UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS BIOLÓGICAS

DEPARTAMENTO DE BIODIVERSIDAD, ECOLOGÍA Y EVOLUCIÓN

TESIS DOCTORAL

Comparative transcriptomics and gene discovery in caecilian amphibians

Transcriptómica comparada y descubrimiento de genes en cecilias

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

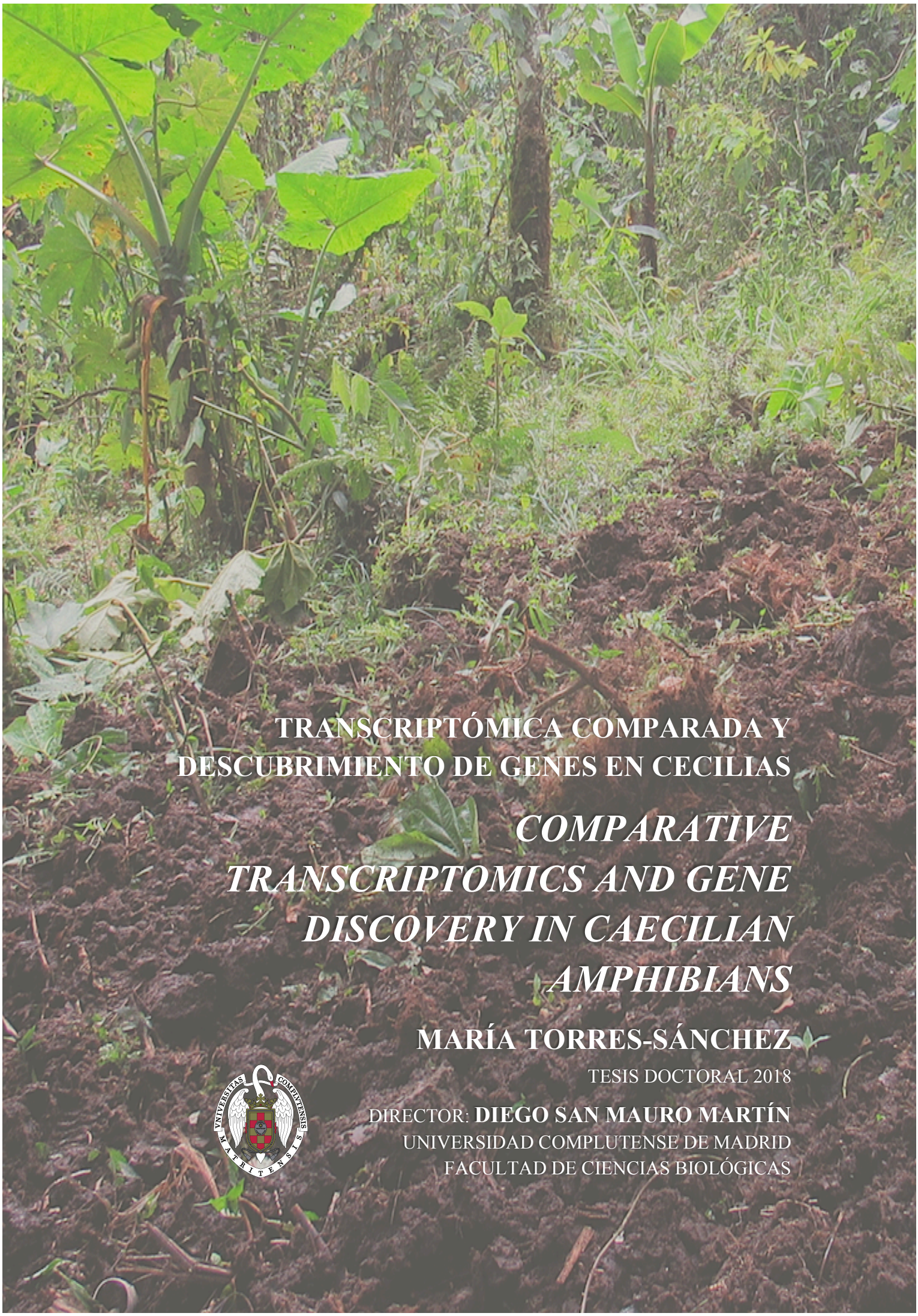
PRESENTADA POR

María Torres-Sánchez

DIRECTOR

Diego San Mauro Martín

Madrid, 2018



TRANSCRIPTÓMICA COMPARADA Y DESCUBRIMIENTO DE GENES EN CECILIAS

COMPARATIVE TRANSCRIPTOMICS AND GENE DISCOVERY IN CAECILIAN AMPHIBIANS

MARÍA TORRES-SÁNCHEZ

TESIS DOCTORAL 2018

DIRECTOR: **DIEGO SAN MAURO MARTÍN**
UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE CIENCIAS BIOLÓGICAS

UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS BIOLÓGICAS
DEPARTAMENTO DE BIODIVERSIDAD, ECOLOGÍA Y EVOLUCIÓN

TRANSCRIPTÓMICA COMPARADA Y DESCUBRIMIENTO DE GENES EN CECILIAS

COMPARATIVE TRANSCRIPTOMICS AND GENE DISCOVERY IN CAECILIAN AMPHIBIANS

TESIS DOCTORAL DE:

MARÍA TORRES-SÁNCHEZ

DIRECTOR:
DR. DIEGO SAN MAURO MARTÍN

MADRID, 2018

Cover Front

Field journey 2017, Dracula Reserve, Carchi, Ecuador

Chapter illustration

Drawing of caecilian gene expression

UNIVERSIDAD
COMPLUTENSE
MADRID

FACULTAD DE CIENCIAS BIOLÓGICAS
DEPARTAMENTO DE BIODIVERSIDAD, ECOLOGÍA Y EVOLUCIÓN

TRANSCRIPTÓMICA COMPARADA Y DESCUBRIMIENTO DE GENES EN CECILIAS

COMPARATIVE TRANSCRIPTOMICS AND GENE DISCOVERY IN CAECILIAN AMPHIBIANS

Memoria presentada por María Torres-Sánchez para optar al grado de Doctor en Biología bajo la dirección del Doctor Diego San Mauro Martín de la Universidad Complutense de Madrid

Madrid, 2018

La doctoranda

María Torres Sánchez

V^{o} B^{o} del director

Dr. Diego San Mauro Martín

La presente Tesis Doctoral ha sido financiada por el Ministerio de Economía y Competitividad (MINECO) mediante el proyecto de investigación CGL2012-40082, la beca predoctoral asociada al mismo BES-2013-062723 y las estancias breves de investigación EEBB-I-15-09665, EEBB-I-16-11395 y EEBB-I-17-12039.

To my grandparents, Generosa and Jesús

"…knowledge of sequences could contribute much to our understanding of living…"
- Frederick Sanger

Acknowledgements

Without all of you (you know who you are), it would not have been possible to take the first steps of this journey; journey that has only begun and in which I hope to have your companionship for many years to come.

I wish firstly to thank my supervisor for his continued trust and support. Thank you for opening the doors of the scientific world to me. I would like to continue thanking to all my research stay supervisors and collaborators, especially to Chris, Mark and Dave for being my giants and let me standing on their shoulders. I extend my gratitude to all the institutions that hosted me during these years.

Words would be little to express my feeling and gratitude for all those who contributed in many ways to the completion of this academic stage, since the early start of my education. Thanks for have taught me to be a scientist. Hours with you in classroom, field, congresses, courses, lab and office were remarkable moments. I am grateful for all our enlightening conversations, and debates about biological questions and future projects. Explore and share this world with you is being the most exciting experience.

Beyond Biology, many of you are forever part of my other family and I am really grateful for sharing with you this adventure. Best ideas happen in leisure time, it is always great to share with you (my friends), messages, talks, teas, beers, meals, walks, wildlife watching and life in general.

Last but not least, I am indebted to my family for their unconditional love and encouragement. Thank you for always fostering my curiosity.

Finally, I am more than grateful to Iván, for his advice, patience and confidence. Your passion in science and life is a truly source of inspiration.

In the very end, I cannot forget to thank to my incredible studied animals. Caecilians, you are breaking my heart!

INDEX

Abstract

Many aspects of biological diversity and their life mechanisms remain unknown and understudied. With the advent of the genomic era, RNA sequencing (RNA-seq) has become one of the most powerful tools to unravel the secrets of biological adaptation and diversity in all species through their particular gene expression profiles. Here, we studied comparatively the genes expressed in different tissues of several species of one of the least known group of vertebrates, the caecilians (order Gymnophiona). Caecilians are fossorial, limbless, tropical amphibians that constitute the sister group of frogs and salamanders. Little is known about this enigmatic animal group. To improve the understanding of caecilian ecology and evolution, we have analysed caecilian genomic functional elements at three levels: across other vertebrates, across caecilian species and among caecilian tissue types. Our study provides valuable insights about the expansion of gene machineries in vertebrates, points out protein-coding genes involved in the specific evolutionary adaptations of caecilian amphibians, and highlights important functional elements in the caecilian skin tissue type. To our knowledge, this is the first large-scale genomic characterization of the genetic functional elements of this secretive vertebrate group, and it provides the basis for future research on the molecular elements underlying the remarkable biology of caecilian amphibians.

Muchos aspectos de la biodiversidad continúan siendo todo un misterio, bien porque todavía no se han llegado a comprender a pesar de los esfuerzos de estudio, o bien porque se encuentran poco estudiados. Con el auge de las tecnologías de secuenciación masiva y herramientas bioinformáticas asociadas, se ha abierto una oportunidad sin precedentes en investigación biológica. De entre las metodologías de secuenciación masiva destaca la secuenciación del ARN que ha irrumpido como una de las metodologías con mayor potencialidad para desentrañar, mediante el estudio de perfiles de expresión génica, los secretos del funcionamiento y adaptación de las especies. En esta tesis hemos analizado la expresión génica de diferentes tejidos de varias especies de uno de los grupos de vertebrados más desconocido, las cecilias, orden Gymnophiona (ápodos). Las cecilias forman junto a ranas y salamandras la clase de tetrápodos Amphibia. La falta de información sobre este orden de anfibios está vinculada con la confusión inicial con otros animales morfológicamente similares, su distribución tropical y principalmente su particular modo de vida fosorial, el cual ha dificultado su estudio mediante metodologías zoológicas clásicas. En aras de ampliar el conocimiento acerca de la ecología y evolución de las cecilias, hemos comparado los genes codificantes expresados en las muestras de las cecilias a tres niveles: entre vertebrados, entre especies de cecilias y entre los distintos tejidos de trabajo. Nuestros estudios apuntan hacia la expansión de las familias génicas de vertebrados en cecilias, aportan información sobre innovaciones moleculares a lo largo de la evolución de estos anfibios y señalan varios genes con importantes funciones en su piel. Esta tesis es el primer estudio de caracterización genómica en cecilias y sienta la bases para futuras investigaciones explorando los elementos genómicos que se esconden detrás de la biología de estos enigmáticos anfibios.

Denantes os incribeis avances en prol da comprensión da vida, a biodiversidade segue a agochar enchentes misterios. O florecemento de tecnoloxías de secuenciación masivas e ferramentas bioinformáticas asociadas abren unha cantidade inimaxinable de diferentes abordaxes de estudo en todas as disciplinas científicas da Bioloxía. Unha delas é o uso da secuenciación do ARN que a través da análise da expresión xénica ten a potencialidade de revelar as bases moleculares do funcionamento e adaptacións das especies. Nesta tese de doutoramento exploramos os patróns de expresión xénica de tecidos distintos de varias especies dun dos grupos animais máis descoñecido, as cecilias (orden Gymnophiona). Xunto a ras, sapos, limpafontes e píntegas, as cecilias forman a clase de tetrápodos Amphibia. As cecilias son anfibios, ápodos, vermiformes, fosoriais que atópanse exclusivamente nos trópicos. Co obxectivo de ampliar o coñecemento ecolóxico e evolutivo neste misterioso grupo, levamos a cabo análises comparativas a tres niveis: entre vertebrados, entre as especies de cecilias de estudo e entre os seus tecidos. Os nosos traballos amosan unha expansión do número de familias xénicas en vertebrados, sinalan innovacións moleculares vinculadas aos diferentes fitos na evolución das cecilias e identifican xenes con importantes funcións na pel destes anfibios. Esta tese conforma a primeira caracterización xenómica das cecilias e presenta as bases moleculares para futuras investigacións, expandindo o coñecemento sobre vertebrados e especificamente sobre as cecilias.

Synthesis

Background and comprehensive introduction

Almost half a century has gone by since the first genome, heredity information of the species, was sequenced. Sanger and collaborators sequenced the bacteriophage φX174 in 1977 (1). Since then, DNA sequencing (elucidation of the order of nucleic acids in polynucleotide chains) has undergone, and still does, important technological changes and improvements mainly towards increasing the amount of data and reducing time and cost of sequencing, and giving rise to high-throughput sequencing (HTS) technologies (2). In parallel, a good amount of bioinformatics tools has been developed as well. Once again, it is an era of exploration, discovery, collection and catalogue but this time for millions of unanalysed nucleotide sequences.

The study of genomic data of any living organism has become a major focus of biology presenting a world of research opportunities, particularly promising in evolutionary and ecological fields. Different experiments and strategies of sequencing can be followed to address different research questions. Among these methodologies stand out massively parallel complementary DNA (cDNA) sequencing or RNA sequencing (RNA-Seq). RNA-Seq is the state-of-the-art transcriptomic methodology where the whole amount of transcripts (RNAs) from a sample is isolated and sequenced by HTS technologies and analysed with several bioinformatics tools (3,4). Mainly, we will refer to one type of RNAs, the messenger RNAs (mRNAs) that in a simple way represents the molecular bridge between DNA and proteins, and are considered genomic functional elements. RNA-seq is providing significant increase in knowledge of quantitative and qualitative analyses of the biology of transcripts, not only for model species but also for non-model species or species with absence of a reference sequenced genome. This method covers a wide variety of applications, from multiple gene expression and regulation studies through phylogenomic analyses and right up to single nucleotide polymorphism (SNP) identification. Even though there are challenging analytical steps, RNA-Seq is presenting itself as a very powerful tool with the potential to improve understanding about biological diversity and to unravel different chapters of the genomic book of the species, the book that encodes life. Knowledge about life of the different species that inhabit Earth has always been biased (5). The bias in the study of biodiversity affects mostly to species that have

secretive habits and lifestyle, and therefore can also be difficult to study in the field. These less well-known species can receive a major benefit from the use of HTS technologies and especially RNA-Seq to improve the understanding of their life history.

Caecilian amphibians (order Gymnophiona) are among such least known major tetrapod lineages. In 1735, Albertus Seba first described caecilians in his work entitled Thesaurus and misclassified them as snakes. Caecilians are one of the extant orders of the class Amphibia, along with frogs and salamanders that present exclusively tropical distribution (6). This enigmatic group is characterised by being fossorial animals with elongate limbless bodies, reduced visual and hearing systems and with sensory tentacles close to the snout. At odds with their order name (Gymnophiona = "naked snakes"), several species present fish-like scales in dermal pockets of their annulated bodies. The large majority have terrestrial adult life and are found in moistly soils, but the species of one family (Typhlonectidae) are fully aquatic or semi-aquatic and among them there is found the largest lungless amphibian, (7). Like other amphibian groups, caecilians have a huge variety of reproductive strategies, presenting internal fecundation mediated by the elongation and differentiation of the external part of the gut in males. Different modes of parental care have been observed including maternal skin feeding. While being an ancient, specialized group with at least 250 million years of independent evolution from the other extant amphibian orders, there are only 206 recognized species thus far. Despite their biological interest, caecilian amphibians have been neglected in many research disciplines including genome-wide characterizations.

In this project, we have sequenced by RNA-Seq the transcripts of several tissue samples form five different species of caecilians (*Rhinatrema bivittatum* Cuvier in Guérrin-Méneville, 1838, *Caecilia tentaculata* Linnaeus, 1758, *Typhlonectes compressicauda* Duméril & Bibron, 1841, *Microcaecilia unicolor* Duméril, 1861, and *Microcaecilia dermatophaga* Wilkinson, Sherratt, Starace & Gower, 2013) in order to study and compare caecilian genomic functional elements at three levels: across other vertebrates, across the five caecilian species and among caecilian tissues types.

Aims

To obtain genomic references for caecilian amphibians from newly generated transcriptomic data (chapter 1).

To study genomic functional elements in the caecilian transcriptomes by comparison with other vertebrates (chapter 1).

To identify key genomic functional elements in the evolution of caecilian amphibians (chapter 2).

To characterise the particular genomic functional elements in the caecilian amphibian skin tissue type (chapter 3).

General view of materials and methods

Source material of this RNA-Seq project was tissue samples of 9 different tissue types (skin, foregut, muscle, liver, kidney, lung, heart, spleen, and testis) from 7 specimens of 5 different species of caecilian amphibians from French Guiana. Species were chosen to represent 4 of the 10 currently caecilian amphibian families (2 specimens of *R. bivittatum* from Rhinatrematidae, 1 specimen of *C. tentaculta* from Caeciliidae, 1 specimen of *T.compressicauda* from Typhlonectidae, and 2 specimens of *M.unicolor* and 1 specimen of *M.dermatophaga* from Siphonopidae). These species represent degrees of evolutionary divergence and a range of different ecologies being *T.compressicauda* aquatic, *R. bivittatum* shallow terrestrial and the other three species found in deeper layers of the soil (6).

RNA-Seq experimental steps could divide in pre-sequencing steps and post-sequencing steps linked to wet and dry laboratories respectively (4). The main pre-sequencing steps are the sample acquisition, RNA extraction and quantity-quality control. After the acquisition of the tissue samples, we carried out more than one hundred of RNA extractions from different samples, checked the amount of extracted RNA and tested the degree of degradation of these RNA molecules using as indicator the integrity of the ribosomal RNAs (Figure 1). A total of 40 samples were selected to sequence on Illumina Hiseq2000 platform with previous selection of poly-A (tail of nucleotides with repeats of the base adenine) transcripts (poly-A tail is characteristic of all mRNAs and other populations of RNAs). Post-sequencing steps or bioinformatics analyses mainly comprise quality control, raw sequences (reads) processing, *de novo* assembly (for samples from species with absence of a sequenced genome as our case), read alignment and quantification, annotation, differential expression analysis and other comparative sequence analyses (for detail information of pre-sequencing steps see subsection sample preparation and high-throughput sequencing of the materials and methods of chapter 1; detailed information about the particular post-sequencing steps used in the different analyses could find in material and methods sections of all chapters).

Overall results and discussion

As a subset of the genome, putative protein-coding genes of the five studied species of caecilian amphibians were identified (chapter 1) providing some of them valuable insights about caecilian ecology and evolution (chapter 1, 2 and 3). Caecilians contain, as might be expected, a set of genes that are shared with other vertebrate species, members of known vertebrate gene families, showing no bias towards any vertebrate group, and being most of them expressed in all the sampled tissues (chapter 1). Some of these genes were used for reconstructing the evolutionary history of the analysed species as well as for testing molecular evolution along the phylogenetic branches of the studied caecilian species identifying caecilian functional elements under positive selection (chapter 1 and chapter 2 respectively). Genes with tissue-specific expression were counted in lesser proportion as members of known vertebrate gene families. Genes with no assignment to known vertebrate gene families were grouped and designated as potential novel gene families. The study between the two types of gene families pointed out a probable expansion of the genetic machinery of caecilian genes with skin tissue-specific expression (chapter 1). Profound study of skin expression led to highlight several skin specialisations (chapter 3).

All our results rely on the assembly and annotations of the sequenced samples being both controversial steps (8–10). Original RNAs of the samples were fragmented before sequencing due to technological restrictions and consequently RNA molecules must be reconstructed afterwards. Even with sequencing and bioinformatics methodologies attempting to guarantee the accuracy of the assembly of the sequenced fragments, some assembly errors are possible and chimeric molecules could be rebuilt. Automatic annotations of the reconstructed molecules are mainly carried out by sequence similarity searches and depend on the information used (databases), driving in some cases to misidentification and inaccurate functional assignments because similar sequence does not necessarily imply similar function. Large-scale molecular studies like ours provide valuable information about the genomic mechanisms of particular species and they establish a research framework for undertaking deeper analyses base on specific elements or experimental designs.

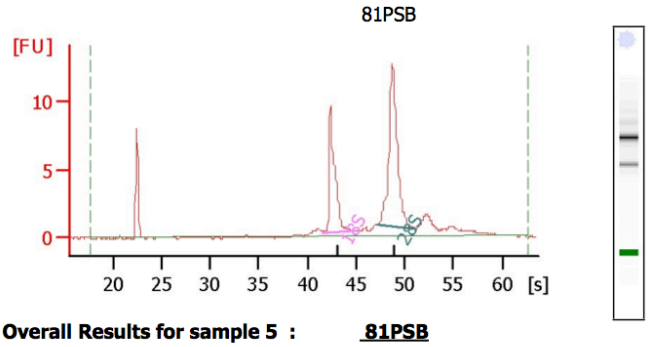
References

1. Sanger F., Air G. M., Barrell B.G., Brown N. L., Coulson A. R., Fiddes C. A., Hutchison C. A., Slocombe P. M. and Smith M. 1977. Nucleotide sequence of bacteriophage phi X174 DNA. Nature. 265: 687–95.
2. Heather J. M. and Chain B. 2016. The sequence of sequencers: The history of sequencing DNA. Genomics. 107: 1–8.
3. Wang Z., Gerstein M. and Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10: 57–63.
4. Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szczesniak M. W., Gaffney D. J., Elo L. L., Zhang X. and Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17: 13.
5. Troudet J., Grandcolas P., Blin A., Vignes-Lebbe R. and Legendre F. 2017. Taxonomic bias in biodiversity data and societal preferences. Sci Rep. 7: 9132.
6. Wilkinson M. 2012. Caecilians. Curr Biol. 22(17).
7. Nussbaum R. A. and Wilkinson M. 1995. A New Genus of Lungless Tetrapod: A Radically Divergent Caecilian (Amphibia: Gymnophiona). Proc R Soc B Biol Sci. 261: 331–5.
8. Paszkiewicz K. and Studholme D. J. 2010. De novo assembly of short sequence reads. Brief Bioinform. 11(5): 457–72.
9. Loewenstein Y., Raimondo D., Redfern O. C., Watson J., Frishman D., Linial M., Orengo C., Thornton J. and Tramontano A. 2009. Protein function annotation by homology-based inference. Genome Biol. 10(2): 207.
10. Brylinski M. and Skolnick J. 2010. Comparison of structure-based and threading-based approaches to protein functional annotation. Proteins Struct Funct Bioinforma. 78(1): 118–34.

Tables and Figures

Figure 1

Electropherogram from the quality-quantity analysis of eukaryote total RNA of one of our caecilian samples (posterior skin form *C. tentaculta*) on BioAnalyzer.



Overall Results for sample 5 : 81PSB

RNA Area:	60,8
RNA Concentration:	50 ng/µl
rRNA Ratio [28s / 18s]:	1,7
RNA Integrity Number (RIN):	10 (B.02.07)
Result Flagging Color:	
Result Flagging Label:	RIN:10

Fragment table for sample 5 : 81PSB

Name	Start Time [s]	End Time [s]	Area	% of total Area
18S	41,55	44,73	14,4	23,7
28S	47,27	50,91	24,1	39,6

Chapter 1

Transcriptomic landscapes indicate expansion of vertebrate gene families in caecilian amphibians

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

Abstract

RNA sequencing (RNA-seq) has become one of the most powerful tools to unravel the genomic basis of biological adaptation and diversity in all organisms. 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 characterise multi-tissue transcriptomes for five species of caecilians amphibians that represent a broad level of diversity across the order. We identified vertebrate homologous elements of caecilian functional genes of varying tissue specificity that indicate a great expansion of known vertebrate gene families in caecilians, especially for the skin. A supertree analysis of phylogenomic data containing 1,955 single-copy orthologous genes recovered phylogenetic relationships among the five caecilians and other major lineages of vertebrates in agreement with current vertebrate 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.

Introduction

High-throughput sequencing (HTS) technologies and associated bioinformatics are transforming the study of evolutionary and comparative genetics, offering an unprecedented opportunity to characterise 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 comprehensively analyse RNA molecules, the transcriptome, on a massively parallel scale (4,5). The transcriptome is a snapshot in time of the set of genes expressed 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 RNA data from a wide range of tissues (from different individuals and/or stages of development) in order 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 recognised families (Caeciliidae, Rhinatrematidae, Siphonopidae, and Typhlonectidae) and a range of ecologies and degrees of evolutionary divergence (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 206 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 characterisation of caecilian genomes using transcriptomic landscapes generated with RNA-seq. We use two complementary approaches to investigate features of caecilian protein-coding gene 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 indicates a great expansion of known vertebrate gene families in caecilians, differentially across tissue types. Second, we infer the phylogenetic relationships of the five sampled caecilian species and the same set of 51 vertebrates based on orthologous genes. This study provides new information about the functional elements of the genome and phylogenomics of caecilians, as well as protein-coding gene evolution in vertebrates.

Materials and methods

Sample preparation and high-throughput sequencing

This study includes data from five caecilian species: *Rhinatrema bivittatum* Cuvier in 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 maintained specimens anesthetized with tricaine methanesulfonate (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 4ºC overnight (to allow the solution to thoroughly penetrate the tissue) and stored at -20ºC. Numbers of specimens and of tissues sampled per species and voucher information are given in Table 1 and Supplementary Table S1.

RNA was isolated using the RNeasy Fibrous Tissue Mini Kit (Qiagen) following the manufacturer's instructions, and performing 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 selected 40 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 1 for details). 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.

Raw data processing and de novo assembly

RNA-seq raw reads 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 1). 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) in order to explore parameter settings and run times.

De novo species-specific assemblies were performed with Trinity r20140717 (25) using default settings with 60 Gb of RAM (--max_memory 60G) with a prior *in silico* normalization done using Trinity (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 of completeness of each caecilian 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).

Multigene family analysis

Contigs of the five species-specific caecilian transcriptomes containing ORFs were aligned against predefined vertebrate-specific gene families (veNOGs) from the EggNOG 4.1 database (34) using BLAST, blastp tool version 2.2.28, (35) 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. We classified all caecilian annotations 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 (and at least 90 reads to other tissues). 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 homologs of the caecilian protein-coding genes on the vertebrate taxonomy tree from the NCBI taxonomy database was generated and visualised using phyloT and ITOL (36) respectively. The vertebrate taxonomy tree was built using the taxonomic Ids of the species that are included in the EggNOG database.

Where possible, caecilian gene families were annotated with the same function as the vertebrate gene families with the best BLAST match in EggNOG identified above. Transcripts with no hits to the known vertebrate gene families in EggNOG were clustered using CD-HIT 4.6.4 (37) with a 90% identity threshold and classified as putative novel caecilian gene families. Of these, we calculated the number of tissues in which any gene family was expressed (as described earlier). Additionally, to characterise the different tissues in a more restrictive approach than simple tissue presence classification, tissue specificity was postulated when 95% of total read counts belonged to a single tissue for both unknown gene family and veNOG gene family annotated contigs. To test if there was a greater number of tissue-specific novel genes than expected by chance, the relative abundance of known vertebrate gene families versus those of putative novel caecilian gene families were compared using a two-tailed Fisher's exact test conducted with R 3.3.0 (38), with the null hypothesis that there was no difference in the number of tissue-specific novel genes. Finally, our characterisation of the tissue specificity expression was completed with

the inference of protein-protein interactions (PPIs) and functional enrichment paths using STRING (39) with the option of auto-detect organism for the known vertebrate gene families; and the Pfam (40) annotation of the putative novel caecilian gene families using HMMER 3.0 (41) with default parameters.

Orthology prediction and phylogenomic analysis

To carry out a phylogenomic analysis we needed to identify single-copy genes from across the vertebrates, including our caecilian samples. To do this we used OrthoFinder 0.2.8 (42) 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 this analysis we filtered out any orthologous groups (orthogroups) that were not in single-copy. Multiple-sequence alignments were performed individually for each of the resulting single-copy orthogroups using MAFFT 7.245 (43) with default settings, and individual gene trees were inferred using approximately-maximum-likelihood with FastTree 2.1.8 (44) and the JTT+CAT model of amino acid substitutions (45). We reconstructed a supertree using ASTRAL 4.10.11 which provides statistically consistent species tree inference from gene trees subject to incomplete lineage sorting (46, 47), and computed posterior probabilities and quartet support for the internal branches of the main recovered topology.

Results

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 1263–1884 (Supplementary Table S2). 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 Figure 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 (Table 1).

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).

Vertebrate multigene family analysis

Annotated caecilian genes that are homologous also with those for vertebrates in the EggNOG database are expressed in most of the (up to nine) sampled caecilian tissue types, with only a small proportion being tissue specific. This pattern is very similar when comparing the pooled caecilian sample (all five species) with each of the 51 EggNOG database vertebrates, with no obvious phylogenetic pattern (Figure 1). The number of caecilian contigs having 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.

A total of 177 known vertebrate and 493 novel caecilian gene families exhibit tissue-specific expression (Table 2). A significantly greater number of novel caecilian genes were expressed only in skin. In contrast, caecilian spleen transcripts had significantly lower than expected tissue-specific novel gene families. Among the tissue-specific known vertebrate gene families, we found significant PPIs and functional enrichment paths for some caecilian tissues (foregut, kidney, liver, spleen and testis, see Supplementary Table S3). Additionally, 143 different protein domains were identified in the tissue-specific novel caecilian gene families (Supplementary Table S4), including 15 structural and functional domains occurring exclusively in the skin, these including some diverse proteases, lipoprotein and amino acid storage receptors, and toxin-like domains.

Orthology prediction and phylogenomic analysis

We obtained a total of 23,761 orthologous groups or clusters, of which 1,955 were single-copy orthogroups comprising genes from at least four vertebrate taxa. The number of single-copy 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 then built from the gene trees under a multi-species coalescent model, maximising the number of induced quartet trees (the Supertree is presented in Supplementary Figure 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 orthologous groups recovered the main known topology of this subset of the Tree of Life (Supplementary Figure 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) molecular analyses (10,48).

Discussion

Transcriptomic landscapes reveal massive gene family expansion in caecilians

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). 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 protein annotation, between 57% and 77%, was obtained in the veNOG database of EggNOG, 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 numbers of protein-coding genes could be overestimated. An additional problem is that the transcriptomes are not solely composed of protein-coding genes. Recently, it has been demonstrated that almost the entire genome is transcribed (49). Accordingly, caecilian contigs that are not protein-coding genes are postulated to be long non-coding RNAs and potentially important regulatory elements.

In order to investigate and quantify the importance of the new genomic records for caecilians, we grouped the protein-coding sequences into multigene families. If caecilians did not have novel genes, it would be expected that the vast majority of their genes would belong to some already described, known vertebrate gene family. However, our results indicate that less than half of the caecilian gene families belong to known vertebrate gene families, indicating that caecilians have likely undergone massive gene family expansion. Given the sparse taxon sampling and the fact that some of theses genes that do not belong to a known vertebrate family are annotated, most of them as homologs of X. tropicalis, at least some of these gene families could

be amphibian rather than caecilian specific. The absence of homologs of these caecilian gene families in the other vertebrate species from the EggNOG database might reflect gene loss events (50,51), or alternatively faster sequence evolution in some caecilian genes. Either way, caecilians likely have many functional elements unknown in other vertebrates.

Greater tissue-specific gene innovation in caecilian skin

The analysis of skin-specific gene families of caecilians demonstrates that vertebrate skin gene families remain poorly characterized.. We detected no significant PPIs or functional enrichment pathways in caecilian skin from the description of the known vertebrate gene families associated with this tissue type, which could mean that these genes are not well known and have unknown, innovative functions and interactions. This is different to other caecilian tissue types such as foregut, where we found PPIs and enrichment in functional elements related to nutrient absorption (GO:0007586, see Supplementary Table S3). On the other hand, the novel caecilian gene families expressed in skin were annotated with protein domains, and these putative novel gene families could be associated causally with specializations of caecilian skin (52). Skin tissue forms the barrier between the organism and the environment both physically and at the (bio)chemical level. It is genetically and physiologically very active throughout an animal's life. Amphibian skin is multifunctional with additional roles in respiration and water regulation, and in defence against predators and pathogens (53,54). The defensive properties of amphibian skin rely mainly on biochemical substances secreted from specialized skin granular glands (55,56). These secretions can contain numerous bioactive components, including alkaloids, biogenic amines, peptides, and proteins (57), some of which have been isolated and studied, particularly in anurans (frogs and toads) and salamanders (58–60). The diversity of functions and biochemical activity 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 frogs and salamanders (15–19) and the sustained contact between the skin and soil for most caecilian species. Indeed, some of the protein domains found exclusively in the skin-specific novel gene families, such as proteases and toxin-like domains (Asp_protease_2, gag-asp_proteas,

Toxin_TOLIP, Trypsin, UPAR_LY6, see Supplementary Table S4) point to novel caecilian skin defensive mechanisms.

The maternal skin of many caecilian species plays another role: in provision of nutrition to newborns (maternal dermatophagy; 61,62). This behavior is present in several of the species sampled in this study (observed in Microcaecilia dermatophaga, likely also present in Microcaecilia unicolor and Caecilia tentaculata; 10). This phenomenon is especially interesting for understanding the evolution of viviparity because it is possibly a precursor of the oviduct feeding by foetuses that occurs in viviparous caecilians (62). Maternal dermatophagy involves structural and histochemical changes in the mothers' epidermis, it becomes hypertrophied and heavily invested with lipids (61), and hence expanded gene machinery is likely needed. Lipoprotein receptor and amino acid storage receptor (Ldl_recept_a, PhaP_Bmeg, see Supplementary Table S4) are other protein domains 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 (63). Scales are absent in other extant amphibians but are present, concealed in dermal pockets, in many caecilians (all except Typhlonectes compressicauda of those sampled in our study).

Further data and analyses are required to identify the taxonomic distribution, diversity and function of these putative skin-specific genes. 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. For example, caecilians differ from other amphibians in that males have a copulatory organ formed from the eversible final part of the gut (64), as well as other autapomorphies of the sperm and internal fertilization specialisations such as the Müllerian gland and the ejaculate (65). All this may be reflected in their genomes.

Phylogenomic utility of orthogroups derived from caecilian transcriptomes

Unlike multigene families containing both orthologous and paralogous genes, exclusively orthologous groups of genes are more straightforward for use in phylogenomics and the study of evolutionary processes that depend upon inferred phylogenetic relationships (66). Our results indicate that combining the information from putative orthologous genes using supertrees is adequate to reconstruct the phylogenetic relationships among the sampled caecilians, and vertebrates in general.

As with other studies that have characterised transcriptomes, this study has a strong descriptive component (9), 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 individual de novo transcriptomes of caecilian amphibians presented here could be improved by additional sequencing of different tissues, individuals, developmental stages, and species (for instance, 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 remains a limiting step for research on many caecilian species (67), and dedicated fieldwork will likely be required to investigate broadly the genomic potential of this neglected, but important group of vertebrates.

Concluding remarks

Genome science has irreversibly changed the landscape of biological research. Understanding life processes and their 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 5,000 eukaryotic complete genomes available on the NCBI database, only five of them (some not fully avaible) are of amphibian species: *Ambystoma mexicanum* Shaw & Nodder, 1798, *Nanorana parkeri* Stejneger, 1927, *Rana catesbeiana* Shaw, 1802, *X. laevis* Daudin, 1802 and *X. tropicalis*. Despite the great effort made by initiatives such as the Genome 10K Project (68,69) and other genome-scale studies (e.g., Xenbase, 33; Salamander Genome project, 70), amphibians are the major group of vertebrates with fewest genomic resources available, and, importantly, none for the order Gymnophiona (71). 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. In order to start filling this gap, here we have reported transcriptomic data for five caecilian amphibian species. 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.

Data availability

Tissue-specific RNA-seq reads and species-specific de novo transcriptome assemblies are available from NCBI through BioProject ID number PRJNA387587.

References

1. Mardis E. R. 2008. The impact of next-generation sequencing technology on genetics. Trends in Genetics. 24(3): 133–41.
2. Rokas A. and Abbot P. 2009. Harnessing genomics for evolutionary insights. Trends Ecol Evol. 24(4): 192–200.
3. da Fonseca R. R., Albrechtsen A., Themudo G. E., Ramos-Madrigal J., Sibbesen J. A., Maretty L., Zepeda-Mendoza M. L., Campos P. F., Heller R. and Pereira R. J. 2016. Next-generation biology: Sequencing and data analysis approaches for non-model organisms. Mar Genomics. 30: 1–11.
4. Nagalakshmi U., Waern K., and Snyder M. 2010. RNA-seq: A method for comprehensive transcriptome analysis. Curr Protoc Mol Biol. 4(11): 1–13.
5. Conesa A., Madrigal P., Tarazona S., Gomez-Cabrero D., Cervera A., McPherson A., Szczesniak M. W., Gaffney D. J., Elo L. L., Zhang X. and Mortazavi A. 2016. A survey of best practices for RNA-seq data analysis. Genome Biol. 17(1):13.
6. Wang Z., Gerstein M. and Snyder M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 10(1):57–63.
7. Ozsolak F. and Milos P. M. 2011. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 12(2): 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(12): 2731–44.
9. Ekblom R. and Galindo J. 2010. Applications of next generation sequencing in molecular ecology of non-model organisms. Heredity. 107(1): 1–15.
10. San Mauro D., Gower D. J., Müller H., Loader S. P., Zardoya R., Nussbaum R. A., Wilkinson M. 2014. Life-history evolution and mitogenomic phylogeny of caecilian amphibians. Mol Phylogenet Evol. 73(1): 177–89.
11. Wilkinson M. 2012. Caecilians. Curr Biol. 22(17).
12. Wilkinson M., San Mauro D., Sherratt E. and Gower D. J. 2011. A nine-family classicification of caecilians (Amphibia: Gymnophiona). Zootaxa. 2874: 41–64.
13. Kamei R. G., San Mauro D., Gower D. J., Van Bocxlaer I., Sherratt E., Thomas

A., Babu S., Bossuyt F., Wilkinson M. and Biju S. D. 2012. Discovery of a new family of amphibians from northeast India with ancient links to Africa. Proc Biol Sci. 279(1737): 2396–401.

14. Darrel R. F. 2016. Amphibian Species of the World: Version 6.0. AMNH. http://research.amnh.org/herpetology/amphibia/index.html.
15. Roelants K., Gower D. J., Wilkinson M., Loader S. P., Biju S. D., Guillaume K., Moriau L. and Bossuyt F. 2007. Global patterns of diversification in the history of modern amphibians. Proc Natl Acad Sci. 104(3): 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(2): 492–508.
17. San Mauro D. 2010. A multilocus timescale for the origin of extant amphibians. Mol Phylogenet Evol. 56(2): 554–61.
18. Pyron R. A. 2011. Divergence time estimation using fossils as terminal taxa and the origins of lissamphibia. Syst Biol. 60(4): 466–81.
19. Marjanović D. and Laurin M. 2013. An updated paleontological timetree of lissamphibians, with comments on the anatomy of Jurassic crown-group salamanders (Urodela). J. Hist Biol. 26(4): 535–50.
20. Mueller O., Lightfoot S. and Schroeder A. 2004. RNA Integrity Number (RIN) – Standardization of RNA Quality Control Application. Agil Appl Note. 1–8.
21. Schmieder R. and Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 27(6): 863–4.
22. Andrews S. 2010. FastQC: A quality control tool for high throughput sequence data. 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 UsingPhred. I. Accuracy Assessment. Genome Res. 8(3): 175–85.
24. Kornobis E., Cabellos L., Aguilar F., Frías-López C., Rozas J., Marco J., Zardoya R. 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., Levin J. Z., Thompson D. A., Amit I., Adiconis X., Fan L., Raychowdhury R., Zeng Q., Chen Z., Mauceli E., Hacohen N., Gnirke A., Rhind N., di Palma F., Birren B. W., Nusbaum C.,

Lindblad-Toh K., Friedman N. and Regev A. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 29(7): 644–52.

26. Haas B. J., Papanicolaou A., Yassour M., Grabherr M., Blood P. D., Bowden J., Couger M. B., Eccles D., Li B., Lieber M., Macmanes M. D., Ott M., Orvis J., Pochet N., Strozzi F., Weeks N., Westerman R., William T., Dewey C. N., Henschel R., Leduc R. D., Friedman N. and Regev A. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 8(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(3):R25.
28. Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25(16): 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(2):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(1): 465.
31. Moreton J., Izquierdo A. and Emes R. D. 2016. Assembly, assessment, and availability of De novo generated eukaryotic transcriptomes. Front Genet. 6 (361): 1–9.
32. Parra G., Bradnam K. and Korf I. 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23(9): 1061–7.
33. Karpinka J. B., Fortriede J. D., Burns K. A., James-Zorn C., Ponferrada V. G., Lee J., Karimi K., Zorn A. M. and Vize P. D. 2015. Xenbase, the Xenopusmodel organism database; new virtualized system, data types and genomes. Nucleic Acids Res. 43(D1): D756–63.
34. Powell S, Forslund K, Szklarczyk D, Trachana K, Roth A, Huerta-Cepas J, Gabaldón T., Rattei T., Creevey C., Kuhn M., Jensen L. J., Von Mering C. and Bork P. 2014. EggNOG v4.0: Nested orthology inference across 3686 organisms. Nucleic Acids Res. 42(D1): 231–9.

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(3): 403–10.
36. Letunic I. and Bork P. 2007. Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics. 23(1): 127–8.
37. Fu L., Niu B., Zhu Z., Wu S. and Li W. 2012. CD-HIT: Accelerated for clustering the next-generation sequencing data. Bioinformatics. 28(23): 3150–2.
38. R Development Core Team. 2016. R: A Language and Environment for Statistical Computing. R Found Stat Comput Vienna Austria.
39. Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K. P., Kuhn M., Bork P., Jensen Lars J. and Von Mering C. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43(D1): D447–52.
40. Finn R. D., Coggill P., Eberhardt R. Y., Eddy S. R., Mistry J., Mitchell A. L., Potter S. C., Punta M., Qureshi M., Sangrador-Vegas A., Salazar G. A., Tate J. and Bateman A. 2016. The Pfam protein families database: Towards a more sustainable future. Nucleic Acids Res. 44(D1): D279–85.
41. Sean R. E. 2010. HMMER: biosequence analysis using profile hidden Markov models http://hmmer.janelia.org/.
42. Emms D. M. and Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16(1): 157.
43. Katoh K. and Standley D. M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 30(4): 772–80.
44. 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(7): 1641–50.
45. Le S. Q., Lartillot N. and Gascuel O. 2008. Phylogenetic mixture models for proteins. Philos Trans R Soc Lond B Biol Sci. 363(1512): 3965–76.
46. 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(17): 541–8.

47. Sayyari E. and Mirarab S. 2016. Fast coalescent-based computation of local branch support from quartet frequencies. Mol Biol Evol. 33(7): 1654–1668.
48. Maciel A. O., Sampaio M. I. C., Hoogmoed M. S. and Schneider H. 2016. Phylogenetic relationships of the largest lungless tetrapod (Gymnophiona, Atretochoana) and the evolution of lunglessness in caecilians. Zool Scr. 46(3): 255–63.
49. Mercer T. R., Dinger M. E. and Mattick J. S. 2009. Long non-coding RNAs: insights into functions. Nat Rev Genet. 10(3): 155–9.
50. Prachumwat A. and Li W. H. 2008. Gene number expansion and contraction in vertebrate genomes with respect to invertebrate genomes. Genome Res. 18(2): 221–32.
51. Albalat R. and Cañestro C. 2016. Evolution by gene loss. Nat Rev Genet. 17(7): 379–91.
52. Eckes B, Krieg T, Niessen CM. 2010. Biology of the skin. Therapy of Skin Diseases: A Worldwide Perspective on Therapeutic Approaches and Their Molecular Basis. 3–14.
53. Duellman WE, Trueb L. 1994. Biology of Amphibians. Johns Hopkins University Press, Baltimore, MD.
54. Clarke B. T. 1997. The natural history of amphibian skin secretions, their normal functioning and potential medical applications. Biol Rev Camb Philos Soc. 72(3): 365–79.
55. Toledo RC, Jared C. 1995. Cutaneous granular glands and amphibian venoms. Comparative Biochemistry and Physiology. Part A: Physiology. 1–29.
56. 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(1): 125–30.
57. Lazarus L. H. and Attila M. 1993. The toad, ugly and venomous, wears yet a precious jewel in his skin. Prog Neurobiol. 41(4): 473–507.
58. 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(2): 125–30.
59. 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.

60. Meng P., Yang S., Shen C., Jiang K., Rong M. and Lai R. 2013. The first salamander defensin antimicrobial peptide. PLoS One. 8(12): 83044.
61. Kupfer A., Müller H., Antoniazzi M. M., Jared C., Greven H, Nussbaum R. A. and Wilkinson M. 2006. Parental investment by skin feeding in a caecilian amphibian. Nature. 440(7086): 926–9.
62. 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(4): 358–61.
63. Taylor, E. H. 1972. Squamation in caecilians, with an atlas of scales. The University of Kansas Science Bulletin 49: 989–164.
64. Gower D. J. and Wilkinson M. 2002. Phallus morphology in caecilians (Amphibia, Gymnophiona) and its systematic utility. Bull Nat Hist Museum Zool Ser. 68(2): 143–54.
65. 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(3): 191–202.
66. Gabaldón T. and Koonin E. V. 2013. Functional and evolutionary implications of gene orthology. Nat Rev Genet. 14(5): 360–6.
67. Gower D. J. and Wilkinson M. 2005. Conservation biology of caecilian amphibians. Conserv Biol. 19(1): 45–55.
68. Genome 10K Community of Scientists. 2009. Genome 10K: A proposal to obtain whole-genome sequence for 10000 vertebrate species. Journal of Heredity. 100(6): 659–74.
69. Koepfli K., Paten B., Genome 10K Community of Scientists and O'Brien S. J. 2015. The Genome 10K Project: a way forward. Annu Rev Anim Biosci. 3: 57–111.
70. Smith J. J., Putta S., Walker J. A., Kump D. K., Samuels A. K., Monaghan J. R., Weisrock D. W., Staben C. and Voss S. R. 2005. Sal-Site: integrating new and existing ambystomatid salamander research and informational resources. BMC Genomics. 6: 181.

71. Shaffer H. B., Gidiş M., McCartney-Melstad E., Neal K. M., Oyamaguchi H. M., Tellez M. and Toffelmier E. M. 2015. Conservation genetics and genomics of amphibians and reptiles. Annu Rev Anim Biosci. 3: 113–38.

Tables and Figures

Table 1

Information on the species-specific caecilian transcriptome assemblies and their annotation. N = number of specimens, T = number of tissues, % CEGs = Percentage completeness core eukaryotic genes, KVGF = known vertebrate gene families.

Species	N	T	Contigs	% CEGs	Protein-coding genes	Annotated veNOG best hits	Annotated protein-coding genes in KVGF
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

Table 2

Novel tissue-specific genes in caecilians. The number of transcriptomes determined for each tissue, and the tissue-specific gene families (known vertebrate and caecilian-specific) 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 putative novel caecilian gene families.

	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
Putative novel caecilian gene families	38	14	46	59	11	30	130	8	167	493
P value	0. 2771	0.7932	0.3880	0.6812	1	0.2428	**1.3e-05**	**0.0064**	0.0818	-

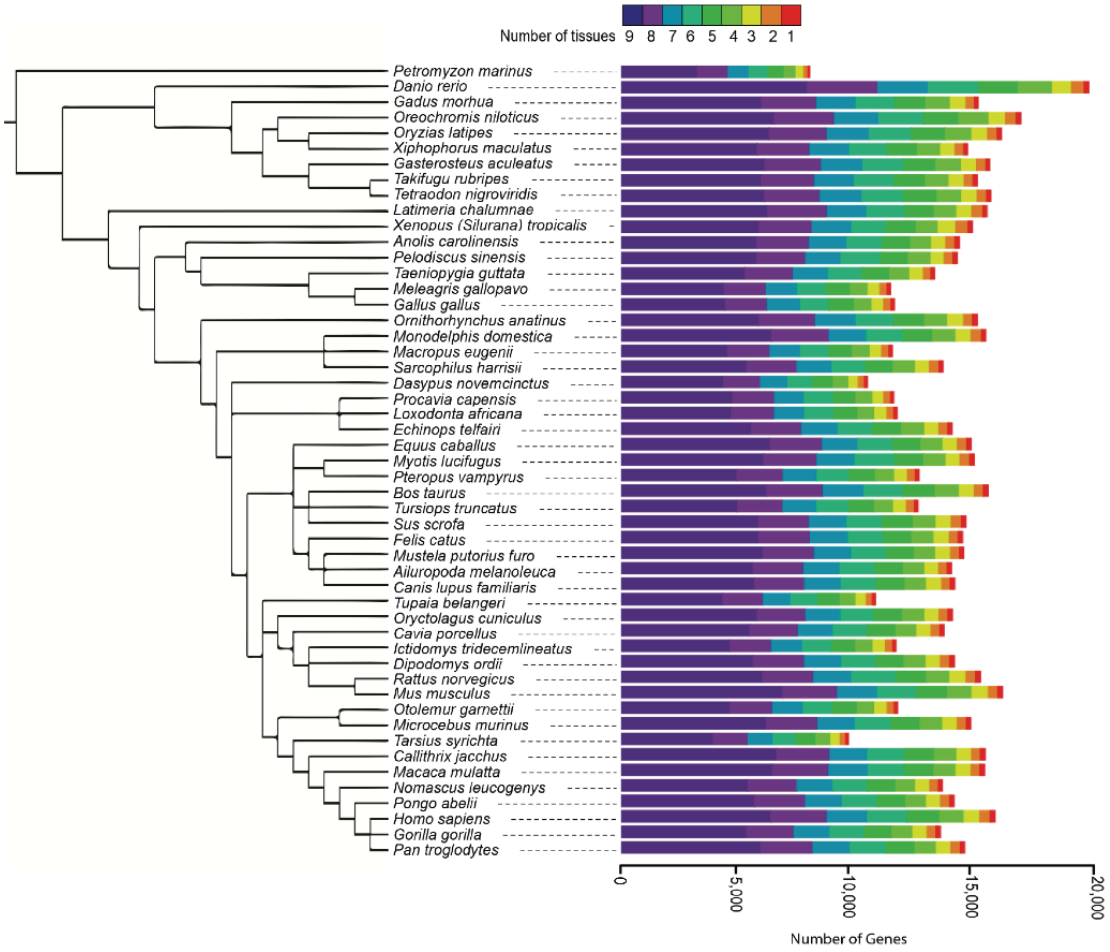


Figure 1

Numbers of annotated genes in the 51 vertebrate species available on the EggNOG database that are expressed in caecilians (pooled for the five sampled species-specific transcriptomes), mapped onto a vertebrate phylogeny inferred from the NCBI's taxonomic identifications. For each vertebrate taxon compared with caecilians, the number of annotated genes in common is subdivided to show the number of caecilian tissue types in which those genes are expressed.

Supplementary material

Table S1

Sample information and quality of RNA extractions for each transcriptome sequenced.

Species	Specimen voucher	Sex	Tissue	RIN value	Number of reads
Caecilia tentaculata (Caeciliidae)	MW 10281	Male	Spleen	9.9	49,757,992
			Foregut	8.8	54,057,168
			Heart	9.8	56,478,032
			Kidney	9.6	51,051,562
			Liver	9.1	57,568,844
			Muscle	9.5	58,911,078
			Posterior Skin	8.7	43,278,838
			Skin	10	40,889,140
			Testis	9.5	56,700,716
			Lung	9.5	46,260,502
Microcaecilia dermatophaga (Siphonopidae)	MW 10280	? (juvenile)	Kidney	9.5	42,750,256
			Liver	9.3	56,029,426
			Posterior Skin	9.2	39,517,710
			Skin	9.6	43,291,236
Microcaecilia unicolor (Siphonopidae)	MW 3338	Female	Kidney	8.9	52,616,356
			Liver	8.2	58,062,338
			Skin	7.8	32,957,690
	MW 10282	Female	Foregut	9.6	47,417,282
			Liver	8.9	55,237,106
			Muscle	8.6	53,570,730
			Posterior Skin	8.2	49,216,256
			Skin	8.9	40,774,934
			Lung	8.9	48,841,504
Rhinatrema bivittatum (Rhinatrematidae)	MW 3339	Male	Kidney	9.2	48,775,398
			Liver	8.8	51,635,654
			Skin	9.7	39,705,618
			Testis	8.9	53,962,268
	MW 10279	Female	Spleen	10	56,676,152
			Foregut	8.6	47,607,116
			Liver	8.8	47,561,092
			Muscle	9.4	52,900,438
			Skin	8.3	39,106,530
			Lung	9.1	51,182,452
Typhlonectes compressicauda (Typhlonectidae)	MW 10283	Male	Foregut	9.8	53,680,574
			Heart	9.4	53,159,450
			Kidney	9.3	45,034,380
			Liver	9.8	51,838,608
			Posterior Skin	10	54,023,400
			Skin	9.9	36,051,312
			Lung	8.7	44,973,848

Table S2

Metrics of the species-specific transcriptomes. Putative protein-coding genes found in each caecilian species-specific transcriptome and number of contigs with open reading frames (ORFs) are shown.

Species	Contigs	N50	Median contig length	Mean contig length	Alignment percentage	ORFs' isoforms	Unique ORFs
Caecilia tentaculata	142,502	1884	429	932.82	96.01	63,540	27,384
Microcaecilia dermatophaga	106,298	1784	426	903.73	97.78	42,510	22,058
Microcaecilia unicolor	146,348	1587	355	850.91	96.93	59,355	26,302
Rhinatrema bivittatum	201,584	1713	398	857.76	96.33	83,643	34,654
Typhlonectes compressicauda	134,394	1263	357	713.87	96.25	59,151	27,603

Table S3

Detection of protein-protein interactions (PPIs) and functional enrichment paths in the tissue-specific known vertebrate gene families.

		Foregut	Kidney	Liver	Spleen	Testis
Known vertebrate gene families		19	21	18	11	80
Number of nodes		17	20	11	11	75
Number of edges		5	5	6	3	19
PPI enrichment p-value		2.76e-06	1.31e-05	0.00258	4.75e-05	4.82e-07
Functional enrichment	**Biological Process GO (#)**	GO:0007586 (8)	GO:0098656 (6); GO:1903825 (5); GO:0003333 (4); GO:0046942 (5); GO:0015889 (2)	GO:0030193 (3); GO:0051918 (2); GO:0072376 (3)	GO:0004252 (4)	-
	KEGG (#)	-	-	-	04972 (3)	-

Table S4

Results of the Pfam annotation of the putative novel caecilian gene families.

Tissue	Putative novel caecilian gene families	Annotated novel caecilian gene families	Number of protein family domains	Protein family domain
Foregut	38	7	11	Dimer_Tnp_hAT Snapin_Pallidin TMF_DNA_bd SRCR Chromo CBX7_C RVT_1 Dam DAP10 Adeno_E3_CR2 PEARLI-4
Heart	14	4	6	DUF4749 BORCS8 ReosigmaC TMEM190 TEX29 LAP2alpha
Kidney	46	9	22	FISNA Tropomyosin ATG16 IncA Apolipoprotein BMFP Tektin XhlA EzrA SlyX zf-C2H2 zf-met Ephrin_rec_like Cadherin_pro IGF V-set I-set Ig Ank CH Exo_endo_phos 2 LAP2alpha
Liver	59	18	22	Dynein_heavy Pkinase A2M_recep E1_DerP2_DerF2 NinD KRAB zf-met zf-C2H2 zf-trcl Zn-ribbon_8 KASH Spectrin Mod_r BCAS2 LAP2alpha adh_short

				RVT_1 A_deaminase DUF3268 CENP-P Hydrolase Transposase_22
Lung	11	2	2	LAP2alpha zf-C2H2
Muscle	30	6	39	IncA ATG16 EMP24_GP25L DUF4600 Fib_alpha CLZ FlaC_arch BRE1 XhlA Nsp1_C Reo_sigmaC DUF1664 Spc7 ABC_tran_CTD EzrA Laminin_II Apolipoprotein DUF812 CENP-F_leu_zip YjcZ EspB Muted KASH_CCD Spectrin TMF_DNA_bd IFT57 Prefoldin ADIP ERM Jnk-SapK_ap_N MscS_porin SlyX TOBE_2 TSP_C Dimer_Tnp_hAT DUF4413 LAP2alpha DUF3584 Kre28
Skin	130	26	29	Nup192 Ank Trypsin DUF2630 Chromo zf-C2H2 zf-RVT DUF4061 UPAR_LY6 Toxin_TOLIP gag-asp_proteas Asp_protease_2 EF-hand DUF1151 C1-set KIX_2 LAP2alpha Tup_N DUF4381

				Spectrin Ldl_recept_a DUF4407 DUF4630 RVT_1 PhaP_Bmeg DUF724 DDE_Tnp_4 OmpH TCTP
Spleen	8	2	3	Cystatin ATP1G1_PLM_MAT8 SQAPI
Testis	167	26	44	7tm_3 DNA_RNApol_7kD zf-LYAR zf-trcl zf-met zf-C2H2 CD225 AIP3 CpXC Ferrochelatase Kazal Ig I-set V-set Reo_sigmaC Serpin Spc7 Tmemb_cc2 fn3 NPV_P10 DUF1664 PXA Nexin_C DUF1518 DUF812 DUF2046 Myb_DNA-bind_5 GTP_EFTU G-alpha YhfH LIM FYDLN_acid Glyco_transf_11 RVT_1 Pur_ac_phosph_N TIL Pkinase Kinase-like FTA2 Haspin_kinase Kdo Peptidase_M14 UPF0564 MatE

Table S5

The number of orthologous genes found in each species (out of 1,995). Caecilian species are highlighted in bold type.

	Class	Taxon occupancy
Caecilia tentaculata	**Amphibia**	**333**
Microcaecilia dermatophaga	**Amphibia**	**240**
Microcaecilia unicolor	**Amphibia**	**314**
Rhinatrema bivittatum	**Amphibia**	**357**
Typhlonectes compressicauda	**Amphibia**	**304**
Xenopus tropicalis	Amphibia	134
Gallus gallus	Aves	204
Meleagris gallopavo	Aves	165
Taeniopygia guttata	Aves	215
Petromyzon marinus	Cephalaspidomorphi	83
Latimeria chalumnae	Coelacanthiformes	246
Ailuropoda melanoleuca	Mammalia	536
Bos taurus	Mammalia	477
Callithrix jacchus	Mammalia	701
Canis lupus familiaris	Mammalia	520
Cavia porcellus	Mammalia	426
Dasypus novemcinctus	Mammalia	370
Dipodomys ordii	Mammalia	387
Echinops telfairi	Mammalia	421
Equus caballus	Mammalia	464
Felis catus	Mammalia	493
Gorilla gorilla	Mammalia	966
Homo sapiens	Mammalia	888
Ictidomys tridecemlineatus	Mammalia	430
Loxodonta africana	Mammalia	448
Macaca mulatta	Mammalia	648
Macropus eugenii	Mammalia	248
Microcebus murinus	Mammalia	553
Monodelphis domestica	Mammalia	318
Mus musculus	Mammalia	521
Mustela putorius	Mammalia	422
Myotis lucifugus	Mammalia	389
Nomascus leucogenys	Mammalia	794
Ornithorhynchus anatinus	Mammalia	299
Oryctolagus cuniculus	Mammalia	448
Otolemur garnettii	Mammalia	451
Pan troglodytes	Mammalia	818
Pongo abelii	Mammalia	876
Procavia capensis	Mammalia	433
Pteropus vampyrus	Mammalia	551
Rattus norvegicus	Mammalia	523
Sarcophilus harrisii	Mammalia	339
Sus scrofa	Mammalia	462
Tarsius syrichta	Mammalia	401
Tupaia belangeri	Mammalia	439
Tursiops truncates	Mammalia	545
Anolis carolinensis	Reptilia	185
Pelodiscus sinensis	Reptilia	220
Danio rerio	Teleostei	341
Gadus morhua	Teleostei	302
Gasterosteus aculeatus	Teleostei	327
Oreochromis niloticus	Teleostei	272
Oryzias latipes	Teleostei	280
Takifugu rubripes	Teleostei	204
Tetraodon nigroviridis	Teleostei	255
Xiphophorus maculatus	Teleostei	346

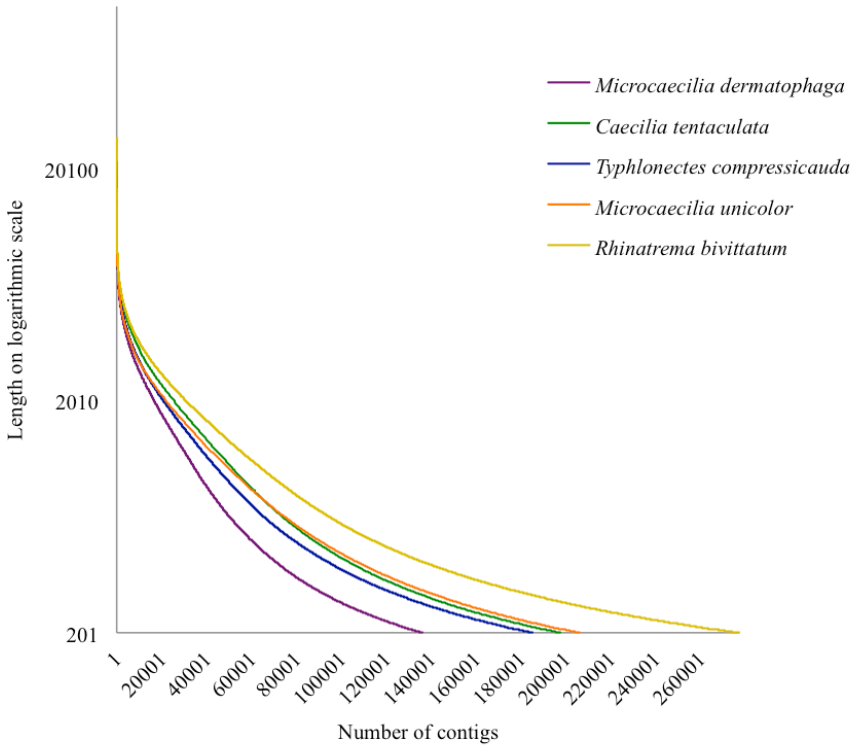


Figure S1

Contig lengths for the five specific-species caecilian transcriptomes.

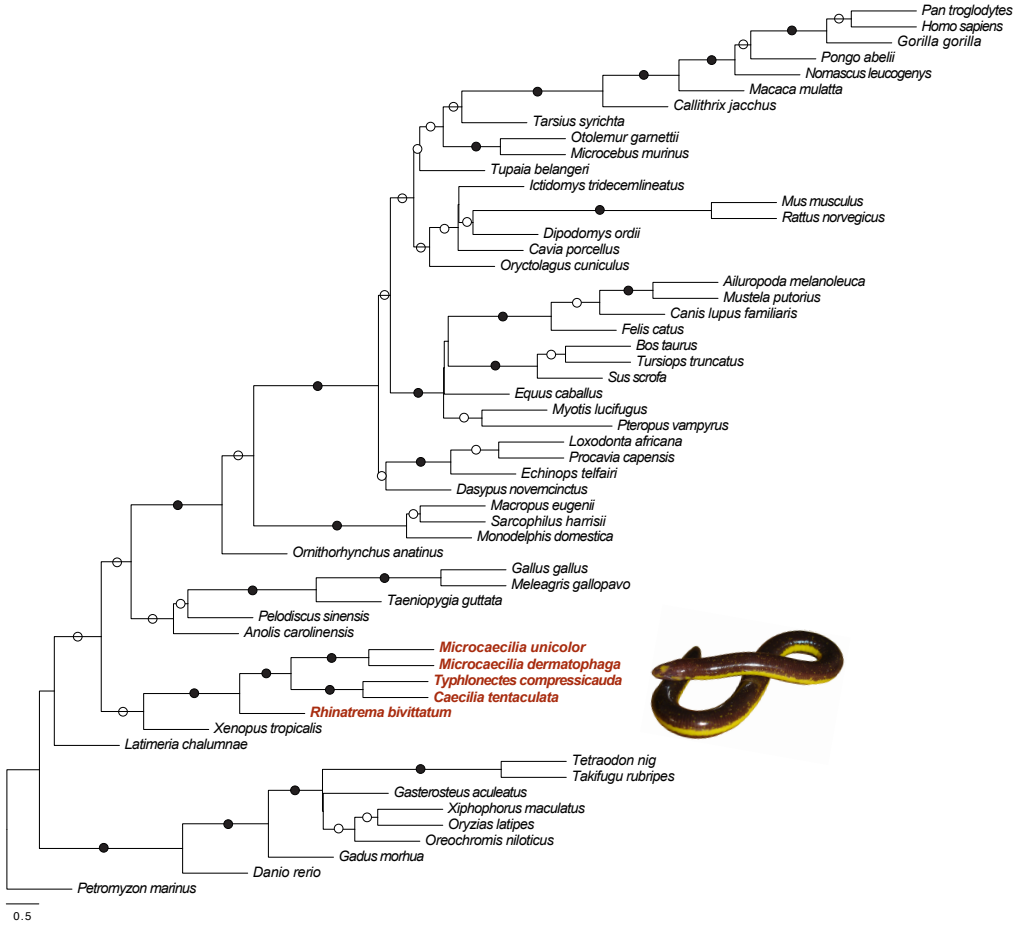


Figure S2

Supertree of vertebrates reconstructed from 1,955 orthologous gene trees using ASTRAL. Filled bullets on branches denote 'good' support as measured by both posterior probabilities (≥0.95) and quartet percentages (≥70%) for the respective internal branches. Open bullets denote 'good' support for posterior probabilities (≥0.95) but not for quartet percentages (<70%). Absence of bullet on a branch denotes lower support as measured by both posterior probabilities (<0.95) and quartet percentages (<70%). Scale bar indicates substitutions per site. The five sampled caecilian species are highlighted. Picture shows a specimen of *Rhinatrema bivittatum* from Angouleme, French Guiana (photo by Diego San Mauro).

Chapter 2

Behind the scenes: molecular innovations during caecilian amphibian evolution

María Torres-Sánchez, David J. Gower, David Alvarez-Ponce, Christopher J. Creevey, Mark Wilkinson, Diego San Mauro

Abstract

All evolutionary changes hide genetic variation. The emergence of molecular innovations may be one of these genetic changes. At the molecular level, innovations are evidenced by the study of the different ratios of nucleotide substitutions in a positive selection framework. These innovations can be correlated with specific adaptations of the organisms to unravel their radiation and their particular current lifestyle. Among the least known vertebrate species are caecilian amphibians (order Gymnophiona). Caecilians are limbless tropical animals adapted to live in soil. Little or nothing is known about the molecular changes that caecilians overcame to adapt to fossorial life. In this study, we analysed 8540 orthologous genes from five species of caecilian amphibians and the frog *Xenopus tropicalis* in order to identify putative molecular innovations at different times of the caecilian amphibian evolution. We found a total of 167 genes that present positive selection signatures. These genes provide valuable insights about ancestral and more recent innovations in caecilian amphibians and about the trends of molecular evolution in vertebrates.

Introduction

Phenotypic evolutionary changes, including those associated with adaptive radiation and the exploitation of novel environments, ultimately have a molecular basis that can involve a variety of genetic changes, including gene gain, loss or other innovations (1–3). With the massive amount of genomic data becoming available, a better understanding of molecular genetics and the evolutionary mechanisms underpinning biodiversity has become attainable. Molecular evolutionary processes can be investigated by studying regulatory and/or functional elements of genomes. In protein-coding genes natural selection can be investigated by comparing rates of nucleotide substitutions (non-synonymous [dN] and synonymous [dS]). The ratio between these rates, omega ($\omega = dN/dS$), provides a means of identifying selective pressure in proteins (4).

The radiation of vertebrates is in part explained by the presence of genetic innovations (5–7), with their new functions involved in adaptations to different environments. One of these environments is the soil, which presents several restrictive conditions, including low levels of light, low airborne transmission of sound and scent, hypercapnia and hypoxia. In addition, many microorganisms (fungi, protozoans, bacteria) and diverse invertebrates (often pathogenic) abound in especially humid and thermally stable soils (8). While it may seem a challenging environment, several different groups of vertebrates are well adapted to live in soil (9,10), including one of the oldest lineages of extant terrestrial vertebrates, the caecilian amphibians.

Caecilians (order Gymnophiona) are highly specialized amphibians in which adult forms of most species burrow in soil. Most other amphibians that spend time in soil take advantage of pre-existing holes and tunnels and feed and breed above ground (11). In contrast, most adult terrestrial caecilians are highly fossorial, dedicated burrowers that feed and breed within moist soils (12). Given that living in soil is a derived condition among amphibians, it is likely that the evolution of caecilians has been strongly influenced by adaptation to this environment. For example, several morphological features of caecilians are likely adaptations to life in soil, such as their modified skull architecture for head-first burrowing and feeding underground (13),

elongate limbless bodies with modified axial musculature (14), reduced visual and hearing systems, and novel sensory tentacles (15,16). Very little is known about the molecular changes associated with the evolutionary origin and radiation of caecilians, providing an as yet unexploited opportunity to further explore patterns of molecular change in vertebrate evolution (17).

Recently, reference transcriptomes for five species of caecilians have been generated (Chapter 1), enabling analyses of adaptive molecular evolution of this major group of vertebrates. Here, we compare for the first time rates of nucleotide substitutions in orthologous protein-coding genes of these caecilian transcriptomes in order to identify genes under positive selection. We identify some probable molecular innovations plausibly involved in adaptation to living in soil, and other molecular adaptations that we hypothesize to be correlated to specific traits of this amphibian group.

Materials and methods

Genomic data

The source data of this study were the protein-coding gene sequences (both nucleotide and amino-acid level) from reference transcriptomes of five caecilian species (*Rhinatrema bivittatum* Cuvier in Guérrin-Méneville, 1838, *Caecilia tentaculata* Linnaeus, 1758, *Typhlonectes compressicauda* Duméril & Bibron, 1841, *Microcaecilia unicolor* Duméril, 1861, and *M. dermatophaga* Wilkinson, Sherratt, Starace & Gower, 2013; Chapter 1) as well as those for the frog *Xenopus tropicalis* Gray, 1864, the only amphibian currently represented in the Ensembl database (18). For each *X. tropicalis* gene, the longest isoform coding region was chosen for analysis, and BLAST searches (blastp tool, version 2.2.28; E-value < 10^{-10}; ref. 19) conducted against the transcriptomes of each of the caecilian species. Likewise, each caecilian protein-coding gene was used as a query in a BLAST search against the *X. tropicalis* proteome. Pairs of best reciprocal hits were considered orthologs. Only *X. tropicalis* genes with putative orthologs in all five caecilian species were used in downstream analyses.

For each group of orthologs, the inferred amino acid sequences were aligned using PRANK (20). Given the sensitivity of positive selection analyses to alignment error, we carried out thorough filtering of the alignments. First, Gblocks version 0.91b (21) was used to remove problematic regions. Second, (as in refs. 22,23) two *ad hoc* sliding window filters (of 15 and 5 residues) were used to eliminate regions coding for amino acids that are unique to one species (with 10 or more amino acid singletons, or where all five were singletons, respectively) because such regions are often annotation errors. The resulting amino acid sequence alignments were used to guide the alignment of the protein-coding genes.

Tests of positive selection

To infer positive selection, we performed branch-site model tests (24,25) for every group of orthologous genes and for every branch of the phylogeny based on chapter 1 and literature (Figure 1), except for that of the outgroup *X. tropicalis*, using the CODEML program in PAML 4.6 (26). The branch-site model test (model A and null

model A; 27) assumes that only a fraction of sites might have undergone positive selection and only along a single *a priori* identified branch (foreground lineage) on the phylogeny. The test assumes four classes of sites: codons that are conserved ($\omega < 1$), codons that are evolving neutrally ($\omega = 1$), and codons under positive selection in the foreground branch but conserved (2a) or neutral (2b) on the other (background) branches ($\omega > 1$). Model A was implemented with a default starting value (0.4) for ω, and used as the alternative hypothesis for the Likelihood Ratio Test (LRT). The null model of the LRTs was the null model A with ω fixed at 1 for sites under positive selection on the foreground branch (2a and 2b sites). P-values for the LRTs were computed using the χ^2 distribution with one degree of freedom, and divided by two (26). Multiple-testing corrections were conducted following Benjamini and Hochberg's method in order to control for a false discovery rate (FDR) using the program R (28). Orthologs with a corrected p-value < 0.1 and $\omega > 1$ for the foreground branch (2a and 2b sites) were assumed to be genes under positive selection.

Gene ontology annotation and network analysis

For each of the putative orthologs groups inferred to be under positive selection, we obtained the associated gene ontology (GO) terms from the *X. tropicalis* annotation using the BioMart data-mining tool (Ensembl release 89; 18). We summarized and visualized the common GO terms of the selected genes and their frequencies of occurrence using REVIGO applying 0.7 % allowed similarity (by the semantic similarity method) and using the whole UniProt database to define the size of each GO term (29). Finally, protein-protein interactions (PPIs) and functional enrichment paths were inferred using STRING (30) with *X. tropicalis* as the reference organism and default settings.

Results and discussion

General view of caecilian innovations

We found 8540 one-to-one orthologous gene groups (orthogroups) among the transcribed protein-coding genes of the five sampled caecilian species and the frog outgroup. Using branch-site models, we detected signals of adaptive molecular evolution along all branches in the sampled caecilian tree (Figure 1). Numbers of genes with evidence of sites under positive selection ($\omega > 1$) are presented in Table 1 (see Supplementary Table S1 for more details). Our analyses identified 167 protein-coding genes that bear signals of having been under positive selection in the evolution of caecilians.

This is almost certainly a substantial underestimate given our conservative selection of orthogroups (present in every species, including *X. tropicalis*; no paralogs; stringent filtering). Despite the relaxed Type-I error rate (0.1), the stringent filtering lends some confidence that we have minimized false positives due to alignment artefacts to which the methods are known to be sensitive (31), and that the identified genes constitute potential molecular innovations of Gymnophiona.

Just two of the nine branches account for almost 50% of the protein-coding genes with signatures of positive selection, the branch that subtends the clade comprising all sampled caecilians, referred to subsequently as the "Gymnophiona branch" (branch 1 in Figure 1: 50 genes, 29.94%) and the terminal branch subtending *M. dermatophaga* (branch 6 in Figure 1: 33 genes, 19.76%). No significant PPIs were found for any of the sets of genes under positive selection on each branch, and only one pathway on the Gymnophiona branch presents evidence of functional enrichment linked to four genes considered to be involved in extracellular matrix interactions (Figure 3). The vast majority of the genes inferred to have been positively selected were associated with GO terms for 256 different biological processes, 76 cellular components and 173 molecular functions (Table 1 and Supplementary Table S1). Figure 2 and Supplementary Figures S1-S8 show network graphs of terms related to biological processes for each analysed branch. Gene ontologies are continuously redefined and even though they are considered global generalizations and taxon-neutral, several GO

terms present taxon constraints being species specific (32,33). Despite that, valuable insights into molecular innovations in Gymnophiona can be extracted from the GO annotations.

Cellular component domains (GO:0016020 and GO:0016021) are the most common terms assigned to the positively selected genes, and many of these genes are also associated with extracellular biological process terms. This high prevalence indicates an important role for the cell membrane and its integral components during caecilian molecular genetic evolution. One wave of gene innovations associated with the origins of major tetrapod groups is proposed to be related to the regulation of extracellular signaling (17) and our results suggest that innovations in functional membrane elements are likely an additional important genetic aspect of vertebrate macroevolution. Others GO terms of the positive selected protein-coding genes in caecilians have been already reported as adaptations in fossorial animals, such as genes involved in oxidation-reduction processes (34–36). These animals adaptively converge on different specialized levels, including the genetic level. Some other genes under positive selection in our analyses were related to processes and functions that might be involved with specific caecilian traits as discussed in the following sections.

Ancient genetic toolkit for caecilians

The largest number of protein-coding genes inferred to have been under positive selection (50 genes) was found on the Gymnophiona branch. These 50 protein-coding genes are involved in 96 biological processes based on their GO annotation (Table 1, Figure 2 and Supplementary Table S1), including several processes related to development (*lamc1*, *tet2*, *nup153*, *tacc2*, SPG11, see Supplementary Table S1). Among these elements, there is a component of the extracellular glycoprotein matrix of the membrane, the laminin subunit gamma 1 (*lamc1*), that is essential for basement membrane assembly during embryogenesis (37). Several developmental processes are associated with *lamc1* by GO terms. Additionally, *lamc1* is one of the four elements of the single detected functional enrichment, which is linked to extracellular matrix interaction mechanisms such as cell adhesion and cell-to-cell communication (ECM-receptor interaction, KEEG pathway ID: 04512; see Figure 3). Among other functions, *lamc1* is related with light perception (GO:0050908) and retinal

development (GO:0031290). Compared with other amphibians, caecilians are rod-only monochromats with small eyes covered by skin and sometimes also bone (38). Light is not only important for visual perception, but also plays other important roles controlling, for example, circadian rhythms, which is vital for synchronization of biological cycles (39). We hypothesize that molecular innovation in *lamc1* might be involved in adaptation of circadian rhythms underground.

Biological process terms related to oxidation-reduction (redox) are also associated with several protein-coding genes inferred to be under positive selection on the Gymnophiona branch (*sod3*, *akr1a1*, *qsox1*, CP, see Supplementary Table S1). Environmental conditions could have driven the emergence of molecular innovations to tolerate chronic low oxygen (O_2) and high carbon dioxide (CO_2) levels that characterise life in soil (8). At higher levels, CO_2 is converted to acid by ionic dissociation and can cause oxidative stress, in turn related to disease and ageing (40). Additionally, O_2 deprivation can affect synaptic transmission and ultimately cause cell death by cytosolic accumulation of calcium ions (Ca^{2+}; 41). As stated in its GO term description, rabphilin-3A (encoded by *rph3a*, see Supplementary Table S1) is a protein involved in the regulation of synaptic vesicle traffic that mediates the release of a neurotransmitter when Ca^{2+} cytosolic levels rise. Redox process innovations might contribute to the development of better protective mechanisms to increased cytotoxic threats in the edaphic atmosphere.

Molecular innovations on the Gymnophiona branch appear to have provided diverse mechanisms for meeting the challenges of soil environments. Molecular innovations including other branches are more recent evolutionary changes within Gymnophiona. The following subsections explore some of the biological processes that are associated with several putative positive selected protein-coding genes in a specific branch or in several independent branches of our analyses.

Collagen scales

Five protein-coding genes annotated as collagen chains, were found under positive selection in several branches (*col4a2*, COL17A1, *col4a1, col12a1* and COL5A2). Collagen chains are structural proteins classified under different types, and main

components in general of skin, connective tissues, bone, teeth and epitheliums (42). Many caecilian amphibians present collagenous scales with no clear function (43). In the absence of verification, we hypothesise that these protein-coding genes might be collagen chains involved in the formation of the caecilian scales and could protected the intern organs for the soil pressure on the caecilians bodies underground.

Lipid metabolism

Lipid metabolism and fatty acid metabolism are biological processes associated with several positively selected genes (*acot2*, *gdpd5*, *plpp1*, *elovl5*, *sptlc3*, *cyp17a1*, *lcat*, *asah1*, *cers6*, see Supplementary Table S1) on different branches within Gymnophiona. Lipids have very diverse biological functions and play important roles in energy storage, signaling, and the formation of barriers in the cell membrane. They are also involved in other vital roles in caecilians, including the provision of nutrition to developing fetuses and/or newborn during oviductal and/or skin feeding (44,45). Some of these genes might be related to the synthesis, transformation and/or storage of lipids for these traits.

Pigmentation or depigmentation?

Another gene that drew our attention was linked to pigmentation by the GO term: GO:0043473, which is inferred to have been under positive selection along the *Microcaecilia* branch (branch number 4 in Figure 1). This protein-coding gene is annotated as a tetraspanin (*tspan36*, see Supplementary Table S1). Tetraspanins are a large family of transmembrane proteins (38 homologous in vertebrates) that are involved in diverse biological processes acting as organizers in the membranes of all kinds of animal cells (46). The functions of all the tetraspanins are not well known but some members of this family have been associated with pigment cell interactions and pigment pattern formation (47). Despite spending all or most of their lives in soil, many caecilian species are brightly coloured, perhaps aposematically in some cases (48). Caecilian species exhibit a range of colours and patterns within *Microcaecilia* with no clear ancestral state. They are more dedicated burrowers and being more fossorial might have consequences for pigment evolution including the molecular innovation in *tspan36*.

The immune system

Several components of the caecilian immune system are inferred to have been under positive selection along different branches (*tet2*, *masp1*, *enpp3*, *yes1*, *fyn*, see Supplementary Table S1). They are not surprising innovations, genes related to the immune system are likely involved in evolutionary arms races against aggressors (parasites and/or predators) and consequently under positive selection (49). Innovations in the immune system of caecilians could be related causally to the particular challenges of living in moist soils with constant physical contact with microbial rich substrate. Amphibians, survivors of the Earth's last four mass extinctions, are facing an unprecedentedly high risk of extinction that seems to be linked, in part, to challenges to their immune systems (50,51). Immune system mechanisms are in need of better understanding.

Concluding remarks

Molecular adaptive evolution of caecilians is found associated mostly with protein-coding gene products with a membrane or extracellular location and are consistent with the general view of molecular evolution (52). These genes presented low levels of conservation and connectivity (no significant PPIs and only one pathway with functional enrichment were found) and are expected to be difficult to annotate. The protein-coding genes found to have been under positive selection in our analyses, are prevalent membrane components and 12 of them are uncharacterised thus far. The 167 genes inferred to have been under positive selection in our analyses are candidate genes of diversifying selection as a result of the adaptation to the life underground. Further experiments are required to test the function of these protein-coding genes in caecilians and identify their actual role in biological processes. The inclusion of representatives of additional caecilian linages in future studies could provide deeper insights of the selective pressure in the different caecilian species associated to their adaptation to particular habitats and life styles.

References

1. Wagner A. 2011. The molecular origins of evolutionary innovations. Trends in Genetics. 27(10): 397–410.
2. Bergthorsson U., Andersson D. I. and Roth J. R. 2007. Ohno's dilemma: Evolution of new genes under continuous selection. Proc Natl Acad Sci. 104(43): 17004–9.
3. Conant G. C. and Wolfe K. H. 2008. Turning a hobby into a job: How duplicated genes find new functions. Nat Rev Genet. 9(12): 938–50.
4. Yang Z. 2008. Adaptive Molecular Evolution. Handbook of Statistical Genetics. Third Edition. 375–406.
5. Meyer A. and Schartl M. 1999. Gene and genome duplications in vertebrates: The one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Curr Opin Cell Biol. 11(6): 699-704.
6. Ohno S. 1970. Evolution by Gene Duplication. Springer-Verlag.
7. Taylor J. S. and Raes J. 2004. Duplication and Divergence: The Evolution of New Genes and Old Ideas. Annu Rev Genet. 38(1): 615–43.
8. Moreira F. M. S, Huising E. J., Bignell D. E. and Senwo Z. 2012. A Handbook of Tropical Soil Biology: Sampling & Characterization of Below-ground Biodiversity. Soil Sci Soc Am J. 76(1): 309.
9. Nevo E. 1979. Adaptive convergence and divergence of subterranean mammals. Annu Rev Ecol Syst. 10: 269–308.
10. Wake M. H. 1993. The skull as a locomotor organ. In: Hanken J. and Hall B. K., editors. The Skull: Functional and Evolutionary Mechanisms. Chicago: The University of Chicago Press. 197–240.
11. Wells K. D. 2007. The Ecology and Behavior of Amphibians. University of Chicago Press.
12. Wilkinson M. 2012. Caecilians. Curr Biol. 22(17).
13. Sherratt E., Gower D. J., Klingenberg C. P. and Wilkinson M. 2014. Evolution of Cranial Shape in Caecilians (Amphibia: Gymnophiona). Evol Biol. 41(4): 528–45.
14. O'Reilly J. C., Summers A. P. and Ritter D A. 2000. The Evolution of the Functional Role of Trunk Muscles During Locomotion in Adult Amphibians.

Am Zool. 40(1): 123–35.

15. Wake M. H. 1985. The comparative morphology and evolution of the eyes of caecilians (Amphibia, Gymnophiona). Zoomorphology. 105(5): 277–95.
16. Maddin H. C. and Sherratt E. 2014. Influence of fossoriality on inner ear morphology: Insights from caecilian amphibians. J Anat. 225(1): 83–93.
17. Lowe C. B., Kellis M., Siepel A., Raney B. J., Clamp M., Salama S. R., Kingsley D. M., Lindblad-Toh K. and Haussler D. 2011. Three Periods of Regulatory Innovation During Vertebrate Evolution. Science. 333(6045): 1019–24.
18. Yates A., Akanni W., Amode M. R., Barrell D., Billis K., Carvalho-Silva D., Cummins C., Clapham P., Fitzgerald S., Gil L., Girón C. G., Gordon L., Hourlier T., Hunt S. E., Janacek S. H., Johnson N., Juettemann T., Keenan S., Lavidas I., Martin F. J., Maurel T., McLaren W., Murphy D. N., Nag R., Nuhn M., Parker A., Patricio M., Pignatelli M., Rahtz M., Riat H. S., Sheppard D., Taylor K., Thormann A., Vullo A., Wilder S. P., Zadissa A., Birney E., Harrow J., Muffato M., Perry E., Ruffier M., Spudich G., Trevanion S. J., Cunningham F., Aken B. L., Zerbino D. R. and Flicek P. 2016. Ensembl 2016. Nucleic Acids Res. 44(D1): D710–6.
19. Altschul S. F., Gish W., Miller W. T., Myers E. W. and Lipman D. J. 1999. Basic local alignment search tool. J Mol Biol. 215(3): 403–10.
20. Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods Mol Biol. 1079: 155–70.
21. Talavera G. and Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol. 56(4): 564–77.
22. Luisi P., Alvarez-Ponce D., Pybus M., Fares M. A., Bertranpetit J. and Laayouni H. 2015. Recent positive selection has acted on genes encoding proteins with more interactions within the whole human interactome. Genome Biol Evol. 7(4): 1141–54.
23. Chakraborty S. and Alvarez-Ponce D. 2016. Positive Selection and Centrality in the Yeast and Fly Protein-Protein Interaction Networks. Biomed Res Int. 4658506.
24. Yang Z. 2002. Inference of selection from multiple species alignments. Curr

Opin Genet Dev. 12(6): 688-94.

25. Zhang J., Nielsen R. and Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22(12): 2472–9.
26. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24(8): 1586–91.
27. Yang Z. and Dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. Mol Biol Evol. 28(3): 1217–28.
28. R Development Core Team. 2016. R: A Language and Environment for Statistical Computing. R Found Stat Comput Vienna Austria.
29. Supek F., Bošnjak M., Škunca N. and Šmuc T. 2011. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 6(7).
30. Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K. P., Kuhn M., Bork P., Jensen Lars J. and Von Mering C. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43(D1): D447–52.
31. Anisimova M. and Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 24(5): 1219–28.
32. Rhee S. Y., Wood V., Dolinski K. and Draghici S. 2008. Use and misuse of the gene ontology annotations. Nat Rev Genet. 9(7): 509–15.
33. Huntley R. P., Sawford T., Martin M. J. and O'Donovan C. 2014. Understanding how and why the Gene Ontology and its annotations evolve: the GO within UniProt. Gigascience. 3(1): 4.
34. Fang X., Seim I., Huang Z., Gerashchenko M. V., Xiong Z., Turanov A. A., Zhu Y., Lobanov A. V., Fan D., Yim S. H., Yao X., Ma S., Yang L., Lee S. G., Kim E. B., Bronson R. T., Šumbera R., Buffenstein R., Zhou X., Krogh A., Park T. J., Zhang G., Wang J. and Gladyshev V. N. 2014. Adaptations to a Subterranean Environment and Longevity Revealed by the Analysis of Mole Rat Genomes. Cell Rep. 8(5): 1354–64.
35. Davies K. T. J, Bennett N. C., Tsagkogeorga G., Rossiter S. J. and Faulkes C. G. 2015, Family wide molecular adaptations to underground life in African

mole-rats revealed by phylogenomic analysis. Mol Biol Evol. 32(12): 3089–107.

36. Kim B., Kang S., Ahn D., Kim J., Ahn I., Lee W., Cho J., Min G. and Park H. 2017. First Insights into the Subterranean Crustacean Bathynellacea Transcriptome : Transcriptionally Reduced Opsin Repertoire and Evidence of Conserved Homeostasis Regulatory Mechanisms. PLoS One. 1–22.
37. Shim C., Kwon H.B. and Kim K. 1996. Differential expression of laminin chain-specific mRNA transcripts during mouse preimplantation embryo development. Mol Reprod Dev. 44(1): 44–55.
38. Mohun S. M., Davies W. L., Bowmaker J. K., Pisani D., Himstedt W., Gower D. J., Hunt D. M. and Wilkinson M. 2010. Identification and characterization of visual pigments in caecilians (Amphibia: Gymnophiona), an order of
39. LeGates T. A., Fernandez D. C. and Hattar S. 2014. Light as a central modulator of circadian rhythms, sleep and affect. Nat Rev Neurosci. 15(7): 443–54.
40. Davalli P., Mitic T., Caporali A., Lauriola A. and D'Arca D. 2016. ROS, Cell Senescence, and Novel Molecular Mechanisms in Aging and Age-Related Diseases. Oxid Med Cell Longev. 3565127.
41. Bhosale G., Sharpe J. A., Sundier S. Y. and Duchen M. R. 2015. Calcium signaling as a mediator of cell energy demand and a trigger to cell death. Ann N Y Acad Sci. 1350(1): 107–16.
42. Lodish H., Berk A., Zipursky S., Matsudaira P., Baltimore D. and Darnell J. 2000. Molecular Cell Biology. (Section 22.3: Collagen: The fibrous Proteins of the Matrix). New York: W. H. Freeman.
43. Zylberberg L., Castanet J. and De Ricqles A. 1980. Structure of the dermal scales in gymnophiona (Amphibia). J Morphol. 165(1): 41–54.
44. Kupfer A., Müller H., Antoniazzi M. M., Jared C., Greven H, Nussbaum R. A. and Wilkinson M. 2006. Parental investment by skin feeding in a caecilian amphibian. Nature. 440(7086): 926–9.
45. 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(4): 358–61.

46. Hemler M.E. 2005. Tetraspanin functions and associated microdomains. Nat Rev Mol Cell Biol. 6: 801–11.
47. Inoue S., Kondo S., Parichy D. M. and Watanabe M. 2914. Tetraspanin 3c requirement for pigment cell interactions and boundary formation in zebrafish adult pigment stripes. Pigment Cell Melanoma Res. 27(2): 190–200.
48. Wollenberg K. C. and Measey G. J. 2009. Why colour in subterranean vertebrates? Exploring the evolution of colour patterns in caecilian amphibians. J Evol Biol. 22(5): 1046–56.
49. Sirisinha S. 2014. Evolutionary insights into the origin of innate and adaptive immune systems: Different shades of grey. Asian Pac J Allergy Immunol. 32(1): 3-15
50. Carey C., Cohen N. and Rollins-Smith L. 1999. Amphibian declines: An immunological perspective. Dev Comp Immunol. 23(6): 459–72.
51. Collins J. P. and Storfer A. 2003. Global amphibian declines: Sorting the hypotheses. Diversity and Distributions. 9(2): 89–98.
52. Aris-Brosou S. 2005. Determinants of adaptive evolution at the molecular level: The extended complexity hypothesis. Mol Biol Evol. 22(2): 200–9.

Tables and Figures

Table 1

Number of genes under positive selection

Foreground branch	Branch number	Genes under positive selection (FDR < 10%)	Genes with description	Genes with GO	Biological process domains	Molecular function domains	Cellular component domains
Gymnophiona	1	50	46	43	96	84	75
Teresomata	2	8	8	7	13	16	16
Rhinatrema bivittatum	3	17	17	15	31	29	22
Microcaecilia	4	13	12	11	34	33	19
Caecilia+ Typhlonectes	5	15	14	15	28	35	19
Microcaecilia dermatophaga	6	33	30	31	74	72	44
Microcaecilia unicolor	7	16	15	15	48	56	28
Typhlonectes compressicauda	8	18	17	16	34	32	27
Caecilia tentaculata	9	7	6	7	23	15	16

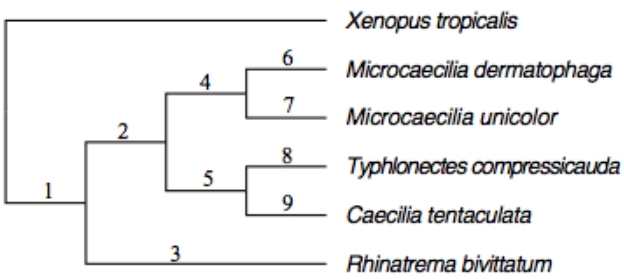


Figure 1

Phylogenetic tree used in the tests of positive selection. Branches used as foreground branches in the different tests are indicated with numbers as follows: 1: Gymnophiona branch, 2: Teresomata branch, 3: *R. bivittatum* branch, 4: *Microcaecilia* branch, 5: *Caecilia+Typhlonectes* branch, 6: *M. dermatophaga* branch, 7: *M. unicolor* branch, 8: *T. compressicauda* branch and 9: *C. tentaculata* branch. Phylogeny based on chapter 1 and literature.

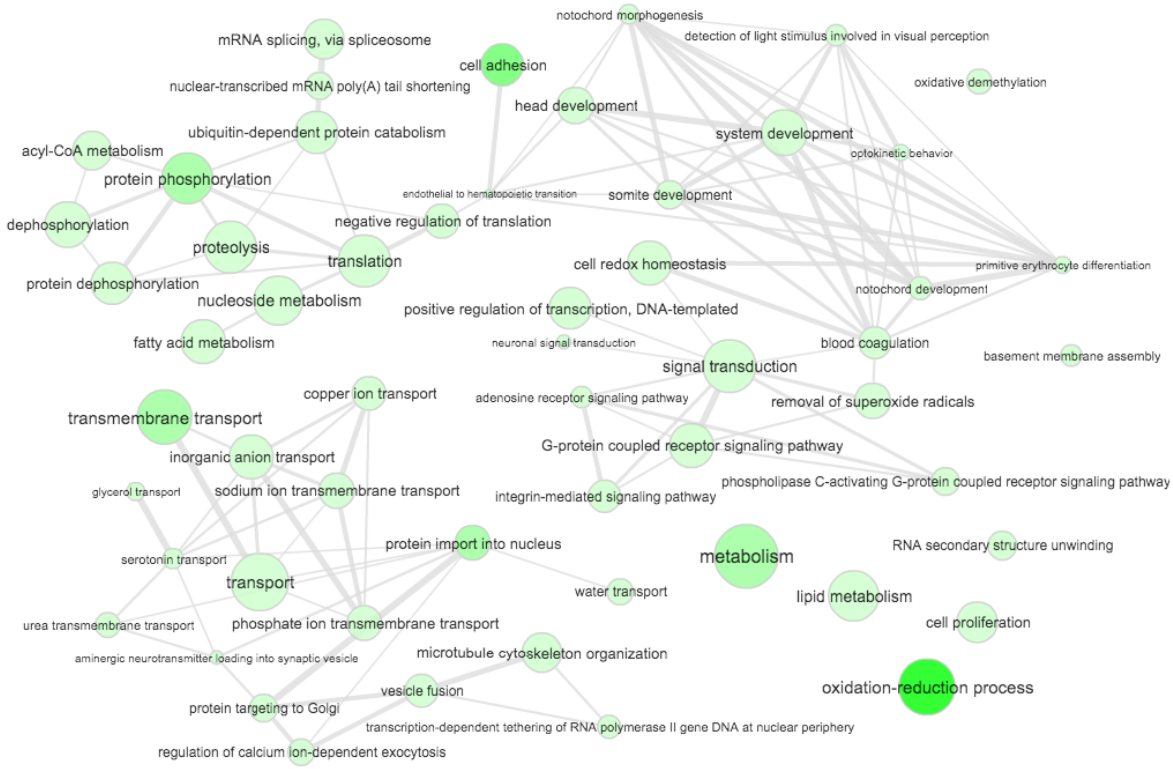


Figure 2

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the Gymnophiona branch (branch 1 in Figure 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

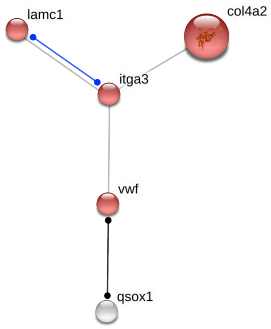


Figure 3

Protein-protein interaction (PPi) network predicted from the positive selected genes of the Gymnophiona branch (branch 1) that are involved in the ECM-receptor interaction pathway with a binding interaction (blue line) between *lamc1* and *itga3*, and a reaction interaction (black line) between *vwf* and *qsox1* (this last protein-coding gene is a second shell of interactions).

Supplementary material

Table S1

Description of genes inferred to have been under positive selection in caecilian evolution (ω* for sites under positive selection on the foreground branch, 2a and 2b sites).

Foreground branch	Gene symbol	Gene description	GO terms	*Xenopus* gene ID	ω*	FDR P-value
Gymnophiona	***acot2***	acyl-CoA thioesterase 2	GO:0005737; GO:0006631; GO:0006637; GO:0016790; GO:0047617	ENSXETG00000000057	107.57442	0.081956824
Gymnophiona	***wdr1***	WD repeat domain 1	-	ENSXETG00000000629	999	0.060673056
Gymnophiona	***slc34a2***	solute carrier family 34 member 2	GO:0005436; GO:0005737; GO:0005886; GO:0005903; GO:0015321; GO:0016020; GO:0016021; GO:0016324; GO:0030643; GO:0031982; GO:0035435; GO:0035725; GO:0044341	ENSXETG00000000954	8.33671	0.081956824
Gymnophiona	***sod3***	superoxide dismutase 3	GO:0004784; GO:0005507; GO:0005615; GO:0005737; GO:0006801; GO:0008270; GO:0016491; GO:0019430; GO:0046872; GO:0055114	ENSXETG00000002122	143.96716	0.075170199
Gymnophiona	***col4a2***	collagen type IV alpha 2 chain	GO:0005201; GO:0005576; GO:0005578; GO:0005581; GO:0005604	ENSXETG00000002635	98.85225	0.082470139
Gymnophiona	***akr1a1***	aldo-keto reductase family 1 member A1	GO:0008106; GO:0016491; GO:0055114	ENSXETG00000003499	300.07976	0.060673056
Gymnophiona	***als2cl***	ALS2 C-terminal like	-	ENSXETG00000003686	54.6686	0.050457662
Gymnophiona	***nup155***	nucleoporin 155kDa	GO:0000972; GO:0005643; GO:0006405; GO:0006606; GO:0006913; GO:0017056; GO:0036228; GO:0044611	ENSXETG00000004785	40.12923	0.081956824
Gymnophiona	***c10orf35***	chromosome 10 open reading frame 35	GO:0016020; GO:0016021	ENSXETG00000006461	418.15868	0.082538498
Gymnophiona	***ddx17***	DEAD-box helicase 17	GO:0000166; GO:0003676; GO:0004004; GO:0004386; GO:0005524; GO:0010501; GO:0016787; GO:0045893	ENSXETG00000006900	999	0.035291841

Gymnophiona	***adamts7***	ADAM metallopeptidase with thrombospondin type 1 motif 7	GO:0004222; GO:0005578; GO:0006508; GO:0008233; GO:0008237; GO:0008270; GO:0031012; GO:0046872	ENSXETG00000007838	22.57471	0.045769604
Gymnophiona	**-**	uncharacterised	GO:0008146; GO:0016740	ENSXETG00000009265	87.87366	0.050457662
	nckipsd	NCK interacting protein with SH3 domain	-	ENSXETG00000009319	59.17284	0.088559456
Gymnophiona	***esyt1***	extended synaptotagmin-like protein 1	GO:0008289; GO:0016020; GO:0016021; GO:0031227; GO:0044232	ENSXETG00000009481	55.12752	0.021371019
Gymnophiona	***msn***	moesin	GO:0003779; GO:0005737; GO:0005856; GO:0008092; GO:0019898	ENSXETG00000009770	4.80253	0.075170199
Gymnophiona	***aqp9***	aquaporin 9	GO:0005215; GO:0005372; GO:0006810; GO:0006833; GO:0015105; GO:0015168; GO:0015204; GO:0015698; GO:0015793; GO:0016020; GO:0016021; GO:0071918	ENSXETG00000010861	999	0.03351781
Gymnophiona	***slc22a31***	solute carrier family 22 member 31	GO:0008514; GO:0015711; GO:0016020; GO:0016021; GO:0022857; GO:0055085	ENSXETG00000011276	178.82387	0.071653906
Gymnophiona	***rph3a***	rabphilin 3A	GO:0005509; GO:0005544; GO:0005886; GO:0006886; GO:0006906; GO:0016020; GO:0017137; GO:0017158; GO:0019905; GO:0030276; GO:0046872; GO:0048791; GO:0070382; GO:0098793	ENSXETG00000011467	999	3.09E-36
Gymnophiona	***lamc1***	laminin subunit gamma 1	GO:0001654; GO:0005604; GO:0007411; GO:0007420; GO:0007517; GO:0007519; GO:0007634; GO:0030903; GO:0031290; GO:0048570; GO:0048731; GO:0048854; GO:0050908; GO:0061053; GO:0070831	ENSXETG00000012525	23.15197	0.060673056
Gymnophiona	***tet2***	tet methylcytosine dioxygenase 2	GO:0030097; GO:0030099; GO:0030218; GO:0060319; GO:0070989; GO:0098508	ENSXETG00000014101	100.54979	0.081956824
Gymnophiona	***gstcd***	glutathione S-transferase C-terminal domain	GO:0005737	ENSXETG00000014108	131.16589	0.081956824
Gymnophiona	***nup153***	nucleoporin 153kDa	GO:0001525; GO:0005487; GO:0005622; GO:0006405; GO:0006606; GO:0008139; GO:0008270; GO:0017056;	ENSXETG00000014197	999	0.045769604

			GO:0046872			
Gymnophiona	***gdpd5***	glycerophosphodiester phosphodiesterase domain containing 5	GO:0006629; GO:0008081; GO:0008889; GO:0016020; GO:0016021	ENSXETG00000015053	31.09229	0.082470139
Gymnophiona	***tacc2***	transforming acidic coiled-coil containing protein 2	GO:0000226; GO:0005737; GO:0008283; GO:0015630; GO:0021987	ENSXETG00000015587	103.81618	0.081956824
Gymnophiona	***klhdc10***	kelch domain containing 10	-	ENSXETG00000016301	56.55156	0.066797583
Gymnophiona	***golga1***	golgin A1	GO:0000042; GO:0005794	ENSXETG00000016840	50.29481	0.042081125
Gymnophiona	***pigr***	polymeric immunoglobulin receptor	GO:0016020; GO:0016021	ENSXETG00000017102	54.9727	0.081956824
Gymnophiona	***gigyf1***	GRB10 interacting GYF protein 1	-	ENSXETG00000018415	63.29151	0.000103512
Gymnophiona	***cul9***	cullin 9	GO:0006511; GO:0008270; GO:0031625; GO:0046872	ENSXETG00000018504	47.26087	0.071653906
Gymnophiona	***cdhr2***	cadherin related family member 2	GO:0005509; GO:0005886; GO:0007155; GO:0007156; GO:0016020; GO:0016021	ENSXETG00000019629	52.04677	0.00251902
Gymnophiona	***hprt1***	hypoxanthine phosphoribosyltransferase 1	GO:0009116	ENSXETG00000019768	19.00688	0.081956824
Gymnophiona	***cgn***	cingulin	GO:0003774; GO:0016459	ENSXETG00000020726	39.28213	0.045769604
Gymnophiona	***itga3***	integrin subunit alpha 3	GO:0007155; GO:0007229; GO:0008305; GO:0016020; GO:0016021	ENSXETG00000021920	32.10395	0.045769604
Gymnophiona	***p2ry11***	purinergic receptor P2Y G-protein coupled 11	GO:0001973; GO:0004871; GO:0004930; GO:0007165; GO:0007186; GO:0007200; GO:0016020; GO:0016021; GO:0023041; GO:0035589; GO:0045028; GO:0045031	ENSXETG00000022059	40.28705	0.045769604
Gymnophiona	***ptprh***	protein tyrosine phosphatase receptor type H	GO:0004725; GO:0005001; GO:0006470; GO:0016020; GO:0016021; GO:0016311; GO:0016791; GO:0035335	ENSXETG00000022920	184.64403	0.021371019
Gymnophiona	**SPEN**	spen family transcriptional repressor	GO:0000166; GO:0000398; GO:0003676; GO:0005634	ENSXETG00000023114	64.05463	0.081956824
Gymnophiona	***qsox1***	quiescin sulfhydryl oxidase 1	GO:0003756; GO:0005615; GO:0016020; GO:0016021; GO:0016491; GO:0016971; GO:0016972; GO:0030173; GO:0045454; GO:0055114	ENSXETG00000023156	105.32195	0.021371019
Gymnophiona	***vwf***	von Willebrand	GO:0005578; GO:0007155;	ENSXETG00000023591	999	0.060673056

		factor	GO:0007596; GO:0007599			
Gymnophiona	***cdk12***	cyclin-dependent kinase 12	GO:0004672; GO:0005524; GO:0006468	ENSXETG00000023695	4.32061	0.060673056
Gymnophiona	***tbrg4***	transforming growth factor beta regulator 4	GO:0004672; GO:0006468	ENSXETG00000023866	999	0.057479166
Gymnophiona	***tcf19***	transcription factor 19	-	ENSXETG00000024079	999	0.081956824
Gymnophiona	**SPG11**	spatacsin vesicle trafficking associated	GO:0007399; GO:0007409	ENSXETG00000025297	93.96152	0.077360076
Gymnophiona	***rps13***	ribosomal protein S13	GO:0003735; GO:0005730; GO:0005840; GO:0006412; GO:0022627; GO:0070181	ENSXETG00000026454	999	0.081956824
Gymnophiona	***gsto2***	glutathione S-transferase omega 2	GO:0004364; GO:0005737; GO:0008152	ENSXETG00000026602	204.42538	0.082470139
Gymnophiona	**TNRC6A**	trinucleotide repeat containing 6A	GO:0000166; GO:0000289; GO:0003676; GO:0017148; GO:0035278	ENSXETG00000030437	7.70315	0.069829147
Gymnophiona	**CP**	ceruloplasmin	GO:0004322; GO:0005507; GO:0005634; GO:0005737; GO:0006825; GO:0006879; GO:0016491; GO:0046872; GO:0055114	ENSXETG00000031159	61.27401	0.082470139
Gymnophiona	**-**	uncharacterised	GO:0003995; GO:0008152; GO:0016491; GO:0016627; GO:0050660; GO:0055114; GO:0060322	ENSXETG00000031271	87.2058	0.001074924
Gymnophiona	**-**	uncharacterised	-	ENSXETG00000033245	999	0.077360076
Gymnophiona	**COL17A1**	collagen type XVII alpha 1 chain	GO:0005578; GO:0016020; GO:0016021	ENSXETG00000033563	136.30841	0.009786183
Gymnophiona	**-**	uncharacterised	GO:0005887; GO:0006837; GO:0015222; GO:0015842; GO:0016020; GO:0016021; GO:0055085; GO:0098793	ENSXETG00000033569	139.35786	0.045769604
Teresomata	***fam3b***	family with sequence similarity 3 member B	-	ENSXETG00000005180	998.98757	0.058383372
Teresomata	***aoc3***	amine oxidase copper containing 3	GO:0005507; GO:0007601; GO:0008131; GO:0009308; GO:0016020; GO:0016021; GO:0016491; GO:0046872; GO:0048038; GO:0055114	ENSXETG00000012588	998.99982	0.068228787
Teresomata	***mbd5***	methyl-CpG binding domain	GO:0003677; GO:0003682; GO:0005634; GO:0010369	ENSXETG00000018214	998.99978	0.049582247

		protein 5				
Teresomata	***hgs***	hepatocyte growth factor-regulated tyrosine kinase substrate	GO:0005622; GO:0006886; GO:0046872	ENSXETG00000019701	999	0.039962285
Teresomata	***masp1***	mannan-binding lectin serine peptidase 1	GO:0001755; GO:0001867; GO:0004252; GO:0005509; GO:0005576; GO:0005615; GO:0005737; GO:0006508; GO:0006956; GO:0008233; GO:0008236; GO:0016787; GO:0046872	ENSXETG00000019757	482.51659	2.34E-22
Teresomata	***pcdh7***	protocadherin 7	GO:0005509; GO:0005886; GO:0007155; GO:0007156; GO:0016020; GO:0016021	ENSXETG00000022281	999	0.032846913
Teresomata	***tnc***	tenascin C	GO:0007155; GO:0031012; GO:0042127	ENSXETG00000023938	557.19204	3.94E-05
Teresomata	***sypl1***	synaptophysin-like protein 1	GO:0005215; GO:0006810; GO:0008021; GO:0016020; GO:0016021; GO:0030285	ENSXETG00000025677	999	0.039962285
Rhinatrema bivittatum	***plpp1***	phospholipid phosphatase 1	GO:0005886; GO:0005887; GO:0006629; GO:0006644; GO:0007165; GO:0008195; GO:0016020; GO:0016021; GO:0016311; GO:0042577; GO:0046839	ENSXETG00000000375	998.96886	0.087167945
Rhinatrema bivittatum	***mtg2***	mitochondrial ribosome-associated GTPase 2	GO:0000287; GO:0003924; GO:0005525	ENSXETG00000002001	38.67929	0.034033894
Rhinatrema bivittatum	***clic3***	chloride intracellular channel 3	GO:0005254; GO:0006821; GO:1902476	ENSXETG00000003974	998.98992	0.05538756
Rhinatrema bivittatum	***cenpa***	centromere protein A	GO:0000775; GO:0000776; GO:0000777; GO:0000786; GO:0003677; GO:0005634; GO:0005694; GO:0046982	ENSXETG00000005197	999	0.05538756
Rhinatrema bivittatum	***atp1a2***	ATPase Na+/K+ transporting alpha 2 polypeptide	GO:0000166; GO:0001947; GO:0001966; GO:0005391; GO:0005524; GO:0005623; GO:0006810; GO:0006811; GO:0006813; GO:0006814; GO:0007368; GO:0007507; GO:0007519; GO:0010084; GO:0010248; GO:0016020; GO:0016021; GO:0016787; GO:0042044; GO:0046872; GO:0051480; GO:0060047; GO:0061371; GO:0090662	ENSXETG00000008125	48.43624	0.079634778

Rhinatrema bivittatum	***rcn1***	reticulocalbin 1	GO:0005509	ENSXETG00000008174	58.64572	0.087167945
Rhinatrema bivittatum	***nuf2***	NUF2; NDC80 kinetochore complex component	GO:0000775; GO:0007067; GO:0007507	ENSXETG00000010463	11.1057	0.087167945
Rhinatrema bivittatum	**COL5A2**	collagen type V alpha 2 chain	GO:0005201	ENSXETG00000010784	259.99922	0.011840021
Rhinatrema bivittatum	***rpl13a***	ribosomal protein L13a	GO:0003729; GO:0003735; GO:0005840; GO:0006412; GO:0015934; GO:0022625; GO:0030529	ENSXETG00000014144	5.33779	7.96E-15
Rhinatrema bivittatum	***rcc2***	regulator of chromosome condensation 2	GO:0001755	ENSXETG00000014793	998.99973	0.032134195
Rhinatrema bivittatum	***anxa2***	annexin A2	GO:0004859; GO:0005509; GO:0005544; GO:0008092; GO:0043086	ENSXETG00000015289	83.87076	0.08170457
Rhinatrema bivittatum	***anxa6***	annexin A6	GO:0001778; GO:0005509; GO:0005544	ENSXETG00000015832	23.63599	0.067225054
Rhinatrema bivittatum	***hdgf***	hepatoma-derived growth factor	-	ENSXETG00000018516	99.10972	0.072192303
Rhinatrema bivittatum	***yipf1***	Yip1 domain family member 1	GO:0005794; GO:0016020; GO:0016021; GO:0017137	ENSXETG00000019983	999	0.000519394
Rhinatrema bivittatum	***tbc1d31***	TBC1 domain family member 31	-	ENSXETG00000023189	310.36313	0.037596663
Rhinatrema bivittatum	***parp14.2***	poly (ADP-ribose) polymerase family member 14 gene 2	GO:0000166; GO:0003676; GO:0003950; GO:0016740; GO:0016757	ENSXETG00000023399	827.88309	0.08170457
Rhinatrema bivittatum	***tnc***	tenascin C	GO:0007155; GO:0031012; GO:0042127	ENSXETG00000023938	105.06599	0.064561005
Caecilia + Typhlonectes	***aqp3***	aquaporin 3	GO:0005215; GO:0006810; GO:0016020; GO:0016021	ENSXETG00000002151	998.99918	0.094399642
Caecilia+ Typhlonectes	***fadd***	Fas associated via death domain	GO:0007165; GO:0042981; GO:0043065	ENSXETG00000003799	52.29035	0.050859898
Caecilia + Typhlonectes	***efemp1***	EGF containing fibulin-like extracellular matrix protein 1	GO:0005006; GO:0005509; GO:0007173; GO:0031012	ENSXETG00000006076	14.04564	0.007001
Caecilia + Typhlonectes	***utp14a***	UTP14A small subunit processome component	GO:0005730; GO:0006364; GO:0030490; GO:0032040	ENSXETG00000007465	22.93783	0.094399642
Caecilia +	***parp9***	poly(ADP-	GO:0003950	ENSXETG00000007985	999	0.094399642

Typhlonectes		ribose) polymerase family member 9				
Caecilia + *Typhlonectes*	***enpp3***	ectonucleotide pyrophosphatase/ phosphodiesterase 3	GO:0003676; GO:0003824; GO:0004528; GO:0004551; GO:0005044; GO:0006898; GO:0006955; GO:0008152; GO:0016020; GO:0016021; GO:0016787; GO:0030247; GO:0046872; GO:0090305	ENSXETG00000008244	998.99977	0.094399642
Caecilia + *Typhlonectes*	***ybx1***	Y-box binding protein 1	GO:0003676; GO:0003677; GO:0003723; GO:0006355; GO:0008190; GO:0045947; GO:0048025; GO:0050686; GO:0051236; GO:1900364	ENSXETG00000013436	999	0.050859898
Caecilia + *Typhlonectes*	***acat1***	acetyl-CoA acetyltransferase 1	GO:0003824; GO:0008152; GO:0016740; GO:0016746; GO:0016747	ENSXETG00000014477	999	0.094399642
Caecilia + *Typhlonectes*	***axl***	AXL receptor tyrosine kinase	GO:0000166; GO:0004672; GO:0004713; GO:0005524; GO:0006468; GO:0016020; GO:0016021; GO:0016301; GO:0016310; GO:0016740; GO:0018108	ENSXETG00000018708	41.86018	0.053497776
Caecilia + *Typhlonectes*	***pfn2***	profilin 2	GO:0003779; GO:0030036; GO:0030833	ENSXETG00000020090	999	6.92E-05
Caecilia + *Typhlonectes*	***exog***	endo/ exonuclease (5'-3'); endonuclease G-like	GO:0003676; GO:0016787; GO:0046872	ENSXETG00000021000	999	0.094399642
Caecilia + *Typhlonectes*	***tmem27***	transmembrane protein 27	GO:0006508; GO:0008237; GO:0008241; GO:0016020; GO:0016021	ENSXETG00000022466	999	0.01469052
Caecilia + *Typhlonectes*	***scarb2***	scavenger receptor class B member 2	GO:0004872; GO:0005764; GO:0016020; GO:0016021	ENSXETG00000024116	53.04016	0.004245511
Caecilia + *Typhlonectes*	***itgb1***	integrin subunit beta 1	GO:0004872; GO:0007155; GO:0007160; GO:0007229; GO:0008305; GO:0016020; GO:0016021	ENSXETG00000026716	376.16418	0.006565164
Caecilia + *Typhlonectes*	**-**	uncharacterised	GO:0016020; GO:0016021; GO:0046983; GO:0061588	ENSXETG00000031447	217.23189	0.094399642
Caecilia tentaculata	***dsc3***	desmocollin 3	GO:0002159; GO:0005509; GO:0005886; GO:0007155; GO:0007156; GO:0007507; GO:0016020; GO:0016021; GO:0055113; GO:0060027	ENSXETG00000004721	21.38483	0.019034822
Caecilia	**-**	uncharacterised	GO:0004134; GO:0004135;	ENSXETG00000013185	11.74808	0.011498941

tentaculata			GO:0003824; GO:0005978; GO:0005980			
Caecilia tentaculata	***ppil4***	peptidylprolyl isomerase like 4	GO:0000166; GO:0000413; GO:0003676; GO:0003755; GO:0006457	ENSXETG00000021385	999	0.052032424
Caecilia tentaculata	***pfkp***	phosphofructokinase; platelet	GO:0003872; GO:0005524; GO:0005737; GO:0006002; GO:0006096; GO:0061615	ENSXETG00000021922	998.99901	2.13E-08
Caecilia tentaculata	***tubgcp6***	tubulin gamma complex associated protein 6	GO:0000226; GO:0000922; GO:0000923; GO:0005200; GO:0005737; GO:0005813; GO:0005815; GO:0005856; GO:0005874; GO:0007020; GO:0007126; GO:0008274; GO:0031122; GO:0043015; GO:0051011; GO:0051298; GO:0051415; GO:0090307	ENSXETG00000022264	998.99915	0.019034822
Caecilia tentaculata	***man2a1***	mannosidase; alpha class 2A member 1	GO:0000139; GO:0003824; GO:0004553; GO:0004559; GO:0005975; GO:0006013; GO:0006491; GO:0006517; GO:0008270; GO:0015923; GO:0016020; GO:0016021; GO:0030246	ENSXETG00000026530	998.9581	7.40E-13
Caecilia tentaculata	***trmt10c***	tRNA methyltransferase 10C	GO:0005739; GO:0008033	ENSXETG00000029921	16.75471	0.071059442
Typhlonectes compressicauda	***f2***	coagulation factor 2 thrombin	GO:0004252; GO:0005509; GO:0005576; GO:0006508; GO:0007596; GO:0008233; GO:0008236; GO:0016787	ENSXETG00000001982	39.56098	0.082333407
Typhlonectes compressicauda	***col4a1***	collagen type IV alpha 1	GO:0005201; GO:0005576; GO:0005578; GO:0005581; GO:0005604	ENSXETG00000002637	30.19389	0.082333407
Typhlonectes compressicauda	***slc30a10***	solute carrier family 30 member 10	GO:0005385; GO:0005886; GO:0006812; GO:0008324; GO:0010043; GO:0016020; GO:0016021; GO:0055085; GO:0061088; GO:0071577; GO:0098655	ENSXETG00000002721	50.94049	0.026827163
Typhlonectes compressicauda	***camkmt***	calmodulin-lysine N-methyltransferase	GO:0005737; GO:0018022; GO:0018025	ENSXETG00000002754	268.68298	0.095269185
Typhlonectes compressicauda	***klkb1***	kallikrein B1	GO:0004252; GO:0005576; GO:0006508	ENSXETG00000005867	12.67365	0.095269185
Typhlonectes compressicauda	***mios***	missing oocyte meiosis regulator homolog	-	ENSXETG00000007293	889.60098	0.072501743

Typhlonectes compressicauda	***polr2a***	polymerase RNA II	GO:0001055; GO:0003677; GO:0003899; GO:0005665; GO:0006351; GO:0006366; GO:0016740; GO:0016779	ENSXETG00000012465	998.99995	0.000307004
Typhlonectes compressicauda	***prkag3***	protein kinase AMP-activated gamma 3 non-catalytic subunit	GO:0016301; GO:0016310	ENSXETG00000013879	998.99942	0.062065169
Typhlonectes compressicauda	***cwc22***	CWC22 homolog spliceosome-associated protein	GO:0000398; GO:0003723; GO:0071006; GO:0071013	ENSXETG00000014099	114.97026	0.030115536
Typhlonectes compressicauda	***ate1***	arginyltransferase 1	GO:0004057; GO:0005737; GO:0016598; GO:0016740; GO:0016746	ENSXETG00000015591	998.99868	0.035809617
Typhlonectes compressicauda	***myh4***	myosin heavy chain 3 embryonic skeletal muscle	GO:0000166; GO:0003774; GO:0003779; GO:0005524; GO:0016459	ENSXETG00000016248	998.99998	0.003513682
Typhlonectes compressicauda	***thoc5***	THO complex 5	-	ENSXETG00000016419	998.99992	0.095269185
Typhlonectes compressicauda	***arhgap33***	Rho GTPase activating protein 33	GO:0005096; GO:0005938; GO:0007165; GO:0007264; GO:0015629; GO:0035091; GO:0043547	ENSXETG00000017543	999	0.095269185
Typhlonectes compressicauda	***-***	uncharacterised	GO:0001775; GO:0001971; GO:0005576	ENSXETG00000018913	999	0.05271803
Typhlonectes compressicauda	***clcn3***	chloride channel voltage-sensitive 3	GO:0005216; GO:0005247; GO:0005623; GO:0005887; GO:0006810; GO:0006811; GO:0006821; GO:0016020; GO:0016021; GO:0034220; GO:0044070; GO:0045794; GO:0055085; GO:0072320; GO:1902476; GO:1903959	ENSXETG00000023146	998.99997	0.030115536
Typhlonectes compressicauda	***fam13a***	family with sequence similarity 13 member A	GO:0007165	ENSXETG00000023661	273.35518	0.086409921
Typhlonectes compressicauda	**ADGRG6**	adhesion G protein-coupled receptor G6	GO:0004888; GO:0004930; GO:0007166; GO:0007186; GO:0016020; GO:0016021	ENSXETG00000030163	269.96578	0.086409921
Typhlonectes compressicauda	**DSG2**	desmoglein 2	GO:0005509; GO:0005886; GO:0007155; GO:0007156; GO:0016020; GO:0016021	ENSXETG00000034243	16.30646	0.035809617
Microcaecilia	***pinx1***	PIN2/TERF1 interacting telomerase inhibitor 1	GO:0003676; GO:0005730; GO:0010521; GO:0051974	ENSXETG00000000688	109.0809	0.00136397
Microcaecilia	***col4a2***	collagen; type	GO:0005201; GO:0005576;	ENSXETG00000002635	56.26914	0.052045712

		IV alpha 2	GO:0005578; GO:0005581; GO:0005604			
Microcaecilia	***fam3b***	family with sequence similarity 3 member B	-	ENSXETG00000005180	67.2091	0.002717397
Microcaecilia	***iqsec2***	IQ motif and Sec7 domain 2	GO:0005086; GO:0030036; GO:0032012; GO:0043547	ENSXETG00000007177	1.86049	5.17E-41
Microcaecilia	***ddx24***	DEAD-box helicase 24	GO:0000166; GO:0003676; GO:0004004; GO:0004386; GO:0005524; GO:0010501; GO:0016787	ENSXETG00000010314	70.67374	0.089979786
Microcaecilia	***mrps7***	mitochondrial ribosomal protein S7	GO:0006412	ENSXETG00000012510	999	0.052045712
Microcaecilia	***elovl5***	ELOVL fatty acid elongase 5	GO:0005783; GO:0005789; GO:0006629; GO:0006631; GO:0006633; GO:0006636; GO:0009922; GO:0016020; GO:0016021; GO:0016740; GO:0019367; GO:0019368; GO:0030425; GO:0030497; GO:0042759; GO:0042761; GO:0042995; GO:0043025; GO:0097447; GO:0102336; GO:0102337; GO:0102338	ENSXETG00000015994	742.62964	0.058323282
Microcaecilia	***ca5b***	mitochondrial carbonic anhydrase VB	GO:0004089; GO:0005739; GO:0006730; GO:0008270; GO:0046872; GO:2000021	ENSXETG00000016594	95.57838	0.00136397
Microcaecilia	***yes1***	YES proto-oncogene 1 Src family tyrosine kinase	GO:0000166; GO:0004672; GO:0004713; GO:0004715; GO:0005102; GO:0005524; GO:0006468; GO:0007169; GO:0016301; GO:0016310; GO:0016477; GO:0016740; GO:0030154; GO:0031234; GO:0034334; GO:0038083; GO:0042127; GO:0045087; GO:0045859; GO:0046777; GO:0060027	ENSXETG00000019176	2.81761	8.07E-05
Microcaecilia	***basp1***	brain abundant membrane attached signal protein 1	-	ENSXETG00000021380	999	0.021733743
Microcaecilia	***tspan36***	tetraspanin 36	GO:0005887; GO:0007166; GO:0016020; GO:0016021; GO:0043473	ENSXETG00000022371	543.60023	0.052045712
Microcaecilia	***acp1***	acid phosphatase 1	GO:0003993; GO:0004725; GO:0004726; GO:0005737; GO:0006470; GO:0035335	ENSXETG00000027987	366.58784	0.052045712

Microcaecilia	**-**	uncharacterised	GO:0004252; GO:0006508; GO:0008233; GO:0008236; GO:0016787	ENSXETG00000033306	998.99948	0.022116997
Microcaecilia unicolor	***col12a1***	collagen type XII alpha 1	GO:0005615	ENSXETG00000003603	136.70977	9.56E-35
Microcaecilia unicolor	***cat 2***	catalase gene 2	GO:0004096; GO:0004601; GO:0005739; GO:0005777; GO:0006979; GO:0016491; GO:0020037; GO:0042542; GO:0042744; GO:0046872; GO:0055114; GO:0098869	ENSXETG00000003981	56.00761	0.000183547
Microcaecilia unicolor	***fabp2***	intestinal fatty acid binding protein 2	GO:0005215; GO:0005504; GO:0006810; GO:0008289	ENSXETG00000004045	999	0.002260627
Microcaecilia unicolor	***lamp2***	lysosomal-associated membrane protein 2	GO:0005764; GO:0005765; GO:0016020; GO:0016021	ENSXETG00000004476	998.9942	0.001628235
Microcaecilia unicolor	***dhx36***	DEAH-box helicase 36	GO:0000166; GO:0003676; GO:0004004; GO:0004386; GO:0005524; GO:0005737; GO:0006396; GO:0008026; GO:0016787	ENSXETG00000007768	27.42023	0.067761385
Microcaecilia unicolor	***sptlc3***	serine palmitoyltransferase long chain base subunit 3	GO:0003824; GO:0008152; GO:0009058; GO:0016020; GO:0016021; GO:0016740; GO:0030170	ENSXETG00000008083	296.87483	0.002260627
Microcaecilia unicolor	***erbb3***	erb-b2 receptor tyrosine kinase 3	GO:0000166; GO:0004672; GO:0004714; GO:0004716; GO:0005524; GO:0005622; GO:0006468; GO:0007169; GO:0016020; GO:0016021; GO:0018108; GO:0023014; GO:0035556	ENSXETG00000009463	15.59687	0.087485616
Microcaecilia unicolor	***pih1d2***	PIH1 domain containing 2		ENSXETG00000010194	658.6695	0.01025015
Microcaecilia unicolor	**COL5A2**	collagen type V alpha 2 chain	GO:0005201	ENSXETG00000010784	999	0.014683217
Microcaecilia unicolor	***tarbp2***	TAR RNA binding protein 2	GO:0003723; GO:0003725; GO:0005737; GO:0006417; GO:0016442; GO:0030422; GO:0030423; GO:0031047; GO:0031054; GO:0035197; GO:0035198; GO:0035280; GO:0042803; GO:0046782	ENSXETG00000012644	999	0.022252093
Microcaecilia unicolor	***cyp17a1***	cytochrome P450 family 17 subfamily A member 1	GO:0004497; GO:0004508; GO:0005506; GO:0006694; GO:0007548; GO:0016491; GO:0016705; GO:0020037; GO:0042448; GO:0046872;	ENSXETG00000015229	42.68226	0.061514343

			GO:0047006; GO:0047442; GO:0055114; GO:1903449			
Microcaecilia unicolor	***cybrd1***	cytochrome b reductase 1	GO:0000293; GO:0010039; GO:0016020; GO:0016021; GO:0031526; GO:0055114	ENSXETG00000018825	999	0.024919723
Microcaecilia unicolor	***fyn***	FYN proto-oncogene Src family tyrosine kinase	GO:0000166; GO:0004672; GO:0004713; GO:0004715; GO:0005102; GO:0005524; GO:0006468; GO:0007169; GO:0016301; GO:0016310; GO:0016477; GO:0016740; GO:0030154; GO:0031234; GO:0038083; GO:0042127; GO:0045087	ENSXETG00000021344	840.37469	1.65E-06
Microcaecilia unicolor	***srpk3***	SRSF protein kinase 3	GO:0000245; GO:0004672; GO:0004674; GO:0005524; GO:0005634; GO:0005737; GO:0006468; GO:0035556; GO:0050684	ENSXETG00000023173	998.99805	1.62E-08
Microcaecilia unicolor	***stx3***	syntaxin 3	GO:0000149; GO:0005484; GO:0005622; GO:0005886; GO:0006886; GO:0006887; GO:0008021; GO:0016020; GO:0016021; GO:0016081; GO:0016192; GO:0031201; GO:0031629; GO:0048278; GO:0061025; GO:0098793	ENSXETG00000023730	176.41269	0.002516965
Microcaecilia unicolor	**-**	uncharacterised	GO:0004252; GO:0006508; GO:0008233; GO:0008236; GO:0016787	ENSXETG00000033306	46.69779	0.024919723
Microcaecilia dermatophaga	**-**	uncharacterised	GO:0004252; GO:0006508; GO:0008233; GO:0008236; GO:0016020; GO:0016021; GO:0016787	ENSXETG00000000063	58.0989	0.093774823
Microcaecilia dermatophaga	**-**	uncharacterised	GO:0005768; GO:0030100	ENSXETG00000000295	477.93978	0.094598672
Microcaecilia dermatophaga	***dnajc21***	DnaJ heat shock protein family	GO:0003676; GO:0008270; GO:0046872	ENSXETG00000000706	82.97368	0.081176472
Microcaecilia dermatophaga	***mrc1***	mannose receptor C type 1	GO:0004888; GO:0005887; GO:0007165; GO:0016020; GO:0016021	ENSXETG00000001366	31.14407	0.008226138
Microcaecilia dermatophaga	***hsdl2***	hydroxysteroid dehydrogenase like 2	-	ENSXETG00000002228	149.55633	0.027606049
Microcaecilia dermatophaga	***mmp2***	matrix metallopeptidase 2	GO:0001945; GO:0004222; GO:0006508; GO:0008233; GO:0008237; GO:0008270; GO:0016787; GO:0031012; GO:0031290; GO:0046872	ENSXETG00000002801	56.88637	0.027606049
Microcaecilia	***lcat***	lecithin-	GO:0006629; GO:0008374	ENSXETG00000003085	262.6812	0.063274157

dermatophaga		cholesterol acyltransferase				
Microcaecilia dermatophaga	***rock1***	Rho-associated coiled-coil containing protein kinase 1	GO:0000166; GO:0004672; GO:0004674; GO:0005524; GO:0005622; GO:0006468; GO:0007266; GO:0016301; GO:0016310; GO:0016740; GO:0017048; GO:0017049; GO:0030036; GO:0035556; GO:0046872; GO:0051492; GO:0051493; GO:2000114	ENSXETG00000003151	19.86355	0.094598672
Microcaecilia dermatophaga	***tead4***	TEA domain family member 4	GO:0001085; GO:0003677; GO:0003700; GO:0005634; GO:0005667; GO:0006351; GO:0006355; GO:0035329; GO:0043565; GO:0044212; GO:0045944; GO:0048568	ENSXETG00000003395	37.88286	0.013817142
Microcaecilia dermatophaga	***col12a1***	collagen type XII alpha 1	GO:0005615	ENSXETG00000003603	16.58674	0.072180873
Microcaecilia dermatophaga	***cyp8b1***	cytochrome P450 family 8 subfamily B member 1	GO:0004497; GO:0005506; GO:0005783; GO:0005789; GO:0008397; GO:0016020; GO:0016021; GO:0016491; GO:0016705; GO:0020037; GO:0046872; GO:0055114	ENSXETG00000006173	4.79268	0.027234823
Microcaecilia dermatophaga	***adamts13***	ADAM metallopeptidase with thrombospondin type 1 motif 13	GO:0004222; GO:0005578; GO:0006508; GO:0008237; GO:0008270; GO:0031012; GO:0046872	ENSXETG00000006882	60.36062	0.073264625
Microcaecilia dermatophaga	***-***	uncharacterised	GO:0008168; GO:0032259	ENSXETG00000008551	999	0.094598672
Microcaecilia dermatophaga	***pdgfd***	platelet derived growth factor D	GO:0005161; GO:0005615; GO:0007596; GO:0008083; GO:0008284; GO:0014068; GO:0016020; GO:0030335; GO:0031954; GO:0043406; GO:0048008; GO:0070374	ENSXETG00000010500	999	0.098652291
Microcaecilia dermatophaga	***rad51ap1***	RAD51 associated protein 1	GO:0003690; GO:0003697; GO:0003723; GO:0005634; GO:0006281	ENSXETG00000011389	998.99536	0.080010268
Microcaecilia dermatophaga	***asah1***	N-acylsphingosine amidohydrolase 1	GO:0005764; GO:0006629	ENSXETG00000012463	10.4016	0.001107763
Microcaecilia dermatophaga	***tarbp2***	TAR RNA binding protein 2	GO:0003723; GO:0003725; GO:0005737; GO:0006417; GO:0016442; GO:0030422; GO:0030423; GO:0031047; GO:0031054; GO:0035197;	ENSXETG00000012644	999	0.085852866

			GO:0035198; GO:0035280; GO:0042803; GO:0046782			
Microcaecilia dermatophaga	**LYZ**	lysozyme	GO:0003796	ENSXETG00000013041	34.81264	0.025923145
Microcaecilia dermatophaga	***cfp***	complement factor properdin	-	ENSXETG00000013748	999	0.092276399
Microcaecilia dermatophaga	***tm2d2***	TM2 domain containing 2	GO:0016020; GO:0016021	ENSXETG00000015155	383.93701	0.094598672
Microcaecilia dermatophaga	***trip11***	thyroid hormone receptor interactor 11	GO:0000042; GO:0005622	ENSXETG00000015833	94.95153	0.032016564
Microcaecilia dermatophaga	***cers6***	ceramide synthase 6	GO:0003677; GO:0005634; GO:0005783; GO:0016020; GO:0016021; GO:0046513; GO:0050291	ENSXETG00000016207	13.45155	0.072180873
Microcaecilia dermatophaga	***golga1***	golgin A1	GO:0000042; GO:0005794	ENSXETG00000016840	392.3178	0.009081857
Microcaecilia dermatophaga	***tspan9***	tetraspanin 9	GO:0005887; GO:0007166; GO:0016020; GO:0016021	ENSXETG00000016985	162.49378	0.080010268
Microcaecilia dermatophaga	***tcf7l2***	transcription factor 7-like 2	GO:0003677; GO:0005634; GO:0005667; GO:0006357; GO:0008013; GO:0016055; GO:0021986; GO:0035462; GO:0043565; GO:0044212; GO:0044333; GO:0060070; GO:0060729; GO:2001237	ENSXETG00000018735	999	1.13E-07
Microcaecilia dermatophaga	***rplp2***	ribosomal protein large P2	GO:0002181; GO:0003735; GO:0005622; GO:0005840; GO:0006414; GO:0022625; GO:0030529; GO:0043009	ENSXETG00000019024	1.90298	0.027234823
Microcaecilia dermatophaga	***aldh1a1***	aldehyde dehydrogenase 1 family member A1	GO:0008152; GO:0016491; GO:0016620; GO:0018479; GO:0055114	ENSXETG00000019615	999	0.027606049
Microcaecilia dermatophaga	***pfkp***	phosphofructokinase platelet	GO:0003872; GO:0005524; GO:0005737; GO:0006002; GO:0006096; GO:0061615	ENSXETG00000021922	323.77006	4.59E-05
Microcaecilia dermatophaga	***fam3c***	family with sequence similarity 3 member C	GO:0005576; GO:0007275	ENSXETG00000022730	108.57494	0.032016564
Microcaecilia dermatophaga	***folr1***	folate receptor 1	GO:0005542; GO:0008517; GO:0015884	ENSXETG00000023968	63.04414	0.094598672
Microcaecilia dermatophaga	***sox17a***	SRY-box 17 alpha	GO:0003677; GO:0005634; GO:0006351; GO:0006355; GO:0007275; GO:0007369; GO:0007492; GO:0008013; GO:0016055; GO:0035469; GO:0043565; GO:0045893;	ENSXETG00000025005	88.54354	1.37E-08

			GO:0045944; GO:0061371; GO:0070121			
Microcaecilia dermatophaga	**NSL1**	MIS12 kinetochore complex component	GO:0000070; GO:0000444	ENSXETG00000030886	228.51919	0.027606049
Microcaecilia dermatophaga	***zcchc2***	zinc finger CCHC domain containing 2	GO:0003676; GO:0008270; GO:0035091	ENSXETG00000032980	999	0.037285469

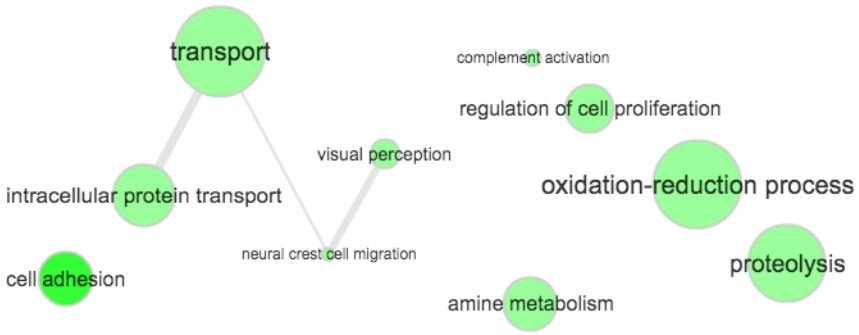


Figure S1

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the Teresomata branch (branch 2 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

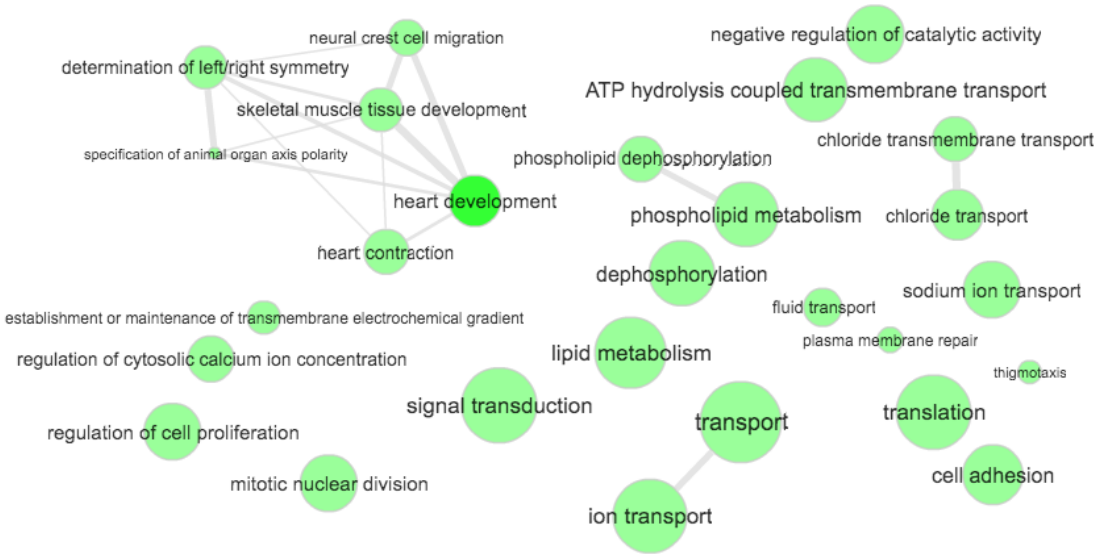


Figure S2

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the *R. bivittatum* branch (branch 3 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

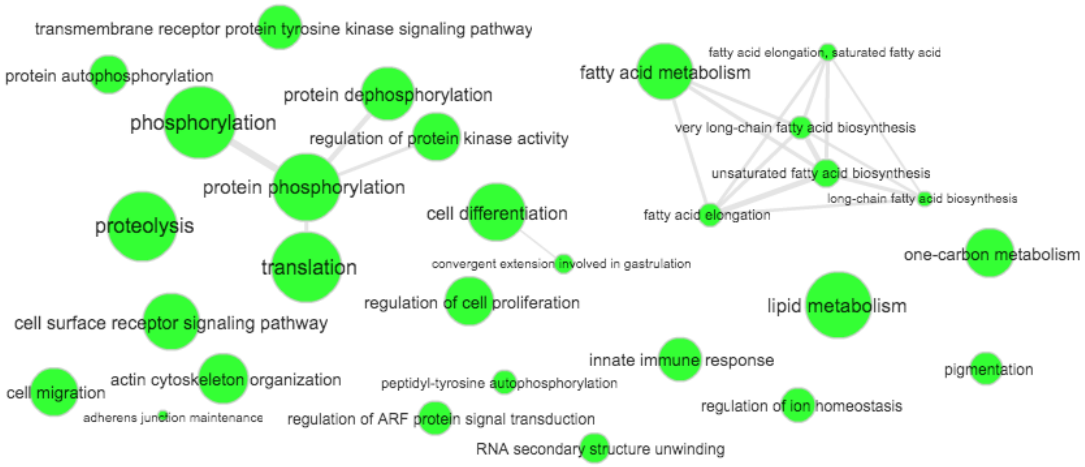


Figure S3

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the *Microcaecilia* branch (branch 4 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

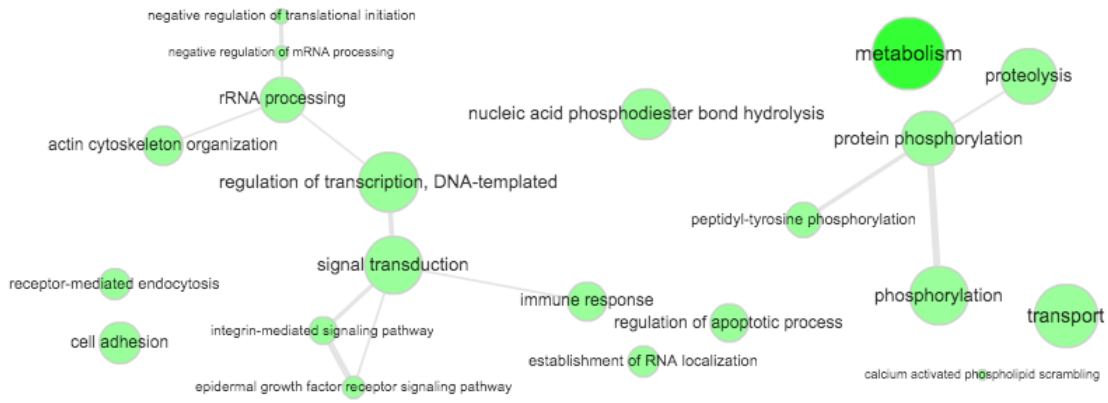


Figure S4

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the *Caecilia+Typhlonectes* branch (branch 5 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

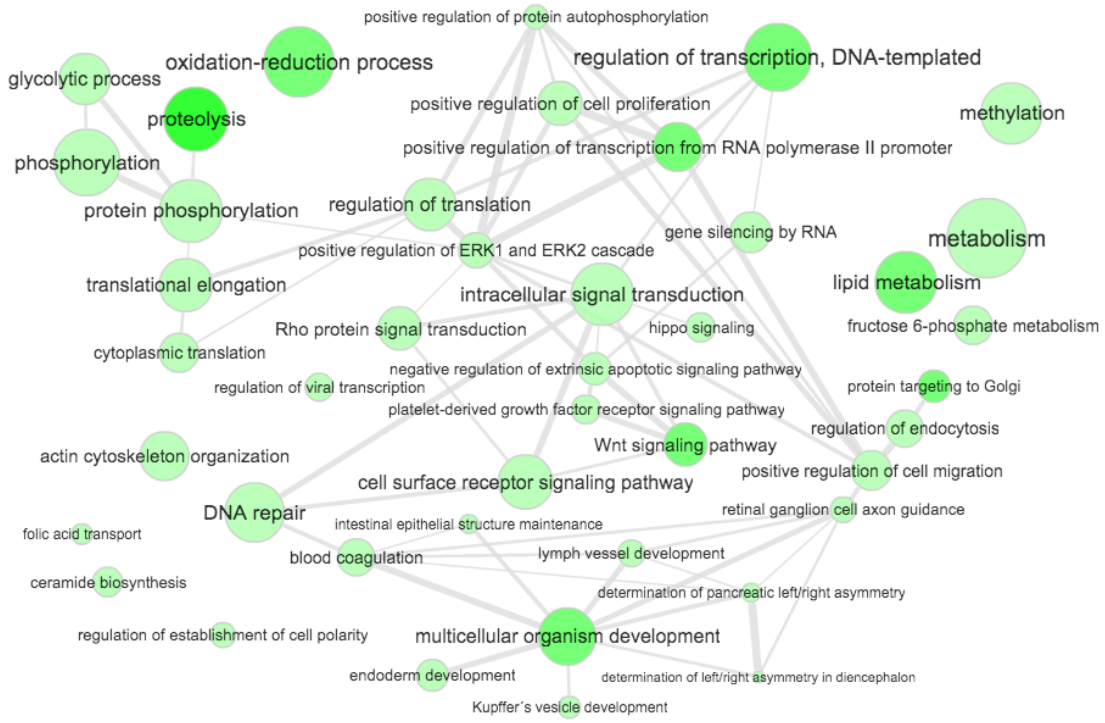


Figure S5

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the *M. dermatophaga* branch (branch 6 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

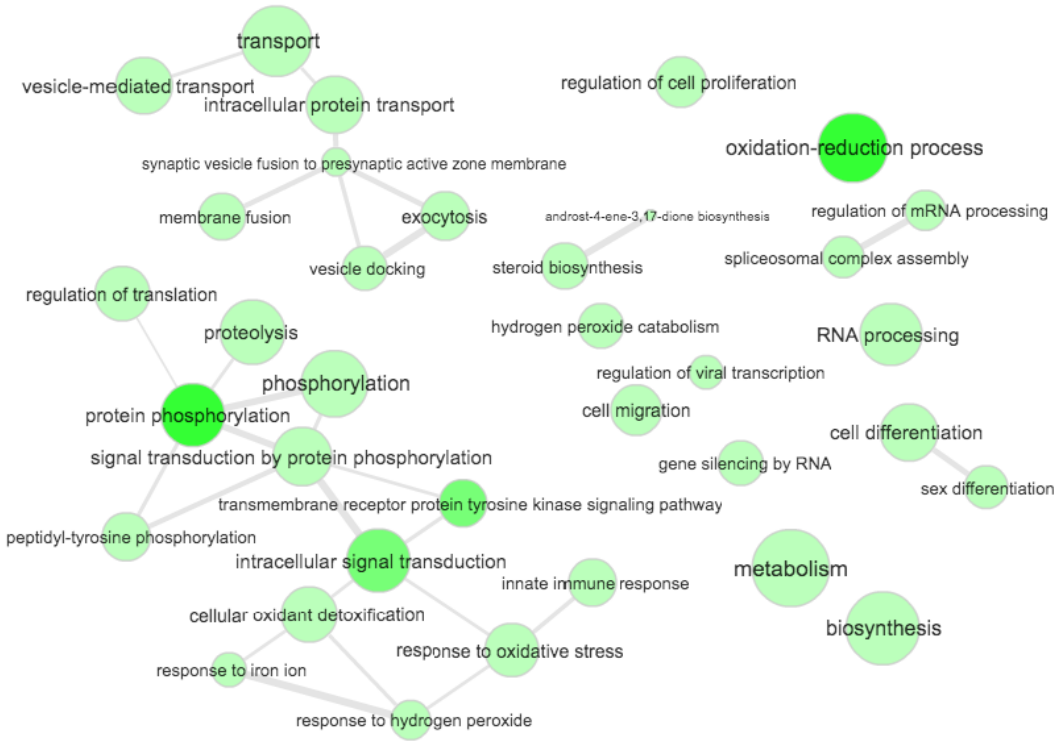


Figure S6

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the *M. unicolor* branch (branch 7 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

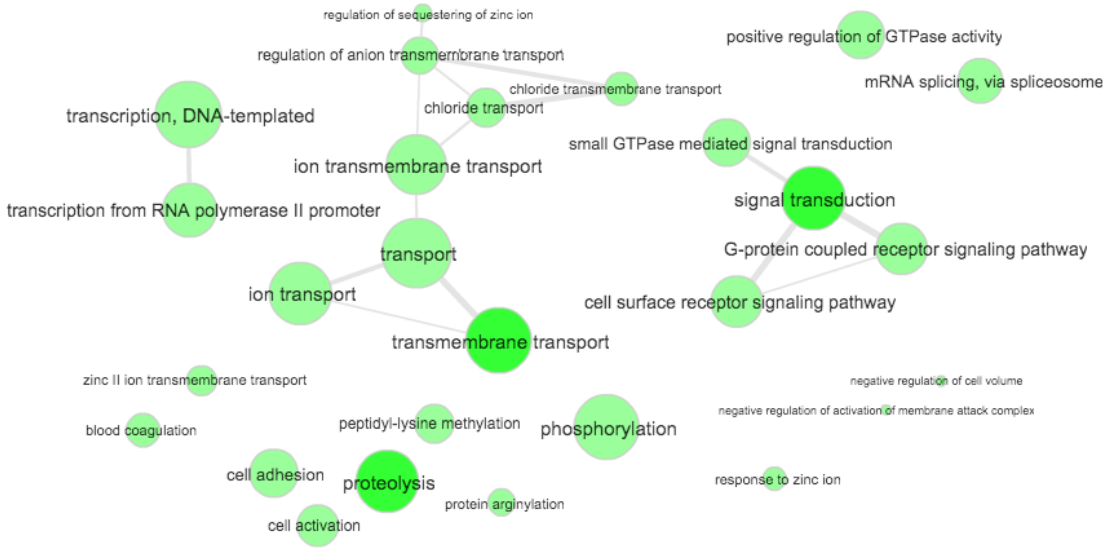


Figure S7

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the *T. compressicauda* branch (branch 8 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

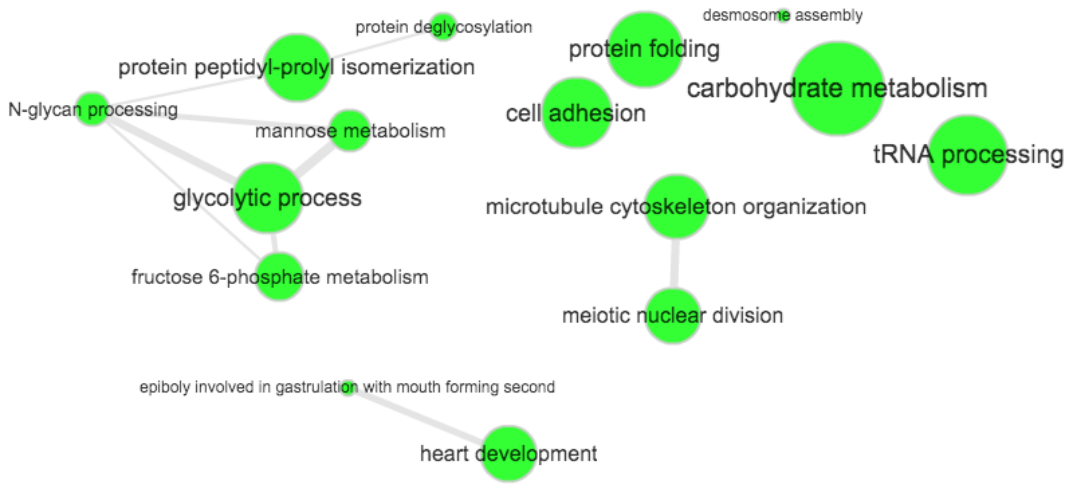


Figure S8

Network of the biological process domains of the gene ontologies (GOs) from the genes inferred to have been under positive selection on the *C. tentaculata* branch (branch 9 in Fig. 1). Circle size is related to the percentage of genes annotated with the GO term. Color intensity of the GO term circles is related to the number of genes associated to each GO term (darker color indicates greater number of genes inferred to have been under positive selection linked to GO term and higher circle size higher number of genes with the same GO in the UniProt database).

Chapter 3

Chemical defence and communication underground? Insights into skin specialisations of caecilian amphibians from gene expression profiles

María Torres-Sánchez, David J. Gower, Christopher J. Creevey, Kim Roelants, Mark Wilkinson, Diego San Mauro

Abstract

Skin is the largest organ of the vertebrate body, which performs many different important functions from protection to communication. To carry out its diverse functions, skin presents several specialisations. Skin tissue type and its specialised structures have their own distinct structural and chemical properties that are reflected in a different gene expression patterns in their cells. The study of gene expression in the skin provides information about animal ecology and its biotic and abiotic interactions. Here, we analyse the gene expression of the skin tissue type and eight different tissues of one of the most neglected vertebrate groups, the caecilian amphibians (order Gymnophiona). Caecilians are the sister group of frogs and salamanders and all of them exhibit moist, permeable skins with cutaneous, mucous and granular, glands. We identified 59 protein-coding genes with enriched expression in the caecilian skin and annotated several putative antimicrobial and pheromone peptides that are expressed in the studied dermal tissue type. Our study provides information about the molecular basis involved in caecilian vital functions such as protection, locomotion, defence, communication and reproduction. Our molecular large-scale characterisation of the caecilian skin provides information about the ecological role and evolution of the skin in vertebrates and particularly in caecilian amphibians.

Introduction

The study of particular tissues has been enhanced by the widespread use of the high-throughput sequencing technologies in transcriptomics and proteomics. All cells of one organism contain the same genetic information but different tissues are functionally distinct and present their characteristic gene expression patterns (1). Genes express in cells of one particular tissue depend not only on life punctual conditions but also on the cells history and/or organogenesis. Accordingly, a specific tissue is displaying a particular gene expression profile linked to the received signals during the embryonic development to perform its specific function. Cell embryological program and memory makes it possible to compare gene expression among tissues from related species helping to underline the particular mechanisms involved in their functionality.

The outer barrier of the surface of animal bodies is in the front line of ecological interactions with both abiotic and biotic elements. In vertebrates, the skin is this covering that interfaces with the environment through its particular gene expression profile. Skin is the largest organ in vertebrates and exhibits multiple functions with diverse specialised structures across species, including glands, scales, feathers and fur (2). When it comes to amphibians, the skin is a moist thin permeable tissue with multiple different types of exocrine glands. Amphibian skin conducts several functions and is involved in vital processes for the survival of the organisms in their own habitat. Glands of amphibian skin produce many biologically active compounds being some of them crucial for ecological interactions and part of complex traits of chemical defence and communication (3,4). From the amphibian chemical cocktail, we highlight antimicrobial peptides (AMPs) and peptide pheromones that are produced by many amphibian species for their importance in defence and communication mechanisms respectively. The first isolated AMP in amphibians was the bombinin from the skin secretion of *Bombina variegata* Linnaeus, 1758 (5). Since then, a larger number of AMPs from the skin of many frogs and salamanders have been identified. On the other hand, sodefrin, isolated from the newt *Cynops pyrrhogaster* Boie, 1826, was the first peptide pheromone identified (6) in amphibians.

Despite of the advances in the characterisation of the amphibian skin from frogs and salamanders, little is known about the gene expression in the skin of the most mysterious amphibian order, the caecilians (order Gymnophiona). Caecilians are a fossorial limbless amphibian group that live mainly in tropical soils (7). Reference transcriptomes for five species of caecilians from diverse tissues including skin have been generated (Chapter 1). Preliminary studies of caecilian amphibian transcriptomes pointed out the uniqueness of the caecilian skin and its potential production of chemicals involved in ecological interactions (Chapter 1). Here, we thoroughly analyse tissue expression patterns of the five caecilian transcriptomes in order to characterise the skin expression profile and identify functional elements involved in possible chemical interactions. We pursue achieving a better understanding of the functionality of caecilian skin, more generally amphibian skin, and the complex ecological interactions in which it is involved.

Materials and Methods

Source data of this study were the protein-coding gene sequences for five caecilian species (*Rhinatrema bivittatum* Cuvier in Guérrin-Méneville, 1838*, Caecilia tentaculata* Linnaeus, 1758*, Typhlonectes compressicauda* Duméril & Bibron, 1841*, Microcaecilia unicolor* Duméril, 1861, and *M. dermatophaga* Wilkinson, Sherratt, Starace & Gower, 2013) from reference species-specific transcriptomes and their raw reads (Chapter 1). The species-specific transcriptomes were generated from multiple tissue types: skin (separate midbody and posterior skin samples for most species), liver, lung, kidney, foregut, testis, heart, spleen, axial muscle (see Supplementary Table S1 for experimental design details).

In order to characterise caecilian skin gene expression, we conducted three different analyses: differential tissue expression, and annotation of genes encoding antimicrobial peptides and pheromones. Protein-coding genes of the five species-specific caecilian transcriptomes were aligned against manually annotated and reviewed proteins (Swiss-Prot) from the UniProt database (8) using the BLAST (9) blastp tool version 2.2.28, applying an arbitrary e-value threshold of 1e-20 that was deemed appropriate relative to the size of the database. Only genes with common annotation across all the transcriptomes were used in subsequent analysis. Gene expression levels were estimated using the counts of reads mapping to each assembly with HTSeq 0.6.1 (10). Expression values per gene in different tissues of each species-specific transcriptome were scaled by the mean of the total expression of the gene in all transcriptomes corrected by the mean of the total expression in the different tissues of the gene in its species-specific transcriptome. Variance-mean estimates were calculated for each tissue sample after normalisation of gene expression levels based on a negative binomial distribution, using the Bioconductor package DESeq2 (11). The tissue sample variance-means were subjected to principal components analysis (PCA). Genes differentially expressed in skin were identified as those with one logarithmic (log2) unit of fold change difference in variance-mean between skin (midbody + posterior) and non-skin tissue samples and with adjusted p-values < 0.05. We obtained gene ontologies (GOs) for those genes with positive logarithmic fold change (up-regulated genes) in skin. GO terms and their adjusted p-

values were summarized and visualized using REVIGO with 0.4 % allowed similarity as measured by semantic similarity and the whole UniProt database to define the size of each GO term (12). Protein-protein interactions (PPis) and functional enrichments within the up-regulated genes were sought using STRING (13) with default parameters.

Antimicrobial peptide (AMP) annotation for genes expressed in skin was carried out by aligning against three different datasets: ADP3 database (14), DADP database (15) and the output sequences from a UniProt search for andersonin, cathelicidin, cecropin and magainin (8), using the BLAST (9) blastp tool version 2.2.28, applying an e-value threshold of 1e-5 given the smaller size of the target databases. Pheromone annotation for genes expressed in skin was performed by aligning against the output sequences from a UniProt search for sodefrin, splendipherin and aphrodisin (8), using the BLAST (9) blastp tool version 2.2.28, applying an e-value threshold of 1e-5. We tested the null hypothesis of no difference in levels of AMP or of peptide pheromone gene expression between midbody and posterior skin using Wilcoxon signed-rank tests (with R: (17)) of transcripts per million (TPM) expression values calculated using RSEM with default parameters (16).

Results

A total of 2624 protein-coding genes have UniProt annotations that are the same across each of the five caecilian transcriptomes. Correlation among tissue samples for variance-means of scaled and normalised gene expression levels are shown in Figure 1. Liver, muscle and lung samples are each clustered by tissue type, indicating high correlation between gene expression values in these tissues among the different species. Skin comprises two groups of samples with closely correlated gene expression levels: (1) those for *R. bivittatum* and *M. unicolor* midbody skin and for *M. dermatophaga* midbody and posterior skin, and (2) those from *M. unicolor* posterior skin and *C. tentaculata* and *T. compressicauda* midbody and posterior skin.

The first six principal components (PCs) of the PCA together explain 45.15% of the total variance of the gene expression levels (Supplementary Table S2), with each subsequent PC each explaining < 5% of the variance. The fourth PC (5.66% of the variance) explains variance among expression levels according to tissue type (Figure 2), with skin having the highest positive values along this axis and liver having high negative values. Along this axis, lung and foregut samples are most similar to skin.

We identified 246 genes with differential expression values in skin (Figure 3). Among these, 59 are up-regulated in skin (Figure 3 and 4, and Supplementary Table S3) with 12 having positive logarithmic values of fold change > 4 (ATP13A4, BPIFC, CLDN4, DLX3, FAT2, KRT75, KRT80, *pou3f1, plcA*, TFAP2C, *tfap2e*, ZNF750). The GO terms for the skin up-regulated genes (Supplementary Table S3) are summarized and visualized in network graphs in Figure 5. Besides constitutive cellular processes, skin up-regulated genes are involved in processes such as epidermis development, epithelial cell migration, circadian rhythm, pathogenesis and secretion (Figure 5A). Binding is the predominant molecular function of the skin up-regulated genes (Figure 5B), and these genes carry out their functions in different cell compartments (Figure 5C). The enrichment analysis found no evidence of protein-protein interactions (p-value = 0.0947) for the skin up-regulated genes.

Annotation of the protein-coding genes resulted in identification of 91 putative AMPs (Table 1; Supplementary Table S4) and 43 putative peptide pheromones (Supplementary Table S5) from best BLAST matches in caecilian skin, expressed differently across the sampled species (Figure 6). Approximately one third of the protein-coding genes annotated as encoding AMPs occurred in the skin of all five sampled caecilian species (28 AMPs, Supplementary Table S4). In contrast, none of the protein-coding genes annotated as encoding peptide pheromones were common to all sampled species, and more than 80% of them are species-specific (35 peptide pheromones, Supplementary Table S5) and belong to sodefrin precursor-like factor (SPF) proteins. AMP gene expression is significantly higher in posterior than midbody skin for all four species for which this comparison was possible (Table 1). Peptide pheromone gene expression is not significantly different in midbody and posterior skin.

Discussion

Exploring the molecular basis of the skin is crucial to understand the ecological mechanisms in which species are involved. In this study, we have analysed the gene expression profile for different tissue types of 5 species of caecilians amphibians in order to identify protein-coding genes involved in caecilian skin adaptation and specialisation. According to the results, caecilian skin presents a special and truly distinct expression pattern across the analysed tissue types (Figure 1, 2, 3 and 4). Skin tissue samples are found correlated in two groups (Figure 1), separated perhaps by a strong phylogenetic signal (greater skin sample size would be required to test this hypothesis).

Gene expression appears to have diverse variation sources (we were studying 9 different tissue types from 5 caecilian species, see Supplementary Table S1) and less than half of the variance of the gene expression of the tissue samples has been captured in the first 6 principal components of our analysis (Supplementary Table S2). The gene expression variance relying on the fourth principal component reminds us of a germ layer classification from tissue organogenesis past (18). Skin tissue type has epithelial and mesenchymal components and is representative of ectoderm and mesoderm derived tissues. The skin tissue samples are found in our principal component analysis distally separated from liver tissue samples that are originated from endoderm layer. The most related samples to the skin, in terms of gene expression variance in the fourth component, are lung and foregut having, both tissues, an external epithelium cover (Figure 2). It may reflect an established pattern from the embryological developmental program geared to confront the challenges of external interactions in these tissues and in particular in the skin tissue type.

Skin exhibits a significant differential expression profile, meaning that the expression level of some genes allows distinguishing between the skin tissue type and non-skin tissue types (Figure 3). Among the 59 skin enriched genes, 12 present high values of logarithmic fold change and are annotated as homologs of transcription factors (*pou3f1*, TFAP2C, *tfap2e* and ZNF750), lyase (*plcA*), cation transporting ATPase

(ATP13A4), cadherin (FAT2), claudin (CLDN4), keratins (KRT75 and KRT80), homeobox protein (DLX3) and bactericidal/permeability-increasing protein (BPIFC).

DLX3 and BPIFC overcome seven positive units of logarithmic fold change and are the highest expressed genes in the caecilian skin. The first one has a crucial role in the differentiation of hair follicles in mammals (20) and its expression underlines the presence of similar specialisations in the caecilian skin. Several caecilian species, including four of the species of this study, present fish type scales in dermal pockets with an uncertain function. Our hypothesis is that DLX3-like caecilian peptide might be involved in the differentiation of caecilian scales and dermis development. From its UniProt description BPIFC is an endogenous bactericidal part of the innate immune system, and might be part of the defence mechanisms of the skin. KRT75 and KRT80 are two type II alpha-keratins present a logarithmic fold change for the skin of 4.74 and 4.65 respectively. Keratins are a family of fibrous proteins involved in cornification which main function is epithelial protection from harmful external damage and stress (19). The high expression of KRT75, KRT80, DLX3 and BPIFC in the skin highlights the important role of theses protein-coding genes in caecilian amphibians.

In GO molecular function terms, the vast majority of the skin up-regulated genes have binding functions, including ion, lipid, nucleic acid and protein binding (Figure 5 B). The high presence of binding elements implies that caecilian skin is active in the synthesis of molecules involved in these binds. The remaining protein-coding genes are modifying proteins (transferases, kinases, peptidases, hydrolases) or are proteins with structural activity.

The term of structural molecule activity conferring elasticity (GO:0005198) is related to three skin up-regulated protein-coding genes, mentioned above the keratins and the transmembrane protein claudin-4 (CLDN4, logarithmic fold change = 5.04, see Figure 3 and Supplementary Table S3). KRT75, KRT80 and CLDN4 could be built as a fibrous structure and help to preserve the integrity of the caecilian skin underground during their hydraulic movement in soils (21).

Regarding GO cellular component terms, the skin up-regulated genes are undertaking their functions broadly in all general cellular compartments (Figure 5 C). Nevertheless, among them we find protein-coding genes that are part of two highly specific cellular components, cornified envelopes (GO:0001533) and blebs (GO:0032059). Sciellin (SCEL, logarithmic fold change = 2.37, see Supplementary Table S3) is the protein-coding gene up-regulated in the caecilian skin that according to its GO description is involved in the formation of cornified envelopes and presumably related to KRT75, KRT80 and CLDN4 or even to the caecilian scales. Pannexin-1 (PANX1, logarithmic fold change = 1.75, see Supplementary Table S3) is the annotation of the caecilian skin protein-coding gene that is related to the blebs. PANX1 is a channel that connects intracellular and extracellular space and seems to be involved in the protrusion of plasmatic membrane (bleb). The function of blebs is not well known although are common in apoptosis (22,23).

Finally, the genes that are enriched in the skin are involved in several GO biological process terms relating not only to the maintenance of the basic cellular mechanism but also to specific processes pointing to skin specialisation, such as epidermis development (GO:0008544), epithelial cell migration (GO:0010631) and pathogenesis (GO:0009405, see Figure 5 A and Supplementary Table S3).

Our general study of the caecilian skin expression is completed with the analyses of AMPs and peptide pheromone annotations. There are previous evidences from secretions, protein-domain annotations of this transcriptomic data source that indicate that caecilians produce AMPs and peptide pheromones (Chapter 1) and from description of chemosensory organs (24), respectively. This study is the first thorough characterisation of the production of chemical peptides in caecilian amphibians. A total of 134 protein-coding genes from the five caecilian transcriptomes with expression in the skin were annotated as chemical peptides, belonging to AMPs or to peptide pheromones (Figure 6, Table 1, Supplementary Table S4 and S5). We analysed the production patterns of these chemicals across the five studied species of caecilians amphibians. We identified 43 different peptide pheromones, each species express around 10 of them (11 peptide pheromomnes in *R. bivittatum*, 11 as well in *C. tentaculata*, 10 in *T. compressicauda*, 12 in *M. unicolor* and 12 *M. dermatophaga*, see

Supplementary Table S5). The vast majority of peptide pheromones were annotated as SPF proteins (all except one protein-coding gene that are expressed in the skin of *C. tentaculata* were annotated as aphrodisin). SPF proteins have a wider presence in salamander species and belong to the same gene family of sodefrin that is courtship pheromone produced by male salamanders (6,25). Our results show a potential production of a multiple pheromone cocktail, being the vast majority sodefrin-like peptides, in caecilians that are synthesizing for both sexes (see Supplementary Table S1 for samples sex information). Male and female pheromone production could be an adaptation to overcome the difficulties to find a partner underground. Besides, these potential cocktails seem to be high species-specific not finding a common peptide pheromone annotation for the all five studied species (Figure 6). Caecilians amphibians have an enormous variety of reproductive modes including viviparity, oviparity with larvae and oviparity with direct development (26). It would not be surprising that mate attraction, as part of the reproduction trait, will exhibit great variation and specialisation, also taking into account that an erroneous mate could be very costly to animal fitness.

We found 91 different protein-coding genes annotated as AMPs, around 55 are expressed in each species (59 AMPs in *R. bivittatum*, 57 in *C. tentaculata*, 57 as well in *T. compressicauda*, 56 in *M. unicolor* and 52 *M. dermatophaga*, see Table 1 and Supplementary Table S4). Several of these AMPs annotations are only known from some specific animal species. Magainins and andersonins are unique to different frogs lineages (27,28) and cecropins are found exclusively in insects (29). The presence of these peptides in caecilian amphibians could be explained by convergent adaptation. Other remote possibility it is that these AMPs were acquired from the diet by sequestration and storage (30).

In contrast with peptide pheromones, we found several common AMP annotations for the five caecilian species (Figure 6). Chemical defence seems to be less specific than chemical communication and design to fight against common hazards. But also many caecilians AMPs are species-specific and could be acting to face the challenges of different ecological conditions of the environments in where caecilians are found. There are fully fossorial, subfossorial and fully aquatic species. Besides, terrestrial

forms occur in highly seasonal subtropical regions to per-humid tropical forest. Finally, we found an overall prevalence of AMP expression in posterior regions of the caecilian skin (Table 1). This high expression of chemical toxins in terminal part of the bodies of caecilian amphibians could be a strategy to avoid predation when the escape action fails.

In summary, we have explored molecular basis, chemical defence and communication of the skin of five caecilian amphibians using species-specific reference transcriptomes. We have identified many protein-coding genes with probable specific skin functions likely linked to adaptive responses. In order to assert the ecological particular role of the highlighted protein-coding genes in caecilian amphibians, further studies are needed. Sequence similarity does not always imply same function. Nevertheless, this study provides molecular information about skin mechanisms in caecilian amphibians opening the possibility of further studies and shedding light on the ecological role and evolution of skin tissue type in amphibians and in vertebrates.

References

1. Alberts B., Johnson A. B., Lewis A. J., Raff M., Roberts K., and Walter P. 2002. Molecular Biology off the cell (An Overview of Gene Control). New York: Garland Science. 1–38.
2. Bereiter-Hahn J., Matoltsy A. G. and Richards K. S. 1984. Biology of the integument 2 Vertebrates. Springer-Verlag Berlin Heidelberg GmbH.
3. Clarke B. T. 1997. The natural history of amphibian skin secretions, their normal functioning and potential medical applications. Biol Rev Camb Philos Soc. 72(3): 365–79.
4. Haslam I. S., Roubos E. W., Mangoni M. L., Yoshizato K., Vaudry H., Kloepper J. E., Pattwell D. M., Maderson P. F. A. and Paus R. 2014. From frog integument to human skin: Dermatological perspectives from frog skin biology. Biol Rev. 89(3): 618–55.
5. Csordás A. and Michl H. 1970. Isolation and Structure Elucidation of an Hemolytic Polypeptide from the Defensive Secretion European Bombina species (in German). Monatshefte für Chemie. 101(1): 182–9.
6. Kikuyama S. and Toyoda F. 1999. Sodefrin: a novel sex pheromone in a newt. Rev Reprod. 4: 1–4.
7. Wilkinson M. 2012. Caecilians. Curr Biol. 22(17).
8. Apweiler R. 2004. UniProt: the Universal Protein knowledgebase. Nucleic Acids Res. 32(90001): 115D–119.
9. 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(3): 403–10.
10. Anders S., Pyl P. T. and Huber W. 2015. HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics. 31(2): 166–9.
11. Love M. I., Huber W. and Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12): 550.
12. Supek F., Bošnjak M., Škunca N. and Šmuc T. 2011. Revigo summarizes and visualizes long lists of gene ontology terms. PLoS One. 6(7).
13. Szklarczyk D., Franceschini A., Wyder S., Forslund K., Heller D., Huerta-Cepas J., Simonovic M., Roth A., Santos A., Tsafou K. P., Kuhn M., Bork P.,

Jensen Lars J. and Von Mering C. 2015. STRING v10: Protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res. 43(D1): D447–52.

14. Wang G., Li X. and Wang Z. 2016. APD3: The antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. 44(D1): D1087–93.
15. Novković M., Simunić J., Bojović V., Tossi A. and Juretić D. 2012. DADP: The database of anuran defense peptides. Bioinformatics. 28(10): 1406–7.
16. Li B. and Dewey C. N. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 12(1): 323.
17. R Development Core Team. 2016. R: A Language and Environment for Statistical Computing. R Found Stat Comput Vienna Austria.
18. Gilbert S. 2007. Developmental Biology. Dev Biol. 311(2): 691.
19. Bragulla H. H. and Homberger D. G. 2009. Structure and functions of keratin proteins in simple, stratified, keratinized and cornified epithelia. J Anat. 214(4): 516–59.
20. Hwang J., Mehrani T., Millar S. E. and Morasso M. I. 2008. Dlx3 is a crucial regulator of hair follicle differentiation and cycling. Development. 135(18): 3149–59.
21. O'Reilly J. C., Summers A. P. and Ritter D. A. 2000. The Evolution of the Functional Role of Trunk Muscles During Locomotion in Adult Amphibians. Am Zool. 40(1): 123–35.
22. Chekeni F. B., Elliott M. R., Sandilos J. K., Walk S. F., Kinchen J. M., Lazarowski E. R., Armstrong A. J., Penuela S., Laird D. W., Salvesen G. S., Isakson B. E., Bayliss D. A. and Ravichandran K. S. 2010. Pannexin 1 channels mediate 'find-me' signal release and membrane permeability during apoptosis. Nature. 467(7317): 863–7.
23. Kalra H., Drummen G. P. C. and Mathivanan S. 2016. Focus on extracellular vesicles: Introducing the next small big thing. Int J Mol Sci. 17(2): 170.
24. Schmidt A. and Wake M. H. 1990. Olfactory and vomeronasal systems of caecilians (Amphibia: Gymnophiona). J Morphol. 205(3): 255–68.
25. Van Bocxlaer I., Maex M., Treer D., Janssenswillen S., Janssens R., Vandebergh W., Proost P. and Bossuyt F. Beyond sodefrin: evidence for a

multi-component pheromone system in the model newt Cynops pyrrhogaster (Salamandridae). Sci Rep. 21880.

26. 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(3): 191–202.

27. Roelants K., Fry B. G., Ye L., Stijlemans B., Brys L., Kok P., Clynen E., Schoofs L., Cornelis P. and Bossuyt F. 2013. Origin and Functional Diversification of an Amphibian Defense Peptide Arsenal. PLoS Genet. 9(8).

28. 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(2): 125–30.

29. Yi H. Y., Chowdhury M., Huang Y. D. and Yu X. Q. 2014. Insect antimicrobial peptides and their applications. Appl Microbiol Biotechnol. 98(13): 5807-22.

30. Savitzky A. H., Mori A., Hutchinson D. A., Saporito R. A., Burghardt G. M., Lillywhite H. B. and Meinwald J. 2012. Sequestered defensive toxins in tetrapod vertebrates: Principles, patterns, and prospects for future studies. Chemoecology. 22(3): 141-158

Tables and Figures

Table 1

Annotated AMPs and p-values for Wilcoxon signed-rank test of differences between AMP expression levels in midbody and posterior skin samples. * indicates custom made databases for subset of entries for these UniProt terms (see Materials and Methods). For *R. bivittatum* only data for midbody skin were available.

Database	*Rhinatrema bivittatum*	*Caecilia tentaculata*	*Typhlonectes compressicauda*	*Microcaecilia unicolor*	*Microcaecilia dermatophaga*
APD3	36	41	37	32	30
DADP	3	3	3	3	2
Andersonin*	3	3	4	4	3
Cathelicidin*	10	4	7	12	11
Cecropin*	5	5	5	4	5
Magainin*	2	1	1	1	1
p-value	-	1.31e-14	1.551e-12	< 2.2e-16	< 2.2e-16

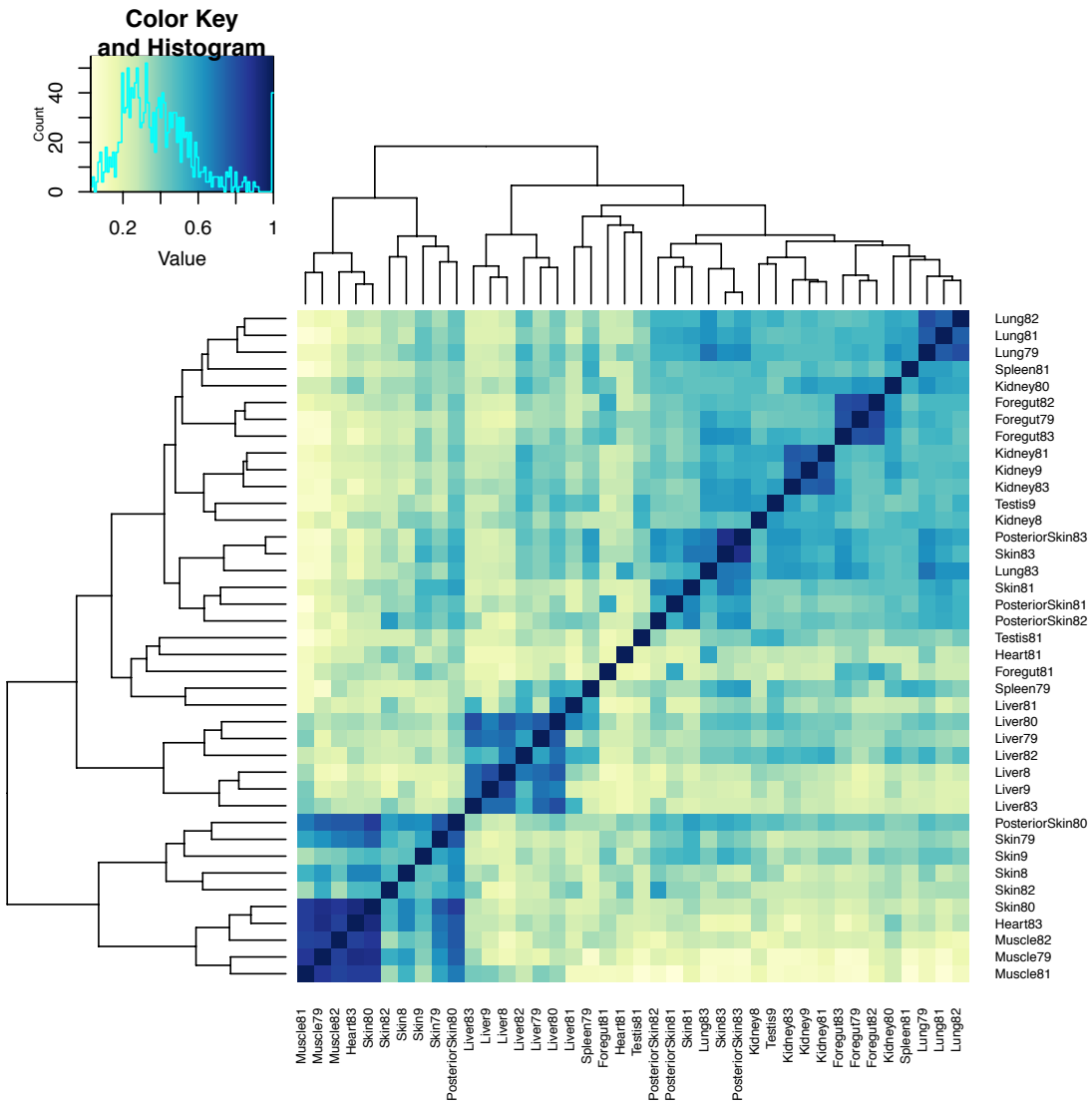


Figure 1

Heatmap showing correlation between variance-mean expression levels for protein-coding genes in different caecilian tissue samples.

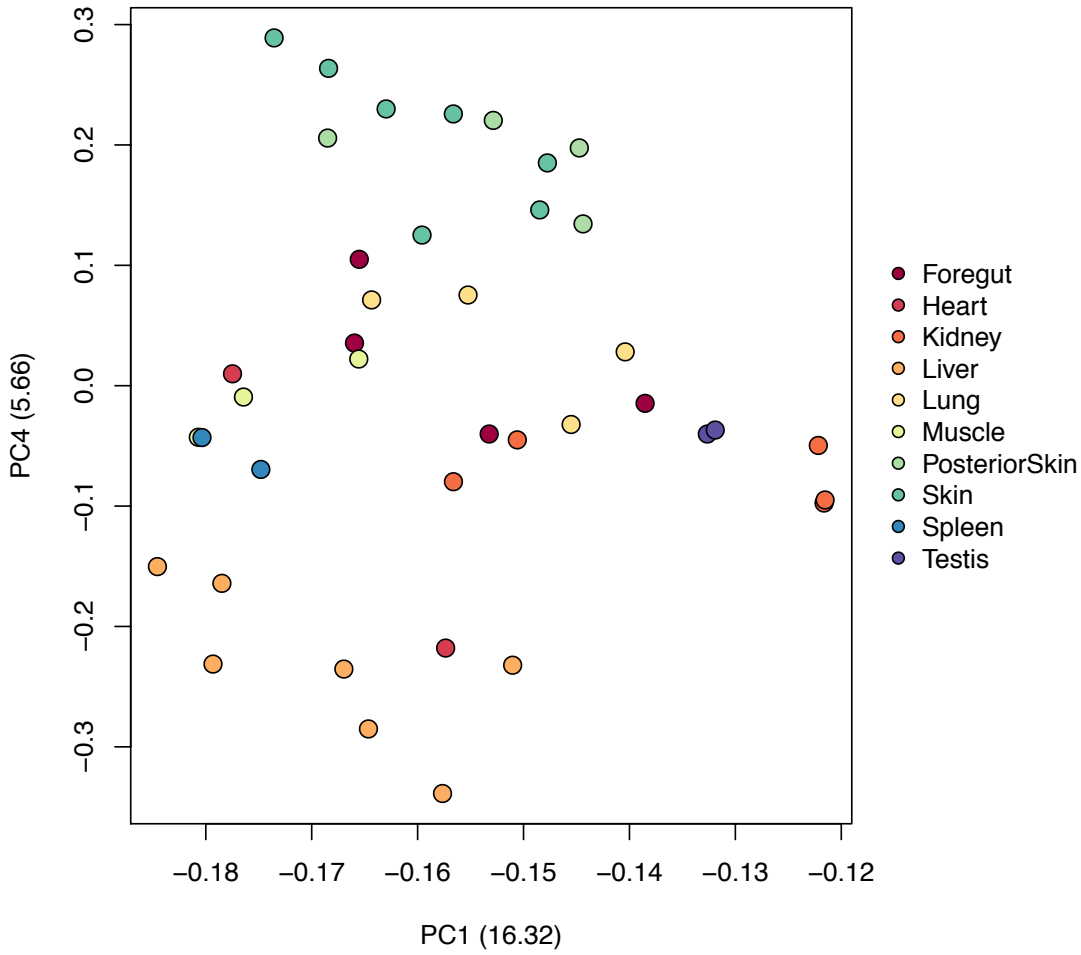


Figure 2

PCA plot of PC1 versus PC4 showing variance among gene expression levels in various tissue types across the five sampled caecilian species.

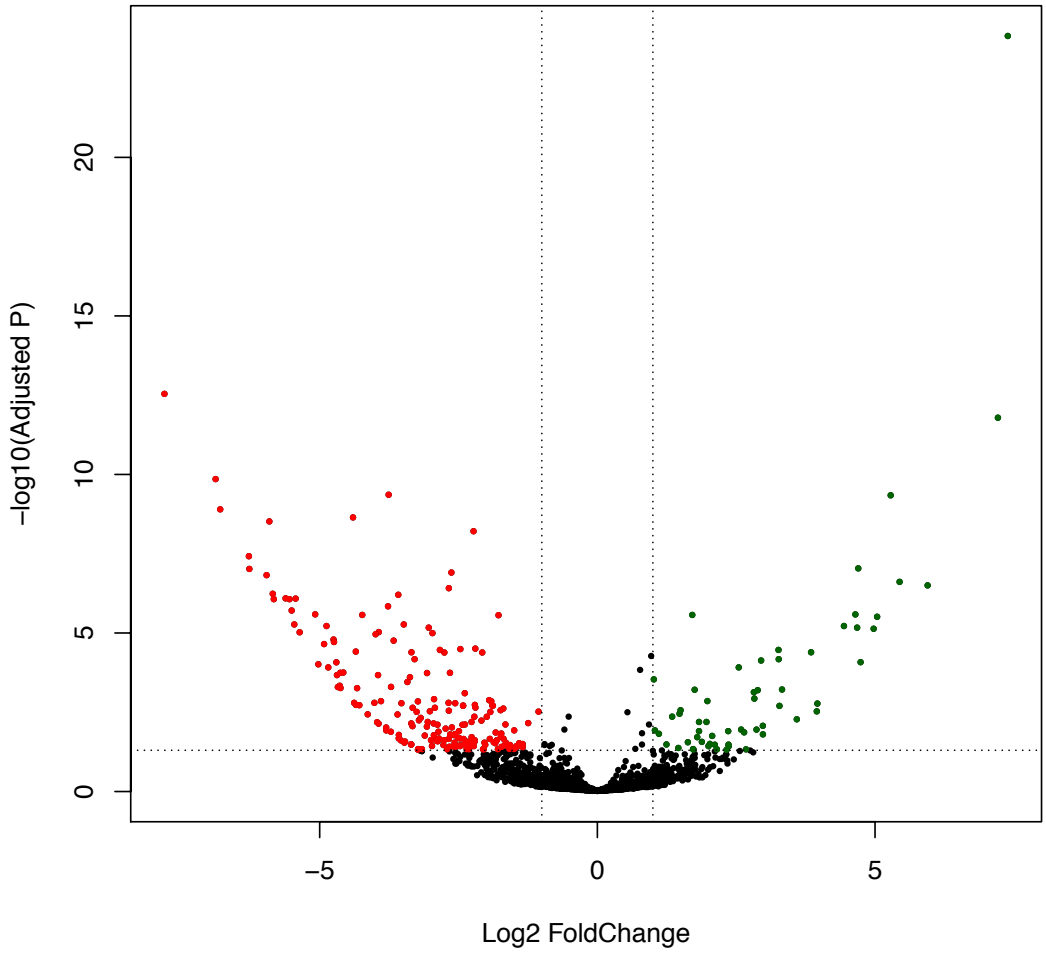


Figure 3

Protein-coding genes differentially expressed in caecilian skin. The plot shows the magnitude of difference in expression levels between skin and non-skin tissues, with red dots indicating significantly down- and green dots significantly up-regulated (enriched) genes.

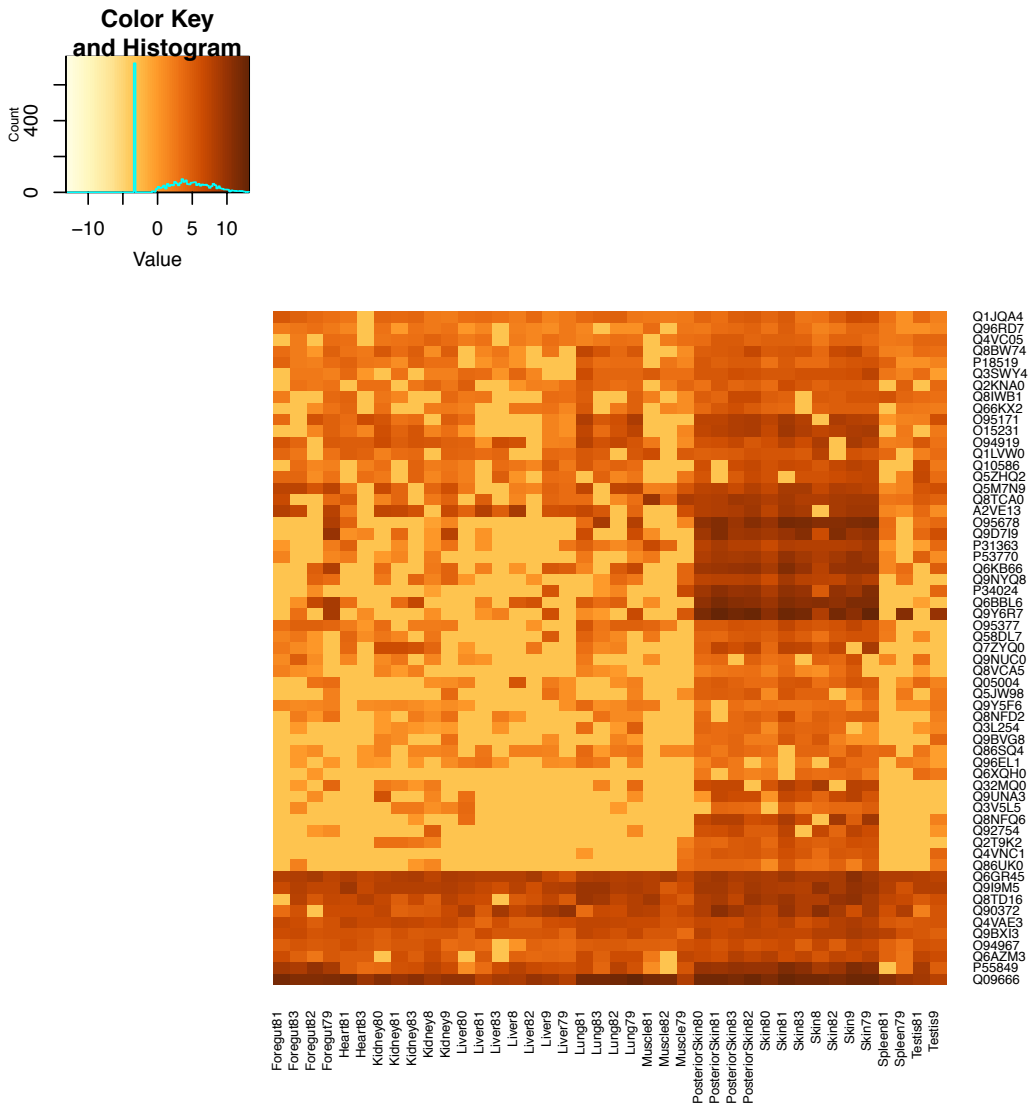


Figure 4

Heatmap showing expression levels of skin up-regulated genes in the sampled caecilian tissues.

Figure 5

Network graphs for GO domains (A: biological process, B: molecular function and C: cellular component) of skin up-regulated genes. Greater colour intensity indicates more significant p-value (of difference in expression between skin and non-skin) and circle is positively correlated with number of genes with the same GO in the UniProt database.

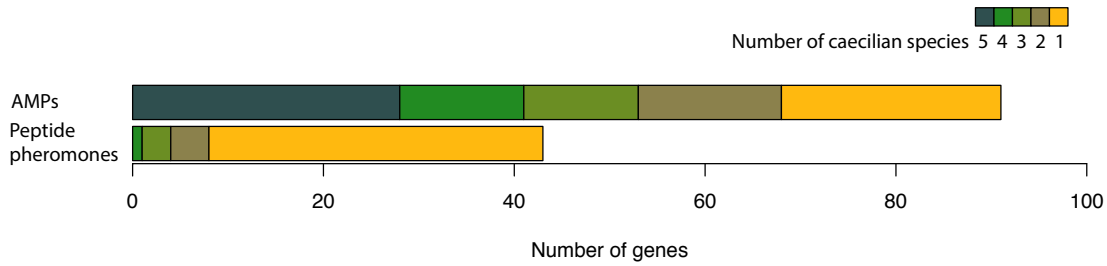


Figure 6

Expressed genes annotated as encoding peptides and their presence in the skin of the five sampled caecilian species.

Supplementary material

Table S1

Experimental design for differential expression analysis and associated sample information.

Species	Life style	Sex	Tissue	Sample name	Experimental design condition
Rhinatrema bivittatum	Terrestrial	Male	Kidney	Kidney9	Non-skin
			Liver	Liver9	Non-skin
			Skin	Skin9	Skin
			Testis	Testis9	Non-skin
		Female	Foregut	Foregut79	Non-skin
			Liver	Liver79	Non-skin
			Lung	Lung79	Non-skin
			Muscle	Muscle79	Non-skin
			Skin	Skin79	Skin
			Spleen	Spleen79	Non-skin
Caecilia tentaculata	Terrestrial	Male	Foregut	Foregut81	Non-skin
			Heart	Heart81	Non-skin
			Kidney	Kidney81	Non-skin
			Liver	Liver81	Non-skin
			Lung	Lung81	Non-skin
			Muscle	Muscle81	Non-skin
			Posterior skin	PosteriorSkin81	Skin
			Skin	Skin81	Skin
			Spleen	Spleen81	Non-skin
			Testis	Testis81	Non-skin
Typhlonectes compressicauda	Aquatic	Male	Foregut	Foregut83	Non-skin
			Heart	Heart83	Non-skin
			Kidney	Kidney83	Non-skin
			Liver	Liver83	Non-skin
			Lung	Lung83	Non-skin
			Posterior skin	PosteriorSkin83	Skin
			Skin	Skin83	Skin
Microcaecilia unicolor	Terrestrial	Young female	Foregut	Foregut82	Non-skin
			Liver	Liver82	Non-skin
			Lung	Lung82	Non-skin
			Muscle	Muscle82	Non-skin
			Posterior skin	PosteriorSkin82	Skin
			Skin	Skin82	Skin
		Female	Kidney	Kidney8	Non-skin
			Liver	Liver8	Non-skin
			Skin	Skin8	Skin
Microcaecilia dermatophaga	Terrestrial	Unknown	Kidney	Kidney80	Non-skin
			Liver	Liver80	Non-skin
			Posterior skin	PosteriorSkin80	Skin
			Skin	Skin80	Skin

Table S2

Principal components (PC) values of the gene expression variance.

	PC1	PC2	PC3	PC4	PC5	PC6
Captured percentage of gene expression variance	16.32	6.67	6.33	5.66	5.16	5.01

Table S3

Description of significantly up-regulated genes in the caecilian skin transcriptomes.

Uniprot ID	Gene name	Protein description	GO terms	Adjusted p-value	Log_2 Fold Change
A2VE13	MAL2	Multispan transmembrane protein	GO:0001766; GO:0008104; GO:0016021; GO:0016324; GO:0019911; GO:0042552; GO:0045056; GO:0045121; GO:0070062	0.046814558	2.152141712
O15231	ZNF185	Zinc finger protein 185	GO:0005737; GO:0005856; GO:0005925; GO:0008270	0.000606255	3.325495301
O94919	ENDOD1	Endonuclease domain-containing 1 protein	GO:0002576; GO:0003676; GO:0004519; GO:0005576; GO:0005829; GO:0016020; GO:0046872; GO:0070062	0.027163336	1.882290321
O94967	WDR47	WD repeat-containing protein 47	GO:0005737; GO:0005874; GO:0007275	0.004348627	1.346953179
O95171	SCEL	Sciellin	GO:0001533; GO:0005737; GO:0008544; GO:0009790; GO:0030216; GO:0046872; GO:0070062	0.036957689	2.367525528
O95377	GJB5	Gap junction beta-5 protein	GO:0005922; GO:0007154; GO:0008544; GO:0016021; GO:0060707; GO:0060708; GO:0060713; GO:1905867	0.000736441	2.818280184
O95678	KRT75	Keratin, type II cytoskeletal 75	GO:0002244; GO:0005198; GO:0005829; GO:0005882; GO:0031424; GO:0045095; GO:0070062; GO:0070268	8.31E-05	4.739755569
P18519	TNFRSF16	Nerve growth factor receptor	GO:0005031; GO:0005516; GO:0005634; GO:0005886; GO:0005887; GO:0006919; GO:0006954; GO:0006955; GO:0007266; GO:0007411; GO:0010977; GO:0015026; GO:0031625; GO:0032496; GO:0032922; GO:0042127; GO:0042981; GO:0043005; GO:0043121; GO:0048406; GO:0097190; GO:1900182; GO:1902895; GO:1903588	0.001402807	1.981215918
P31363	*pou3f1*	POU domain, class 3, transcription factor 1-A	GO:0003700; GO:0005634; GO:0006351; GO:0006357; GO:0007420; GO:0043565	9.16E-08	4.697728521
P34024	*plcA*	Phosphatidylinositol diacylglycerol-lyase	GO:0004436; GO:0005576; GO:0005737; GO:0008081;	3.15E-07	5.946929132

			GO:0009405; GO:0016042		
P53770	DLX3	Homeobox protein DLX-3	GO:0005634; GO:0006355; GO:0007275; GO:0043565	1.48E-24	7.390548223
P55849	*dsc1*	Desmocollin-1	GO:0005509; GO:0005886; GO:0007156; GO:0016021; GO:0030057; GO:0070062	7.37E-05	2.948641112
Q05004	NXPE1	Brush border protein	GO:0005576	0.011153655	2.588884021
Q09666	AHNAK	Desmoyokin	GO:0003723; GO:0005634; GO:0005737; GO:0005765; GO:0005829; GO:0005886; GO:0005925; GO:0015629; GO:0016020; GO:0030315; GO:0031982; GO:0042383; GO:0043034; GO:0043484; GO:0044291; GO:0044548; GO:0045296; GO:0051259; GO:0070062; GO:0097493; GO:1901385	0.032811518	1.244850455
Q10586	DBP	D site-binding protein	GO:0000977; GO:0001077; GO:0001889; GO:0005634; GO:0006357; GO:0007275; GO:0007623; GO:0045944	0.000636255	2.886685174
Q1JQA4	TSPAN15	Tetraspanin-15	GO:0005887; GO:0007166; GO:0009986; GO:0019899; GO:0031902; GO:0051604; GO:0070062; GO:0090002; GO:0097197	0.002730037	1.498620555
Q1LVW0	*btbd11a*	Ankyrin repeat and BTB/POZ domain-containing protein BTBD11-A	GO:0000786; GO:0003677; GO:0005634; GO:0005737; GO:0016021; GO:0019005; GO:0030162; GO:0031625; GO:0042787; GO:0043161; GO:0046982; GO:0060395	0.045919504	1.723118039
Q2KNA0	SPECC1L	Cytospin-A	GO:0005737; GO:0005815; GO:0005819; GO:0005921; GO:0007026; GO:0007049; GO:0016477; GO:0030036; GO:0030835; GO:0031941; GO:0051301	0.019297852	1.798481517
Q2T9K2	*tfap2e*	Transcription factor AP-2-epsilon	GO:0003677; GO:0003700; GO:0005634; GO:0006351	6.02E-06	4.440094364
Q32MQ0	ZNF750	Zinc finger protein 750	GO:0001046; GO:0001077; GO:0005634; GO:0005730; GO:0008544; GO:0030154; GO:0043231; GO:0045944; GO:0046872; GO:1990841	2.44E-07	5.442262925
Q3L254	WNT7B	Protein Wnt-7b	GO:0005109; GO:0005578; GO:0005615; GO:0016055; GO:0030182; GO:0045165; GO:0070307	0.033257584	2.361343934

Q3SWY4	LRRN4CL	LRRN4 C-terminal-like protein	GO:0016021	0.000121558	2.545750467
Q3V5L5	MGAT5B	Alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyltransferase B	GO:0000139; GO:0005794; GO:0006487; GO:0016021; GO:0030144; GO:0046872	0.008455543	2.979659158
Q4VAE3	*tmem65*	Transmembrane protein 65	GO:0003231; GO:0005739; GO:0005743; GO:0005886; GO:0014704; GO:0016021; GO:1903779	0.011779563	1.036585843
Q4VC05	BCL7A	B-cell CLL/lymphoma 7 protein family member A	GO:0045892	0.003519799	1.479851356
Q4VNC1	ATP13A4	Probable cation-transporting ATPase 13A4	GO:0005388; GO:0005524; GO:0005886; GO:0005887; GO:0006874; GO:0019829; GO:0034220; GO:0043231; GO:0046872	6.78E-06	4.677358853
Q58DL7	ARHGEF9	Collybistin	GO:0005089; GO:0005829; GO:0035023	0.001172687	2.827209183
Q5JW98	FAM26D	Protein FAM26D	GO:0005261; GO:0005887; GO:0034220	0.001990421	3.279102681
Q5M7N9	*esyt3*	Extended synaptotagmin-3	GO:0005886; GO:0006869; GO:0008289; GO:0016021; GO:0031227; GO:0044232; GO:0046872	0.031547858	2.026833989
Q5ZHQ2	LGALSL	Galectin-related protein	GO:0005737; GO:0030246	0.012470091	2.355960573
Q66KX2	*cadm4*	Cell adhesion molecule 4	GO:0007155; GO:0016021	0.017628461	2.067869795
Q6AZM3	*reep4*	Receptor expression-enhancing protein 4	GO:0005783; GO:0005789; GO:0005874; GO:0006998; GO:0007084; GO:0008017; GO:0016021; GO:0051301	0.02797932	1.629959292
Q6BBL6	CLDN4	Claudin-4	GO:0005198; GO:0005887; GO:0005923; GO:0016327; GO:0016338; GO:0042802; GO:0061436	3.08E-06	5.037737827
Q6GR45	*eif6*	Eukaryotic translation initiation factor 6 (eIF-6)	GO:0003743; GO:0005730; GO:0005737; GO:0042256; GO:0043022	0.00028972	1.017060671
Q6KB66	KRT80	Keratin, type II cytoskeletal 80	GO:0005198; GO:0005737; GO:0005829; GO:0005882; GO:0031424; GO:0045095; GO:0045111; GO:0070268	2.60E-06	4.646101698
Q6XQH0	*gal3st2*	Galactose-3-O-sulfotransferase 2	GO:0001733; GO:0008146; GO:0009101; GO:0009247; GO:0016020; GO:0016021;	0.046814558	2.676687744

			GO:0032580; GO:0050694; GO:0051923		
Q7ZYQ0	*foxi1e*	Forkhead box protein I1-ema	GO:0003700; GO:0005634; GO:0006351; GO:0007275; GO:0042664; GO:0043565; GO:0045893; GO:0048335	0.015743687	2.981023402
Q86SQ4	ADGRG6	Adhesion G-protein coupled receptor G6	GO:0004930; GO:0005518; GO:0005622; GO:0005886; GO:0007005; GO:0007166; GO:0007186; GO:0010579; GO:0014037; GO:0016021; GO:0019933; GO:0022011; GO:0042552; GO:0043236; GO:0050840; GO:0060347	0.00639498	1.965019155
Q86UK0	ABCA12	ATP-binding cassette sub-family A member 12	GO:0005102; GO:0005319; GO:0005524; GO:0005737; GO:0005743; GO:0005829; GO:0005886; GO:0006869; GO:0010875; GO:0016021; GO:0019725; GO:0031424; GO:0032940; GO:0033700; GO:0034040; GO:0034191; GO:0035627; GO:0042626; GO:0043129; GO:0043231; GO:0045055; GO:0048286; GO:0055085; GO:0055088; GO:0061436; GO:0072659; GO:0097209; GO:2000010	4.05E-05	3.848213513
Q8BW74	*hlf*	Hepatic leukemia factor	GO:0000977; GO:0001077; GO:0001228; GO:0005634; GO:0035914; GO:0043565; GO:0045944; GO:0048511	0.047379797	1.730771243
Q8IWB1	ITPRIP	Inositol 1,4,5-trisphosphate receptor-interacting protein	GO:0005886; GO:0016020	0.033439292	2.114786378
Q8NFD2	ANKK1	Ankyrin repeat and protein kinase domain-containing protein 1	GO:0004674; GO:0005524	6.69E-05	3.265571376
Q8NFQ6	BPIFC	BPI fold-containing family C protein	GO:0001530; GO:0005543; GO:0005615	1.62E-12	7.212949827
Q8TCA0	LRRC20	Leucine-rich repeat-containing protein 20	-	0.013581506	2.642562647
Q8TD16	BICD2	Protein bicaudal D homolog 2	GO:0000042; GO:0005635; GO:0005642; GO:0005643; GO:0005794; GO:0005829; GO:0005856; GO:0005886; GO:0006890; GO:0017137;	0.00639498	1.830649916

			GO:0031410; GO:0034452; GO:0051028; GO:0051642; GO:0051959; GO:0070840; GO:0072385; GO:0072393		
Q8VCA5	*tmprss4*	Transmembrane protease serine 4	GO:0004252; GO:0005044; GO:0016021	3.42E-05	3.261649905
Q90372	QNR-71	Protein QNR-71	GO:0016021	0.012470091	1.830499398
Q92754	TFAP2C	Transcription factor AP-2 gamma	GO:0000122; GO:0000977; GO:0001047; GO:0001077; GO:0001078; GO:0003677; GO:0003700; GO:0005634; GO:0005654; GO:0005739; GO:0005829; GO:0006357; GO:0007267; GO:0008584; GO:0040029; GO:0042127; GO:0045944; GO:0046983	7.35E-06	4.975393316
Q96EL1	FAM212A	PAK4-inhibitor INKA1	GO:0005634; GO:0005737; GO:0019901; GO:0021915; GO:0030291; GO:0070062	0.036957689	2.001293676
Q96RD7	PANX1	Pannexin-1	GO:0002020; GO:0002931; GO:0005102; GO:0005262; GO:0005783; GO:0005789; GO:0005886; GO:0005921; GO:0006812; GO:0006816; GO:0007267; GO:0016020; GO:0016021; GO:0022840; GO:0032059; GO:0033198; GO:0034214; GO:0043234; GO:0044325; GO:0046982; GO:0050717; GO:0050718; GO:0051015; GO:0055077; GO:0097110	0.000615936	1.751947004
Q9BVG8	KIFC3	Kinesin-like protein KIFC3	GO:0003777; GO:0005524; GO:0005794; GO:0005813; GO:0005871; GO:0005874; GO:0005915; GO:0007018; GO:0007030; GO:0007601; GO:0008017; GO:0008569; GO:0016887; GO:0030659; GO:0045218; GO:0070062; GO:0090136	0.011153655	2.86600906
Q9BXI3	NT5C1A	Cytosolic 5'-nucleotidase 1A	GO:0000166; GO:0000287; GO:0005829; GO:0006195; GO:0008253; GO:0009116; GO:0009128; GO:0046085; GO:0046135	2.70E-06	1.709244848
Q9D7I9	*tgm5*	Transglutaminase-5	GO:0003810; GO:0005737; GO:0018149; GO:0046872	0.005268197	3.591491589
Q9I9M5	*fzd1*	Frizzled-1	GO:0004930; GO:0005886; GO:0007275; GO:0016021;	0.015121931	1.108554378

			GO:0042813		
Q9NUC0	SERTAD4	SERTA domain-containing protein 4	GO:0005634	0.048330757	2.326310075
Q9NYQ8	FAT2	Protocadherin Fat 2	GO:0005509; GO:0005634; GO:0005886; GO:0005913; GO:0007156; GO:0010631; GO:0016021; GO:0070062	4.54E-10	5.281769654
Q9UNA3	A4GNT	Alpha-1,4-N-acetylglucosaminyltransferase	GO:0000139; GO:0005975; GO:0006493; GO:0008375; GO:0009101; GO:0016020; GO:0016021; GO:0016266; GO:0050680	0.00168146	3.961880549
Q9Y5F6	PCDHGC5	Protocadherin gamma-C5	GO:0005509; GO:0005887; GO:0007155; GO:0007156; GO:0007267; GO:0007399; GO:0070062	0.042191149	1.458789538
Q9Y6R7	FCGBP	IgGFc-binding protein	GO:0070062	0.002961722	3.951361199

Table S4

Antimicrobial peptide annotation (APD31, DADP2, Uniprot terms: Andersonin3, Cathelicidin4, Cecropin5 and Magainin6) and occurrence in the five sampled caecilian species.

AMP ID	Gene name	Protein description	*Rhinatrema bivittatum*	*Caecilia tentaculata*	*Typhlonectes compressicauda*	*Microcaecilia unicolor*	*Microcaecilia dermatophaga*
AP00140[1]	SK84	Glycine-rich AMP		X	X	X	X
AP00208[1]	P80230	Inhibitory polypeptide	X	X	X	X	X
AP00294[1]	eNAP-1	Alpha defesin	X	X	X	X	X
AP00400[1]	P80952	Skin peptide tyrosine-tyrosine			X	X	
AP00429[1]	NK-lysin	Effector peptide of cytotoxic T/NK cells		X	X	X	
AP00481[1]	*Kaliocin-1*	Lactoferrin	X	X			
AP00489[1]	*Hipposin*	Histone-derived AMP	X	X	X	X	X
AP00536[1]	*Luxuriosin*	AMP with Kunitz domain				X	X
AP00612[1]	*Chrombacin*	Sulfated phosphorylated peptide	X	X		X	X
AP00630[1]	*Amoebapore A*	Saposin-like protein	X				X
AP00812[1]	*Enkelytin*	Proenkephalin-A neuropeptide		X		X	
AP01157[1]	*Ixodidin*	Cys-rich AMP	X	X	X	X	X
AP01339[1]	BHP	Hemoglobin peptide		X	X	X	
AP01340[1]	*Naegleriapore A*	Saposin-like protein	X	X	X		X
AP01372[1]	CXCL14	Chemokine			X		X
AP01373[1]	*Thrombocidin1*	Chemokine		X	X		
AP01374[1]	*Thrombocidin2*	Chemokine			X	X	
AP01474[1]	NPY	Neuropeptide Y	X	X			
AP01476[1]	CGRP	Calcitonin gene-related neuropeptide		X	X		
AP01477[1]	VIP	Vasoactive neuropeptide	X	X	X	X	X
AP01479[1]	*Adrenomedullin*	Neuropeptide	X	X	X	X	X
AP01522[1]	*Ap*	Antifungal peptide		X	X	X	
AP01540[1]	SP-BN	N-terminal region of surfactant protein B saposin-like	X	X	X	X	X
AP01575[1]	TCP	Thrombin-derived C-terminal peptide	X	X	X	X	X
AP01580[1]	*Elafin*	Elastase-specific inhibitor	X				
AP01646[1]	*gcLEAP-2*	Grass carp liver-expressed antimicrobial peptide-2		X			
AP01676[1]	*Abeta42*	Beta-amyloid peptide	X	X	X	X	X
AP02012[1]	YFGAP	Yellowfin tuna GAPDH-related AMP	X	X	X	X	X

AP02017[1]	hGAPDH	Glyceraldehyde -3-phosphate dehydrogenase		X			
AP02030[1]	*cgUbiquitin*	Hemolytic peptide	X	X	X	X	X
AP02068[1]	*Ang1*	Murine angiogenin 1			X		
AP02069[1]	*Ang4*	Murine angiogenin 4	X				
AP02070[1]	*RegIIIgamma*	Secreted C-type lectin AMP	X	X	X	X	X
AP02071[1]	*RegIIIalpha*	Secreted C-type lectin AMP	X	X	X	X	X
AP02075[1]	CCL20	Macrophage inflammatory protein-3alpha	X				
AP02076[1]	CXCL1	Chemokine		X	X	X	
AP02078[1]	CXCL3	Chemokine	X	X	X	X	X
AP02080[1]	CXCL10	Chemokine		X	X	X	X
AP02081[1]	CXCL11	Chemokine		X			
AP02082[1]	CXCL12	Chemokine			X		
AP02083[1]	CXCL13	Chemokine		X	X	X	
AP02084[1]	XCL1	Lymphotactin chemokine	X	X	X		
AP02087[1]	CCL11	Eotaxin	X				
AP02088[1]	CCL13	Chemokine	X				
AP02090[1]	CCL18	Chemokine				X	X
AP02091[1]	CCL19	Chemokine	X		X		X
AP02092[1]	CCL25	Chemokine	X	X			
AP02095[1]	SLPI	Secretory leukocyte protease inhibitor	X	X	X	X	X
AP02096[1]	UBI	Murine peptide	X	X	X	X	X
AP02122[1]	pCM1	Fragment of hglyrichin	X				
AP02185[1]	CXCL6	Chemokine	X	X			X
AP02186[1]	CCL28	Chemokine	X	X	X	X	
AP02188[1]	mCCL28	Chemokine	X	X			X
AP02195[1]	Chemerin	Retinoic acid receptor responder protein 2	X	X	X	X	X
AP02230[1]	HMGN2	High mobility group nucleosomal binding domain 2	X	X	X		X
AP02257[1]	*Lysozyme*	Lectin-binding enzyme	X	X	X	X	X
2965[2]	*Buforin 2*	Histone H2A derived AMP	X	X	X	X	X
P80952[2]	-	Skin peptide tyrosine-tyrosine	X	X	X	X	
P83578[2]	-	Proteinase inhibitor PSKP-1		X			
P86282[2]	*Phylloseptin Bu-1*	-	X		X	X	X
D2K8I9[3]	-	Andersonin-9 AMP	X	X	X	X	X
E3SZM1[3]	-	Andersonin-8a peptide		X			
E3SZM2[3]	-	Andersonin-8b peptide	X		X	X	X
E3SZM5[3]	-	Andersonin-11 peptide	X	X	X	X	X
E3SZM6[3]	-	Andersonin-7 peptide			X	X	
K7GIB1[4]	-	Uncharacterized protein cathelicidin-like			X		
A0A067QGP8[4]	L798_02828	Cathelicidin-B1	X				X

A0A096N3S1[4]	-	Uncharacterized protein cathelicidin-like	X	X	X	X	X
D3ZMP7[4]	LOC689081	Cystatin	X				
F7C0D9[4]	KNG1	Kininogen 1	X		X		X
F7C0Z0[4]	KNG1	Kininogen 1				X	
G1PML8[4]	KNG1	Kininogen 1				X	
G3VCA4[4]	CST7	Cystatin					X
G3VCA5[4]	CST7	Cystatin			X	X	
H3A036[4]	CST7	Cystatin	X	X	X	X	X
I3IVB6[4]	-	Uncharacterized protein cathelicidin-like	X				
I3JY77[4]	LOC100708655	Fetuin B	X	X		X	X
K7FKZ8[4]	-	Uncharacterized protein cathelicidin-like					X
K7FPA2[4]	-	Uncharacterized protein cathelicidin-like				X	
K7GI21[4]	-	Uncharacterized protein cathelicidin-like	X		X	X	X
K7GID5[4]	-	Uncharacterized protein cathelicidin-like	X		X	X	X
M7BBJ0[4]	UY3_13360	Uncharacterized protein cathelicidin-like	X			X	X
Q5M8F3[4]	*cst7*	Cystatin		X		X	X
R0LBE5[4]	*Anapl_03194*	Kininogen 1				X	
P82115[5]	*prtA*	Serralysin	X	X	X		X
Q94527[5]	Rel CG11992	Nuclear factor NF-kappa-B p110 subunit	X	X	X	X	X
B4R1J7[5]	*Anp*	Andropin	X	X	X	X	X
Q589Y5[5]	*ohsp1*	Serine protease	X	X	X	X	X
Q7PT80[5]	REL2 1270310	AGAP006747-PA	X	X	X	X	X
P05223[6]	-	Preprocaerulein type I	X				
Q45TR8[6]	-	Ubiquitin	X	X	X	X	X

Table S5

Peptide pheromones annotation and occurrence in the five sampled caecilian species.

Pheromone ID	Gene name	Protein description	*Rhinatrema bivittatum*	*Caecilia tentaculata*	*Typhlonectes compressicauda*	*Microcaecilia unicolor*	*Microcaecilia dermatophaga*
A0A0A0QT03	*Cloa_05*	Sodefrin-like factor beta isoform 05	X				
A0A0A0QTC2	*Cloa_19*	Sodefrin-like factor alpha isoform 19				X	
A0A0A0QTD8	*Cloa_02*	Sodefrin-like factor beta isoform 02		X			
A0A0A0QU60	*Cloa_08*	Sodefrin-like factor beta isoform 07	X				
A0A0A0QU84	*Cloa_26*	Sodefrin-like factor beta isoform 24				X	
A0A0A0QUZ5	*Cloa_10*	Sodefrin-like factor beta isoform 10		X			
A0A0A0QVR4	*Cloa_09*	Sodefrin-like factor alpha isoform 09	X	X		X	
A0A0A0QVU0	*Cloa_05*	Sodefrin-like factor beta isoform 04		X			
A0A0A0QVV1	*Cloa_30*	Sodefrin-like factor alpha isoform 28					X
A0A0A0QVV6	*Cloa_01*	Sodefrin-like factor beta isoform 01	X				X
A0A0B5GR37	-	Sodefrin-like factor 9	X				
A0A0B5H1E4	-	Sodefrin-like factor 1	X				
A0A0B5H3N9	-	Sodefrin-like factor 20		X			
A0A0B5H6P8	-	Sodefrin-like factor 15				X	X
A0A0E3KK02	SPF	Sodefrin-like factor	X		X	X	X
A0A0E3KK06	SPF	Sodefrin-like factor					X
A0A0E3N0I6	SPF	Sodefrin-like factor			X		
A0A0E3N2L5	SPF	Sodefrin-like factor	X	X		X	
A0A0E3N2L9	SPF	Sodefrin-like factor		X			
A0A0E3N3L3	SPF	Sodefrin-like factor		X			
A0A0E3N4Q2	SPF	Sodefrin-like factor		X			
A0A0F7JG78	-	Sodefrin-like factor			X		
A0A0F7JHQ7	-	Sodefrin-like factor			X		
A0A0F7JHR5	-	Sodefrin-like factor					X
A0A0F7JJU1	-	Sodefrin-like factor				X	
A0A125S9K3	SPF1	Sodefrin-like factor	X				
A0A125S9K5	SPF8	Sodefrin-like factor			X	X	X

A0A125S9K7	SPF2	Sodefrin-like factor					X	X
A0A125S9K8	SPF7	Sodefrin-like factor					X	
A0A125S9L0	SPF4	Sodefrin-like factor	X					
A0A125S9L4	SPF10	Sodefrin-like factor				X		
A0A125S9L5	SPF11	Sodefrin-like factor						X
A0A140IHG1	-	Sodefrin-like factor						X
A0A140IHG6	-	Sodefrin-like factor				X		X
A0A140IHH1	-	Sodefrin-like factor				X		
A0A140IHH2	-	Sodefrin-like factor			X			
A0A172AZC5	-	Sodefrin-like factor				X		
B2CM93	-	Sodefrin-like factor				X		
G3IMK8	I79_025148	Aphrodisin			X			
Q2EFD2	-	Sodefrin-like factor	X					
Q4FAD1	SPF1	Sodefrin-like factor					X	
Q4FAE1	SPF1	Sodefrin-like factor					X	
Q4FAG1	SPF1	Sodefrin-like factor						X

Conclusions

Genome-wide studies are revolutionizing biological sciences, and genomic resources of the entire diversity of species are needed for comparative analysis. One of the major gaps of the genomic information of vertebrates is found in caecilian amphibians. In this study we start filling this gap with five reference transcriptomes for five species of caecilians.

Comparisons of our reference transcriptomes with a database with the information of 51 other vertebrates has uncovered the incompleteness of vertebrate gene families and pointed out important unknown functional genomic elements for caecilians and/or amphibians, especially in the skin.

The study of adaptive evolution at the molecular level has unraveled several elements that were under positive selection in caecilian amphibians for some evolutionary time epoch. These elements likely underlie the particular biology of caecilian amphibians, being probably related to their fossorial habits, life history, and interactions with other organisms of the same or different species.

Skin expression analysis revealed the uniqueness of skin tissue type in caecilians. Elements associated with many vital functions in caecilians, such as movement, communication, and defence, are expressed in the skin.

RNA-seq is a powerful tool with many applications, as stated by this research study. Broad-scale transcriptome studies provide a useful platform for functional analysis in order to explore less well-known and enigmatic species like caecilian amphibians.

Curriculum vitae

MARÍA TORRES SÁNCHEZ
Personal information
Email: torressanchez.maria@gmail.com
Skype: María Torres Sánchez (archcortegada)
Date of birth: 16th of December of 1989

Professional experience

2014 - 2018 (4 years). **PhD candidate researcher** (official code: BES-2013-062723) in the Spanish National project: Comparative transcriptomics and gene discovery in the caecilian skin (official code: CGL2012-40082). Principal Investigator: Dr. Diego San Mauro, Complutense University of Madrid (UCM) and University of Barcelona (UB).

Research stays:

2017. Four month stay (official code: EEBB-I-17-12039) in the Natural History Museum (NHM), London, UK - supervisors: Dr. Mark Wilkinson and Dr. David Gower.

2016. Four month stay (official code: EEBB-I-16-11395) in the University of Nevada, Reno (UNR), US - supervisor: Dr. David Alvarez-Ponce.

2015. Four month stay (official code: EEBB-I-15-09665) in the Institute of Biological Environmental and Rural Sciences (IBERS) of Aberystwyth University, Wales, UK - supervisor: Prof. Chris Creevey.

Teaching collaborations:

2016/2017: Genetic diversity and population structure (35 hours) - MSc in Physical Anthropology: Evolution and Biodiversity in humans, Complutense University of Madrid (UCM), University of Alcalá (UAH) and Autonomous University of Madrid (UAM), supervisor: Dr. Antonio González Martín

New technologies in zoological research (20 hours) - MSc in Zoology, Complutense University of Madrid (UCM), supervisor: Dr. María Saura Álvarez.

MSc thesis co-mentoring: Molecular evolution of mudskipper aquaporins, Héctor Lorente Martínez (5 hours) - MSc in Evolutionary biology, Complutense University of Madrid (UCM), supervisors: Dr. Diego San Mauro and Dr. Ainhoa Agorreta.

2017/2018: New technologies in zoological research (20 hours) - MSc in Zoology, Complutense University of Madrid (UCM), supervisor: Dr. María Saura Álvarez.

2013 (3 months). **Molecular and Phytopathology laboratory technician**, assistance in detection of amphibian chytrid fungus and avian malaria, Research center for the Biodiversity Conservation (BIOCAMB) of Indoamerican Technological University of Quito (UTI).

Academic career

2015 - present. PhD candidate in Biology, Complutense University of Madrid (UCM) - supervisor: Dr. Diego San Mauro.

2015 - 2017. MSc in Bioinformatics and Biostatistics, Open University of Catalunya (UOC) and University of Barcelona (UB) - 60 ECTS.
MSc thesis: Epigenetic and embryological implications of the DNMT1 gene in marsupials - supervisor: Dr. David Alvarez-Ponce.

2012 - 2013. MSc in Biodiversity and Conservation in tropical areas, Menéndez Pelayo International University (UIMP) and Spanish National Research Council (CSIC) - 60 ECTS.
MSc thesis: Diversity of the family Ptilodactylidae (Coleoptera) along an elevational gradient in the Napo river basin, Ecuador - supervisors: Dr. Juan Manuel Guayasamín and Dr. Andrea Encalada.

2007 - 2012. BSc in Biology double specialization in Biotechnology and Molecular Biology, and Environmental Science, University of Santiago de Compostela (USC) - 361 ECTS.
BSc thesis: Proteomics: methods and applications - supervisor: Prof. Gonzalo Álvarez Jurado.

2004 - 2009. Professional Degree in Music specialization in Piano, Music Conservatory of Vilagarcía de Arousa (Spain).

Relevant complementary formation

Courses and Workshops

2017. CEE Autumn Symposium and Mixer, Linnean Society of London.

2017. UCL Genetics Short Course: Methods for Analysing Complex Trait GWAS Data.

2017. DNA-Barcoding course, Natural history museum of Paris (MNHN) and Distributed European school of taxonomy (DEST).

2017. AdaptNET workshop: Bioinformatics for Adaptation Genomics.

2017. Practical Course in Scientific journalism and divulgation, Complutense University of Madrid (OTRI-UCM).

2016. EMBO Practical Course: Regulatory small and long ncRNAS: Durat et lucet, Centre for Integrative Biology (CIBIO), University of Trento (UNITN).

2015. Workshop in Microbiological Identification by Mass Spectrometry MALDI Biotyper, Complutense University of Madrid (UCM).

2015. Poznan Summer School of Bioinformatics: Modern approaches to RNA analyses, Adam Mickiewicz University in Poznan (UAM).

2015. NERC-MDIBL Environmental Genomics course, University of Birmingham.

2013. NESCent Theory and Practice of Phylogenetic Inference course, Indoamerican Technological University of Quito (UTI).

2012. Workshop in Young Proteomics Researchers, Spanish Society of Proteomics (SEProt), University of Santiago de Compostela (USC).

Internships and other activities

2017. Volunteer monitoring amphibians and reptiles in the Illas Átlanticas national park, Galicia. Spanish Herpetological association (AHE).

2017. Field journey, discovering amphibian and reptile diversity, Dracula Reserve, Carchi, Ecuador. San Francisco University of Quito and Natural History Museum of London (USFQ & NHM).

2013. Volunteer monitoring sea turtles and sharks, Foundation Equilibrio Azul Manabí, Ecuador.

2012. Research assistant in training, Hidrobiological Station of Encoro do Con (EHEC), University of Santiago de Compostela (USC).

2011. Laboratory technician in training, Marine Control Institute (INTECMAR).

2010 - 2012. First year academic student advisor, University of Santiago de Compostela (USC).

Publications and Communications

Papers

2017. María Torres-Sánchez, David J. Gower, Christopher J. Creevey, Kim Roelants, Mark Wilkinson, Diego San Mauro. Chemical defence and communication underground? Insights into skin specialisations of caecilian amphibians from gene expression profiles. (In preparation)

2017. María Torres-Sánchez, David J. Gower, Christopher J. Creevey, David Alvarez-Ponce, Mark Wilkinson, Diego San Mauro. Behind the scenes: molecular innovations during caecilian amphibian evolution. (In preparation)

2017. Héctor Lorente, María Torres-Sánchez, Ainhoa Agorreta, Diego San Mauro. Molecular evolution of mudskipper aquaporins. (In preparation)

2017. David Alvarez-Ponce, María Torres-Sánchez, Felix Feyertag, Asmita Kulkarni, Taylen Nappi. Molecular evolution of DNMT1 in vertebrates: duplications in marsupials followed by positive selection. *BMC Evolutionary Biology*. (Submitted)

2017. María Torres-Sánchez, Christopher J. Creevey, Etienne Kornobis, David J. Gower, Mark Wilkinson, Diego San Mauro. Transcriptomic landscapes indicate expansion of vertebrate gene families in caecilian amphibians. *Proceedings of the Royal Society B: Biological Sciences*. (R1 under review)

Congress and Conferences

2017. María Torres-Sánchez. Chemical war underground? Antimicrobial peptides in caecilian amphibians. Venom Day, Bangor (Oral communication).

2017. María Torres-Sánchez, Christopher J. Creevey, David J. Gower, Mark Wilkinson, Diego San Mauro. The secret life of caecilian amphibians.

Symposium 13, Progress in Neotropical Caecilian Biology, XI Latin American Congress of Herpetology, Quito (Oral communication).

2017. María Torres-Sánchez and Diego San Mauro. Challenges of a powerful tool: RNA-seq for non-model species, the case of caecilian amphibians. Symposium 14, Applying genomic-scale approaches to the study of Neotropical amphibians and reptiles, XI Latin American Congress of Herpetology, Quito (Oral communication).

2017. María Torres-Sánchez, Christopher J. Creevey, David J. Gower, Mark Wilkinson, Diego San Mauro. Comparison of amino acid replacement ratios in vertebrates: insights into life history traits of caecilian amphibians (Gymnophiona). Symposium 13 Investigating ecological and evolutionary process with NGS, XIV MEDECOS & XIII AEET meeting, Sevilla (Oral communication).

2016. María Torres-Sánchez. Transcriptómica en cecilias (orden Gymnophiona). III Congress of New technologies in zoological research, Complutense University of Madrid (UCM), Madrid (Spanish oral communication).

2015. María Torres-Sánchez, Christopher J. Creevey, David J. Gower, Mark Wilkinson, Diego San Mauro. Comparative transcriptomics in caecilian amphibians: a multigene family approach. Annual meeting of Society for Molecular Biology and Evolution, Viena (Poster).

Projects, Grants and Awards

2017. Travel grant for XI Latin American Congress of Herpetology by Gans Collections and Charitable Fund (1,400 $).

2017. Personnel involves in the Labex CEBA (Center for the study of biodiversity in Amazonia) project: Secretomes and microbiomes of caecilian skin - CAECILISKIN. Principal Investigator: Dr. Mark Wilkinson. Natural History Museum (NHM), London, UK (10,000 €).

2014. PhD fellowship by Ministry of Economy and Competitiveness of Spain (official code: BES-2013-062723).
Research stays (official code of the financial support):
2017. EEBB-I-17-12039 (6,330 €)
2016. EEBB-I-16-11395 (6,280 €)
2015. EEBB-I-15-09665 (6,330 €)