Seminars in Cell & Developmental Biology 63 (2017) 2–12

Contents lists available at ScienceDirect

Seminars in Cell & Developmental Biology

journal homepage: www.elsevier.com/locate/semcdb

Review

Molecular basis for the genome engagement by Sox proteins

Linlin Hou[a,b,c], Yogesh Srivastava[a,b,c], Ralf Jauch[a,b,c,*]

[a] Genome Regulation Laboratory, Drug Discovery Pipeline, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China
[b] Key Laboratory of Regenerative Biology, South China Institute for Stem Cell Biology and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China
[c] Guangdong Provincial Key Laboratory of Stem Cell and Regenerative Medicine, South China Institute for Stem Cell Biology and Regenerative Medicine, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China

ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form 9 August 2016
Accepted 9 August 2016
Available online 10 August 2016

Keywords:
Sox
Transcription factors
Gene regulation
DNA binding
Cellular reprogramming

ABSTRACT

The Sox transcription factor family consists of 20 members in the human genome. Many of them are key determinants of cellular identities and possess the capacity to reprogram cell fates by pioneering the epigenetic remodeling of the genome. This activity is intimately tied to their ability to specifically bind and bend DNA alone or with other proteins. Here we discuss our current knowledge on how Sox transcription factors such as Sox2, Sox17, Sox18 and Sox9 'read' the genome to find and regulate their target genes and highlight the roles of partner factors including Pax6, Nanog, Oct4 and Brn2. We integrate insights from structural and biochemical studies as well as high-throughput assays to probe DNA specificity in vitro as well as in cells and tissues.

Contents

* Corresponding author at: Genome Regulation Laboratory, Drug Discovery Pipeline, Guangzhou Institutes of Biomedicine and Health, Chinese Academy of Sciences, Guangzhou 510530, China.
E-mail address: ralf@gibh.ac.cn (R. Jauch).

http://dx.doi.org/10.1016/j.semcdb.2016.08.005

1. Introduction

Transcription factor (TF) proteins determine cellular identities and direct embryonic development by selectively binding to genomic DNA to orchestrate gene expression programs. Amongst the ~21,000 human protein-coding genes about 1600 encode for sequence-specific DNA binding TF proteins [1,2]. Most TFs belong to gene families comprising a handful up to several hundred members in mammalian genomes. Members of such gene families are termed paralogs and evolved by the expansion of ancestral genes through gene or genome duplications [3]. The gene encoding for the TF Sry (sex-determining region Y) was discovered following an intense search for the testis-determining factor on the Y-chromosome [4,5]. The sequence conservation between mouse and human Sry genes is restricted to a region of 79 amino acids. This sequence motif encodes for a special version of the high-mobility group (HMG) box also found in a class of ubiquitous and highly abundant non-histone DNA binding proteins [6]. In the original Sry study, four more homologous genes were isolated from autosomal loci of mouse 8.5 d.p.c. (days post coitum) cDNA libraries corresponding to Sox1-4 [4,7]. The unifying feature of these genes is the Sry-like HMG box; hence this gene family was termed Sox. Additional members were subsequently detected and cloned in a wide array of tissues taking advantage of sequence signatures within the HMG box [7–10]. With the availability of whole genome sequences it became clear that the mouse and human genomes each encodes for 20 Sox genes [11]. Based on the sequence identity of the HMG box, the Sox genes are classified into 8 groups denoted SoxA to SoxH with 1–3 members each [12].

The Sox TFs were soon found to constitute essential molecules with key roles during virtually all phases of embryonic development and the fate determination of many cell types as summarized in a number of excellent reviews [13–23]. The prominence of the gene family received a further elevation when one of its members, Sox2, was found to be a core component of TF cocktails with the ability of converting mouse and human somatic cells to induced pluripotent stem cells (iPSCs) [24–26]. Most Sox TFs are highly pleiotropic as they bind and regulate different gene sets in different cellular contexts. Sox2, for example, acts in a staggeringly diverse array of cell and tissue types including pluripotent stem cells, neural lineages, lung tissue, the eye and the ear [27]. Yet, what endows Sox proteins with this versatility and developmental plasticity largely remains elusive. Moreover, Sox proteins are reported to function as 'pioneer' factors. That is, they are able to bind compact transcriptionally silent chromatin and to recruit non-pioneer TFs to drive cell fate conversions. In this review, we discuss recent progress in the understanding of the biochemical basis for DNA and chromatin recognition by Sox proteins, mechanisms for the partnership of Sox proteins with other TFs and mechanisms for their pioneering activity.

2. Principles of DNA recognition

2.1. Structural basis

Evidence that the Sox HMG box enables DNA binding was first provided for the SRY protein. Binding was found to be sequence specific with a preference for a $C_1T_2T_3T_4G_5T_6$ C_7-like motif [28,29]. This core-motif was later verified to be the preferred binding sequence for all 20 Sox proteins although there can be subtle variations especially in the flanks of the element and a substantial degeneracy is tolerated [30,31]. As several sex-reversing mutations of SRY profoundly reduced the affinity for DNA, it was immediately clear that DNA binding is of critical functional importance [28]. The first structural view on the DNA recognition by Sox TFs was provided by the group of Marius Clore with the nuclear magnetic resonance (NMR) structure of the HMG box of human SRY bound to a $G_1T_2T_3T_4G_5T_6G_7C$ dsDNA in 1995 [32]. In the same year, the NMR structure of the HMG box of the related Lef-1 protein bound to $CACCC_1T_2T_3T_4A_5A_6GCTC$ was reported [33]. More recently, crystal structures of Sox2 [34], Sox17 [35], Sox4 [36] and Sox18 [37] bound to cognate DNA elements were published and a Sox9/DNA complex was deposited to the protein data bank (PDB-id 4S2Q). Altogether, these studies provided valuable insights into the molecular basis for the DNA recognition by Sox proteins. The HMG box folds into an L-shaped 'boomerang' structure constructed of three alpha helices and extended N- and C-terminal tails with an irregular strand-like configuration (Fig. 1A, B). Two hydrophobic clusters stabilize the fold, which include a conserved set of aromatic amino acids such as Phe10, Trp13 and Trp41 and Phe52 (HMG numbering according to reference [12] used throughout this manuscript). The short and long arms of the 'L' have also been denoted as major and minor 'wings' [38–40]. The shorter major wing encompasses the bulk of the amino acids and is composed of helices 1, 2 and the N-terminal turn of helix 3. The minor wing consists of the remainder of helix 3 and the extended N-terminus, which packs against helix 3. Contrary to most other TFs that bind to the major groove of the DNA, the Sox HMG binds to the minor groove of the DNA and its binding induces an overall bend of 60–70° (Fig. 1C). All base pairs of the CATTGT core motif are directly contacted by amino acids via base-specific interactions (Fig. 1D, E). Several of the contact residues emanate from the $R_5PMNAF_{10}MVW$ Sox signature motif at the N-terminus of the HMG box. The $F_{10}M_{11}$ dipeptide constitutes a wedge that intercalates between the central $T_3A_3'T_4A_4'$ base pair forcing the kinking of the DNA. Notably, all residues engaged in base-specific DNA interactions are invariant amongst the 20 Sox TFs (Fig. 1E). Only some residues mediating non-specific interactions with the DNA backbone show conservative replacements such as residues 2 and 15. Overall, the highly positively charged DNA binding surface exhibits a strong evolutionary conservation, whereas interfaces pointing away from the DNA are variable amongst Sox TFs (Fig. 1F, G). Therefore, monomeric forms of all 20 Sox TFs are expected to bind DNA in an identical fashion. Nevertheless, with the availability of a growing number of structures some variations at the Sox/DNA interface have been observed. For example, Arg18 and Asn30 can undergo a concerted conformational switch [36]. Moreover, Arg5, His29 and Tyr72 can structurally re-orient to better accommodate changes in the sequence of the DNA binding element [37]. However, these changes are dictated by the chemical environment provided by the DNA sequence and do not reflect differences inherent to individual Sox TFs. Thus, other mechanisms must account for the multitude of non-redundant and cell-type specific functions of Sox TFs.

2.2. Differences between Sox TFs and other HMG box proteins

Sox TFs belong to the HMG box superfamily of proteins, which is evolutionarily ancient with members present in unicellular eukaryotes such as yeast species. The HMG superfamily can be broadly divided into two groups based on the mechanism of DNA interaction [41,42]. First, there are sequence specific HMG boxes (ssHMGs) including the Sox TFs, the Tcf/Lef TFs and the yeast mating protein MATA. Second, there are the non-sequence specific HMG boxes (nsHMG) including HMG1, HMG2, the SSRP1 subunit of FACT (facilitates chromatin transcription), the mitochondrial TFAM/mtTF1 and UBF1 [41,42]. It appears plausible that sequence specificity evolved after the divergence of ssHMG and nsHMG groups however this question is not ultimately resolved. In an alternative scenario, nsHMGs lost sequence specificity while sequence specificity was present in the ancestral protein.

Lef-1 of the Tcf/Lef family exhibits a similar fold and DNA binding mechanism as Sox proteins but has a number of features distin-