Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Public-data re-examination of SOX6 as a general pre-mRNA splicing factor

Kenji Ohe

oekenji@gmail.com

Fukuoka University https://orcid.org/0000-0002-1958-7572

Research Article

Keywords: SOX6, SRY, SOX9, pre-mRNA splicing, intron retention, reproducibility, public data reanalysis

Posted Date: November 16th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11725154/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Public-data re-examination of SOX6 as a general pre-mRNA splicing factor

Kenji Ohe

Department of Clinical Pharmacotherapy, Faculty of Pharmaceutical Sciences, Fukuoka University, Fukuoka, Japan. ORCID 0000-0002-1958-7572. Correspondence: ohekenji@fukuoka-u.ac.jp

Abstract

In 2002 we reported that SOX6, SRY and the SOX9 HMG domain act in pre-mRNA splicing, and that immunodepletion of SOX6 from HeLa nuclear extract impaired splicing of several substrates (Ohe et al., PNAS 2002). Here the author, a co-first author of that study, tests four predictions of the claim that SOX6 is a general splicing factor using only public data. (1) In mouse epiblast stem cell-derived oligodendrocyte precursor cells (OPCs) and oligodendrocytes (OLs) with CRISPR knockout of Sox6 by two independent guides (GSE181952), editing of Sox6 was confirmed in the RNA-seq reads, yet splicing efficiency over 21,371–29,140 introns did not decrease. The knockout-minus-control difference in pooled splicing efficiency was −0.03% in OPCs (exact permutation $p = 0.661$) and +0.09% in OLs ($p = 0.190$), with no intron reaching FDR < 0.05. The same pipeline applied to pladienolide B-treated mouse hepatocyte organoids (GSE249801) detected a 7.6-fold rise in pooled intron retention and 4,281 introns with increased retention at FDR < 0.05. (2) In DepMap CRISPR screens (26Q3), 9 of 1,361 cell lines depended on SOX6, and its top co-dependencies did not include spliceosome genes. (3) SOX6 and SRY were not detected in any of 62 human RNA-interactome datasets compiled in RBP2GO 2.0, whereas U1-70K was detected in 39. (4) The Human Protein Atlas annotates SOX6 and SOX9 to the nucleoplasm, not to nuclear speckles. SOX6 mRNA in HeLa was 0–2.5 counts per million across four public datasets. These data do not support a general requirement for SOX6 in constitutive splicing in the cell types examined.

Keywords: SOX6; SRY; SOX9; pre-mRNA splicing; intron retention; reproducibility; public data reanalysis

Introduction

SOX proteins share an HMG box that binds and bends DNA, and most of their known functions are transcriptional. In 2002 we proposed a second function in pre-mRNA splicing [1]. That study reported four observations. SOX6 and SRY co-localized with SC-35 in nuclear speckles and redistributed with it when splicing was blocked. Antibodies against SOX6 inhibited both steps of splicing of β-globin and fushi-tarazu pre-mRNAs and

reduced spliceosomal B and C complexes. SOX6-immunodepleted HeLa nuclear extract was defective in the first step of splicing. Recombinant SOX6 HMG domain, SRY, or SOX9 HMG domain restored activity at nanomolar concentrations. From these results we suggested that SOX6 "may function as a general splicing factor" [1].

The 2002 experiments were done in vitro and with polyclonal antibodies. Data that did not exist then now allow the claim to be tested in cells. If SOX6 were a general splicing factor, four predictions follow:

1. Loss of SOX6 in cells that express it should lower splicing efficiency across many introns.
2. Cell lines should depend on SOX6 for growth, as they do on core spliceosomal proteins, and its loss should co-vary with loss of spliceosome genes.
3. SOX6 should be captured in proteome-wide surveys of RNA-bound proteins.
4. SOX6 should be annotated to nuclear speckles in systematic imaging, and should be expressed in HeLa cells, whose nuclear extract was used in 2002.

The author was a co-first author of reference [1]. This report tests the four predictions with public data, without new experiments.

Methods

Datasets

Use	Accession	System	Groups (n)
Sox6 knockout	GSE181952 [2]	Mouse EpiSC-derived OPCs and OLs, CRISPR	OPC: non-targeting 3, sg1 2, sg2 3; OL: non-targeting 3, sg1 3, sg2 3
Positive control	GSE249801 [3]	Mouse hepatocyte organoids, pladienolide B vs DMSO	3 vs 3
SOX6 mRNA in HeLa	GSE299983, GSE325258, GSE348256, GSE242279	HeLa or HeLa S3, control and other arms	3–6 per series

All RNA-seq libraries were paired-end 150 nt, poly(A)-selected and unstranded. Runs SRR15424154–SRR15424170 (GSE181952) and SRR27164224–SRR27164229 (GSE249801) were taken from the European Nucleotide Archive. Run-to-sample assignments were

read from the SRA run tables.

Read subsets and alignment

To fit the available compute (2 CPU cores, 8 GB memory), the first 5,000,000 read pairs of each GSE181952 run and the first 4,000,000 pairs of each GSE249801 run were used. The same depth was used within each dataset. Reads were aligned with HISAT2 2.2.3 [4] with default settings to the GRCm39 primary assembly (RefSeq GCF_000001635.27; 21 chromosomes and chrM). The index was built without splice-site annotation. Overall alignment rates were 95.4–96.3% (GSE181952) and 95.0–96.4% (GSE249801). BAM files were sorted and indexed with samtools 1.25.

Intron set

Introns were taken from RefSeq Select mRNA transcripts in the NCBI GRCm39 annotation (release RS_2024_02), one transcript per gene: 184,520 introns. An intron was kept when it was at least 100 nt long and did not overlap any exon of any annotated transcript on either strand. This leaves 147,430 introns whose reads cannot come from an alternative exon or an overlapping gene, which matters because the libraries are unstranded.

Quantification

Only primary alignments with MAPQ ⩾ 10 were counted. For each intron and sample, J was the number of reads with a gap exactly matching the intron. L and R were the numbers of reads whose aligned blocks covered the first and last intronic nucleotide, that is, reads continuous across the 5′ or 3′ splice site. B was the number of reads overlapping the intron body, defined as the intron minus 10 nt at each end. Counts were made with featureCounts 2.1.1 [5] at read level with multi-overlap allowed. Against samtools bedcov on one sample, L and R were identical (Pearson r = 1.000) and body depth agreed with Spearman ρ = 0.998. From these counts:

$U = (L + R) / 2$; SE = $J / (J + U)$; $b = 150 \times B / \ell$; IR = $b / (b + J)$

SE is splicing efficiency at the splice sites. IR is an intron-retention ratio comparable to that of IRFinder [6], and ℓ is the body length. Pooled SE and IR per sample were computed from the sums of J, U and b over all introns passing the filter. Introns were analysed when J + U ⩾ 20 in every sample of a comparison.

Statistics

Group differences were tested by exact permutation over every relabelling of the samples with the group sizes kept. There are 56 relabellings for n = 3 vs 5, 84 for 3 vs 6, 20 for 3 vs 3 and 10 for 3 vs 2. The smallest attainable two-sided p values are therefore 0.018,

0.012, 0.100 (the complementary labelling gives the same absolute statistic) and 0.100. Per intron, a Welch t test on logit-transformed SE or IR (pseudocount 0.5) was followed by Benjamini–Hochberg correction. The number of introns with $q < 0.05$ was compared with the same count under every relabelling. The pre-specified primary statistic was the median over introns of the per-intron group difference. At this depth most introns had no unspliced reads, so per-intron SE was 1 for most introns and the median difference was 0 in all comparisons. Pooled per-sample SE and IR, and the mean per-intron difference, were therefore added. They were added because of this zero inflation, before the knockout results were inspected for direction, and are reported here as secondary analyses.

Knockout validation

For each Sox6 coding exon (RefSeq Select), reads with MAPQ ⩾ 10 were counted, along with the fraction carrying an insertion, a deletion, or a soft clip of 5 nt or more that started inside the exon. Sox6 coding-region reads per million mapped reads were also computed. Exons are numbered in ascending genomic order. Sox6 is on the minus strand, so exon 1 here is the 3′-most coding exon.

Other resources

DepMap Public 26Q3+Score (Chronos) and DEMETER2 RNAi results for SOX6 were read from the gene page [7]. RNA-binding evidence was taken from RBP2GO 2.0 [8], as Listing Count, the number of proteome-wide RNA-interactome datasets in which a protein was detected. Subcellular localization and cell-line expression were taken from the Human Protein Atlas [9]. For HeLa, gene-level count tables deposited with each GEO series were converted to counts per million (CPM). All resources were accessed on 13 November 2026.

Results

Sox6 is edited in the knockout samples

Both guides left a mark on Sox6 in the RNA-seq reads (Table 1). With sg1, 11–60% of reads over the 7th and 8th coding exons carried an indel or soft clip, against 0–3% in non-targeting controls. With sg2, coverage of the 4th coding exon fell in OPCs, and 12.5–40.5% of reads over that exon were edited in OLs. Sox6 coding-region reads per million were lower in every knockout group.

Table 1. Sox6 editing and expression in GSE181952.

Cell type	Group	n	Sox6 CDS reads per million	Edited reads at the target exon
OPC	Non-targeting	3	39–90	exons 7–8: 0–3%
OPC	sg1	2	34–36	exons 7–8: 11–60%
OPC	sg2	3	26–33	exon 4: 2–10 reads (control 23–37)
OL	Non-targeting	3	157–164	exon 8: 0%; exon 4: 0–2%
OL	sg1	3	16–84	exon 8: 42–53%
OL	sg2	3	77–91	exon 4: 12.5–40.5%

Splicing efficiency does not fall after Sox6 knockout

In OPCs, pooled SE was 0.9942 in controls and 0.9939 after knockout (difference −0.03%, exact permutation p = 0.661). In OLs it was 0.9928 and 0.9936 (+0.09%, p = 0.190). No intron reached q < 0.05 for SE or IR in any knockout comparison (Table 2). The two guides did not move IR in the same direction in OPCs (+18.4% for sg1, −2.8% for sg2). In OLs, both guides lowered IR (−13.0% and −15.9%), the opposite of the direction the 2002 claim predicts. Among non-targeting OPC replicates alone, pooled IR ranged from 0.0057 to 0.0072, a spread of about 25%.

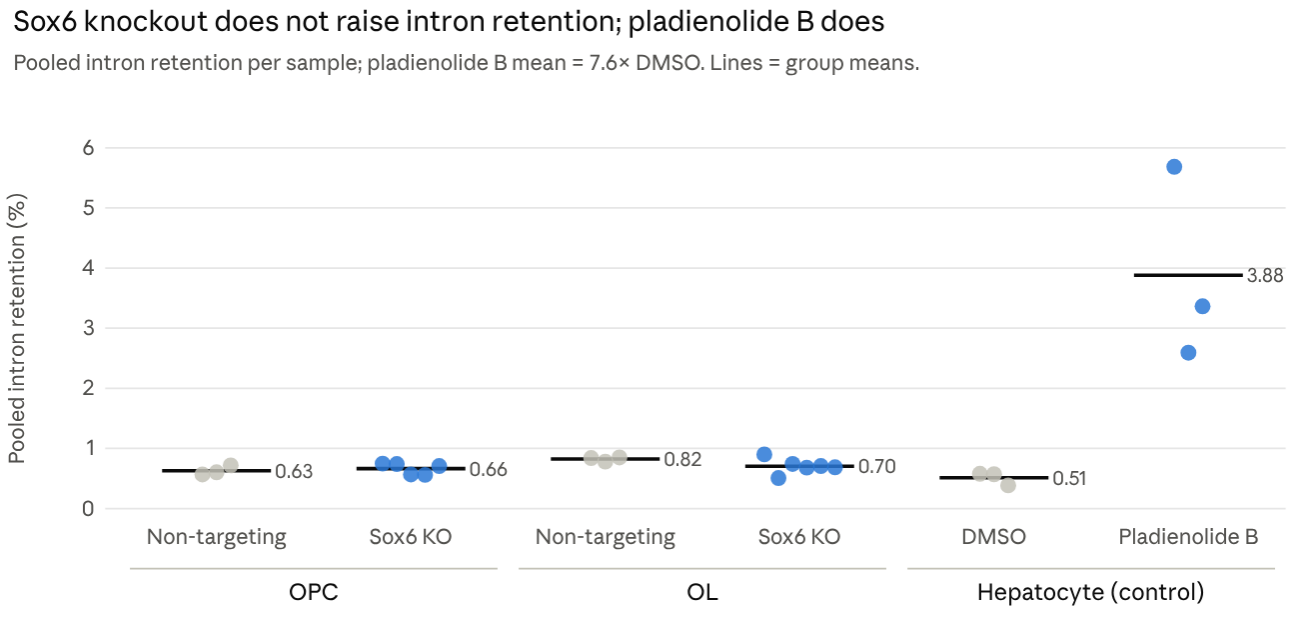
The same pipeline detected the effect of the SF3B inhibitor pladienolide B. Pooled IR rose from 0.0051 to 0.0388 (7.6-fold), pooled SE fell by 2.2%, and 4,281 introns had higher IR at q < 0.05 against 72 lower. Every treated sample lay outside the range of every control sample (Figure 1).

Table 2. Pooled splicing efficiency (SE) and intron retention (IR), control vs perturbation. p = exact permutation test.

Comparison	Introns analysed	Pooled SE, control vs test	ΔSE (relative)	p	Pooled IR, control vs test	ΔIR (relative)	p	Introns with IR q < 0.05 (up/down)
OPC, NTC vs KO (3 vs 5)	29,140	0.9942 vs 0.9939	−0.03%	0.661	0.0063 vs 0.0066	+5.5%	0.768	0/0
OPC, NTC vs sg1 (3 vs 2)	31,012	0.9941 vs 0.9932	−0.09%	0.200	0.0064 vs 0.0075	+18.4%	0.200	0/0

Comparison	Introns analysed	Pooled SE, control vs test	ΔSE (relative)	p	Pooled IR, control vs test	ΔIR (relative)	p	Introns with IR q < 0.05 (up/down)
OPC, NTC vs sg2 (3 vs 3)	30,528	0.9941 vs 0.9943	+0.01%	0.700	0.0064 vs 0.0062	−2.8%	0.600	0/0
OL, NTC vs KO (3 vs 6)	21,371	0.9928 vs 0.9936	+0.09%	0.190	0.0082 vs 0.0070	−14.5%	0.167	0/0
OL, NTC vs sg1 (3 vs 3)	22,654	0.9926 vs 0.9934	+0.08%	0.600	0.0084 vs 0.0073	−13.0%	0.600	0/0
OL, NTC vs sg2 (3 vs 3)	27,136	0.9925 vs 0.9934	+0.09%	0.100	0.0084 vs 0.0071	−15.9%	0.100	0/0
Hepatocyte, DMSO vs pladienolide B (3 vs 3)	9,026	0.9960 vs 0.9738	−2.22%	0.100	0.0051 vs 0.0388	+659.5%	0.100	4,281/72

Figure 1. Pooled intron retention per sample.



GSE181952 (5M read pairs per sample) and GSE249801 (4M pairs per sample); pooled IR over introns passing J + U ⩾ 20 in every sample of each comparison

KO values (sg1 and sg2 together) overlap the non-targeting controls and lie on both sides of them: 0.56–0.75% vs 0.57–0.72% in OPCs and 0.51–0.90% vs 0.78–0.85% in OLs. All three pladienolide B samples (2.6–5.7%) lie above all DMSO samples (0.39–0.58%).

Cell lines rarely depend on SOX6

In DepMap Public 26Q3+Score, 9 of 1,361 cell lines were classed as dependent on SOX6 in CRISPR screens (Chronos), and SOX6 was labelled strongly selective. No line was dependent in RNAi screens (0 of 710, DEMETER2). Gene effects were centred near zero. The five strongest CRISPR co-dependencies were SOX10 (r = 0.22), IL12RB2 (0.21), SRGAP2 (0.21), C1orf167 (0.20) and IL18 (0.19). The five strongest in RNAi were MORC3 (0.36), MLC1 (0.34), ZNF415 (0.33), CLSTN2 (0.31) and FMO1 (0.31). None is a spliceosomal protein. Many cell lines express little SOX6 mRNA, so this test does not separate non-essentiality from non-expression (see Limitations).

SOX6 and SRY are absent from RNA-interactome surveys

RBP2GO 2.0 compiles 62 human and 18 mouse proteome-wide datasets of RNA-bound proteins.

Table 3. RBP2GO Listing Count: datasets in which each protein was detected.

Protein	Human (of 62)	Mouse (of 18)
U1-70K (SNRNP70), positive control	39	12
SOX9	2	1
SOX2	2	not checked
SOX6	0	0
SRY	0	not checked
SOX5	0	not checked
SOX11	0	not checked

SOX6 is nucleoplasmic and nearly absent from HeLa mRNA

The Human Protein Atlas places SOX6 in the nucleoplasm (reliability "supported", antibody HPA001923; A-431, CACO-2 and U2OS cells) and SOX9 in the nucleoplasm ("enhanced", three antibodies, siRNA-validated). Neither is annotated to nuclear speckles. SRY has no localization data. In HPA's A-431 cells the SOX6 antibody stains the nucleoplasm although SOX6 mRNA is 0.0 nTPM there. HPA reports SOX6 mRNA at 0.0 nTPM in HeLa and no SOX6 peptides in HeLa proteomics, against 0.4 nTPM in HEK293 and 2.8 nTPM in K-562.

Across four independent GEO series, SOX6 mRNA in HeLa was low but not always zero (Table 4). It was highest in HeLa S3, the suspension subline widely used for nuclear extracts. In the same libraries, SNRNP70 was 58–278 CPM.

Table 4. SOX6 mRNA in public HeLa RNA-seq (gene-level counts per million).

Series	Cell line	Samples	SOX6 CPM	SNRNP70 CPM
GSE299983	HeLa S3	6	1.6–2.5	249–272
GSE325258	HeLa (TP53 knockout clones)	6	0.0–0.9	165–278
GSE242279	HeLa	6	0.07–0.65	58–98
GSE348256	HeLa	4	0.0–0.05	204–236

Discussion

None of the four predictions of a general splicing role for SOX6 was met in these data. Knockout of Sox6 by two guides, with editing visible in the reads, did not lower splicing efficiency in mouse OPCs or OLs. A pipeline that detected pladienolide B-induced intron retention in every treated sample found no intron with altered retention after Sox6 loss. Cell lines rarely depend on SOX6, and its co-dependencies are not spliceosomal. SOX6 is absent from 62 human RNA-interactome datasets in which U1-70K appears 39 times. It is not annotated to nuclear speckles.

These results bear on the cellular claim, not on the 2002 biochemistry as such. The 2002 data showed that anti-SOX6 antibodies inhibited splicing and that SOX6-depleted extract was rescued by recombinant HMG-box proteins [1]. Two features of the present data suggest alternative readings of those experiments. First, SOX6 mRNA in HeLa is 0–2.5 CPM, at least 100-fold below SNRNP70 in the same libraries, and HPA detects no SOX6 peptides in HeLa. A protein that scarce is unlikely to be a limiting splicing factor in HeLa extract, which raises the possibility that the antisera also recognized another protein. Second, rescue by three different HMG-box proteins at nanomolar concentrations is also consistent with a non-specific nucleic-acid-binding or chaperone effect of HMG domains on the extract.

The data do not exclude a regulated, cell-type-specific or substrate-specific role in splicing for SOX6, SRY or SOX9. The 2002 report itself speculated that SRY might control splicing of particular pre-mRNAs in the developing gonad [1]. Testing that would need RNA-seq after SRY or SOX9 loss in gonadal cells, with alternative-splicing analysis rather than the constitutive-splicing measures used here.

Limitations

- Depth: one subset of 5M read pairs per sample (4M for the control). Most introns had no unspliced reads, which made the pre-specified median statistic uninformative. Effects on weakly expressed introns could be missed.
- Sample size: n = 2–3 per guide. A 3 vs 3 design cannot reach $p < 0.10$ by exact permutation. The conclusions rest on effect sizes and on the positive control, not on p values.
- Knockout is shown at the read level (indels, reduced exon coverage and mRNA), not at the protein level. Edited fractions of 11–60% imply mixed populations.
- Cell types: mouse OPCs and OLs only. HeLa itself, the 2002 system, could not be tested this way because it expresses almost no SOX6.
- The positive control is from another cell type and laboratory, and was sequenced to a lower depth.
- DepMap values were read from the gene page. Restricting the dependency analysis to SOX6-expressing lines, and comparing with core spliceosome genes, needs the full DepMap matrix.
- Spliceosome proteomics of purified B and C complexes were not checked, because the supplementary files could not be reached from the analysis environment.
- The HeLa CPM values come from gene-level tables made by different pipelines and are not length-normalized.

Declarations

Data availability. All data are public: GEO GSE181952, GSE249801, GSE299983, GSE325258, GSE348256 and GSE242279; DepMap; RBP2GO 2.0; the Human Protein Atlas. Run accessions are given in Methods.

Code availability. Scripts for read subsetting, alignment, intron set construction, quantification, statistics and knockout validation will be deposited at Zenodo [DOI to be added].

Competing interests. The author declares no competing interests.

Funding. JSPS KAKENHI 18K09212; Fukuoka University (167005, 197010); Kyushu University Hospital ARO Center for Next-generation Medicine (A147). [Author to confirm that these apply to this work.]

Use of AI tools. Data retrieval, analysis code and a first draft of the text were produced with Claude (Anthropic) under the author's direction. The author checked the analyses and is responsible for the content.

Author's note. The author was a co-first author of reference [1].

References

1. Ohe K, Lalli E, Sassone-Corsi P. A direct role of SRY and SOX proteins in pre-mRNA splicing. Proc Natl Acad Sci USA. 2002;99:1146–1151. doi:10.1073/pnas.022645899
2. GEO GSE181952 (SubSeries of GSE197319). Transient gene melting governs the timing of oligodendrocyte maturation [RNA-seq]. [Primary publication to be added.]
3. GEO GSE249801 (SubSeries of GSE249804). Study of the effects of RNA splicing inhibitor Pladienolide B on hepatocyte proliferation in vitro. [Primary publication to be added.]
4. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019;37:907–915.
5. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–930.
6. Middleton R, Gao D, Thomas A, et al. IRFinder: assessing the impact of intron retention on mammalian gene expression. Genome Biol. 2017;18:51.
7. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a cancer dependency map. Cell. 2017;170:564–576. Dempster JM, Boyle I, Vazquez F, et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biol. 2021;22:343. DepMap Public 26Q3+Score, https://depmap.org
8. Caudron-Herger M, Jansen RE, Wassmer E, Diederichs S. RBP2GO: a comprehensive pan-species database on RNA-binding proteins, their interactions and functions. Nucleic Acids Res. 2021;49:D425–D436. Cantarella S, Neffe et al. RBP2GO 2.0. Nucleic Acids Res. 2025;gkaf1258.
9. Thul PJ, Åkesson L, Wiking M, et al. A subcellular map of the human proteome. Science. 2017;356:eaal3321. Human Protein Atlas, https://www.proteinatlas.org
10. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.