scientific reports

OPEN

Evolutionary diversification of epidermal barrier genes in amphibians

Attila Placido Sachslehner & Leopold Eckhart✉

The epidermal differentiation complex (EDC) is a cluster of genes encoding components of the skin barrier in terrestrial vertebrates. EDC genes can be categorized as S100 fused-type protein (SFTP) genes such as *filaggrin*, which contain two coding exons, and single-coding-exon EDC (SEDC) genes such as *loricrin*. SFTPs are known to be present in amniotes (mammals, reptiles and birds) and amphibians, whereas SEDCs have not yet been reported in amphibians. Here, we show that caecilians (Amphibia: Gymnophiona) have both SFTP and SEDC genes. Two to four SEDC genes were identified in the genomes of *Rhinatrema bivittatum*, *Microcaecilia unicolor* and *Geotrypetes seraphini*. Comparative analysis of tissue transcriptomes indicated predominant expression of SEDC genes in the skin of caecilians. The proteins encoded by caecilian SEDC genes resemble human SEDC proteins, such as involucrin and small proline-rich proteins, with regard to low sequence complexity and high contents of proline, glutamine and lysine. Our data reveal diversification of EDC genes in amphibians and suggest that SEDC-type skin barrier genes have originated either in a common ancestor of tetrapods followed by loss in Batrachia (frogs and salamanders) or, by convergent evolution, in caecilians and amniotes.

The evolutionary transition from water to land was associated with the evolution of a proficient skin barrier in terrestrial vertebrates[1,2]. The barrier function of the skin depends largely on epithelial cells, i.e. keratinocytes, which proliferate in the basal layer of the epidermis and undergo differentiation in the suprabasal layers thereof[3]. Keratinocyte differentiation involves expression of high amounts of keratin intermediate filaments to enforce the cytoskeleton and other genes which serve specific barrier-associated functions, such as cross-linking of proteins, building intercellular connections and anti-microbial defense[2,4].

The epidermal differentiation complex (EDC)[5–7] is a cluster of genes which are expressed in the late phase of keratinocyte differentiation. The genes of the EDC can be categorized according to their exon–intron organisation and the domain structure of the encoded proteins. S100A genes, consisting of one non-coding and two protein-coding exons, are located at the borders of the EDC. The same exon–intron organization is found in genes encoding S100 fused-type proteins (SFTPs), whereas single-coding-exon EDC (SEDC) genes[8] have one non-coding and one coding exon. S100A's are involved in calcium-associated signal transduction[9], whereas SFTP and SEDC genes encode components of proteinaceous intracellular structures that provide mechanical stability to the outermost, cornified cell layers of the epidermis. The EDC was originally identified in the human and mouse genomes[5–7] and mutations in specific EDC genes, such as *filaggrin* (*FLG*) and *late cornified envelope* (*LCE*) *3B/3C* were found to be associated with prevalent skin diseases, atopic dermatitis and psoriasis, respectively[10,11]. Subsequently, EDC gene clusters were also identified in other mammals[12–14], birds[8,12,15,16], crocodiles[17], turtles[18], squamates[8,19], sphenodonts[20] and amphibians[21]. Among amphibians, only a small portion of the phylogenetic diversity of this clade was covered by previous research[22] because genome sequences were available from no other amphibians than frogs (Anura). The new availability of genome sequences of caecilians and a salamander allows us to extend the comparative analysis of the EDC in amphibians in the present study.

Caecilians (Amphibia: Gymnophiona) represent a phylogenetic lineage of amphibians which have diverged approximately 320 million years ago from batrachians (frogs and salamanders)[22,23]. They are limbless, have a burrowing lifestyle and can be found in tropical regions of America, Africa and Asia. The outermost cell layer of the epidermis of caecilians is cornified and covered by a layer of mucus produced by abundant skin glands[24,25]. In addition to these features shared with other amphibians, caecilian skin contains mineralized dermal scales[26]. Recently, genome sequences of three species of caecilians, i.e., the two-lined caecilian (*Rhinatrema bivittatum*), the Gaboon caecilian (*Geotrypetes seraphini*) and a representative of the family Siphonopidae (*Microcaecilia*

Skin Biology Laboratory, Department of Dermatology, Medical University of Vienna, Vienna, Austria. ✉email: leopold.eckhart@meduniwien.ac.at

unicolor), were reported[27–29]. However, genes regulating epidermal differentiation and particularly the EDC have not yet been characterized in caecilians.

Here, we analyzed the gene composition of the EDC of caecilians and compared it to the EDC of other amphibians and humans. We show that the EDC of caecilians differs from that of frogs and salamanders by the presence of an EDC gene class previously identified only in amniotes.

Results

Comparison of the EDC in caecilians, other amphibians and amniotes.   The EDC of three caecilians, *Rhinatrema bivittatum*, *Microcaecilia unicolor* and *Geotrypetes seraphini*, and two batrachians, the tropical clawed frog (*Xenopus tropicalis*) and the axolotl (*Ambystoma mexicanum*) was investigated for the presence of *SFTP* and *SEDC* genes. To identify *SFTP* and *SEDC* homologs, we followed a published approach involving tBLASTn searches using the S100 domains as queries for *SFTP* genes and de novo gene prediction from nucleotide sequences for finding *SEDC*'s[8]. The strategy is outlined in detail in the "Materials and methods" section. *SEDC* and *SFTP* genes were defined by their characteristic exon–intron structures with *SEDC* genes being comprised of a non-coding and a protein-coding exon and *SFTP* genes contain a non-coding and two coding exons (Fig. 1A). The presence of a non-coding exon upstream of the exon containing the coding sequence was confirmed by the identification of exon-spanning RNA-seq reads in the skin transcriptomes of caecilians (Fig. 1B). The location of the predicted *SFTP* and *SEDC* genes are summarized in Tables S1 and S2. Amino acid sequences of SFTP and SEDC proteins encoded by these genes are shown in Supplementary Fig. S1.

SFTP genes were identified in all species of amphibians investigated, whereas *SEDC*s were found only in caecilians. Two *SEDC* genes are present in *Rhinatrema bivittatum* and *Geotrypetes seraphini,* and 4 *SEDC*s are present in *Microcaecilia unicolor* (Fig. 1C). Each of the 3 species of caecilians investigated and *Xenopus tropicalis* have 2 *SFTP* genes. Three *SFTP*s were identified in the axolotl (Fig. 1C).

The locus of *SEDC*s in caecilians is not syntenic with that of *SEDC*s in humans and other amniotes[19], because *SEDC*s of caecilians are located between *S100A16* and *SFTP* genes, whereas *SEDC*s of amniotes are located between the *SFTP* genes and *S100A11* (Fig. 1C).

SFTP and SEDC genes of caecilians are expressed at higher levels in the skin than in other organs.   To compare the expression levels of EDC genes in different organs of caecilians, we performed a tBLASTn search for EDC genes in the transcriptomes of the organs of *Rhinatrema bivittatum*. SFTP1 (Fig. 2A), SFTP2 (Fig. 2B), SEDC1 (Fig. 2C) and SEDC2 (Fig. 2D) were abundantly detected in the skin but not, or only at minute levels, in other organs whereas a house-keeping gene was detected at similar levels in all organs (Fig. 2E). As the transcriptomes of all organs contain comparable numbers of RNA-sequencing reads (Fig. 2F), we conclude that EDCs genes of *Rhinatrema bivittatum* are predominantly expressed in the skin.
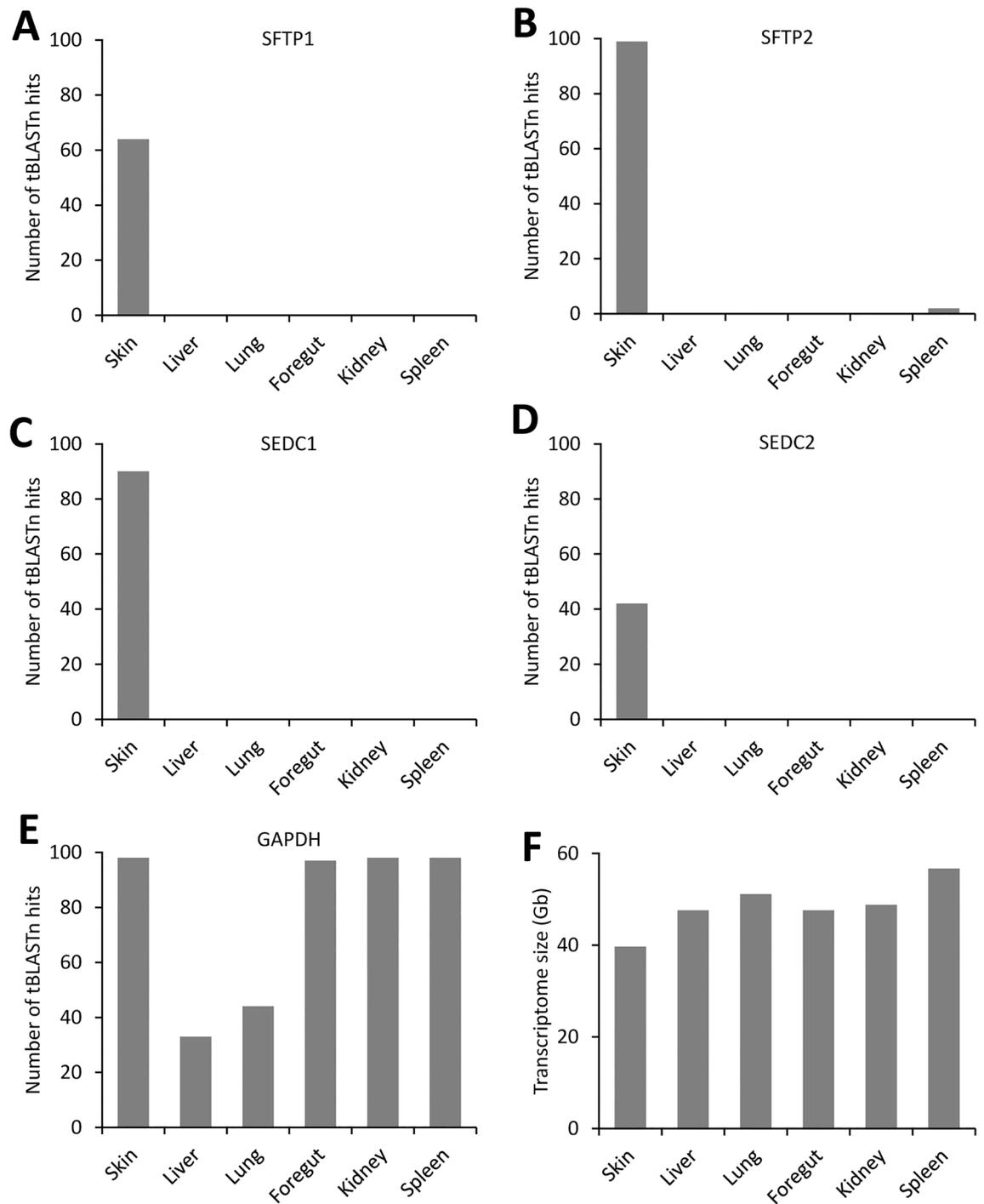
SEDC proteins of caecilians have high contents of proline, glutamine and lysine.   The amino acid sequences of SEDC proteins of caecilians are characterized by low sequence complexity (Fig. 3A, Supplementary Fig. S1) and the presence of sequence repeats (Fig. 3B, Supplementary Fig. S2). The amino acid contents of both SEDC and SFTPs of caecilians as well as those of SFTPs of *X. tropicalis* are biased towards proline, glutamine and lysine (Fig. 3C). Proline disrupts or terminates secondary structures in proteins, and glutamine and lysine are the targets of transglutamination, a process critical for epidermal cornification[4]. In addition, all three amino acids are associated with a high propensity to form disordered protein segments[30], suggesting that SEDC proteins and the carboxy-terminal domain of SFTPs of caecilians are intrinsically disordered proteins, thus resembling their counterparts in mammals[31–33]. Sequence alignment showed similarities between SEDC proteins of caecilians and small proline-rich proteins (SPRRs) of mammals (Fig. 3D). However, the sequence similarity was largely due to the similarly biased amino acid composition of these proteins and therefore did not represent strong evidence for homology. For this reason, amino acid sequence alignments of SEDC proteins were highly ambiguous, and we refrained from building phylogenetic models based on these alignments. By contrast, the sequences of the S100 domains at the amino-terminus of SFTPs could be aligned well, and molecular phylogenetics supported shared ancestry of SFTPs of amphibians and amniotes (Figs. S3, S4).

Discussion

The results of this study show that the organization of the EDC of caecilians (Gymnophiona) differs from that of representatives of frogs (Anura) and salamanders (Caudata) as only caecilians have *SEDC* genes. This finding indicates that the evolution of *SEDC*s was more complex than that of *SFTP*s, which exist in all major taxa of terrestrial vertebrates but not in fish[21], including lungfish (Supplementary Fig. S5), and therefore have evolved in a common ancestor of tetrapods[21]. To infer the evolutionary histories of *SEDC*s, we compared the presence or absence and the chromosomal arrangement of *SEDC*s relative to *SFTP* and *S100A* genes in the main taxa of terrestrial vertebrates (Fig. 4). The distribution of *SEDC* genes among modern amphibians and amniotes suggests two alternative evolutionary scenarios for the origin of *SEDC*s (Fig. 4). In scenario 1, *SEDC*s originated in a common ancestor of tetrapods and were conserved in amniotes and caecilians whereby a translocation of *SEDC* relative to *SFTP* genes occurred in either one of the two clades and the primordial *SEDC* gene was lost in batrachians (frogs and salamanders) (Fig. 4A). In scenario 2, *SEDC*s originated independently in the phylogenetic lineage leading to caecilians and in the phylogenetic lineage leading to amniotes (Fig. 4B). Due to the low sequence complexity of SEDCs, faithful sequence alignments required for phylogenetic analysis are not possible and it remains unclear which of the two scenarios is more likely to be correct. A hypothetical pathway of gene recombination and mutation events underlying an evolutionary trajectory from *S100A* to *SFTP* and *SEDC* genes is schematically depicted in Supplementary Fig. S6. Independently of the question about the relationship
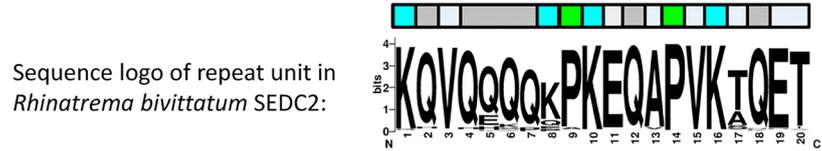
Figure 1. The EDC of caecilians contains SEDC genes. (**A**) Exon–intron structures of SEDC (single-coding-exon EDC) and S100 fused-type protein (SFTP) genes. Exons are shown as boxes in which the protein coding sequence (cds) is shaded black. (**B**) Alignment of the nucleotide sequences of the SEDC1 gene of *Rhinatrema bivittatum* (Rb) (GenBank accession number NC_042630.1, nucleotide positions are indicated above the sequence) and an RNA-seq read from the skin of this species (GenBank accession number: SRR5591419, experiment SRX2848310, read: gnl|SRA|SRR5591419.19486855.2). The TATA box is underlined. Only the first and last 5 nucleotides of the intron are shown. The sequence gap is indicated by "//". Non-coding sequences of exons are indicated by a black box and coding sequences are shown with white fonts on black background. The amino acid sequence of the translation product is shown underneath the nucleotide sequence. (**C**) Structure of the EDC in amphibians in comparison to the human EDC. The genes of the chromosomal segments bordered by the conserved genes *S100A11* and *S100A16* in *Rhinatrema bivittatum* (two-lined caecilian), *Microcaecilia unicolor* (a common caecilian from French Guayana), *Geotrypetes seraphini* (Gaboon caecilian), *Xenopus tropicalis* (tropical clawed frog), *Ambystoma mexicanum* (axolotl) and *Homo sapiens* (human) are schematically depicted by arrows pointing in the direction of gene transcription. SEDC genes are shown as red arrows. SFTP genes are shown as blue arrows. Grey arrows indicate *S100A11* and *S100A16* genes. White arrows indicate other S100 genes located between *S100A11* and *S100A16* in amphibians. The dashed line in the human EDC indicates that only a subset of SEDC and SFTP genes and no other S100A genes are shown for lack of space. The total numbers of human SEDC and SFTP genes are indicated.

of *SEDC*s of caecilians and amniotes, the presence of SEDCs in caecilians demonstrates that the evolutionary diversification of epidermal barrier genes is not restricted to amniotes[8], but has also occurred in the amphibian clade of terrestrial vertebrates.

Our finding that *SEDC* genes of caecilians are expressed in the skin but not, or only at low levels, in other organs, indicates that these EDC genes, like their counterparts in amniotes, are specifically important for the body's barrier against the environment, i.e. the primordial function of the skin[3,34]. Each species of caecilians investigated has at least two *SEDC* genes with distinct amino acid sequences (Fig. 3A). Further studies are

Figure 2. Semiquantitative analysis of EDC gene expression in tissue transcriptomes of *Rhinatrema bivittatum*. (**A**–**E**) Sequence fragments of the predicted proteins as described in the "Methods" section were used as queries for tBLASTn analysis. The accession numbers for the transcriptomes were as follows: skin (SRX2848310), liver (SRX2848294), lung (SRX2848293), foregut (SRX2848291), kidney (SRX2848286), spleen (SRX2848287). Default settings of tBLASTn were used except for deactivation of the filter for low complexity regions. tBLASTn hits with 100% sequence identity to the query were counted. *GAPDH* was investigated as a house-keeping gene (**E**). Note that this analysis allows comparison of expression levels of a particular gene in different organs, but not comparison between expression levels of different genes. (**F**) Size of tissue transcriptomes. *Gb* Gigabases.

Figure 3. Amino acid sequences of caecilian EDC proteins are rich in glutamine, lysine and proline. (**A**) Amino acid sequences of *Rhinatrema bivittatum* SEDC proteins. (**B**) Sequence logo of the repeat unit in *Rhinatrema bivittatum* SEDC2 based on the alignment of repeats shown in Supplementary Fig. S2. (**C**) Amino acid contents of EDC proteins of caecilians and *Xenopus tropicalis*. For comparison, the values for two human EDC proteins (SPRR1A, an SEDC protein, GenBank accession NP_001186757.1, and trichohyalin, TCHH, an SFTP, GenBank accession number NP_009044.2) are shown on the right. (**D**) Amino acid sequence alignment of *Rhinatrema bivittatum* SEDC1 and human SPRR1A. Glutamine, lysine and proline residues are highlighted by grey, blue and green shading, respectively.
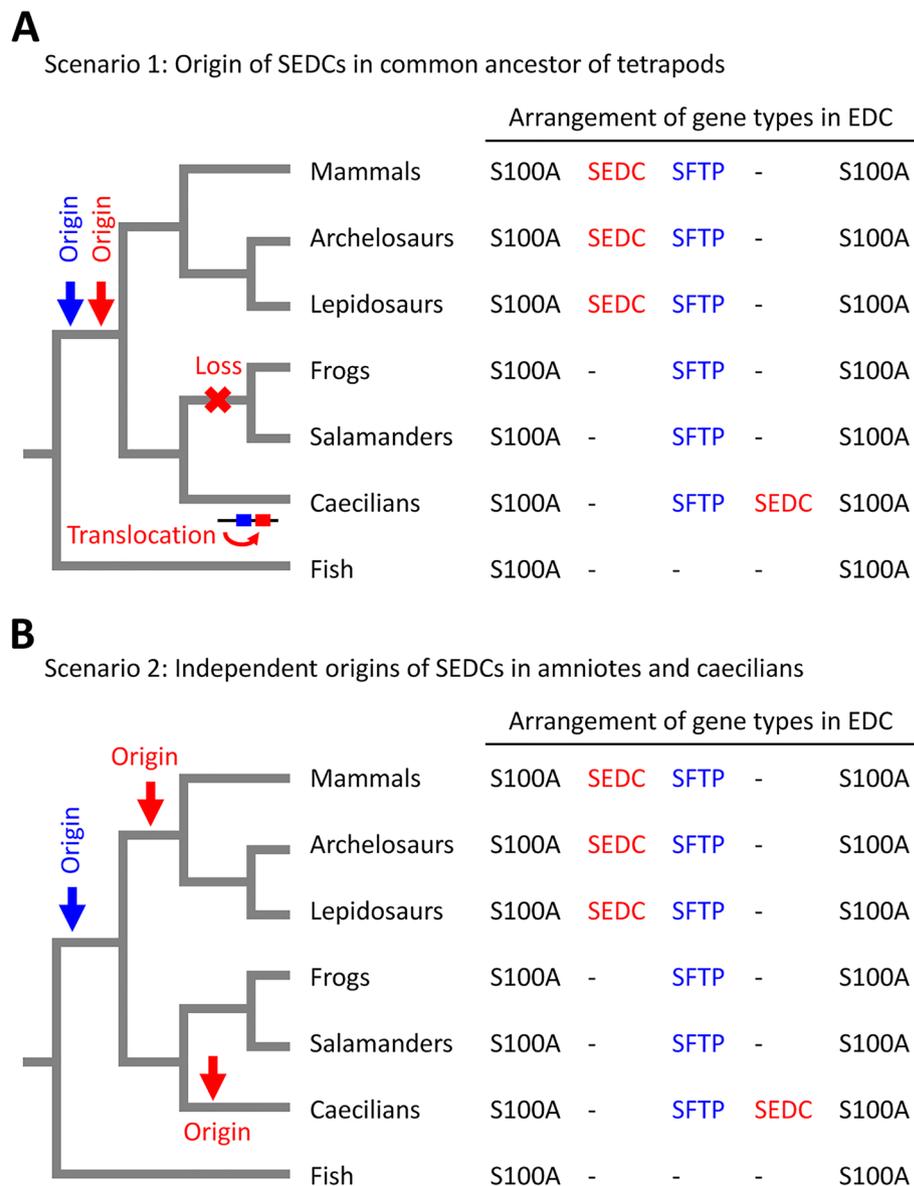
Figure 4. Scenarios for the evolution of SEDC genes in tetrapods. Two alternative scenarios (**A**, **B**) for the evolution of SEDC genes in terrestrial vertebrates are schematically depicted. Both scenarios are compatible with the arrangement of gene types in the epidermal differentiation complex (EDC) of caecilians and other main taxa of vertebrates. The relative arrangement of SEDC, SFTP and S100A genes within the EDC of each taxon is indicated on the right.

required to determine the expression pattern of individual *SEDC* genes in the various layers of the skin, during development of the skin and at different stages of the shedding cycle of the epidermis. It will be interesting to test whether individual *SEDC*s and *SFTP*s have unique or shared expression patterns in caecilians.

The amino acid sequences of SEDC proteins of caecilians resemble those of SPRRs and other proline-rich EDC proteins of amniotes[14,18]. The high contents of proline, glutamine and lysine indicate that these proteins lack secondary structures[30], thus representing intrinsically disordered proteins. The high contents of glutamine and lysine suggest that SEDCs serve as substrates for transglutamination, the main function of involucrin, loricrin and other mammalian SEDCs during epidermal cornification[4]. Interestingly, *Rhinatrema* SEDC2 contains many stretches of 3–5 consecutive glutamine residues, thereby resembling human involucrin. Besides these sequence similarities, there are also important differences between SEDCs of caecilians and amniotes. Sequence motifs that are conserved at the amino- and carboxy-termini of many SEDCs of amniotes[8,17], are not present in the SEDCs of caecilians. Furthermore, it is noteworthy that amniotes have a much greater diversity of SEDC proteins, including cornified beta-proteins (CBPs), also known as beta-keratins[8,35], glycine-rich proteins such as loricrin[36] and cysteine-rich proteins present in feathers[37,38]. The relatively low cysteine content of SEDC proteins of caecilians suggests that protein cross-linking through disulfide bonds is not critical for epidermal cornification in

these amphibians. The limited number and sequence diversity of SEDCs of caecilians corresponds to the rather uniform morphology of the epidermis of caecilians relative to amniotes.

The cornification of the epidermis is of great importance for caecilians and other amphibians[39]. The cornified layer protects the body in a terrestrial environment, it serves in feeding offspring of the caecilian, *Boulengerula taitanus*[40] and it is the site of infection with chytrid fungi, which are implicated in the global decline of populations of amphibians[41,42]. As the molecular basis for these functions and interactions of amphibian epidermis are largely unknown, the results of the present study contribute to a better understanding of the biology of amphibian skin.

Methods

Ethics statement.   No studies involving live animals were performed. Genomes and transcriptomes of vertebrates were investigated exclusively using sequences available in public databases.

Identification of EDC genes.   EDC genes were identified in the genome sequences of *Rhinatrema bivittatum*, GenBank accession number NC_042630.1, *Microcaecilia unicolor*, GenBank accession number NC_044044.1, and *Geotrypetes seraphini*, GenBank accession number NC_047099.1, submitted by Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK (Tables S1, S2)[27–29]. The EDC was also analyzed in the genomes of the tropical clawed frog (*Xenopus tropicalis*, GenBank accession number NC_030684.2, submitted by University of California, Berkeley, CA, USA) and the axolotl (*Ambystoma mexicanum*, GenBank accession number CM010927.2, submitted by Max Planck Society/University of Kentucky, Lexington, KY, USA). Furthermore, the locus of S100 genes on chromosome 8 of the West African lungfish (*Protopterus annectens*) was analyzed (GenBank accession number NC_056741.1, whole-genome assembly released by the Northwestern Polytechnical University, Xian, China)[43].

To identify EDC genes in the currently available genome sequences of amphibians, we followed a published protocol with modifications[8]. Briefly, the nucleotide sequence between the *S100A* genes, which are located on the borders of the EDC, was searched for *SFTP* and *SEDC* genes. *SFTP* genes were identified by tBLASTn search using the S100 domains of *Xenopus* SFTPs[21] as queries. Genes in which the coding sequence extended for more than 400 nucleotides over the end of the S100 domain region without coding for another defined structural domain were classified as *SFTP* genes. Of note, genes encoding two consecutive S100 domains, such as the *S100A11* genes of *Rhinatrema bivitattum* and *Xenopus tropicalis*, were classified as *S100A* genes. Candidate *SEDC* genes were searched by translating the entire nucleotide regions located between *SFTP* genes and the next *S100A* gene on each side of the *SFTP* genes. Open reading frames encoding polypeptides of 50 or more amino acid residues with a composition biased towards either glycine and serine (loricrin-like), glutamine (involucrin-like), proline (SPRR-like) or cysteine (EDCRP-like) were analyzed further. The open reading frames were considered *SEDC*-like if they were preceded by a *bona fide* splice acceptor motif within 50 nucleotides upstream of the putative start codon. To test if the predicted genes are transcribed, we performed BLAST searches in transcriptomes of caecilians that were available in the single-read archive (SRA) of the GenBank. The non-coding exon 1 of *SFTP* and *SEDC* genes was identified by alignment of transcript reads against the genomic DNA upstream of the coding region, followed by verification of the presence of a TATA box and a splice donor sequence. Some exon–intron borders of EDC genes of caecilians were correctly predicted by the gene prediction algorithm of the GenBank. However, the protein-coding potential of most genes was not identified by this algorithm. Accordingly, we corrected gene predictions where necessary (Figs. S1, S2). Orthology of genes was assessed using the criteria of shared local synteny and best reciprocal sequence similarity in BLAST.

Analysis of amino acid sequences encoded by EDC genes.   The predicted EDC genes were translated with the Expasy Translate tool[44] (https://web.expasy.org/translate/, last accessed on 3 March 2022). The resulting amino acids were aligned with MUSCLE, version 3.8.425 using default parameters at https://www.ebi.ac.uk/Tools/msa/muscle/ or MultAlin[45] followed by manual optimization, highlighting of sequence motifs and repeats. Percentage content of amino acids for each predicted sequence was determined with the Expasy Protparam tool[44] (https://web.expasy.org/protparam/, last accessed on 3 March 2022). Sequence logos were created with the Weblogo tool at https://weblogo.berkeley.edu/logo.cgi (last accessed on 3 March 2022).

Analysis of EDC gene expression in tissues of *Rhinatrema bivittatum*.   The expression of EDC genes was validated by tBLASTn searches in the tissue transcriptomes of *Rhinatrema bivittatum* (Skin: SRX2848310, liver: SRX2848294, lung: SRX2848293, foregut: SRX2848291, kidney: SRX2848286, spleen: SRX2848287). The following sequences were used as queries: *Rhinatrema bivittatum* SFTP1: amino acid positions 41–60, *Rhinatrema bivittatum* SFTP2: amino acid positions 1–40, *Rhinatrema bivittatum* SEDC1: amino acid positions 1–50, *Rhinatrema bivittatum* SEDC2: amino acid positions 41–60 (Supplementary Fig. S1), *Rhinatrema bivittatum* GAPDH (XP_029432574): amino acid positions 211–336. The tBLASTn search was run with default parameters, but the option "filter for low complexity regions" was not used. The number of hits in each tissue was determined, whereby only hits with 100% sequence identity to the query were counted.

Molecular phylogenetics.   Sequences of *SFTP* and *S100A16* genes for the phylogenetic analysis were downloaded from NCBI GenBank. The sequences were aligned using MAFFT (multiple alignment using fast fourier transform) (Version 7.427)[46] with the parameters "-maxiterate" set to 1000 and "-localpair". Sequences were trimmed manually with aliview[47], and only the sequences of S100 domains were aligned for the phylogenetic analysis (Supplementary Fig. S3). The model for amino acid replacement was calculated using prottest (Version 3.0)[48]. All available matrices (setting "-all-matrices") and models with rate variation among sites (set-

ting "-all-distributions") were included. The likelihood of the predicted models was assessed with the Akaike information criterion (setting "-sort A")[49]. The selected amino acid substitution model for the SFTP phylogeny was LG[50]. Maximum likelihood tree and bootstrap analysis (setting "-b 100") were performed using PHYML (Version 20120412). Tree topology, branch length, and rate parameters were optimized (setting "-o tlr")[51]. Phylogenetic trees (Supplementary Fig. S4) were visualized and annotated with FigTree (http://tree.bio.ed.ac.uk/software/figtree/, last accessed on 3 March 2022).

Data availability

All data generated or analyzed during this study are included in this published article and its Supplementary Information files.

References

1. Alibardi, L. Adaptation to the land: The skin of reptiles in comparison to that of amphibians and endotherm amniotes. *J. Exp. Zool. B Mol. Dev. Evol.* **98**, 12–41 (2003).
2. Matsui, T. & Amagai, M. Dissecting the formation, structure and barrier function of the stratum corneum. *Int. Immunol.* **27**, 269–280 (2015).
3. Eckhart, L. & Zeeuwen, P. L. J. M. The skin barrier: Epidermis vs environment. *Exp. Dermatol.* **27**, 805–806 (2018).
4. Candi, E., Schmidt, R. & Melino, G. The cornified envelope: A model of cell death in the skin. *Nat. Rev. Mol. Cell Biol.* **6**, 328–340 (2005).
5. Mischke, D., Korge, B. P., Marenholz, I., Volz, A. & Ziegler, A. Genes encoding structural proteins of epidermal cornification and S100 calcium-binding proteins form a gene complex ("epidermal differentiation complex") on human chromosome 1q21. *J. Invest. Dermatol.* **106**, 989–992 (1996).
6. Henry, J. *et al.* Update on the epidermal differentiation complex. *Front. Biosci.* **17**, 1517–1532 (2012).
7. Kypriotou, M., Huber, M. & Hohl, D. The human epidermal differentiation complex, cornified envelope precursors, S100 proteins and the "fused genes" family. *Exp. Dermatol.* **21**, 643–649 (2012).
8. Strasser, B. *et al.* Evolutionary origin and diversification of epidermal barrier proteins in amniotes. *Mol. Biol. Evol.* **31**, 3194–3205 (2014).
9. Eckert, R. L. *et al.* S100 proteins in the epidermis. *J. Invest. Dermatol.* **123**, 23–33 (2004).
10. Palmer, C. N. *et al.* Common loss-of-function variants of the epidermal barrier protein filaggrin are a major predisposing factor for atopic dermatitis. *Nat. Genet.* **38**, 441–446 (2006).
11. de Cid, R. *et al.* Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nat. Genet.* **41**, 211–215 (2009).
12. Vanhoutteghem, A., Djian, P. & Green, H. Ancient origin of the gene encoding involucrin, a precursor of the cross-linked envelope of epidermis and related epithelia. *Proc. Natl. Acad. Sci. USA* **105**, 15481–15486 (2008).
13. Jiang, Y. *et al.* The sheep genome illuminates biology of the rumen and lipid metabolism. *Science* **344**, 1168–1173 (2014).
14. Holthaus, K. B., Lachner, J., Ebner, B., Tschachler, E. & Eckhart, L. Gene duplications and gene loss in the epidermal differentiation complex during the evolutionary land-to-water transition of cetaceans. *Sci. Rep.* **11**, 12334 (2021).
15. Mlitz, V. *et al.* Trichohyalin-like proteins have evolutionarily conserved roles in the morphogenesis of skin appendages. *J. Invest. Dermatol.* **134**, 2685–2692 (2014).
16. Davis, A. & Greenwold, M. J. Evolution of an epidermal differentiation complex (EDC) gene family in birds. *Genes (Basel)* **12**, 767 (2021).
17. Holthaus, K. B. *et al.* Comparative analysis of epidermal differentiation genes of crocodilians suggests new models for the evolutionary origin of avian feather proteins. *Genome Biol. Evol.* **10**, 694–704 (2018).
18. Holthaus, K. B. *et al.* Comparative genomics identifies epidermal proteins associated with the evolution of the turtle shell. *Mol. Biol. Evol.* **33**, 726–737 (2016).
19. Holthaus, K. B. *et al.* Identification and comparative analysis of the epidermal differentiation complex in snakes. *Sci. Rep.* **7**, 45338 (2017).
20. Holthaus, K. B., Alibardi, L., Tschachler, E. & Eckhart, L. Identification of epidermal differentiation genes of the tuatara provides insights into the early evolution of lepidosaurian skin. *Sci. Rep.* **10**, 12844 (2020).
21. Mlitz, V., Hussain, T., Tschachler, E. & Eckhart, L. Filaggrin has evolved from an "S100 fused-type protein" (SFTP) gene present in a common ancestor of amphibians and mammals. *Exp. Dermatol.* **26**, 955–957 (2017).
22. Wilkinson, M. Caecilians. *Curr. Biol.* **22**, R668–R669 (2012).
23. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A resource for timelines, timetrees, and divergence times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).
24. Duellman, W. E. & Trueb, L. *Biology of Amphibians* (The Johns Hopkins University Press, 1994).
25. Mauricio, B. *et al.* Morphology of the cutaneous poison and mucous glands in amphibians with particular emphasis on caecilians (*Siphonops annulatus*). *Toxins (Basel).* **13**, 779 (2021).
26. Arun, D., Akbarsha, M. A., Oommen, O. V. & Divya, L. Light and transmission electron microscopic structure of skin glands and dermal scales of a caecilian amphibian *Gegeneophis ramaswamii*, with a note on antimicrobial property of skin gland secretion. *Microsc. Res. Tech.* **82**, 1267–1276 (2019).
27. Torres-Sánchez, M. *et al.* Multi-tissue transcriptomes of caecilian amphibians highlight incomplete knowledge of vertebrate gene families. *DNA Res.* **26**, 13–20 (2019).
28. Torres-Sánchez, M. *et al.* What lies beneath? Molecular evolution during the radiation of caecilian amphibians. *BMC Genom.* **20**, 354 (2019).
29. Torres-Sánchez, M., Wilkinson, M., Gower, D. J., Creevey, C. J. & San Mauro, D. Insights into the skin of caecilian amphibians from gene expression profiles. *BMC Genom.* **21**, 515 (2020).
30. Campen, A. *et al.* TOP-IDP-scale: A new amino acid scale measuring propensity for intrinsic disorder. *Protein Pept. Lett.* **15**, 956–963 (2008).
31. Shamilov, R., Robinson, V. L. & Aneskievich, B. J. Seeing keratinocyte proteins through the looking glass of intrinsic disorder. *Int. J. Mol. Sci.* **22**, 7912 (2021).
32. Avecilla, A. R. C. & Quiroz, F. G. Cracking the skin barrier: Liquid–liquid phase separation shines under the skin. *JID Innov.* **1**, 100036 (2021).
33. Quiroz, F. G. *et al.* Liquid–liquid phase separation drives skin barrier formation. *Science* **367**, eaax9554 (2020).
34. Hsu, Y. C., Fuchs, E. Building and maintaining the skin. *Cold Spring Harb. Perspect. Biol.* **14**, a040840 (2022).

35. Holthaus, K. B., Eckhart, L., Dalla Valle, L. & Alibardi, L. Review: Evolution and diversification of corneous beta-proteins, the characteristic epidermal proteins of reptiles and birds. *J. Exp. Zool. (Mol. Dev. Evol.)* **330**, 1–16 (2019).
36. Davis, A. C., Greenwold, M. J. & Sawyer, R. H. Complex gene loss and duplication events have facilitated the evolution of multiple loricrin genes in diverse bird species. *Genome Biol. Evol.* **11**, 984–1001 (2019).
37. Strasser, B., Mlitz, V., Hermann, M., Tschachler, E. & Eckhart, L. Convergent evolution of cysteine-rich proteins in feathers and hair. *BMC Evol. Biol.* **15**, 82 (2015).
38. Lachner, J. *et al.* Immunolocalization and phylogenetic profiling of the feather protein with the highest cysteine content. *Protoplasma* **256**, 1257–1265 (2019).
39. Fox, H. The skin of amphibia. Epidermis. In *Biology of the Integument* Vol. 2 (eds Bereither-Hahn, J. *et al.*) 78–110 (Springer, 1986).
40. Kupfer, A. *et al.* Parental investment by skin feeding in a caecilian amphibian. *Nature* **440**, 926–929 (2006).
41. Gower, D. J. *et al.* Batrachochytrium dendrobatidis infection and lethal chytridiomycosis in caecilian amphibians (Gymnophiona). *EcoHealth* **10**, 173–183 (2013).
42. Fisher, M. C. & Garner, T. W. J. Chytrid fungi and global amphibian declines. *Nat. Rev. Microbiol.* **18**, 332–343 (2020).
43. Wang, K. *et al.* African lungfish genome sheds light on the vertebrate water-to-land transition. *Cell* **184**, 1362–1376 (2021).
44. Artimo, P. *et al.* ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res.* **40**, W597-603 (2012).
45. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890 (1988).
46. Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* **30**, 3059–3066 (2002).
47. Larsson, A. AliView: A fast and lightweight alignment viewer and editor for large datasets. *Bioinformatics* **30**, 3276–3278 (2014).
48. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–2265 (2011).
49. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike* (eds Parzen, E. *et al.*) 199–213 (Springer, 1998).
50. Le, S. Q. & Gascuel, O. An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320 (2008).
51. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.* **52**, 696–704 (2003).

Acknowledgements

Author contributions

A.P.S. and L.E. conceived the study, performed genome and transcriptome sequence analyses and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18053-7.

Correspondence and requests for materials should be addressed to L.E.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.