Phylogenetic and Structural Analysis of the Pluripotency Factor Sex-Determining Region Y box2 Gene of *Camelus dromedarius* (cSox2)

Abdullah Alawad[1], Sultan Alharbi[1], Othman Alhazzaa[1], Faisal Alagrafi[1], Mohammed Alkhrayef[1], Ziyad Alhamdan[1], Abdullah Alenazi[1], Hasan Al-Johi[2], Ibrahim O. Alanazi[2] and Mohamed Hammad[1,3]

[1]National Center for Stem Cell Technology, King Abdulaziz City for Science and Technology, Riyadh, KSA. [2]National Center for Genomic Technology, King Abdulaziz City for Science and Technology, Riyadh, KSA. [3]SAAD Research and Development Center, Clinical Research Laboratory and Radiation Oncology, SAAD Specialist Hospital, Al Khobar, KSA.

ABSTRACT: Although the sequencing information of Sox2 cDNA for many mammalian is available, the Sox2 cDNA of *Camelus dromedaries* has not yet been characterized. The objective of this study was to sequence and characterize Sox2 cDNA from the brain of *C. dromedarius* (also known as Arabian camel). A full coding sequence of the Sox2 gene from the brain of *C. dromedarius* was amplified by reverse transcription PCRjmc and then sequenced using the 3730XL series platform Sequencer (Applied Biosystem) for the first time. The cDNA sequence displayed an open reading frame of 822 nucleotides, encoding a protein of 273 amino acids. The molecular weight and the isoelectric point of the translated protein were calculated as 29.825 kDa and 10.11, respectively, using bioinformatics analysis. The predicted cSox2 protein sequence exhibited high identity: 99% for *Homo sapiens*, *Mus musculus*, *Bos taurus*, and *Vicugna pacos*; 98% for *Sus scrofa* and 93% for *Camelus ferus*. A 3D structure was built based on the available crystal structure of the HMG-box domain of human stem cell transcription factor Sox2 (PDB: 2 LE4) with 81 residues and predicting bioinformatics software for 273 amino acid residues. The comparison confirms the presence of the HMG-box domain in the cSox2 protein. The orthologous phylogenetic analysis showed that the Sox2 isoform from *C. dromedarius* was grouped with humans, alpacas, cattle, and pigs. We believe that this genetic and structural information will be a helpful source for the annotation. Furthermore, Sox2 is one of the transcription factors that contributes to the generation-induced pluripotent stem cells (iPSCs), which in turn will probably help generate camel induced pluripotent stem cells (CiPSCs).

KEYWORDS: *Camelus dromedarius*, Sox2, sequence analysis, bioinformatics, 3D model, camel-induced pluripotent stem cells (CiPSCs)

CITATION: Alawad et al. Phylogenetic and Structural Analysis of the Pluripotency Factor Sex-Determining Region Y box2 Gene of *Camelus dromedarius* (cSox2). *Bioinformatics and Biology Insights* 2016:10 111–120 doi: 10.4137/BBI.S39047.

TYPE: Original Research

RECEIVED: February 09, 2016. **RESUBMITTED:** May 15, 2016. **ACCEPTED FOR PUBLICATION:** May 21, 2016.

ACADEMIC EDITOR: Thomas Dandekar, Associate Editor

PEER REVIEW: Eight peer reviewers contributed to the peer review report. Reviewers' reports totaled 2,165 words, excluding any confidential comments to the academic editor.

FUNDING: The study was sponsored by the National Center for Stem Cell Technology (NCSCT) and King Abdulaziz City for Science and Technology (KACST). Special thanks are due to SAAD Research & Development Center at SAAD Specialist Hospital for the support. This work was supported by a 2012 National Science, Technology and Innovation Plan (NSTIP) Translational Stem Cell Research (TSCR) grant (no. 33–837). The authors also gratefully acknowledge the financial support from KACST under NSTIP grant (no. 32–685). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: Authors disclose no potential conflicts of interest.

CORRESPONDENCE: stemcell@kacst.edu.sa

Introduction

Transcription factors (TFs) are often regarded as being composed of a sequence-specific DNA-binding domain and a functional domain. Thus, such domain acts to aim the TF to particularly regulatory regions in the genome based on its affinity for a certain DNA sequence and the transacting domain then carries out regulatory effects on the appropriate gene.[1,2] Sox2 is a member of the Sox family transcription factors and has conserved high mobility group-box (HMG-box) DNA-binding domain, which has been identified in mice and humans.[3,4] It was first described by the discovery of the mammalian testis-determining factor.[5,6] Sox2 is classified as a member of the SoxB1 group (also known as SoxNeuro), which also includes Sox1 and Sox3. Though Sox1, Sox2, and Sox3 proteins share about 80% sequence similarity and are functionally redundant, Sox2 can apply different functions in a biologically context-dependent manner and is essential for embryonic development.[7] For example, in early murine neural progenitors, Sox2 is demonstrated to interact with the brain-specific Pit-Oct-Unc (POU) domain transcription factor to activate the neural progenitor cells.[8] Moreover,[9] Sox2 is over-expressed in stem cells and different kinds of cells, including brain cells.[10,11] Therefore, it is considered one of the significant reprogramming factors that is involved in the remodeling of the cell fate.[9]

Although considerable progress has been made, the role of Sox2 in stem cells (SCs), self-renewal, and pluripotency is still not fully understood.

The induced pluripotent stem cells (iPSCs) from rat,[12] rhesus monkey,[13] pig,[14] marmoset,[15] dog,[16] and human[17] have invigorated regenerative medicine research and enabled apparently unlimited applications. However, the derivation of

ciPSCs from camel somatic cells has not yet been reported in spite of its distinctive value as a better model for many human pathological conditions compared with small animals. To ensure successful reprogramming, a thorough knowledge of transcription factors that reprogram cell fate is necessary. Despite sequencing the whole genomes of both the Arabian camel and the Bactrian camel,[18–20] there is lack of sequencing information studies that target the pluripotency genes of camel.

Although the Sox2 protein is highly conserved in a variety of distinct species from bacteria to mammals, there are no reports about the Sox2 protein of the Arabian camel. In this study, we sequenced and identified the mRNA of cSox2 gene using bioinformatics approaches. Public datasets were used to construct the phylogenetic tree using the available amino acid sequences. In addition, we examined the evolutionary conserved domains of cSox2. The present study aimed to (1) sequence the mRNA of cSox2 gene, (2) predict its amino acid sequence, (3) use the homology-based method to identify homology in the regulatory domains for cSox2 and Sox2 in six different species, (4) construct the phylogenetic tree of cSox2 with six mammalian species using multiple sequence alignment of Sox2 proteins under study, and (5) model and contrast existing mammalian homologues with the predicted cSox2 3D structure.

Our results based on full-length mRNA, homology, read sequencing quality, and comparative genetic analysis suggested that we have successfully sequenced Sox2 mRNA in the Arabian camel that matched known coding sequences in other mammalian species. To the best of our knowledge, the data presented here represented novel cSox2 mRNA sequence data as well as its 3D-modeling protein and we believe this genetic and structural information will become a helpful resource for the annotation. Our work based on comparative cSox2 mRNA will have significant impact on iPSCs research, since we have sequenced and described one of the reprogramming transcriptional factors, which is the backbone of the iPSCs technology.

Materials and Methods

Sample collection. Camel brain tissue was obtained from an adult male camel slaughtered at the main slaughterhouse in Southern Riyadh. Brain tissue samples were taken from different parts of the camel brain, which were then immersed in RNAlater solution (Qiagen) to protect them against RNA degradation; they were then stored at –20 °C. Strains of *Escherichia coli* were cultured in the Luria-Bertain (LB) medium with 100 mg/mL ampicillin, unless otherwise indicated. This study was approved by the Local Ethics Committee in KACST.

Primer design. Primers (Table 1) were designed according to the data from the Arabian camel genome project (http://camel.kacst.edu.sa/) using Primer-BLAST at GeneBank website (http://www.ncbi.nlm.nih.gov/tools/primer-blast/).

Table 1. Checklist of primers used in this study.

PRIMERS	PRIMER SEQUENCE	PRODUCT (bp)
cSOX2 F1	AACCAGAAGAACAGCCCGGA	901
cSOX2 R1	TGAAAATTTCTTCCCTCTCCCC	
cSOX2 F2	TGAACGCCTTCATGGTGTGG	858
cSOX2 R2	TTTCTTCCCTCTCCCCCTCC	
Camel-β-Actin F	GCCATGGATGACGATATTGCT	1150
Camel-β-Actin R	GGAACGTAACTAAGTCCGCC	

Beta actin was used as an endogenous control. A couple of primers were examined using Amplifx 1.7.1 (http://crn2m.univ-mrs.fr/pub/amplifx-dist) in order to determine the optimized annealing temperatures to generate PCR products constituting a complete coding sequence that was subjected to sequencing. The sequence, amplification product length of every primer pair is presented in Table 1.

Isolation of RNA and synthesis of cDNA. About 50 mg of brain tissues were collected from male camels. RTL lysis buffer (Qiagen) complemented with 1% 2-mercaptoethanol was used as a medium for homogenization. E.Z.N.A. kit (Omega Bio-Tek) was used to isolate the total RNA per the manufacturer's instructions. Nanodrop spectrophotometer (NanoDrop; ThermoScientific) was used to quantify samples at 260 nm and the quality of RNA sample was evaluated using denaturing formaldehyde agarose gel (1%) electrophoresis. Total RNAs (2 µg) were reverse transcribed to single-stranded cDNA by ImProm-II Reverse Transcription System (Promega), per the manufacturers' suggestions, with the next cycling conditions: 96 °C for 1 minutes, followed by 40 cycles at 94 °C for 30 seconds, 65 °C for 30 seconds, and 72 °C for 1 minute.

PCR. Gradient PCR was adopted using descending annealing temperatures from 60 to 50 °C with a typical 25 µL reaction volume as follows:

- GoTaqGreen Master Mix (Promega): 50% reaction volume
- cDNA: 20% reaction volume
- forward primer (5 pmol): 4% reaction volume
- reverse primer (5 pmol): 4% reaction volume
- nuclease free water: 22% reaction volume

The PCR conditions were as follows: one cycle at 95 °C for 2 minutes, 25 cycles at 94 °C for 30 seconds, 60–50 °C for 45 seconds, 72 °C for 105 seconds, and 72 °C for 5 minutes for final extension. Electrophoresis on 1.2% agarose gel was conducted to verify PCR products (Supplementary Fig. 1).

Sequencing of DNA and prediction of amino acid sequence. The cSox2 complete coding sequence was generated by the 3730XL series platform sequencer at KACST. The primers pairs, cSox2F1/cSox2R1 and cSox2F2/cSox2R2, were used to amplify 901 and 858 bp cDNA fragments by

PCR. They were then used to sequence using 3730XL DNA Sequencer; Geneious 7.1.7[21] was consequently used to analyze nucleotide sequences in forward and reverse directions. The similarity of the obtained sequence was examined in the GenBank database using the BLASTN algorithm on the NCBI Blast server (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Multiple sequence alignment and phylogenetic relationship analysis. Geneious 7.1.7 software was utilized to *in silico* translate the cSox2 mRNA to the deduced cSox2 amino acids sequence, which was then contrasted with the current sequences in Protein Database at NCBI using the BLASTP algorithm on the NCBI blast server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The cSox2-predicted amino acid sequence was used as a template to identify homologous mammalian sequences in the PSI-BLAST search in the NCBI protein database. Multiple sequence alignment was conducted using the ClustalW alignment of the Geneious 7.1.7 software for six analogous sequences from the closest mammalian species. The Geneious software displays different residues in different colors. Arabian camel and other mammalian Sox2 amino acid sequences were aligned and phylogenetic trees were constructed using the BLOSUM62 matrix. We used bootstrap resampling, which was repeated 1,000 times in order to measure the reliability of each obtained topological trees.

cSox2 secondary and 3D structure. The secondary structure of cSox2 was predicted using the Geneious 7.1.7 software, while the 3D structure was predicted using both the Swiss-model server[22] and Protean 3D program (Lasergene 12; DNASTAR).[22]

Globular and disordered regions in the cSox2 protein. In order to identify ordered "globular" and disordered regions of the cSox2 protein, we used the GlobPlot 2.3 server[23] at the globplot.embl.de website. The Russell/Linding set was chosen in which the structures of α-helices and β-sheets are assigned as globular regions (GlobDoms), whereas the structures of random coils and turns are as disordered regions (Disorder). This method can predict a novel propensity based on the disorder prediction algorithm.

ANCHOR analysis. In order to predict binding sites within disordered regions of cSox2, the ANCHOR web server[24] at http://anchor.enzim.hu has been used. This method depends on the pairwise energy estimation method developed for the general disorder prediction method and is based on the hypothesis that long-disordered regions contain local potential binding sites. The IUP server presents a novel algorithm for predicting such regions from amino acid sequences by estimating their total pairwise inter-residue interaction energy.

Results

Sequence identity of cSox2. The similarity of the obtained sequence was examined in the GenBank database using the BLASTN algorithm on the NCBI Blast server (http://blast.ncbi.nlm.nih.gov/Blast.cgi). The cSox2 analysis with nucleotide BLAST exhibited its close similarity (99%–94%) with other mammals' Sox2 mRNAs: 99% with alpacas (*Vicugna pacos*), 99% with wild bactrian camels (*Camelus ferus*), 97% with cattle (*Bos taurus*), 97% with humans (*Homo sapiens*), 96% with pigs (*Sus scrofa*), and 94% with the house mice (*Mus musculus*). The full sequence contained 822 nucleotides (Fig. 1) and is considered the first cSox2 mRNA sequence.

Amino acid composition of cSox2. cSox2 mRNA sequence encoded a cSox2 protein of 273 amino acids (Fig. 1). The cSox2 protein analysis conducted by the Protean program (Lasergene 12) showed that it contains 72 charged amino acids (26.37%), 61 hydrophobic amino acids (20.49%), 18 acidic amino acids (6.59%), 29 basic amino acids (10.62%), and 84 polar amino acids (30.77%). Moreover, the distribution of hydrophilic and hydrophobic amino acids of cSox2 using the Eisenberg's method[25] was used (Supplementary Fig. 2), which shows that the cSox2 protein has more hydrophobic amino acids at its N-terminal tail than at its C-terminal tail. The expected isoelectric point (pI) was found to be 10.11. The N-terminal of the sequence was considered M (Met). The chemical composition of the predicted cSox2 protein is illustrated in Table 2 and Figure 1. As shown in Table 2, the cSox2 protein is rich in Serine (S) residue, which constitutes 12.5%

Table 2. PROTEAN analysis of the expected chemical composition of the cSox2 protein.

AMINO ACID	NUMBER COUNT	% BY FREQUENCY
A (Ala)	22	8.1
C (Cys)	1	0.4
D (Asp)	8	2.9
E (Glu)	10	3.7
F (Phe)	2	0.7
G (Gly)	27	9.9
H (His)	11	4.0
I (Ile)	4	1.5
K (Lys)	14	5.1
L (Leu)	19	7.0
M (Met)	22	8.1
N (Asn)	11	4.0
P (Pro)	21	7.7
Q (Gln)	15	5.5
R (Arg)	15	5.5
S (Ser)	34	12.5
T (Thr)	10	3.7
V (Val)	11	4.0
W (Trp)	3	1.1
Y (Tyr)	13	4.8
Negatively charged amino acids	18	–
Positively charged amino acids	29	–

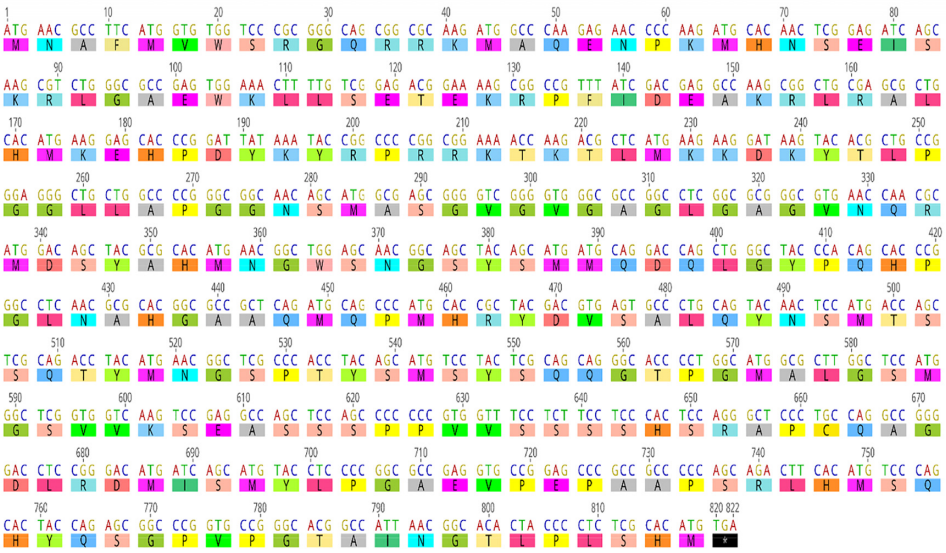

Figure 1. Complete nucleotide sequence encoding cSox2 and its predicted amino acids.
Note: *Termination codon.

of the total amino acids. It has been shown that this amino acid is rich in transcription factors.[26]

Multiple sequence alignment and phylogenetic analysis. When the predicted sequence of amino acids of cSox2 was contrasted with the top similar sequences from six species using the BLASTP algorithm on the NCBI Blast server (http://blast.ncbi.nlm.nih.gov/Blast.cgi), the relative percentage identities were ranging from 99% to 93%: 99% for *Homo sapiens, Mus musculus, Bos taurus, and Vicugna pacos;* 98% for *Sus scrofa*; and 93% for *Camelus ferus* (Table 3). The multiple alignments of amino acid sequences used for this analysis are presented in Figure 2.

The amino acid sequence of cSox2 was aligned with that of six mammalian species by ClustalW alignment

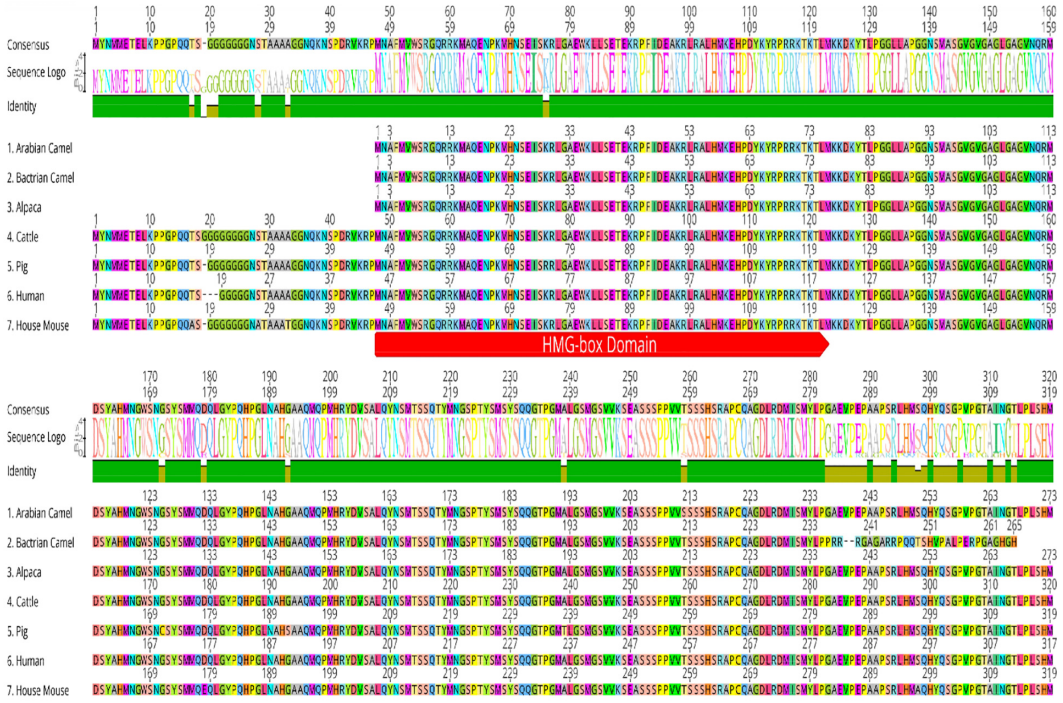

Figure 2. Amino acid sequence alignment of the cSox2 protein with Sox2 proteins of six species. The alignment was generated with the Geneious 7.1.7 multiple sequence alignment software. Residues were color coded according to their conservancy. Red line shows the HMG-box domain.

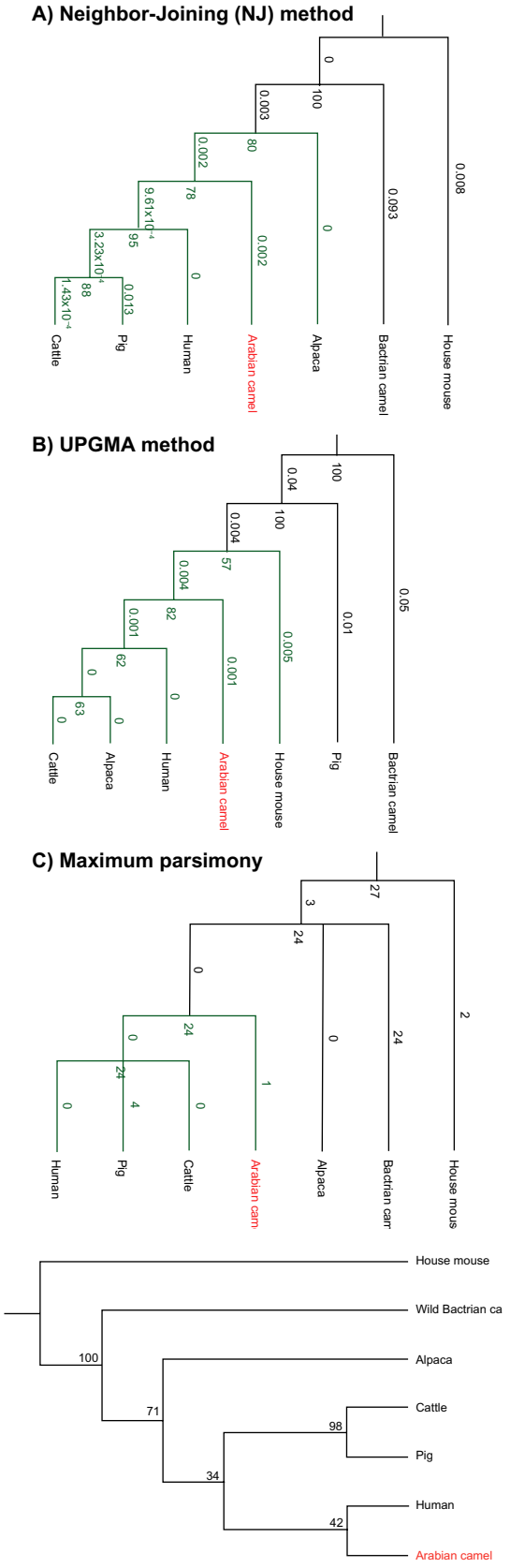

Figure 3. The rooted phylogenetic trees of cSox2 and other six species. (**A**) An additive tree constructed using the Neighbor-Joining method. (**B**) An ultrametric tree using the UPGMA method calculated for seven taxa under study. (**C**) Using the Maximum Parsimony method.
Note: The numbers next to each node represent a measure of support for the node (confidence level for obtained phylogenetic tree).

using Geneious 7.1.7 software.[21] In general, the amino acid alignment of the cSox2 and six mammalian species has shown that the *N*-terminus is more conserved than the *C*-terminus. A conserved sequence of about 75 residues (red line) revealed the HMG-box domain (Fig. 2).

Predict globular and disordered regions in the cSox2 protein. In order to identify the ordered "globular" and disordered regions of the cSox2 protein, we used the GlobPlot 2.3 server[23] at the globplot.embl.de website. The Russell/Linding set was chosen in which the structures of α-helices and β-sheets are assigned as globular regions (GlobDoms), whereas the structures of random coils and turns are as disordered regions (Disorder). This method can predict a novel propensity based on the disorder prediction algorithm. Residue ranges for disordered regions (blue) and globular regions (green) are shown at the bottom of Figure 5.

Predict ordered and disordered region within cSox2. We also used four different predictors in order to determine ordered (structured) and disordered (unstructured) regions within the cSox2 protein (Fig. 6). These predictors are VLXT,[27] VL3,[28] VSL2B,[28] and P-FIT.[29] They use amino acids sequence as inputs and give a structured order or disorder as outputs.

ANCHOR analysis. In order to search potential binding sites within disordered regions of the cSox2 protein, we used the ANCHOR algorithm available at http://anchor.enzim.hu.[30]

cSox2 secondary and 3D structure modeling contrasted with human Sox2. The Geneious 7.1.7 software generated a prediction of the secondary structure of cSox2 that was contrasted with hSox2 (Fig. 8). The predicted structure proposed that this protein holds close similarity to its human counterpart. The predicted structure also suggested that the cSox2 protein is composed of 6 α-helices, 18 β-sheets, 20 coils (black), and 23 turns (brown).

Discussion

In this study, we provided the first report on the full-length cDNA and deduced the protein sequence of the Sox2 gene from *C. dromedaries* (Fig. 1). The open reading frame is composed of 822 nucleotides, which are similar to those from other mammalian species. The predicted amino acid sequence of the open reading frame deduced a protein of 273 residues with the molecular weight of 29.825 kDa. The homologous comparison between the cSox2 protein with other mammalian species was greater than 90% (Table 3).

The phylogenetic trees for the predicted amino acid sequence of whole cSox2 and six of the highly similar mammalian Sox2 were constructed using three different methods (Fig. 3). All the three methods confirmed that the common ancestor of the Arabian camel has a further evolutionary distance from the root than the bactrian camel and alpaca. Figure 3A shows that pig and cattle diverged from their ancestor at a later time than the Arabian camel, whereas alpacas and cattle are found to have the closest relationship

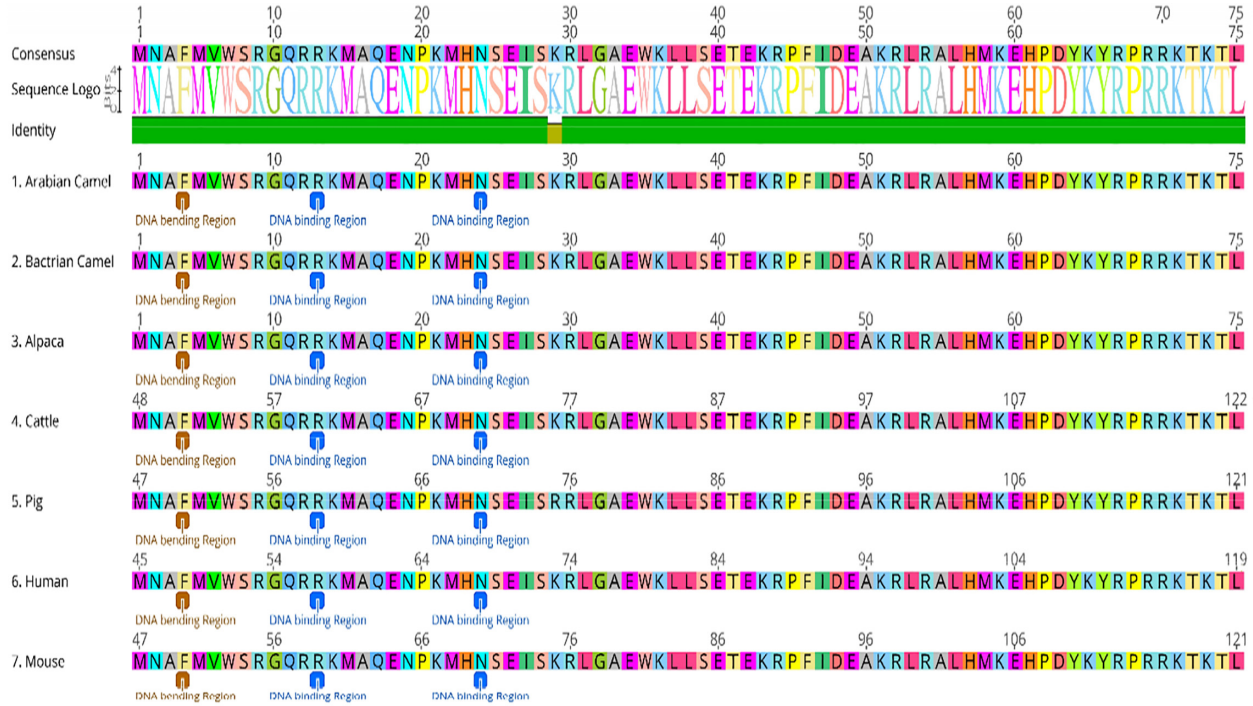

Figure 4. Showing the multiple sequence alignment of the cSox2 HMG-box domain with those from other species.

to each other using the UPGMA method (Fig. 3B). Our results also revealed that cSox2 is clustered with Sox2 from humans, alpacas, pigs, and cattle. Bactrian camel and mouse were segregated in the early evolution. Figure 3C show that pigs, cattle, and humans constitute the multifurcating internal node. Arabian camel, alpacas, cattle, pigs, and humans were more closely related to each other than they are to the other two taxa. In Figure 3, the branches represent evolutionary lineages changing over time. The ancestors of Arabian camels, alpacas, and humans existed prior to the ancestors of cattle and pigs, and time is approximately flowing from up to down. For both NJ and UPGMA methods, all internal nodes are boostrapply supported by more than 50% and the Jukes-Cantor model was applied.

We confirmed that cSox2 predicted protein has the HMG-box domain, which contains highly conserved DNA contact amino acids. For example, in the cSox2 HMG-box domain, Arg^{13} (R^{13}) and Asn^{22} (N^{22}) form hydrogen bonding with DNA (Fig. 4). The degree of sequence identity of the HMG-box domain of cSox2 to other species under study was 100% except for pigs. In addition, the cSox2 HMG-box domain contains a nonpolar DNA intercalating Phe^{4}/F^{4} residue at its N-terminus, which is responsible for DNA bending.[31] The ability of Sox2 HMG-box domain to bend DNA is required for its function as a transcription factor.

Protein structure and function regions are often divided into two sub-regions. The first contains the globular regions (ordered domains) such as HMG-box domains. The second consists of non-globular regions (disordered domains) such as SH3 ligands. An initial step toward developing such a structural protein is to optimize the target selection by identifying its domains and consequently increasing the spanning of the protein folds and its structure space. However, it has been reported that many functional protein segments are localized outside the globular domains in regions that are intrinsically disordered/unfolded[32] Regular and

Table 3. Comparison of cSox2 with other Sox2 proteins from various, mostly similar, mammals.

SPECIES	(Ref. Seq)	NUMBER OF AMINO ACIDS	COVERAGE (%)	E-VALUE	IDENTITY
Camelus ferus (Bactrian Camel)	EPY80590	265	90%	$9e^{-170}$	93%
Vicugna pacos (Alpaca)	XP 006201129	273	100%	0.0	99%
Bos Taurus (Cattle)	NP 001098933	320	100%	0.0	99%
Sus scrofa (Pig)	NP 001116669	319	100%	0.0	98%
Homo sapiens (Human)	NP 003097	317	100%	0.0	99%
Mus musculus (Mouse)	NP 035573	319	100%	0.0	99%

Note: The comparison included number of amino acid residues, percent identity and E-value.

irregular secondary structures of proteins play an important role in predicting functional sites of proteins as well as their 3D structures. Proteins that have regular secondary structures are often described as globular proteins (ordered proteins), whereas those that lack regular secondary structures and high degree of flexibility are classified as disordered proteins (unstructured). However, a number of reports (more than 100) of intrinsically unstructured/disordered proteins have indicated that such regions may contain functional sites (also known as linear motifs)[33] or they may become ordered under specific conditions when they bind to another molecule.[34,35] As it is shown in Figure 5, N-terminal part of the cSox2 protein is predicted to be ordered, whereas its C-terminal part is predicted to be disordered. The cSox2 protein contains more disordered regions than the globular one and there are six predicted disordered regions within the cSox2 protein: 84–109, 118–147, 170–197, 203–223, 239–246, and 253–272 and there is only one predicted globular domain: 2–83.

Furthermore, we used the JRONN method[36,37] that is based on the Regional Order Neural Network (RONN) analysis method using the Protean program in order to predict disordered regions of cSox2 (Supplementary Fig. 3). As a result, cSox2 can be classified as an intrinsically disordered protein. It has been suggested that all the pluripotent-stem-cell–inducing proteins reveal high amounts of disordered regions, indicating that there is an intrinsic requirement for these transcription factors to be highly flexible and thus to be able to interact with other proteins and DNA.[38]

In addition, analysis of the intrinsic disorder is also important in the case of cSoxc2, as about 80% of cancer-associated proteins predicted to have large regions of disorder.[39] In order to predict the antigenicity of the cSox2 protein, we utilized the Jameson–Wolf method[40] using the Protean program. This approach compares the percentage of the amino acids in the average composition of the cSox2 protein to the percentage of each amino acid present in known antigenic determinants. More than nine regions in cSox2 were predicted as antigenic regions, from which six regions show higher antigenicity values >1.2 located at the termini of the cSox2 protein (Supplementary Fig. 4).

In Figure 6, all amino acids/regions with disorder disposition higher than 0.5 score are predicted to be disordered. We used these four meta-predictors because they used different predictive approaches and emphasized different features of the sequence. In general, the graph revealed that of the 273 amino acids of the cSox2 protein, more than 90% were in the disordered regions (above the threshold of 0.5). As shown in Figure 6, the VLXT predictor (Grey), whose accuracy reached 70% and integrated three different predictors, clearly showed that six regions within the cSox2 protein, 1–10, 20–30, 90–95, 110–130, 210–240, and 260–273, had a higher tendency of being structured and flanked by disordered regions. The accuracy of this predictor was low to predict short regions (<10 amino acids) of disorder. However, this predictor had significant advantages in finding potential binding sites in proteins.[41]

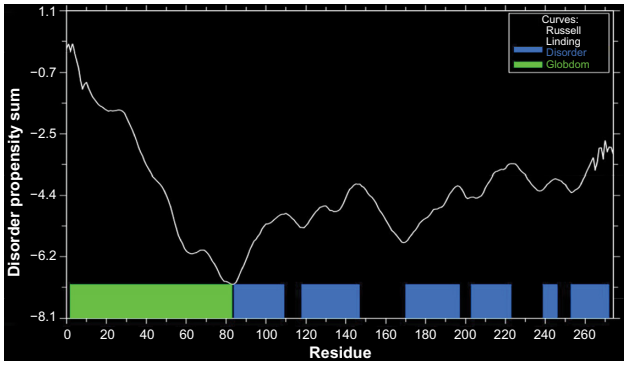

Figure 5. Glob Plot analysis. Blue boxes are disordered regions and green boxes are ordered regions in the cSox2 protein.

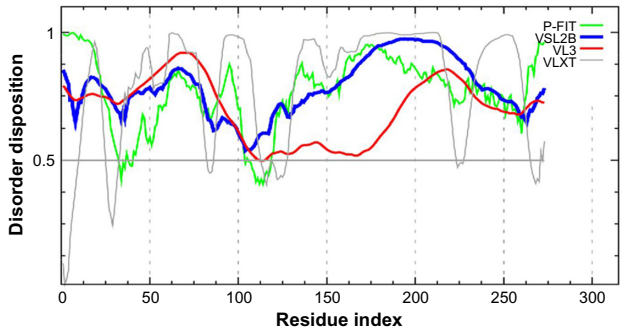

Figure 6. Disorder predictions for 273 amino acids of cSox2.
Notes: The green line is disorder prediction from P-FIT; blue line is prediction of VSL2B; the red line is prediction of VL3; and the grey line is the prediction of VLXT.

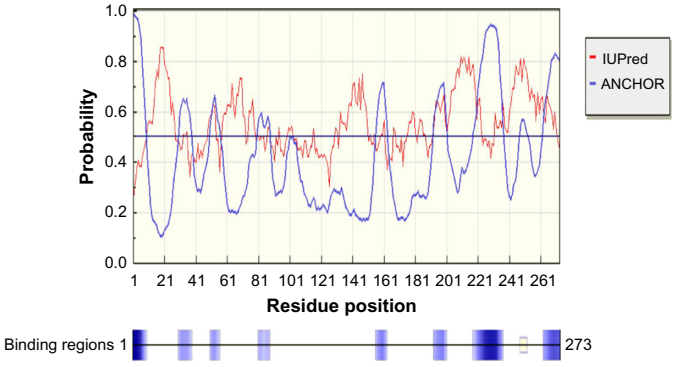

Figure 7. ANCHOR plot. Prediction of protein-binding regions in the disordered cSox2 protein. Blue boxes are binding regions.

The VL3 predictor had higher accuracy in predicting longer unstructured regions[41] The predictor (red) predicted that the whole cSox2 protein to be mostly disordered protein. Similarly, the VSL2B predictor (blue) used the result of sequence alignments from PSI-blast and secondary structure prediction from PHD and PSI-Pred; therefore, it was the most accurate predictor. Both VL3 and VSL2B predicted the cSox2 protein to be disordered.

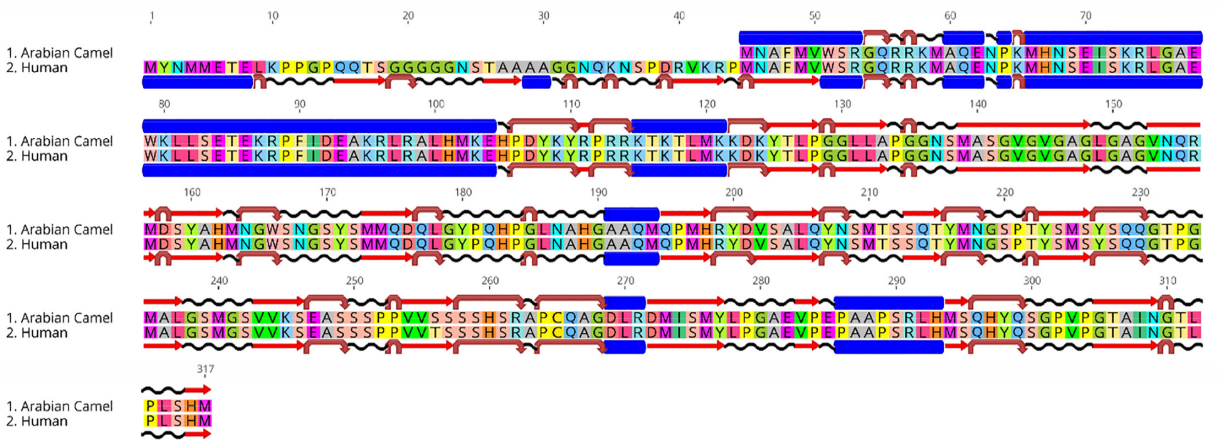

Figure 8. The predicted secondary structure sites of the cSox2 and hSox2 protein sequences. Red arrows and blue cylinders represent β-sheet and α-helix respectively.

The P-FIT predictor, also known as the meta-predictor, is a combination of several individual predictors. This predictor used a collection of results from many individual predictors as its input. In Figure 6, the general trend of the P-FIT predictor (green) was very similar to the VLXT predictor except at N- and C-terminal regions in which the VLXT predictor showed a stronger effect of termini. Most differences between P-FIT and VLXT predictors were 7 sharp dips found by the VLXT predictor near AA3 (A), AA31 (L), AA98 (A), AA122 (W), AA134 (L), AA230 (I), and AA263 (A). These dips usually indicate the molecular recognition feature (MoRF) region within the protein, which in general has a much higher content of aliphatic and aromatic amino acids than disordered regions.[42]

To summarize the results in Figure 6 and Supplementary Figure 5, all predictors accurately predicted the N- terminal of the cSox2 protein to be ordered. Moreover, they strongly predicted the region of the amino acids from 100 to 125 to be an ordered region. This region has been predicted as an α-helix structure using Protean 3D (Fig. 9D).

Figure 7 shows eight potential binding sites within cSox2 proteins (blue boxes): 1–10, 30–40, 50–60, 80–90, 150–160, 190–200, 220–240, and 265–273. The VLXT predictor and ANCHOR-indicated binding sites are often completely or partially overlapping each other. Four of the eight potential binding sites from the ANCHOR predictor overlapped with the VLXT predictor (Figs. 6 and 7).

Another server used for predicting intrinsically unstructured/disordered region of cSox2 is the IUP server (red) Figure 7 (red line). Regions above 0.5 thresholds were predicted to be disordered. Based on the assumption by IUP, the sequences do not fold due to their inability to form sufficient stabilizing inter-residue interactions, which classified the cSox2 protein as unstable protein. Moreover, we examined the stability and instability of cSox2 using the approach of Guruprasad,[43] which revealed that most of the cSox2 protein consisted of unstable regions (Supplementary Fig. 6).

Proteins with highly similar residues have a higher tendency to form similar 3D structures. As an experimental structure of cSox2 is unavailable, aligning to known structures is required. The crystal structure of the HMG-box domain of human stem cell transcription factor Sox2 (PDB: 2LE4) with 81 residues was the best match with the 3D structure of cSox2 with 75 residues using the Swiss model server homology structure modeling (Fig. 9A, 9B). An obtained HMG-box domain of the cSox2 protein consisted of 75 residues and had a characteristic L-shaped fold consisting of three α-helices with an angle of ~80° between the arms. The long arm contained the C-terminal strand and helix III, while the short arm with the N-terminal strand was composed of helices I and II. The length of the loop between helices I and II was longer than that between II and III. Hydrophilic and hydrophobic amino acids were present within the loop between helices I and II. The presence of N-terminal and/or C-terminal tails composed of disordered strands of basic and/or acidic residues was suggested to enhance DNA binding and bending[44] The similarities between the HMG-box domain of cSox2 and hSox2 were studied by superimposing their structures using Protean 3D (Fig. 9C). The overall root mean square deviation (RMSD) between the cSox2 protein and cSox2 protein structures was 0.047.

The main secondary structure elements were HMG-box domains of both cSox2 and hSox2 encompassing residues 1 to 75 and 45 to 119, respectively. We generated the de novo 3D model of cSox2 (273 residues) in the Arabian camel. The 3D-predicted structure of the cSox2 protein correctly predicted the HMG-box domain (Fig. 9D). The C-score, TM-score, and RMSD score, which measure the quality of the predicted modeling structure of cSox2, were −4.59, ±0.07, and 17.57 respectively.

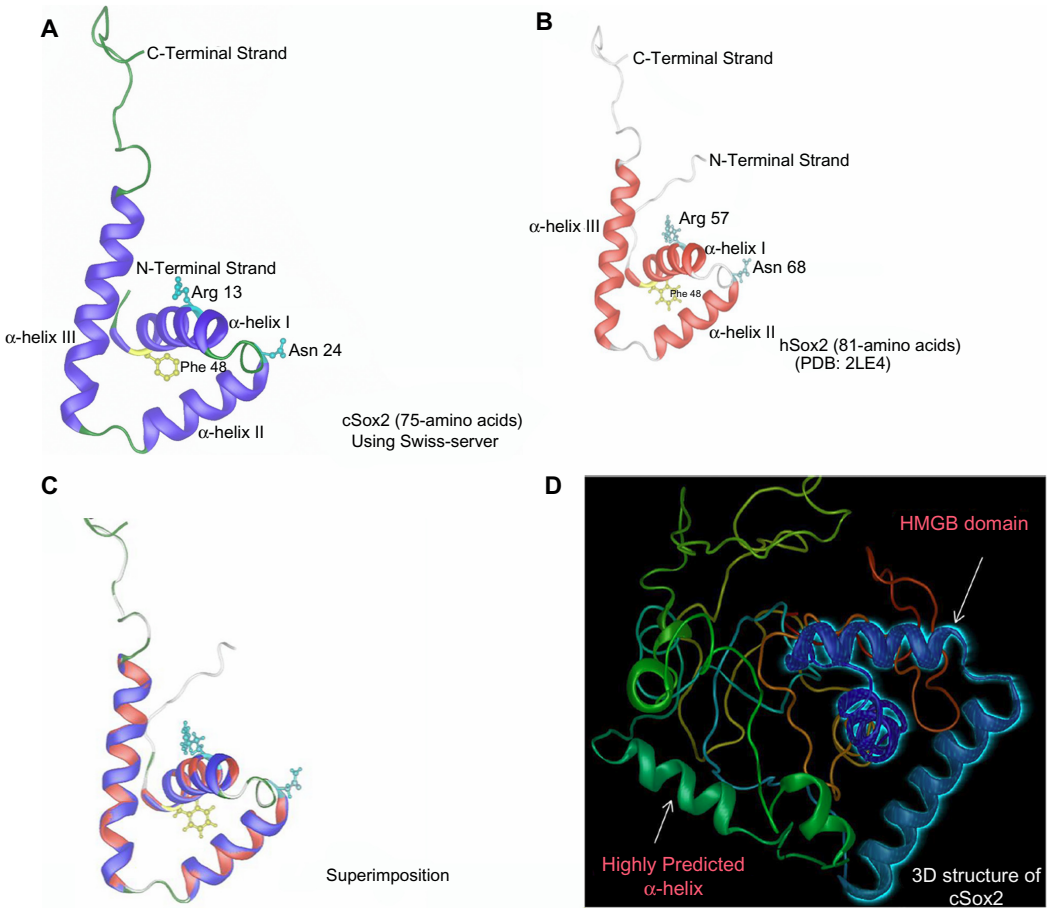

Figure 9. 3D structure of the HMG-box domain. (**A**) Predicted for cSox2. (**B**) Experimental for hSox2. (**C**) Stereo ribbon representation of the predicted 3D structures of the HMG-box domain of cSox2 (blue) and the superimposition with hSox2 (red). (**D**) predicted 3D structure model for cSox2 indicates the HMG-box domain.

To conclude, we sequenced and matched the Sox2 mRNA of the Arabian camel with other mammalians' corresponding coding sequences. This study also generated its 3D model that is very critical for annotation and eventually might contribute to iPSCs research.

Author Contributions

Generated and analyzed the data: AOA, SA. Wrote the first draft of the manuscript: SA, MH. Contributed to the writing of the manuscript: AOA, SA, OA, FA, MA, ZA, AA, HA, IA, MH. Agreed with manuscript results and conclusions: All authors. Jointly developed the structure and arguments for the paper: SA, MH. Made critical revisions and approved final version: SA, MH. All authors reviewed and approved the final manuscript.

Supplementary Material

Supplementary Figure 1. Agarose gel electrophoresis.

Supplementary Figure 2. The Hydrophobicity of cSOX2 protein.

Supplementary Figure 3. The Disorder (JRONN) method.

Supplementary Figure 4. Antigenicity of cSOX2.

Supplementary Figure 5. Output from the DisEMBL web server.

Supplementary Figure 6. The stability and instability of cSOX2 protein.

REFERENCES

1. Chlon TM, Dore LC, Crispino JD. Cofactor-mediated restriction of GATA-1 chromatin occupancy coordinates lineage-specific gene expression. *Mol Cell*. 2012;47:608–21.
2. Tsang AP, Visvader JE, Turner CA, et al. FOG, a multitype zinc finger protein, acts as a cofactor for transcription factor GATA-1 in erythroid and megakaryocytic differentiation. *Cell*. 1997;90:109–19.
3. Soullier S, Jay P, Poulat F, Vanacker JM, Berta P, Laudet V. Diversification pattern of the HMG and SOX family members during evolution. *J Mol Evol*. 1999;48:517–27.
4. Štros M, Launholt D, Grasser KD. The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. *Cell Mol Life Sci*. 2007;64:2590–606.
5. Gubbay J, Collignon J, Koopman P, et al. A gene mapping to the sex-determining region of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. *Nature*. 1990;346:245–50.
6. Sinclair AH, Berta P, Palmer MS, et al. A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif. *Nature*. 1990;346:240–4.
7. Wegner M. All purpose sox: the many roles of sox proteins in gene expression. *Int J Biochem Cell Biol*. 2010;42:381–390.
8. Tanaka S, Kamachi Y, Tanouchi A, Hamada H, Jing N, Kondoh H. Interplay of SOX and POU factors in regulation of the nestin gene in neural primordial cells. *Mol Cell Biol*. 2004;24:8834–46.
9. Simara P, Motl JA, Kaufman DS. Pluripotent stem cells and gene therapy. *Translational Res*. 2013;161:284–92.
10. Brazel CY, Limke TL, Osborne JK, et al. Sox2 expression defines a heterogeneous population of neurosphere-forming cells in the adult murine brain. *Aging Cell*. 2005;4:197–207.

11. Ferri AL, Cavallaro M, Braida D, et al. Sox2 deficiency causes neurodegeneration and impaired neurogenesis in the adult mouse brain. *Development*. 2004;131:3805–19.
12. Liao J, Cui C, Chen S, et al. Generation of induced pluripotent stem cell lines from adult rat cells. *Cell Stem Cell*. 2009;4:11–5.
13. Liu H, Zhu F, Yong J, et al. Generation of induced pluripotent stem cells from adult rhesus monkey fibroblasts. *Cell Stem Cell*. 2008;3:587–90.
14. Esteban MA, Xu J, Yang J, et al. Generation of induced pluripotent stem cell lines from Tibetan miniature pig. *J Biol Chem*. 2009;284:17634–40.
15. Wu Y, Zhang Y, Mishra A, Tardif SD, Hornsby PJ. Generation of induced pluripotent stem cells from newborn marmoset skin fibroblasts. *Stem Cell Res*. 2010;4:180–8.
16. Luo J, Suhr ST, Chang EA, et al. Generation of leukemia inhibitory factor and basic fibroblast growth factor-dependent induced pluripotent stem cells from canine adult somatic cells. *Stem Cells Dev*. 2011;20:1669–78.
17. Takahashi K, Tanabe K, Ohnuki M, et al. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell*. 2007;131:861–72.
18. Al-Swailem AM, Shehata MM, Abu-Duhier FM, et al. Sequencing, analysis, and annotation of expressed sequence tags for Camelus dromedarius. *PLoS One*. 2010;5:e10720.
19. Wu H, Guang X, Al-Fageeh MB, et al. Camelid genomes reveal evolution and adaptation to desert environments. *Nat Commun*. 2014;5:5188.
20. Bactrian Camels Genome Sequencing and Analysis Consortium, Jirimutu, Wang Z, et al. Genome sequences of wild and domestic bactrian camels. *Nat Commun*. 2012;3:1202.
21. Kearse M, Moir R, Wilson A, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*. 2012;28:1647–9.
22. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics*. 2006;22:195–201.
23. Linding R, Russell RB, Neduva V, Gibson TJ. GlobPlot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res*. 2003;31:3701–8.
24. Dosztányi Z, Mészáros B, Simon I. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*. 2009;25:2745–6.
25. Eisenberg D, Weiss RM, Terwilliger TC. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc Natl Acad Sci U S A*. 1984;81:140–4.
26. Nakamura T, Alder H, Gu Y, et al. Genes on chromosomes 4, 9, and 19 involved in 11q23 abnormalities in acute leukemia share sequence homology and/or common motifs. *Proc Natl Acad Sci USA*. 1993;90:4631–5.
27. Dunker AK, Lawson JD, Brown CJ, et al. Intrinsically disordered protein. *J Mol Graph Model*. 2001;19:26–59.
28. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics*. 2006;7:208.
29. Xue B, Dunbrack RL, Williams RW, Dunker AK, Uversky VN. PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim et Biophys Acta*. 2010;1804:996–1010.
30. Mészáros B, Simon I, Dosztányi Z. Prediction of protein binding regions in disordered proteins. *PLoS Comput Biol*. 2009;5:e1000376.
31. Štros M. HMGB proteins: interactions with DNA and chromatin. *Biochim et Biophys Acta*. 2010;1799:101–13.
32. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol*. 1999;293:321–31.
33. Fukuchi S, Sakamoto S, Nobe Y, et al. IDEAL: intrinsically disordered proteins with extensive annotations and literature. *Nucleic Acids Res*. 2012;40:D507–11.
34. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*. 2002;11:739–56.
35. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. *Biochemistry*. 2002;41:6573–82.
36. Yang ZR, Thomson R, McNeil P, Esnouf RM. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics*. 2005;21:3369–76.
37. Troshin PV, Procter JB, Barton GJ. Java bioinformatics analysis web services for multiple sequence alignment—JABAWS:MSA. *Bioinformatics*. 2011;27:2001–2.
38. Xue B, Oldfield CJ, Van Y-Y, Dunker AK, Uversky VN. Protein intrinsic disorder and induced pluripotent stem cells. *Mol Bio Syst*. 2012;8:134–50.
39. Iakoucheva LM, Brown CJ, Lawson JD, Obradović Z, Dunker AK. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J Mol Biol*. 2002;323:573–84.
40. Jameson BA, Wolf H. The antigenic index: a novel algorithm for predicting antigenic determinants. *Comput Appl Biosci*. 1988;4:181–6.
41. Cheng Y, et al. Mining α-helix-forming molecular recognition features with cross species sequence alignments. *Biochemistry*. 2007;46:13468–77.
42. Mohan A, et al. Analysis of molecular recognition features (MoRFs). *J Mol Biol*. 2006;362:1043–59.
43. Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence. *Protein Eng*. 1990;4:155–61.
44. Malarkey CS, Bestwick M, Kuhlwilm JE, Shadel GS, Churchill MEA. Transcriptional activation by mitochondrial transcription factor A involves preferential distortion of promoter DNA. *Nucleic Acids Res*. 2011;40(2):614–24.