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.

Reported RNA sequence preferences of SOX2 do not survive matched controls, whereas its crosslinking localises to coding exons

Kenji Ohe

oekenji@gmail.com

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

Research Article

Keywords: SOX2, RNA-binding protein, PAR-CLIP, RNA immunoprecipitation, SELEX, reanalysis, crosslinking specificity, coding exons

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11694032/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.

Reported RNA sequence preferences of SOX2 do not survive matched controls, whereas its crosslinking localises to coding exons

Kenji Ohe

Department of Clinical Pharmacotherapy, Faculty of Pharmaceutical Sciences, Fukuoka University, Fukuoka, Japan

Correspondence: ohekenji@fukuoka-u.ac.jp

Abstract

SOX2 is cited as an RNA-binding protein on the basis of two independent 2020 studies: RNA immunoprecipitation (RIP) in mouse embryonic stem cells, which described high-affinity, low-specificity binding to double-stranded RNA, and PAR-CLIP in HEK293T together with RNA SELEX, which described a GC-rich CCCY core motif and a CGCG side motif. Both datasets are public. I asked which parts of this evidence survive when control RNA-binding proteins are processed through an identical pipeline, and when the raw SOX2 reads are reprocessed rather than taken as deposited peak calls.

Three published claims do not survive. The two datasets do not identify the same targets: length- and abundance-corrected RIP targets show no enrichment among PAR-CLIP-bound genes (Mantel–Haenszel odds ratio 0.891, 95% CI 0.800–0.991), and neither species nor cell type accounts for this. Under one pipeline applied to twelve PAR-CLIP datasets from six studies, the CCCY odds ratio for SOX2 (1.16–1.22) is lower than for the unrelated proteins FXR1 and FXR2 (1.69–1.78, non-overlapping intervals), while the same pipeline recovers the known preferences of ZFP36 and FMRP and returns null results for MATR3. And in the SELEX pool, whose GC content is 0.670, GCCCY enrichment falls from 5.68-fold against uniform sequence to 1.36-fold against a composition-matched background ($p = 0.26$).

One feature survives, reduced in magnitude. Scoring T-to-C conversions against a background drawn from the same aligned reads, and against the non-T-to-C mismatch rate as an internal specificity control, gives an apparent thirtyfold excess of specificity in coding exons over introns. But 42% to 46% of conversion events lie in structured non-coding RNA or outside annotated genes, and 34 genes account for half of one replicate even after those classes are removed. With them removed and each gene limited to a fixed contribution, the contrast falls to roughly fourfold and the odds ratio against read placement is 1.08–1.16 for exonic against 0.81–0.88 for intronic positions, reproducibly in both libraries; mappability, thymidine content, splice-junction soft-clipping, common variants and local GC content do not account for it. Exonic targets are also not enriched among the genes SOX2 occupies in human embryonic stem cells (odds ratio 0.879, 95% CI 0.753–1.027). The public data therefore support a modest compartment preference rather than a sequence preference, and the reported motifs require identically processed control proteins before they can be attributed to SOX2.

Introduction

A growing literature reports that sequence-specific DNA-binding transcription factors also bind RNA. SOX proteins have been implicated in RNA processing since SRY, SOX6 and SOX9 HMG domains were shown to be required for pre-mRNA splicing in vitro [1]; SOX9 has since been assigned roles in alternative splicing that are separable from its transcriptional activity [12], and RNA binding has been surveyed across the family [11]. For SOX2 specifically, two 2020 studies reached different conclusions from different methods. One, using formaldehyde and UV RIP in mouse embryonic stem cells (mESC), described high-affinity, low-specificity binding to double-stranded RNA [2]. The other, using PAR-CLIP in HEK293T together with RNA SELEX, described a GC-rich CCCY core motif and a CGCG side motif [3]. Both datasets are public, and both are cited as evidence that SOX2 is an RNA-binding protein.

Two features of this evidence base invite re-examination. First, the two studies have not been compared to each other at the level of their target lists, so it is not known whether they describe the same molecular event. Second, neither included an unrelated RBP processed through the same pipeline. The need for such controls is not hypothetical. A systematic reanalysis of 33 PAR-CLIP datasets found that covalently crosslinked background binding is common, reproducible and apparently universal, that biological replicates cannot remove it because it is itself reproducible, and that background overlap is highest for weakly binding proteins, for proteins whose RNA-binding status is putative, and for low-read libraries [4]. The recommended remedy is an empirically measured negative control, such as a sequenced FLAG-GFP immunoprecipitation. The SOX2 PAR-CLIP deposit contains no such sequenced control, and SOX2 falls squarely in the category the authors identified as most exposed.

Here I ask which parts of the published evidence survive two additions: control RBPs analysed under one pipeline, and reprocessing of the raw SOX2 reads so that crosslinking can be examined before peak calling. I do not ask whether SOX2 binds RNA at all, which these data cannot settle.

Results

The two published SOX2 datasets recover different RNA pools

The GSE142013 series provides gene-level UMI counts for eight matched immunoprecipitation/input pairs (four formaldehyde-crosslinked, four UV-crosslinked) in mESC, allowing enrichment to be computed without reprocessing raw reads. Restricting to 13,491 genes with at least 10 reads in every input, per-pair log2(IP/input) showed essentially no relationship with input abundance (Spearman rho −0.015 to −0.112 across the eight pairs). Genes in the top 1,000 by fRIP enrichment had a median input abundance indistinguishable from the bottom 1,000 (4.96 versus 4.98 on a log2 CPM scale; overall median 4.67). Reported SOX2 RNA targets are therefore not simply abundant transcripts.

Transcript length, however, was strongly associated with enrichment. The correlation between fRIP enrichment and log10 gene length was rho = +0.510; the median length of the top 1,000 targets was 84,524 nt against 5,636 nt for the bottom 1,000 (overall median 25,590 nt). The corresponding correlation for UV-RIP was weaker (+0.163).

Mapping genes to mouse genome annotation and computing union-exon and intronic lengths showed that the length bias is an intron bias. Before length correction, fRIP enrichment correlated with intronic fraction at rho = +0.406; after LOWESS removal of the log10-length trend, the correlation fell to +0.008 (p = 0.347), while the length correlation itself fell to +0.001. The equivalent intronic-fraction correlation for UV-RIP was +0.065 before correction. Formaldehyde, which also crosslinks protein to protein and can recover entire nascent ribonucleoprotein territories, thus recovers intron-containing pre-mRNA preferentially; UV crosslinking, which captures direct contacts, does so far less.

Assigning the 16,964 (replicate 1) and 55,830 (replicate 2) published SOX2 PAR-CLIP peaks to GRCh37 genes gave 12,007 genes with at least one peak. Peak counts correlated weakly with gene length (rho = +0.189 and +0.144), and peak scores essentially not at all (+0.060 and −0.064). Ordering the four measurements by the fraction of a transcript each recovers produces a monotone series: fRIP (+0.510) > UV-RIP (+0.163) > PAR-CLIP peak counts (+0.144) > PAR-CLIP peak scores (−0.064). The same ordering holds for intronic fraction, with PAR-CLIP scores correlating negatively (rho = −0.125, p = 1.8 × 10^−41).

A reproducible RIP signal nevertheless exists and is not an artefact of length. Agreement between the two crosslinkers survived length correction and slightly improved (Spearman rho between mean fRIP and mean UV-RIP enrichment: +0.612 before, +0.635 after). Overlap of the top 500 genes remained 9.7-fold above chance after correction, and 2,616 genes were enriched in all eight samples (expected 53 by chance).

The two datasets do not identify the same targets, and species and cell type do not explain it

Human–mouse orthologue pairs were taken from NCBI gene_orthologs (17,090 pairs), giving 11,743 genes comparable between the mESC RIP data and the HEK293T PAR-CLIP data, of which 5,326 carried peaks in both PAR-CLIP replicates. Length-corrected RIP consensus targets showed no enrichment among PAR-CLIP-bound genes (Fisher OR 0.978, p = 0.70). Stratifying by quintiles of human gene length gave odds ratios of 0.817, 0.839, 0.881, 0.940 and 0.940, with a Mantel–Haenszel common OR of 0.891 (95% CI 0.800–0.991; test of homogeneity p = 0.91). As a continuous measure, length-corrected RIP enrichment correlated negatively with PAR-CLIP peak score (rho = −0.116, p = 1.5 × 10^−27).

Expression differences do not account for this. Using HEK293 RNA-seq (GSE122425) and mESC expression from the RIP inputs, the comparison was repeated on genes expressed above thresholds of 0.5, 1, 5 and 20 units in both cell types; Mantel–Haenszel common odds ratios were 0.946, 0.936, 0.734 and 1.234, with confidence intervals including unity throughout. Restricting instead to genes expressed in only one cell type gave the same result. The failure to overlap is not a missing-orthologue or missing-expression artefact:

95.2% of RIP consensus targets have a human orthologue expressed in HEK293 (median 5.07 RPKM).

Two published target lists for the same protein therefore do not identify the same genes, and the difference is not explained by the transcriptomes of the two cell types. What does explain it is taken up in the Discussion.

The reported sequence features are not distinguishable from those of control RBPs

Twelve PAR-CLIP datasets from six independent studies were processed identically: all peaks read, 4,000 sampled at random without regard to chromosome order, a fixed 21-nt window taken around each peak centre, and five within-gene shuffled windows per peak as background. Logistic regression modelled peak status on GC content (per 10%), presence of CCCY, CGCG and GGAC, and transcript region.

Under this design, SOX2 gave CCCY odds ratios of 1.22 (95% CI 1.09–1.35) and 1.16 (1.04–1.29) for the two replicates. FXR1 and FXR2, which have no reported relationship to a CCCY element, gave 1.78 (1.52–2.08) and 1.69 (1.47–1.93), with confidence intervals that do not overlap those of SOX2. CGCG odds ratios were below unity for SOX2 (0.70 and 0.50) but varied in both direction and magnitude across controls (0.40 to 17.37), so no consistent baseline exists against which the SOX2 value can be read.

The pipeline is not insensitive. ZFP36, an AU-binding protein, returned a GC odds ratio of 0.23 per 10% and CCCY and CGCG odds ratios of 5.33 and 17.37. FMRP returned GGAC odds ratios of 2.84 and 2.39, consistent with its published WGGA element [5]. MATR3, with 133,089 clusters, returned odds ratios near unity for every feature in both replicates (CCCY 0.95 and 0.99, CGCG 0.90 and 1.26, GGAC 1.03 and 0.95, GC 1.00 and 1.00), demonstrating that the design does not manufacture enrichment. MOV10 K530A reproduced MOV10 wild type closely, providing an internal technical control.

Two features of the design deserve note. Because PAR-CLIP peak files are sorted by coordinate, reading only the first n peaks of a file confines large datasets to the lowest-numbered chromosomes; every peak was therefore read before sampling. And the control datasets are unevenly sourced, four of the twelve coming from a single study.

A de novo motif in the SOX2 data is attributable to a co-occurring FMRP-like activity

Unbiased 5-mer analysis of SOX2 peaks returned CGGAC, TGGAC, GGACG and related sequences as the most enriched (up to 5.19-fold over shuffled background), above any CCCY-containing 5-mer, and reproducibly in both replicates. GGAC occurrences clustered at the peak centre (2.93-fold enrichment within ±5 nt versus 1.31-fold beyond ±30 nt; Fisher OR 2.212, p = 9.4 × 10^−219), the profile expected of a binding determinant rather than of a compositional bias. This motif was not reported in the original study; it is a de novo observation here, and the following paragraphs assign it to a source other than SOX2.

The motif is not m6A-related. Using single-base m6A sites in HEK293 (GSE162356; 12,234 sites in 5,895 genes), genes with GGAC-containing peaks appeared enriched for

m6A (OR 1.973), but stratification by peak count reduced this to a Mantel–Haenszel common OR of 1.122 (95% CI 1.024–1.229). The sequence contexts also differ: m6A sites were dominated by AGACT (27.5%) and GGACT (22.4%), that is, DRACH with a constrained 3' position, whereas GGAC within PAR-CLIP peaks was preceded by T (37.5%), A (25.5%) or G (24.5%) with no 3' constraint.

Instead, TGGAC and AGGAC together account for 63% of GGAC contexts in SOX2 peaks, and WGGA is the published FMRP recognition element [5]. Under the uniform pipeline, FMRP gave higher GGAC odds ratios than SOX2 (2.84 and 2.39 versus 1.77 and 2.27). Comparing peak coordinates directly, 18.4% of SOX2 replicate-2 peaks overlap FMRP isoform-1 peaks and 14.5% overlap isoform-7 peaks, against 0.3% for coordinate-shifted controls preserving chromosome, strand, interval length and density (52.7- and 54.6-fold). The corresponding overlaps with FXR2, FXR1 and ZFP36 were 5.9%, 1.0% and 0.3%. The GGAC signal in the SOX2 dataset therefore cannot be attributed to SOX2 rather than to a co-occurring FMRP-like activity or to a shared property of the transcripts both experiments recover. Whether the two experiments share transcripts because of co-immunoprecipitation or because both recover the same abundant exonic pool is not resolved by these data.

The published SELEX core motif is explained by the base composition of the selected pool

The 51 unique 25-nt sequences recovered after twelve rounds of SELEX with purified His6-SOX2 were extracted from the published supplementary table. The pool has a GC content of 0.670, well above the 0.50 expected of the starting library, confirming selection for GC-rich sequence.

Compared with uniform random sequence, GCCCY appears enriched 5.68-fold. Compared with random sequence of the same base composition as the selected pool, the enrichment is 1.38-fold and not significant ($p = 0.26$); CCCY is at 0.83-fold ($p = 0.46$). CGCG remains enriched 2.71-fold ($p = 4.4 \times 10^{-9}$), as does CCC at 1.70-fold ($p = 2.0 \times 10^{-12}$). The SELEX data thus support a GC-rich preference and a residual CGCG preference, but the published GCCCY core is accounted for by composition.

GGAC is absent from all 51 sequences (expected frequency 0.110 under matched composition, $p = 0.0056$), and WGGA appears in one (0.020 versus 0.097 expected). The de novo motif found in the cellular PAR-CLIP data is therefore not selected by purified SOX2 in vitro, independently supporting the assignment made in the preceding section.

Finally, the 51 sequences are not independent observations. Median pairwise Hamming distance is 16 of 25, significantly below the 18 expected for composition-matched random sequence (Mann–Whitney $p = 4.8 \times 10^{-55}$), and single-linkage clustering at a distance of 8 yields 34 clusters from the 51 sequences, the largest containing 6. A statement that a motif occurs in all 51 sequences carries correspondingly less weight.

The two experiments used to define the motifs do not support the same one: PAR-CLIP retains CCCY but reverses CGCG after adjustment, whereas SELEX retains CGCG but loses GCCCY.

What survives: a modest exonic preference within a dataset with a large background fraction

The three preceding sections concern sequence. To ask where on the transcript SOX2 crosslinks, the raw reads of both PAR-CLIP replicates (SRR7253210 and SRR7253211) were reprocessed, since the deposit contains only PARalyzer cluster calls. Reads proved to be adapter-trimmed before deposition (median length 25 and 28 nt; 1.8% and 1.3% of reads carried the adapter), and 1,448,852 and 1,702,115 reads aligned uniquely to GRCh37 (overall alignment rates 72.42% and 65.61%). T-to-C conversions were called per read, giving 691,254 and 915,494 conversion events on assembled chromosomes, together with 501,168 and 418,287 mismatches of every other type.

Two background models were used. The first shuffles each site to a random position within its gene. The second, which additionally controls for mappability and for local read placement, draws one random thymidine position from each aligned read, so that the background occupies exactly the positions the sequencing and alignment steps were able to reach. Two measures are reported throughout: the odds ratio of a region against the read-matched background, which asks whether conversions are redistributed relative to where reads land, and the ratio of T-to-C events to all other mismatches within the same region, which asks how far the T-to-C signal there stands above the local mismatch background. The two answer different questions and, as shown below, differ in magnitude.

Mismatches of types other than T-to-C arise from sequencing error, misalignment and genetic variation, and they behaved in the direction expected for repeat-rich, low-complexity sequence where alignment degrades: intron odds ratios of 2.435 and 2.364 against the read-matched background, and CDS odds ratios of 0.152 and 0.131. Their distribution is therefore the opposite of the T-to-C distribution, which makes their ratio usable as an internal, per-region specificity measure.

Taken at face value, the event-level tallies give a large effect: specificity ratios of 29.5 and 23.9 in coding exons against 0.98 and 1.56 in introns, with read-matched odds ratios of 1.739 and 1.531 for CDS against 0.976 and 1.075 for introns. Three further controls show that most of this magnitude does not survive.

First, the composition of the signal. Structured non-coding RNA and its pseudogenes account for 31.5% and 21.0% of all conversion events, dominated by 5S ribosomal RNA pseudogenes (129,649 and 146,940 events), transfer RNA, ribosomal RNA, small nuclear RNA and vault RNA; a further 14.6% and 20.6% of events fall outside any annotated gene. Between 42% and 46% of the T-to-C signal therefore lies in the classes that an immunoprecipitation recovers non-specifically. This is the background fraction that the systematic survey of PAR-CLIP datasets predicted would be largest for proteins of uncertain RNA-binding status [4], here quantified for this dataset.

Second, the concentration of the signal on few loci. Eleven genes account for half of all replicate-1 conversion events and the ten highest for 49.0%. After removing structured non-coding RNA and unassigned events, 34 genes still account for half of replicate 1, against 1,631 for replicate 2, and the ten highest for 40.9% against 10.5%: the two libraries differ substantially in complexity. Capping each gene's contribution at a fixed num-

ber of events equalises genes and asks whether, position for position, exonic sequence is more likely to be crosslinked.

Third, the two corrections together. With structured non-coding RNA and unassigned events removed and each gene capped at ten events per class, the specificity ratios are 19.0 and 15.0 in coding exons, 16.5 and 12.9 in non-coding exons, and 4.1 and 3.5 in introns; read-matched odds ratios are 1.111 and 1.078 for CDS, 1.163 and 1.087 for non-coding exon, and 0.814 and 0.881 for intron (Table 1). Raising the cap to 100 gives 26.4 and 24.4 in coding exons against 3.7 and 5.6 in introns. The ordering is reproducible between replicates at every cap, but the exonic-to-intronic contrast falls from roughly thirty-fold to roughly fourfold, and the odds ratio against read placement is modest throughout. Restricting instead to protein-coding and long non-coding genes gives the same values.

Local sequence composition does not account for the regional difference. The mean GC content of 51-nt windows around background positions is nearly identical in coding exons and introns (0.5454 versus 0.5442 for replicate 1; 0.5426 versus 0.5267 for replicate 2), so the two regions are not separated by GC at the scale the reads sample. Stratifying by GC decile left the region odds ratios unchanged (CDS 1.766 and 1.561; intron 0.970 and 1.066), and a model fitting region and GC together returned region effects of the same size (CDS versus intron 1.708 and 1.432). In that model GC itself was negatively associated with being a conversion site (0.921 and 0.911 per 10%), the opposite sign from the GC odds ratio obtained at the level of PARalyzer clusters against shuffled windows (1.13–1.15 per 10%). Because thymidine density falls as GC rises and conversions can only be scored at thymidines, neither coefficient can be read as an RNA sequence preference, and no claim about GC preference is made from the PAR-CLIP data in either direction.

Three further alternative explanations were excluded. Mappability and local expression are controlled by the read-matched background. Thymidine content does not differ in the required direction (mean thymidines per read 7.15 in CDS versus 7.18 in intron for replicate 1, and 7.10 versus 7.05 for replicate 2); non-coding exonic reads have the highest thymidine content, 8.67 and 8.58, with intermediate conversion rates. Misplacement of splice-junction reads by local alignment does not account for it, since mean soft-clipped length is not higher in CDS than in introns (1.27 versus 1.32, and 1.14 versus 1.64). Common genetic variation does not account for it: removing all 13,200,832 common single-nucleotide variants in dbSNP build 151 eliminated 0.13% to 0.66% of conversion events, with the largest proportion in introns, and moved every specificity ratio by less than 5%.

Because 4-thiouridine labels newly synthesised RNA, whereas mature mRNA accumulates over a longer interval, the labelling step is expected to favour intron-containing nascent transcripts. The observed direction is the opposite, which makes the exonic result conservative with respect to this bias.

Reproducibility between the two libraries is adequate at the level of genes and good at the level of specificity. Conversion counts per gene correlate at Spearman rho +0.811 across all 20,082 genes detected in either library and +0.704 among the 10,506 with at least ten events, but only +0.405 among the 960 with at least a hundred, so the highest-count loci are the least reproducible. Sixty-five of the hundred highest-count genes are shared. The per-gene specificity ratio, in contrast, correlates at rho +0.846 (p = 1.7 ×

10^−105, 381 genes with at least twenty conversions and five other mismatches): which loci carry a clean T-to-C signal is consistent between libraries even where the counts are not.

That measure separates the high-count loci into two groups. One has specificity ratios far above the genome-wide value and appears in both libraries: EYA3 (87.9 and 84.2), EYA4 (90.6), NFIB (63.1), H3C2 (55.3), FAM107B (52.7 and 51.6), TDRD9 (41.5 and 31.6), EEF1A1 (28.2), WDR74 (24.2) and CDH1 (9.9); ZNF704 reaches 377 but on only twenty comparison events and is not a stable estimate. The other has high counts and background-like specificity: DPP10 (0.5), BCAS3 (0.3), MTMR2 (0.6), CCDC91 (1.0) and PCLO (1.1). The first group is the set of loci this experiment recovered with high T-to-C specificity, and it includes two paralogues of the same family, EYA3 and EYA4. That group cannot be called SOX2-specific: without a sequenced negative immunoprecipitation there is no measurement of what the same protocol recovers in the absence of the protein, and two of its members belong to classes that immunoprecipitations recover non-specifically, a replication-dependent histone gene and a very highly expressed translation factor.

Within-gene shuffling gave the same direction for the raw conversion sites as for the deposited cluster calls, and the two differ much less than had been expected: intron folds of 0.56 and 0.63 for raw sites against 0.53 and 0.43 for clusters, with CDS folds of 3.69 and 3.57 against 3.43 and 3.90. Raising the minimum number of conversion events per site from 1 to 10 left the intronic fold between 0.39 and 0.71 and never moved it towards intronic enrichment. Whatever exonic preference exists is therefore present at the level of individual crosslink sites and is not created by the clustering step.

Exonic PAR-CLIP targets are not the transcripts of SOX2-occupied loci

If SOX2 engaged RNA while bound at its genomic sites, its RNA targets should be drawn preferentially from the genes it occupies. The exonic localisation described above predicts the opposite, because mature mRNA is separated from the site of transcription. This prediction is testable with existing data.

SOX2 chromatin immunoprecipitation data for undifferentiated human embryonic stem cells (HUES64) were taken from GSE61475 [17], which provides hg19 peak calls for two independently prepared samples (10,624 and 6,211 peaks). Genes were called occupied when a peak was present in both samples, under two definitions: a peak overlapping any transcription start site ±2 kb, and a peak overlapping the gene body ±10 kb, the latter being the more appropriate definition for a factor that binds predominantly distal elements. The assignment recovers the expected loci: NANOG, POU5F1, SOX2 itself, LIN28A, DPPA4, LEFTY1, SALL4 and FOXD3 are all reproducibly occupied.

Among 10,213 protein-coding and long non-coding genes with HEK293 expression of at least 1 RPKM, exonic PAR-CLIP targets (defined as at least one exonic peak in each replicate; 3,577 genes) showed no enrichment among SOX2-occupied genes. Stratifying by quintiles of HEK293 expression gave a Mantel–Haenszel common odds ratio of 0.944 (95% CI 0.627–1.422; homogeneity p = 0.80) for the promoter definition and 0.879 (95% CI 0.753–1.027; homogeneity p = 0.34) for the gene-body definition (Table 3). Logistic regression adjusting simultaneously for log10 HEK293 expression and log10 gene length gave an odds ratio of 0.920 for exonic target status (95% CI 0.610–1.389, p = 0.69); in the

same model only gene length predicted occupancy (1.53 per log10, 95% CI 1.09–2.14, p = 0.015). Replacing the exposure with any PAR-CLIP peak gave 1.049 (0.661–1.664) and with intronic peaks 0.792 (0.544–1.152). The number of exonic peaks per gene did not differ between occupied and unoccupied genes (median 0 versus 0 for replicate 1, 2 versus 2 for replicate 2; Mann–Whitney p = 0.40 and 0.64).

The RNA targets recovered by PAR-CLIP are therefore not the transcripts of the loci SOX2 occupies. Two qualifications bound this. First, power: with 200 occupied genes under the promoter definition the data exclude an odds ratio above approximately 1.4 but not a small positive effect; the gene-body definition, with 1,655 occupied genes, is tighter and its point estimate lies below unity. Second, the two assays come from different cells, chromatin immunoprecipitation from HUES64 and PAR-CLIP from HEK293T, and the GSE115273 deposit contains no chromatin immunoprecipitation from the same overexpression system, so where transfected SOX2 binds the HEK293T genome is unknown. A limited form of this confounding can be excluded: SOX2-occupied genes are not depleted of HEK293 expression as a class (43.8% below 1 RPKM against 47.4% of unoccupied genes), so the absence of enrichment is not simply a consequence of occupied genes being silent in HEK293. The canonical pluripotency loci are the exception, and an informative one: NANOG, POU5F1, LIN28A, DPPA4 and ZFP42 have HEK293 expression of 0 to 0.2 RPKM and therefore cannot appear as PAR-CLIP targets in this system at all, whatever SOX2 does to them in a pluripotent cell.

A comparison that cannot be made with the available data

Deposited conversion-site files for other proteins are intron-enriched when analysed as provided (OR 2.98 for MOV10, 2.84 for MOV10 K530A, 2.74 for UPF1, and 2.01–3.13 for MATR3 cluster centres), which invites the inference that clustering shifts signal from introns to exons in general. Those values come from each study's own pipeline, whereas the SOX2 values above come from the pipeline described here, so the comparison is confounded by processing. No such pipeline-matched comparison is attempted, and the general claim is not made.

Discussion

What the public data support, and what they do not

Three published claims about SOX2 and RNA do not survive the controls applied here. The two independent target lists do not identify the same genes, and the discrepancy is not attributable to species or cell type. The CCCY odds ratio obtained for SOX2 under a uniform pipeline is lower than that obtained for two unrelated proteins with no reported CCCY element, and the CGCG odds ratio has no consistent sign across controls. In the SELEX pool, the published GCCCY core is accounted for by the base composition of the selected sequences, and the 51 sequences are not independent.

One feature survives, and it concerns compartment rather than sequence, but it is smaller than the uncorrected tallies suggest. Taken at face value, the T-to-C specificity ratio is thirtyfold higher in coding exons than in introns. Once structured non-coding RNA and

unassigned events are removed, and once each gene is limited to a fixed contribution so that no locus dominates, the contrast is roughly fourfold (coding exon 15–19, intron 3.5–4.1) and the odds ratio against read placement is 1.08–1.16 for exonic against 0.81–0.88 for intronic positions. These corrections are not optional: between 42% and 46% of the conversion signal lies in structured non-coding RNA or outside annotated genes, and after filtering, 34 genes still account for half of one replicate. Mappability, thymidine content, splice-junction soft-clipping, common variants and local GC content were each tested and do not account for the residual difference, and the direction of the 4-thiouridine labelling bias runs against it. The statement supported by the public data is therefore that exonic positions in spliced RNA are modestly more likely to be crosslinked than intronic positions, and that the T-to-C signal in introns stands much closer to the generic mismatch background than it does in exons. The larger figure should not be quoted.

Nothing here shows that SOX2 lacks RNA sequence preference, and nothing here establishes one either. A GC-rich preference is supported by the composition of the SELEX pool, where the selected sequences reach 0.670 against the 0.50 expected of the starting library, and the residual CCC enrichment against a composition-matched background is significant (1.71-fold, p = 1.9 × 10^−12). The PAR-CLIP data do not add to this: the GC odds ratio obtained at the level of clusters against shuffled windows (1.13–1.15 per 10%) and the GC coefficient obtained at the level of individual conversion sites against a read-matched background (0.921 and 0.911 per 10%) have opposite signs, and both are confounded by the fall in thymidine density as GC rises, so neither is interpretable as a binding preference. The published EMSA experiments, in which substituting a CCC core with UUU, GGG or AAA reduced binding without altering predicted secondary structure, are not addressed by a reanalysis of sequencing data and are consistent with the SELEX residual. What the reanalysis shows is narrower: the specific motifs GCCCY and CGCG are not jointly supported, each failing in one of the two experiments used to define them, and neither can be separated from the behaviour of unrelated proteins analysed identically.

The leading unexcluded explanation for the disagreement between datasets

Species and cell type were tested and rejected as explanations for the failure of the two target lists to agree. That test matched the expression of candidate target transcripts; it did not match the state of the protein.

The SOX2 PAR-CLIP was performed on HEK293T cells transfected to express SOX2, a cell line that does not express it endogenously. Three consequences follow. The stoichiometry of SOX2 to its endogenous partners, including OCT4 and its genomic binding sites, is not physiological, so the fraction of SOX2 not engaged by a partner — the fraction available to bind RNA — is abnormally large. The available RNA pool is the HEK293 transcriptome, so transcripts specific to pluripotent cells cannot appear among the targets. And the HMG box is a non-sequence-specific nucleic-acid-binding domain, so an excess of unengaged protein is the condition under which low-specificity binding would be expected. The direction of this effect is not determined: overexpression may generate binding that does not occur at physiological concentration, or it may reveal binding that is masked by genomic occupancy in cells that express SOX2 natively.

This is a better-supported explanation for the disagreement than species or cell type, and it is not excluded by any analysis presented here. It also connects to the background problem identified for PAR-CLIP generally [4]: background overlap was found to be highest for putative RBPs, for weak binders and for low-read libraries, and the recommended control is a sequenced negative immunoprecipitation. The SOX2 deposit contains no sequenced negative control; the FLAG-GFP control in the original work was assessed by blotting only, and only two SOX2 samples are deposited. The dataset thus lacks precisely the measurement that would quantify how much of its signal is background.

Determining the RNA targets of endogenous SOX2 in human iPSC

The experiment the public data cannot substitute for is crosslinking of endogenous SOX2 in a cell type that expresses it. Human induced pluripotent stem cells (iPSC) are the appropriate setting, both because SOX2 is expressed there at its physiological level and because the function attributed to its RNA binding — reprogramming — is defined in that system. Three design elements are required, none of which is present in the data analysed here.

The first is physiological stoichiometry with a fixed immunoprecipitation reagent. Knocking a small epitope tag into the endogenous SOX2 locus by homology-directed repair keeps the protein at its own expression level, removes the antibody-validation problem, and removes the transfection excess that complicates the HEK293T dataset. Appending a degron at the same locus supplies the matching depletion condition from the same cell line and the same clone.

The second is a sequenced negative control carried through the identical pipeline. The eCLIP design, in which a size-matched input from the same lysate is sequenced alongside the immunoprecipitate [13], provides this by construction. The degron-depleted or SOX2-knockout sample provides the stronger version: an immunoprecipitation performed with the same reagent in cells lacking the antigen, which measures what the protocol recovers in the absence of the protein. A sample omitting ultraviolet irradiation, and an RNase titration, establish that recovery is crosslink- and RNA-dependent. For a protein whose RNA-binding status is the question at issue, these controls carry more weight than the choice of protocol: the systematic survey of PAR-CLIP background found the largest background fractions precisely among proteins of uncertain RNA-binding status [4].

The third is compartment resolution. The exonic localisation reported here predicts that SOX2 contacts mature mRNA rather than nascent transcripts. That prediction is testable by performing the same crosslinking on separated chromatin-associated, nucleoplasmic and cytoplasmic fractions. If the signal is nucleoplasmic or cytoplasmic and absent from the chromatin-associated fraction, the interaction is not co-transcriptional, and the hypothesis that SOX2 engages RNA while occupying its genomic sites is excluded. If instead the signal is chromatin-associated yet still exonic, the interaction would have to occur on transcripts that have already been spliced but not yet released, a more restrictive claim that would require direct demonstration.

Two analytical points follow from the present work. If 4-thiouridine is used, the read-matched background and the non-T-to-C mismatch specificity ratio applied here transfer directly, and candidate sites should be ranked by that ratio rather than by cluster score,

since the ratio is the quantity that distinguishes crosslinking from the local mismatch background. Short labelling pulses are necessary, because 4-thiouridine inhibits ribosomal RNA synthesis and is poorly tolerated by pluripotent cells over long intervals. If 254 nm crosslinking is used instead, there is no conversion signature, and the internal specificity control must be re-derived: the deletion rate at the crosslink position in iCLIP and eCLIP serves the analogous role, and its regional behaviour should be characterised on well-established RNA-binding proteins before it is applied to a protein under test.

Comparing an RNA interactome with the SOX2 cistrome and with the other reprogramming factors

The DNA side of this comparison is already well mapped in pluripotent cells, which makes several specific contrasts available once matched RNA data exist. All of them require that the cistrome and the interactome be measured in the same cell line, ideally the same tagged clone, so that expression differences do not drive the comparison.

Is the RNA interaction cis or trans? This question is partly answered already. Exonic PAR-CLIP targets show no enrichment among the genes SOX2 occupies in human embryonic stem cells, under either a promoter or a gene-body definition of occupancy and after adjustment for expression and gene length. The result is consistent with contact on mature mRNA away from chromatin, but it is bounded by the fact that the two assays were performed in different cells. The comparison that would settle it is chromatin immunoprecipitation and CLIP from the same tagged iPSC clone, which also removes the possibility that transfected SOX2 occupies a different genome-wide site set than endogenous SOX2 does.

Do the two activities read the same sequence? The SOX2 HMG box binds the minor groove of A/T-rich DNA and bends it, and acts through a composite element in which its site is closely spaced with the OCT4 octamer [14], an arrangement that underlies much of the pluripotency cistrome [15]. The RNA preference indicated by both the SELEX pool composition (GC content 0.670) and the PAR-CLIP GC odds ratios reported here is GC-rich, that is, the opposite base composition. If a single nucleic-acid-binding surface were responsible for both, the two consensus sequences would be expected to resemble each other more than they do. Determining whether the HMG box or another region of the protein contacts RNA is therefore a prerequisite for interpreting any target list, and point mutants that abolish DNA binding while retaining the fold provide the discriminating reagent: comparing chromatin immunoprecipitation and CLIP in the same mutant separates the two activities experimentally rather than by inference.

Is RNA binding a property of this factor or of the reprogramming factor set? RNA-binding capacity appears to be widespread among transcription factors, mediated in many cases by arginine-rich motifs rather than by canonical RNA-binding domains, and capable of affecting gene regulation [16]. OCT4, KLF4 and MYC bind DNA through a POU domain, zinc fingers and a basic helix–loop–helix leucine zipper respectively, all of which are nucleic-acid-binding folds, so there is no reason to expect SOX2 to be unusual in this respect. What has not been done is a uniform test: whether the reprogramming factors bind RNA in cells, with what specificity, and whether the specificity differs between them, has not been addressed with matched controls and one pipeline. Processing all four factors

from the same tagged iPSC background, with the same depletion controls, would answer it. The present reanalysis provides the control framework for that comparison but not the data.

Does the RNA interaction have a function? The functional claim in the literature is that concurrent DNA and RNA binding facilitates reprogramming [3]. Testing it requires a separation-of-function mutant: a SOX2 variant that retains DNA binding and the composite element with OCT4 but loses RNA binding, assayed for reprogramming efficiency in a defined system. Without such a mutant, a correlation between RNA targets and reprogramming outcome cannot be distinguished from the DNA-binding activity of the same protein. Identifying the RNA-contacting residues, which the comparison above would begin, is the step that makes this mutant designable. The read-out should be the direction of the effect rather than its presence: if RNA binding assists reprogramming the mutant will be less efficient, whereas if abundant RNA instead sequesters SOX2 away from chromatin the same mutant will be more efficient, and the two possibilities are distinguished by sign.

Limitations

The central limitation is the absence of a sequenced negative control, and the two libraries differ substantially in complexity (Table 2), so replicate-level inference rests on agreement between two unlike samples. The absolute magnitude of every event-level quantity depends on how far the dominant loci are allowed to contribute, which is why the capped values are reported as the primary estimates. A reader who prefers the uncapped tallies will read a much larger effect from the same data, and the choice between them cannot be settled without a negative control. For the same reason, the high-specificity loci named here are what this experiment recovered, not what SOX2 is shown to bind.

Several analytical limitations follow from the use of deposited data. Cluster definitions differ between studies, and although a fixed 21-nt window was imposed to mitigate this, upstream pipeline differences cannot be fully removed. Control datasets are unevenly distributed across RNA-binding domain classes, with four of twelve from a single study, and GGAC preference did not track domain class. The cross-species comparison relies on one-to-one orthologue mapping and on a single HEK293 expression dataset. Reads were subsampled to five million per replicate before alignment, which limits absolute site counts but not the proportions on which the regional analyses depend.

Two specific points require care in any replication of this analysis. Mitochondrial and unplaced sequence were excluded from the event-level analysis: the mitochondrial genome is almost entirely coding and contributes reads with elevated mismatch rates, so including it inflates the coding-exon mismatch background and distorts the specificity ratio. And the attribution of the SOX2–FMRP peak overlap to co-immunoprecipitation rather than to shared transcript properties is not resolved by these data.

Methods

Data

SOX2 fRIP and UV-RIP gene-level UMI counts for eight matched IP/input pairs in mESC were taken from GSE142013. SOX2 PAR-CLIP peak calls, made with PARalyzer [6] on hg19, for two replicates were taken from GSE115273 [3] (BioSamples SOX2_PAR-CLIP_rep1 and rep2), and the corresponding raw reads from SRR7253210 and SRR7253211 via the European Nucleotide Archive. Control PAR-CLIP peak calls were taken from GSE39682 [5] (FMRP isoform 1 and isoform 7, FXR1, FXR2), GSE53184 [7] (ZFP36), GSE48245 [8] (MOV10 wild type, MOV10 K530A and UPF1, as individual T-to-C conversion sites) and GSE262647 [9] (MATR3 replicates 1 and 2). One further dataset (GSE84134) was processed but excluded because the deposited file names did not permit unambiguous assignment of the immunoprecipitated protein. Single-base m6A sites in HEK293 were taken from GSE162356 [10], and HEK293 expression from GSE122425 (mean RPKM of the three HEK293 libraries, excluding the HEK293NK libraries). SOX2 chromatin immunoprecipitation peak calls for undifferentiated HUES64 human embryonic stem cells, on hg19, were taken from GSE61475 (GSM1505766 and GSM1505768, two independently prepared samples). The genome build of these files was confirmed by the presence of peaks at the hg19 NANOG promoter and their absence at the corresponding hg38 coordinates. The 51 SELEX sequences were extracted from Supplementary Table 2 of reference [3]. Common single-nucleotide variants were taken from UCSC snp151Common for hg19 (13,200,832 variants of class "single").

Annotation

Human annotation was GCF_000001405.25_GRCh37.p13 and mouse annotation GCF_000001635.27_GRCm39, both from NCBI, restricted to assembled chromosomes. Genome sequence was GRCh37.p13 (UCSC hg19). m6A sites, which are reported on GRCh38, were assigned to genes using GCF_000001405.40_GRCh38.p14; the build was confirmed by checking the motif strings recorded in the deposited file against GRCh37 sequence, which matched in 0 of 400 cases. Orthologues were taken from NCBI gene_orthologs restricted to taxa 9606 and 10090, joined to gene symbols and Ensembl identifiers through NCBI gene_info.

Raw read processing

The first five million reads of each run were taken in file order, which is flowcell-tile order and unrelated to genomic coordinate. Residual adapter was removed with cutadapt (TGGAATTCTCGGGTGCCAAGG, minimum length 15, quality cutoff 20, no ambiguous bases); the deposited reads proved to be already trimmed. Alignment used bowtie2 in local mode with parameters permitting mismatches in the seed (-N 1 -L 16 --score-min G,15,6), chosen so that T-to-C conversions do not cause alignment failure, followed by retention of uniquely aligned primary alignments (samtools view -q 10 -F 0x904).

Conversion-site calling and regional classification

For each aligned read, reference bases over the aligned span were compared with read bases. A T-to-C event was recorded where the reference base was T and the read base C

on the forward strand, or A and G respectively on the reverse strand; all other mismatches were recorded as the comparison class. Regional labels were assigned at base resolution by painting per-chromosome, per-strand arrays with gene spans, then exons, then CDS, so that the precedence is CDS, non-coding exon, intron, intergenic. Mitochondrial and unplaced sequence were excluded.

Background models

Two backgrounds were used. Within-gene shuffling placed each site at a uniformly random position within its assigned gene, preserving strand, with 5 to 20 shuffles depending on set size. The read-matched background drew one uniformly random thymidine position (adenine on the reverse strand) from the aligned span of each read, so that the background reflects read placement, mappability and local expression. Odds ratios compare the observed site distribution with the read-matched background by Fisher exact test. The specificity ratio is the count of T-to-C events divided by the count of all other mismatches within the same region.

Filtering, gene equalisation and composition controls

Each conversion event, comparison mismatch and background position was additionally assigned to the largest annotated gene containing it on the same strand, or recorded as unassigned. Structured non-coding RNA was identified by gene biotype (rRNA, tRNA, snRNA, snoRNA, scaRNA, misc_RNA, scRNA, sRNA, vault_RNA, telomerase_RNA, ribozyme, RNase_P_RNA, RNase_MRP_RNA, Y_RNA, srpRNA and the mitochondrial equivalents) and, for pseudogene copies of these classes, by symbol prefix (RNA5S, RNA5-8S, RNA18S, RNA28S, RNA45S, RN7SL, RN7SK, RNU, RNY, TRNA, RMRP, RPPH1, VTRNA, SNOR, SCARNA). Gene equalisation capped each gene's contribution to each class of event at a fixed number, applied identically to conversions, comparison mismatches and background positions; caps of 10, 50 and 100 were reported alongside the uncapped tallies. Local sequence composition was controlled by computing the GC content of a 51-nt window centred on each position from the reference sequence, binning it at 0.25, 0.35, 0.45, 0.55, 0.65 and 0.75, computing Mantel–Haenszel odds ratios stratified by bin, and fitting a binomial model of site-versus-background status on region indicators and GC content together. Reproducibility between libraries was assessed by Spearman correlation of per-gene conversion counts and of per-gene specificity ratios, and by the overlap of ranked gene lists.

RIP enrichment

Counts were restricted to genes with at least 10 reads in all eight inputs (13,491 genes), converted to counts per million, and enrichment computed per pair as log2((IP + 1)/(input + 1)). Length correction used LOWESS regression of enrichment on log10 gene length (frac = 0.3) with the fitted values subtracted. Gene length, union-exon length and intronic length were computed from the GRCm39 annotation.

Uniform motif pipeline

For each dataset, all peaks on assembled chromosomes were read, and 4,000 were sampled at random without replacement using a fixed seed; datasets with fewer peaks were

used in full. A fixed window of 21 nt centred on each peak centre was extracted in transcript orientation. Background consisted of five windows of identical width placed at uniformly random positions within the same gene, same strand. Logistic regression modelled peak status on GC content scaled per 10%, binary indicators for CCCY (CCC[CT]), CGCG and GGAC, and region indicators with CDS as reference. De novo analysis counted 5-mer frequencies in foreground and background and ranked by ratio. Positional analysis used windows of ±50 nt around peak centres.

Chromatin occupancy and its comparison with RNA targets

Chromatin immunoprecipitation peaks were assigned to genes under two definitions. A gene was promoter-bound if a peak overlapped an interval of ±2 kb around any of its annotated transcription start sites, and body-bound if a peak overlapped its gene span extended by 10 kb on each side. A gene was called occupied only if it met the definition in both samples. PAR-CLIP peaks were assigned to the gene containing the peak centre on the same strand, and classified as exonic if the centre fell within a CDS or an exon on that strand and intronic otherwise. The comparison was restricted to protein-coding and long non-coding genes with a HEK293 RPKM value, and then to genes expressed at or above 1 RPKM; thresholds of 0.5 and 5 were also examined. Enrichment was assessed by Fisher exact test, by the Mantel–Haenszel common odds ratio across quintiles of HEK293 expression and separately across quintiles of gene length, and by logistic regression of occupancy on exonic target status with log10 HEK293 expression and log10 gene length as covariates.

Peak overlap

Overlap was computed as the proportion of SOX2 peaks intersecting at least one peak of the comparison set on the same chromosome and strand. The null expectation was obtained by shifting all comparison intervals by a random offset of 50–100 kb per chromosome and strand, preserving interval length, count and relative spacing, averaged over five shifts.

SELEX reanalysis

The 51 sequences were parsed from the supplementary PDF and verified to be 25 nt each, corresponding to the randomised region only. Background sets of 20,000 sequences were generated either with the base composition of the selected pool or uniformly, under a fixed random seed; motif folds move in the second decimal place with the seed, while the clustering and the Hamming comparison do not depend on it. Motif frequencies were compared by exact binomial test against the composition-matched background. Independence was assessed by pairwise Hamming distance against composition-matched random sequence (Mann–Whitney U) and by single-linkage clustering at a distance threshold of 8.

Statistics

Correlations are Spearman unless stated otherwise. Two-by-two comparisons used Fisher exact tests, and stratified comparisons the Mantel–Haenszel common odds ratio with a test of homogeneity; with the event counts involved, that homogeneity test rejects at neg-

ligible differences and the stratum-specific values are reported alongside the pooled estimate. Confidence intervals are 95%. Analyses used Python with pandas, numpy, scipy, statsmodels and pysam, with bowtie2, samtools and cutadapt for read processing.

Use of large language models

The analysis scripts and the draft text of this manuscript were prepared with the assistance of a large language model (Claude, Anthropic). The author set the questions, decided which controls and comparisons to run, inspected the output of every analysis step, and checked each number reported here against the script outputs archived at Zenodo. The author takes full responsibility for the content.

Tables

Table 1. Regional distribution of SOX2 PAR-CLIP T-to-C conversions, before and after removing non-specific classes and equalising genes. Specificity ratio, T-to-C events divided by all other mismatches in the same region; OR, odds ratio against the read-matched background. "All events" uses every conversion on assembled chromosomes. "Filtered" removes structured non-coding RNA and its pseudogenes and events outside annotated genes. "Filtered, capped" additionally limits each gene to ten events per class, so that no locus dominates. The exonic-to-intronic contrast falls from roughly thirtyfold to roughly fourfold, and the odds ratio against read placement is modest in every column.

Rep	Region	Specificity ratio: all events	Filtered	Filtered, cap 100	Filtered, cap 10	OR vs background: all	Filtered, cap 10
1	CDS	29.5	29.5	26.4	19.0	1.739	1.111
1	Non-coding exon	5.0	8.1	16.0	16.5	1.564	1.163
1	Intron	0.98	1.29	3.66	4.14	0.976	0.814
2	CDS	23.9	23.9	24.4	15.0	1.531	1.078
2	Non-coding exon	7.7	17.1	20.5	12.9	1.468	1.087
2	Intron	1.56	2.66	5.57	3.54	1.075	0.881

Table 2. Composition of the T-to-C signal and agreement between the two libraries. The upper block is per library; the lower block reports agreement between them.

Quantity	Replicate 1	Replicate 2
Conversion events on assembled chromosomes	691,254	915,494
In structured non-coding RNA or its pseudogenes	31.5%	21.0%
Outside any annotated gene	14.6%	20.6%
Genes accounting for half of all events	11	138
Genes accounting for half, after filtering	34	1,631
Ten highest-count genes, after filtering	40.9%	10.5%

Agreement between libraries	Genes compared	Spearman rho
Conversion counts per gene, all genes detected	20,082	+0.811
Conversion counts per gene, ≥ 10 events	10,506	+0.704
Conversion counts per gene, ≥ 100 events	960	+0.405
Specificity ratio per gene	381	+0.846
Overlap of the 100 highest-count genes	—	65 of 100

Table 3. Exonic PAR-CLIP targets are not enriched among SOX2-occupied genes. Genes with HEK293 expression of at least 1 RPKM (n = 10,213). Occupancy requires a peak in both HUES64 chromatin immunoprecipitation samples. Mantel–Haenszel odds ratios are stratified by quintiles of HEK293 expression; the logistic model adjusts for log10 HEK293 expression and log10 gene length simultaneously.

Exposure	Occupancy definition	Occupied genes	Estimate	OR	95% CI	p
Exonic peak in both replicates	Promoter ±2 kb	200	Mantel–Haenszel	0.944	0.627–1.422	0.78
Exonic peak in both replicates	Gene body ±10 kb	1,655	Mantel–Haenszel	0.879	0.753–1.027	0.11
Exonic peak in either replicate	Promoter ±2 kb	200	Mantel–Haenszel	1.021	0.678–1.537	0.92
Any PAR-CLIP peak	Promoter ±2 kb	200	Mantel–Haenszel	1.049	0.661–1.664	0.84
Intronic peak in either replicate	Promoter ±2 kb	200	Mantel–Haenszel	0.792	0.544–1.152	0.22
Exonic peak in both replicates	Promoter ±2 kb	200	Logistic, adjusted	0.920	0.610–1.389	0.69
log10 gene length (same model)	Promoter ±2 kb	200	Logistic, adjusted	1.525	1.087–2.141	0.015

Figure legends

Figure 1. The two published SOX2 datasets recover different RNA pools. (A) Enrichment versus log10 gene length for fRIP, UV-RIP and PAR-CLIP, with LOWESS fits.

(B) Forest plot of length-bias Spearman coefficients ordered by the fraction of transcript each method recovers. (C) fRIP enrichment versus intronic fraction before and after length correction. (D) PAR-CLIP peak score versus intronic fraction.

Figure 1 The two published SOX2 datasets recover different RNA pools

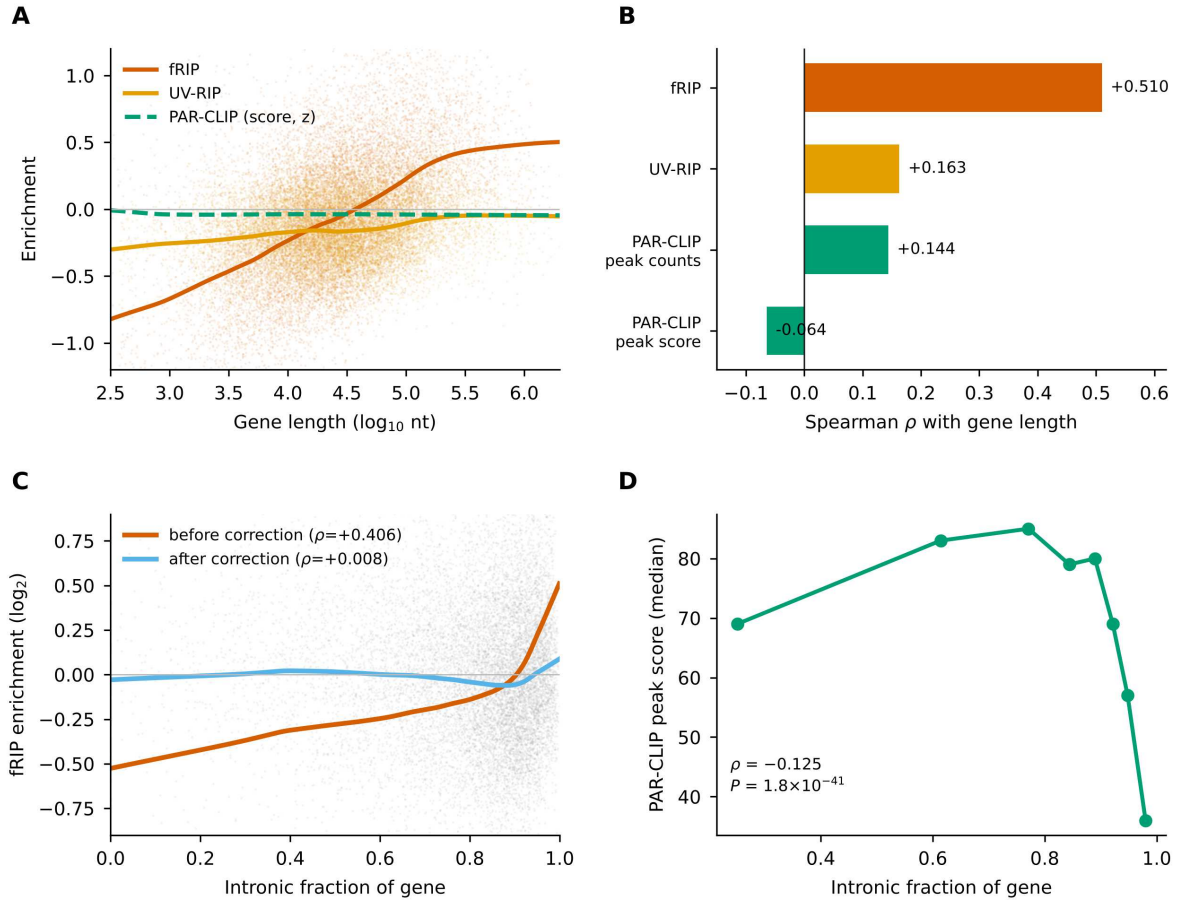


Figure 2. The two datasets do not identify the same targets, and species and cell type do not explain it. (A) Overlap of top fRIP and UV-RIP targets at three thresholds, with chance expectation. (B) Odds ratios for RIP consensus targets among PAR-CLIP-bound genes by gene-length quintile, with Mantel–Haenszel summary. (C) The same odds ratios under increasing expression thresholds in both cell types. (D) Distribution of HEK293 expression for RIP consensus targets and non-targets.

Figure 2 The two datasets do not identify the same targets

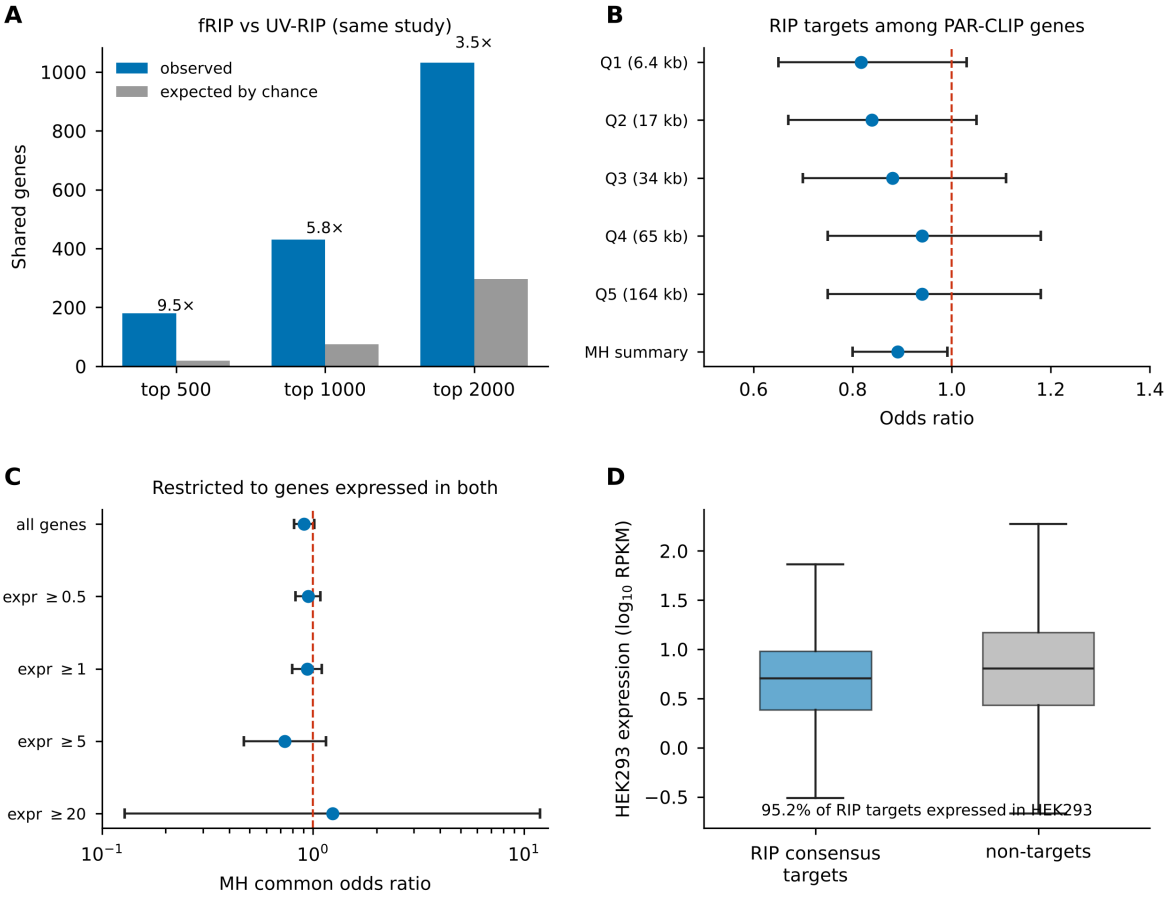


Figure 3. Reported sequence features are not distinguishable from control RBPs, and the strongest de novo motif belongs to a different protein. (A) Forest plot of GC- and region-adjusted odds ratios for CCCY across twelve PAR-CLIP datasets processed through one pipeline; SOX2 in blue, FXR1, FXR2 and FMRP in green, ZFP36 (positive control) in pink and MATR3 (null control) in light blue. (B) The same for GGAC. (C) Positional profile of GGAC around SOX2 replicate-2 peak centres, against a within-gene shuffled background; the shaded band marks ±5 nt. (D) Proportion of SOX2 replicate-2 peaks overlapping the peaks of each control dataset, against the expectation from coordinate-shifted controls. Odds ratios for CGCG and for GC content are given in Supplementary Figure S1.

Figure 3 Reported sequence features are not distinguishable from control RBPs

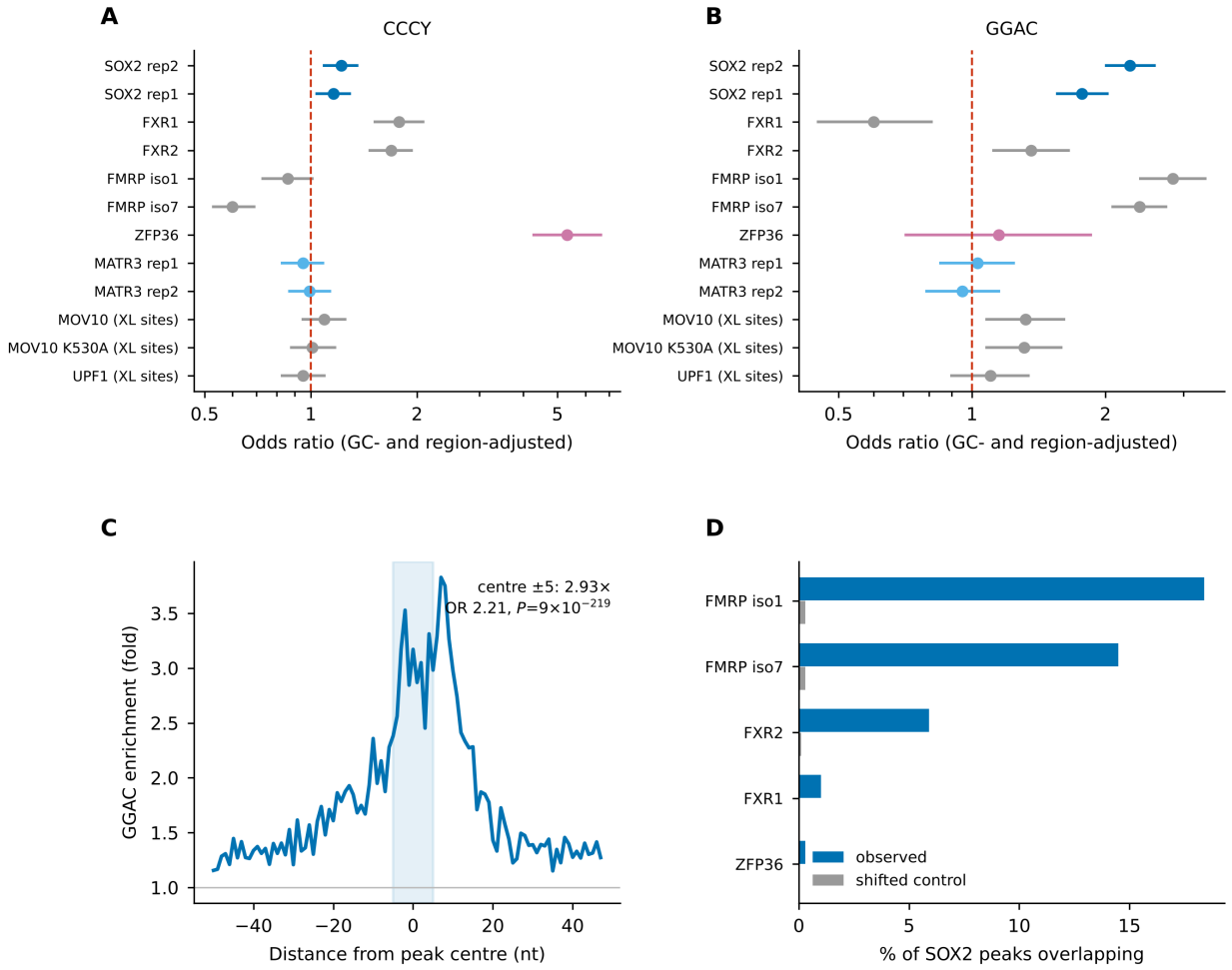


Figure 4. Reanalysis of the published SELEX pool. (A) Base composition of the 51 selected sequences against the starting library expectation. (B) Motif frequencies against uniform and composition-matched backgrounds. (C) Pairwise Hamming distance distribution against composition-matched random sequence. (D) Single-linkage clustering of the 51 sequences.

Figure 4 The published SELEX core motif is explained by pool GC content

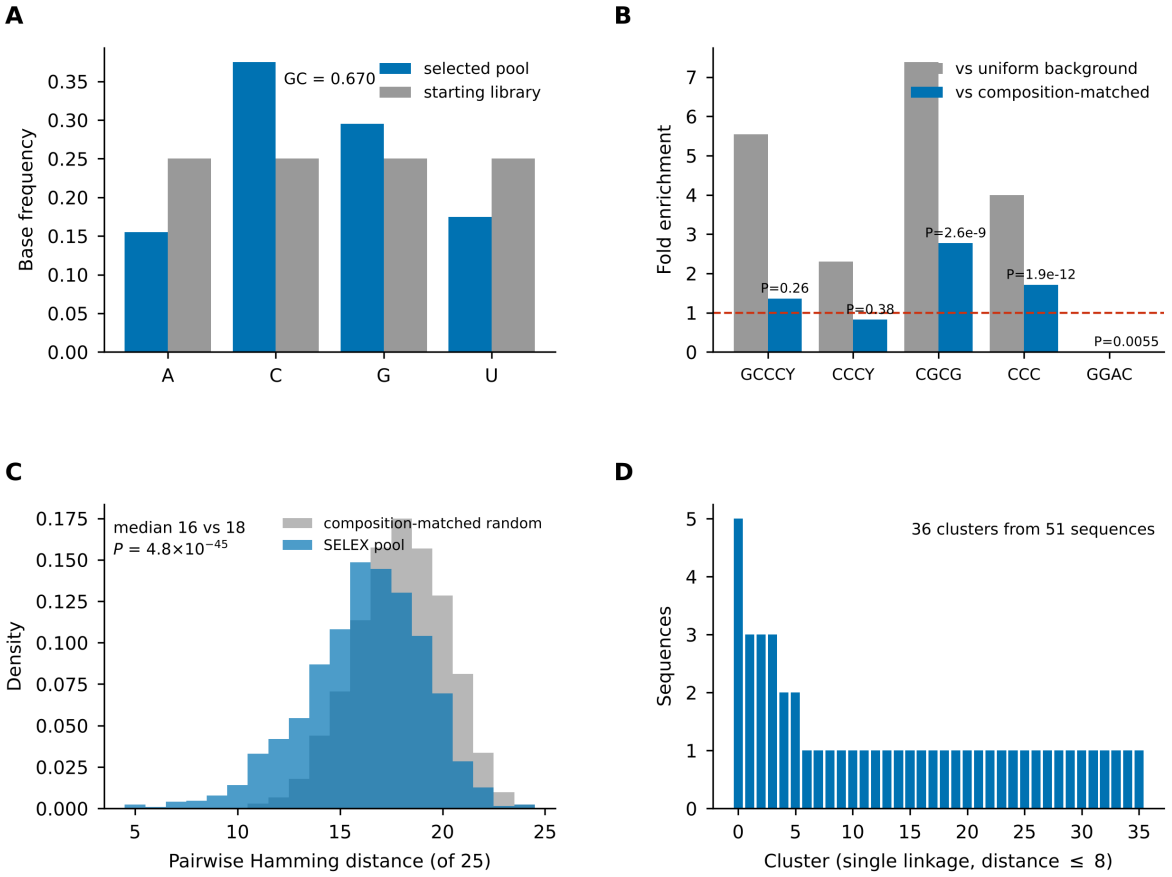
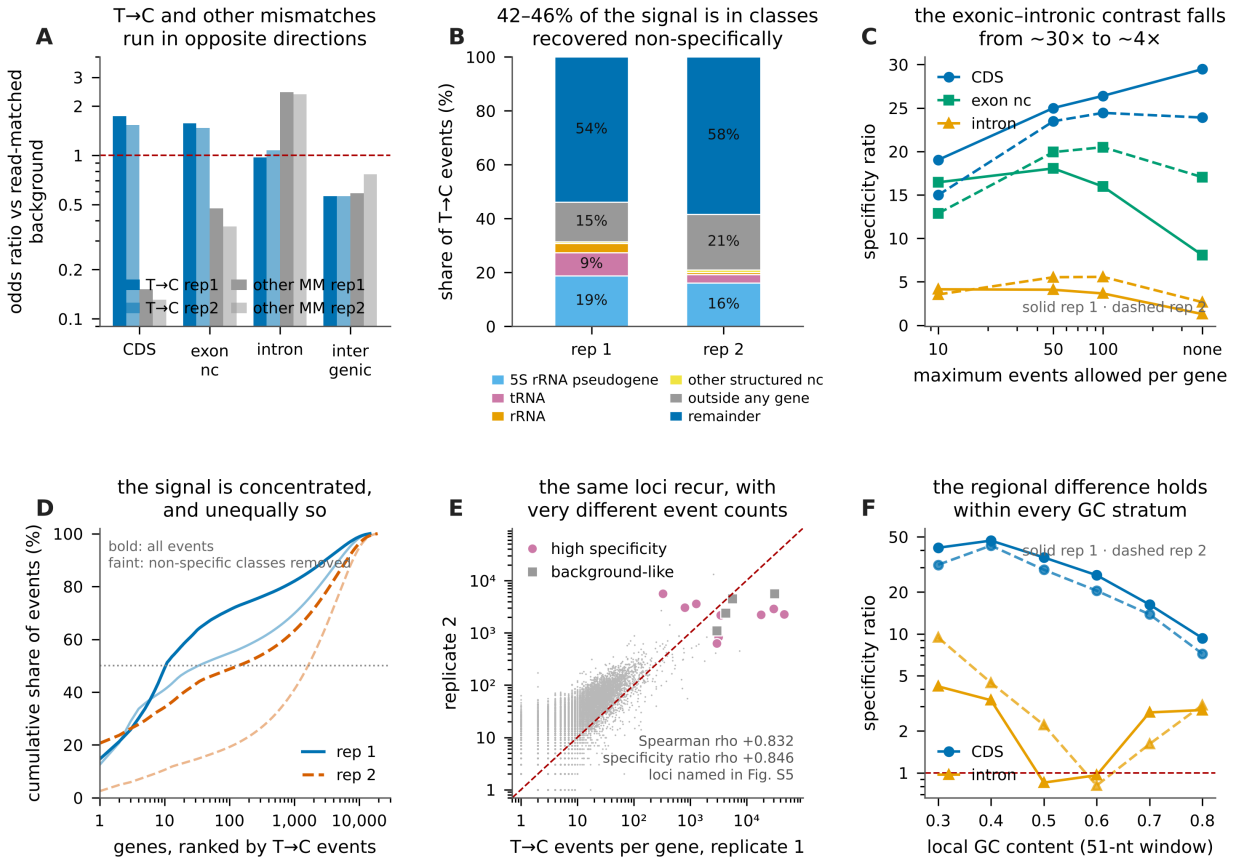


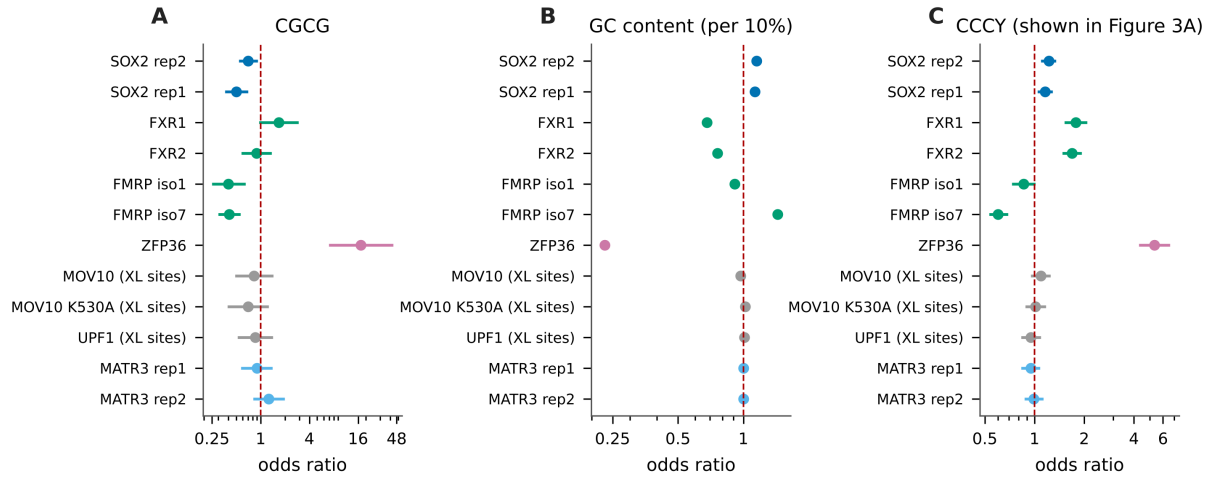
Figure 5. The exonic preference is modest once non-specific classes are removed and genes are equalised. (A) Regional odds ratios for T-to-C conversions and for all other mismatches against the read-matched background, both replicates, showing opposite directions. (B) Composition of the conversion signal: structured non-coding RNA by class, events outside annotated genes, and the remainder, for both replicates. (C) Specificity ratio by region as a function of the per-gene cap, after removal of structured non-coding RNA and unassigned events, showing the fall of the exonic-to-intronic contrast from roughly thirtyfold to roughly fourfold. (D) Cumulative share of conversion events against genes ranked by event count, before (bold) and after (faint) removing the non-specific classes; the dotted line marks 50%. (E) Conversion counts per gene between replicates, with loci of high and of background-like specificity ratio highlighted; the diagonal is equality. Names are given in Supplementary Figure S5. (F) Specificity ratio by region within strata of local GC content, showing that sequence composition does not account for the regional difference; computed on all events without a per-gene cap.



Supplementary figure legends

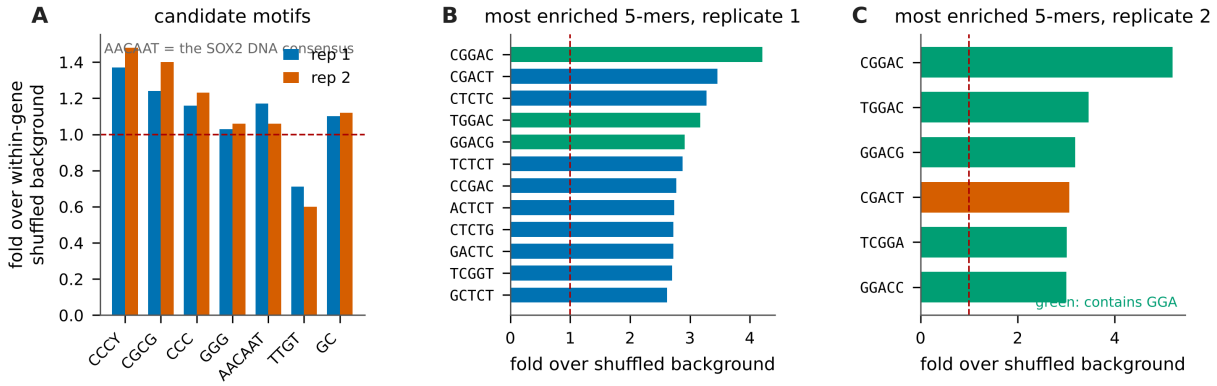
Supplementary Figure S1. Sequence-feature odds ratios across twelve PAR-CLIP datasets. (A) CGCG and (B) GC content per 10%, from the same GC- and region-adjusted logistic model as Figure 3; (C) CCCY, repeated from Figure 3A for comparison. Colours as in Figure 3. CGCG varies in both direction and magnitude across the control datasets, so the SOX2 values cannot be read against a stable baseline. The GC panel shows the cluster-level estimate only; the opposing site-level estimate is reported in the text, and neither is interpretable as a sequence preference.

Supplementary Figure S1 Sequence-feature odds ratios across twelve PAR-CLIP datasets



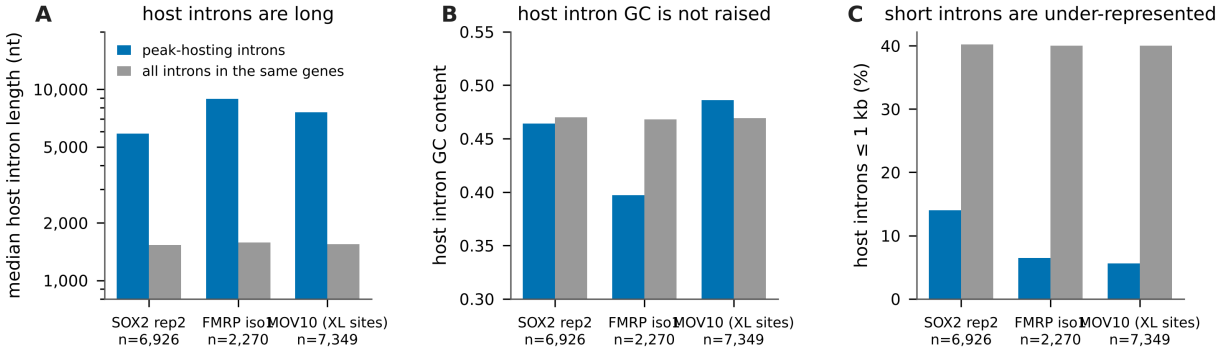
Supplementary Figure S2. Motif enrichment in SOX2 PAR-CLIP peaks. (A) Fold enrichment of candidate motifs over a within-gene shuffled background, both replicates. AACAAT, the SOX2 DNA consensus, is included and is not enriched. (B, C) The most enriched 5-mers in each replicate; those containing GGA are coloured green. The ranking is headed by GGAC-containing 5-mers in both replicates, above any CCCY-containing 5-mer.

Supplementary Figure S2 Motif enrichment in SOX2 PAR-CLIP peaks



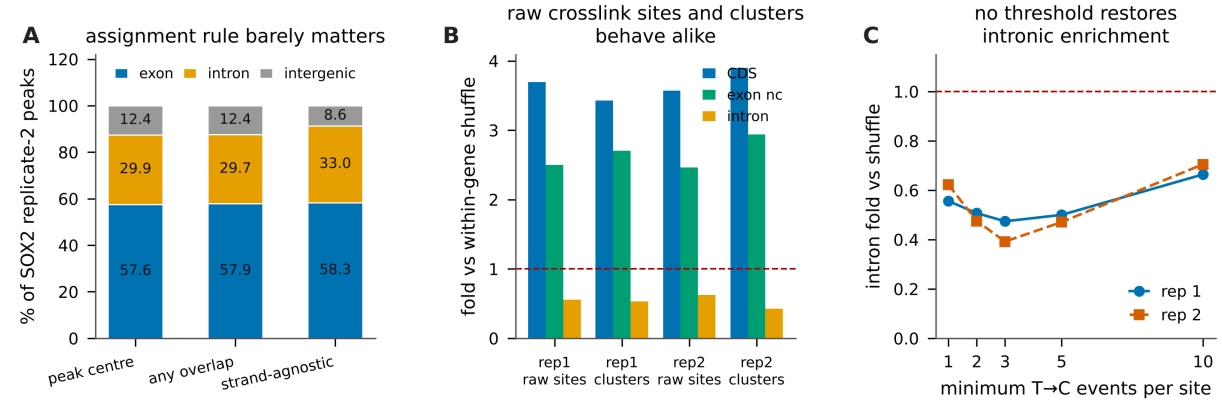
Supplementary Figure S3. Intronic peaks lie in long introns of ordinary GC content. For intronic peaks of SOX2, FMRP isoform 1 and MOV10, the host introns are compared with all introns of the same genes: (A) median host intron length, (B) host intron GC content, (C) percentage of host introns of 1 kb or less. Intronic peaks are not concentrated in the short, GC-rich introns that would confound a GC-based analysis; host intron GC is indistinguishable from background for SOX2 (0.464 versus 0.470, p = 0.4).

Supplementary Figure S3 Intronic peaks lie in long introns of ordinary GC content



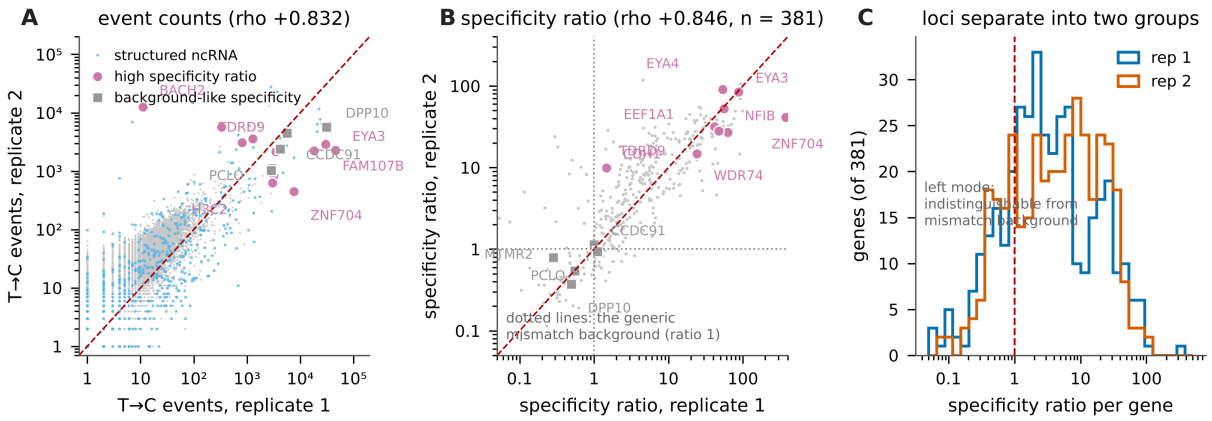
Supplementary Figure S4. Robustness of the regional analysis. (A) Regional proportions of SOX2 replicate-2 peaks under three assignment rules. (B) Fold enrichment against a within-gene shuffled background for raw T-to-C sites and for deposited PARalyzer clusters, both replicates, showing that clustering does not create the regional pattern. (C) Intronic fold as a function of the minimum number of conversion events required per site; no threshold moves the value towards intronic enrichment.

Supplementary Figure S4 Robustness of the regional analysis



Supplementary Figure S5. Per-gene reproducibility between the two SOX2 PAR-CLIP libraries. (A) T-to-C events per gene, replicate 1 against replicate 2; structured non-coding RNA in light blue, loci of high specificity ratio in pink, loci of background-like specificity in grey. Counts correlate (Spearman rho +0.832 after excluding unassigned events and structured non-coding RNA) but the highest-count loci lie far from the diagonal, differing between libraries by up to twentyfold. (B) Specificity ratios for the 381 genes with at least twenty conversions and five other mismatches in both libraries; these correlate more tightly (rho +0.846). Dotted lines mark the generic mismatch background. (C) Distribution of per-gene specificity ratios in both libraries. The distribution is bimodal: one group of loci sits at or below the mismatch background and another an order of magnitude above it, and the two libraries agree on which loci fall where. Without a sequenced negative control, the upper group identifies what this experiment recovered cleanly, not what SOX2 is shown to bind.

Supplementary Figure S5 Per-gene reproducibility between the two SOX2 PAR-CLIP libraries



Supplementary tables

Supplementary Table S1. Datasets analysed. The column that matters most for interpretation is the last: no PAR-CLIP dataset analysed here, including the SOX2 dataset, contributes a sequenced negative immunoprecipitation.

Accession	Protein / samples	Cells	Assay	Unit analysed	n	Sequenced negative control
GSE115273	SOX2, two replicates (transfected)	HEK293T	PAR-CLIP	PARalyzer clusters; raw reads SRR7253210, SRR7253211	16,964 and 55,830 clusters	None (the FLAG-GFP control was assayed by blotting only)
GSE39682	FMRP isoform 1 and 7, FXR1, FXR2	HEK293 (Flp-In T-REx, per ref 5)	PAR-CLIP	clusters	4,000 sampled per dataset	None deposited
GSE53184	ZFP36	HEK293 (Flp-In T-REx)	PAR-CLIP	clusters	4,000 sampled	None deposited
GSE48245	MOV10, MOV10 K530A, UPF1	HEK293	PAR-CLIP	T-to-C conversion sites	3,910–3,936 sampled	None deposited
GSE262647	MATR3, two replicates	HEK293 (deposit states "Kidney")	PAR-CLIP	clusters (133,089 total)	3,935 and 3,945 sampled	None deposited
GSE142013	SOX2 fRIP and UV-RIP	mouse embryonic stem cells	RIP	gene-level UMI counts	8 matched IP/input pairs	matched inputs (not an IP control)
GSE61475	SOX2 chromatin occupancy	HUES64 human ES cells	ChIP-seq	peaks	10,624 and 6,211	not applicable
GSE122425	HEK293 transcriptome	HEK293	RNA-seq	gene-level RPKM	3 libraries	not applicable
GSE162356	m6A sites	HEK293	single-base m6A	sites	12,234 sites in 5,895 genes	not applicable
—	SELEX pool	in vitro, His6-SOX2	SELEX	25-nt randomised region	51 unique sequences	not applicable

One further PAR-CLIP series (GSE84134) was processed and excluded because the deposited file names did not permit unambiguous assignment of the immunoprecipitated protein.

Supplementary Table S2. Regional specificity ratios and odds ratios under every filter and cap. Table 1 reports a subset of these columns; the full table is provided as `filtered_capped_region.csv`. Filters: A, all events; B, structured non-coding RNA and its pseudogenes removed; C, B plus events outside annotated genes removed; D, protein-coding and long non-coding genes only. Caps limit each gene's contribution to each event class.

Data availability

All data analysed are public: GSE142013, GSE115273 (including SRR7253210 and SRR7253211), GSE39682, GSE53184, GSE48245, GSE262647, GSE162356, GSE122425, GSE61475, UCSC snp151Common, and Supplementary Table 2 of reference [3]. Analysis code, including the scripts that fetch every input and generate every figure, is archived at Zenodo (DOI: 10.5281/zenodo.24078850).

Funding

This work received no specific funding.

Competing interests

None declared.

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.
2. Holmes ZE, Hamilton DJ, Hwang T, Parsonnet NV, Rinn JL, Wuttke DS, Batey RT. The Sox2 transcription factor binds RNA. Nat Commun. 2020;11:1805.
3. Hou L, Wei Y, Lin Y, et al. Concurrent binding to DNA and RNA facilitates the pluripotency reprogramming activity of Sox2. Nucleic Acids Res. 2020;48:3869–3887.
4. Friedersdorf MB, Keene JD. Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. Genome Biol. 2014;15:R2.
5. Ascano M, Mukherjee N, Bandaru P, et al. FMRP targets distinct mRNA sequence elements to regulate protein expression. Nature. 2012;492:382–386.
6. Corcoran DL, Georgiev S, Mukherjee N, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol. 2011;12:R79.
7. Mukherjee N, Jacobs NC, Hafner M, Kennington EA, Nusbaum JD, Tuschl T, et al. Global target mRNA specification and regulation by the RNA-binding protein ZFP36. Genome Biol. 2014;15:R12.
8. Gregersen LH, Schueler M, Munschauer M, Mastrobuoni G, Chen W, Kempa S, et al. MOV10 is a 5′ to 3′ RNA helicase contributing to UPF1 mRNA target degradation by translocation along 3′ UTRs. Mol Cell. 2014;54:573–585.
9. Islam Z, Polash A, Suzawa M, Chim B, Kuhn S, Sultana S, et al. MATRIN3 deficiency triggers autoinflammation via cGAS-STING activation. bioRxiv; 2024 (preprint). Data: GSE262647.
10. Hu L, Liu S, Peng Y, Ge R, Su R, Senevirathne C, et al. m6A RNA modifications are measured at single-base resolution across the mammalian transcriptome. Nat Biotechnol. 2022;40:1210–1219.
11. Ghafoori SM, Sethi A, Petersen GF, et al. RNA binding properties of SOX family members. Cells. 2024;13:1202.

12. Girardot M, Bayet E, Maurin J, Fort P, Roux P, Raynaud P. SOX9 has distinct regulatory roles in alternative splicing and transcription. Nucleic Acids Res. 2018;46:9106–9118.
13. Van Nostrand EL, Pratt GA, Shishkin AA, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). Nat Methods. 2016;13:508–514.
14. Ambrosetti DC, Basilico C, Dailey L. Synergistic activation of the fibroblast growth factor 4 enhancer by Sox2 and Oct-3 depends on protein–protein interactions facilitated by a specific spatial arrangement of factor binding sites. Mol Cell Biol. 1997;17:6321–6329.
15. Boyer LA, Lee TI, Cole MF, Johnstone SE, Levine SS, Zucker JP, et al. Core transcriptional regulatory circuitry in human embryonic stem cells. Cell. 2005;122:947–956.
16. Oksuz O, Henninger JE, Warneford-Thomson R, Zheng MM, Erb H, Vancura A, et al. Transcription factors interact with RNA to regulate genes. Mol Cell. 2023;83:2449–2463.e13.
17. Tsankov AM, Gu H, Akopian V, Ziller MJ, Donaghey J, Amit I, et al. Transcription factor binding dynamics during human ES cell differentiation. Nature. 2015;518:344–349.