Journal Homepage: www.cellmolbiol.org

Cellular and Molecular Biology

Original Article

Genome-wide identification: molecular characterization and evolutionary aspects of Sox genes in *Nile tilapia*

Muhammad Farhan Khan[1,2#], Mehwish Sultana[3#], Shakeela Parveen[3], Wardah Hassan[4], Muhammad Tayyab[5], Mashahour Fawwaz Alenazi[6], Alanazi Khalid Zabena[7], Youhou Xu[1], Zibin Hong[8], Peng Zhu[1*], Laiba Shafique[1*]

[1] Guangxi Key Laboratory of Beibu Gulf Marine Biodiversity Conservation, Beibu Gulf University, Qinzhou 535011, China
[2] Department of Chemistry, Gomal University, Dera Ismail Khan 29050, Pakistan
[3] Department of Zoology, Government Sadiq College Women University, Bahawalpur 63100, Pakistan
[4] Department of Zoology, Emerson University, Multan
[5] Department of Zoology, University of Agriculture Faisalabad, Faisalabad, Pakistan
[6] College of Medicine, Shaqra University, P.O. box 13343 Riyadh 7396, Saudi Arabia
[7] King Khaled General Hospital, Hafer Al-Batin, Saudi Arabia
[8] Department of Reproductive Medicine, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, 106 Zhongshan 2nd Road, Guangzhou, 510080, China

Article Info

Article history:

Received: October 10, 2024
Accepted: January 19, 2025
Published: February 28, 2025

Use your device to scan and read the article online

Abstract

Nile tilapia has become one the most significant species in global aquaculture due to its exceptional adaptability, rapid growth and high reproductive capacity. Role of Sox genes in reproduction and development made attention to further investigate the role of these genes. Based on *N. tilapia* importance in aquaculture industry and role of Sox genes in the development of tissues and organs during embryogenesis, this study systematically analyzed Sox genes functionality in *N. tilapia* by using computational tools. In our study, phylogenetic analysis revealed that *N. tilapia* is most closely related to blue *tilapia* compared to other species. Sox genes are conserved in nature and share both acidic and basic properties as well as thermostable and hydrophobic in nature. The subcellular localization in *N. tilapia* indicated that majority of the Sox proteins are expressed in the Nucleus and Cytoplasm. Enrichment analysis explains the Sox genes' role in cell differentiation, and biosynthesis process and acts as a molecular functional regulator. Significant differences in transcription factor binding sites were observed, highlighting the potential role of these regulatory regions in the regulation of Sox genes in *N. tilapia*. First time it is reported that Sox genes in *N. tilapia* have four major recombinant breakpoints that revealed phylogenetic segregation across several recombination fragments. In this primer, we aim to provide the reader with a comprehensive overview of Sox gene family in *N. tilapia* and to provide the functional properties of Sox genes for better follow-up in upcoming experiments for futuristic research.

Keywords: Nile tilapia, Sox genes, Thermostable, Hydrophobic, Duplications.

1. Introduction

Nile tilapia has become a corner stone in global aquaculture, primarily due to its high reproductive capacity and adaptability in diverse farming environments [1]. This species is known for its ability to thrive in various environmental conditions, ranging from freshwater to brackish water, which has allowed it to be cultivated in diverse geographical regions around the world. The economic impact of tilapia farming is particularly evident in developing countries where it provides a reliable source of affordable protein to address food security challenges and has become widely cultivated due to its adaptability in various farming conditions [2]. The increasing demand for tilapia has made it a keystone species in the aquaculture industry with production continuously rising to meet global needs [3].

Additionally, the scalability of tilapia farming from small backyard ponds to large commercial operations, makes it adaptable to different socio-economic contexts which further enhances its global importance [4]. On the environmental front *N. tilapia*'s resilience to various environmental conditions, including suboptimal water quality, positions it as a sustainable option in aquaculture. Its ability to thrive in environments with low dissolved oxygen levels and its resistance to common aquatic diseases reduce the need for extensive chemical inputs, thereby lowering the cost and encouraging tilapia farming [5].

Sox gene family comprises 32 identified genes that

[*] Corresponding author.
E-mail address: yijianrudi@163.com (P.Zhu); laibazoologist@gmail.com (L. Shafique).
[#] These authors contributed equally

Doi: http://dx.doi.org/10.14715/cmb/2025.71.2.8

play a pivotal role in various developmental and reproductive processes in fishes. Sox genes are transcriptional factors that regulate critical aspects of embryonic development, organogenesis and cell differentiation. Significance of Sox genes is further underscored by their involvement in neurogenesis and early embryogenesis as observed in various aquaculture species where these genes have undergone expansion due to whole genome duplication events.

Sox gene family is of paramount importance in fishes, particularly in *N. tilapia* where these genes are crucial for various developmental processes [6], including sex determination, neurogenesis and cell differentiation [7]. Sox 1a, Sox 1b, Sox 2, Sox3, Sox8a and Sox 8b are integral to neural development with functions ranging from maintaining stem cell pluripotency to influencing neural and reproductive system development [8]. Sox 9a, Sox 9b and also Sox30, Sox32 are particularly critical in sex determination with different expression patterns contributing to gonadal differentiation [9]. Other genes such as Sox 4b, Sox5, Sox 6a and Sox 6b regulate cell fate and contribute to the formation of tissues like cartilage and muscles [10]. Sox 7, Sox 17 and Sox 18 are involved in vascular and endoderm formation which are essential for heart, blood vessels and digestive system collectively, the identified Sox genes in *N. tilapia* are indispensable for species growth, development and reproductive functions with each gene contributing uniquely to the biological processes that ensure the species survival and adaptability [11]. In contrast, teleosts exhibit a much larger Sox gene family with 29 members reported in *Oncorhynchus mykiss* [12], 27 genes in *Oreochromis niloticus* [7], and 26 genes in *Collichthys lucidus* [13] and *Danio rerio* [14]. On the other hand, the Sox gene family appears to be more conserved in invertebrates, with 7 members in *Patinopecten yessoensis* [15] and 8 members in *Drosophila melanogaster* [16]. However, despite these advances, a critical gap remains in our knowledge of Sox genes in *N. tilapia*.

The current study examines the evolutionary dynamics and molecular characterization of Sox gene family in *N. tilapia*. Understanding these genes' roles and mechanisms will provide valuable insights into the species development, growth and reproductive functions. By leveraging the advanced genomic and bioinformatics tools, this study seeks to map the intricate genetic networks governed by Sox genes. These genes offer novel insights into their role in critical biological processes such as sex determination, neurogenesis and organogenesis. Findings of this study are expected to have significant implications for improving *N. tilapia* breeding, health and reproduction. Furthermore, this study will contribute to the broader understanding of gene regulation and evolutionary adaption and potential applications in *N. tilapia* and other aquatic species. Future research will explore the environmental factors that influence the Sox genes expression and their potential implications in sustainable aquaculture.

2. Materials and Methods

2.1. Identification of Sox genes in *N. tilapia*

The sequences of *N. tilapia* that belong to the Sox gene family were obtained by using the databases of NCBI (https://www.ncbi.nlm.nih.gov/), with indication of accession numbers presented in (Supplementary Table S1). To evaluate potential Sox genes across different aquatic species, we utilized the conserved HMG box domain protein, a DNA binding protein, by employing the Hidden Markov Model (HMM) profile obtained from the Pfam database. To obtain predicted protein-coding variations, we employed local BlASTP software with E-value set at 105 [1].

2.2. Phylogenetic analysis and multiple sequence alignment analysis

In order to acquire Sox genes sequences for the species which are *Oreochromis niloticus*, *Oreochromis aureus*, *Pelmatolapia mariae*, *Danio rerio*, *Ctenopharyngdon Idella* and *Labeo rohita*, the NCBI database was used. Sox genes sequences for the subsequent species we employed ClustalW for aligning gene sequences for each represented species by utilizing the MEGA software version (v.11). Subsequently, we constructed a neighbor-joining (NJ) phylogenetic tree using MEGA11, setting the Bootstrap value to 1000 replicates. We then employed ITOL to visualize and present the tree using a single aligned file [17]. Using Multiple Align Show (https://www.bioinformatics.org/sms/multi_align.html), we aligned the Sox genes of *N. tilapia* to detect the sequence alterations or insertions and deletions or indels [18].

2.3. Physicochemical properties and structural analysis

By using the ProtParam program, we physiochemically characterized *N. tilapia* Sox proteins and indulged the information on the following parameters of physiochemical properties as molecular weight (MW), isoelectric point (pI), aliphatic index (AI), number of amino acids (A.A.) and instability index (II). Moreover, conserved protein motifs in *N. tilapia*'s Sox proteins were analyzed using the MEME suite, which can identify up to 10 MEME motifs. Verification of the conserved domains was performed through the NCBI CDD database (https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). Employing the web servers Cello-life (http://cello.life.nctu.edu.tw/) and Wolf Psort (https://wolfpsort.hgc.jp/), the subcellular localization of *N. tilapia*'s Sox proteins was examined [19].

2.4. Gene duplication analysis and chromosomal location of Sox gene family

We analyzed the whole-genome dataset to determine the chromosomal lengths and positions of Sox genes in *N. tilapia*. The gene locations on the chromosomes were precisely mapped using the MCScanX tool, based on the genome annotation file serving as reference [20].

2.5. 3D-Structure analysis

Amino acid sequences of all Sox proteins in *N. tilapia* were yielded in Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2), in order to construct three-dimensional structure of each Sox protein and analyze their secondary structure features. Fold recognition and homology modeling were also calculated [21].

2.6. Analysis using Scan Prosite and functional enrichment analysis

The online tool Scan Prosite was utilized to compute structural and functional variations within specific protein domains [22]. We utilized the Prosite motif library to search for motifs by uploading a protein sequence via the ScanProsite web application (https://prosite.expasy.org/scanprosite) [23]. We used the Shiny Go online tool for making gene function graphs in enrichment analysis [24].

2.7 Identification and retrieval of Transcription Factor Binding sites (TFBSs)

We uploaded the genomic data TFBIND online tool (https://tfbind.hgc.jp/), which analyzed the 100 bp upstream region of the highly predicted gene locations to identify potential transcription factor binding sites, utilizing the TRANSFAC R.3.4 weight matrix [25].

2.8 Analysis of recombination breakpoints in the Sox gene family

We employed Genetic Algorithm Recombination Detection (GARD) for identification of the recombination breakpoints in multiple sequence-aligned in *N. tilapia*'s Sox genes [26], we used the following approach to identify the segment-specific phylogenies. When the maximum number of breakpoints B is specified to infer the method searches for B or less in the sequence alignment. For each segment, we employed a maximum likelihood model as this approach is appropriate for derivation of phylogenies for possibility of each non-recombinant segment and [27], estimation of the appropriate by utilization of the related data metrics, such as Akaike Information criteria (AIC) [28].

2.9 Statistical analysis

For descriptive analysis, bioinformatic tools have been applied to the Sox genes dataset.

3. Results

3.1. Phylogenetic analysis and multiple sequence analysis

Phylogenetic analysis revealed the link between closely related species and shared common ancestors. We determined the evolutionary history of the Sox gene family through molecular phylogenetic analysis by employing the neighbor-joining (NJ) method, with bootstrap consensus values calculated for each node. Based on the homologous gene sequences, the amino acid sequences from the following species (*Oreochromis niloticus*, *Oreochromis aureus*, *Pelmatolapia mariae*, *Danio rerio*, *Ctenopharyngdon Idella* and *Labeo rohita*) were examined. Based on the molecular phylogenetic analysis, these sequences were subsequently categorized into three clades: Clade-A, Clade-B, and Clade-C as demonstrated in Figure 1. Phylogenetic analysis findings show that more sequence similarities were observed among the *Oreochromis niloticus*, *Oreochromis aureus* and *Pelmatolapia mariae*. Multiple sequences of *Oreochromis niloticus and Danio rerio* alignments showed similarities, differences, and indels based on the presented data (Supplementary Figure 1).

3.2. Physiochemical attributes of Sox genes in *N. tilapia*

Using the NCBI database, data of 22 Sox genes were acquired from the genome of *N. tilapia.* Physicochemical properties of these genes were investigated, such as molecular weight (MW), instability index (II), isoelectric point (pI), and grand average of hydropathicity (GRAVY) and aliphatic index (AI) as demonstrated in (Table 1). All *N. tilapia* Sox proteins have lengths ranging from 238 to 853. Molecular weight of *N. tilapia* Sox proteins which are ranged from 26577.37 D to 93458.27 D. The isoelectric point (pI) ranged from 5.47 to 9.78. Except for Sox1a, Sox1b, Sox2, Sox3, Sox5, Sox13, Sox14, Sox18 and Sox19. All Sox genes were acidic in nature. In the aliphatic index values, the Sox13 peptides were observed thermostable at high temperatures in *N. tilapia* aliphatic index (AI) value greater than 65, demonstrating their thermostability at elevated temperatures. This indicates that all Sox proteins have lower GRAVY values and are hydrophobic (Table 1).

3.3 Structural characterization of Sox genes in *N. tilapia*

The gene structure and conserved motifs in *N. tilapia*'s Sox gene family were examined for deeper understanding (Figure 2). For the Sox gene family in *N. tilapia*, 10 conserved motifs were estimated (Figure 2-A to 2-D), where MEME-1 and MEME-4 were identified as Sox domain using Pfam due to their large number of amino acids (50) as shown in (Figure 2-B & Table 2). Results have also been verified by comparison with the NCBI CDD database (Figure 2-C). Furthermore, the Sox pro-peptide family domain was identified in all homologs of the Sox family in addition to the Sox domain. Furthermore, the analysis of gene structure showed that there are notable differences in the structure of downstream and upstream non-translated regions (UTRs), exons and introns (Figure 2-D).

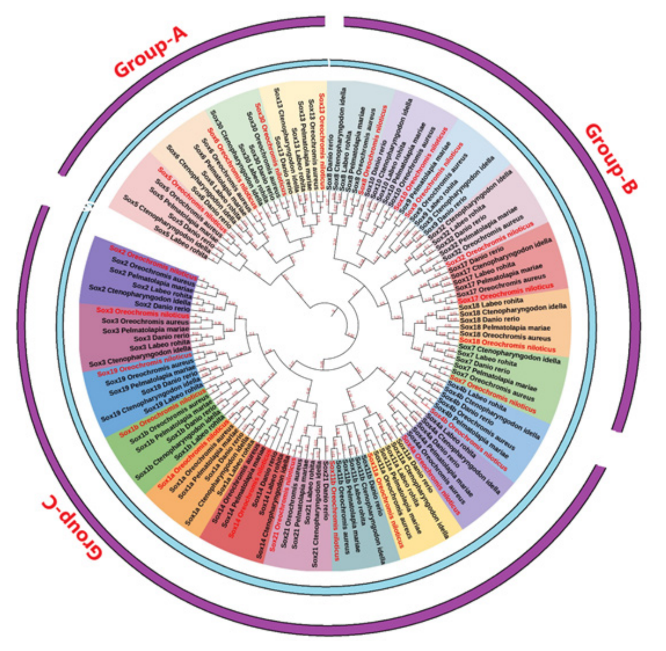

Fig. 1. Phylogenetic analysis of Sox gene family.

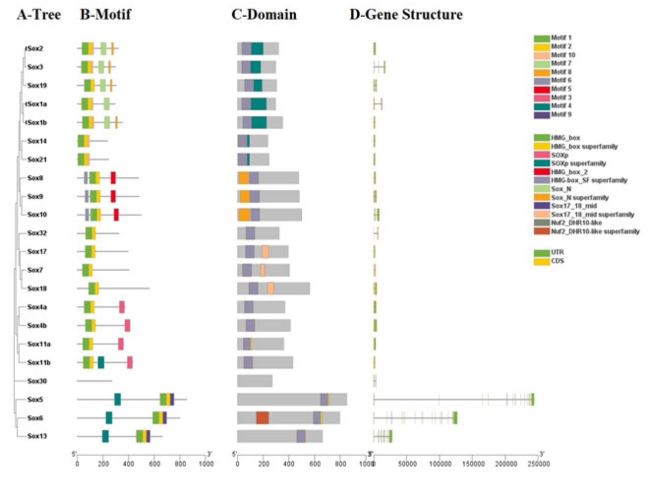

Fig. 2. Structural characterization of Sox genes in *N. tilapia.*

Table 1. Physiochemical properties of Sox genes in *N. tilapia.*

Variants	Chr ID	MW(DA)	AA	PI	AI	II	Gravy
Sox1a	LG16	32408.19	299	9.78	47.42	53.19	-0.824
Sox1b	LG23	37679.34	354	9.70	54.92	61.01	-0.665
Sox2	LG17	35132.80	322	9.74	45.16	58.56	-0.818
Sox3	LG2	33491.00	300	9.63	51.80	69.45	-0.807
Sox4a	LG22	40390.48	371	6.33	47.68	71.31	-0.862
Sox4b	LG11	44244.36	414	6.54	50.24	68.18	-0.756
Sox5	LG17	93458.27	853	8.00	59.55	70.93	-0.875
Sox6	LG1	88767.97	799	6.48	64.64	62.12	-0.777
Sox7	LG15	44127.72	407	6.13	56.14	56.00	-0.716
Sox8	LG8	52505.95	479	6.52	49.54	68.89	-0.932
Sox9	LG4	53669.87	484	6.20	50.79	71.40	-1.001
Sox10	LG4	53481.49	503	6.45	50.78	60.72	-0.782
Sox11a	LG19	40384.91	363	5.47	56.47	63.39	-0.796
Sox11b	LG15	48303.17	433	5.47	50.32	74.73	-1.004
Sox13	LG5	746301.18	664	8.94	67.71	73.14	-0.900
Sox14	LG23	26577.37	238	9.68	58.32	58.67	-0.668
Sox17	LG9	44068.91	397	6.23	59.40	50.91	-0.840
Sox18	LG20	33660.08	307	9.61	46.78	58.11	-0.839
Sox19	LG3	27127.21	248	9.74	56.05	58.93	-0.533
Sox21	LG16	30129.42	273	5.68	60.66	56.00	-0.679
Sox30	LG10	30129.42	273	5.68	60.66	56.00	-0.679
Sox32	LG9	36881.09	327	6.99	52.60	73.02	-0.854

Table 2. 10 significantly conserved motifs within the Sox gene family of *N. tilapia.*

MEME Motif	Sequence of amino acid	Width	Pfam Domain
1	IKRPMNAFMVWSKDZRRKLAQZNPDMHNAEJSKRLGKRWKLLSESEKRPF	50	HMG
2	IEEAERLRAQHMKDYPDYKYRPRRKKKTL	29	-
3	SFEEGSLGSHFEFPDYCTPELSEMIAGDWLEATFSDLVFTY	41	-
4	KKLAASQMEKQRQQMELARQQQEQIARQQQLLQQQHKINLLQQQIQQVQ	50	-
5	NIDFGNVDIGELSTDVIANIDPFDVBEFDQYLPPNSH	37	-
6	EDERFPVCIRDAVSQVLKGYDWTLVPM	27	-
7	HHHNPHNPQPMHRYDMSALQYSPISNSQSYMNASPTGYGGI	41	-
8	APSGDLRDMISMYLP	15	-
9	VDGKKLRIGEYKAMMRSRRQEMRQYFSVGQ	30	-
10	LKKDKYSLPGGLL	13	-

3.4. Gene duplication, chromosomal distribution and cellular distribution

Analyzing the precise position of each gene on the chromosomes aids and clarifies the relationships between Sox gene pairs, which in turn influence the size, positioning and orientations of associated genomic elements. The data, depicted in Figures 3A and 3B, show the locations and duplication events related to Sox genes**.** For the purpose of obtaining the evolutionary history of *N. tilapia*'s Sox gene family, duplication events were examined. Tandem duplication of the Sox9-Sox10 gene pair was observed on chromosome LG4, while seven gene pairs were identified as segmental duplications. Additionally, the subcellular localization of Sox proteins in *N. tilapia* indicated that majority of these proteins are expressed in both Nucleus and Cytoplasm as shown in (Figure 4).

3.5. Scan PROSITE and enrichment analysis

We employed PROSITE for the identification of functional and structural residues associated with PROSITE and ProRule signatures in the selected Sox proteins. This method has the ability to detect intra-domain features, such as binding sites, active sites and disulfide bridges. The accuracy of functional predictions was enhanced by combining motif recognition specificity with profile sensitivity. HMG_B and Sox C were identified in *N. tilapia* as shown in Figure 5, which demonstrates a graphical display of Sox protein hits and feature predictions based on domain analysis from the Scan PROSITE database. Gene enrichment explains the functional properties of selected genes. Sox genes have high mobility group and domain. These genes also have transcription regulator complex and have ability to bind DNA. Sox genes also play significant role in cell differentiation, and biosynthesis process and act as a molecular functional regulator as represented in Figure 6.

3.6. 3-D structure of Sox genes in *N. tilapia*

Figure 7 demonstrates the prediction of 3-D structure model prediction and secondary structure as presented in

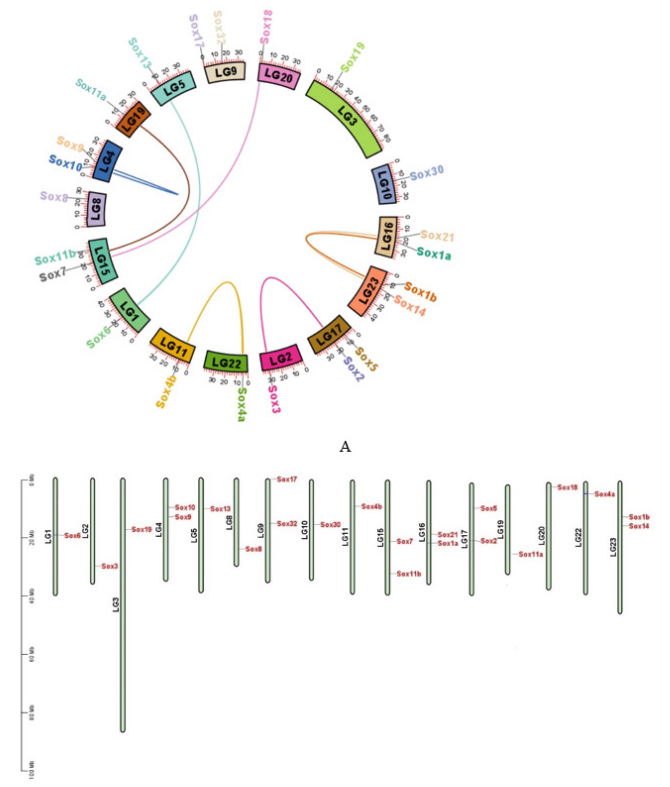

Fig. 3. Sox Gene Duplication Events and Chromosomal Segregation in *N. tilapia*. A: Sox gene duplication events B: Sox genes segregate on different chromosomes at specific locations.

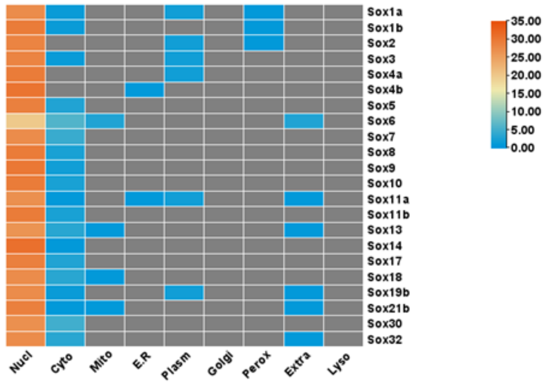

Fig. 4. Subcellular localization of Sox proteins in *N. tilapia*.

Table 3, which was obtained for the identification of *N. tilapia* proteins. Sox13 and Sox14 which are closely related in *N. tilapia*, exhibited comparable quantities of secondary structural components such as β-helices and β-sheets and disorder rate. Sox genes share an alpha helix range between (7-25) and the beta sheet highest values also vary in Sox gene family. Most of the genes shared a residues value of 79. The highest disorder value was found in the Sox 9 and Sox 10 genes.

3.7. Transcription factor binding sites

Transcription factor binding sites are activators or repressors that activate or inhibit cellular activity on the basis of gene nature. Transcription factor binding to specific genomic locations is fundamental to transcriptional regulation in cells. In this study, we examined the binding sites for five transcription factors TATA, OCT1, GATA, YY1, and STAT within Sox gene family of *N. tilapia*. The distribution pattern of these transcription factor binding sites (TFBSs) in *N. tilapia* as follows: GATA > YY1 > OCT1 > STAT > TATA. In contrast, the pattern observed in *Danio rerio* was GATA > OCT1 > YY1 > STAT > TATA. Overall, there were more TFBSs in *Danio rerio* as compared to *N.*

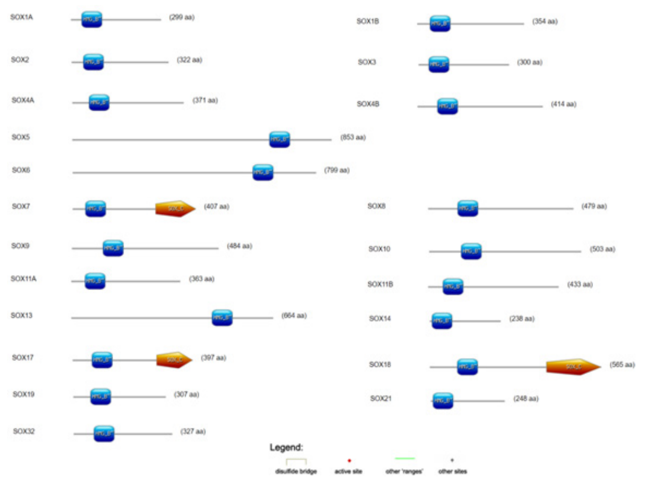

Fig. 5. Scan PROSITE analysis of Sox genes in *N. tilapia*.

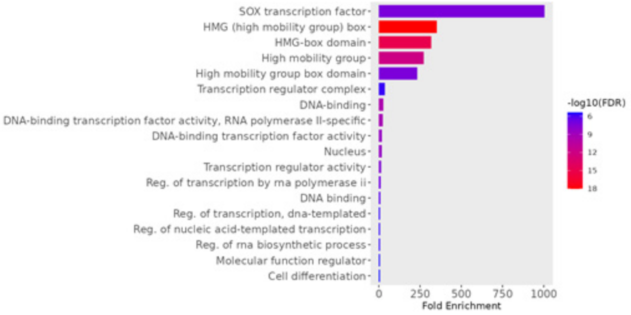

Fig. 6. Enrichment analysis.

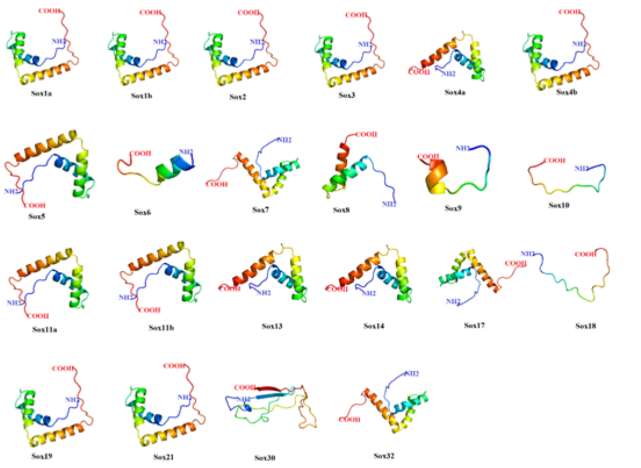

Fig. 7. Secondary Structure of Sox genes in *N. tilapia*.

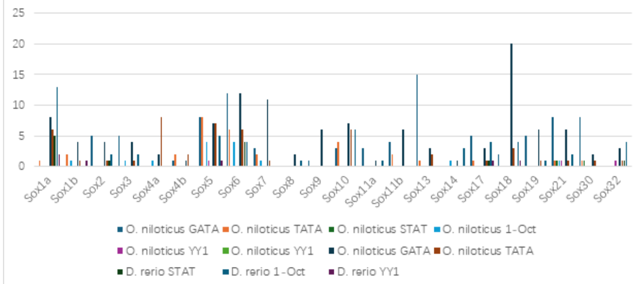

Fig. 8. Transcription factor binding sites (GATA, TATA, STAT, OCT1 and YY1).

Table 3. Secondary structure of Sox proteins in *N. tilapia*.

Sr.no	Name of Protein	Alpha Helix	Beta Strand	Disorder	Residues
1	Sox1a	22	0	79	79
2	Sox1b	19	0	84	79
3	Sox2	17	0	82	79
4	Sox3	16	0	79	79
5	Sox4a	13	0	75	76
6	Sox4b	11	0	79	76
7	Sox5	13	0	75	76
8	Sox6	26	0	81	16
9	Sox7	15	0	80	81
10	Sox8	12	0	86	46
11	Sox9	13	0	87	13
12	Sox10	10	1	87	18
13	Sox11a	13	0	77	76
14	Sox11b	14	0	76	76
15	Sox13	23	1	75	70
16	Sox14	23	1	75	70
17	Sox17	12	1	78	81
18	Sox18	11	1	82	25
19	Sox19b	17	0	81	79
20	Sox21b	25	0	72	79
21	Sox30	7	27	47	75
22	Sox32	20	1	81	80

tilapia (Figure 8).

3.8. Recombination analysis

We conducted GARD to identify recombination breakpoints for detection of fragmented sequences, and subsequent analysis that revealed phylogenetic segregation across several recombination fragment trees: Tree1 (1-618), Tree 2 (619-733), Tree 3 (734-957), Tree 4(958-1704) and Tree 5 (1705-2689) (Supplementary Figure 2). GARD evaluated 7176 models to identify evidence of recombination breakpoints, locating 2176 putative breakpoints with up to 4 inferred breakpoints per model. Notably, the genetic algorithm examined just 0.00% of these breakpoints (Supplementary Figure 3). By calculating the likelihood of identifying a breakpoint at a specific site across all alignment points using the standardized Akaike weights of the model, we determined the model-averaged support for the breakpoint sites. This analysis was in agreement with the best-fitting model. Using a genetic technique for multiple breakpoints investigations the multiple sequence alignment of eight Sox gene nucleotide sequences showed the five main recombination breakpoints and several minor breakpoints across different locations (Figure 9).

4. Discussion

Technological advancements in next-generation sequencing and other high-efficiency genome sequencing tools have greatly improved our ability to analyze genetic diversity, including single nucleotide polymorphisms (SNPs) and their functional impact on specific phenotype traits. This capability allows for a deeper understanding of animal genetics at the molecular level [29]. Candidate gene studies make it easier to examine the genetic resources in aquatic species, allowing for the identification of functional genes and their relationships to characteristics like productivity, adaptability and disease resistance [30]. Comparative genomics aids in identifying new genes and understanding the processes that regulate these genes which are crucial for investigating commercially significant physiological features in aquaculture species. This study offers valuable and novel insights for aquaculture industry and future research endeavors.

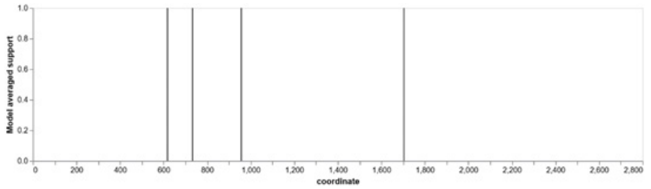

Fig. 9. Recombination Analysis of Sox genes in *N. tilapia*.

4.1. Phylogenetic analysis

Our phylogenetic analysis indicates that the Sox gene family of *N. tilapia* is closely related to blue tilapia and shows greater sequence similarity with other species. New molecular phylogenies provide support for natural groups that were unanticipated by previous studies. Likewise, studies on the fish Tree of Life have been reported in 2013 based on a comprehensive molecular phylogeny [31]. Phylogenetic analysis revealed that valuable insights into genetic diversity, functional biology and evolutionary history closely related species in the clade, similar studies on DMRT gene in grass carp have depicted the closely related species relationship in the same clade [32].

4.2. Physiochemical attributes

Evaluating the physicochemical properties of proteins associated from various gene families is essential for understanding their functions and characteristics. In the species studied, Sox genes shared both acidic and basic properties as indicated by their isoelectric point (pI). Previous studies have discussed the structural characteristics of globular proteins and their thermostability using the aliphatic index (AI) to predict protein stability at different temperatures [32,33]. The Aliphatic index (AI) measures the relative volume of aliphatic side chains (valine, leucine, alanine and isoleucine) within the selected protein. A higher aliphatic index is associated with increased thermostability of globular protein. Therefore, proteins that remain stable at higher levels of temperatures are considered to have a higher aliphatic index (AI) [1,32]. Since the aliphatic index (AI) of Sox peptides from *N. tilapia*, it was concluded that these peptides were thermostable. However, differences in isoelectric point values were observed, leading to instability in some proteins. Of the five proteins analyzed, a few exhibited clear instability. The GRAVY value provides the prediction about interaction of water and protein and is calculated by taking "total protein's hydropathy values" divided by "protein's length". Through GRAVY value, it is easy to determine that proteins have hydrophobic or hydrophilic characteristics. If GRAVY value of selected protein is negative it means the protein is hydrophilic in nature and if GRAVY value is negative, then nature of protein is considered to be hydrophobic [34]. All Sox proteins exhibited lower GRAVY values indicating their hydrophobic nature.

4.3. Structural analysis

Protein sequence data can uncover significant evolutionarily conserved regions critical for various biological processes. Multiple sequence alignment is a fundamental technique for identifying conserved regions and also obtaining the essential information that is necessary for functional and structural analysis of the proteins [35]. In this study, the MEME tool was used to identify 10 motifs in *N. tilapia*'s Sox gene family, to analyze the protein sequence properties of Sox genes, while Sox genes have HMG domain in motif 1. This analysis provides insights into regular expression patterns of these conserved motifs. The structural and formational organization is influenced by these common motifs. Our study indicates that the Sox domains of *N. tilapia* contain the three and four MEME motifs, respectively. Same approach has been implemented in other studies to analyze biological roles and conserved features of motifs in different species genes [35].

4.4. Gene duplication and chromosomal distribution

To acquire genetic variations or novel genes, we used various gene duplication methods which are retro position, crossing over and chromosomal duplication or genome. In order to achieve best findings in evolution functional processes we used genetic variation and enhanced functions as these mechanisms have important role in evolution. Recognition of gene duplication dynamics and trajectories is important as reported in previous studies about both features of evolutionary forces which are genome-wide and localized. The above-mentioned procedures provide insights into interaction and connection which regulate genetic diversity and adaptability during process of evolution either intra-specific or inter-specific [36]. It is difficult to determine gene duplications' pace but utilization of emergence of redundant genetic variations through mutations and selection pressure has been observed through functional impacts. Consequently, these variables affect the evolution of selected genes, the retention and differentiation of duplicated genes and their ability to develop new functions. In a similar number of generations, the transmission of duplicated genes is facilitated by comparing them to functional copies which mitigates unwanted changes. This increased mutation rate can result in a gene acquiring new functions or becoming more complex in biological processes compared to their original counterparts, potentially resulting in the evolution of new roles or new adaptions. Previous studies in ice fish have demonstrated that defects in a duplicated digestive gene led to the development of an antifreeze gene, while the duplication can also result in creation of entirely different genes, such as those responsible for snake venom production [37]. To unravel the genetic evolution of *N. tilapia* gene family, this investigation examined duplication events within this gene family. Seven segmental duplications and one tandem duplication were predicted in Sox gene family. Likewise, studies also reported in TGF gene family in *N. tilapia* [1]. Furthermore, it seems that there is a correlation between the retention rate and the divergence of two paralogs. On the 10 WG-duplicated *Sox* genes in teleostean genomes, 2 are duplicated in all species analyzed: *Sox4* and *Sox9*, of which paralogs a and b possess closely related and low dN/dS ratios. This observation can be explained if we consider a rapid sub-functionalization event just after the WGD. If so, the maintenance of the 2 paralogs is then absolutely necessary to ensure the ancestral function of the gene and the survival of the organism. On the contrary, genes presenting paralogs with a more divergent evolution, at the molecular level, are more often detected as singleton [38].

4.5. Analysis of scan PROSITE and enrichment analysis

PROSITE analysis was conducted to identify specific residues crucial for activities and interactions of Sox proteins, including those involved in disulfide bridges, active sites, binding sites and structural functions correlation. In comparison to previously mentioned predictions, this analysis refines the findings by enhancing profile sensitivity and specificity in determining motifs utilizing ProRule and PROSITE signature. Similar methods have been employed in other studies, such as *Bufo bufo* was used to identify the key characteristics of proteins and their implications for the species [1].

4.6. Protein structural configuration

From multiply aligned homologous sequence, a protein secondary structure prediction method is given which has ability to calculate, an overall per residue three-state accuracy of 70.1%. There are two main objectives, one is to acquire maximum accuracy for the identification of a set of concepts that is significant for prediction by following linear statistics and the second is to understand the folding process. Secondary structure prediction provides insights about sequence edge effects, residue conformational propensities, position of insertions and deletions in aligned homologous sequence, moments of conservation, residue ratio, filtering, secondary structure feedback effects, autocorrelation and moments of hydrophobicity. Precise use of edge effects, moments of conservation, and auto-correlation are new in the present study. Concepts that have been used in prediction have significance and it was determined by step-by-step procedure by adding information and evaluating the weights in discrimination function. Simple structure of the prediction permits the procedure to repeat easily. Prediction of accuracy is easy and predictable [39]. Results can be related to distinct functions that chaperones have to perform and cellular responses that are visible in each species. Further transcriptome-level research is required to validate these findings. Current results are consistent with previous studies [21], [30].

4.7. Transcription factor binding sites

Transcriptional regulatory network of an organism contains various steps which most important step is the identification of every transcription factor and all of its DNA binding sites. Many procedures can be used to find transcription factor binding sites in which consensus sequences and position-specific scoring matrices are important. Additionally, methods by which the average number of nucleotides have been calculated that match between a putative site and all known sites could be utilized. These approaches can be expanded naturally by employing pairwise nucleotide dependencies and preposition content of information [40]. Present study indicated that more TFBSs in *Danio rerio* as compared to *N. tilapia* were calculated. The following studies have reported transcription-binding sites in different species [1,21,41].

4.8. Recombination analysis

In our study, four key potential breakpoints in the nucleotide sequences of Sox genes were identified using recombination analysis conducted through GARD. These breakpoints enable Sox gene family to perform a wide range of tasks in different species. Evolutionary factors control the variations in the following sequences which have impact on the observed functional diversity [26]. It is inferred that, when compared to the model that uses the Akaike Information Criterion (AIC) scores to take into consideration different topologies in the segment arrangement despite assuming a uniform tree for all partitions; a minimum of one, the breakpoints indicate topological incongruence. Evaluating this variance might provide information about the underlying evolutionary processes (or biological processes) that have emerged and could provide insight into particular features of the species tree. Evaluating this issue within a methodological framework is also important as it might provide flaws and ambiguities in phylogenetic inference that influence the observed topological incongruence of species tree and gene tree [26].

Sox genes have multifaced functions in growth, breeding and cell differentiation [28]. In the process of cell diffraction, immune response and gonadal development, Sox genes will be effective tools for enhancing aquaculture production by manipulation or selective breeding strategies. Our study explores comprehensive knowledge about molecular foundation of Sox genes in *N. tilapia*, which become a valuable source for future genomic, phylogenetic and evolutionary studies.

Whole genome data have become increasingly important across a variety of research fields, including evolutionary developmental biology, social ecology and phylogenetics as well as more applied areas often referred to as translational genomics. Various techniques and approaches are employed to comprehend gene evolution from these diverse perspectives. However, comparative genomics remains a relatively new tool in this area of research. By conducting comparative gene study of Sox genes, we explore their characteristics and functions in detail. Sox genes play an important role in sex differentiation and determination. Our study aimed to predict the role of these genes in *N. tilapia*. Phylogenetic analysis revealed that *N. tilapia* is closely related to blue tilapia. Sox genes are conserved, shared acidic, basic properties, thermostable and hydrophobic in nature. One tandem and seven segmental duplications were predicted. Sox genes predominantly expressed their proteins in the Nucleus and Cytoplasm. Enrichment analysis explains the Sox gene's significant role in cell differentiation, and biosynthesis process and act as a molecular functional regulator. Sox genes have four major recombinant breakpoints that revealed phylogenetic segregation across several recombination fragments. Significant differences observed in TFB sites revealed their potential role of regulatory regions of Sox genes in differential transcriptional and translational efficiencies of Sox genes in *N. tilapia*. Hence, additional investigations are mandatory to corroborate these findings and reveal putative mechanism of action behind these effects.

Conflict of interest

None

Ethical approval

Not applicable

Informed consent

Not applicable

Funding

This project is granted by the Guangxi Talent Grant (81QMKYQD12) and the High Talent Research Start-up Program Beibu Gulf University (81-23KYQD02).

Authors contribution

MFK and MS wrote original draft, SP wrote methodology, WH and MT performed analyses, AMFA data curation, AKZA visualization and software, YX and ZH revised manuscript. PZ supervised study, LS conceived idea, wrote and revised original manuscript.

References

1. Khan MF, Parveen S, Sultana M, Zhu P, Xu Y, Safdar A, Shafique L (2024). Evolution and Comparative Genomics of the Transforming Growth Factor-β-Related Proteins in Nile Tilapia. Mol Biotechnol 6:1-5 doi: 10.1007/s12033-024-01263-x.
2. El-Sayed AM, Fitzsimmons K (2023) From Africa to the world—The journey of Nile tilapia. Rev Aquac 15(S1):6-21. doi: 10.1111/raq.12738.
3. Prabu E, Rajagopalsamy C, Ahilan B, Jeevagan I, Renuhadevi M (2019). Tilapia–an excellent candidate species for world aquaculture: a review. Annu Res Rev Biol 31(3):1-14. doi: 10.9734/arrb/2019/v31i330052.
4. El-Sayed AFM (2006). Tilapia culture. CABI pub. doi: 10.1079/9780851990149.00
5. Gu DE, Yu FD, Yang YX, Xu M, Wei H, Luo D, Hu YC (2019). Tilapia fisheries in Guangdong Province, China: Socio-economic benefits, and threats on native ecosystems and economics. Fis Manag Ecol 26(2):97-107. doi: 10.1111/fme.12330.
6. Angelozzi M, Lefebvre V (2019). SOXopathies: growing family of developmental disorders due to SOX mutations. Trends Genet 35(9):658–71. doi: 10.1016/j.tig.2019.06.003.
7. Wei L, Yang C, Tao W, Wang D (2016). Genome-wide identification and transcriptome-based expression profiling of the Sox gene family in the Nile tilapia (Oreochromis niloticus). Int J Mol Sci 17(3):270. doi: 10.3390/ijms17030270
8. Wan H, Han K, Jiang Y, Zou P, Zhang Z, Wang Y (2019). Genome-wide identification and expression profile of the sox gene family during embryo development in large yellow croaker, lari-

michthys crocea. DNA Cell Biol 38(10):1100-11. doi: 10.1089/dna.2018.458

9. Cresko WA, Yan YL, Baltrus DA, Amores A, Singer A, Rodríguez-Marí A, Postlethwait JH (2003). Genome duplication, subfunction partitioning, and lineage divergence: Sox9 in stickleback and zebrafish. Dev Dyn 228(3): 480-489. doi: 10.1002/dvdy.10424. doi: 10.1002/dvdy.10424
10. Cnaani A, Lee BY, Ozouf-Costaz C, Bonillo C, Baroiller JF, D'Cotta H, Kocher T (2007). Mapping of sox2 and sox14 in tilapia (*Oreochromis* spp.). Sex Dev 1(3): 207-210. doi: 10.1159/000102109.
11. Wegner M (1999). From head to toes: the multiple facets of Sox proteins. Nucleic Acids Res 27(6): 1409-1420. doi: 10.1093/nar/27.6.1409.
12. Ma F, Zou Y, Ma R, Chen X, Ma L (2022). Evolution, characterization and expression analysis of Sox gene family in rainbow trout (*Oncorhynchus mykiss*). Czech J Anim Sci 67(4). doi: 10.17221/4/2022-CJAS.
13. Liu F, Zhou L, Zhang J, Wang Y, Wang Z, Liu X, Cai M (2022). Genome-wide identification and transcriptome-based expression profiling of the Sox gene family in the spinyhead croaker (*Collichthys lucidus*). J Fish Biol 100(1): 15-24. doi: 10.1111/jfb.14913.
14. Hu Y, Wang B, Du H (2021). A review on sox genes in fish. Rev Aquacult 13(4): 1986-2003. doi: 10.1111/raq.12554.
15. Yu J, Zhang L, Li Y, Li R, Zhang M, Li W, Xie X, Wang S, Hu X, Bao Z (2017). Genome-wide identification and expression profiling of the SOX gene family in a bivalve mollusc *Patinopecten yessoensis*. Gene 627: 530-537. doi: 10.1016/j.gene.2017.07.013.
16. Crémazy F, Berta P, Girard F (2001). Genome-wide analysis of Sox genes in *Drosophila melanogaster*. Mech Dev 109(2): 371-375. doi: 10.1016/S0925-4773(01)00529-9.
17. Letunic I, Bork P (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res 49(W1): W293-W296. doi: 10.1093/nar/gkab301.
18. Fu Y, Khan MF, Wang Y, Parveen S, Sultana M, Liu Q, Shafique L (2024). In silico analysis: Molecular characterization and evolutionary study of CLCN gene family in buffalo. Genes 15(9): 1163. doi: 10.3390/genes15091163.
19. Riaz Z, Hussain M, Parveen S, Sultana M, Saeed S, Ishaque U, Faiz Z, Tayyab M (2024). In silico analysis: Genome-wide identification, characterization and evolutionary adaptations of bone morphogenetic protein (BMP) gene family in *Homo sapiens*. Mol Biotechnol 66(11): 3336-3356. doi: 10.1007/s12033-023-00944-3.
20. Hassan FU, Deng T, Rehman MS, Rehman ZU, Sarfraz S, Mushahid M, Rehman SU (2024). Genome-wide identification and evolutionary analysis of the FGF gene family in buffalo. J Biomol Struct Dyn 42(19): 10225-1036. doi: 10.1080/07391102.2023.2256861.
21. Sultana M, Tayyab M, Sunil, Parveen S, Hussain M, Saeed S, Riaz Z, Shabbir S (2024). In silico molecular characterization of TGF-β gene family in *Bufo bufo*: genome-wide analysis. J Biomol Struct Dyn 1-5. doi: 10.1080/07391102.2024.2313168.
22. Sigrist CJ, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N (2010). PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38(Suppl 1): D161-D166. doi: 10.1093/nar/gkp885.
23. Ghosh A, Gao L, Thakur A, Siu PM, Lai CW (2017). Role of free fatty acids in endothelial dysfunction. J Biomed Sci 24: 1-5. doi: 10.1186/s12929-017-0357-5.
24. Rashmi R, Nandi C, Majumdar S (2021). Evolutionarily conserved regions of THAP9 transposase reveal new motifs for subcellular localization and post-translational modification. bioRxiv 2021-08. doi: 10.1101/2021.08.01.454642.
25. Oshima S, Turer EE, Callahan JA, Chai S, Advincula R, Barrera J, Shifrin N, Lee B, Yen B, Woo T, Malynn BA (2009). ABIN-1 is a ubiquitin sensor that restricts cell death and sustains embryonic development. Nature 457(7231): 906-909. doi: 10.1038/nature07575.
26. Kosakovsky Pond SL, Posada D, Gravenor MB, Woelk CH, Frost SD (2006). Automated phylogenetic detection of recombination using a genetic algorithm. Mol Biol Evol 23(10): 1891-1901. doi: 10.1093/molbev/msl051.
27. Sugiura N (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. Commun Stat Theory Methods 7(1): 13-26. doi: 10.1080/03610927808827599.
28. Abdullah M, Rehman MS, Rehman MS, AlKahtane AA, Al-Hazani TM, Hassan FU, Rehman SU (2023). Genome-wide identification, evolutionary and mutational analysis of the buffalo sox gene family. Animals 13(14): 2246. doi: 10.3390/ani13142246.
29. Witte JS (2010). Genome-wide association studies and beyond. Annu Rev Public Health 31(1): 9-20. doi: 10.1146/annurev.publhealth.012809.103723.
30. Rehman SU, Hassan FU, Luo X, Li Z, Liu Q (2021). Whole-genome sequencing and characterization of buffalo genetic resources: recent advances and future challenges. Animals 11(3): 904. doi: 10.3390/ani11030904.
31. Betancur-R R, Wiley EO, Arratia G, Acero A, Bailly N, Miya M, Lecointre G, Orti G (2017). Phylogenetic classification of bony fishes. BMC Evol Biol 17: 1-40. doi: 10.1186/s12862-017-0958-3.
32. Parveen S, Khan MF, Sultana M, Rehman SU, Shafique L (2024). Molecular characterization of doublesex and Mab-3 (DMRT) gene family in *Ctenopharyngodon idella* (grass carp). J Appl Genet 1-2. doi: 10.1007/s13353-024-00924-6.
33. Ikai A (1980). Thermostability and aliphatic index of globular proteins. J Biochem 88(6): 1895-1898. doi: 10.1093/oxfordjournals.jbchem.a133168.
34. Lanneau D, Brunet M, Frisan E, Solary E, Fontenay M, Garrido C (2008). Heat shock proteins: essential proteins for apoptosis regulation. J Cell Mol Med 12(3): 743-761. doi: 10.1111/j.1582-4934.2008.00273.x.
35. Neuwald AF (2016). Gleaning structural and functional information from correlations in protein multiple sequence alignments. Curr Opin Struct Biol 38: 1-8. doi: 10.1016/j.sbi.2016.04.006.
36. Magadum S, Banerjee U, Murugan P, Gangapur D, Ravikesavan R (2013). Gene duplication as a major force in evolution. J Genet 92(1): 155-161. doi: 10.1007/s12041-013-0212-8.
37. Lynch VJ (2007). Inventing an arsenal: adaptive evolution and neofunctionalization of snake venom phospholipase A2 genes. BMC Evol Biol 7: 2. doi: 10.1186/1471-2148-7-2.
38. Voldoire E, Brunet F, Naville M, Volff JN, Galiana D (2017). Expansion by whole genome duplication and evolution of the sox gene family in teleost fish. PLoS One 12(7): e0180936. doi: 10.1371/journal.pone.0180936.
39. King RD, Sternberg MJ (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. Protein Sci 5(11): 2298-2310. doi: 10.1002/pro.5560051116.
40. Osada R, Zaslavsky E, Singh M (2004). Comparative analysis of methods for representing and searching for transcription factor binding sites. Bioinformatics 20(18): 3516-3525. doi: 10.1093/bioinformatics/bth438.
41. Rehman SU, Feng T, Wu S, Luo X, Lei A, Luobu B, Hassan FU, Liu Q (2021). Comparative genomics, evolutionary and gene regulatory regions analysis of casein gene family in *Bubalus bubalis*. Front Genet 12: 662609. doi: 10.3389/fgene.2021.662609.