Pathology & Oncology Research (2020) 26:693–706
https://doi.org/10.1007/s12253-019-00581-9

ORIGINAL ARTICLE

Bioinformatics Analysis Makes Revelation to Potential Properties on Regulation and Functions of Human Sox2

Jianguo Zhang[1] · Jianzhong Zhang[2] · Wenqi Chen[1] · Huiyu Li[1] · Meiying Li[1] · Lisha Li[1]

Received: 11 February 2018 / Accepted: 15 January 2019 / Published online: 2 February 2019

Abstract

Sex determining region Y-box 2 (Sox2) is a transcription factor that is essential for maintaining self-renewal or pluripotency of undifferentiated embryonic stem cells. The expression and distribution of Sox2 in tumor tissues have been extensively recorded, which are related to the progression and metastasis of tumor. However, a complete mechanistic understanding of Sox2 regulation and function remains to be studied. Herein, we show new potential properties of Sox2 regulation and functions from bioinformatics analysis. We use numerous algorithms to characterize the Sox2 gene promoter elements and the Sox2 protein structure, physio-chemical, localization properties and its evolutionary relationships. The expression of Sox2 is regulated by a diverse set of transcription factors and associated with the levels of methylation of CpG Islands in promoters. The structural properties of Sox2 indicate that Sox2 expresses as a stem cell marker in a variety of stem cells. Sox2 together with other transcription factors or proteins regulate the expression of downstream target genes, which makes a great difference to the biological function of stem cells. Not only stem cells, Sox2 also play an important role in tumor cells. In conclusion, this information from bioinformatics analysis will help to understand Sox2 regulation and functions better in future attempts.

Keywords Bioinformatics analysis · SOX2 · Protein-protein interactions · Proteomics · Protein regulation

Introduction

The Sox gene family is composed of a class of SRY related genes, which encode a family of transcription factors that bind to the minor groove in DNA. The Sox gene family consisting of at least 20 members are divided into 8 groups (from A to H), based on their HMG sequence identity in humans [1]. Many Sox genes are involved in sex determination, some are also important in processes such as neuronal development [2]. Sox2 as one of the main members of the Sox family is essential for maintaining self-renewal or pluripotency of undifferentiated embryonic stem cells. Sox2 has a critical role in maintenance of embryonic and neural stem cells. Transcription factor Sox2 can regulate gene expression by interacting with other proteins. Sox2 is a critical downstream target of fibroblast growth factor signaling, which mediates self-renewal of trophoblast stem cells [3].

Sox2 gene is located on 3q26.33, its expression is mainly found in stem cells and many different kinds of cancer cells, including glioma, breast cancer, colorectal cancer and etc. [4]. Over-expression of Sox2 strongly enhanced the growth of tumor cells. It has been reported that Sox2 was involved in the regulation of cell behavior in a variety of pathways, such as Wnt/β-catenin, PI3K-AKT-Mtor, MAP4K4/JNK and etc. [5]. Although some functions of Sox2 are still known, a complete understanding of Sox2 regulation and function remains to be studied. Bioinformatics analysis to predict regulatory mechanism of the gene and protein expression greatly solves these problems.

Bioinformatics is an interdisciplinary field, which combines molecular biology and genetics, computer science, mathematics, statistics to develop methods and software tools for processing and understanding biological data [6]. In the field of genetics and genomics, it helps to sequence and annotate the genome and its observed mutations. Sequence analysis is the analysis of DNA and protein sequences to find functional clues, including homologous identification, multi-sequence alignment, search sequence patterns and evolution

✉ Lisha Li
lilisha@jlu.edu.cn

[1] The Key Laboratory of Pathobiology, Ministry of Education, Norman Bethune Medical College, Jilin University, Changchun, China

[2] Department of Medicine, Qingdao University, Qingdao, China

analysis and other sub-problems, which helps to explain the biological meaning and function of the gene and protein. In addition, protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. Knowledge of this structure is vital in understanding the function of the protein. Moreover, network analysis seeks to understand the relationships within biological networks such as metabolic or protein-protein, small molecular interaction networks. Therefore, bioinformatics tools can aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology as well as, at a more integrative level, analyzing and cataloguing of the biological pathways and networks that are an important part of systems biology [7].

In this study, we used bioinformatics tools to examine the Sox2 sequence to characterize the gene promoter, CpG islands, potential transcriptional factors binding sites (TFBS), encoded protein structure and its subcellular localization, secondary and tertiary structures, and even evolutionary relationships. These characteristics will help define the basis for Sox2 regulation and differential expression in stem cell and cancer. These various bioinformatics tools are among the common tools of molecular biology helping investigators finding leads to investigate genes/proteins.

Materials and Methods

The following prediction methods were used for Sox2 regulatory elements, structure and function: Promoter:- Neural Network Promoter Prediction (http://www.fruitfly.org/seq_tools/promoter.html) [8]; CpG island:- EMBOSS (http://www.ebi.ac.uk/Tools/emboss/) and MethPrimer 2.0 (http://www.urogene.org/methprimer2) [9, 10];TFBS:- PROMO (http://alggen.lsi.upc.es/cgi-bin/promo_v3/promo/promoinit.cgi?dirDB=TF_8.3) [11]; The relatively molecular, amino acid sequences, protein relatively molecular quality, mass of amino acids, theoretical isoelectric point, PI, half-life, unstable factor, the total average hydrophilic:- ProtParam (http://www.expasy.org/tools/protparam.html) [12]; Polarity, hydrophilicity, refractivity:-Protscale (http://www.expasy.org/tools/protscale.html) [12];The secondary structure:- SOPMA (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_sopma.html) and GOR4(http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html) [13, 14]; Signal peptide cutting locus:- SignalP4.1 Server (http://www.cbs.dtu.dk/services/SignalP/) [15]; Nuclear localization signal prediction:- NLSTRADAMUS (http://www.moseslab.csb.utoronto.ca/NLStradamus/) [16];Subcellular location:- TargetP 1.1 Server (http://www.cbs.dtu.dk/services/TargetP/) and PSORT II (https://psort.hgc.jp/form2.html) [17, 18]; Protein structure and function:- InterPro (http://www.ebi.ac.uk/interpro/) [19];Transmembrane domain:- TMHMM (http://www.cbs.dtu.dk/services/TMHMM/) [20]; Advanced structure:- SWISS-MODEL (http://www.expasy.ch/swissmod/SWISS-MODEL.html) [21]; Evolutionary tree:- Clustalx program (http://www.clustal.org/download/current/) and Tree view (http://www.taxonomy.zool-ogy.gla.ac.uk/rod/rod.html) [22, 23]; Protein-protein interactions:- String (https://string-db.org/) [24]; Phylogenetic tree:- BLAST(https://blast.ncbi.nlm.nih.gov/Blast.cgi) and MEGA7(https://mega.co.nz/) [25, 26]; Signaling pathway:- KEGG (http://www.genome.jp/kegg/) [27].

Results

Properties of the TATA-Box, GC-Box, CAAT-Box Motifs in the 5′ Regulatory Region of Sox2

To determine whether there are TATA-box, GC-box, CAAT-box motifs in the 5' regulatory region of human Sox2 and to which the transcription factors TBP, SP1 and CBF can bind. We BLAST searched Sox2 mRNA (NM_003106) and human genomic sequences between −2000 bp to 200 bp from the transcription start site, for the motifs TATAWAW (where W represents A or T), GGGCGG and CCAAT. We identified two GC-box fragments, and one TATA-, but no CCAAT, suggesting that expression and transcriptional activity of Sox2 is regulated by SP1 and TBP.

Promoter, CpG Island and TFBS Prediction in the 5′ Regulatory Region of Sox2

The Sox2 promoter sequence and TFs that bind to this sequence determine the temporal and spatial expression pattern of the gene. Therefore, defining of TFBS is important in the study of gene regulation. We used online programs, such as neural network promoter prediction, EMBOSS, MethPrimer 2.0 and PROMO to predict promoters, CpG Islands and TFBS in 5' regulatory region sequences of human Sox2. We identified four promoters shown in Table 1. CpG Islands are regions with a high frequency of CpG sites. Though objective definitions for CpG islands are limited, the usual formal definition is a region with at least 200 bp, a GC percentage greater than 50%, and an observed-to-expected CpG ratio greater than 60%, but we cannot find CpG Island in this way. Hackenberg M et al. Find that the shortest length of CpG Island can be shorter than 200 bp, as long as it has the meaning of CpG Island [28]. Therefore, we reduced the length of CpG Island. Finally, we find two CpG islands. These CpG islands have different length and location as shown in Fig. 1a.

Table 1 Sox2 promoter site prediction using online Neural Network Promoter Prediction program

Start sites	End sites	Score	Promoter Sequence
382	432	0.88	aagccctttataaaaaagaaatggcatcaggttttttttctttattccc
631	681	0.96	ggagagcggctaaaattaccctcttgggtcctgggcgggcaagattcctg
1689	1739	0.96	gtttaaagaaaaaaaaacccacgtagtcttagtgctgtttacccacttcc
1851	1901	0.87	cccccggcctcccccgcgcggccggcggcgcgggaggccccgccccctttt

MethPrimer 2.0 also identified two different CpG islands (Fig. 1b). PROMO program predicted 77 potential TFBS with a score higher than 85 points, 70 potential TFBS with a score higher than 90 points, 46 potential TFBS with a score higher than 95 points and 19 potential TFBS with a score over 99 points (Fig. 2). Together, these findings suggest that Sox2 expression is associated with the levels of methylation of CpG islands in its promoters. Different transcription start sites makes Sox2 transcription in a different way, and then produce a variety of different biological functions of a transcription product.

The Amino Acid Sequence and Physicochemical Properties of Sox2 Protein

To predict the protein structure of Sox2, we used ProtParam online software (http://au.expasy.org/tools/) to analyze amino acid composition, molecular formula, molecular weight and isoelectric point. Sox2 consists of 317 amino acids with 20 different amino acids, including alanine (8.2%), glycine (11.0%), leucine (6.3%), serine (11.4%) and Methionine (7.9%) (Table 2a). Sox2 has the following properties: protein formula $C_{1467}H_{2321}N_{443}O_{457}S_{26}$, molecular weight 34.30982 kDa, theoretical PI 9.74, estimated half-life in vitro 30 h, instability index 58.73 and the hydrophilic residue ratio is lower than that of the hydrophobic residues. The threshold value of Sox2 hydrophilicity is highest at 31–44, 53–77 and 80–126 bp (Fig. 3a and Table 2b). The threshold value of Sox2 polarity is highest at 6–10, 36–44, 53–126, 158–160, 197–202 and 268–273 bp (Fig. 3b and Table 2b). The threshold value of turn-back coefficient is highest at 5–8, 37–125, 156–164, 174–180, 195–204, 225–227, 272–276, 293–297 bp (Fig. 3c and Table 2b). The overlapping range of these parameters is shown in Table 2b at 37–44, 53–77, 80–125 bp. Taken together, these data suggest that Sox2 mainly contains hydrophilic amino acid and polar amino acids, functional overlapping structure ranges from 37 to 125 bp, and that Sox2 protein maybe a Hydrophilic and unstable protein.

Secondary Structure Prediction for Sox2 Protein

The arrangement of atoms in space of the main polypeptide chain (α helix, extended strand, β turn and random coil) determines basic protein secondary structure. Determination of this arrangement can help predict functions, and protein modifications. We used ExPASy-SOPMA and GOR4 secondary structure prediction module to calculate Sox2 secondary structure, and to draw the structure model. The prediction results can be shown as a peak figure or the diagram can be simplified to show as the random of coiled and folded areas. The SOPMA method identified 95 (29.97%) α helix, 32 (10.09%) extended strands, 26(8.2%) β turns and 164 (51.74%) random coils and irregular coiled and folded structures located mainly at 7–25, 32–44, 48–57, 102–151, 219–237, 246–259 and 296–317 bp between the peak figure and simplified diagram (Fig. 4a and Table 3). The GOR4 method identified 82 (25.87%) α helix, 72 (22.71%) extended strands, and 163 (51.42%) random coils and irregular coiled and folded structures located mainly at 1–24, 31–46, 105–153, 158–189, 198–282 and 297–317 bp (Fig. 4b and Table 3). The comparison of two secondary structure prediction results is shown in Table 3. Sox2 secondary structure mainly consists of the irregular curl overlapping areas at 7–24,32–44, 105–153, 219–237, 246–259 and 297–317 bp, suggesting that these areas are mainly composed of α helix, extended strand and random coil structure. Taken together, the functional domain of Sox2 protein is likely to be limited to these overlapped areas.

Analysis of Signal Peptide Cleavage Site, Subcellular Location, Transmembrane Domains in the Sox2 Protein

Signal peptides direct protein localization in cells and usually consists of 15–30 N-terminal amino acid residues. To analyze the Sox2 signal peptide, we first used the Antheprot signal peptide cutting locus analysis module and the result showed Sox2 may have two signal peptide cutting locus in Fig. 5a. Consistent with previous results, the Sox2 isoelectric point is approximately near 10.0, and the physiological state of Sox2 molecules is closest to pH 9.08, with a positive charge of 10.065, as shown in Fig. 5b.

We also used SignalP4.1 to predict the signal peptide and its cleavage site. As shown in Fig. 5c and Table 4, there is no signal peptide cleavage site. We also used NLStradamus, a simple Hidden Markov Model for nuclear localization signal prediction, to show that there was one nuclear localization signal sequences, suggesting that Sox2 enter the nucleus. TargetP1.1 predicts Sox2 to be located in the other localization within cell (92.5%) (Table 5), and PSORT II Prediction

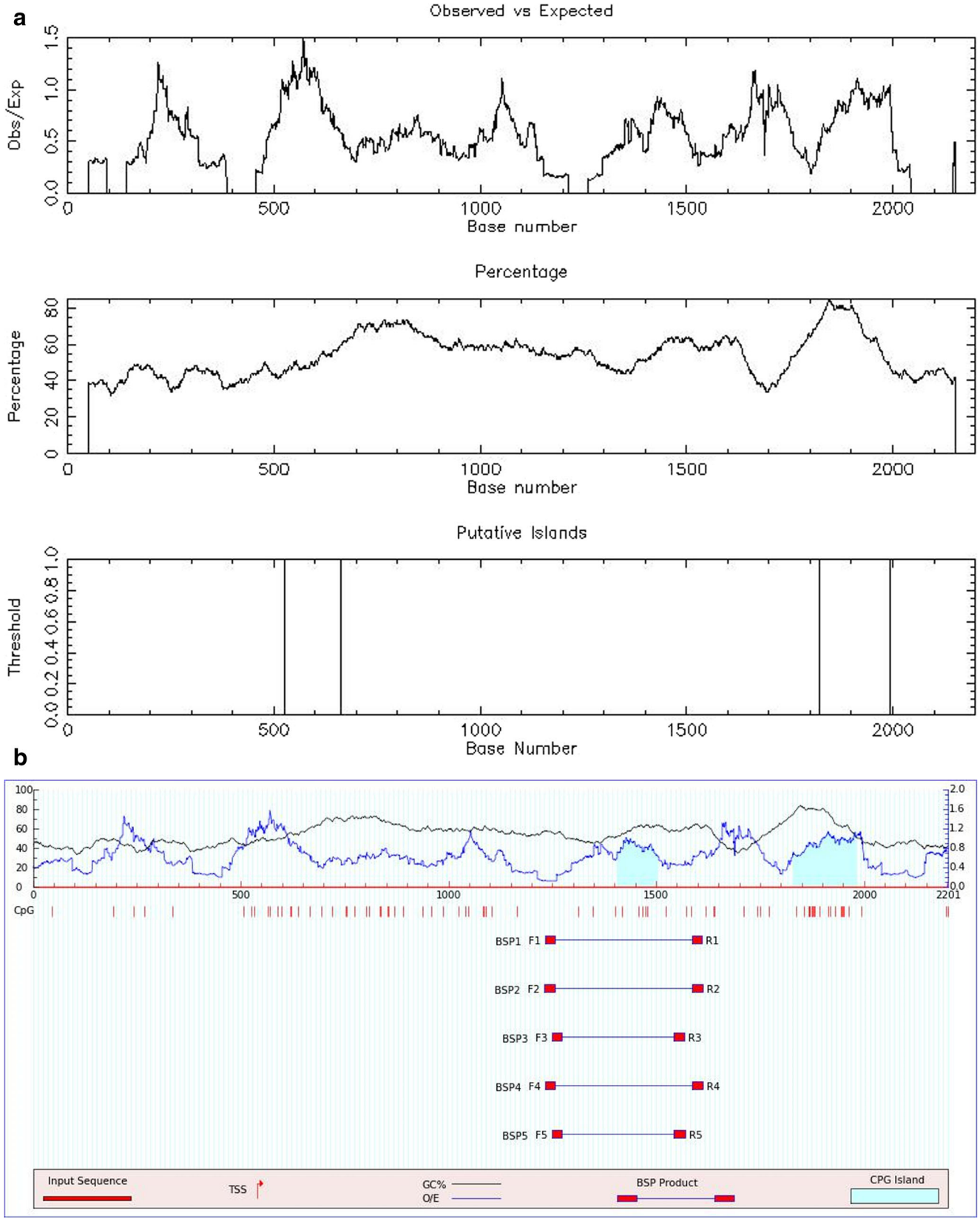

Fig. 1 CpG Island prediction for Sox2 using two prediction programs. **a** CpG island prediction using online EMBOSS. These CpG Islands were 136 bp in length (located at 527–662 bp), 171 bp (1823–1993 bp). **b** CpG island prediction using MethPrimer 2.0 program, Urogene program identified two different CpG islands of 99 bp (1407–1502 bp) and 136 bp (1829–1984 bp)

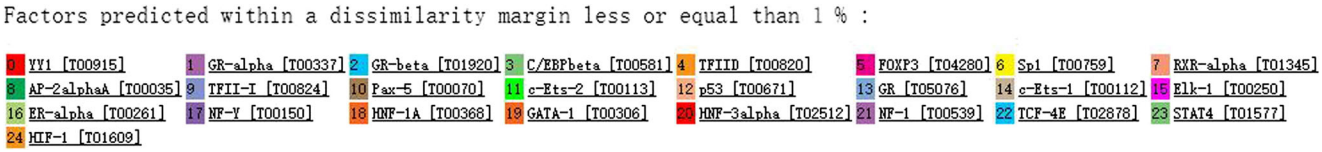

Fig. 2 There are 24 potential TFBSs prediction for Sox2 with a score of 99 using PROMO program prediction

indicated that Sox2 may locate to the nuclear (73.9%), cytoskeletal (8.7%), vesicles of secretory system (4.3%), plasma membrane (4.3%), Golgi (4.3%) and cytoplasmic (4.3%) (Table 6). Therefore, we considered that Sox2 is a nuclear protein that corresponds to its function.

Prediction of Sox2 Protein Domain and Function Site

To predict Sox2 structural domain and important functional sites, we used InterPro software. The results show that Sox2 belongs to the high mobility group box domain superfamily

Table 2 The basic properties of Sox2 analyzed by using ProtParam online software

A. Amino acids composition of Sox2			
Amino acids	Abbreviation	Number	Conposition (%)
Alanine	Ala(A)	26	8.20%
Arginine	Arg(R)	17	5.40%
Asparagine	Asn(N)	15	4.70%
Aspartate	Asp(D)	9	2.80%
Cystine	Cys(C)	1	0.30%
Glutamine	Gla(Q)	18	5.70%
Glutamate	Glu(E)	12	3.80%
Glycine	Gly(G)	35	11.00%
Histidine	His(H)	11	3.50%
Isoleucine	Ile(I)	4	1.30%
Leucine	Leu(L)	20	6.30%
Lysine	Lys(K)	17	5.40%
Methionine	Met(M)	25	7.90%
Phenylalanine	Phe(F)	2	0.60%
Proline	Pro(P)	26	8.20%
Serine	Ser(S)	36	11.40%
Threonine	Thr(T)	14	4.40%
Tryptophan	Trp(W)	3	0.90%
Tryosine	Try(Y)	14	4.40%
Valine	Val(V)	12	3.80%
B. Molecular formula, molecular weight, isoelectric point and other basic properties of Sox2			
Parameters		Predication results	
Formula		C1467H2321N443O457S26	
Molecular weight		34,309.82	
Theoretical pI		9.74	
Total number of atoms		4714	
Number of amino acids		317	
Total number of negatively charged residues		21	
Total number of positively charged residues		34	
Estimated half-life(mammalian reticulocytes, in vitro)		30 h	
Instability index		58.73	
Hydrophilic		31–44, 53–77, 80–126	
Polarity		6–10, 36–44, 53–126, 158–160, 197–202, 268–273	
Turn-back coefficient		5–8, 37–125, 156–164, 174–180, 195–204, 225–227, 272–276, 293–297	
The predicted results the overlapping area		37–44, 53–77, 80–125	

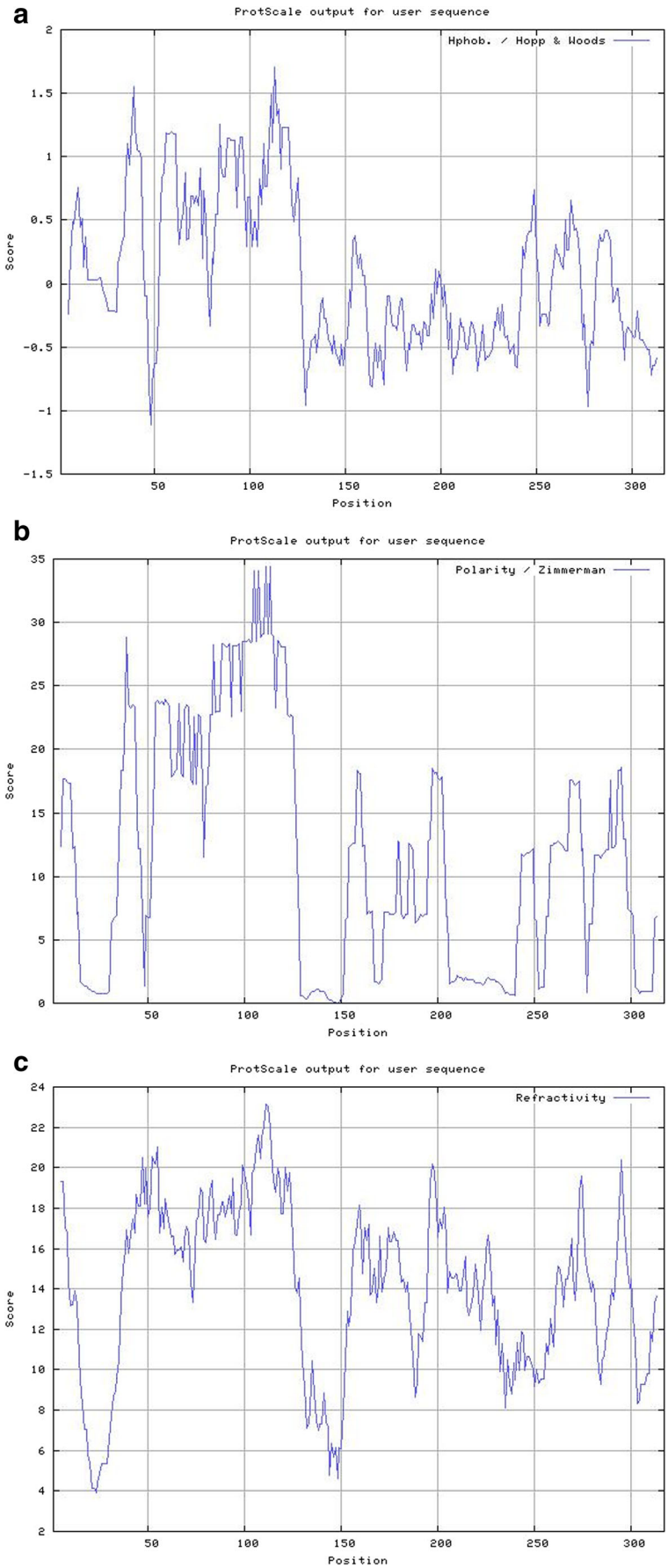

Fig. 3 Nucleotide and deduced amino acid sequences of human Sox2 analyzed using protscale. **a** Hydrophilicity plot analysis of Sox2. **b** Polarity analysis of Sox2. **c** Refractivity analysis of Sox2

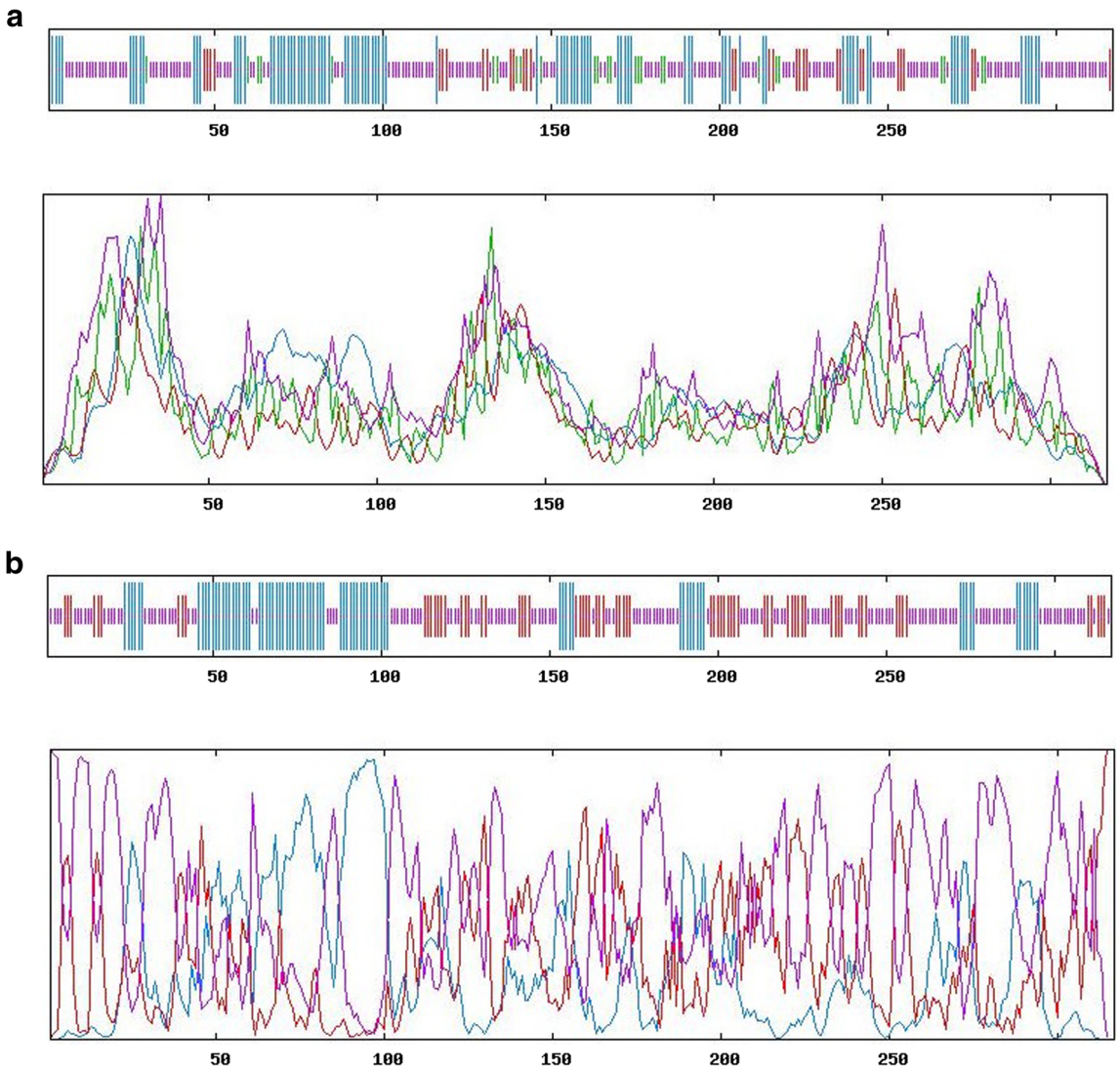

Fig. 4 Secondary structure analysis of Sox2 protein using SOPMA (**a**, up: schematic illustration; down: the peak figure) and GOR4 (**b**, up: schematic illustration; down: the peak figure) prediction software. Helix (blue), spiral structure; Sheet (red), folding; Turn (green), corner structure; Coil (purple), irregular curly structure

(IPR036910) and Sox (IPR22097) as shown in Fig. 6. Its function is associated with structural molecular activity (GO: 0003677) and that action contributes to the regulation of transcription and DNA-templated (GO: 0006355).

Transmembrane domain usually denotes a transmembrane segment of single alpha helix of a transmembrane protein. More broadly, a transmembrane domain is any membrane-spanning protein domain. We used TMHMM Server 2.0 to predict the transmembrane domain in Sox2 and the result showed Sox2 is not have a transmembrane domain (Fig. 7).

Advanced Structure of Human Sox2 Protein

Analysis of the detailed structures of Sox2 will further our understanding of its biological role. To obtain a simulation map of high-level protein structure, the amino acid sequence was analyzed using Swiss-model server. Sox2 protein and its

Table 3 Comparison of Sox2 secondary structure prediction results between SOPMA and GOR4 methods

Secondary structure prediction methods	Prediction results
SOPMA	7–25, 32–44, 48–57, 102–151, 219–237, 246–259, 296–317
GOR4	1–24, 31–46, 105–153, 158–189, 198–282, 297–317
The predicated results of the overlapping area	7–24, 32–44, 105–153, 219–237, 246–259, 297–317

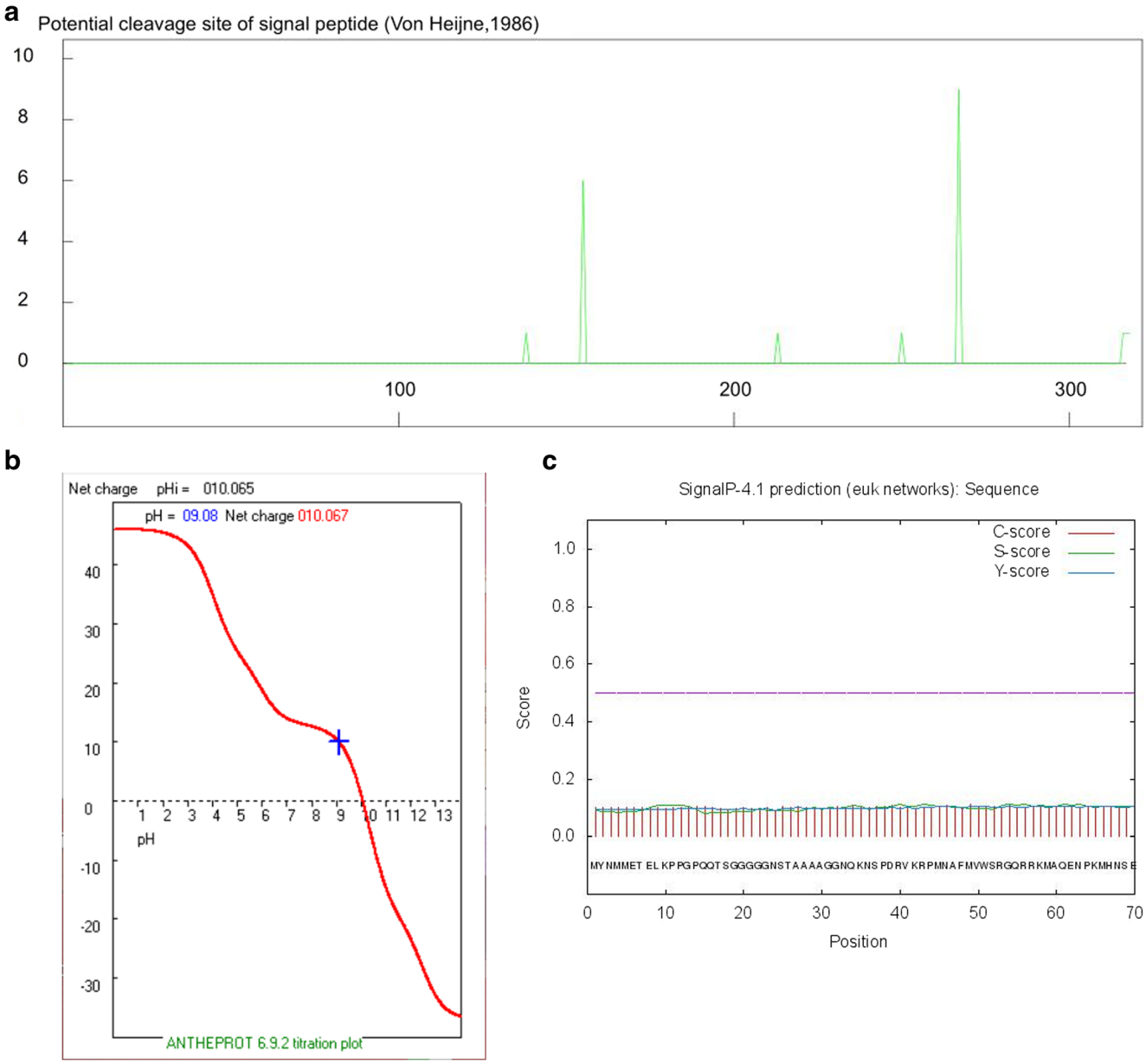

Fig. 5 Signal peptide cleavage site and titration curve of Sox2 using Antheprot and SignalP analysis. **a** Antheprot signal peptide cleavage site, subcellular localization prediction using TargetP1.1 and PSORT II Prediction. **b** Titration curve of Sox2 using Antheprot analysis. **c** Curve of Sox2 protein signal peptide using SignalP-4.1 predictions

structure database template 2le4.1.A, has 69.14% amino acid sequence, which is derived from the template sex-determining region Y protein that GMQE is 0.19 and QMEAN4 is −2.74, which 3D structure as shown in Fig. 8a-d.

Table 4 Results of Sox2 protein signal peptide using SignalP-4.1 euk predictions

Measure	Position	Value	Cutoff	Signal peptide
Max.C	49	0.111		NO
Max.Y	54	0.107		NO
Max.S	56	0.114		NO
Mean S	1–53	0.098		NO
D	1–53	0.102	0.45	NO

Table 5 Subcellular localization prediction of SOX2 using TargetP1.1

Name	Length	mTP	SP	other	Loc	RC
Sequence	317	0.045	0.097	0.925	_	1
cut off		0.000	0.000	0.000		

[i] The location assignment is based on the predicted presence of any SOX2 N-terminal presequences: mitochondrial targeting peptide (mTP) or secretory pathway signal peptide; other, other localization within cells. Number of query sequences: 1, cleavage site predictions are not included using Non-Plant networks

Table 6 Sox2 protein may locate in the nucleus (73.9%), cytoskeleton (8.7%), vesicles of secretory system (4.3%), plasma membrane (4.3%), Golgi (4.3%) and cytoplasm (4.3%) using PSORT II Prediction

K = 9/23
73.9%: nucleus
8.7%: cytoskeleton
4.3%: vesicles of secretory system
4.3%: plasma membrane
4.3%: Golgi
4.3%: cytoplasm
> > prediction for QUERY is nuc (k = 23)

The Evolutionary Tree of Sox Family Amino Acid Sequence Rendering System and Homology Analysis of Human Sox2 Protein Sequences

With Clustalx program construct phylogenetic tree, and with the TreeView software on the system of evolutionary tree edit and comparison, the family system evolutionary tree of Sox protein amino acid sequence was drawn. All members of Sox are clustered closely except Sox3, Sox14, SRY, Sox15, Sox11, Sox5a, Sox5f, Sox5c, Sox5b, Sox13, Sox7, Sox8 and Sox30, 14 members of the others exist in pairs and Sox2 in particular is paired with Sox1, suggesting they are the closest (Fig. 9a). Through the relevant data in the library collection, download sequence similarity information encoding protein of the species to build the system tree. The evolutionary history was inferred using the Neighbor-Joining method and the optimal tree with the sum of branch length = 0.05158097 is shown. The tree is drawn to scale, with branch lengths in the same units as those of the evolutionary distances used to infer the phylogenetic tree (Fig. 9b).

Known and Predicted Protein-Protein Interactions with Sox2

STRING is a database of known and predicted protein interactions that are derived from four sources, including "genomic context," "high-throughput experiments," "coexpression," and "previous knowledge." In this study, many factors showed protein-protein interactions with Sox2 and were displayed in the network view (Fig. 10). The network nodes will display the details about the protein. Based on the above description, we could find that TDGF1, POU5F1, DPPA4, LIN28A, NANOG, KLF4, ASLL4, CTNNB1, FGF2 and STAT3 were shown to have protein-protein interactions with Sox2. Interestingly, Nanog is a pivotal transcription factor in embryonic stem (ES) cells and is essential for maintaining the pluripotency and self-renewal of ES cells, which is similar to the function of Sox2 [29]. Moreover, Nanog and Sox2 participate in the formation of core transcriptional network [30].

Sox2 Related Signaling Pathway Analysis

To identify Sox2-related signaling pathway, we focused on Sox2-related molecular network information in KEGG pathway maps. Sox2 is related to two signaling pathways including ko04390, ko04550 and has been identified as related to Septo-optic dysplasia, Anophthalmia and microphthalmia (A/M). Septo-optic dysplasia is a heterogeneous condition with optic nerve hypoplasia, dysgenesis of the septum pellucidum, and pituitary hypofunction. Septal dysplasia has been confirmed to be associated with the lack of Sonic hedgehog (Shh) in the hypothalamus, and Sox 2 is a dose-dependent regulator of Shh transcription that directly bind and activate a

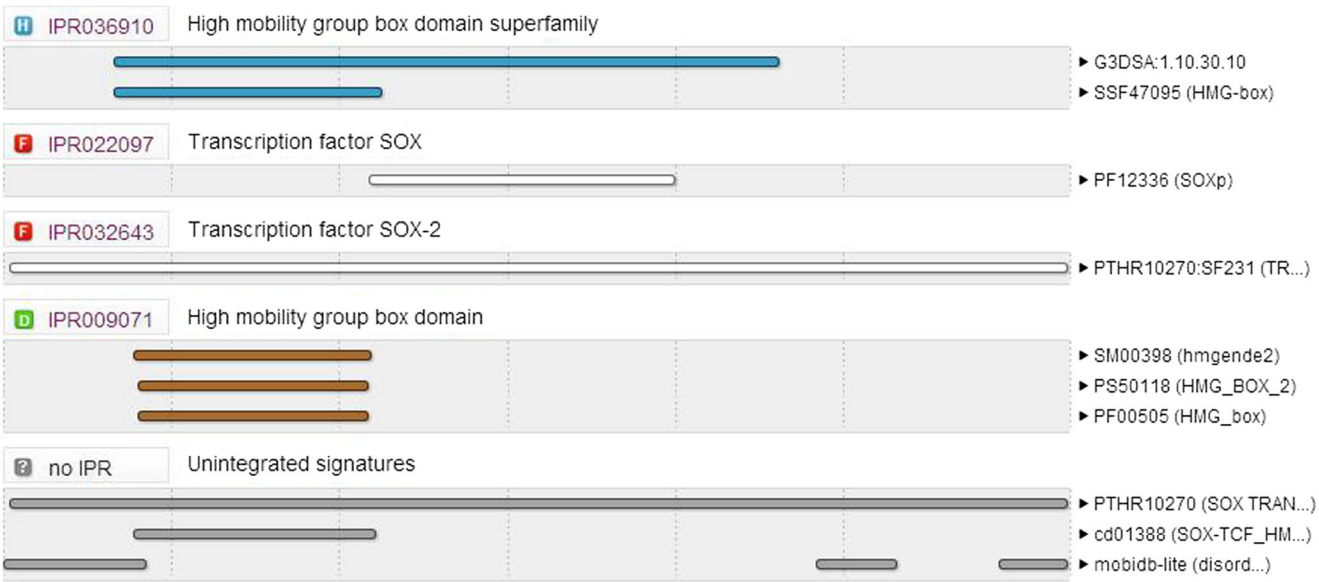

Fig. 6 Function of Sox2. **a** Sox2 belongs to the High mobility group box domain superfamily (IPR036910) and Sox (IPR22097) according to the InterPro software

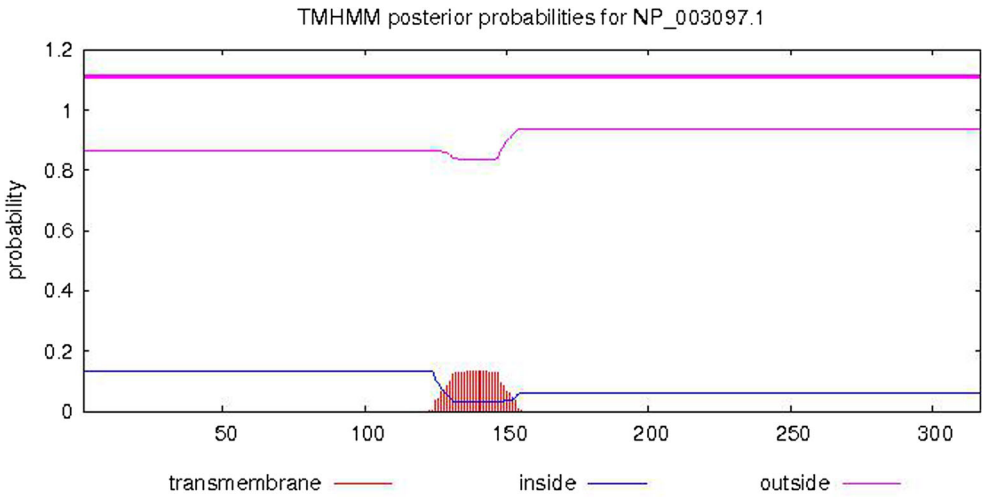

Fig. 7 Transmembrance domain analysis of Sox2 using TMHMM Server 2.0

long-range Shh forebrain enhancer [31]. Therefore, the lack of Sox2 may be related to Septo-optic dysplasia. A/M can be defined as an absence or reduced size of the globe in the orbit. The occurrence of A / M is related to the mutation of Sox2 gene, but the exact reason is still unclear.

The transcriptional coactivator yes-associated protein 1 (YAP1), which is the oncogenic component of the Hippo signaling pathway (ko04390). It has been reported that the effects of YAP1 were mediated through the embryonic stem cell transcription factor Sox2. YAP1 could transcriptionally induce Sox2

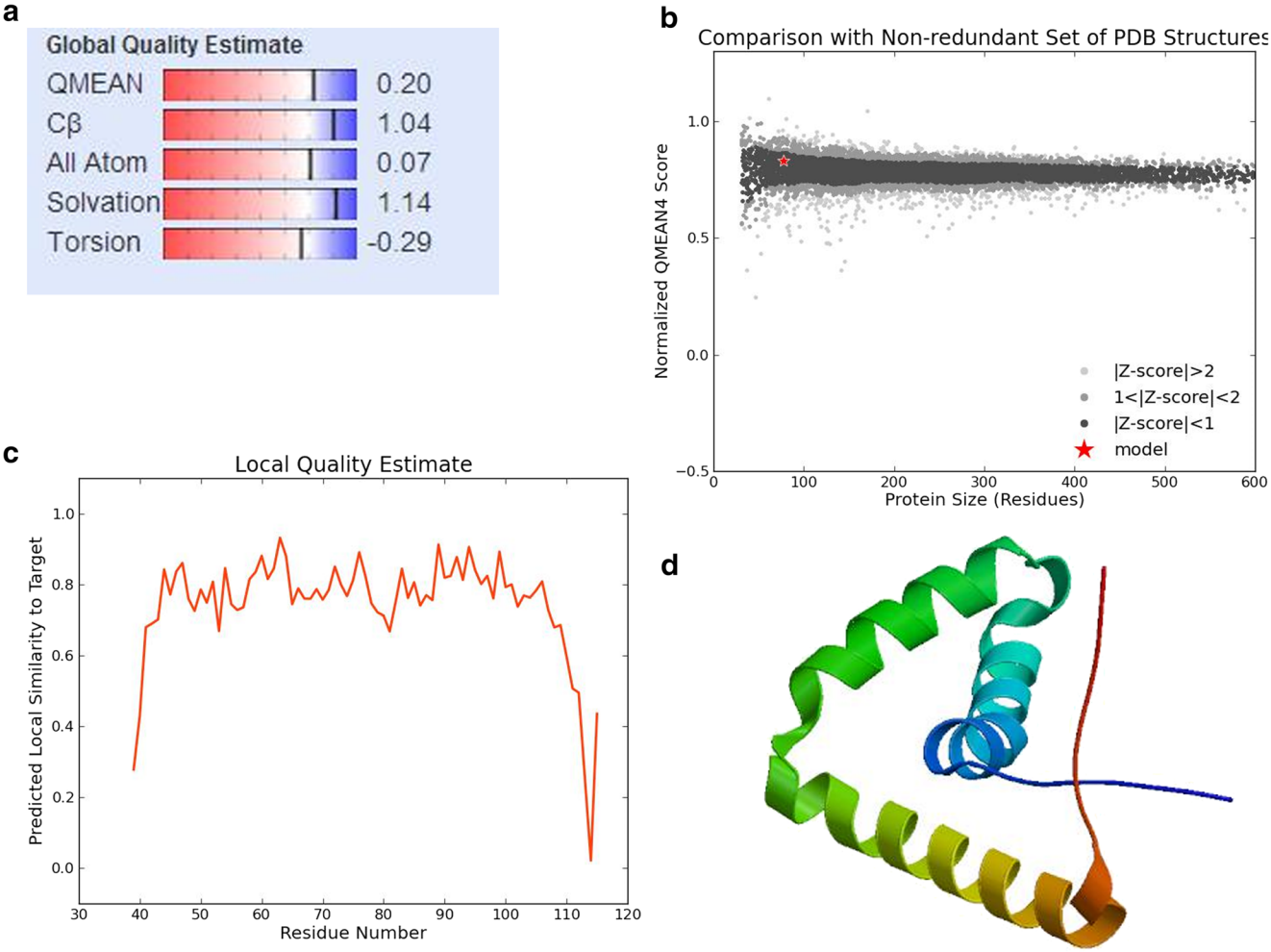

Fig. 8 Predicted three-dimensional structure of Sox2 with Swissmodel server. **a** Sox2 protein and its structure database template 2le4.1.A. **b** Sox2 protein has 69.14% amino acid sequence. **c** GMQE is 0.19 and QMEAN4 is −2.74. **d** The predicted 3D structure of Sox2

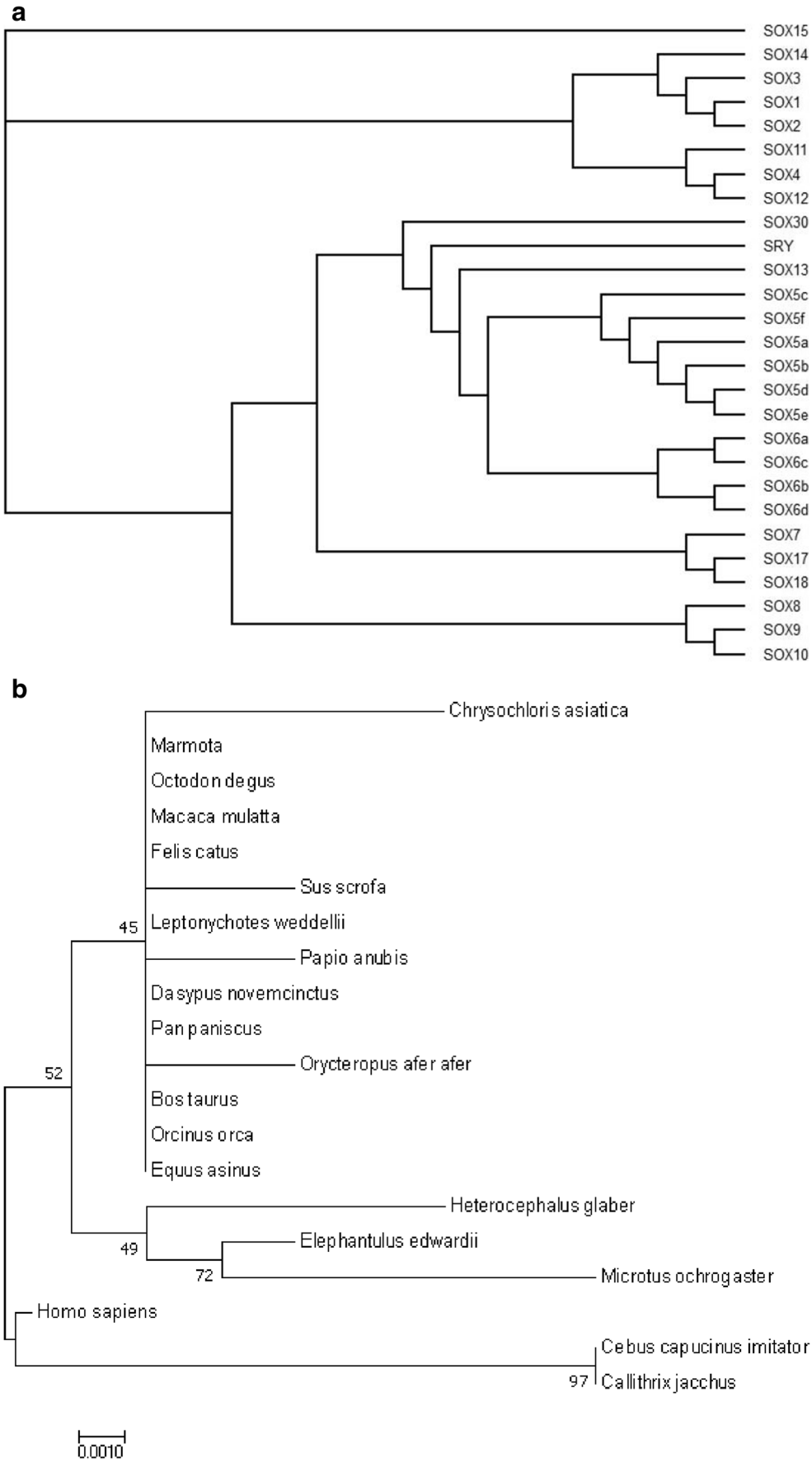

Fig. 9 Evolutionary tree of the SOX family. **a** The family system evolutionary tree of SOX protein amino acid sequence was drawn. **b** The phylogenetic tree of Sox2.Evolutionary analyses were conducted in MEGA7.The evolutionary distances were computed using the Poisson correction method and are in the units of the number of amino acid substitutions per site. The analysis involved 20 amino acid sequences. All positions containing gaps and missing data were eliminated. There were 312 positions in the final dataset

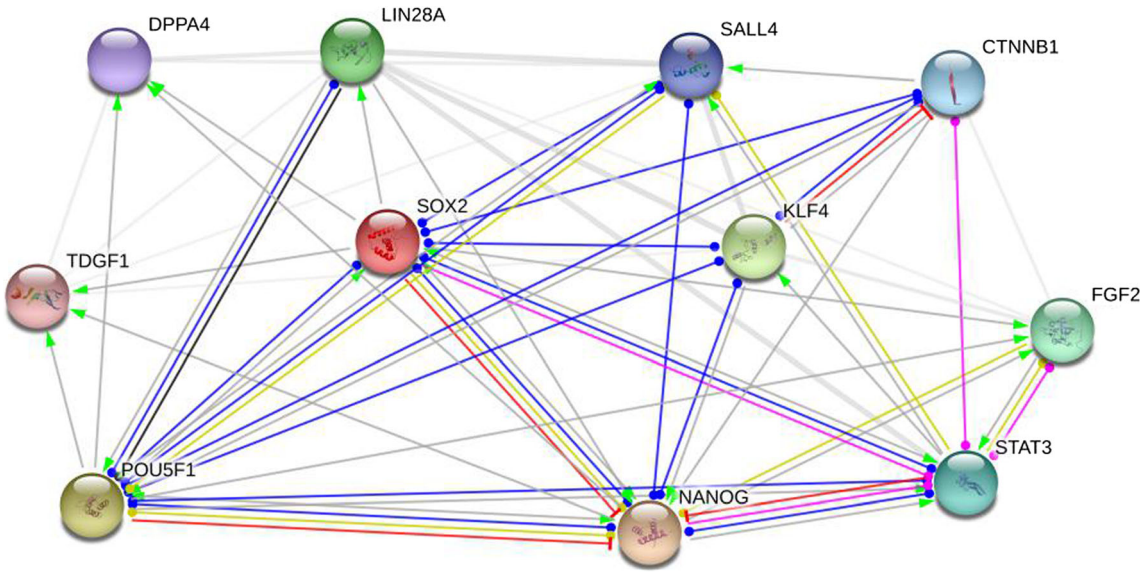

Fig. 10 Known and predicted protein-protein Interactions with Sox2.The edges represent the predicted functional associations and are drawn with up to 7 differently colored lines that represent the existence of the 7 types of evidence used in predicting the associations: (1) red line: fusion evidence; (2) green line: neighborhood evidence; (3) blue line: co-occurrence evidence; (4) purple line: experimental evidence; (5) yellow line: text-mining evidence; (6) light blue line: database evidence; (7) black line: coexpression evidence

through a physical interaction with Oct4 [32]. Meanwhile Sox2 could promote YAP1 expression by binding to an intronic enhancer element. Thus, Sox2 and YAP1 reinforce each other's expression to maintain stemness and tumorigenicity [33] (Fig. 11).The other signaling pathway is ko04550. Through this signaling pathway, Sox2 establishes connections with other signaling molecules, regulates stem cell development, homeostasis maintenance and aging processes. For example, Wnt/β-catenin signaling pathway promotes the expression of differentiation genes, and the expression of Wnt is decreased in hippocampal astrocytes during aging, thereby driving neural progenitor cells to quiescence [34]. Interestingly, Sox2 reduced Wnt signal by up-regulation of APC and GSK 3β and down-regulation of Fzd in osteoblasts [35]. Sox2 knockdown by small interfering RNAs (siRNAs) in two-cell embryo mostly leads to a decrease in Sox2 level at morulae stage, developmental arrest at the morulae/blastocyst transition, and inability to form trophectoderm [36]. Not just stem cells, the research on the relationship between Sox2 and tumor has been carried out in the experimental and clinical research. For example, breast cancer, prostate cancer, oral squamous cell carcinoma, ovarian epithelial cancer, liver cancer etc. [37–41].

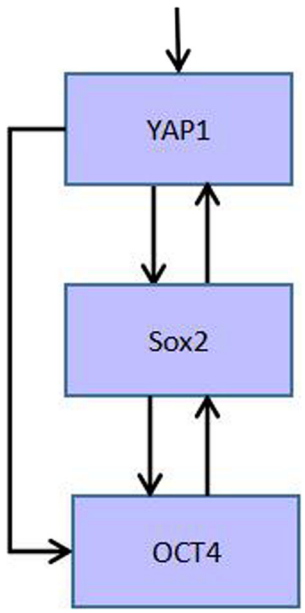

Fig. 11 Sox2-related signaling pathways identified using KEGG pathway searches

Discussion

Sox2 encodes a 34-kDa transcription factor on chromosome 3q26.33. Sox2 has been shown to be expressed differentially in various tumor tissues and cells [42, 43], and to a certain degree its alteration can inhibit or promote tumor cell growth, apoptosis, invasion, migration and EMT in vitro, and methylation of the gene may be involved in tumorigenesis [5, 44–46]. Although we have clarified some functions of Sox2, more questions remain to be answered. Why it is differentially expressed in different cells, which signal pathway relates to it and how it is regulated, suggesting alternative ways are needed to answer the questions. Bioinformatics may be a way complementary to experimental biology in understanding Sox2 regulation and function, which necessarily must be validated by experiments in vitro or in vivo.

We studied the 5' regulatory region of Sox2 using bioinformatics tools and found that it contained two GC-boxes and one TATA-box, two CpG islands. In the 5′ regulatory region

sequence 77 and 19 potential TFBS were predicted with a score of 85–99 and diverse TFs such as SP1 are predicted to bind to these TFBS. These results are supported by published literatures, e.g. Tomin J and Martinez-Cruzado L documented that the associations between Sox2 expression and mRNA levels of SP1 in sarcoma [47], as well as transcription factors such as P53 and P21 [48], the expression of Sox2 is associated with methylating CpG islands; treatment with 5-azacitidine and Dimethyl Sulfoxide induced marked decreases in the levels of methylation of CpG islands in the promoters of Sox2 [49, 50]. In addition, it is reported that some TFBS, including those for YB-1, HMGA2, ARID3B, FOXM1 and MSX2 have been shown to contribute to Sox2 regulation [51–54].

The amino acid sequence of Sox2 was also analyzed by bioinformatics methods. The results showed that the isoelectric point was 9.74; alanine, glycine, leucine, serine and Methionine were the most abundant amino acids; Sox2 was an unstable protein, which has Hydrophilic amino acids. The main predicted secondary structures of Sox2 are α helices, extended strands and random coils in our results. Our Sox evolutionary tree shows Sox2 and Sox1 are most closely related, which is similar to a report by Tenley C. Archer et al. and it is reported that continuous expression of Sox1 and Sox2 in transgenic embryos represses neuron differentiation and inhibits anterior development while increasing cell proliferation [55], suggesting they have common functions. Sox2 related signaling pathway analysis results obtained from these tools are used to support Sox2 promotes YAP1 expression in osteosarcoma [33].

In summary, these results reveal that Sox2 gene may have diverse transcription start sites and its transcription is regulated by DNA methylation and transcription factors such as SP1. Additionally, Sox2 may act as a transcription factor in the nucleus to regulate the expression of other genes.

Acknowledgements This study was funded by the National Natural Science Foundation of China (Grant No. 31150007, 31201052), Jilin Province Science and Technology Development Program for Young Scientists Fund (Grant No. 20190103094JH), and Science and Technology Projects of the Education Department of Jilin Province (Grant No. [2016]445).

Compliance with Ethical Standards

Competing Interests The authors declare that they have no competing interests.

References

1. Sarkar A, Hochedlinger K (2013) The sox family of transcription factors: versatile regulators of stem and progenitor cell fate. Cell Stem Cell 12:15–30
2. Rex M, Church R, Tointon K, Ichihashi RMA, Mokhtar S, Uwanogho D, Sharpe PT, Scotting PJ (1998) Granule cell development in the cerebellum is punctuated by changes in Sox gene expression. Mol Brain Res 55:28–34
3. Adachi K, Nikaido I, Ohta H, Ohtsuka S, Ura H, Kadota M, Wakayama T, Ueda HR, Niwa H (2013) Context-dependent wiring of Sox2 regulatory networks for self-renewal of embryonic and trophoblast stem cells. Mol Cell 52:380–392
4. Lin BY, Huang XF, Han X, Foltz G (2011) SOX2 (SRY (sex determining region Y)-box 2). Atlas Genet Cytogenet Oncol Haematol 15(12):1054–1057
5. Zhao X, Sun B, Sun D et al (2015) Slug promotes hepatocellular cancer cell progression by increasing sox2 and nanog expression. Oncol Rep 33:149–156
6. Can T (2014) Introduction to bioinformatics. Methods Mol Biol 1107:51–71
7. Eck RV, Dayhoff MO (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science 152:363–366
8. Reese MG (2001) Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. Comput Chem 26:51–56
9. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R (2015) The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 43: W580–W584
10. Li LC, Dahiya R (2002) MethPrimer: designing primers for methylation PCRs. Bioinformatics 18:1427–1431
11. Messeguer X, Escudero R, Farre D, Nunez O, Martinez J, Alba M (2002) PROMO: detection of known transcription regulatory elements using species-tailored searches. Bioinformatics 18:333–334
12. Wilkins MR, Gasteiger E, Bairoch A et al (1999) Protein identification and analysis tools in the ExPASy server. Methods Mol Biol 112:531–552
13. Geourjon C, Deleage G (1995) SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. Comput Appl Biosci : CABIOS 11:681–684
14. Garnier J: GOR secondary structure prediction method version IV. Methods Enzymol, RF Doolittle Ed. 266: 540-553, 1998
15. Petersen TN, Brunak S, von Heijne G, Nielsen H (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8:785–786
16. Nguyen Ba AN, Pogoutse A, Provart N, Moses AM (2009) NLStradamus: a simple hidden Markov model for nuclear localization signal prediction. BMC Bioinf 10:202
17. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005–1016
18. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585–W587
19. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang HY, Dosztányi Z, el-Gebali S, Fraser M, Gough J, Haft D, Holliday GL, Huang H, Huang X, Letunic I, Lopez R, Lu S, Marchler-Bauer A, Mi H, Mistry J, Natale DA, Necci M, Nuka G, Orengo CA, Park Y, Pesseat S, Piovesan D, Potter SC, Rawlings ND, Redaschi N, Richardson L, Rivoire C, Sangrador-Vegas A, Sigrist C, Sillitoe I, Smithers B, Squizzato S, Sutton G, Thanki N, Thomas PD, Tosatto SCE, Wu CH, Xenarios I, Yeh LS, Young SY, Mitchell AL (2017) InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45: D190–D199
20. Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden

Markov model: application to complete genomes. J Mol Biol 305: 567–580
21. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Cassarino TG, Bertoni M, Bordoli L, Schwede T (2014) SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252–W258
22. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG (2007) Clustal W and Clustal X version 2.0. Bioinformatics 23:2947–2948
23. Saldanha AJ (2004) Java Treeview-extensible visualization of microarray data. Bioinformatics 20:3246–3248
24. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 43:D447–D452
25. Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL (2008) NCBI BLAST: a better web interface. Nucleic Acids Res 36:W5–W9
26. Kumar S, Stecher G, Tamura K (2016) MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol 33:1870–1874
27. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res 44:D457–D462
28. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martínez-Aroza J, Oliver JL (2006) CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinf 7:446
29. Wu Y, Guo Z, Wu H, Wang X, Yang L, Shi X, et al (2012) SUMOylation represses Nanog expression via modulating transcription factors Oct4 and Sox2. PLoS One 7(6):e39606
30. Orkin SH (2005) Chipping away at the embryonic stem cell network. Cell 122:828–830
31. Zhao L, Zevallos SE, Rizzoti K, Jeong Y, Lovell-Badge R, Epstein DJ (2012) Disruption of SoxB1-dependent sonic hedgehog expression in the hypothalamus causes septo-optic dysplasia. Dev Cell 22:585–596
32. Bora-Singhal N, Nguyen J, Schaal C, Perumal D, Singh S, Coppola D, Chellappan S (2015) YAP1 regulates OCT4 activity and SOX2 expression to facilitate self-renewal and vascular mimicry of stem-like cells. Stem Cells 33:1705–1718
33. Verma NK, Gadi A, Maurizi G, Roy UB, Mansukhani A, Basilico C (2017) Myeloid zinc finger 1 and GA binding protein co-operate with Sox2 in regulating the expression of yes-associated protein 1 in cancer cells. Stem Cells 35(12):2340–2350
34. Miranda CJ, Braun L, Jiang YY, Hester ME, Zhang L, Riolo M, Wang H, Rao M, Altura RA, Kaspar BK (2012) Aging brain microenvironment decreases hippocampal neurogenesis through Wnt-mediated survivin signaling. Aging Cell 11:542–552
35. Seo E, Basu-Roy U, Zavadil J, Basilico C, Mansukhani A (2011) Distinct functions of Sox2 control self-renewal and differentiation in the osteoblast lineage. Mol Cell Biol 31:4593–4608
36. Keramari M, Razavi J, Ingman KA, Patsch C, Edenhofer F, Ward CM, et al (2010) Sox2 is essential for formation of trophectoderm in the preimplantation embryo. PLoS One 5(11):e13952
37. Piva M, Domenici G, Iriondo O, Rábano M, Simões BM, Comaills V, Barredo I, López-Ruiz JA, Zabalza I, Kypta R, Vivanco MM (2014) Sox2 promotes tamoxifen resistance in breast cancer cells. EMBO Mol Med 6:66–79
38. Du J, Li B, Fang Y et al (2015) Overexpression of class III β-tubulin, Sox2, and nuclear Survivin is predictive of taxane resistance in patients with stage III ovarian epithelial cancer. BMC Cancer 15:536
39. Li D, Zhao L-N, Zheng X-L et al (2014) Sox2 is involved in paclitaxel resistance of the prostate cancer cell line PC-3 via the PI3K/Akt pathway. Mol Med Rep 10:3169–3176
40. Wen W, Han T, Chen C, Huang L, Sun W, Wang X, Chen SZ, Xiang DM, Tang L, Cao D, Feng GS, Wu MC, Ding J, Wang HY (2013) Cyclin G1 expands liver tumor-initiating cells by Sox2 induction via Akt/mTOR signaling. Mol Cancer Ther 12:1796–1804
41. Chou M-Y, Hu F-W, Yu C-H, Yu C-C (2015) Sox2 expression involvement in the oncogenicity and radiochemoresistance of oral cancer stem cells. Oral Oncol 51:31–39
42. Rudin CM, Durinck S, Stawiski EW, Poirier JT, Modrusan Z, Shames DS, Bergbower EA, Guan Y, Shin J, Guillory J, Rivers CS, Foo CK, Bhatt D, Stinson J, Gnad F, Haverty PM, Gentleman R, Chaudhuri S, Janakiraman V, Jaiswal BS, Parikh C, Yuan W, Zhang Z, Koeppen H, Wu TD, Stern HM, Yauch RL, Huffman KE, Paskulin DD, Illei PB, Varella-Garcia M, Gazdar AF, de Sauvage FJ, Bourgon R, Minna JD, Brock MV, Seshagiri S (2012) Comprehensive genomic analysis identifies SOX2 as a frequently amplified gene in small-cell lung cancer. Nat Genet 44:1111–1116
43. Chen S, Li X, Lu D et al (2013) SOX2 regulates apoptosis through MAP4K4-survivin signaling pathway in human lung cancer cells. Carcinogenesis 35:613–623
44. Ji J, Zheng PS (2010) Expression of Sox2 in human cervical carcinogenesis. Hum Pathol 41:1438–1447
45. Huang X, Xiong M, Jin Y et al (2016) Evidence that high-migration drug-surviving MOLT4 leukemia cells exhibit cancer stem cell-like properties. Int J Oncol 49:343–351
46. Wang L, Yang H, Lei Z, Zhao J, Chen Y, Chen P, Li C, Zeng Y, Liu Z, Liu X, Zhang HT (2016) Repression of TIF1gamma by SOX2 promotes TGF-beta-induced epithelial-mesenchymal transition in non-small-cell lung cancer. Oncogene 35:867–877
47. Tomin J, Martinez-Cruzado L, Santos L et al (2016) Inhibition of SP1 by the mithramycin analog EC-8042 efficiently targets tumor initiating cells in sarcoma. Oncotarget 7:30935–30950
48. Marques-Torrejon MA, Porlan E, Banito A et al (2013) Cyclin-dependent kinase inhibitor p21 controls adult neural stem cell expansion by regulating Sox2 gene expression. Cell Stem Cell 12:88–100
49. Alonso MM, Diez-Valle R, Manterola L, Rubio A, Liu D, Cortes-Santiago N, Urquiza L, Jauregi P, de Munain AL, Sampron N, Aramburu A, Tejada-Solís S, Vicente C, Odero MD, Bandrés E, García-Foncillas J, Idoate MA, Lang FF, Fueyo J, Gomez-Manzano C (2011) Genetic and epigenetic modifications of Sox2 contribute to the invasive phenotype of malignant gliomas. PLoS One 6: e26740
50. Li X, Wang YK, Song ZQ, Du ZQ, Yang CX (2016) Dimethyl sulfoxide perturbs cell cycle progression and spindle Organization in Porcine Meiotic Oocytes. PLoS One 11:e0158074
51. Jung K, Wu F, Wang P, Ye X, Abdulkarim BS, Lai R (2014) YB-1 regulates Sox2 to coordinately sustain stemness and tumorigenic properties in a phenotypically distinct subset of breast cancer cells. BMC Cancer 14:328
52. Chien CS, Wang ML, Chu PY, Chang YL, Liu WH, Yu CC, Lan YT, Huang PI, Lee YY, Chen YW, Lo WL, Chiou SH (2015) Lin28B/Let-7 regulates expression of Oct4 and Sox2 and reprograms oral squamous cell carcinoma cells to a stem-like state. Cancer Res 75:2553–2565
53. Wu QQ, Zhang LS, Su P, Lei X, Liu X, Wang H, Lu L, Bai Y, Xiong T, Li D, Zhu Z, Duan E, Jiang E, Feng S, Han M, Xu Y, Wang F, Zhou J (2015) MSX2 mediates entry of human pluripotent stem cells into mesendoderm by simultaneously suppressing SOX2 and activating NODAL signaling. Cell Res 25:1314–1332
54. Lee Y, Kim KH, Kim DG, Cho HJ, Kim Y, Rheey J, Shin K, Seo YJ, Choi YS, Lee JI, Lee J, Joo KM, Nam DH (2015) FoxM1 promotes stemness and radio-resistance of glioblastoma by regulating the master stem cell regulator Sox2. PLoS One 10:e0137703
55. Archer TC, Jin J, Casey ES (2011) Interaction of Sox1, Sox2, Sox3 and Oct4 during primary neurogenesis. Dev Biol 350:429–440