Independent Pseudogenizations and Losses of Sox15 During Amniote Diversification Following Asymmetric Ohnolog Evolution

Yusaku Ogita

Kitasato University

Kei Tamura

Kitasato University

Shuuji Mawaribuchi

National Institute of Advanced Industrial Science and Technology

Nobuhiko Takamatsu

Kitasato University

Michihiko Ito (✉ ito@sci.kitasato-u.ac.jp)

Kitasato University

Research Article

Keywords: ohnolog, ortholog, 2R-WGD, pseudogene, gene loss, neofunctionalization, relax, dN/dS, marsupial, reptile

Posted Date: February 1st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-157747/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Independent pseudogenizations and losses of *sox15* during amniote diversification following asymmetric ohnolog evolution

Yusaku Ogita¹, Kei Tamura¹, Shuuji Mawaribuchi², Nobuhiko Takamatsu¹, & Michihiko Ito^{1*}

¹Department of Bioscience, School of Science, Kitasato University, 1-15-1 Kitasato, Minamiku Sagamihara, Kanagawa 252-0373, Japan

²Cellular and Molecular Biotechnology Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Central 6, 1-1-1 Higashi, Tsukuba, Ibaraki 305-8566, Japan

* Corresponding author, Michihiko Ito

E-mail address; ito@sci.kitasato-u.ac.jp

Abstract

Background

Four ohnologous genes (*sox1*, *sox2*, *sox3*, and *sox15*) were generated by two rounds of whole-genome duplication in a vertebrate ancestor. In eutherian mammals, *Sox1*, *Sox2*, and *Sox3* participate in central nervous system (CNS) development. *Sox15* functions in skeletal muscle regeneration, and has little functional overlap with the other three ohnologs. In contrast, *Xenopus* frog and zebrafish orthologs of *sox15* as well as *sox1-3s* are expressed and function in CNS development. We previously reported that *Sox15* is involved in mouse placental development as neofunctionalization, but is pseudogenized in marsupial opossum. These findings suggest that *sox15* might have evolved with unusual gene fates during vertebrate evolution. However, knowledge concerning *sox15* in other vertebrate lineages is scant. Our purpose was to clarify the fate and molecular evolution of *sox15* during vertebrate evolution.

Results

We searched for *sox15* orthologs in various vertebrate lineages by homology and synteny analyses using vertebrate genome databases. Interestingly, *sox15* was independently pseudogenized at least twice during species diversity in marsupial mammals. Moreover, we observed independent gene loss of *sox15* at least twice during reptile evolution in squamates and crocodile-bird diversification. Codon-based phylogenetic tree and selective analyses revealed the highest dN/dS value for *sox15* among the four ohnologs during jawed vertebrate evolution. The finding was supported by the high values in cartilaginous fishes, anuran amphibians, and amniotes. The high dN/dS value of *sox15* may have been mainly caused by a relaxed selection. Marsupial and squamate *sox15* may have evolved under more relaxed selection than those of eutherian mammals and testudine reptiles, respectively.

Conclusions

The findings revealed an asymmetric evolution of *sox15* among the four ohnologs during vertebrate evolution. Notably, independent pseudogenizations and losses of *sox15* were observed during marsupial and reptile evolution, respectively. Both might have been caused by strong relaxed selection. The drastic gene fates of *sox15*, including neofunctionalization and pseudogenizations/losses during amniote diversification, might be caused by a release from evolutionary constraints. We discuss why *sox15* has evolved under relaxed selection, considering the possible escapes from some constraints there could have been during its molecular evolution.

Keywords

ohnolog, ortholog, 2R-WGD, pseudogene, gene loss, neofunctionalization, relax, dN/dS, marsupial, reptile,

Background

In mammals, the *Sex-determining region Y* (*Sry*)-type high mobility group (HMG) box (*Sox*) family of genes includes approximately 20 members. The family is divided into eight groups (A–H) based on sequence identity in the DNA-binding domain HMG box [1, 2]. Interestingly, groups A and G contain only one member, *Sry* and *Sox15*, respectively. In contrast, most *Sox* groups are comprised of three closely related genes. Notably, *Sry* and *sox15* may have diverged independently from the *sox3* orthologs in common ancestors of therian mammals and vertebrates, respectively [3, 4]. *Sox3* shares higher sequence identity with *Sox1* or *Sox2* than *Sox15* in mammals, with group B1 comprising three members (*Sox1*, *Sox2*, and *Sox3*) [1, 5]. Vertebrate ohnologs are paralogs generated through the two rounds of whole-genome duplication (2R-WGD) in the common ancestor of vertebrates [6–8]. Synteny analysis data has revealed ohnologous relationships for these four members. The two ancestral genes *sox1/2* and *sox3/15* emerged from *soxB1* in the common ancestor of vertebrates in the first round of WGD, followed by the generation of these four genes in the second round of WGD [4]. Mammalian *Sox15* has orthologous relationships with teleost *sox19a/b* and amphibian *soxD* (now termed *sox15*), although *sox19a/b* and *soxD/sox15* belong to groups B1 and I, respectively [4]. The pairwise sequence identities between *sox15*, *sox19a/b*, and *soxD/sox15* are not particularly high, which likely resulted in the different gene names conferred during an era of limited genome information.

Both zebrafish *sox19a/b* and the African clawed frog *Xenopus laevis* *soxD/sox15* have high expression levels and function in central nervous system (CNS) development [4, 9, 10]. There are no reports of mammalian *sox15* involvement in neurogenesis. In contrast, mammalian *Sox15* is expressed in embryonic stem cells and satellite cells involved in muscle regeneration [11, 12]. Moreover, we described the specific expression of *Sox15* in the placenta during mouse embryogenesis in placenta-derived trophoblast giant cells and placental stem cells, and that *Sox15*

became pseudogenized in the marsupial opossum [5, 13, 14]. Based on the findings of *Sox15* orthologs in mammals and actinopterygian (bony) fish and amphibians, we concluded that *Sox15* evolved as a neofunctionalization-type gene for placental formation and/or myogenesis in the ancestor of eutherian mammals [5]. Therefore, there appears to be little functional overlap between *sox15* and the *sox19a/b* or *soxD/sox15* ortholog.

The collective findings indicate that *sox15* orthologs appear to have evolved during vertebrate evolution based on their expression and functions. However, little is known about *sox15* in aves, reptiles, non-eutherian mammals, non-anuran amphibians, and chondrichthyan fishes. To clarify the molecular evolution of *sox15* orthologs, we retrieved orthologous sequences of *sox15* from genome databases of amniote vertebrates, urodele/apoda amphibians, and chondrichthyan fishes, and performed synteny analysis. Notably, we identified independent losses and pseudogenizations of *sox15* in reptiles, including birds and marsupial mammals, respectively. The fates of *sox15* with independent pseudogenization/loss and neofunctionalization represent a rare example of gene fate during vertebrate evolution, although *bmp16* was recently reported as an example of independent gene losses among the three ohnologs [15, 16]. We then examined the evolutionary analysis and found that the highest dN/dS value of *sox15* among the four ohnologs and its independent pseudogenizations/loses during amniote diversification under relaxed selection.

Results

Independent pseudogenizations of *Sox15* during marsupial mammalian evolution

We previously reported that *Sox15* is a pseudogene in the marsupial opossum, although *Sox15* functions in eutherian mammals [13, 14]. Presently, to examine whether pseudogenization of *Sox15* occurred in the common ancestor of marsupials, we searched for *Sox15* orthologs in seven other marsupial genomes by genome browser and tblastn analysis using the amino acid (aa) sequences encoded by eutherian *Sox15* (Table S1) and performed synteny analysis using the *Fxr2* and *Mpdul* genes adjacent to *Sox15*. Annotated orthologs of *Sox15* or *Sox15*-like sequences were identified between *Fxr2* and *Mpdul* in each genome of all seven species. The *Sox15*-like sequence of Tammar wallaby (*Macropus eugenii*) did not contain its HMG box-coding region, while those of both Tasmanian devil (*Sarcophilus harrisii*) and fairy possum (*Gymnobelideus leadbeateri*) contained pseudogenization signatures evident as in-frame stop codons and frame shift mutations in the HMG box-encoding regions (Fig 1, 2a, and S1). However, the *Sox15*-like sequence of the other four species—common brushtail possum (*Trichosurus vulpecula*), thylacine (*Thylacinus cynocephalus*), koala (*Phascolarctos cinereus*), and wombat (*Vombatus ursinus*)—contained an intact 75 aa HMG box-encoding region (Fig 1, 2a, and S1). The predicted aa sequences from the koala and wombat *sox15* orthologs comprising two exons and one intron shared high sequence identities with that of eutherian mice (Fig. S2). In addition, their mRNA expressions in both species were found (https://www.ncbi.nlm.nih.gov/nuccore/XM_020966486.1 and https://www.ncbi.nlm.nih.gov/nuccore/XM_027857251.1). Importantly, two observations indicated that the pseudogenizations of the *Sox15* orthologs must have independently arisen in the three distinct ancestors of Tasmanian devil, fairy possum, and opossum. Firstly, the sequences adjacent to the in-frame stop codon and the frameshift mutation sites differed from each other.

Secondly, the most recent common ancestors of marsupials—the order *Dasyuromorphia* containing Tasmanian devil and thylacine, and the order *Phalangeriformes* containing fairy possum and common brushtail possum—harbored non-pseudogenized *Sox15* (Fig. 2a). These results indicate the independent pseudogenizations of *Sox15* during marsupial divergence.

Absence of pseudogenizations/losses of *Sox15* in both eutherian and monotreme mammals

We previously reported the potential neofunctionalization role of mouse *Sox15* in placental development [5, 13, 14]. In addition, Lee et al. [11] reported the involvement of *Sox15* in myogenesis. Based on our finding of *Sox15* pseudogenization during marsupial divergence, we searched for *Sox15* orthologs in NCBI gene databases from as many different eutherian mammalian genomes as possible. Annotations for *Sox15* were found in all 148 genomes examined. All the *Sox15* sequences contained the predicted open reading frames (ORFs). In addition, all the eutherian mammals, except polar bear (*Ursus maritimus*), harbored *Sox15* between *Mpdu1* and *Fxr2* (Table S2). There was a 1018 bp gap around the *Sox15* sequence in the *U. maritimus* genome database. These findings suggest the presence of *Sox15* as a neofunctionalization-type gene in the common ancestor of eutherian mammals, resulting in no or almost no pseudogenizations/losses of *Sox15* during eutherian evolution.

To clarify the fate of *Sox15* in monotreme mammals, we also searched for *Sox15* in the platypus genome, although its genome database has no annotation of *sox15*. A *Sox15*-like sequence encoding an intact 75 aa HMG box was identified between *Mpdu1* and *Fxr2* (Fig. 1).

Independent gene losses of *sox15* during reptilian evolution

Independent pseudogenizations of *sox15* during marsupial evolution were identified

(Fig. 1). Next, we searched for *sox15* orthologs in reptilian lineages using Chinese softshell turtle (*Pelodiscus sinensis*) or the common wall lizard (*Podarcis muralis*). The search revealed the aa sequences of *sox15* and its adjacent *fxr2* and *mpdul* genes as tblastn queries on genome databases of 34 squamata (lizard and snake), 22 testudines (turtle and tortoise), four crocodilian, and 503 bird species (Table S1).

In the 34 squamata genomes, the tblastn hits identified *sox15* orthologs in 17 species from five lineage families, including teiidae, lacertidae, agamidae, viperidae, and colubridae. No *sox15*-like sequences were found in the other 17 species (Fig. 2b and Table S2). Blastn-based synteny analysis using Easyfig [17] indicated that *sox15* was lost in nine of the 17 species from four lineage families: gekkonidae, varanidae, pythonidae, and elapidae (Fig. 1, 2b, S3a-o, and Table S2). We categorized the other eight as “unknown” because *fxr2* and/or *mpdul* orthologs were not detected (Fig. 2b, and Table S2).

Among the 22 testudines species, *sox15* orthologs were identified in 17 species. There were no *sox15*-like sequences in the other five species (Fig. 2c and Table S2). Easyfig analysis revealed that *sox15* was lost in the Pinta Island giant tortoise (*Chelonoidis abingdonii*) belonging to the testudinoidea superfamily (Fig. 2c, S3r-y, and Table S2). In the other four species, *sox15* was categorized as “unknown” (Fig. 2c and Table S2). In the four crocodilian genomes, no *sox15* orthologs were found on the tblastn search (Fig. 2c, S3p, q, and Table S2).

Easyfig analysis indicated that *sox15* was lost in *Alligator sinensis* and *Crocodylus porosus* belonging to the alligatoridae and crocodylidae families, respectively (Fig. 1, 2c and S3p, q). The other two species were categorized as “unknown” (Fig. 2c). No *sox15* orthologs were found in 503 bird genomes (Fig. 2c and Table S2). By synteny analysis using Easyfig, all the bird *sox15* orthologs were categorized as “unknown” because of was missing of *fxr2* (Fig. 1).

Collectively, these findings revealed independent gene losses of *sox15* at least twice during species diversity of squamates and crocodile-birds on reptile evolution.

Absence of gene losses/pseudogenizations of *sox15* in the examined amphibian and actinopterygian fish genomes

Orthologous relationships have been described between *sox15* among teleost fish *sox19a/b* and frog *soxD/sox15* [4]. Thus, we examined the fate of *sox15* orthologs in four amphibian and 12 actinopterygian fish species. All 16 genomes harbored *sox15* orthologs containing the predicted ORF (Fig. S4).

Absence of introns in cartilaginous fish *sox15*

We next examined the fate of *sox15/sox19* in cartilaginous fish (Chondrichthyes) and searched for *sox15* orthologs in eight elasmobranch and two holocephalan genomes (Table S1) using the amino acid sequence of polypterus *Erpetoichthys calabaricus* SOX19 as a query. We found *sox15*-like sequences in three elasmobranch species (*Pristis pectinata*, *Chiloscyllium punctatum*, and *Chiloscyllium plagiosum*), but not in the two holocephalan species. Synteny analysis of the elasmobranch *sox15*-like sequences based on FGENESH gene annotation and phylogenetic tree reconstruction of *soxB1/G* ohnologs indicated that the *P. pectinata* and *C. plagiosum* sequences should correspond to *sox15* orthologs (Fig. S4). In addition, the *C. punctatum* transcriptome database (<https://transcriptome.riken.jp/squalomix/blast/>) revealed that the *sox15* gene could be transcribed (Fig. S5).

Notably, a sequence comparison between the *sox15* transcript and its genomic region in *C. punctatum* indicated that the *sox15* gene consists of one exon, although almost all the *sox15* orthologs in mammals, reptiles, amphibians, and actinopterygian fish examined in this study

consisted of two exons with the same splicing sites. In contrast, no jawless fish orthologs of *sox15* were identified.

***sox15* shows the highest dN/dS value among the four ohnologs and higher dN/dS values in marsupials and reptiles than in eutherians**

Independent pseudogenizations and losses of *sox15* were evident only in amniote divergence and not in anamniotes (Fig. 1, 2, S3, S4, and Table S2). Additionally, *Sox15* in eutherian mammals could be a neofunctionalization-type gene [4]. Why did the drastic gene fates, including gene losses/pseudogenizations and neofunctionalization, occur in *sox15* during amniote divergence? To answer this question, we constructed phylogenetic trees of four *soxB1/G* ohnologue proteins—SOX1, SOX2, SOX3, and SOX15—in jawed vertebrates using maximum likelihood and Bayesian methods (Fig. 3 and S6). Each tree showed a relatively longer branch length in the SOX15 clade than on SOX1, SOX2, or SOX3. Interestingly, longer branch lengths of SOX15 (SOX19) on amniotes and amphibians were observed compared to the lengths on actinopterygian fish and those of SOX1, SOX2, and SOX3 on jawed vertebrates. These findings prompted the idea that there might be different selective pressures on the molecular history of *sox15* by evolutionary lineages in vertebrates, or in contrast to the other three ohnologs. To explore this idea we reconstructed a codon-based phylogenetic tree of jawed vertebrate *soxB1/G* ohnologs and calculated their ratio of nonsynonymous substitution sites per synonymous substitution site (dN/dS; ω) values on several vertebrate lineages (Fig. S7). The ω values of *sox1*, *sox2*, *sox3*, and *sox15* in jawed vertebrates were 0.0085, 0.0053, 0.0103, and 0.0233, respectively (Fig. 4). Likelihood ratio tests (LRTs) revealed that the ω value of *sox15* significantly deviated from those of the other three *soxB1/G* ohnologs ($p < 0.001$; Table 1). We next divided jawed vertebrates into three classes—chondrichthyes (cartilaginous fishes), actinopterygii (bony fishes),

and sarcopterygii (lobe-finned fish and tetrapod species)—and calculated each ω value of *sox15*. Importantly, actinopterygian *sox15* had a lower ω value (0.0074) than that of chondrichthyes (0.0641) and sarcopterygii (0.0443), which was statistically supported by χ^2 tests ($p < 1.0 \times 10^{-7}$; Table 1 and Fig. 4). Because the sarcopterygian class had different evolutionary distances of *sox15* in the codon-based phylogenetic tree, we further divided the class into three groups—coelacanth and non-anuran amphibians (Sar1), anuran amphibians (Sar2), and amniotes (Sar3)—and examined each ω value of *sox15* from a node of sarcopterygian common ancestors. These results and χ^2 test results revealed that the ω value of Sar1 was statistically significantly lower (0.0208) than those of Sar2 (0.0425) and Sar3 (0.0605) ($p < 0.05$; Table 1 and Fig. 4). These findings indicated that the higher ω value of *sox15* than *sox1*, *sox2*, and *sox3* could contribute to those of *sox15* in cartilaginous fishes, anuran amphibians, and amniotes. In particular, reptile *sox15* and marsupial non-pseudogenized *sox15* as well as pseudogenized *sox15* showed much higher values than those of eutherian and monotreme mammals (Fig. 5).

Relaxed selection has participated in an asymmetric evolution of *sox15* among the four ohnologous *soxB1/G* members and is involved in pseudogenizations/losses of *sox15*

To elucidate whether the molecular evolution of *sox15* is under relaxed or intensified selection, we performed RELAX tests of the gene in three classes of jawed vertebrates: chondrichthyes, actinopterygii, and sarcopterygii. *sox3* was used as a reference branch to calculate the relaxation or intensification parameter (k) of *sox15* in each class. The RELAX test data (Table 2) revealed that *sox15* might have evolved under relaxed selection in chondrichthyes and sarcopterygii ($k = 0.031$ and 0.50 , $p < 0.01$), and even in all jawed vertebrates ($k = 0.71$, $p < 0.01$). Although the k value of actinopterygian *sox15* was also lower than 1 ($k = 0.76$), the χ^2 test did not dismiss the null hypothesis of ' $k = 1$ ', suggesting that *sox15* and *sox3* have similarly evolved

under purifying selection in bony fishes. These results indicate that the relaxed selection might have contributed to the higher ω value of *sox15* than that of *sox3* during cartilaginous fish and sarcopterygian evolution, and induced an asymmetric evolution of *sox15* among the four ohnologous *soxB1/G* members during jawed vertebrate evolution. In addition, marsupial non-pseudogenized and squamate *sox15* orthologs were under more relaxed selection than those of eutherian mammals and testudine reptiles, respectively (Table 3, Fig. 5, and S8). Moreover, the pseudogenes of *Sox15* in marsupials were under more relaxed selection than intact ones in marsupials.

Discussion

The three *soxB1* subfamily genes (*sox1*, *sox2*, and *sox3*) and *sox15* share ohnologous relationships in vertebrates, but only *sox15* orthologs do not belong to the *soxB1* subfamily [5]. In this study, we examined the fate of the unique ohnologous member *sox15* during vertebrate evolution. Although we did not identify the *sox15* ortholog in jawless vertebrate genomes, we found an intron-free *sox15* in cartilaginous fish genomes. All the other orthologs of *sox15* examined in this study consisted of two exons. Because vertebrate *sox1*, *sox2*, and *sox3* are single exon genes, Okuda et al. [4] reported that *sox15* should have acquired an intron during vertebrate evolution after or on the second round of WGD. Our results suggest that *sox15* acquired an intron in the ancestor of bony (Osteichthyes) fish after the divergence between cartilaginous and bony fish (Fig. 6).

We previously reported neofunctionalization and pseudogenization of *Sox15* in eutherian mice and marsupial opossums, respectively [5, 13, 14]. Presently, *sox15* was independently pseudogenized during marsupial evolution and lost at least twice during reptile evolution (Figs 1 and 2). In contrast, there have been no reports of losses/pseudogenizations of *sox1*, *sox2*, or *sox3*. Why did only *sox15* orthologs have the drastic gene fates during amniote divergence among the four ohnologs? The selection analysis revealed the highest dN/dS value of *sox15* among the four ohnologs during jawed vertebrate evolution. In addition, the amniote dN/dS value of *sox15* was highest among the four classes (actinopterygii and Sar1-3) in jawed vertebrates (Fig. 4). These findings suggest that the high dN/dS value could be a key factor for the drastic gene fates of *sox15* during amniote divergence. Moreover, the RELAX test data showed that *sox15* appeared to be under more relaxed selection in marsupials and squamata than in eutherians and testudines, respectively (Fig. 4, Fig 5, and Table 1). The findings suggest that the independent losses and pseudogenizations of *sox15* might occur under relaxed selection.

During the molecular evolution of paralogs, especially ohnologs, what important factors for high protein evolution could lead to neofunctionalization and pseudogenization/loss? In general, ohnologs exhibit shared expression and functional redundancy prior to genome duplication, even in allotetraploidization by hybridization between two related species, as suggested by findings from the *Xenopus* frog [18].

Considering the redundancy of expression, although *sox1*, *sox2*, *sox3*, and *sox15* were expressed in the developing CNS in anamniotes, only *sox15* was not present in amniotes [4, 9, 19–21]. Notably, Kuma et al. demonstrated that brain-expressing genes evolve slowly [22]. Moreover, a recent report indicated that the coding sequences of ohnologs from the 2R-WGDs in the vertebrate ancestor are weakly biased for nervous system expression [23]. Therefore, it is possible that the CNS-expressing genes evolved under strong evolutionary constraints, and that *sox15* was released from the expression constraint in the stem of amniotes. Such genes might have been evolutionally constrained by conserved multi-functions of their proteins during vertebrate evolution.

Considering redundancy constraints, expression redundancy and/or partial functional redundancy in the CNS were observed in the four ohnologs in the anamniote zebrafish and *Xenopus* frogs [4, 9]. In contrast, *sox15* was released from expression and/or functional redundancy in the CNS, unlike *sox1*, *sox2*, and *sox3* in mice [21]. Since expression and functional redundancy have often been observed among the ohnologous genes in vertebrates, redundancy constraints may be involved in their mutual conservation during vertebrate evolution. Furthermore, complex developmental processes, such as CNS development, which emerged in the common ancestor of vertebrates, might have needed multiple ohnologous members, such as like *sox1*, *sox2*, *sox3*, and *sox15*, with redundancy constraints.

Conclusions

The evolutionary analyses revealed independent pseudogenizations and losses of *sox15* during marsupial, mammalian, and reptile evolution, respectively, although no pseudogenizations/losses of *sox15* were observed in the other classes of jawed vertebrates. *sox15* shows the highest dN/dS value among the four ohnologous *soxB1/G* members, and higher dN/dS values in marsupials and reptiles than in eutherians. Moreover, we found that relaxed selection has been involved in asymmetric evolution of *sox15* among the four ohnologs, which might have been one of the factors for the *sox15*'s drastic gene fates including pseudogenizations, losses, and neofunctionalization during amniote diversification. We propose that *sox15* might have been released from some evolutionary constraints including expression and/or functional redundancy in the CNS, unlike *sox1*, *sox2*, and *sox3* in the common ancestor of amniotes.

Methods

Identification of *sox15* orthologs from genome databases

Genome sequences from >100 vertebrate species were downloaded in FASTA format from NCBI (<https://www.ncbi.nlm.nih.gov/>). Local databases were created using BLAST (v.2.9.0+; https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=Download). Gene searches using tblastn [24] were performed using the default option. Sequence similarity of the top hit sequence to the SOX15 sequence in the database of the most similar species was analyzed using megablast (https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome) using BLAST, followed by manual synteny analysis. Gene annotation of *sox15* orthologs was performed with Softberry-FGENESH (http://www.softberry.com/berry.phtml?topic=case_study_animal&no_menu=on) [25] based on a species-specific gene-finding parameter [26, 27].

Molecular phylogenetic analysis

The coding sequences of *sox1*, *sox2*, *sox3*, and *sox15* in various vertebrate species were translated to aa sequences. Multiple alignment of results was performed using MAFFT version 7.427 (<https://mafft.cbrc.jp/alignment/software/>) [28]. Multiple alignments of nucleotide sequences were deduced using Pal2nal version 14 (<http://www.bork.embl.de/pal2nal>) [29] with the aa alignments. Ambiguous sites of the alignments were removed using trimAl version 1.2 (<http://trimal.cgenomics.org/>) [30] with option-gappyout. Maximum likelihood trees were inferred by IQ-TREE [31], where the most fitting aa and nucleotide substitution rate was estimated by ModelFinder in part of analysis by IQ-TREE [32]. A Bayesian phylogenetic tree

was inferred by MrBayes 3.2.7a (<https://nbisweden.github.io/MrBayes/download.html>) [33]. Two MCMC chains were run 300,000 times and sampled 100 times to analyze the convergence of the statistics by Tracer version 1.7.1 (<http://beast.community/tracer>). dN/dS was estimated by codeml in Paml4.8 (<http://abacus.gene.ucl.ac.uk/software/paml.html>) [34]. A branch model was used to calculate the dN/dS ratio for each group. RELAX [35] in HYPHY version 2.5.8 (<https://github.com/veg/hyphy>) was used to detect the relaxed or intensified selection for the *sox15* ortholog in each lineage.

Statistical analysis

χ^2 tests were performed using Microsoft Excel for Mac version 16.42 for LRT value in codeml and RELAX analysis.

Availability of data and materials

All data generated or analysed during this study are included in the main manuscript, figures, tables, and supplementary information file. Raw data of phylogenetic tree inference and analyses of codeml and RELAX are described in Fig. S9 on the supplementary information file.

References

1. Wegner M. From head to toes: The multiple facets of Sox proteins. *Nucleic Acids Research*. 1999;27:1409–20.
2. Bowles J, Schepers G, Koopman P. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Dev Biol*. 2000;227:239–55.
3. Foster JW, Marshall Graves JA. An SRY-related sequence on the marsupial X chromosome: Implications for the evolution of the mammalian testis-determining gene. *Proc Natl Acad Sci U S A*. 1994;91:1927–31.
4. Okuda Y, Yoda H, Uchikawa M, Furutani-Seiki M, Takeda H, Kondoh H, et al. Comparative genomic and expression analysis of group B1 sox genes in zebrafish indicates their diversification during vertebrate evolution. *Dev Dyn*. 2006;235:811–25.
5. Ito M. Function and molecular evolution of mammalian Sox15, a singleton in the SoxG group of transcription factors. *Int J Biochem Cell Biol*. 2010;42:449–52.
6. Ohno S. Evolution by gene duplication. 1970.
7. Holland PWH, Garcia-Fernandez J, Williams NA, Sidow A. Gene duplications and the origins of vertebrate development. *Development*. 1994;120 SUPPL.:125–33.
8. Dehal P, Boore JL. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol*. 2005;3.
9. Mizuseki K, Kishi M, Shiota K, Nakanishi S, Sasai Y. SoxD: An essential mediator of induction of anterior neural tissues in *Xenopus* embryos. *Neuron*. 1998;21:77–85.
10. Okuda Y, Ogura E, Kondoh H, Kamachi Y. B1 SOX coordinate cell specification with patterning and morphogenesis in the early zebrafish embryo. *PLoS Genet*. 2010;6:36.
11. Lee H-J, Goring W, Ochs M, Muhlfeld C, Steding G, Paprotta I, et al. Sox15 Is Required for Skeletal Muscle Regeneration. *Mol Cell Biol*. 2004;24:8428–36.

12. Maruyama M, Ichisaka T, Nakagawa M, Yamanaka S. Differential roles for Sox15 and Sox2 in transcriptional control in mouse embryonic stem cells. *J Biol Chem.* 2005;280:24371–9.
13. Yamada K, Kanda H, Tanaka S, Takamatsu N, Shiba T, Ito M. Sox15 enhances trophoblast giant cell differentiation induced by Hand1 in mouse placenta. *Differentiation.* 2006;74:212–21.
14. Yamada K, Kanda H, Aihara T, Takamatsu N, Shiba T, Ito M. Mammalian Sox15 Gene: Promoter Analysis and Implications for Placental Evolution . *Zoolog Sci.* 2008;25:313–20.
15. Feiner N, Meyer A, Kuraku S. Evolution of the vertebrate Pax4/6 class of genes with focus on its novel member, the Pax10 gene. *Genome Biol Evol.* 2014;6:1635–51.
16. Feiner N, Motone F, Meyer A, Kuraku S. Asymmetric paralog evolution between the “cryptic” gene Bmp16 and its well-studied sister genes Bmp2 and Bmp4. *Sci Rep.* 2019;9:1–13. doi:10.1038/s41598-019-40055-1.
17. Sullivan MJ, Petty NK, Beatson SA. Easyfig: A genome comparison visualizer. *Bioinformatics.* 2011;27:1009–10.
18. Session AM, Uno Y, Kwon T, Chapman JA, Toyoda A, Takahashi S, et al. Genome evolution in the allotetraploid frog *Xenopus laevis*. *Nature.* 2016;538:336–43. doi:10.1038/nature19840.
19. Prior HM, Walter MA. Sox genes: Architects of development. *Mol Med.* 1996;2:405–12.
20. Uchikawa M, Kamachi Y, Kondoh H. Two distinct subgroups of Group B Sox genes for transcriptional activators and repressors: Their expression during embryonic organogenesis of the chicken. *Mech Dev.* 1999;84:103–20.
21. Wood HB, Episkopou V. Comparative expression of the mouse Sox1, Sox2 and Sox3 genes from pre-gastrulation to early somite stages. *Mech Dev.* 1999;86:197–201.
22. Kuma K, Iwabe N, Miyata T. Functional constraints against variations on molecules from the tissue level: slowly evolving brain-specific genes demonstrated by protein kinase and

- immunoglobulin supergene families. *Mol Biol Evol.* 1995;12:123–30.
23. Roux J, Liu J, Robinson-Rechavi M. Selective constraints on coding sequences of nervous system genes are a major determinant of duplicate gene retention in vertebrates. *Mol Biol Evol.* 2017;34:2773–91.
 24. Gerts EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biol.* 2006;4:1–14.
 25. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* 2006;7 Suppl 1 Suppl 1:1–12.
 26. Pyron RA, Burbrink FT, Wiens JJ. A phylogeny and revised classification of Squamata, including 4161 species of lizards and snakes. *BMC Evol Biol.* 2013;13:1. doi:10.1186/1471-2148-13-93.
 27. Crawford NG, Parham JF, Sellas AB, Faircloth BC, Glenn TC, Papenfuss TJ, et al. A phylogenomic analysis of turtles. *Mol Phylogenet Evol.* 2015;83:250–7. doi:10.1016/j.ympev.2014.10.021.
 28. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol.* 2013;30:772–80.
 29. Suyama M, Torrents D, Bork P. PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 2006;34 WEB. SERV. ISS.:609–12.
 30. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics.* 2009;25:1972–3.
 31. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.*

2015;32:268–74.

32. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14:587–9.

33. Ronquist F, Teslenko M, Van Der Mark P, Ayres DL, Darling A, Höhna S, et al. Mrbayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 2012;61:539–42.

34. Yang Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24:1586–91.

35. Wertheim JO, Murrell B, Smith MD, Pond SLK, Scheffler K. RELAX: Detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol*. 2015;32:820–32.

Acknowledgements

We would like to thank Editage (www.editage.com) for English language editing. This work was supported by Grant-in-Aid for Scientific Research, Japan Society for the Promotion of Science (18K06389) to MI.

Authorship

Y. O., S. M., and M. I. designed the study; Y. O., K. T., N. T., and M. I. analyzed the data; and Y. O. and M. I. wrote the paper.

Declaration of interests

The authors declare no competing interests.

Figures

Figure 1

Incomplete retention of *sox15* during amniote evolution Summary of syntenic analysis of reptile and mammalian orthologs of *sox15* using 24 amniote species. *H. sapiens*, Homo sapiens; *L. africana*; Loxodonta africana; *M. domestica*, Monodelphis domestica; *T. vulpecula*, Trichosurus vulpecula; *S. harrisii*, Sarcophilus harrisii; *V. ursinus*, Vombatus ursinus; *P. cinereus*, Phascolarctos cinereus; *T.*

cynocephalus, *Thylacinus cynocephalus*; *G. leadbeateri*, *Gymnobelideus leadbeateri*; *O. anatinus*, *Ornithorhynchus anatinus*; *P. sinensis*, *Pelodiscus sinensis*; *G. evgoodei*, *Gopherus evgoodei*; *A. sinensis*, *Alligator sinensis*; *C. porosus*, *Crocodylus porosus*; *N. harrisi*, *Nannopterum harrisi*; *C. moneduloides*, *Corvus moneduloides*; *P. picta*, *Paroedura picta*; *P. muralis*, *Podarcis muralis*; *V. komodoensis*, *Varanus komodoensis*; *V. berus*, *Vipera berus*; *P. vitticeps*, *Pogona vitticeps*; *P. obsoletus*, *Pantherophis obsoletus*; *N. naja*, *Naja naja*; and *L. colubrina*, *Laticauda colubrina*.

Figure 2

Independent pseudogenizations or losses of *sox15* during marsupial or reptilian speciation (a)
Independent pseudogenization of *sox15* during marsupial speciation. (Ψ) indicates pseudogenization of *sox15*. Numbers from the first nucleotide in 225 nucleotide sequences encoding the HMG box are shown in the nucleotide alignment. In-frame stop codons and deletions with frame shift mutation are highlighted by gray boxes. Asterisks denote identical nucleotides among seven species following alignment. MRCA denotes most recent common ancestor. (b, c) Independent losses of *sox15* in two lineages during reptilian speciation: squamata including lizards and snakes (b) and archosauromorpha including testudines, crocodilians, and birds (c). “num. species” denotes the number of species examined in each lineage. Presence or loss of *sox15* is shown as + or -. “unknown” indicates that the existence of *sox15* was not determined in this analysis.

Figure 3

Phylogenetic relationships of vertebrate soxB1/G ohnologous proteins on the maximum likelihood method. A total of 73 aa sequences containing 287 sites were used for this tree inference. The JTT + F + I + G4 model was selected as the best-fit model in this dataset and used for the inference. The invertebrate SOXB1 clade was rooted. Values of the 1000 times ultrafast bootstrap test and the Bayesian posterior probability are shown at each node as "UB/PP." "-" indicates the node was not supported by Bayesian inference.

Figure 4

dN/dS (ω) values of four soxB1/G ohnologs (sox15 and sox1-3s) during vertebrate evolution. A total of 98 gap-containing 831 nucleotide sequences corresponding to 277 codons were used for this tree inference. The GTR + F + R5 model was selected as the best-fit model in this dataset and used for the inference. dN/dS (ω) values were calculated using 264 nucleotide sites with gaps deleted on the same nucleotide alignment as the tree inference. The scale bar indicates nucleotide substitutions per site.

Figure 5

dN/dS (ω) values of *sox15* in mammalian and reptilian lineages. A total of 225 nucleotide sites encoding the HMG box of SOX15 from 26 mammalian and 31 reptile species were used for dN/dS calculation. The right table shows the ω values calculated for each of the ω groups. The colors in species trees of mammals (left) and reptiles (middle) correspond to those in the table. “ Ψ ” denotes pseudogenization of *sox15* in marsupials.

Figure 6

Proposed model for molecular evolution of sox15 and its ohnologous members sox1- 3s during vertebrate evolution.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Tabls13.pdf](#)
- [sox15Supplementary0113.pdf](#)