 of 248 ultra-conserved core eukaryotic genes (CEGs) occur in all five species-specific transcriptomes (Table 1). For the sake of comparison, we checked also the presence of CEGs in two different genome assemblies of *X. tropicalis* and found 225 CEGs in the most recent (v9.0) and 219 in an earlier version (v4.1).

On the basis of the quality of our transcriptome assembly reconstructions, we obtained useful reference genomic records for caecilian amphibians, the first to our knowledge that are broad and diverse in terms of species and tissues sampled. Although the metrics used to assess the quality of assemblies of transcriptomic data are controversial[30] our caecilian transcriptome sequences contain more CEGs than the two genome assemblies of *X. tropicalis* used for comparison, suggesting that our reference species-specific transcriptomes are fairly complete (Table 1). Even so, the generated reference transcriptomes are not fully complete, missing specific genes related to developmental stages and to tissues not sampled in our study. As with estimates for other vertebrates, the number of protein-coding genes identified in the species-specific caecilian transcriptomes is approximately 25,000 (Table 1), and a relatively high percentage of such proteins were annotated, which is also indicative of accurate transcriptome reconstruction. Gene identification is one of the major challenges of *de novo* transcriptome assembly, even for Trinity assembly of paired-end sequence data that enables potentially confounding sources of variation such as alternative splicing and paralogous genes to be overcome.[25] Thus, the numbers of protein-coding genes could be overestimated. An additional problem is that the transcriptomes are not composed solely of transcripts from protein-coding genes. Recently, it has been demonstrated that almost the entire genome is transcribed.[52] Accordingly, caecilian contigs that are not protein-coding genes or degradation products of the same, nor possible chimeras, are postulated to be long non-coding RNAs and potentially important regulatory elements.

3.2. Vertebrate gene families and unclassified gene families

The vast majority of the annotated caecilian genes that are homologous with those vertebrate genes in the EggNOG database, are expressed in most of the (up to nine) sampled caecilian tissue types. This could be interpreted as indicative of constitutive expression and many might be housekeeping genes. Only a small proportion (see Fig. 1, one tissue fraction) of the caecilian genes with matches to EggNOG (and thus annotated) are tissue specific. This same pattern was found when comparing the pooled caecilian sample (all five species) with each of the 51 EggNOG database vertebrates, with no obvious phylogenetic pattern (Fig. 1). The number of caecilian contigs with matches to known vertebrate genes ranged from 17,099 to 19,863 per caecilian species (Table 1), representing 57.32–77.52% (mean 67.70%) of all caecilian protein-coding genes. We found that 38.75–52.91% (mean 46.36%) of the annotated caecilian genes were classified into vertebrate gene families from EggNOG.

To investigate and quantify the importance of the uncharacterized genes in caecilians, we grouped these protein-coding sequences into multigene families and filtered them by excluding clusters with close similarity to genes from the available amphibian genomes. If caecilian genomes did not contain genes novel for vertebrates, it would be expected that the vast majority of their genes would belong to some already described, known vertebrate gene family or have homologous sequences in the reported amphibian genomes. However, our results indicate that less than half of the caecilian gene families belong to known vertebrate gene families. Given the sparse taxon sampling and the currently poor genomic reference record for amphibians, at least some of the unclassified gene families in caecilians could contain genes from other vertebrate taxa or be amphibian rather than caecilian specific. The absence of homologues of these caecilian gene families in other vertebrate species might reflect gene loss events[53,54] or, alternatively, faster sequence evolution in some caecilian genes. Either way, caecilians likely have many functional elements that are novel for vertebrates.

A total of 177 known vertebrate and 422 novel caecilian gene families exhibit tissue-specific expression (Table 2). A significantly greater number of novel caecilian genes were expressed only in skin (P-value = 4.5e-05, Fisher's exact test). In contrast, caecilian spleen transcripts had significantly lower than expected tissue-specific novel gene families (P-value = 0.01935, Fisher's exact test). Among the tissue-specific known vertebrate gene families, we found significantly more predicted protein–protein interactions (PPIs) than expected by chance and functional enrichment of metabolic pathways in five caecilian tissues (foregut, kidney, liver, spleen and testis, see Supplementary Table S3). The functional enrichments observed tend to relate to well-characterized processes in these tissues such as nutrient absorption in the foregut samples (GO: 0007586), organic acid, anion and amino acid transmembrane transport in the kidney samples (GO:1903825, GO:0098656, GO:0003333), and regulation of fibrinolysis in the liver (GO:0051918), (see Supplementary Table S3). In contrast, in skin samples we did not observe significantly more PPIs than expected by chance, or functional enrichment of pathways in the genes with known annotations. This may be because the vast majority (87%) of genes with tissue-specific gene expression in skin did not match any known vertebrate gene families, the highest of any of the tissues examined (Table 2). This analysis suggests that skin-specific vertebrate gene families remain poorly characterized in general and likely have unknown, innovative functions and interactions.

3.3. Skin-specific genes of caecilians

Potentially novel caecilian gene families (those without hits to known genes) expressed in skin were annotated with protein domains that might be associated causally with specializations of caecilian skin.[55,56] From the uncharacterized tissue-specific clusters (108 in the skin), a total of 91 different protein domains were identified (Supplementary Table S4), including 16 domains occurring exclusively in the skin in our analysis, such as diverse proteases, amino acid storage receptors and toxin-like domains. Skin forms the barrier between the organism and the environment both physically and

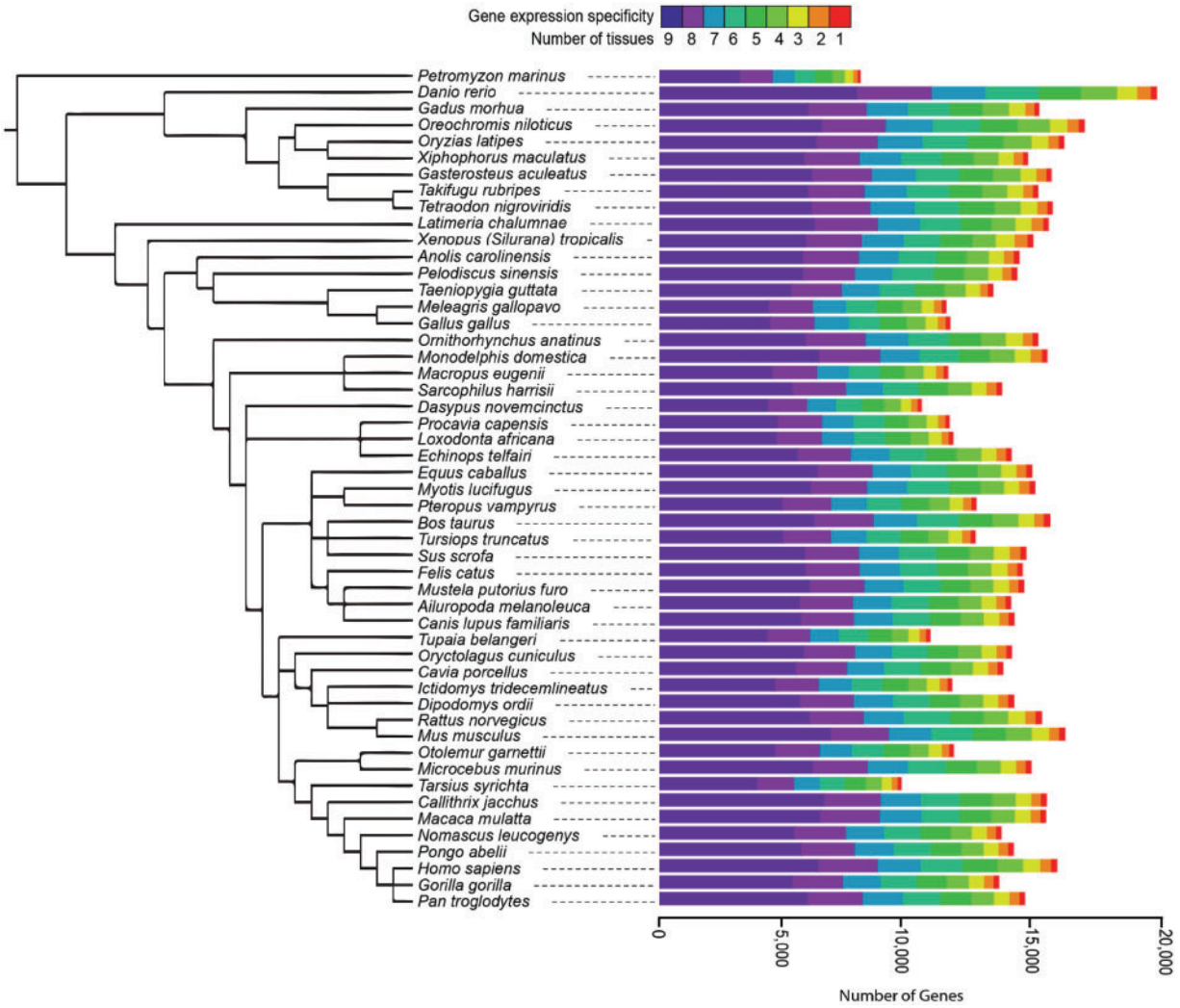

Figure 1. Numbers and tissue presence of the annotated genes found in caecilian transcriptomes. Genes were pooled for the five sampled species-specific transcriptomes and annotated in the 51 vertebrate species available on the EggNOG database, and mapped onto a vertebrate phylogeny inferred from the NCBI's taxids (using phyloT and ITOL). For each vertebrate taxon, the number of caecilian annotated genes is subdivided to show the number of caecilian tissue types in which those genes are expressed.

Table 2. Novel tissue-specific genes in caecilians

	Foregut	Heart	Kidney	Liver	Lung	Muscle	Skin	Spleen	Testis	Total
Number of transcriptomes analysed	4	2	5	7	4	3	11	2	2	40
Known vertebrate gene families	19	4	21	18	3	6	15	11	80	177
Gene families shared with the other sampled non-caecilian amphibians	6	2	6	15	2	3	22	—	25	81
Candidate novel caecilian gene families	32	12	40	44	9	27	108	8	142	422
P-value	0. 2671	0.7887	0.4639	1	1	0.2355	**4.5e-05**	**0.01935**	0.07605	—

The number of transcriptomes determined for each tissue, and the tissue-specific gene families (caecilian gene families that are already known vertebrate gene families, caecilian gene families shared with the other four sampled non-caecilian amphibians, and candidate caecilian-specific gene families) are shown. The last row shows the *P*-value (significant values in bold font) for Fisher's exact test of the difference between the abundance of known vertebrate gene families and those of uncharacterized candidate novel caecilian gene families. Skin tissue includes skin samples from different parts of the body: skin and posterior skin samples, see Supplementary Table S1.

biochemically. It is genetically and physiologically very active throughout an animal's life. Amphibian skin is multifunctional with additional roles in respiration, water regulation, and in defence against predators and pathogens.[57,58] The defensive properties of amphibian skin rely mainly on biochemical substances secreted from specialized skin granular glands.[59] These secretions can contain numerous bioactive components, including alkaloids, biogenic amines, peptides and proteins,[60] some of which have been isolated and

studied, particularly in frogs and salamanders.[61–63] The diversity of functions and biochemical activities of amphibian skin makes it unsurprising that caecilians present specific expression patterns of novel genes, particularly considering their 250+ myr of separate evolutionary history from the other major amphibian lineages[15–19] and the sustained contact between the skin and soil for most caecilian species. Indeed, some of the protein domains found exclusively in caecilian skin-specific novel gene families, such as proteases and toxin-like domains (Asp_protease_2, gag-asp_proteas, Toxin_TOLIP, UPAR_LY6, see Supplementary Table S4) point to novel skin defensive mechanisms.

The maternal skin of some caecilian species plays another unique role: in provision of nutrition to newborns (maternal dermatophagy).[64,65] This behaviour occurs in several of the species sampled in this study (observed in *M. dermatophaga*, likely also present in *M. unicolor* and *C. tentaculata*).[10] This phenomenon is especially interesting for understanding the evolution of viviparity because it is possibly a precursor of the oviduct feeding by fetuses that occurs in viviparous caecilians.[62] Maternal dermatophagy involves structural and histochemical changes in the mothers' epidermis, it becomes hypertrophied and heavily invested with energy reserves,[64] and hence expanded gene machinery is likely needed. Amino acid storage receptor (PhaP_Bmeg, see Supplementary Table S4) is another protein domain found in skin-specific novel gene families that might be related to the unique parental care of caecilian amphibians. A final feature of caecilian skin that makes it so distinctive is the presence of scales.[66] Scales are absent in other extant amphibians but are present, concealed in dermal pockets, in many caecilians (all except *T. compressicauda* of those sampled in our study). Some of the skin-specific gene families with domains of unknown function (DUF, see Supplementary Table S4) might be involved in the production and maintenance of scales.

Further data and analyses are required to identify the taxonomic distribution, diversity and function of these candidate skin-specific gene families. Greater tissue sampling in the future may reveal similar patterns in other tissues, such as testis or gut, that present particularities in caecilians with respect to other amphibians that may be reflected in their genomes. For example, caecilians differ from other amphibians in that males have a copulatory organ formed from the eversible final part of the gut,[67] as well as other autapomorphies of the sperm and internal fertilization specializations such as the Müllerian gland and the ejaculate.[68]

3.4. Phylogenomic dataset

We obtained a total of 23,761 groups or orthogroups, of which 1,955 were groups comprising genes with only one copy in at least four vertebrate taxa. The filtered orthogroups seemingly contain no paralogous genes, at least from the same species, and are straightforward for use in phylogenomics and the study of evolutionary processes that depend upon inferred phylogenetic relationships.[69] The number of analysed genes found in each species is detailed in Supplementary Table S5. For each of the 1,955 orthogroups phylogenetic gene trees were inferred. A supertree was retrieved from the gene trees under a multi-species coalescent model, maximizing the number of induced quartet trees (the supertree is presented in Supplementary Fig. S2). The normalized quartet score of the main topology was 0.798 (i.e. 79.8% of the quartet trees displayed by our gene trees are displayed by the supertree). The supertree constructed from the gene trees of the candidate orthologous groups recovered the main known topology of this subset of the Tree of Life (Supplementary Fig. S2). Branches within the caecilian part of the supertree are well supported as judged by both posterior probabilities and quartet support values. Among the sampled vertebrates, Lissamphibia and Gymnophiona are recovered as monophyletic, and the inferred relationships among the five caecilian species are fully congruent with those inferred in other (non-phylogenomic and phylogenomic) molecular analyses.[10,34] Our results indicate that combining the information from putative orthologous genes using supertree approach is adequate to reconstruct the phylogenetic relationships among the sampled caecilians, and vertebrates in general.

4. Concluding remarks

As with other studies that have characterized transcriptomes,[9] this study has a strong descriptive component, but it has yielded novel discoveries and represents an important turning point for genomic studies in caecilians (and vertebrates), improving prospects for future research. The species-specific *de novo* transcriptomes of caecilian amphibians presented here could be improved by additional sequencing of different tissues, individuals and developmental stages (e.g. the transcriptome of *M. dermatophaga* was built from only four tissue-type samples). In terms of sampling and biological replicates, only the species-specific transcriptomes of *R. bivittatum* and *M. unicolor* were reconstructed using more than one (two) specimen each. Obtaining fresh biological samples has been a limiting step for research on many caecilian species,[70] and dedicated fieldwork will likely be required to investigate broadly the genomic potential of this neglected, but important group of vertebrates.

Genome science has irreversibly changed the landscape of biological research. Understanding life processes and their evolutionary changes by reading the complete set of encoded instructions that each species holds is increasingly becoming a reality. Nonetheless, achieving this goal thoroughly still remains a challenge for most groups of organisms. Of the almost 6,600 eukaryotic genomes available on the NCBI database, only six records are of amphibian species: *A. mexicanum*, *N. parkeri*, *R. catesbeiana*, *Rhinella marina Linnaeus*, 1758, *X. laevis* and *X. tropicalis* (21 September 2018, date last accessed). Despite the great effort made by initiatives such as the Genome 10 K Project[71,72] and other genome-scale studies (e.g. Xenbase,[33] Salamander Genome project[73]), amphibians are the major group of vertebrates with fewest genomic resources available, and, importantly, there are none for the order Gymnophiona.[74] The lack of at least one representative organism of each of the three extant amphibian orders has compromised the diversity of comparable genomic resources for vertebrates, as well as the opportunities for evolutionary and phylogenomic research. To start filling this gap, here we have reported transcriptomic data for five caecilian amphibian species, including first genomic records for three species (*C. tentaculata*, *M. unicolor* and *M. dermatophaga*), and characterized several unclassified candidate gene families with tissue-specific expression, especially in the skin. This provides insights into the evolution of vertebrate protein-coding genes, and further establishes the basis for gene-discovery work as well as investigation of the molecular elements underlying the singular biology of caecilian amphibians.

Acknowledgements

We thank Ainhoa Agorreta, Cristina Frías-López, Julio Rozas, Rafael Zardoya, Kim Roelants, Karen Siu-Ting, Jeff Streicher and Iván de la Hera for insightful comments, help, and advice on this project. We thank Le Comité Scientifique Régional du Patrimoine Naturel for approving our caecilian research in French Guiana, Myriam Virevaire (Direction de l'Environnement, de

l'Aménagement et du Logement, Guyanne), Céline Dupuy and Nicolas Krieger (Direction des Services Vétérinaires de la Guyane, Cayenne) for providing export permits and Jérôme Chave, Patrick Chatelet and Philippe Gaucher (Centre National de la Recherche Scientifique, Cayenne) and Jeannot and Odette (Camp Patawa) for facilitating our research on the caecilian fauna of French Guiana. Two anonymous reviewers gave insightful comments on an earlier version of the manuscript. Computational analyses were performed at the Altamira HPC cluster of the Institute of Physics of Cantabria (IFCA-CSIC), which is part of the Spanish Supercomputing Network.

Accession numbers

Tissue-specific RNA-seq reads and species-specific *de novo* transcriptome assemblies are available from NCBI through BioProject ID number PRJNA387587. SRA database accession numbers are also provided in Supplementary Table S1.

Funding

This work received financial support from the Ministry of Economy and Competitiveness of Spain (RYC-2011-09321 and CGL2012-40082 grants to DSM, BES-2013-062723 FPI predoctoral fellowship and EEBB-I-15-09665 research stay to MTS). Support was also provided by the network of research laboratories working on adaptation genomics (AdaptNET) funded by the Ministry of Economy and Competitiveness of Spain (grant CGL2015-71726-REDT).

Supplementary data

Supplementary data are available at *DNARES* online.

Conflict of interest

None declared.

References

1. Mardis, E. R. 2008, The impact of next-generation sequencing technology on genetics, *Trends Genet.*, **24**, 133–41.
2. Rokas, A. and Abbot, P. 2009, Harnessing genomics for evolutionary insights, *Trends Ecol. Evol. (Amst.)*, **24**, 192–200.
3. da Fonseca, R. R., Albrechtsen, A. and Themudo, G. E. 2016, Next-generation biology: sequencing and data analysis approaches for non-model organisms, *Mar. Genomics*, **30**, 3–11.
4. Nagalakshmi, U., Waern, K. and Snyder, M. 2010, RNA-seq: a method for comprehensive transcriptome analysis, *Curr. Protoc. Mol. Biol.*, **4**, 1–13.
5. Conesa, A., Madrigal, P. and Tarazona, S. 2016, A survey of best practices for RNA-seq data analysis, *Genome Biol.*, **17**, 13.
6. Wang, Z., Gerstein, M. and Snyder, M. 2009, RNA-seq: a revolutionary tool for transcriptomics, *Nat. Rev. Genet.*, **10**, 57–63.
7. Ozsolak, F. and Milos, P. M. 2011, RNA sequencing: advances, challenges and opportunities, *Nat. Rev. Genet.*, **12**, 87–98.
8. Gibbons, J. G., Janson, E. M., Hittinger, C. T., Johnston, M., Abbot, P. and Rokas, A. 2009, Benchmarking next-generation transcriptome sequencing for functional and evolutionary genomics, *Mol. Biol. Evol.*, **26**, 2731–44.
9. Ekblom, R. and Galindo, J. 2011, Applications of next generation sequencing in molecular ecology of non-model organisms, *Heredity (Edinburgh)*, **107**, 1–15.
10. San Mauro, D., Gower, D. J., Müller, H., et al. 2014, Life-history evolution and mitogenomic phylogeny of caecilian amphibians, *Mol. Phylogenet. Evol.*, **73**, 177–89.
11. Wilkinson, M. 2012, Caecilians, *Curr. Biol.*, **22**, R668.
12. Sherratt, E., Gower, D. J., Klingenberg, C. P. and Wilkinson, M. 2014, A nine-family classification of caecilians (Amphibia: Gymnophiona), *Evol. Biol.*, **41**, 528–64.
13. Kamei, R. G., San Mauro, D., Gower, D. J., et al. 2012, Discovery of a new family of amphibians from northeast India with ancient links to Africa, *Proc. Biol. Sci.*, **279**, 2396–401.
14. Darrel, R. F. 2016, Amphibian Species of the World: Version 6.0. AMNH. Available at: http://research.amnh.org/vz/herpetology/amphibia/ (21 September 2018, date last accessed).
15. Roelants, K., Gower, D. J., Wilkinson, M., et al. 2007, Global patterns of diversification in the history of modern amphibians, *Proc. Natl. Acad. Sci. USA.*, **104**, 887–92.
16. Zhang, P. and Wake, D. B. 2009, Higher-level salamander relationships and divergence dates inferred from complete mitochondrial genomes, *Mol. Phylogenet. Evol.*, **53**, 492–508.
17. San Mauro, D. 2010, A multilocus timescale for the origin of extant amphibians, *Mol. Phylogenet. Evol.*, **56**, 554–61.
18. Pyron, R. A. 2011, Divergence time estimation using fossils as terminal taxa and the origins of lissamphibia, *Syst. Biol.*, **60**, 466–81.
19. Marjanović, D. and Laurin, M. 2014, An updated paleontological timetree of lissamphibians, with comments on the anatomy of Jurassic crown-group salamanders (Urodela), *J. Hist. Biol.*, **26**, 535–50.
20. Mueller, O., Lightfoot, S. and Schroeder, A. 2016, RNA integrity number (RIN)—standardization of RNA quality control application, *Agil. Appl. Note*, 5989-1165EN.
21. Schmieder, R. and Edwards, R. 2011, Quality control and preprocessing of metagenomic datasets, *Bioinformatics*, **27**, 863–4.
22. Andrews, S. 2010, FastQC: a quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/
23. Ewing, B., Hillier, L., Wendl, M. C. and Green, P. 1998, Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res.*, **8**, 175–85.
24. Kornobis, E., Cabellos, L., Aguilar, F., et al. 2015, TRUFA: a user-friendly web server for *de novo* RNA-seq analysis using cluster computing, *Evol. Bioinform. Online*, **11**, 97–104.
25. Grabherr, M. G., Haas, B. J., Yassour, M., et al. 2011, Full-length transcriptome assembly from RNA-seq data without a reference genome, *Nat. Biotechnol.*, **29**, 644–52.
26. Haas, B. J., Papanicolaou, A., Yassour, M., et al. 2013, *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis, *Nat. Protoc.*, **8**, 1494–512.
27. Langmead, B., Trapnell, C., Pop, M. and Salzberg, S. 2009, Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol.*, **10**, R25.
28. Li, H., Handsaker, B., Wysoker, A., et al. 2009, The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078–9.
29. Anders, S., Pyl, P. T. and Huber, W. 2015, HTSeq—a Python framework to work with high-throughput sequencing data, *Bioinformatics*, **31**, 166–9.
30. O'Neil, S. T. and Emrich, S. J. 2013, Assessing *de novo* transcriptome assembly metrics for consistency and utility, *BMC Genomics*, **14**, 465.
31. Moreton, J., Izquierdo, A. and Emes, R. D. 2016, Assembly, assessment, and availability of *de novo* generated eukaryotic transcriptomes, *Front. Genet.*, **6**, 1–9.
32. Parra, G., Bradnam, K. and Korf, I. 2007, CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes, *Bioinformatics*, **23**, 1061–7.
33. Karpinka, J. B., Fortriede, J. D., Burns, K. A., et al. 2015, Xenbase, the Xenopusmodel organism database; new virtualized system, data types and genomes, *Nucleic Acids Res.*, **43**, D756–D63.
34. Irisarri, I., Baurain, D., Brinkmann, H., et al. 2017, Phylotranscriptomic consolidation of the jawed vertebrate timetree, *Nat. Ecol. Evol.*, **1**, 1370–8.
35. Altschul, S. F., Gish, W., Miller, W. T., Myers, E. W. and Lipman, D. J. 1990, Basic local alignment search tool, *J. Mol. Biol.*, **215**, 403–10.
36. Powell, S., Forslund, K., Szklarczyk, D., et al. 2014, EggNOG v4.0: nested orthology inference across 3686 organisms, *Nucleic Acids Res.*, **42**, D231–D9.

37. Letunic, I. and Bork, P. 2007, Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation, *Bioinformatics*, **23**, 127–8.
38. Fu, L., Niu, B., Zhu, Z., Wu, S. and Li, W. 2012, CD-HIT: accelerated for clustering the next-generation sequencing data, *Bioinformatics*, **28**, 3150–2.
39. Nowoshilow, S., Schloissnig, S., Fei, J. F., et al. 2018, The axolotl genome and the evolution of key tissue formation regulators, *Nature*, **554**, 50–5.
40. Sun, Y. B., Xiong, Z. J., Xiang, X. Y., et al. 2015, Whole-genome sequence of the Tibetan frog *Nanorana parkeri* and the comparative evolution of tetrapod genomes, *Proc. Natl. Acad. Sci. USA*, **112**, E1257–E62.
41. Hammond, S. A., Warren, R. L., Vandervalk, B. P., et al. 2017, The North American bullfrog draft genome provides insight into hormonal regulation of long noncoding RNA, *Nat. Commun.*, **8**, 1433.
42. R Development Core Team. 2016, *R: A Language and Environment for Statistical Computing*. R Found Stat Comput, Vienna, Austria.
43. Szklarczyk, D., Franceschini, A., Wyder, S., et al. 2015, STRING v10: protein–protein interaction networks, integrated over the tree of life, *Nucleic Acids Res.*, **43**, D447–D52.
44. Finn, R. D., Coggill, P., Eberhardt, R. Y., et al. 2016, The Pfam protein families database: towards a more sustainable future, *Nucleic Acids Res.*, **44**, D279–D85.
45. Sean, R. E. 2010, HMMER: biosequence analysis using profile hidden Markov models. Available at: http://hmmer.janelia.org/
46. Emms, D. M. and Kelly, S. 2015, OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy, *Genome Biol.*, **16**, 157.
47. Katoh, K. and Standley, D. M. 2013, MAFFT multiple sequence alignment software version 7: improvements in performance and usability, *Mol. Biol. Evol.*, **30**, 772–80.
48. Price, M. N., Dehal, P. S. and Arkin, A. P. 2009, Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix, *Mol. Biol. Evol.*, **26**, 1641–50.
49. Le, S. Q., Lartillot, N. and Gascuel, O. 2008, Phylogenetic mixture models for proteins, *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **363**, 3965–76.
50. Mirarab, S., Reaz, R., Bayzid, M. S., Zimmermann, T., Swenson, M. S. and Warnow, T. 2014, ASTRAL: genome-scale coalescent-based species tree estimation, *Bioinformatics*, **30**, i541–i8.
51. Sayyari, E. and Mirarab, S. 2016, Fast coalescent-based computation of local branch support from quartet frequencies, *Mol. Biol. Evol.*, **33**, 1654–68.
52. Mercer, T. R., Dinger, M. E. and Mattick, J. S. 2009, Long non-coding RNAs: insights into functions, *Nat. Rev. Genet.*, **10**, 155–9.
53. Prachumwat, A. and Li, W. H. 2008, Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes, *Genome Res.*, **18**, 221–32.
54. Albalat, R. and Cañestro, C. 2016, Evolution by gene loss, *Nat. Rev. Genet.*, **17**, 379–91.
55. Eckes, B., Krieg, T. and Niessen, C. M. 2010, Biology of the skin, In: Krieg, T., Bickers, D.R. and Miyachi, Y. (Eds.) *Therapy of Skin Diseases: A Worldwide Perspective on Therapeutic Approaches and Their Molecular Basis*, Springer-Verlag: Berlin, Heidelberg, pp. 3–14.
56. Duellman, W. E. and Trueb, L. 1994, *Biology of Amphibians*, Johns Hopkins University Press, Baltimore, MD, USA.
57. Clarke, B. T. 1997, The natural history of amphibian skin secretions, their normal functioning and potential medical applications, *Biol. Rev. Camb. Philos. Soc.*, **72**, 365–79.
58. Toledo, R. C. and Jared, C. 1995, Cutaneous granular glands and amphibian venoms, *Comp. Biochem. Physiol. A: Physiol.*, **111**, 1–29.
59. Chen, T., Farragher, S., Bjourson, A. J., Orr, D. F., Rao, P. and Shaw, C. 2003, Granular gland transcriptomes in stimulated amphibian skin secretions, *Biochem. J.*, **371**, 125–30.
60. Lazarus, L. H. and Attila, M. 1993, The toad, ugly and venomous, wears yet a precious jewel in his skin, *Prog. Neurobiol.*, **41**, 473–507.
61. Roelants, K., Fry, B. G., Norman, J. A., Clynen, E., Schoofs, L. and Bossuyt, F. 2010, Identical skin toxins by convergent molecular adaptation in frogs, *Curr. Biol.*, **20**, 125–30.
62. Huang, L., Li, J., Anboukaria, H., Luo, Z., Zhao, M. and Wu, H. 2016, Comparative transcriptome analyses of seven anurans reveal functions and adaptations of amphibian skin, *Sci. Rep.*, **6**, 24069.
63. Meng, P., Yang, S., Shen, C., Jiang, K., Rong, M. and Lai, R. 2013, The first salamander defensin antimicrobial peptide, *PLoS One*, **8**, e83044.
64. Kupfer, A., Müller, H., Antoniazzi, M. M., et al. 2006, Parental investment by skin feeding in a caecilian amphibian, *Nature*, **440**, 926–9.
65. Wilkinson, M., Kupfer, A., Marques-Porto, R., Jeffkins, H., Antoniazzi, M. M. and Jared, C. 2008, One hundred million years of skin feeding? Extended parental care in a *Neotropical caecilian* (Amphibia: Gymnophiona), *Biol. Lett.*, **4**, 358–61.
66. Taylor, E. H. 1972, Squamation in caecilians, with an atlas of scales, *Univ. Kansas Sci. Bull.*, **49**, 989–164.
67. Gower, D. J. and Wilkinson, M. 2002, Phallus morphology in caecilians (Amphibia, Gymnophiona) and its systematic utility, *Bull. Nat. Hist. Museum Zool. Ser.*, **68**, 143–54.
68. Gomes, A. D., Moreira, R. G., Navas, C. A., Antoniazzi, M. M. and Jared, C. 2012, Review of the reproductive biology of caecilians (Amphibia, Gymnophiona), *South Am. J. Herpetol.*, **7**, 191–202.
69. Gabaldón, T. and Koonin, E. V. 2013, Functional and evolutionary implications of gene orthology, *Nat. Rev. Genet.*, **14**, 360–6.
70. Gower, D. J. and Wilkinson, M. 2005, Conservation biology of caecilian amphibians, *Conserv. Biol.*, **19**, 45–55.
71. Genome 10K Community of Scientists. 2009, Genome 10K: a proposal to obtain whole-genome sequence for 10000 vertebrate species, *J. Hered.*, **100**, 659–74.
72. Koepfli, K.-P., Paten, B. and O'Brien, S. J. and O'Brien S. J. 2015, The Genome 10K Project: a way forward, *Annu. Rev. Anim. Biosci.*, **3**, 57–111.
73. Smith, J. J., Putta, S., Walker, J. A., et al. 2005, Sal-Site: integrating new and existing ambystomatid salamander research and informational resources, *BMC Genomics*, **6**, 181.
74. Shaffer, H. B., Gidiş, M., McCartney-Melstad, E., et al. 2015, Conservation genetics and genomics of amphibians and reptiles, *Annu. Rev. Anim. Biosci.*, **3**, 113–38.