PLOS ONE

RESEARCH ARTICLE

Dimerization and Transactivation Domains as Candidates for Functional Modulation and Diversity of Sox9

Marcos Tadeu Geraldo[1], Guilherme Targino Valente[2], Rafael Takahiro Nakajima[1], Cesar Martins[1]*

1 Integrative Genomics Laboratory, Department of Morphology, Institute of Biosciences, Sao Paulo State University–UNESP, Botucatu, SP, 18618–000, Brazil, **2** Systems Biology and Genomics Laboratory, Department of Bioprocess and Biotechnology, Agronomical Science Faculty, Sao Paulo State University–UNESP, Botucatu, SP, 18610–307, Brazil

* cmartins@ibb.unesp.br

OPEN ACCESS

Citation: Geraldo MT, Valente GT, Nakajima RT, Martins C (2016) Dimerization and Transactivation Domains as Candidates for Functional Modulation and Diversity of Sox9. PLoS ONE 11(5): e0156199. doi:10.1371/journal.pone.0156199

Editor: Klaus Roemer, University of Saarland Medical School, GERMANY

Received: February 25, 2016

Accepted: May 10, 2016

Published: May 19, 2016

Data Availability Statement: Sequence data has been submitted to GenBank: accession number KT201670.

Funding: This study was supported by São Paulo Research Foundation (FAPESP) [grant numbers 2009/05234-4; 2010/13143-6] and National Counsel of Technological and Scientific Development (CNPq) [grant number 475147/2010-3]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Sox9 plays an important role in a large variety of developmental pathways in vertebrates. It is composed of three domains: high-mobility group box (HMG box), dimerization (DIM) and transactivation (TAD). One of the main processes for regulation and variability of the pathways involving Sox9 is the self-gene expression regulation of Sox9. However, the subsequent roles of the Sox9 domains can also generate regulatory modulations. Studies have shown that TADs can bind to different types of proteins and its function seems to be influenced by DIM. Therefore, we hypothesized that both domains are directly associated and can be responsible for the functional variability of Sox9. We applied a method based on a broad phylogenetic context, using sequences of the HMG box domain, to ensure the homology of all the Sox9 copies used herein. The data obtained included 4,921 sequences relative to 657 metazoan species. Based on coevolutionary and selective pressure analyses of the Sox9 sequences, we observed coevolutions involving DIM and TADs. These data, along with the experimental data from literature, indicate a functional relationship between these domains. Moreover, DIM and TADs may be responsible for the functional plasticity of Sox9 because they are more tolerant for molecular changes (higher Ka/Ks ratio than the HMG box domain). This tolerance could allow a differential regulation of target genes or promote novel targets during transcriptional activation. In conclusion, we suggest that DIM and TADs functional association may regulate differentially the target genes or even promote novel targets during transcription activation mediated by Sox9 paralogs, contributing to the subfunctionalization of Sox9a and Sox9b in teleosts.

Introduction

The Sox proteins are involved in many developmental processes [1] across different metazoan groups such as insects, nematodes, amphibians, reptiles, birds and mammals [2]. These

proteins are members of a non-canonical family of DNA-binding domain known as the high-mobility group box (HMG box), consisting of approximately 80 residues in a twisted L-shape structure that binds to the minor groove of DNA [3]. The HMG proteins were discovered as acid-extractable chromatin components of high electrophoretic mobility [4], and its superfamily has been divided into three unrelated families, known as HMGA, HMGB and HMGN based on the systematic reference to their different DNA-binding domains [5, 6]. Canonical HMGs are architectural proteins and do not contain transactivation domains, in contrast to some transcription factors with non-canonical HMG-motifs that contain transactivation domains [7]. The Sox family proteins contain only one non-canonical HMG box domain with only about 20% sequence identity to the canonical DNA-binding domain of HMGB proteins [8]. The first Sox protein discovered was Sry (Sex determining region Y) [9] and since then around 20 Sox proteins have been identified in mice and humans. In general, the Sox proteins have been grouped based on the sequence and structural similarity of their HMG box domain [10].

Sox9 (Sex Determining Region Y-Box 9), a member of the SoxE subgroup (comprised of Sox8, Sox9 and Sox10), plays an important role in a variety of developmental processes of mesoderm (cartilage [11, 12] and male gonad [13–15]), ectoderm (central nervous system [16, 17], neural crest [18, 19] and retina [20]) and endoderm (pancreas [21, 22], liver [22] and intestine [23]). Therefore, Sox9 is a broad regulator and it has been related to different developmental disorders. For instance, heterozygous Sox9 mutations cause campomelic dysplasia, a syndrome characterized by skeletal malformations and associated, in some cases, with XY sex reversal [13, 24]. In addition, it is related to some acquired diseases, such as fibrosis [25], sclerosis [26], tumorigenesis [27] and cancer [28].

Although Sox9 is encoded by a single-copy gene in most vertebrates, studies have shown the existence of duplicate copies in teleosts [29] (hereafter named as *Sox9a* and *Sox9b*). According to recent data of gene expression in *Danio rerio* (zebrafish) and *Oryzias latipes* (medaka), the two copies seem to have undergone a lineage-specific subfunctionalization process [2, 30].

In addition to the HMG box domain, Sox9 is composed of a self-dimerization domain (DIM) [31, 32], and two transactivation domains (TAD)–K2 and PQS [29, 33, 34]. In humans and mice, there is an additional TAD (known as PQA), which enhances the PQS transactivation activity but it is unable to activate transcription alone [34]. Additionally, since PQA is exclusive of mammals, it has been suggested that this domain is related to the SRY sex-determining mechanism of these organisms [35]. These domains from Sox9 have also been found in Sox8 and Sox10, and the annotations used for them are still nonconsensual in the literature. Therefore, we propose the aforementioned nomenclature compiled from different studies related to the SoxE proteins [31, 36–38].

Studies have evaluated the effects of amino acids substitutions, deletions, frameshifts and truncations in Sox9 activity and have determined the functions of each domain [29, 31–34, 37]. Some questions arose based on these works: (i) Is there a functional and/or structural relationship between the domains?; (ii) Which domains can be responsible for the functional diversity observed in Sox9? Regarding the second question, we hypothesized that DIM and TADs are directly associated and responsible for the functional variability observed in Sox9. The premise for this hypothesis is that these domains may interfere and/or mediate the interactions with a distinct number of proteins. We believe that this binding plasticity could be a good regulatory strategy for transcriptional activation and repression.

A detailed analyses of coevolutions of Sox9 residues (including Sox9a and Sox9b) along with selective pressure calculations (Ka/Ks ratio) suggested that DIM and TADs are functionally or structurally associated and can be candidates to modulate the variability of Sox9 function.

Results

Homology relationships of *Sox9*

The phylogenetic reconstruction of the HMG box sequences showed the SoxE subgroup with three well-supported branches comprising Sox8, Sox9 and Sox10 (S1 Fig).

The sequences from the *Sox9* branch were selected for a more detailed evolutionary analysis. An exclusive *Sox9* phylogeny was generated and clearly showed three main clades, corresponding to *Sox9*, *Sox9a* and *Sox9b*, and only the teleost species exhibited the duplicate copies (Fig 1A).

Some inconsistencies were perceived comparing the gene annotations from the searched databases with our results. For instance, *Sox9a* and *Sox9b* were reversed annotated in *Takifugu rubripes* (fugu) and *Clarias gariepinus* (african sharptooth catfish); moreover, these genes were assigned as *Sox9a2* and *Sox9a1*, respectively, in *Monopterus albus* (rice field eel). Other inconsistencies were observed and are shown in Fig 1A. The retrieved sequence from *Cichla monoculus* (peacock bass) was ortholog to *Sox9b*. The degenerate primers used may have annealed only to *Sox9b*, precluding the detection of *Sox9a*.

To overcome some undesired incongruences from phylogenetic analysis, possibly related to the high evolutionary rate of substitutions in sequences, an approach based on the combination of phylogenetic results with genomic array of genes can be helpful to corroborate evolutionary relationships. We already used this approach in a study of the *Foxl2* gene [39]. Therefore, an additional support for the *Sox9* evolutionary relationships was determined using the synteny information of specific genetic markers combined with a hierarchical clustering (Fig 1B) (S1 Table). We observed *Sstr2* as the common genetic marker of the *Sox9* single and duplicate copies in all the vertebrate groups analyzed, and *Rasd1* and *Adap1* as the specific markers of *Sox9a* and *Sox9b* in almost all the teleost groups analyzed. The specific markers of *Sox9a* were primarily *Cops3* and *Pemt* but other makers were also observed, such as *Lmf1*, *Vstm4b*, *Lrrc18b* and *Mapk8b*. The main markers of *Sox9b* were *Abca3*, *Tmc6a*, *Notum1a*, *Myadml2*, *Mafg*, *Pcyt2* and *Cant1a*. An exception was zebrafish *Sox9b*, in which none of those markers were found.

Intra-molecular coevolution within and among the Sox9 domains

The inference of intra-molecular coevolutions indicated possible relations among the DIM, K2 and PQS domains in all the Sox9 homologs analyzed (Fig 2). Sox9a showed a large number of intra and inter-domain coevolutions, especially in K2 and PQS, with a large group of mutual correlations. Although less representative, Sox9b showed a similar profile. Even though Sox9 showed a divergent profile, it also evidenced relations among DIM, K2 and PQS but with a lower degree compared to Sox9a. Besides, Sox9 showed a higher number of coevolving residues that are located outside the annotated domains. Finally, the highly conserved HMG box domain exhibited the lowest number of coevolving amino acids.

The results of Ka/Ks calculation showed that all domains are under purifying selection; K2 had the highest Ka/Ks ratio followed by PQS and DIM, whereas HMG box had the lowest value (S2 Table).

Discussion

Paralogy of *Sox9a* and *Sox9b*: phylogenetic and syntenic approaches

The divergence in gene annotation and the lack of protein tertiary structure of Sox9 difficult the determination of correlations between evolutionary events and their functional implications. Nakamura et al. [30] already settled, based on genomic synteny analyses, the nomenclature of *Sox9* paralogs as *Sox9a* and *Sox9b*. However, these analyses were restricted to a small

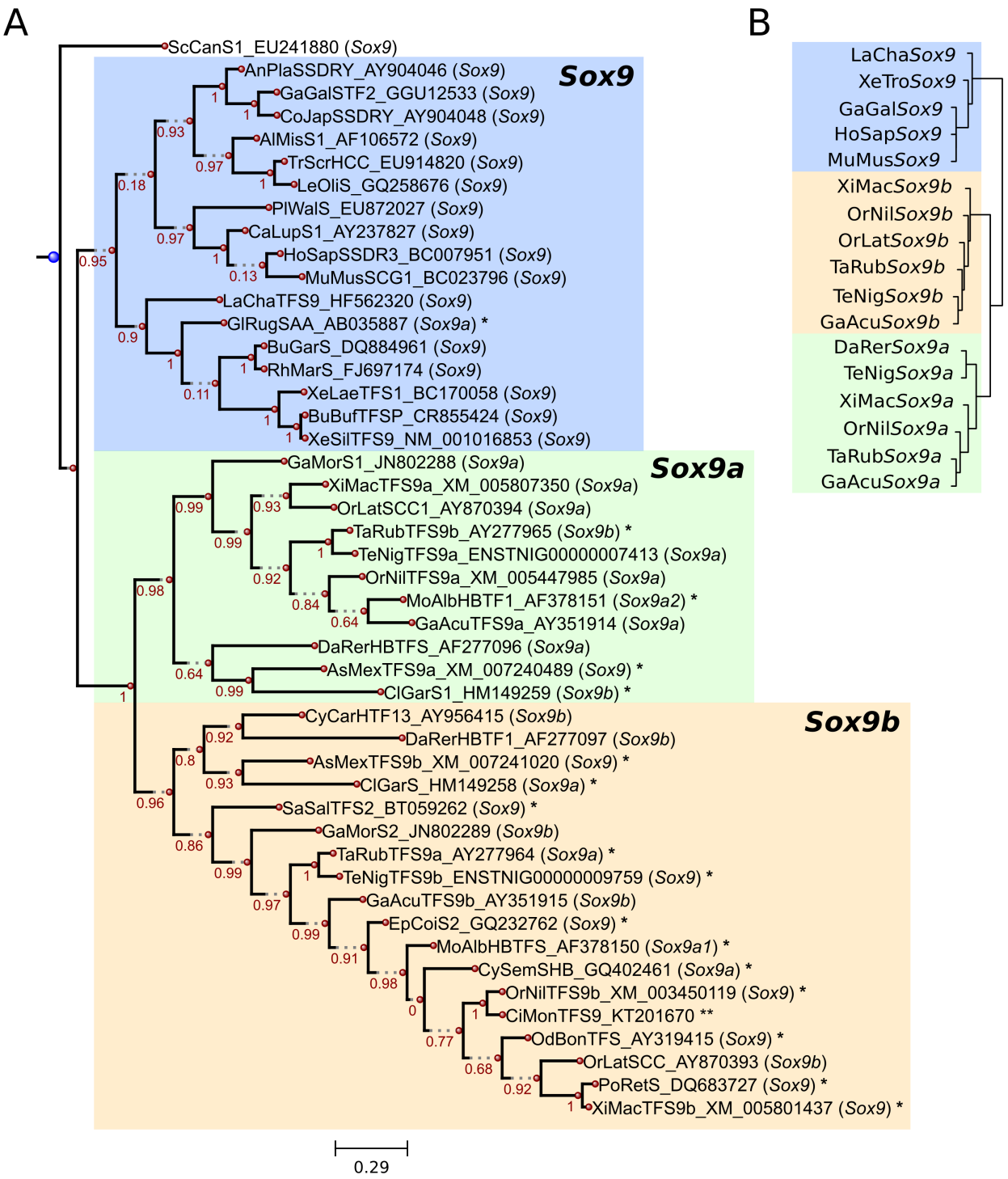

Fig 1. Evolutionary relationships among the *Sox9* vertebrate sequences. (A) Phylogeny of the Sox9 multiple sequence alignment using the maximum likelihood method from the PhyML program. The aLRT (SH-like) values of branch support are shown. The corresponding clades for *Sox9* (blue), *Sox9a* (green) and *Sox9b* (orange) are indicated. The database entries (accession numbers from GenBank or Ensembl) are shown with their corresponding annotations in the database in parenthesis. The single asterisks (*) indicates the divergences between our results and the gene annotation in the corresponding database entry, whereas the double asterisks (**) indicates the sequence obtained from the genome walking technique. The scale bar below the phylogenetic tree indicates the average number of nucleotides substitutions per site. (B) Dendogram of the synteny analysis, based on the closest genetic markers: the clades corresponding to *Sox9* (blue), *Sox9a* (green) and *Sox9b* (orange) were obtained from the hierarchical manhattan clustering method, implemented in the R software.

doi:10.1371/journal.pone.0156199.g001

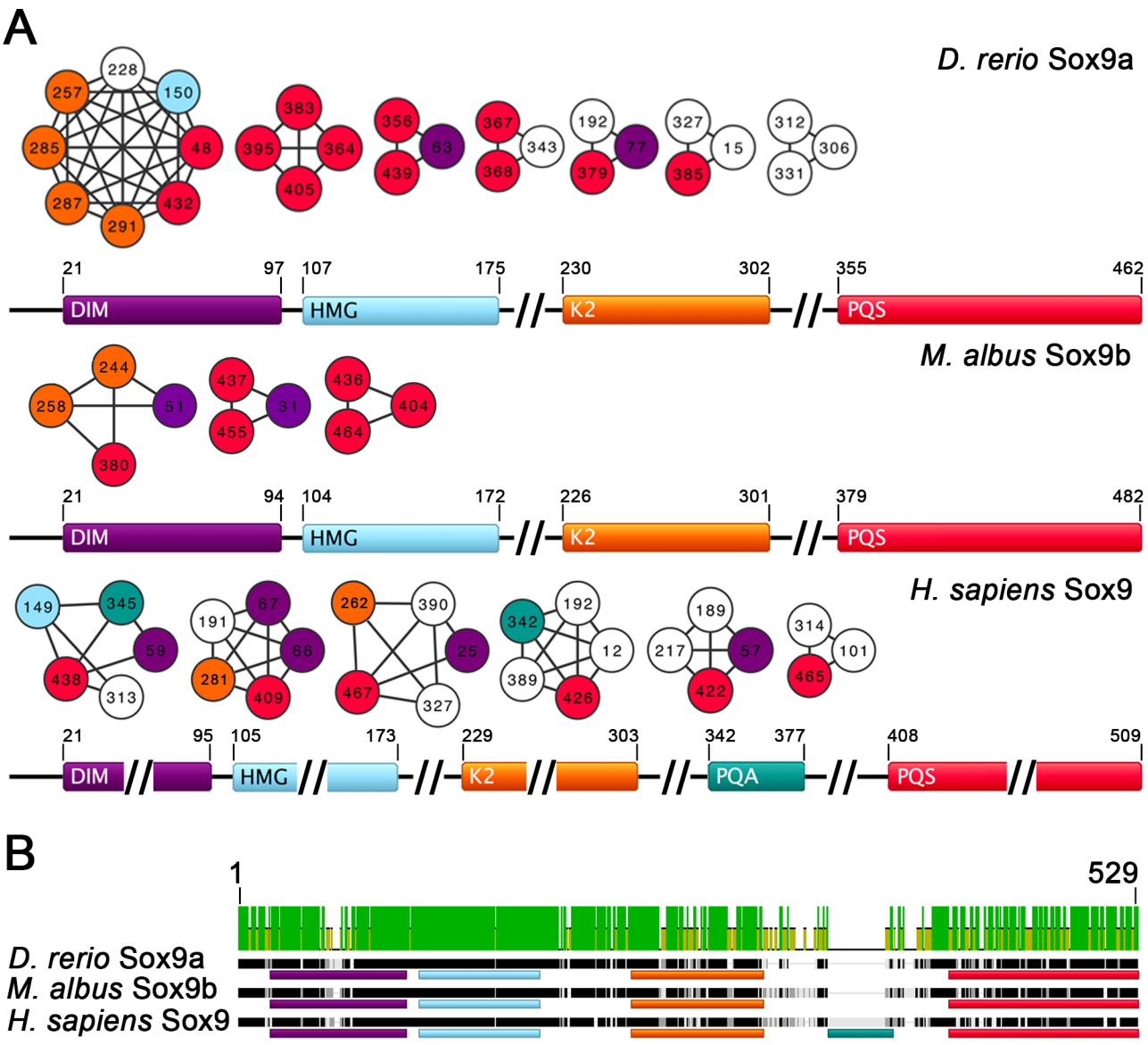

Fig 2. Network of coevolutions in the Sox9 proteins. (A) The coevolutions are shown using *D. rerio* Sox9a, *M. albus* Sox9b and *H. sapiens* Sox9 as the reference sequences. The graph nodes indicate the amino acid with its corresponding position number in the reference sequence. The edges (lines) connect the pair of coevolving amino acids inferred from the CAPS program. Each amino acid is colored based on its localization in the corresponding Sox9 domain: DIM (purple), HMG box (light blue), K2 (orange), PQA (dark cyan) and PQS (red). The range (begin-end) of each domain is also depicted in the scheme. The sign for truncated regions (//) is used for fitting the figure dimensions. (B) The Sox9 multiple protein sequence alignment is indicated. The color of each domain follows the description aforementioned. Upper green bars evidence high conserved sites.

doi:10.1371/journal.pone.0156199.g002

number of sequences and no phylogeny was inferred. In contrast, we gathered a representative data set of *Sox9* sequences, based on a search over different vertebrate genomes. The imbalance in the number of sequences for *Sox9a* and *Sox9b* was due to the data availability at GenBank because only one of the paralog forms was found in the database for some species. Despite of this imbalance, it is evident, by the current data from the searched genomes, that the duplicate copies of Sox9 (referred to as Sox9a and Sox9b) are unique to the teleost lineage. Therefore, they were probably originated from the specific whole genome duplication occurred in this group [40, 41], in agreement with the suggestion of previous studies [2, 30, 42]. Moreover, we

used the synteny information from *Sox9* to generate a dendogram using a hierarchical-clustering algorithm, and the results also supported the paralogy of *Sox9a* and *Sox9b*, reinforcing the conclusions aforementioned.

DIM and TADs as important regions for functional plasticity

Intra-coevolutionary inferences represent an important approach for understanding the evolution and function of proteins because structural and functional relationships or physical interactions between amino acids can be the main causes of their coevolution [43]. Mutations at functionally related sites can change the selective constraints, so coevolution inference is an important complement for molecular selection analysis [43]. Furthermore, it has been observed that physically distant coevolving sites can be essential to maintain the structural and functional stability of the protein [44].

Our coevolution results show possible relationships among all the Sox9 domains from the single and duplicate copies but especially among DIM, K2 and PQS. Primarily, we observed a large group of coevolutions in Sox9a. Studies with single and frameshift mutations/deletions in TADs showed that, for maximum transactivation, K2 and PQS are critical for the proper transcriptional activation/repression of target genes [29, 34]. Therefore, this result is in agreement with our coevolution data that showed pairs of coevolving residues between K2 and PQS. Furthermore, we showed that K2 and PQS are under purifying selection and have the most relaxed selective pressure (primarily K2), reflecting a higher fixation of mutations. In general, most of the protein-protein interactions involved to orchestrate the assembly of the transcriptional machinery are mediated by TAD [45]; therefore, our results of coevolution and selective pressure suggest that this domain can be a good candidate for transcriptional regulation.

The transactivation process, however, depends on the oligomeric state of Sox9. It has been observed that Sox9 can bind to response elements of target genes as a homodimer or monomer, associating with different protein partners to activate or repress the transcriptional machinery [31, 32, 46, 47]. It has been observed that mutations within DIM abrogated the dimerization and ceased or reduced the transactivation activity played by Sox9 for target chondrogenic and sex-determining genes, whereas other functional features of the protein remained unaltered [32]. The most dramatic effect was observed in a 66–75 deletion within DIM, affecting the promoter activation of Amh, a sex-determing gene that only requires a Sox9 monomer for its activation, indicating that DIM is not limited to the dimerization function. We observed in the human Sox9 the residues 66 and 67 from DIM coevolving with the residues 281 and 409 from K2 and PQS, respectively. Altogether, these experimental data and our coevolutionary results indicate that DIM, apart from its dimerization function, may influence directly the transactivation activity.

Sock et al. [32] considered unlikely the direct interference between DIM and TAD (primarily PQS and PQA) because they are physically distant, considering the protein sequence. However, as we mentioned earlier, distant sites in a protein can be functionally or structurally associated; therefore, the transcriptional activation/repression of Sox9 target genes could be strictly related to the TAD affinity, influenced by DIM, to its binding partners. We suppose that these domains are physically close in the tertiary structure, or have some structural feature that a change in one domain may affect the other. This scenario agrees with our coevolutionary and selective pressure analyses, corroborating the relationship between DIM and TAD as a good strategy for the transcriptional regulation of pathways involving Sox9.

The HMG box domain, in contrast, had the lowest rate of mutational fixations and the lowest number of coevolving residues. The high conservation of the HMG box in the Sox9 sequences may be a result of its binding pattern [48], generating a significant constraint in the protein conformation. An exception was observed in Sox9a with a large group of amino acids from K2 and

PQS coevolving with one amino acid from the HMG box. However, no experimental analysis indicates a direct functional or structural involvement between HMG box and TADs. Indeed, mutational analyses within the HMG box have shown that the transactivation activity is only abrogated as a consequence of a reduced affinity or inability of Sox9 to bind DNA [34, 49, 50].

Finally, *Sox9a* and *Sox9b*, after the duplication event in teleosts, seem to have undergone a lineage-specific subfunctionalization, indicated by the similar tissue expression patterns in zebrafish *Sox9a* and medaka *Sox9b* [2, 30]. Although it requires additional data from other teleost species, we suggest that the functional association between DIM and TADs differentially regulate the target genes or even promote novel targets during the transcriptional activation mediated by the Sox9 paralogs. Therefore, this association makes DIM and TADs more flexible to promote a fine gene expression regulation along the subfunctionalization process.

Conclusion

This study correlates evolutionary analyses with functional implications for the Sox9 proteins. First, we reinforce that the event of genome duplication specific of teleosts is the evolutionary process that triggered the evolution of *Sox9a* and *Sox9b*. We also suggest that DIM and TADs are candidates for functional modulation and variability of the Sox9 single and duplicate copies (Sox9a and Sox9b) in vertebrates, allowing the use of well-coordinated regulation strategies that can operate within and among these domains. Furthermore, our study correlates evolutionary analyses with functional implications of Sox9 and highlights that DIM and TAD can be additional players involved in the regulation and diversification of the Sox9 activity. Finally, this strategy could be also used by other transcription factors, such as SoxE proteins.

Materials and Methods

Sequence acquisition and alignment procedures

HMG box domain sequences. To obtain a significant number of sequences and ensure the use of Sox9 homologs, a retrieval methodology based on the HMG box domain was used. This approach ensures a more accurate identification of homologous sequences to the gene of interest and facilitates the choice of outgroup.

Protein sequences of the HMG box family were collected from a large number of metazoans in the Pfam database [51] (accession: PF00505), totaling 4,921 sequences relative to 657 species (S3 Table). Each sequence, identified by an UniProt (http://www.uniprot.org) accession number, was converted to GenBank accession numbers utilizing the gbreader software (Razente HL, Braz ASK, Scott LPB (unpublished)).

To avoid redundancy and ensure the use of representative data, a clustering methodology was performed using the CD-HIT software [52]. The clustering cut-off applied was 99% identity to the HMG box domain (high cut-off because the HMG box domain is highly conserved in metazoan); afterwards, only one representative protein from each cluster was selected for the phylogenetic analysis. This previous procedure resulted in a data set of 779 sequences that were aligned based on a Hidden Markov Model (HMM) profile of the HMG box domain (final alignment length = 69 sites), using the HMMER v.3 software [53].

Sox9 sequences. In addition to the sequences obtained from the UniProt database, Sox9 sequences were obtained based on TBLASTN searches over the nucleotide collection from GenBank (http://www.ncbi.nlm.nih.gov), and the vertebrate genomes from the Ensembl database (http://www.ensembl.org/) [54].

Additionally, an experimental procedure was employed to retrieve the *Sox9* sequence from the neotropical cichlid peacock bass *Cichla monoculus* collected from Balbina hydroelectric station lake, Presidente Figueiredo-AM, Brazil (3°09'57"S, 59°54'44"W), according to the

Brazilian laws for environmental protection (wild collection permit, ICMBio 22984–1 e 32556–2). Voucher specimens (10816-female; 10850-male) were included in the fish collection of INPA-National Institute for Amazonian Researchs, Manaus/AM. The experimental procedure on the teleost was conducted according to the international guidelines of Sao Paulo State University and approved by the Institutional Animal Care and Use Committee (IACUC) (Protocol no. 34/08—CEEA/IBB/UNESP), and the tissue samples are available at Integrative Genomics Laboratory of Sao Paulo State University under the registration numbers 10816 and 10850. The animals were euthanized through immersion in a water bath with benzocaine 250 mg/liter during 10 minutes. The genomic DNA was extracted from muscle and liver tissues using the phenol-chloroform method [55]. *Sox9* was amplified by PCR (polymerase chain reaction) using degenerate primers (F—5' CARGTNYTNAARGGNTAYGA 3' and R—5' CCANAR YTTNCCNARNGTYTT 3') constructed with the Primer3Plus software [56]. For the construction of these primers, Sox9 protein sequences from different vertebrate species were retrieved based on a BLASTP search at the Expasy Proteomic Server (http://ca.expasy.org/), using the Sox9a sequence from *Oreochromis aureus* (EU373500) as query. These sequences were aligned using ClustalW [57], and the primers were constructed for the most conserved regions. The amplicons were 195 base pair (bp), and the full gene sequence (3,327 bp) was obtained by the genome walking technique, using the Genome Walker™ kit (Clontech) according to the manufacturer's protocol. All amplicons were cloned in p-GEM-T plasmid vectors (Promega), and the corresponding clones were sequenced with an ABI Prism 3100 DNA sequencer (Perkin-Elmer), using ABI Prism Big Dye Terminator Cycle Sequencing Ready Reaction kits (Perkin-Elmer). Finally, a prediction of introns, exons and amino acid sequences were performed using the online program Softberry FGENESH (http://www.softberry.com/).

Phylogenetic analyses: HMG box, SoxE and Sox9. Three phylogenetic analyses were conducted separately: (i) HMG box, (ii) SoxE and (iii) *Sox9*. Based on the phylogeny of the HMG box domain, the corresponding branch for the subgroup E was determined (S1 Fig). Within this subgroup, the clustered amino acid sequences for *Sox8*, *Sox9* and *Sox10* were retrieved and aligned using the Mafft algorithm [58]. A maximum likelihood phylogeny was generated to allow the precise identification of the Sox9 homologs and guide the appropriate choice of outgroup. Afterwards, the Sox9 sequences were aligned based on their codons using the Mafft algorithm allocated on the GUIDANCE web server [59]. Finally, the reference data set for the *Sox9* phylogeny was composed of 47 nucleotide sequences, and the alignment length included 2,049 sites (S1 Text). The phylogenetic analysis was performed using only complete and putative *Sox9* sequences.

The choice of the best-fit model of evolution was performed with ProtTest3 [60] and Jmodeltest [61] for the amino acid and nucleotide sequences, respectively.

The phylogenetic reconstruction of *Sox9* was determined by the maximum likelihood method, implemented in the Phyml v3.0.1 [62] software, using the aLRT (SH-like) reliability test [63] (other parameter details, see S4 Table). The visualization and the final tree edition were performed using FigTree v1.3.1 (http://tree.bio.ed.ac.uk/software/figtree/) and ETE2 [64].

Sox9 synteny analysis

To provide an additional support to the homology of the *Sox9* sequences and understand the composition/organization of the genes in the proximity of *Sox9*, the syntenic regions were analyzed in species which genomic information is available at the Genomicus browser [65]. A matrix was built based on the presence and absence of each genetic marker considered in the proximity of the *Sox9* single and duplicate copies. Latter, a hierarchical clustering was performed based on the this matrix information (S1 Table), using the manhattan method available in the R program [66].

Intra-coevolution inference and Ka/Ks ratio calculation

The inference of intra-molecular coevolutions was done using the CAPS software [67]. CAPS measures the correlated evolutionary variation of amino acids to identify the coevolving site pairs. The performance and sensitivity of the method has been examined in different proteins [43] and, according to the authors, more accurate results can be achieved with long ($\geq$ 20 sequences) and populated sequence alignments. The major advantage of CAPS over other methods of coevolutionary inference is the separation of phylogenetic linkage from structural and functional coevolution [43].

Each branch corresponding to *Sox9*, *Sox9a* and *Sox9b* was extracted and submitted, along with its aligned protein sequences, to the CAPS analysis. The complete alignment and subalignments, combined with the removal of specific phylogenetic clades, were generated following the CAPS automated protocol to exclude phylogenetic coevolutions [43]. Finally, a total of 1,000 alignment simulations were generated to remove the stochastic effects of the coevolutions inferred for the *Sox9* single and duplicate copies.

The Ka/Ks ratio calculations were conducted using the sequences of human Sox9 [GenBank:BC007951], zebrafish Sox9a [GenBank:NM_131643] and rice field eel Sox9b [GenBank: AF378150]. The calculations were based on the pairwise alignment (performed with Muscle [68]) of the HMG box, DIM, K2 and PQS domains, using the command line KaKs-Calculator v1.2 [69] with model selection parameter (S2 Table).

Supporting Information

S1 Fig. HMG box and SoxE phylogeny. Maximum-likelihood phylogenetic tree of the HMG box domain (A) and SoxE subgroup (B).
(PNG)

S1 Table. Sox9 genetic markers. The important markers of *Sox9*, *Sox9a* and *Sox9b* are highlighted in blue, green and orange, respectively. The presence (1) or absence (0) of a specific marker is depicted.
(XLS)

S2 Table. Ka/Ks calculation. Values of Ka/Ks for the Sox9 domains, including the p-value and model of evolution.
(PDF)

S3 Table. Sequence retrieval. Amino acid sequences obtained from Uniprot database based on the HMG box domain.
(XLS)

S4 Table. Maximum-likelihood reconstruction tree parameters. Data type, substitution model, tree searching operations, starting tree method and branch support are indicated.
(PDF)

S1 Text. Sox9 multiple codon sequence alignment. Codon based alignment of all sequences used for *Sox9* phylogenetic inference.
(PDF)

Acknowledgments

This study was supported by São Paulo Research Foundation (FAPESP) [grant numbers 2009/05234-4; 2010/13143-6] and National Counsel of Technological and Scientific Development (CNPq) [grant number 475147/2010-3]. We thank Dr C Schneider for providing fish tissue samples.

Author Contributions

Conceived and designed the experiments: MTG GTV CM. Performed the experiments: MTG RTN. Analyzed the data: MTG RTN GTV CM. Contributed reagents/materials/analysis tools: CM. Wrote the paper: MTG RTN GTV CM.

References

1. Wegner M. From head to toes: the multiple facets of Sox proteins. Nucleic Acids Research. 1999; 27 (6):1409–20. doi: 10.1093/nar/27.6.1409 PMID: ISI:000079256800001.
2. Kluver N, Kondo M, Herpin A, Mitani H, Schartl M. Divergent expression patterns of Sox9 duplicates in teleosts indicate a lineage specific subfunctionalization. Development genes and evolution. 2005; 215 (6):297–305. Epub 2005/04/09. doi: 10.1007/s00427-005-0477-x PMID: 15818483.
3. Bewley CA, Gronenborn AM, Clore GM. Minor groove-binding architectural proteins: Structure, function, and DNA recognition. Annual Review of Biophysics and Biomolecular Structure. 1998; 27:105–31. doi: 10.1146/annurev.biophys.27.1.105 PMID: ISI:000074324000006.
4. Goodwin GH, Sanders C, Johns EW. A new group of chromatin-associated proteins with a high content of acidic and basic amino acids. European journal of biochemistry / FEBS. 1973; 38(1):14–9. Epub 1973/09/21. PMID: 4774120.
5. Bustin M. Revised nomenclature for high mobility group (HMG) chromosomal proteins. Trends Biochem Sci. 2001; 26(3):152–3. doi: 10.1016/S0968-0004(00)01777-1 PMID: ISI:000168719800007.
6. Bianchi ME, Agresti A. HMG proteins: dynamic players in gene regulation and differentiation. Curr Opin Genet Dev. 2005; 15(5):496–506. doi: 10.1016/j.gde.2005.08.007 PMID: ISI:000232454200005.
7. Bianchi ME, Beltrame M. Upwardly mobile proteins Workshop: The role of HMG proteins in chromatin structure, gene expression and neoplasia. Embo Rep. 2000; 1(2):109–14. doi: 10.1093/embo-reports/kvd030 PMID: ISI:000209560600001.
8. Lefebvre V, Dumitriu B, Penzo-Mendez A, Han Y, Pallavi B. Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. Int J Biochem Cell B. 2007; 39 (12):2195–214. doi: 10.1016/j.biocel.2007.05.019 PMID: ISI:000250899300006.
9. Berta P, Hawkins JR, Sinclair AH, Taylor A, Griffiths BL, Goodfellow PN, et al. Genetic evidence equating SRY and the testis-determining factor. Nature. 1990; 348(6300):448–50. Epub 1990/11/29. doi: 10.1038/348448A0 PMID: 2247149.
10. Bowles J, Schepers G, Koopman P. Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. Developmental biology. 2000; 227(2):239–55. Epub 2000/11/10. doi: 10.1006/dbio.2000.9883 PMID: 11071752.
11. Bi WM, Deng JM, Zhang ZP, Behringer RR, de Crombrugghe B. Sox9 is required for cartilage formation. Nature genetics. 1999; 22(1):85–9. PMID: ISI:000080096300029.
12. Akiyama H, Chaboissier MC, Martin JF, Schedl A, de Crombrugghe B. The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. J Bone Miner Res. 2002; 17:S142–S. PMID: ISI:000177952800072.
13. Foster JW, Dominguez-Steglich MA, Guioli S, Kwok C, Weller PA, Stevanovic M, et al. Campomelic dysplasia and autosomal sex reversal caused by mutations in an SRY-related gene. Nature. 1994; 372 (6506):525–9. PMID: 7990924
14. Bishop CE, Whitworth DJ, Qin YJ, Agoulnik AI, Agoulnik IU, Harrison WR, et al. A transgenic insertion upstream of Sox9 is associated with dominant XX sex reversal in the mouse. Nature genetics. 2000; 26 (4):490–4. doi: 10.1038/82652 PMID: ISI:000165671700027.
15. Kobayashi A, Chang H, Chaboissier MC, Schedl A, Behringer RR. Sox9 in testis determination. Ann Ny Acad Sci. 2005; 1061:9–17. doi: 10.1196/annals.1336.003 PMID: ISI:000236118500002.
16. Stolt CC, Wegner M. SoxE function in vertebrate nervous system development. The international journal of biochemistry & cell biology. 2010; 42(3):437–40.
17. Wegner M, Stolt CC. From stem cells to neurons and glia: a Soxist's view of neural development. Trends Neurosci. 2005; 28(11):583–8. doi: 10.1016/j.tins.2005.08.008 PMID: ISI:000233213700003.
18. Cheung M, Chaboissier MC, Mynett A, Hirst E, Schedl A, Briscoe J. The transcriptional control of trunk neural crest induction, survival, and delamination. Dev Cell. 2005; 8(2):179–92. doi: 10.1016/j.devcel.2004.12.010 PMID: ISI:000226875500009.
19. Taylor KM, LaBonne C. SoxE factors function equivalently during neural crest and inner ear development and their activity is regulated by SUMOylation. Dev Cell. 2005; 9(5):593–603. doi: 10.1016/j.devcel.2005.09.016 PMID: ISI:000233186700006.

20. Zhu M- Y, Gasperowicz M, Chow RL. The expression of NOTCH2, HES1 and SOX9 during mouse retinal development. Gene Expression Patterns. 2013; 13(3):78–83.

21. Seymour PA, Freude KK, Tran MN, Mayes EE, Jensen J, Kist R, et al. SOX9 is required for maintenance of the pancreatic progenitor cell pool. Proceedings of the National Academy of Sciences. 2007; 104(6):1865–70.

22. Kawaguchi Y. Sox9 and programming of liver and pancreatic progenitors. The Journal of clinical investigation. 2013; 123(123 (5)):1881–6.

23. Mori–Akiyama Y, Van den Born M, Van Es JH, Hamilton SR, Adams HP, Zhang J, et al. SOX9 is required for the differentiation of paneth cells in the intestinal epithelium. Gastroenterology. 2007; 133 (2):539–46. PMID: 17681175

24. Wagner T, Wirth J, Meyer J, Zabel B, Held M, Zimmer J, et al. Autosomal sex reversal and campomelic dysplasia are caused by mutations in and around the SRY-related gene SOX9. Cell. 1994; 79(6):1111–20. PMID: 8001137

25. Hanley KP, Oakley F, Sugden S, Wilson DI, Mann DA, Hanley NA. Ectopic SOX9 mediates extracellular matrix deposition characteristic of organ fibrosis. Journal of Biological Chemistry. 2008; 283 (20):14063–71. doi: 10.1074/jbc.M707390200 PMID: 18296708

26. Bennett MR, Czech KA, Arend LJ, Witte DP, Devarajan P, Potter SS. Laser capture microdissection-microarray analysis of focal segmental glomerulosclerosis glomeruli. Nephron Experimental Nephrology. 2007; 107(1):e30–e40. PMID: 17684420

27. Drivdahl R, Haugk KH, Sprenger CC, Nelson PS, Tennant MK, Plymate SR. Suppression of growth and tumorigenicity in the prostate tumor cell line M12 by overexpression of the transcription factor SOX9. Oncogene. 2004; 23(26):4584–93. PMID: 15077158

28. Wang H, Leav I, Ibaragi S, Wegner M, Hu G-f, Lu ML, et al. SOX9 is expressed in human fetal prostate epithelium and enhances prostate cancer invasion. Cancer research. 2008; 68(6):1625–30. doi: 10.1158/0008-5472.CAN-07-5915 PMID: 18339840

29. Chiang EFL, Pai CI, Wyatt M, Yan YL, Postlethwait J, Chung BC. Two sox9 genes on duplicated zebrafish chromosomes: Expression of similar transcription activators in distinct sites. Developmental biology. 2001; 231(1):149–63. doi: 10.1006/dbio.2000.0129 PMID: ISI:000167271000012.

30. Nakamura S, Watakabe I, Nishimura T, Toyoda A, Taniguchi Y, Tanaka M. Analysis of medaka sox9 orthologue reveals a conserved role in germ cell maintenance. PLoS One. 2012; 7(1):e29982. Epub 2012/01/19. doi: 10.1371/journal.pone.0029982 PMID: 22253846; PubMed Central PMCID: PMC3257256.

31. Bernard P, Tang PY, Dewing P, Harley VR, Vilain E. Dimerization of SOX9 is required for chondrogenesis, but not for sex determination. Hum Mol Genet. 2003; 12(14):1755–65. doi: 10.1093/Hmg/Ddg182 PMID: ISI:000183953700011.

32. Sock E, Pagon RA, Keymolen K, Lissens W, Wegner M, Scherer G. Loss of DNA-dependent dimerization of the transcription factor SOX9 as a cause for campomelic dysplasia. Hum Mol Genet. 2003; 12 (12):1439–47. Epub 2003/06/05. PMID: 12783851.

33. Sudbeck P, Schmitz ML, Baeuerle PA, Scherer G. Sex reversal by loss of the C-terminal transactivation domain of human SOX9. Nature genetics. 1996; 13(2):230–2. Epub 1996/06/01. doi: 10.1038/ng0696-230 PMID: 8640233.

34. McDowall S, Argentaro A, Ranganathan S, Weller P, Mertin S, Mansour S, et al. Functional and structural studies of wild type SOX9 and mutations causing campomelic dysplasia. Journal of Biological Chemistry. 1999; 274(34):24023–30. Epub 1999/08/14. PMID: 10446171.

35. Harley VR. The molecular action of testis-determining factors SRY and SOX9. Novartis Foundation symposium. 2002; 244:57–66; discussion -7, 79–85, 253–7. Epub 2002/05/07. PMID: 11990798.

36. Jo A, Denduluri S, Zhang B, Wang Z, Yin L, Yan Z, et al. The versatile functions of Sox9 in development, stem cells, and human diseases. Genes Dis. 2014; 1(2):149–61. Epub 2015/02/17. doi: 10.1016/j.gendis.2014.09.004 PMID: 25685828; PubMed Central PMCID: PMC4326072.

37. Coustry F, Oh CD, Hattori T, Maity SN, de Crombrugghe B, Yasuda H. The dimerization domain of SOX9 is required for transcription activation of a chondrocyte-specific chromatin DNA template. Nucleic Acids Research. 2010; 38(18):6018–28. doi: 10.1093/Nar/Gkq417 PMID: ISI:000283116600013.

38. Barrionuevo F, Scherer G. SOX E genes: SOX9 and SOX8 in mammalian testis development. International Journal of Biochemistry and Cell Biology. 2010; 42(3):433–6. Epub 2009/08/04. doi: 10.1016/j.biocel.2009.07.015 PMID: 19647095.

39. Geraldo MT, Valente GT, Braz AS, Martins C. The discovery of Foxl2 paralogs in chondrichthyan, coelacanth and tetrapod genomes reveals an ancient duplication in vertebrates. Heredity (Edinb). 2013; 111(1):57–65. Epub 2013/04/04. doi: 10.1038/hdy.2013.19 PMID: 23549337; PubMed Central PMCID: PMC3692314.

40. Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, et al. Zebrafish hox clusters and vertebrate genome evolution. Science. 1998; 282(5394):1711–4. Epub 1998/11/30. PMID: 9831563.

41. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature. 2004; 431 (7011):946–57. Epub 2004/10/22. doi: 10.1038/nature03025 PMID: 15496914.

42. Cresko WA, Yan YL, Baltrus DA, Amores A, Singer A, Rodriguez-Mari A, et al. Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. Developmental Dynamics. 2003; 228(3):480–9. Epub 2003/10/28. doi: 10.1002/dvdy.10424 PMID: 14579386.

43. Fares MA, Travers SA. A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. Genetics. 2006; 173(1):9–23. Epub 2006/03/21. doi: 10.1534/genetics.105.053249 PMID: 16547113; PubMed Central PMCID: PMC1461439.

44. Gloor GB, Martin LC, Wahl LM, Dunn SD. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. Biochemistry. 2005; 44(19):7156–65. Epub 2005/05/11. doi: 10.1021/bi050293e PMID: 15882054.

45. Piskacek S, Gregor M, Nemethova M, Grabner M, Kovarik P, Piskacek M. Nine-amino-acid transactivation domain: establishment and prediction utilities. Genomics. 2007; 89(6):756–68. Epub 2007/05/01. doi: 10.1016/j.ygeno.2007.02.003 PMID: 17467953.

46. Leung VY, Gao B, Leung KK, Melhado IG, Wynn SL, Au TY, et al. SOX9 governs differentiation stage-specific gene expression in growth plate chondrocytes via direct concomitant transactivation and repression. PLoS Genet. 2011; 7(11):e1002356. Epub 2011/11/11. doi: 10.1371/journal.pgen.1002356 PMID: 22072985; PubMed Central PMCID: PMC3207907.

47. Ikeda T, Kamekura S, Mabuchi A, Kou I, Seki S, Takato T, et al. The combination of SOX5, SOX6, and SOX9 (the SOX trio) provides signals sufficient for induction of permanent cartilage. Arthritis and rheumatism. 2004; 50(11):3561–73. Epub 2004/11/06. doi: 10.1002/art.20611 PMID: 15529345.

48. Mertin S, McDowall SG, Harley VR. The DNA-binding specificity of SOX9 and other SOX proteins. Nucleic Acids Research. 1999; 27(5):1359–64. doi: 10.1093/nar/27.5.1359 PMID: ISI:000078990300018.

49. Meyer J, Sudbeck P, Held M, Wagner T, Schmitz ML, Bricarelli FD, et al. Mutational analysis of the SOX9 gene in campomelic dysplasia and autosomal sex reversal: lack of genotype/phenotype correlations. Hum Mol Genet. 1997; 6(1):91–8. Epub 1997/01/01. PMID: 9002675.

50. Kwok C, Weller PA, Guioli S, Foster JW, Mansour S, Zuffardi O, et al. Mutations in SOX9, the gene responsible for Campomelic dysplasia and autosomal sex reversal. Am J Hum Genet. 1995; 57 (5):1028–36. Epub 1995/11/01. PMID: 7485151; PubMed Central PMCID: PMC1801368.

51. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, et al. The Pfam protein families database. Nucleic Acids Res. 2012; 40(Database issue):D290–301. doi: 10.1093/nar/gkr1065 PMID: 22127870; PubMed Central PMCID: PMC3245129.

52. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. Epub 2006/05/30. doi: 10.1093/bioinformatics/btl158 PMID: 16731699.

53. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011; 39(Web Server issue):W29–37. doi: 10.1093/nar/gkr367 PMID: 21593126; PubMed Central PMCID: PMC3125773.

54. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, et al. Ensembl 2013. Nucleic Acids Res. 2013; 41(Database issue):D48-55. doi: 10.1093/nar/gks1236 PMID: 23203987; PubMed Central PMCID: PMC3531136.

55. Sambrook J, Russel DW. Molecular Cloning: A Laboratory Manual. 3rd ed. New York: Cold Spring Harbor Laboratory Press; 2001.

56. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JAM. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Research. 2007; 35:W71–W4. doi: 10.1093/Nar/Gkm306 PMID: ISI:000255311500016.

57. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994; 22(22):4673–80. Epub 1994/11/11. PMID: 7984417; PubMed Central PMCID: PMC308517.

58. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular biology and evolution. 2013; 30(4):772–80. doi: 10.1093/molbev/mst010 PMID: 23329690; PubMed Central PMCID: PMC3603318.

59. Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. GUIDANCE: a web server for assessing alignment confidence scores. Nucleic Acids Res. 2010; 38(Web Server issue):W23–8. Epub 2010/05/26. doi: 10.1093/nar/gkq443 PMID: 20497997; PubMed Central PMCID: PMC2896199.

60. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011; 27(8):1164–5. Epub 2011/02/22. doi: 10.1093/bioinformatics/btr088 PMID: 21335321.

61. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012; 9(8):772. Epub 2012/08/01. doi: 10.1038/nmeth.2109 PMID: 22847109.

62. Guindon S, Gascuel O. Phy ML: A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic biology. 2003; 52(5):696–704. Epub 2003/10/08. PMID: 14530136.

63. Anisimova M, Gascuel O. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. Syst Biol. 2006; 55(4):539–52. Epub 2006/06/21. doi: 10.1080/10635150600755453 PMID: 16785212.

64. Huerta-Cepas J, Dopazo J, Gabaldon T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010; 11:24. Epub 2010/01/15. doi: 10.1186/1471-2105-11-24 PMID: 20070885; PubMed Central PMCID: PMC2820433.

65. Louis A, Muffato M, Roest Crollius H. Genomicus: five genome browsers for comparative genomics in eukaryota. Nucleic Acids Res. 2013; 41(Database issue):D700–5. Epub 2012/11/30. doi: 10.1093/nar/gks1156 PMID: 23193262; PubMed Central PMCID: PMC3531091.

66. Ihaka R, Gentleman R. R: a language for data analysis and graphics. Journal of computational and graphical statistics. 1996; 5(3):299–314.

67. Fares MA, McNally D. CAPS: coevolution analysis using protein sequences. Bioinformatics. 2006; 22 (22):2821–2. doi: 10.1093/bioinformatics/btl493 PMID: ISI:000241958000018.

68. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research. 2004; 32(5):1792–7. Epub 2004/03/23. doi: 10.1093/nar/gkh340 PMID: 15034147; PubMed Central PMCID: PMC390337.

69. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics, proteomics & bioinformatics. 2006; 4(4):259–63. Epub 2007/05/29. doi: 10.1016/S1672-0229(07)60007-2 PMID: 17531802.