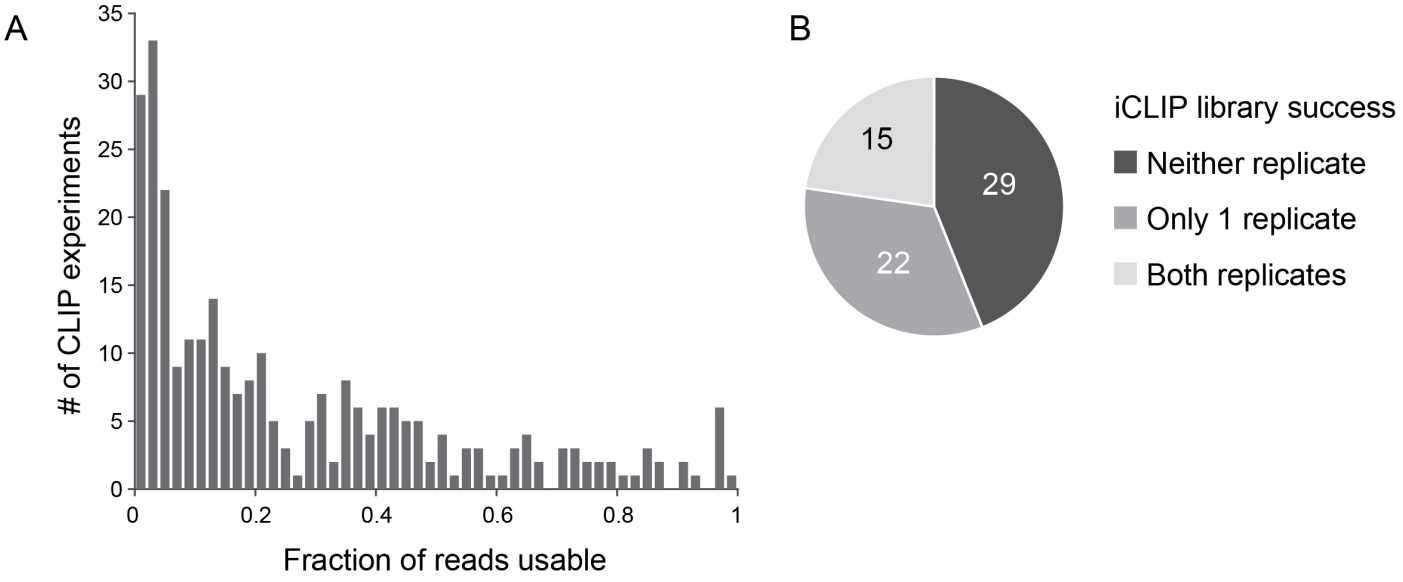
Suppementary Figure 1

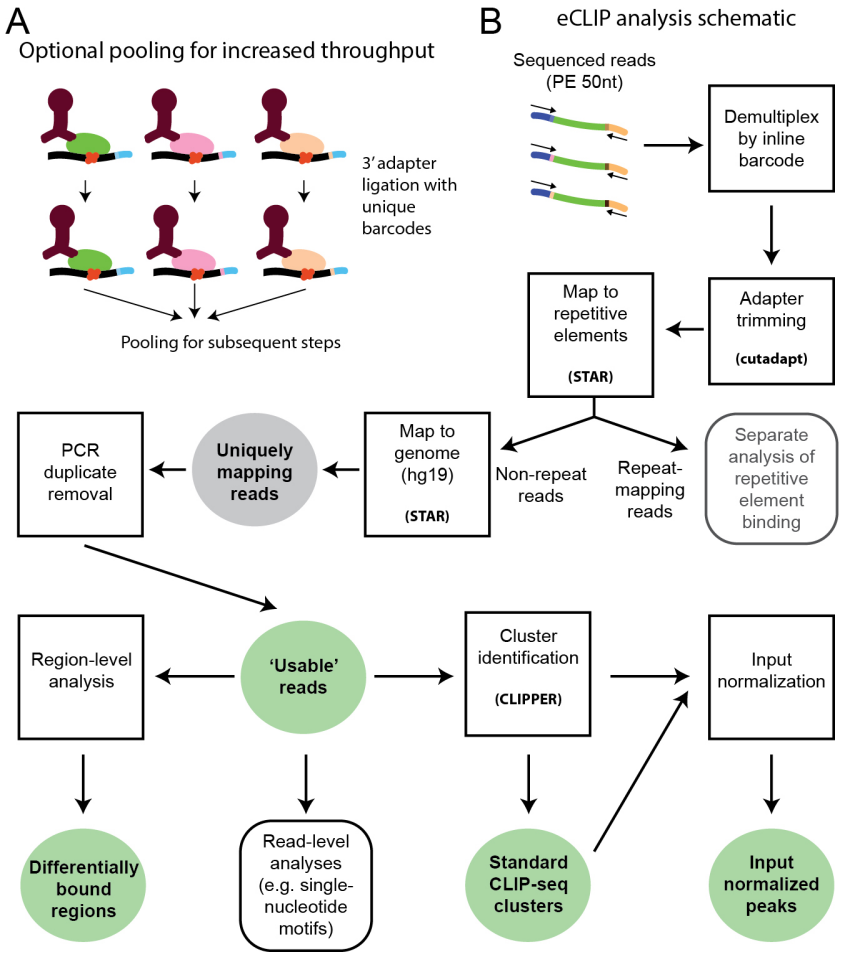


Supplementary Figure 1

Large-scale iCLIP experiments indicate poor efficiency.

(a) The fraction of usable (non-PCR duplicate, uniquely mapping) reads out of uniquely mapped reads is shown for 279 published CLIP experiments: 127 iCLIP (12 performed for the ENCODE consortium as well as 115 published) and 152 other (including PAR-CLIP and HITS-CLIP). Datasets and read-level processing statistics are listed in Supplementary Table 1. Histogram indicates the number of CLIP experiments within the indicated usable fraction bin. (b) Out of 66 iCLIP experiments performed for the ENCODE consortium, only 15 showed successful amplification of library in both biological replicates (all requiring 24-32 cycles of PCR).

Supplementary Figure 2

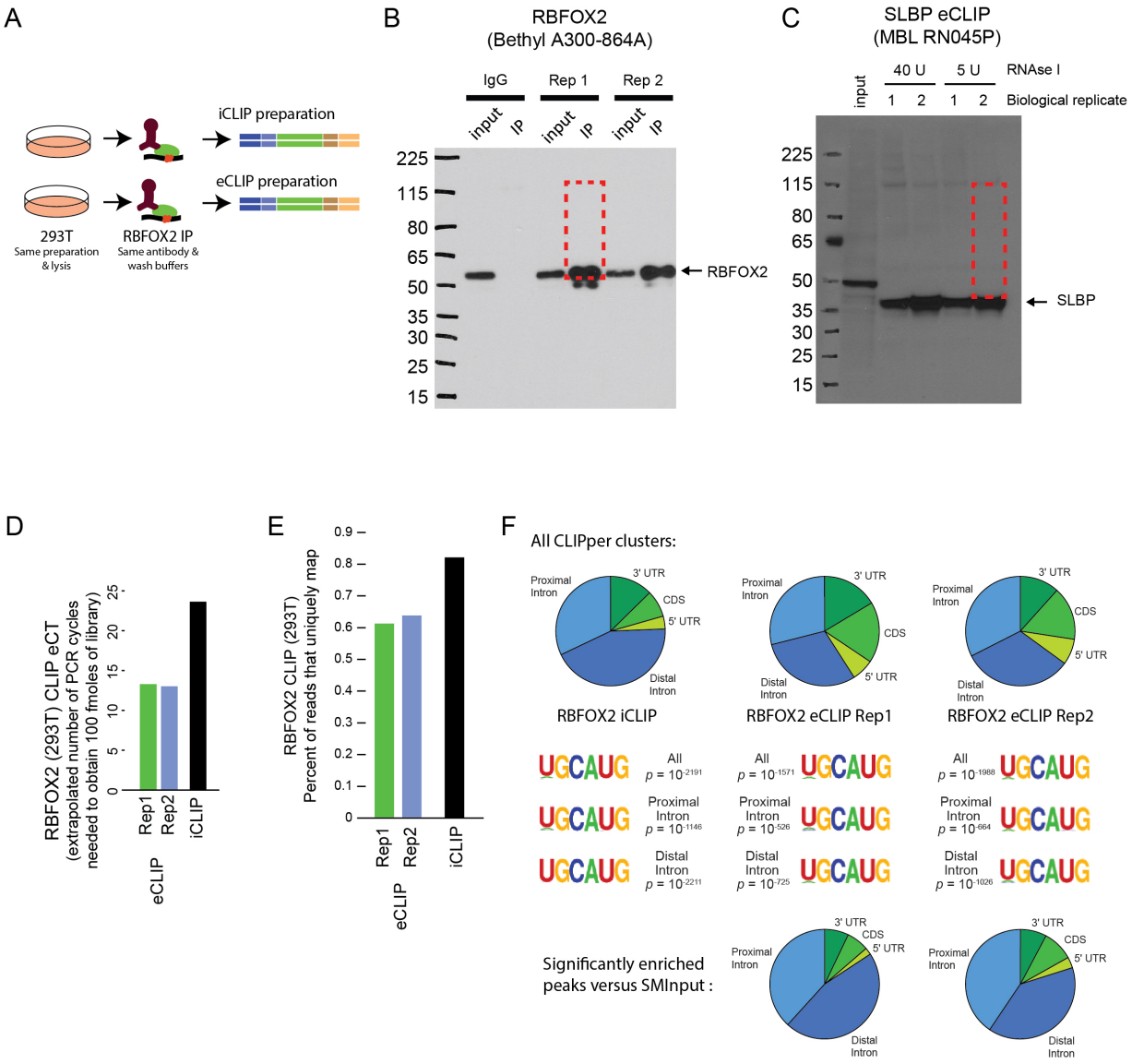


Supplementary Figure 2

Optional sample pooling strategy and eCLIP computational analysis workflow.

(a) At the 3′ RNA adapter ligation step in eCLIP, the RNA adapter includes a barcode sequence, enabling pooling of multiple experiments before the protein gel electrophoresis step. Note that pooled samples must have identical desired cut size on the nitrocellulose membrane, and should have a similar number of RNA molecules (to avoid over- or under-sequencing of individual experiments within the pooled sample). (b) Schematic of eCLIP computational analysis pipeline. Squares indicate processing steps, with processing output used for downstream analyses indicated as filled green circles. Software packages used are indicated in bold.

Supplementary Figure 3

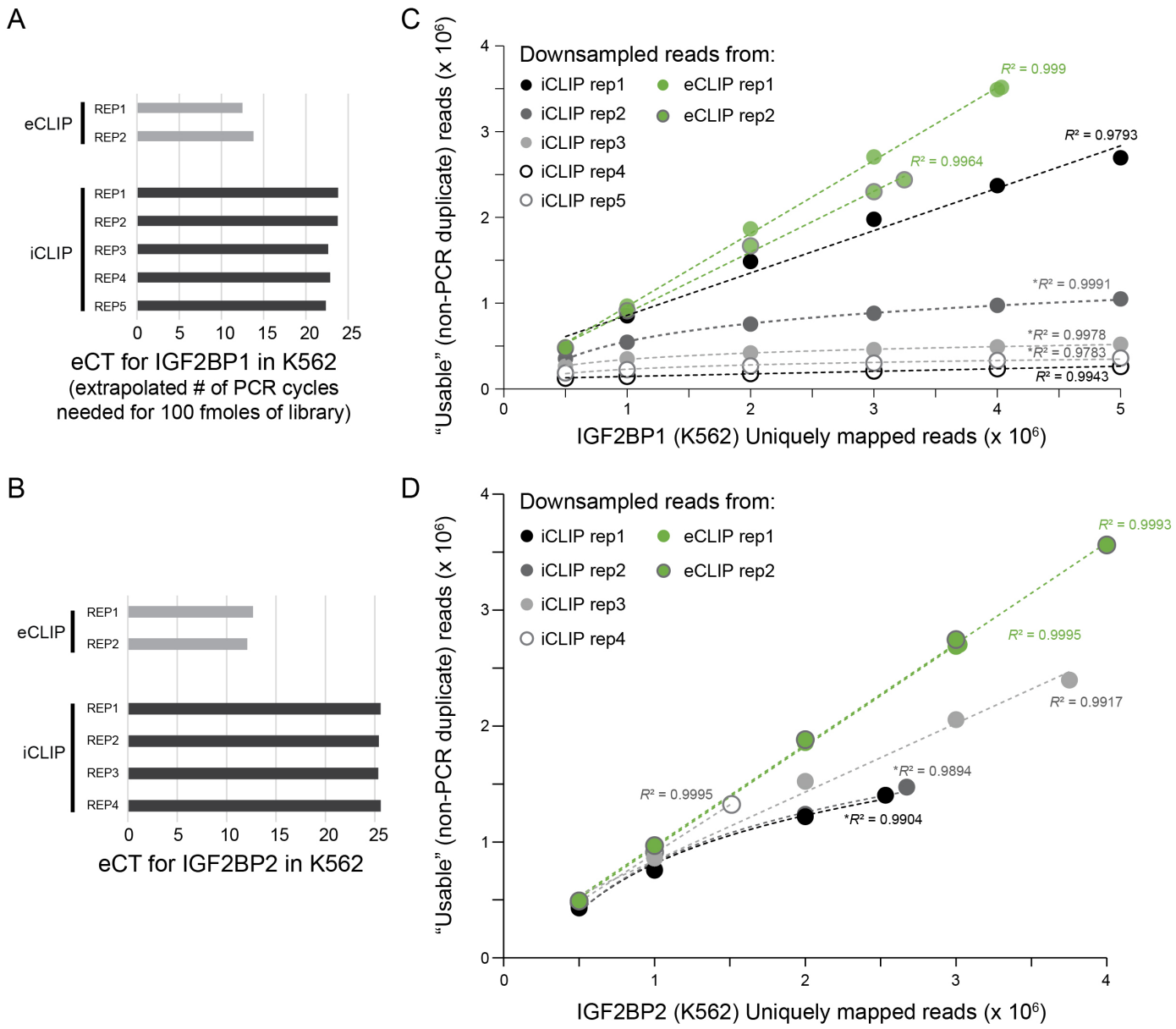


Supplementary Figure 3

eCLIP of RBFOX2 improves library efficiency over iCLIP.

(a) eCLIP and iCLIP were performed using the same RBFOX2 antibody on HEK293T cells. (b) Western blot of RBFOX2 immunoprecipitation during eCLIP. Replicates (Rep1 and Rep2) were performed on 'biological replicate' 293T samples grown and crosslinked ~2 months apart. Red dotted line indicates region excised for eCLIP library preparation. (c) Western blot of SLBP immunoprecipitation during eCLIP, performed with two concentrations (5U or 40U) of RNase I during fragmentation. (d) eCLIP requires decreased amplification compared to iCLIP. To more easily compare across samples, we defined an extrapolated cycle number (eCT) as the number of cycles needed to obtain 100 fmoles of amplified material, extrapolated from the final library volume, final library concentration, and number of PCR cycles done, assuming doubling at each cycle. (e) Fraction of reads that uniquely map to the genome is similar between iCLIP and eCLIP. (f) Peak locations (top) and *de novo* motifs identified by HOMER (middle) show similar signal between iCLIP and eCLIP. Proximal intron indicates the region ≤ 500 nt from the 5′ or 3′ splice site, with the remainder annotated as distal intron. Motifs were identified relative to a background of randomly selected regions from the same annotation class (e.g. CDS exons, proximal introns, etc). Significance indicated is as reported by HOMER. The subset of clusters significantly enriched vs SMInput (≥ 8-fold, $p \leq 10^{-5}$ by Fisher Exact or Chi Square test) show increased intronic localization for both eCLIP replicates (bottom).

Supplementary Figure 4

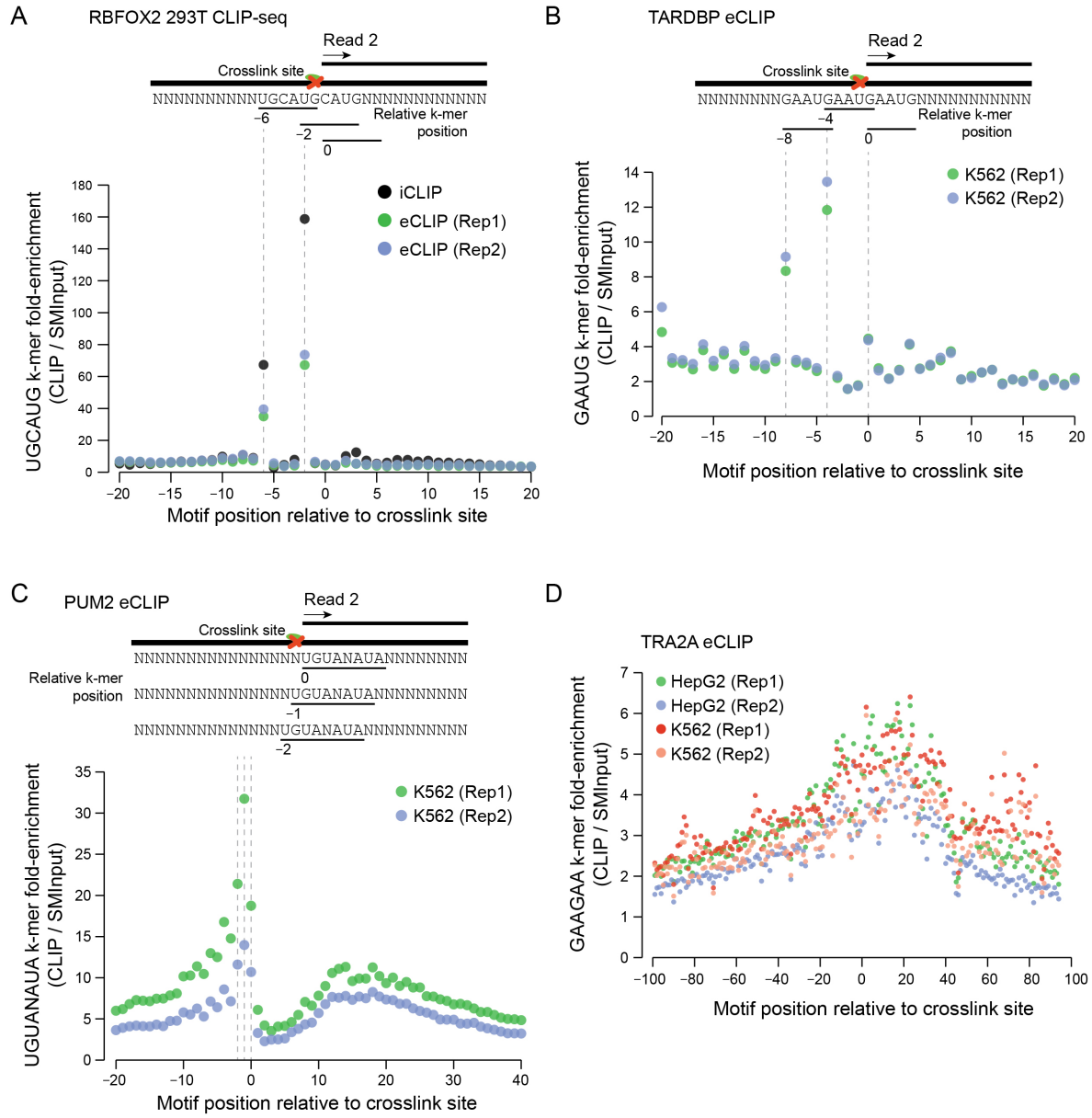


Supplementary Figure 4

eCLIP improves library efficiency over iCLIP for IGF2BP1 and IGF2BP2.

(a-b) eCT shows > 10 cycle improvement for eCLIP over iCLIP of (a) IGF2BP1 and (b) IGF2BP2 in K562 cells. (c-d) eCLIP shows improvement in the fraction of uniquely mapped reads that are usable relative to iCLIP when identical numbers of reads are downsampled from biological replicates of (c) IGF2BP1 and (d) IGF2BP2 in K562 cells. Correlation to regression (R^2) is indicated, where * indicates best fit by logarithmic regression and unlabeled indicates linear regression.

Supplementary Figure 5

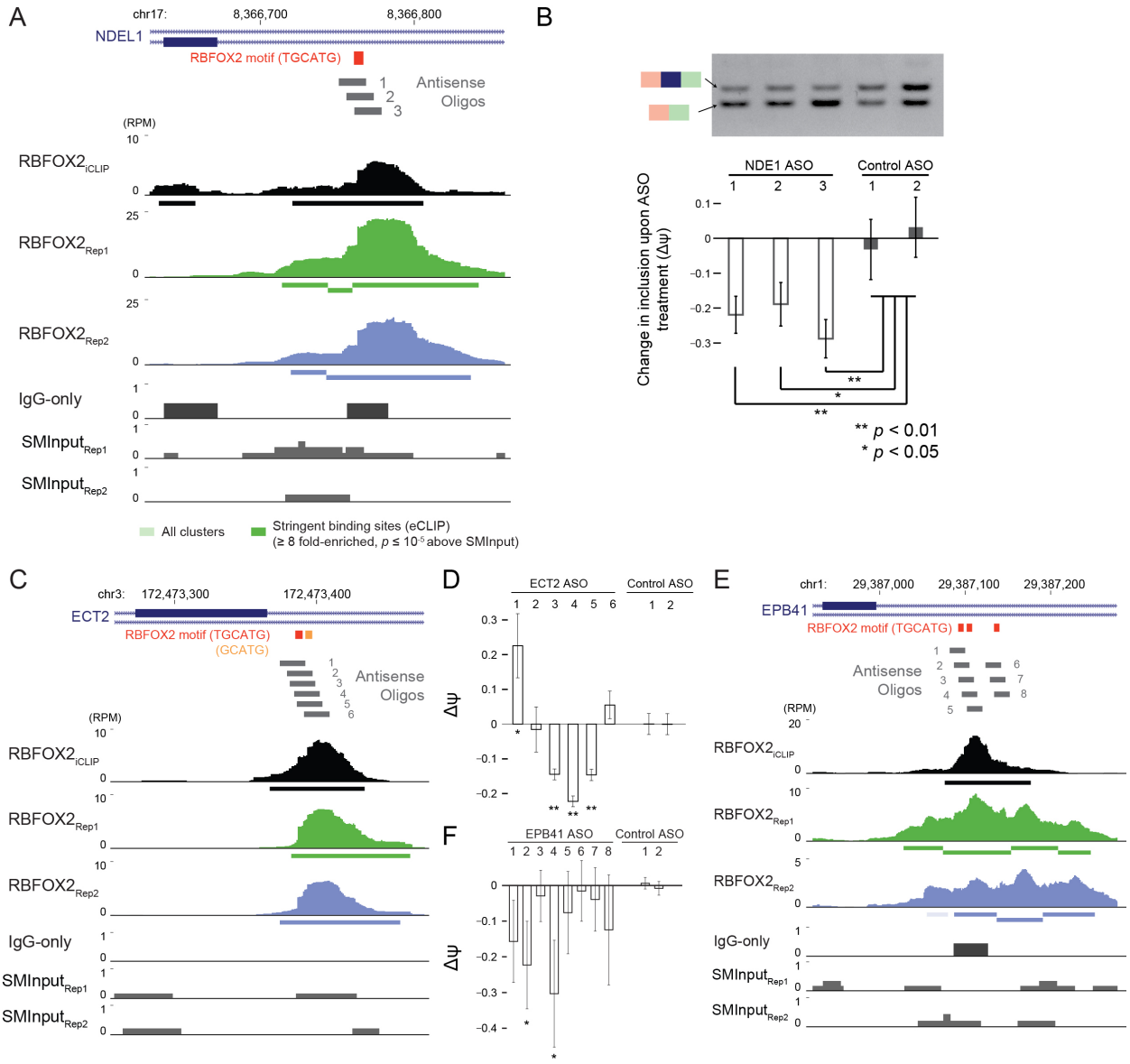


Supplementary Figure 5

Reverse transcriptase termination at crosslink sites leaves stereotypical motif frequencies flanking eCLIP sequence reads.

(a-d) Plots indicate the enrichment of indicated motifs at each position flanking the start position of mapped reads for (a) RBFOX2 eCLIP and iCLIP in 293T cells, (b) TARDBP eCLIP in K562 cells, (c) PUM2 eCLIP in K562 cells, and (d) TRA2A eCLIP in HepG2 and K562 cells. For each dataset, the frequency of the indicated kmer at each position was tallied, and compared against the frequency in paired SMInput to obtain single-nucleotide enrichment profiles. (a) RBFOX2 shows enrichment for crosslinking at G_2 and G_6 positions in both iCLIP and eCLIP, consistent with previous results. (b) TARDBP single-nucleotide profile indicates enrichment for the GAAUG at –8 and –4 nucleotides relative to read start positions. (c) PUM2 indicates UGUANAUA motif at –2, –1, and 0 relative to read start positions. (d) For TRA2A, the canonical GAAGAA motif is highly enriched around read starts but does not show specific fold-enrichment at any particular position, indicating that the majority of termination-inducing crosslinks occur at positions within the RNA that are distant from the sequence-specific site of TRA2A interaction.

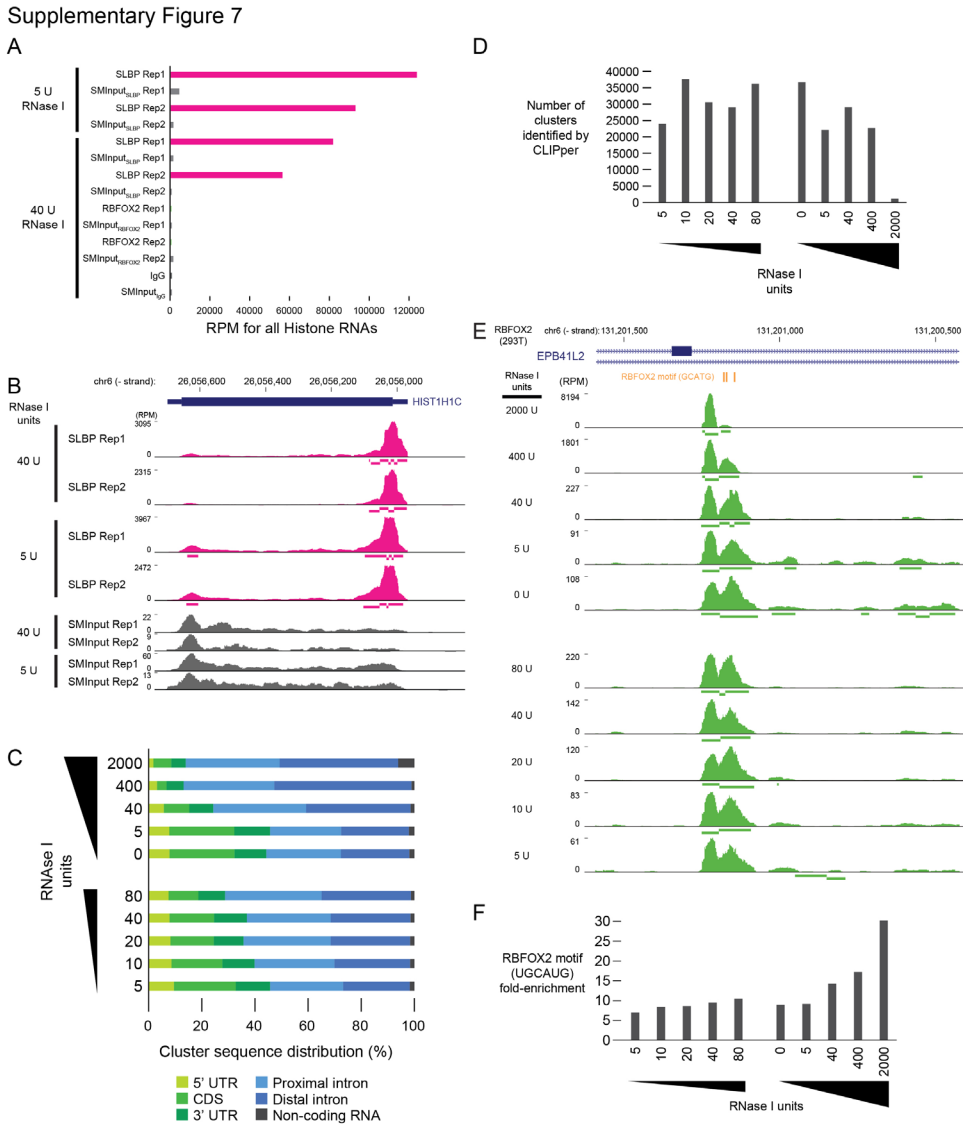
Supplementary Figure 6



Supplementary Figure 6

Functional validation of eCLIP binding sites by antisense oligonucleotide (ASO) blocking.

(a) Tracks indicate read density for iCLIP and eCLIP of RBFOX2 at an RBFOX2 binding site flanking exon 9 of *NDEL1*, and the location of three antisense oligonucleotides (with uniform 2′-*O*-methoxyethyl-modified nucleotides and a phosphorothioate backbone). Darkened bars underneath indicate peaks significantly enriched above SMInput. Read density tracks are normalized to show the number of reads per million total usable reads (RPM). (b) Treatment of 293T cells with *NDEL1*-targeting and control ASOs indicates that blocking RBFOX2 binding increases cassette exon exclusion. Asterisks denote significance determined by Student's *t*-Test performed on the change in percent spliced in (ΔΨ). (c-f) Similar analysis indicates ASO blocking of RBFOX2 binding affects splicing of (c-d) *ECT2* exon 5 and (e-f) *EPB41* exon 16.

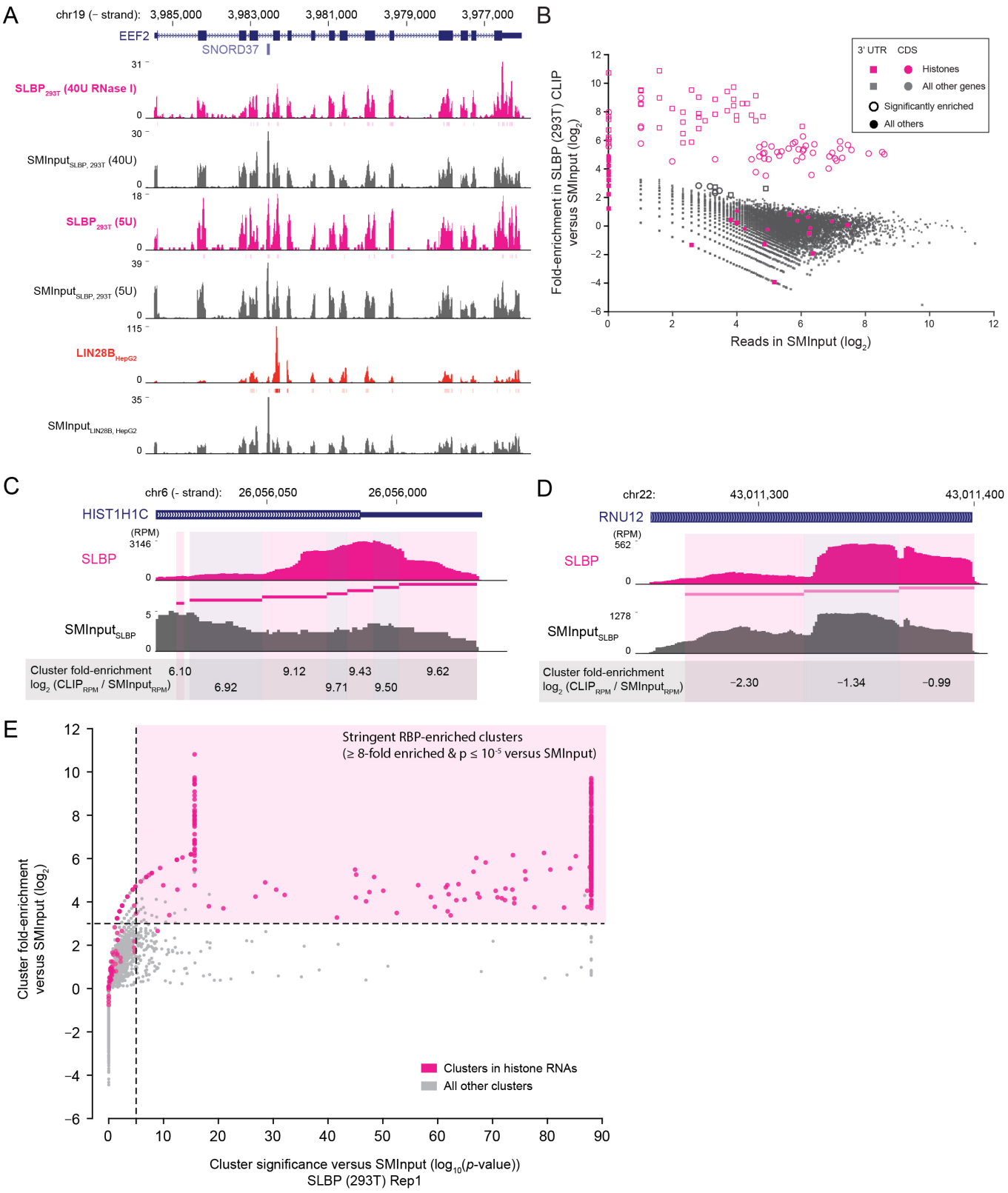


Supplementary Figure 7

Validation of standard eCLIP conditions across fragmentation conditions.

(a) Histone RNA read density is increased in 5 U RNase I digestion relative to 40 U, but both are dramatically increased above SMInput, RBFOX2, and IgG controls. See Supplementary Data for histone gene list. (b) eCLIP of SLBP shows similar enrichment at HIST1H1C 3′UTR with 40 U or 5 U of RNase I fragmentation. (c-f) Multiple analyses indicate similar eCLIP results across a range of 0-2000U RNase I fragmentation conditions for RBFOX2 eCLIP in 293T cells (c) Increased fragmentation (by increased RNase I concentration) slightly increases the fraction of intronic RBFOX2 signal relative to exonic, but intronic regions compromise the majority of bases covered across all conditions. Stacked bars indicate the fraction of bases covered by RBFOX2 clusters (identified by CLIPper) with respect to the indicated RNA transcript regions. (d) Bar graphs indicate the number of clusters identified in RBFOX2 eCLIP fragmentation experiments. Most showed 20,000-40,000 clusters, with the exception of the 2000U condition in which only 1,137 clusters were identified. (e) Read density tracks show eCLIP binding profiles flanking an RBFOX2-dependent cassette exon in *EPB41L2*. With the exception of the 2000 U condition, conditions show similar enrichment patterns and RPM coverage. (f) RBFOX2 motif (UGCAUG) enrichment in CLIPper-identified clusters increases with increasing RNase I fragmentation. Fold-enrichment shown is relative to frequency observed in ten random permutations of cluster sequences.

Supplementary Figure 8



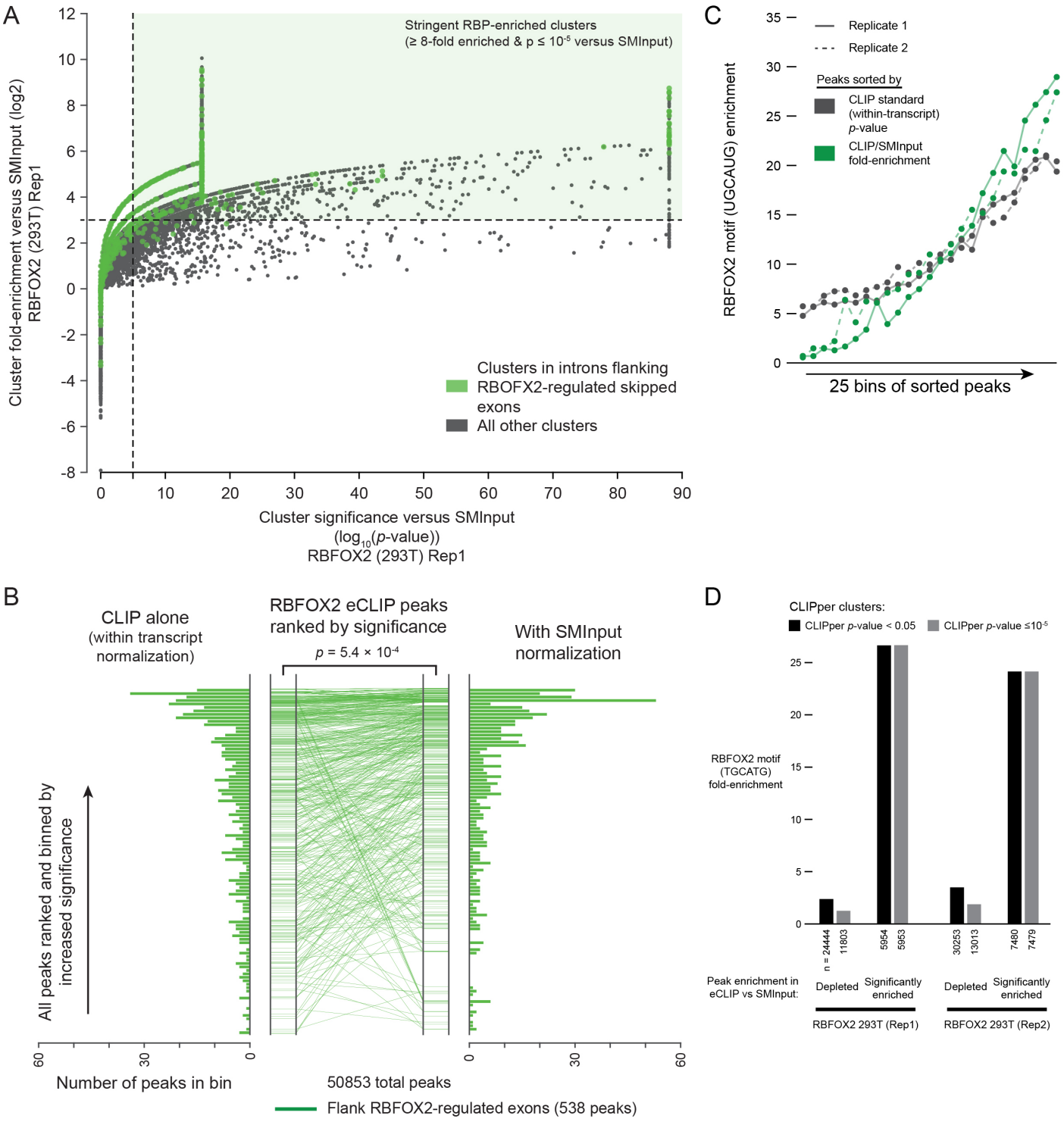
Nature Methods: doi:10.1038/nmeth.3810

Supplementary Figure 8

Paired Size-Matched Input (SMInput) reveals enrichment over common background in CLIP of histone-binding SLBP.

(a) At an abundantly expressed housekeeping gene EEF2 (Eukaryotic Translation Elongation Factor 2), similar read density is observed in eCLIP of histone-binding SLBP as in SMInput, indicating that this signal is not indicative of true binding events. Tracks below read density indicate CLIPper clusters, with darkened clusters indicating clusters significantly enriched above SMInput. Below, exonic-binding LIN28B shows significant binding to exon 5 of *EEF2*, indicating that enriched binding events can be observed above this background. (b) SLBP eCLIP shows specific enrichment for reads in histone coding exon (CDS, circles) and 3′UTR (square) regions relative to paired SMInput. Each point indicates a gene, with the x-position indicating the number of reads observed in SMInput (plus a pseudocount of 1) and the y-position indicating the fold-enrichment in SLBP 293T eCLIP (Rep1). Histone genes are indicated in pink. Significantly enriched regions (fold-enrichment ≥ 4-fold, p-value ≤ 10^{-5} in eCLIP vs SMInput) are indicated by open shapes (Significance is determined by Yates' Chi-Square test, with Fisher's Exact tests when eCLIP or SMInput has < 5 reads). (c-d) Read density (normalized as reads per million (RPM)) is shown for eCLIP of histone processing factor SLBP, along with paired SMInput, for SLBP-enriched target *HIST1H1C* and non-enriched U12 snRNA transcript *RNU12*. Rectangles below SLBP read density track indicate clusters identified with the CLIPper peak identification algorithm, with fold-enrichment in eCLIP indicated below. (e) All CLIPper-identified clusters identified for SLBP 293T eCLIP (Rep1) are plotted based on their fold-enrichment and significance compared to paired SMInput. Significance is determined by Yates' Chi-Square test, with Fisher's Exact tests (minimum p-value = 2.2×10^{-16}) when eCLIP or SMInput has < 5 reads. Only 284 clusters (1.2%) are enriched at least 8-fold with $p \leq 10^{-5}$ by Fisher Exact or Chi Square test in eCLIP (pink shaded box). Clusters overlapping histone genes (indicated in pink) are shifted towards high significance and fold-enrichment.

Supplementary Figure 9

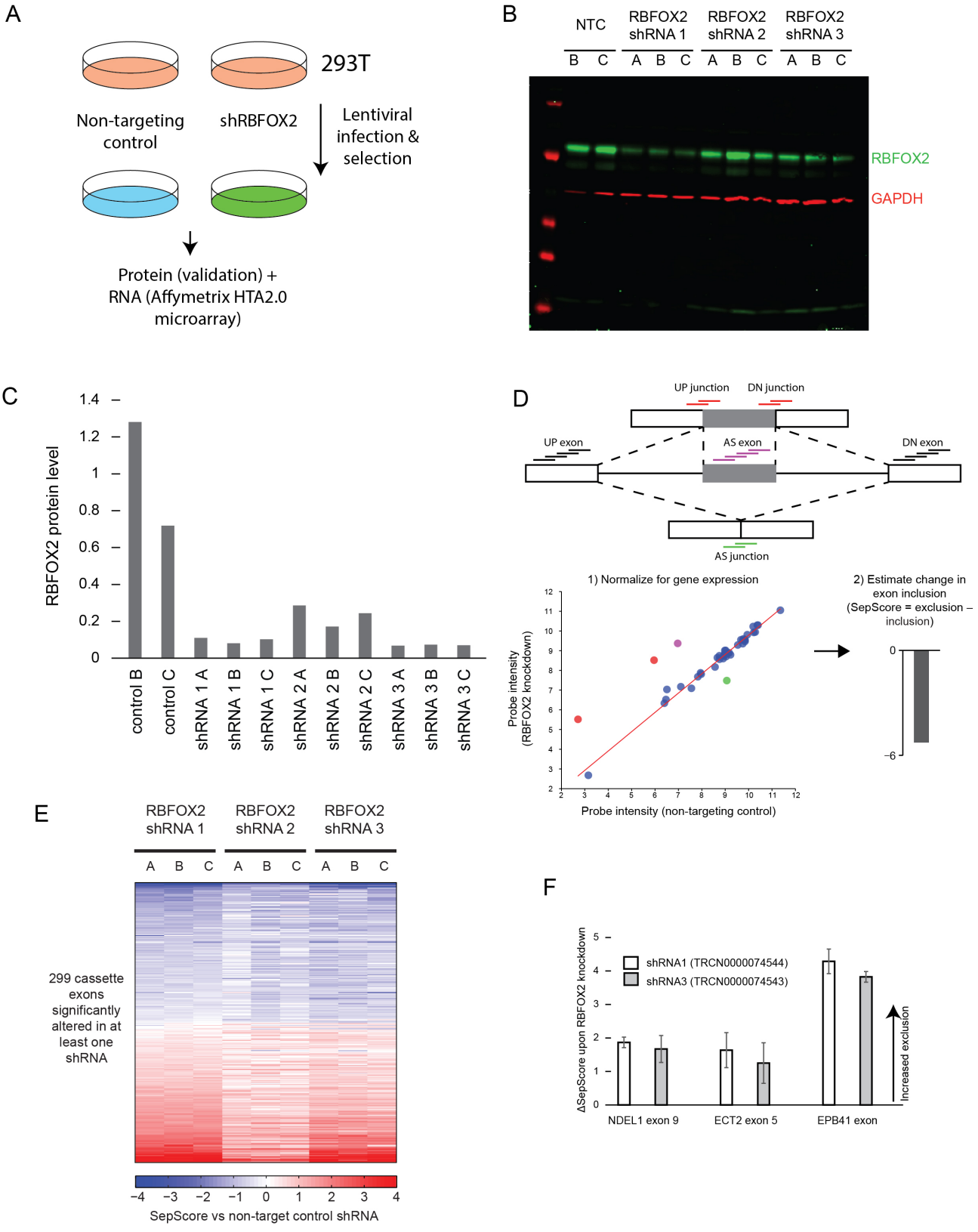


Supplementary Figure 9

Paired Size-Matched Input (SMInput) reveals enrichment over common background in CLIP of splicing regulator RBFOX2.

(a) All CLIPper-identified clusters identified for RBFOX2 293T eCLIP (Rep1) are plotted based on their fold-enrichment and significance compared to paired SMInput. Only 5,954 clusters (7.9%) are enriched at least 8-fold with $p \leq 10^{-5}$ by Fisher Exact or Chi Square test in eCLIP (green shaded box). Clusters overlapping introns flanking a set of 197 exons with RBFOX2 dependent splicing observed from microarray analysis of RBFOX2 knockdown (shRNA 1; Supplementary Fig. 10a-f) are indicated in green. (b) The subset of 50,853 RBFOX2 eCLIP clusters with either pre-normalized (CLIPper) or SMInput normalized *p*-value $\leq 10^{-5}$ were ranked by pre-normalized CLIPper *p*-value (left) or by SMInput normalization (right), as in Figure 2D. (Center) for clusters located in introns flanking RBFOX2-dependent cassette exons (Supplementary Fig. 10), change in rank is indicated by green lines, with significance determined by Kolmogorov-Smirnov test. Histograms indicate the number of RBFOX2-dependent cassette exon-flanking binding sites in each bin for clusters sorted by (left) CLIPper *p*-value, or (right) SMInput-normalized *p*-value. (c) Points indicate the enrichment for the RBFOX2 (UGCAUG) motif in each bin for RBFOX2 eCLIP clusters ranked by SMInput fold-enrichment (green) or pre-normalized CLIPper *p*-value (grey), with Replicate 1 indicated as solid and Replicate 2 as dashed lines. SMInput normalization decreases the frequency of motifs at non- or lowly-enriched clusters (left; indicating down-ranking of false positive clusters), but increases the frequency of motifs at highly enriched clusters (right; indicating up-ranking of true positive clusters). Motif enrichment was determined by counting the number of UGCAUG 6-mers in cluster sequences, and in 10 random permutations of the sequence within each clusters. (d) For the data shown in C, clusters were separated into two bins: 'depleted' clusters with decreased RPM in eCLIP vs SMInput, and 'significantly enriched' clusters with eCLIP read density at least 8-fold enriched and $p \leq 10^{-5}$ relative to SMInput. For both all CLIPper clusters (black), as well as a more stringent subset of only those with CLIPper *p*-value $\leq 10^{-5}$ (grey), depleted clusters show little or no enrichment for RBFOX2 motifs, whereas significantly enriched peaks show > 20-fold enrichment.

Supplementary Figure 10

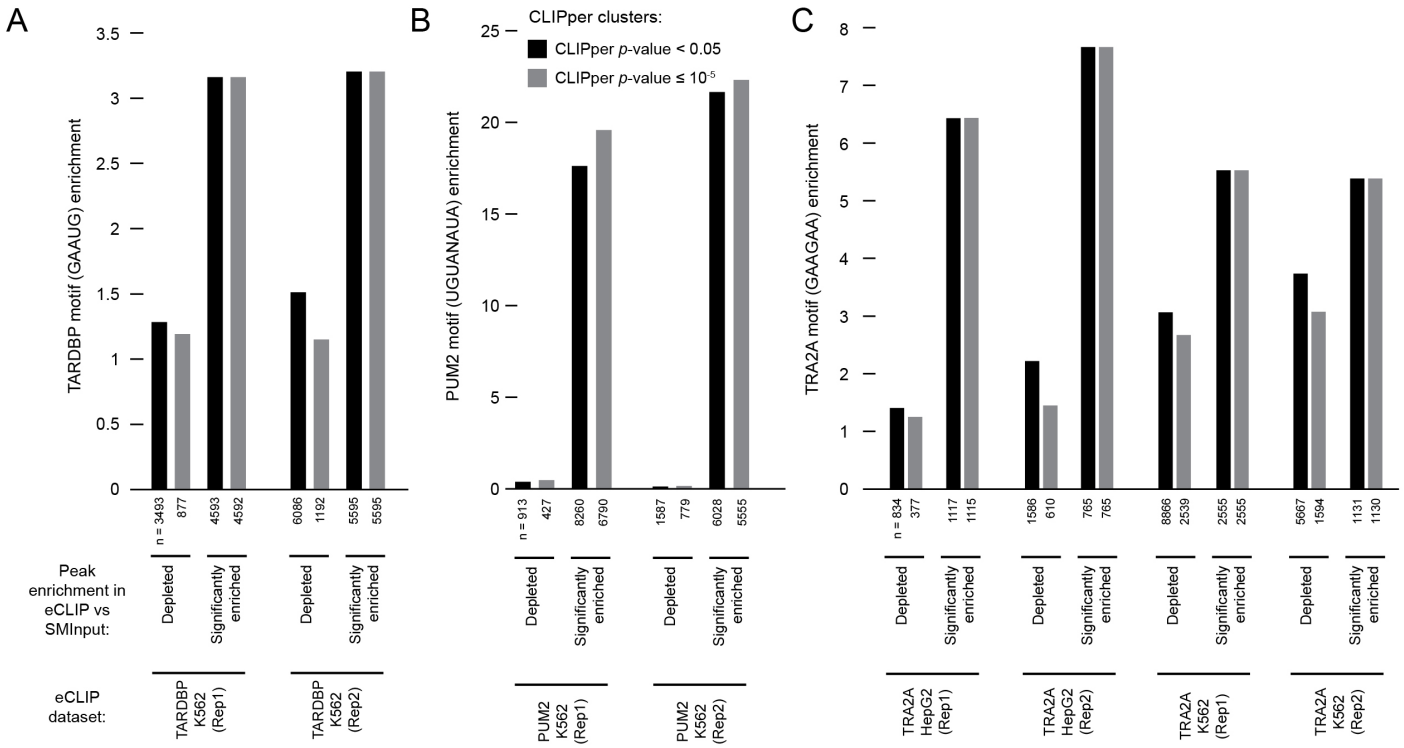


Supplementary Figure 10

Splicing-sensitive microarray analysis identifies RBFOX2-dependent cassette exons.

(a) RBFOX2 knockdown by transduction and selection for shRNA was performed in 293T cells, with splicing profiled by Affymetrix HTA2.0 microarray. Each knockdown was performed in biological triplicate, and each sample was separately prepared and hybridized. (B-C) Validation of RBFOX2 knockdown by western blot for shRNA 1 (TRCN0000074544), shRNA 2 (TRCN0000074546), and shRNA 3 (TRCN0000074543). (b) After lentiviral infection and puromycin selection, 293T cells were lysed in eCLIP lysis buffer, run on standard NuPAGE Novex 4-12% Bis-Tris gel (Thermo Fisher), transferred to PVDF membrane, and imaged on a LiCor Odyssey using RBFOX2 (rabbit A300-864A, Bethyl) and GAPDH (mouse ab8245, Abcam) primary and fluorescent secondary antibodies. (c) Band intensity was quantitated using LiCor ImageStudio Lite software. (d-f) Analysis of splicing-sensitive microarrays identifies RBFOX2-dependent cassette exons. (d) (top) Probes corresponding to cassette exon inclusion (AS exon probes (purple) and UP and DN junction probes (red)) and exclusion (AS junction (green)) were identified for all cassette exons profiled on the array. (bottom left) All probes for each gene were then normalized across samples to obtain residuals. (bottom right) Change in splicing is quantified by calculating a SepScore, defined as the mean residual signal for exclusion probes minus the mean signal for inclusion probes. (e) Heatmap indicates SepScore across all nine knockdown samples (relative to the average of non-target control samples) for the set of 299 events that showed significant change in either inclusion or exclusion probes ($p \leq 0.001$ by *t*-Test), as well as $|\text{SepScore}| \geq 0.5$ for at least one shRNA. Comparison of events significant in any of the three knockdown experiments showed high similarity in splicing change across shRNAs. (f) Splicing analysis SepScore shows increased exclusion for *NDEL1* exon 9, *ECT2* exon 5, and *EPB41* exon 16 upon RBFOX2 knockdown by shRNA.

Supplementary Figure 11

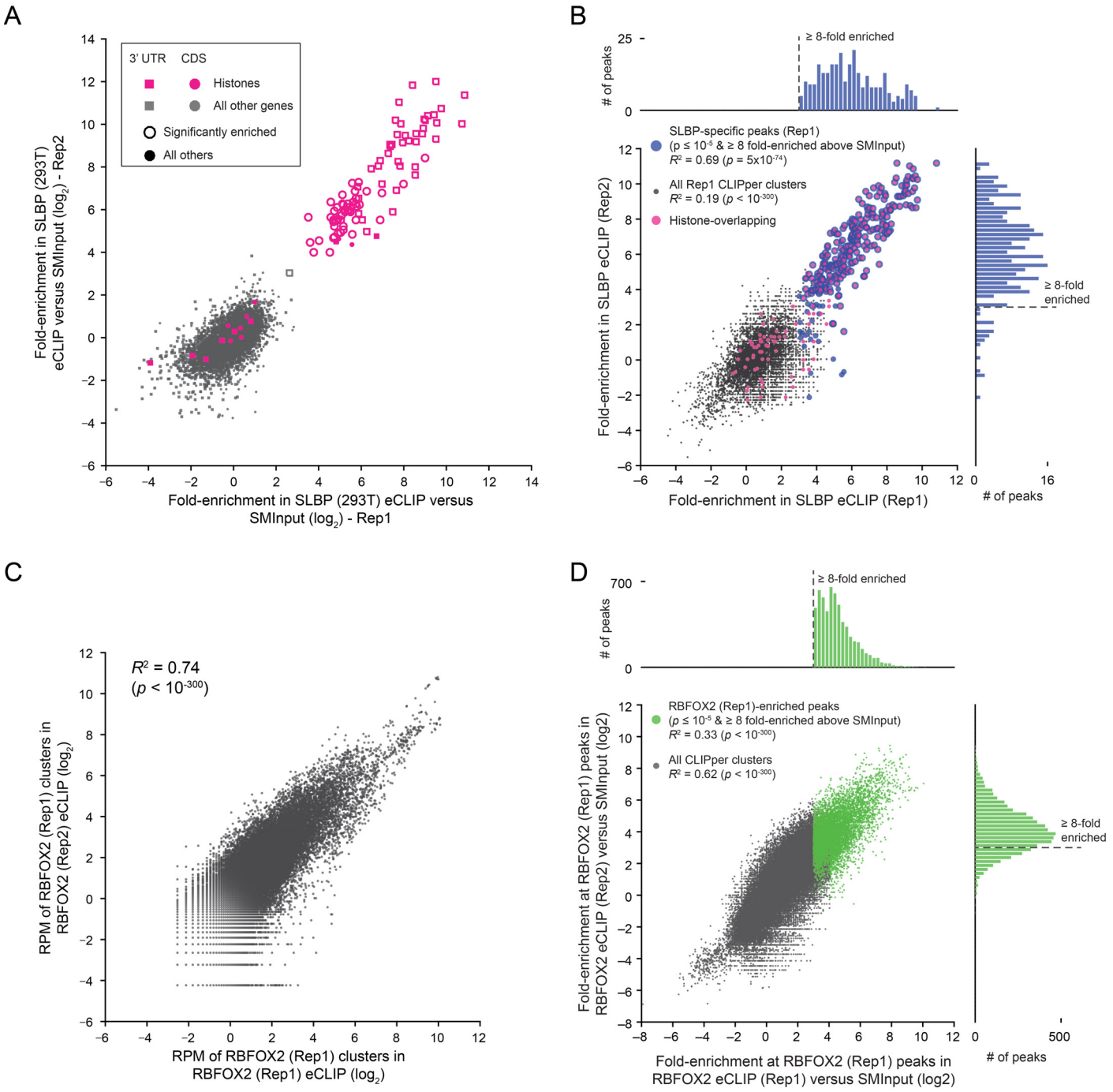


Supplementary Figure 11

SMInput-normalization distinguishes significantly enriched eCLIP peaks which contain known binding motifs from clusters depleted in eCLIP which lack motifs.

(a-c) As shown in Supplementary Figure 9D for RBFOX2, clusters for eCLIP of (a) TARDBP, (b) PUM2, and (c) TRA2A were separated into two bins: 'depleted' clusters with decreased RPM in eCLIP vs SMInput, and 'significantly enriched' clusters with eCLIP read density at least 8-fold enriched and $p \leq 10^{-5}$ relative to SMInput. Shown are motif enrichment for all CLIPper clusters (black), as well as a more stringent subset of only those with CLIPper p-value $\leq 10^{-5}$ (grey).

Supplementary Figure 12



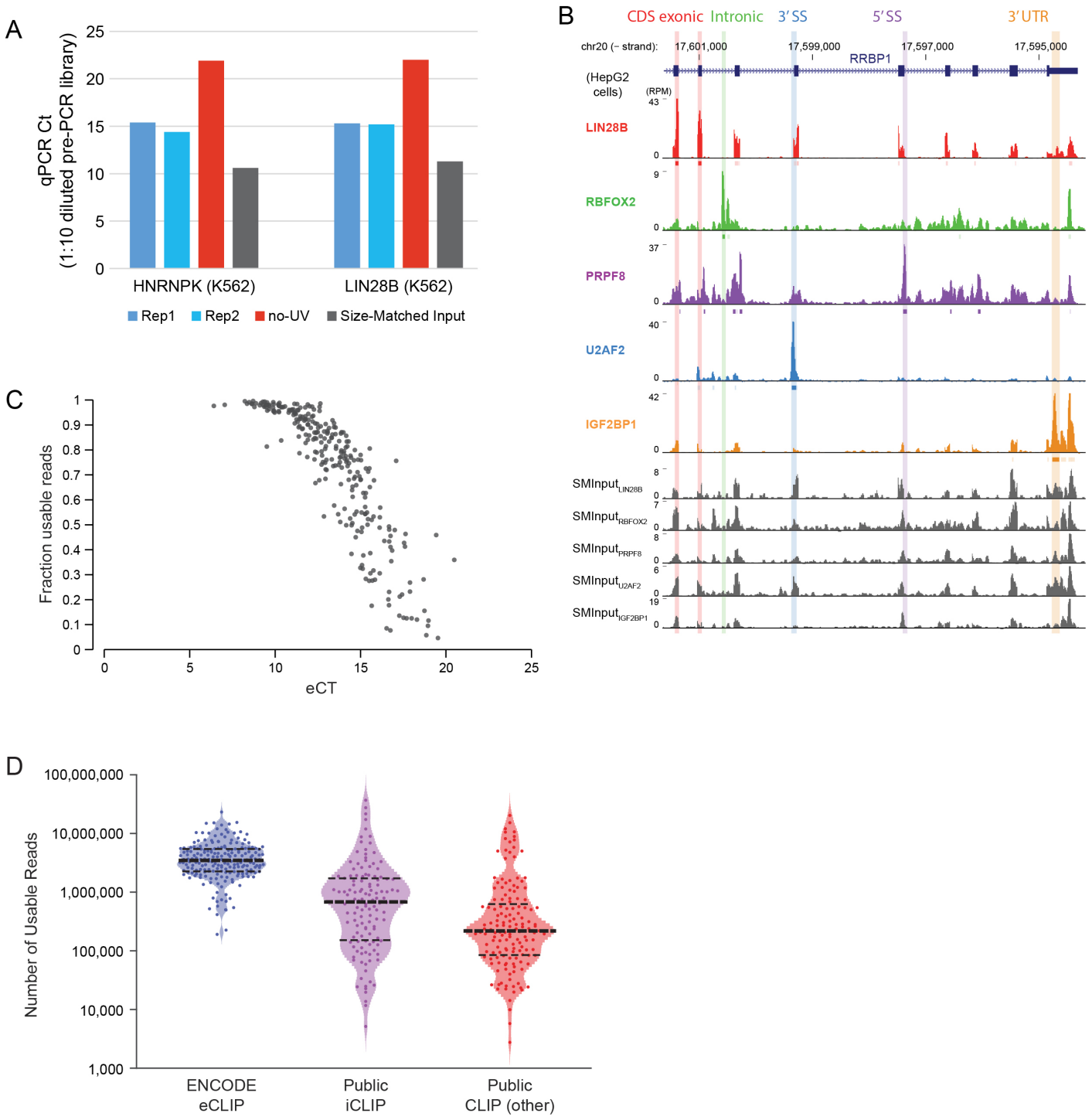
Supplementary Figure 12

eCLIP shows high reproducibility across biological replicates.

(a) SLBP eCLIP fold-enrichment over SMInput at histone RNAs is reproducible across biological replicate experiments. Each point indicates eCLIP fold-enrichment over paired SMInput for the CDS (circle) or 3′UTR (square) of genes profiled in independent biological replicate SLBP eCLIP experiments. Histone genes are indicated in pink, with open circles indicating significantly enriched regions (fold-

enrichment ≥ 4-fold, p-value ≤ 10^{-5} in eCLIP vs SMInput). Both CDS (R^2 = 0.50) and 3′UTR (R^2 = 0.73) show significant correlation (p < 10^{-300}, all significance determined by standard conversion of r values to t-statistic), and show enrichment at most histones. (b) SLBP clusters were identified in Replicate 1, and for each cluster the fold-enrichment was determined for both Replicate 1 and Replicate 2 eCLIP. Histone-overlapping points are indicated in pink, with significantly enriched peaks indicated in blue. Attached histograms show the number of significantly-enriched peaks with specified fold-enrichment in Replicate 1 (top) and Replicate 2 (right). (c) Correlation in read density across biological replicate RBFOX2 eCLIP experiments. Clusters were first identified in Replicate 1, and then each point indicates RBFOX2 eCLIP RPM for Rep1 and Rep2 at these clusters. (d) SMInput-normalized eCLIP peak signal shows high correlation between biological replicate RBFOX2 (and SMInput) experiments. Clusters are identified using CLIPper on Rep1 only, and points indicate fold-enrichment in eCLIP over SMInput for these regions across biological replicates. Green points indicate eCLIP-enriched peaks identified in replicate 1 (p-value ≤ 10^{-5} & fold-enrichment ≥ 8), with the distribution of these peaks across both replicates indicated by attached histograms.

Supplementary Figure 13



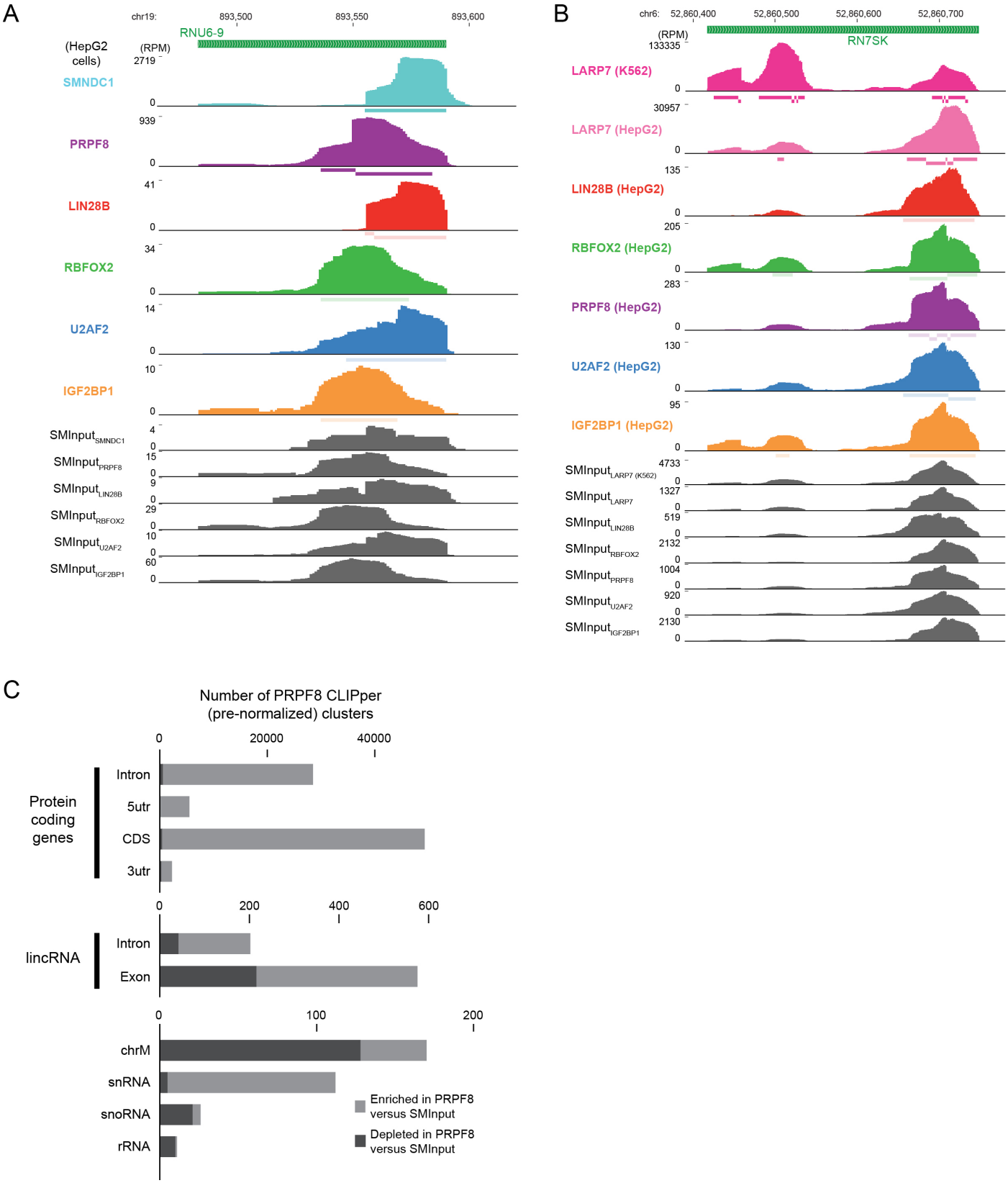
Supplementary Figure 13

Scalable RBP target identification with eCLIP.

(a) Non-crosslinked samples show decreased RNA recovery. Bars indicate Ct value obtained by performing qPCR on 1:10 diluted pre-

PCR (post-adapter ligated) library of HNRNPK and LIN28B eCLIP performed on two UV-crosslinked replicates, one non-crosslinked sample, and the paired SMInput. Increased qPCR Ct (greater required amplification) indicates decreased material obtained from the eCLIP procedure. (b) Distinct RNA binding profiles identified by eCLIP. Five HepG2 eCLIP experiments (along with paired SMInputs) are shown for the ~7kb region at the 3′ end of the RRBP1 gene, with peak calls indicated as boxes below RPM-normalized read density tracks. Significantly enriched peaks are indicated as darkened boxes. (c) Correlation between required amplification and percent of reads that are usable (i.e., not PCR duplicates) for 277 sequenced eCLIP libraries with more than 10^6 uniquely mapped reads. Each point indicates the eCT (extrapolated number of PCR cycles required to obtain 100 fMoles of library (*x*-axis), and the corresponding fraction of usable reads (out of uniquely mapped) obtained after high-throughput sequencing (*y*-axis). (d) eCLIP (204 libraries comprising 102 experiments in biological duplicate) yields increased usable reads with standard sequencing depth compared to 127 published iCLIP datasets or 152 published CLIP experiments. Each dataset is indicated by a point, with smoothed density plots created with the distributionPlot Matlab package with default settings (smoothened using ksdensity with a Normal kernel).

Supplementary Figure 14



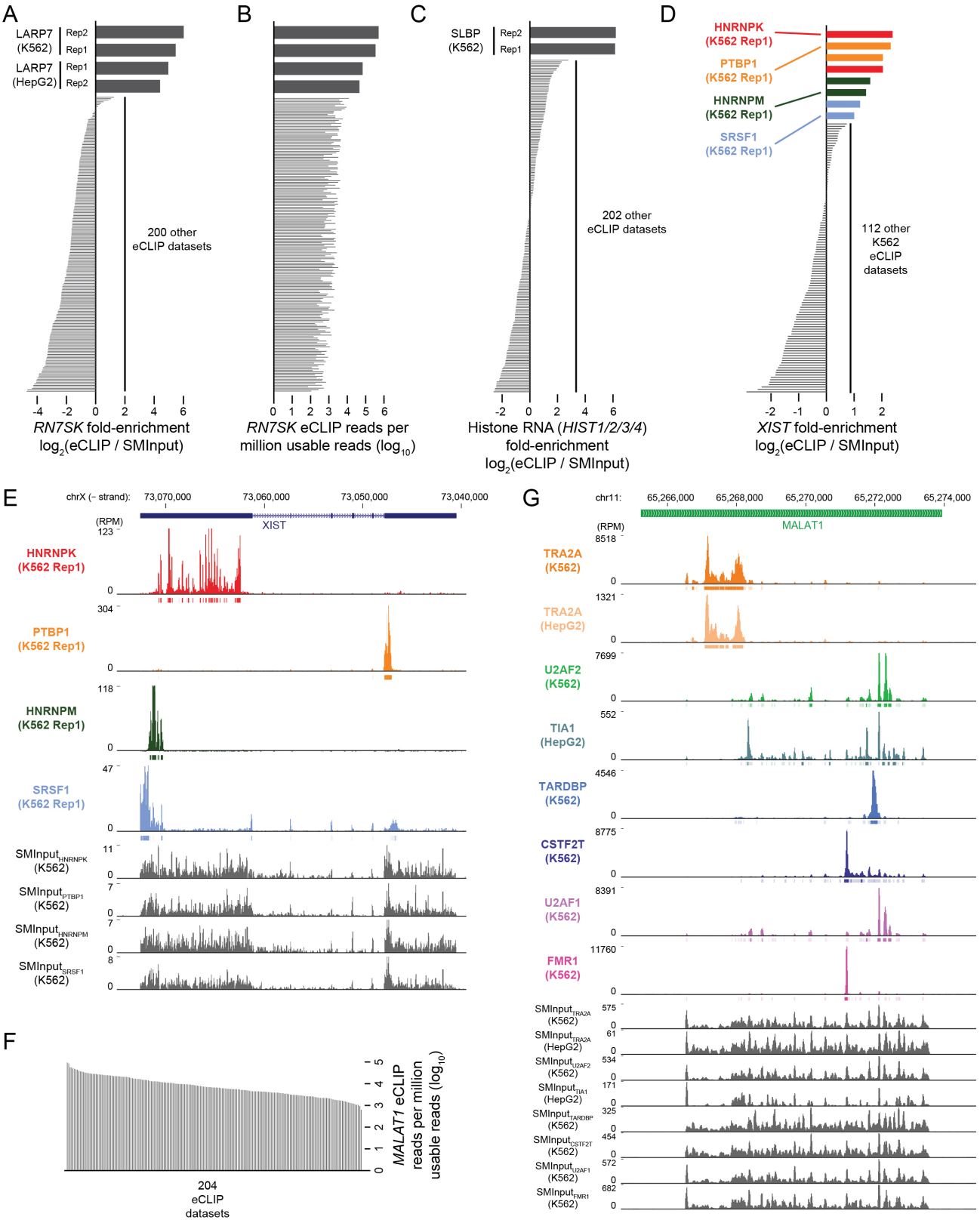
Nature Methods: doi:10.1038/nmeth.3810

Supplementary Figure 14

eCLIP enables distinction between significant binding and common background.

(a-b) Read density tracks indicate eCLIP and SMInput signal at two abundant small RNAs. (a) A U6 snRNA (Gencode ID *RNU6-9*) shows read density across all eCLIP and SMInput samples, including CLIPper-called clusters (light colored bars below tracks). Significantly enriched signal is only observed for eCLIP of SMNDC1 and PRPF8, with significantly enriched peaks indicated below (as darkly colored bars). (b) Similar analysis indicates common background signal at 7SK snRNA (*RN7SK*), but significant enrichment in eCLIP of known 7SK ribonucleoprotein particle component LARP7. (c) Analysis of PRPF8 clusters indicates that whereas the vast majority of intronic and CDS clusters show enrichment in PRPF8 eCLIP relative to SMInput, chrM, snoRNA, and rRNA-overlapping clusters are typically false positives that are depleted in eCLIP. Notably, unlike RBFOX2 (Figure 4a), snRNA clusters identified for PRPF8 are largely enriched in PRPF8 CLIP, consistent with its known role as a core spliceosome component.

Supplementary Figure 15



Nature Methods: doi:10.1038/nmeth.3810

Supplementary Figure 15

RNA-centric view of RNA binding protein association.

(a-b) Sorting all 204 K562 and HepG2 eCLIP datasets by fold-enrichment for 7SK snRNA (*RN7SK*) identifies LARP7 as specifically binding 7SK. (a) Each bar indicates fold-enrichment in eCLIP compared to SMInput for usable reads mapping to the 7SK snRNA. (b) Bars indicate RPM of 7SK snRNA in each eCLIP dataset, before SMInput normalization. 7SK has over 100 reads in nearly all eCLIP experiments, and over 1,000 reads in the majority of experiments (datasets are ordered identically as in (a)). (c) Sorting all eCLIP experiments by fold-enrichment summed over all histone RNAs identifies SLBP as uniquely binding to histone transcripts. (d) Sorting 120 K562 eCLIP experiments by fold-enrichment for *XIST* identifies four proteins with greater than 2-fold enrichment: HNRNPK, PTBP1, HNRNPM, and SRSF1. (e) For the four proteins with enriched binding to XIST identified in (d), read density tracks across *XIST* identify specific regions of binding. Bars below density tracks indicate clusters, with significantly SMInput-enriched clusters indicated by darkened color. (f) Bars indicate RPM across lncRNA *MALAT* for all 204 K562 and HepG2 eCLIP datasets. (g) Tracks show read density (in RPM) across *MALAT1* for eight RBPs indicated in Figure 4c, with paired SMInput datasets below. Boxes indicate CLIPper-identified clusters, with significantly enriched peaks indicated as dark boxes.

Supplementary Table 3. Antisense oligonucleotides used for RBFOX2-blocking experiments.

ASO_name	ASO_sequence	Modifications	Used as control in
NDEL1_ASO1	CCCATGCAGTTAGTAAAA	Uniform MOE, PS	
NDEL1_ASO2	TCAGCCCCATGCAGTTAG	Uniform MOE, PS	
NDEL1_ASO3	CTGAGTCAGCCCCATGCA	Uniform MOE, PS	
ECT2_ASO1	CACATGCAATGAGTTACT	Uniform MOE, PS	NDEL1 (control1)
ECT2_ASO2	CATGCCACATGCAATGAG	Uniform MOE, PS	NDEL1 (control1)
ECT2_ASO3	AGCATGCCACATGCAATG	Uniform MOE, PS	
ECT2_ASO4	TGCAGCATGCCACATGCA	Uniform MOE, PS	
ECT2_ASO5	AGTGCAGCATGCCACATG	Uniform MOE, PS	
ECT2_ASO6	AGGGAAGTGCAGCATGCC	Uniform MOE, PS	
EPB41_ASO1	TGCATGCAAAACCAAATG	Uniform MOE, PS	ECT2 (control1)
EPB41_ASO2	GCAATTGCATGCAAAACC	Uniform MOE, PS	ECT2 (control2)
EPB41_ASO3	TTCATGCAATTGCATGCA	Uniform MOE, PS	
EPB41_ASO4	GTCCCTTCATGCAATTGC	Uniform MOE, PS	
EPB41_ASO5	CTAAAGTCCCTTCATGCA	Uniform MOE, PS	
EPB41_ASO6	AACATGCAAAAGCATTTT	Uniform MOE, PS	
EPB41_ASO7	TCACCAACATGCAAAAGC	Uniform MOE, PS	
EPB41_ASO8	CATGTTCACCAACATGCA	Uniform MOE, PS	
MPZL1_ASO1	GACATGCAATCCTTTCAC	Uniform MOE, PS	ECT2 (control1)
MPZL1_ASO2	TTCAGGACATGCAATCCT	Uniform MOE, PS	ECT2 (control2)
LRRFIP2_ASO1	CACATGCTTGAAAGACAA	Uniform MOE, PS	ECT2 (control1)
LRRFIP2_ASO2	TTAAGCACATGCTTGAAA	Uniform MOE, PS	ECT2 (control2)
ANKRD26_ASO1	TTCATGCAGTGTTAGTAT	Uniform MOE, PS	EPB41 (control1)
ANKRD26_ASO2	ATAAATTCATGCAGTGTT	Uniform MOE, PS	EPB41 (control2)
FAM190Bx_ASO1	ATCATGCAAACTGGAAAA	Uniform MOE, PS	EPB41 (control1)
DOCK7_ASO1	ATCATGCAGTATTAGCTA	Uniform MOE, PS	EPB41 (control1)
DOCK7_ASO2	GAGATATCATGCAGTATT	Uniform MOE, PS	EPB41 (control2)

Eric Van Nostrand / Yeo Lab – contact elvannostrand@ucsd.edu & geneyeo@ucsd.edu

Supplementary Protocol 1: eCLIP-seq Experimental Procedures

Buffers

iCLIP lysis buffer
50 mM Tris-HCl pH 7.4
100 mM NaCl
1% NP-40 (Igepal CA630)
0.1% SDS
0.5% sodium deoxycholate (protect from light)
1:200 Protease Inhibitor Cocktail III (add fresh)

High salt wash buffer
50 mM Tris-HCl pH 7.4
1 M NaCl
1 mM EDTA
1% NP-40
0.1% SDS
0.5% sodium deoxycholate (protect from light)

Wash buffer
20 mM Tris-HCl pH 7.4
10 mM $MgCl_2$
0.2% Tween-20

5X PNK pH 6.5 buffer
350mM Tris-HCl pH 6.5
50mM $MgCl_2$

1X FastAP Buffer
10mM Tris pH 7.5
5mM $MgCl_2$
100mM KCl
0.02% Triton X-100

1x RNA Ligase Buffer
50mM Tris-HCl pH 7.5
10mM $MgCl_2$

PK Buffer
100mM Tris-HCl pH 7.4
50mM NaCl
10mM EDTA

RLT Buffer
Qiagen cat # 79216

Enzymes

Turbo DNase	2 U/µl	LifeTech	AM2239
RNase I	100 U/µl	LifeTech	AM2295
FastAP	1 U/µl	LifeTech	EF0652
Murine RNase Inhibitor	40 U/µl	NEB	M0314L
T4 PNK	10 U/µl	NEB	M0201L
T4 RNA ligase 1 high conc	30 U/µl	NEB	M0437M
Proteinase K	0.8 U/µl	NEB	P8107S
Q5 PCR Master Mix		NEB	M0492L
Protease Inhibitor Cocktail III		EMD Millipore	
AffinityScript reverse transcriptase		Agilent	600107
Exo-SAP-IT		Affymetrix	78201

Beads

Dynabeads M-280 sheep anti-rabbit	10 mg/ml	LifeTech	
Dynabeads Protein G	30 mg/ml	LifeTech	37002D
Dynabeads MyOne Silane	40 mg/ml	LifeTech	
Agencourt AMPure XP beads		Beckman Coµlter	A63881

Nature Methods: doi:10.1038/nmeth.3810

Primers

RNA oligos:

Original RNA adapters:

(RNA_A01 & RNA_B06 create a colorbalanced pair)

RNA_A01 /5phos/rArUrUrGrCrUrUrArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

RNA_B06 /5phos/rArCrArArGrCrCrArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

(RNA_C01 & RNA_D08 create a colorbalanced pair)

RNA_C01 /5phos/rArArCrUrUrGrUrArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

RNA_D08 /5phos/rArGrGrArCrCrArArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

RiL19 /5phos/rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrG/3SpC3/

New RNA adapters (avoids low-complexity issues in HiSeq 2500 cluster identification)

(RNA_X1A & RNA_X1B create a colorbalanced pair)

RNA_X1A /5Phos/rArUrArUrArGrG rNrNrNrNrN rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

RNA_X1B /5Phos/rArArUrArGrCrA rNrNrNrNrN rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

(RNA_X2A & RNA_X2B create a colorbalanced pair)

RNA_X2A /5Phos/rArArGrUrArUrA rNrNrNrNrN rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

RNA_X2B /5Phos/rArGrArArGrArU rNrNrNrNrN rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrGrUrArG/3SpC3/

---> All RNA barcode adapters: 200 uM stock concentration (store at -80C), 20 uM working concentration

RiL19 /5phos/rArGrArUrCrGrGrArArGrArGrCrGrUrCrGrUrG/3SpC3/

(stock 200 uM; working 40 uM)

DNA oligos:

AR17 ACACGACGCTCTTCCGA (stock 200 uM; working 20 uM)

rand103Tr3 /5Phos/NNNNNNNNNNAGATCGGAAGAGCACACGTCTG/3SpC3/ (stock 200 uM; working 80 uM)

(Below we order page-purified)

PCR_F_D501 AATGATACGGCGACCACCGAGATCTACACTATAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

PCR_F_D502 AATGATACGGCGACCACCGAGATCTACACATAGAGGCACACTCTTTCCCTACACGACGCTCTTCCGATCT

PCR_R_D701 CAAGCAGAAGACGGCATACGAGATCGAGTAATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

PCR_R_D702 CAAGCAGAAGACGGCATACGAGATTCTCCGGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATC

(See Illumina customer service letter for D503-508, D703-712; any standard Illumina HT RNA-seq primers work fine)

(stock 100 uM; working 20 uM)

Nature Methods: doi:10.1038/nmeth.3810

Notes

For this protocol (& for ENCODE eCLIP-seq experiments), one 'experiment' is defined as 4 libraries: 2 eCLIP experiments on UV crosslinked biological replicate samples, 1 eCLIP experiment on a non-UV crosslinked sample, and 1 size-matched input control (taken from one of the two UV crosslinked samples). Additionally, an IgG-only IP is run on the Western gel to validate antibody specificity.

For other experiments, one can modify this design to add the paired size-matched input control from the other replicate, remove the non-UV crosslinked sample, or add additional library controls (IgG, FLAG or V5 pulldown on wild-type cells, etc.) if desired.

Although we do not standardly do P32 labeling, we still use the "HOT" and "COLD" membranes nomenclature from iCLIP & other CLIP protocols. COLD = 10% of sample run as a standard Western blot; HOT = 80% of sample run for membrane cutting & RNA isolation

DAY 1

Prepare iCLIP lysis mix

- Pre-chill iCLIP lysis buffer
- Per sample (20 million cells): add **5.5 µl 200x Protease Inhibitor Cocktail III** to **1 mL iCLIP lysis buffer**, mix
    - ** Note: For tissues or cell-types with high endogenous RNAse, add 11 µl **Murine RNase Inhibitor per 1 mL lysis buffer** at this step (works for ES, Neuronal Stem Cell, many tissues). This may need to be further increased for particularly difficult samples (e.g. Pancreas).

Lyse cells (Do this first)

- **Lyse cells:**
    - Retrieve cell pellets from -80 degC freezer, immediately add 1 mL cold **iCLIP lysis mix** to each pellet, pipette to resuspend

      3 Pellets per experiment:

        - Sample 1: IP-A (UV-crosslinked batch #1)
        - Sample 2: IP-B (UV-crosslinked batch #2)
        - Sample 3: nonX-UV (non-UV crosslinked, batch #3)

      IMPORTANT: for ENCODE, Sample 1 and Sample 2 MUST be different biological replicates. The simplest way to do this is to have different culture start date and culture end dates. If dates are similar, you must make sure before starting that the samples actually meet ENCODE criteria for being distinct biological replicate samples.

      Potential sample 4

        - Sample 4: IgG (non-UV-crosslinked batch #3)

      (One 20 M IgG IP is good for 10 IP experiments & can be stored after IP and denaturation in NuPage buffer + DTT)

    - Lyse 15 mins on ice

Couple antibody to magnetic beads (start while lysate on ice)

Nature Methods: doi:10.1038/nmeth.3810

Note: Process IgG identically to antibodies

- **Beads and antibodies:**
  - Use **125 µl beads** per sample
    - rabbit antibodies: use sheep anti-rabbit beads
    - mouse antibodies: use sheep anti-mouse beads
  - Use **10 µg antibody** per sample

- **Prepare beads:**
  - Magnetically separate beads, remove supernatant
  - Wash beads 2x in 500 µl cold **iCLIP lysis buffer**
  - Resuspend beads in 100 µl cold **iCLIP lysis buffer**

- **Bind antibody:**
  - Add antibody (10 µg) to 100 µl washed beads
  - Rotate, room temp, 45 min

RNase treat lysate (while ab+bead binding**):**

- Sonicate in Bioruptor at 'low' setting, 4 degC, 5 min, 30sec on / 30 sec off
- Dilute RNase I in PBS at 1:25 on ice; use 10 µl diluted RNase I per sample
- Add 2 µl **Turbo DNase**, mix Immediately before use,
- Add 10 µl **diluted RNase I**, mix & immediately proceed to next step
- Incubate in Thermomixer at 1200 rpm, 37 degC, 5 mins (exactly), place on ice
- Immediately add 11 µl **Murine RNase Inhibitor**, mix (If added earlier, ignore this step)
- Centrifuge 15,000g, 4 degC, 15min
- Transfer supernatant to a new tube

Capture RBP-RNA complexes on beads

- Wash antibody beads 2x in 500 µl cold **iCLIP lysis buffer**
- Remove 20 µL (2%) of Sample 1, 2, 3 as BACKUP inputs for western; store at 4 degC
- Add remainder to washed antibody beads
- Rotate 4 degC, 2 h or overnight (in cold room)

Step: SAVE INPUT SAMPLES: Remove Input Samples

- Mix samples well
- To new tube, take 20 µL (2%) of Sample 1 (A-Input) for 'HOT' gel, store at 4 degC
- To new tube, take 20 uL (2%) of Sample 1 (A-Input), 2 (B-Input), 3 (NX-Input), 4 (IgG-Input) for COLD gel; store at 4 degC
- To new tube, take 2 uL (0.2%) of Sample 3 (NX-Input) for COLD gel 0.1% input lane; store at 4 degC
- To new tubes, take 5 tubes of 20ul each of sample 4 (IgG-Input) as COLD IgG Input samples.

Wash beads

- Wash 2x with 900µL cold **High salt wash buffer**
- Wash 1x with 500µL cold **Wash buffer**
- Transition to 1xFastAP buffer: add 500 µl cold **Wash buffer**, move through magnet, separate on magnet, add 500 µl **1xFastAP** buffer, mix, remove supernatant
- Wash 1x with 500 µl **1xFastAP buffer**

Nature Methods: doi:10.1038/nmeth.3810

⇨ (If doing **IgG samples**: **pause** the IgG sample here and store on ice in Wash buffer)

FastAP treat beads (all samples except IgG)

- **Prepare FastAP master mix** on ice; 100 µl per sample:
  - H_2O 79 µl
  - 10x FastAP buffer 10 µl
  - Murine RNase Inhibitor 2 µl
  - Turbo DNase 1 µl
  - FastAP enzyme 8 µl

- Mix, add **100 µl** to each sample, incubate in Thermomixer at 1200 rpm, 37 degC, 15 min

PNK treat beads

- While beads are incubating, **prepare PNK master mix** on ice; 300 µl per sample:
  - H_2O 224 µl
  - 5X PNK pH 6.5 buffer 60 µl
  - 0.1 M DTT 3 µl
  - Murine RNase Inhibitor 5 µl
  - Turbo DNase 1 µl
  - T4 PNK enzyme 7 µl

- Mix, add **300 µl** to each sample, incubate in Thermomixer at 1200 rpm, 37 degC, 20 min

Wash beads

- Magnetically separate bead suspension, remove supernatant
- Wash 1x with 500µL cold **Wash buffer**
- Transition to High salt wash buffer: add 500 µl cold **Wash buffer**, move through magnet, separate on magnet, add 500 µl **High salt wash buffer**, move through magnet, remove supernatant
- Transition to Wash buffer: add 500 µl cold **High salt wash buffer**, move through magnet, separate on magnet, add 500 µl **Wash buffer**, move through magnet, remove supernatant
- Wash 1x with 500µL cold **Wash buffer**
- Transition to 1xLigase buffer (no DTT): add 500 µl **Wash buffer**, move through magnet, separate on magnet, add 300 µl **1xLigase buffer (no DTT)**, move through magnet, remove supernatant
- Wash 2X with 300 µl **1xLigase buffer (no DTT)**
- Prepare the 3' ligation master mix
- Just before adding the 3' ligation master mix, briefly spin tubes in minifuge, magnetically separate, remove residual liquid with fine tip

Ligate 3' RNA linker (on-bead)

- **Prepare 3' ligation master mix** on ice; 25 µl per sample:
  - H_2O 9 µl
  - 10x Ligase buffer (no DTT) 3 µl
  - 0.1 M ATP 0.3 µl
  - 100% DMSO 0.8 µl
  - 50% PEG 8000 9 µl

Nature Methods: doi:10.1038/nmeth.3810

  - Murine RNase Inhibitor 0.4 µl
  - RNA Ligase high conc. 2.5 µl

- Mix carefµlly by pipetting or flicking (do not vortex)
- Add **25 µl** to each sample
- To each sample, add 2.5 µl of each of two different **barcoded RNA adapters** to each sample

  Acceptable RNA adapter pairs:

    - A01 + B06
    - C01 + D08
    - A03 + G07
    - A04 + F05
    - X1-A + X1-B
    - X2-A + X2-B
- Incubate at room temperature for 75 min; flick to mix every ~10 min

Wash beads (resume IgG sample here)

- Add 500µL cold **Wash buffer**, magnetically separate, remove supernatant
- Transition to High salt wash buffer: add 500 µl cold **Wash buffer**, move through magnet, separate on magnet, add 500 µl **High salt wash buffer**, move through magnet, remove supernatant
- Wash 1x with 500µL cold **High salt wash buffer**
- Transition to Wash buffer: add 500 µl cold **High salt wash buffer**, move through magnet, separate on magnet, add 500 µl **Wash buffer**, move through magnet, remove supernatant
- Wash 2x with 500µL cold **Wash buffer**

Prepare samples for gel loading

- **IP-Bead samples (HOT and COLD)**:
  **** Note: HOT & COLD are named relative to iCLIP gels; neither is radioactive in eCLIP**
  HOT = CLIP gel – for membrane transfer & RNA isolation
  COLD = WESTERN gel – for western imaging
  - **Remove s/n, a**dd **100** µl cold **Wash buffer**, resuspend beads well
  - Move 20 µl to new tube #1 = **COLD IP-WB samples**
  - Remaining 80 uL = **HOT IP samples**

- For **COLD IP-WB samples:**
  - **COLD IP-WB samples** 20.0 µl

  Add:
  - 4x NuPAGE buffer 7.5 µl
  - 1M DTT 3.0 µl

- For **HOT IP samples:**
  - Place sample on magnet, remove supernatant
  - Resuspend in **elution/loading master mix**; 30 µl per sample:
    - Wash buffer 20.0 µl
    - 4x NuPAGE buffer 7.5 µl
    - 1M DTT 3.0 µl

- For **HOT Input samples**, mix:

Nature Methods: doi:10.1038/nmeth.3810

- Input sample 20.0 µl (Saved in step x)
- 4x NuPAGE buffer 7.5 µl
- 1M DTT 3.0 µl

- For **COLD input samples**, mix:

Input sample:	1x	0.1x
Input sample	20.0 µl /	2.0 µl (Saved in Step: SAVE INPUT SAMPLES)
Wash buffer	0 µl /	18.0 µl
4x NuPAGE buffer	7.5 µl	
1M DTT	3.0 µl	

- For **IgG samples**:
  - Resuspend beads in 100 uL of wash buffer
  - 4x NuPAGE buffer 37.5 µl
  - 1M DTT 15.0 µl

  (Final volume 150 uL -> load 15 uL per well)

- Denature all samples in Thermomixer, 1200 rpm, 70 degC, 10 min
- Cool on ice 1 min, spin briefly in minifuge
- For **all samples**, transfer supernatant to new tube (IP AND Inputs have beads)

Load and run gels

- Load HOT gel (4-12% Bis-Tris, 10-well, 1.5 mm) with (M) pre-stained markers and (m) diluted pre-stained marker (2 uL marker, 2 uL 4x NuPAGE buffer, 6 uL Wash Buffer)

1	2	3	4	5	6	7	8	9	10
M	Input	(m)	A-IP	(m)	B-IP	(m)	NX-IP	M	(m)

Load:

HOT Input: 30 uL volume (30 uL denatured sample = 20 uL input lysate = 2% of input). HOT Input (for library prep) should come from crosslinked samples (either Sample A or Sample B).

IP-NX, IP-A, IP-B: 30 uL volume (80% of IP)

- Load COLD gel gel (4-12% Bis-Tris, -well, 1.5 mm)

1	2	3	4	5	6	7	8	9	10
NX-input (1:10 diluted)	IgG Input	IgG bead	NX-Input	NX-IP	M	A-IP	A-INPUT	B-IP	B-INPUT

Load:

Input & 1:10 input: Load 15 ul, save remaining 15 uL as backup (15ul denatured sample = 10ul lysate = 1% or 0.1% Input respectively)

IP: Load 15 ul, save remaining 15 uL as backup (15ul denatured sample = 10% of IP bead sample).

IgG: Load 15 uL, save remaining

- (All saved samples at -20C)
- Run at 150V in 1xMOPS running buffer, 75 min or until dye front is at the bottom

Nature Methods: doi:10.1038/nmeth.3810

Transfer to membranes

- **Prepare transfer:**
  - (Have pre-prepared COLD (4 deg) transfer buffer with methanol: 1xNuPAGE transfer buffer, 10% methanol
  - **COLD gel**: Prepare PVDF membrane(s): pre-flash 10 s in methanol, move to transfer buffer with methanol
  - **HOT gel**: Prepare Nitrocellulose membrane(s): incubate in transfer buffer for > 1 min
  - Wet sponges and Whatman papers in transfer buffer with methanol
  - Assemble transfer stacks, from bottom to top (black side of stack holder on bottom):

    1x sponge – 2x Whatman paper – gel – membrane – 2x Whatman paper – 1x sponge

    Cold gel: PVDF membrane

    HOT gel: Nitrocellulose membrane
- **Transfer:**
  - overnight 30V (preferred) OR
  - 2 hr 200 mA (if doing this, only hook up one transfer box per power supply)

Nature Methods: doi:10.1038/nmeth.3810

Day 2

- Remove HOT membrane, rinse quickly once with sterile 1X PBS, wrap in Saran wrap, store at -20C

Develop COLD membrane

- Block in 5% milk in TBST, room temp, 30 min
- Probe with primary antibody: 0.2 ug/ml (1:5000 for a 1 mg/ml stock; check antibody) in 5% milk in TBST, room temp, 1 hr.
- Wash 3x with TBST, 5 min
- Probe with secondary antibody: 1:4000 Rabbit TrueBlot HRP in in 5% milk in TBST, room temp, 1 – 3 h
    - (Note: if western fails or signal is low, 1:1000 gives higher signal)
- Wash 3x with TBST, 5 min
- Mix equal volumes of ECL Buffer A + Buffer B (or 40:1 of ECL Plus Buffer 1 to Buffer 2), add to membrane and incubate (mix/rotate) for 1-5 min. (1ml final volume per membrane)
- Develop 30 sec & 5 min, then judge signal (15 min maximum; if 15 sec is still too bright, expose two films

Cut HOT membrane

- Note RBP band on film with respect to prestained protein markers
- Place HOT membrane on clean glass/metal surface
- Using a fresh razor blade, cut lane from HOT membrane from the RBP band to 75 kDa above it
- Slice membrane pieces into ~1-2 mm slices, use a fresh razor blade for each sample
- Transfer slices to Eppendorf tube – place tube on ice if doing many samples
- Collect slices at the bottom of tube (centrifuge if necessary)

Release RNA from membrane

- Prepare **Proteinase K mix** on ice, 200 µl per sample:
    - PK buffer 160 µl
    - Proteinase K 40 µl
- Mix, add **200 µl** Proteinase K mix to membrane slices, incubate in Thermomixer at 1200 rpm, 37 C, 20 min (make sure all membrane slices are submerged)
- Prepare Urea/PK buffer: Dissolve 420 mg Urea in 500 µL PK buffer, then add PK buffer to final volume of 1 mL
- Add **200 µl** Urea/PK buffer to samples, mix, incubate in Thermomixer at 1200 rpm, 37 C, for an additional 20 min

Purify RNA

- Add 400 µL **acid phenol/chloroform/isoamyl alcohol** (pH 6.5), mix well by shaking, incubate in Thermomixer at 1200 rpm, 37 C, 5 min
- Spin briefly in picoFuge, transfer all except membrane slices to Phaselock gel HEAVY tube, incubate in Thermomixer at 1200 rpm, 37 C, 5 min
- Centrifuge at 13000g, 15 min, room temp (gel should have separated phenol and aqueous phases)
- Transfer aqueous layer to new 15 mL conical tube (at least 3 mL volume tube)

Zymo column cleanup (replaces precipitation) – RNA Clean & Concentrator-5 columns (Cat R1016)

Nature Methods: doi:10.1038/nmeth.3810

*** Note: we replaced precipitation with the column preps to avoid phenol/chloroform carryover, which SIGNIFICANTLY inhibits Exo-SAP and can cause adapter dimer libraries. Either can work, but column cleanup is significantly safer for large-scale experiments**

- Add 2 volumes RNA binding buffer (usually 2 x ~400 = 800 uL)
- Add equal volume 100% ethanol & mix (usually ~1200 uL)
- Transfer 750 uL of mixed sample to Zymo-Spin column
- Centrifuge 30 sec & discard flow-through
- Repeat spins by reloading additional 750 uL volume until all sample has been spun through column
- Add 400 uL RNA Prep Buffer, centrifuge for 30 sec, discard flow through
- Add 700 uL RNA Wash Buffer, centrifuge for 30 sec, discard flow through
- Add 400 uL RNA Wash Buffer, centrifuge for 30 sec, discard flow through
- Centrifuge additional 2 mins
- Transfer column to new 1.5 mL tube (avoid getting wash buffer on column)
- **INPUT: Add 10 uL H2O to column, let sit for 1 min, centrifuge for 30 sec**
- **CLIP: Add 10 uL H2O to column, let sit for 1 min, centrifuge for 30 sec**
- **Store at -80 C until RT**

~~(Struck through is previous precipitation version of the protocol)~~

- ~~Add 400 uL **chloroform**, mix well by shaking, centrifuge 13,000g, 1 min, room temp~~
- ~~Transfer aqueous (upper) phase to new tube – AVOID TRANSFERRING ORGANIC PHASE (leave some aqueous phase if necessary, organic phase inhibits later steps)~~
- ~~add 2 µL **GlycoBlue**, 30 µL **3M NaOAc (pH 5.5),** vortex, spin briefly in picoFuge~~
- ~~Add 1 mL cold **100% EtOH**, mix well by inverting, precipitate at -80 C (O/N, or for at least 1h)~~

Day 3

START Inputs only →

- ~~Store CLIP samples at -80 C until RT~~

~~Precipitate input RNA~~

- ~~Centrifuge samples at 13,000g, 15 min, 4 degC~~
- ~~Locate pellet, carefully remove supernatant~~
- ~~Carefully add 750 µL **75% ice-cold EtOH**~~
- ~~Centrifuge at max speed, 5 min, 4 degC~~
- ~~Locate pellet, remove supernatant~~
- ~~Spin briefly in picoFuge, remove residual liquid with fine tip~~
- ~~Air-dry until dry (~10 mins).~~
- ~~Resuspend in **20 µL H_2O**~~

FastAP treat input RNA

- To 10 µL sample, add:
  - 10 uL H_2O
  - 2.5 µl 10X FastAP buffer
  - 0.5 µl RNase Inhibitor
  - 2.5 µl FastAP enzyme

Nature Methods: doi:10.1038/nmeth.3810

- Mix, incubate in Thermomixer at 1200 rpm, 37 C, 15 min

PNK treat input RNA

- Make **PNK master mix**; 75 µl per sample:
    - H_2O 45 µl
    - 5xPNK 6.5 buffer 20 µl
    - 0.1M DTT 1 µl
    - Turbo DNase 1 µl
    - Murine RNase Inhibitor 1 µl
    - PNK enzyme 7 µl

- Mix, add **75 µl** to samples, mix, incubate in Thermomixer at 1200 rpm, 37 C, 20 min

Silane cleanup input RNA

- **Prepare beads:**
    - Magnetically separate 20 µl **MyONE Silane beads** per sample, remove supernatant
    - Wash 1x with 900 µl **RLT buffer**
    - Resuspend beads in 300 µL **RLT buffer** per sample

- **Bind RNA:**
    - Add beads in 300 µl **RLT buffer** to sample, mix
    - Add 10 µL **5M NaCl**
    - Add 615 µL **100% EtOH**
    - Mix, rotate at room temp, 15 min

- **Wash beads:**
    - Magnetically separate, remove supernatant
    - Add 1 mL **75% EtOH**, pipette resuspend and move suspension to **new tube**
    - After 30 s, magnetically separate, remove supernatant
    - Wash 2x with **75% EtOH** (let sit 30 s)
    - Spin briefly in picoFuge, magnetically separate, remove residual liquid with fine tip
    - Air-dry 5 min

- **Elute RNA:**
    - Resuspend in **10 µl H_2O**, let sit for 5 min
    - Magnetically separate
    - Transfer 5 uL of supernatant to new tube (for 3' linker ligation below)
    - Transfer remainder of supernatant to new tube & store at -20 (this is the backup input RNA sample)

3' linker ligate input RNA

- **Anneal adapter:**
    - Take 5 µl of RNA (from above)
    - Add 1.5 µl 100% DMSO
    - Add 0.5 µl **RiL19** adapter
    - Incubate 65 C, 2 min
    - Place on ice >1 min

Nature Methods: doi:10.1038/nmeth.3810

- **Prepare ligation master mix; 13.5 µl per sample:**
    - 10x NEB Ligase Buffer (with DTT) 2.0 µl
    - 0.1M ATP 0.2 µl
    - Murine RNase Inhibitor 0.2 µl
    - 100% DMSO 0.3 µl
    - 50% PEG 8000 8.0 µl
    - RNA Ligase high conc 1.3 µl
    - H_2O 1.5 µl

- Flick/pipette mix, add **13.5 µl** to each sample, flick/pipette-mix, incubate at room temp for 75 min
- Flick to mix every ~15 min

Silane cleanup input RNA

Note: can start next CLIP sample precipitation spin in parallel

- **Prepare beads:**
    - Magnetically separate 20 µl **MyONE Silane beads** per sample, remove supernatant
    - Wash 1x with 900 µl **RLT buffer**
    - Resuspend beads in 61.6 µL **RLT buffer**

- **Bind RNA:**
    - Add beads in 61.6 µl **RLT buffer** to sample, mix
    - Add 61.6 µL **100% EtOH**
    - Pipette mix, leave pipette tip in tube, pipette mix every ~3-5 min for 15 min

- **Wash beads:**
    - Magnetically separate, remove supernatant
    - Add 1 mL **75% EtOH**, pipette resuspend and move to **new tube**
    - After 30 s, magnetically separate, remove supernatant
    - Wash 2x with **75% EtOH** (30 s)
    - Spin briefly in picoFuge, magnetically separate, remove residual liquid with fine tip
    - Air-dry 5 min

- **Elute RNA:**
    - Resuspend in **10 µl H_2O**, let sit for 5 min
    - Magnetically separate, transfer supernatant to new tube

- Possible stopping point (Can store input samples at -80 C until next day)

←END Inputs only

~~(Struck through is previous precipitation version of the protocol)~~

~~Precipitate CLIP RNA (this can be done simultaneously to Silane cleanup above)~~

- ~~Centrifuge samples at 13000g, 15 min, 4 C~~
- ~~Locate pellet, remove supernatant~~
- ~~Carefully add 750 µL 75% ice-cold EtOH~~
- ~~Centrifuge at max speed, 5 min, 4 C~~

Nature Methods: doi:10.1038/nmeth.3810

- ~~Locate pellet, remove supernatant~~
- ~~Spin briefly in picoFuge, remove residual liquid with fine tip~~
- ~~Air-dry until dry (~10 mins)~~
- ~~Resuspend in **10 µL H_2O**~~

All CLIP and INPUT samples are now synchronized.

Reverse transcribe RNA (ALL CLIP and INPUTS)

- **Anneal primer** in 8-well strip tubes:
  - Mix 10µl of RNA with 0.5µl AR17 primer (using Rainin pipette + tips)
  - Heat 65 C for 2 min in pre-heated PCR block, place immediately on ice (do not cool down in PCR block)

- **Prepare reverse transcription master mix** on ice; 10 µl per sample:
  - H_2O 4.0 µl
  - 10x AffinityScript Buffer 2.0 µl
  - 0.1M DTT 2.0 µl
  - dNTPs (**25 mM each**) 0.8 µl
  - Murine RNase Inhibitor 0.3 µl
  - AffinityScript Enzyme 0.9 µl

- Add 10 µl to each sample, mix, incubate 55 C, 45 min in pre-heated PCR block

Cleanup cDNA

- **ExoSAP Treatment**
  - Add 3.5 µl **ExoSAP-IT** to each sample, vortex, spin down
  - Incubate 37 degC for 15 mins on PCR block
  - Add 1 µl **0.5M EDTA**, pipette-mix
- **RNA removal**
  - Add 3 µl of **1M NaOH**, pipette-mix
  - Incubate 70 degC, 12 min on PCR block
  - Add 3 µl of **1M HCl**, pipette-mix (to fix pH)

Silane cleanup cDNA

- **Prepare beads:**
  - Magnetically separate 10 µl **MyONE Silane beads** per sample, remove supernatant
  - Wash 1x with 500 µl **RLT buffer**
  - Resuspend beads in 93 µL **RLT buffer**

- **Bind RNA:**
  - Add beads in 93 µl **RLT buffer** to sample, mix
  - Add 111.6 µL **100% EtOH**
  - Pipette mix, leave pipette tip in tube, pipette mix twice, for 5 min

- **Wash beads:**
  - Magnetically separate, remove supernatant
  - Add 1 mL **80% EtOH**, pipette resuspend and move to **new tube**

Nature Methods: doi:10.1038/nmeth.3810

- After 30 s, magnetically separate, remove supernatant
- Wash 2x with **80% EtOH** (30 s)
- Spin briefly in picoFuge, magnetically separate, remove residual liquid with fine tip
- Air-dry 5 min

- **Elute RNA:**
  - Resuspend in 5 µl 5 mM tris-Cl pH 7.5, let sit for 5 min (do not remove from beads)

5' linker ligate cDNA (on-bead)

- **Anneal linker:**
  - Add 0.8 µl **rand3Tr3** adapter
  - Add 1 µl 100% **DMSO**
  - Heat at 75 degC, 2 min, place immediately on ice for >1 min

- **Prepare ligation master mix** on ice; 12.8 µl per sample:
  - 10x NEB RNA Ligase Buffer (with DTT) 2.0 µl
  - 0.1M ATP 0.2 µl
  - 50% PEG 8000 9.0 µl
  - RNA Ligase high conc 0.5 µl
  - H_2O 1.1 µl

- Flick to mix, spin down, add 12.8 µl to each sample: stir sample with pipette tip, then add master mix slowly with stirring; needs to be homogeneous
- Add another 1 µl **RNA Ligase high conc** to each sample, flick to mix
- Incubate on Thermomixer at 1200 rpm, room temp for 30 s, then put on bench
- Flick, ideally every hour, at least a few times before leaving overnight
- Incubate at room temp overnight

Day 4

Silane cleanup linker-ligated cDNA

- **Prepare beads:**
  - Magnetically separate 5 µl **MyONE Silane beads** per sample, remove supernatant
  - Wash 1x with 500 µl **RLT buffer**
  - Resuspend beads in 60 µL RLT buffer per sample

- **Bind RNA:**
  - Add beads in 60 µl **RLT buffer** to each sample, mix
  - Add 60 µL **100% EtOH**
  - Pipette mix, leave pipette tip in tube, pipette mix twice, for 5 min

- **Wash beads:**
  - Magnetically separate, remove supernatant
  - Add 1 mL **75% EtOH**, pipette resuspend and move to **new tube**
  - After 30 s, magnetically separate, remove supernatant
  - Wash 2x with **75% EtOH** (30 s)
  - Spin briefly in picoFuge, magnetically separate, remove residual liquid with fine tip

Nature Methods: doi:10.1038/nmeth.3810

- Air-dry 5 min

- **Elute RNA:**
  - Resuspend in 27 µl **10mM Tris-HCl pH 7.5**, let sit for 5 min
  - Magnetically separate, transfer **25 µL** sample to new tube

qPCR quantify cDNA

- **Prepare qPCR master mix**; 9 µl per sample:
  - PowerSybr 2x master mix 5.0 µl
  - H_2O 3.6 µl
  - qPCR primer mix 0.4 µl (10 uM each qPCR-grade D5x/D7x mix)

- Mix, dispense into 384-well qPCR plate, add **1 µl 1:10 diluted (in H_2O) cDNA**, seal, mix
- Run qPCR, note Cq values
  - **Cycle # for final PCR will be 3 cycles less than the Ct of the 1:10 diluted sample**

    **** Note: we use the automatically calculated Ct for this; this '3 cycle less' rule may change based on your lab setup, so for the first couple CLIPs it is best to err on the side of 1 or 2 extra PCR cycles. If final libraries are > 50 nM (especially if > 100 nM), you should back off a couple cycles.**

PCR amplify cDNA

- **Prepare PCR** on ice; 50 uL total per sample:
  - 2x Q5 PCR master mix 25.0 µl
  - H_2O 5.0 µl
  - 20 µM right primer (D50x) 2.5 µl
  - 20 µM left primer (D70x) 2.5 µl

- Dispense into 8-well strips, add **12.5 µl CLIP sample + 2.5 H_2O**; for **inputs**, use **10 µl + 5 µl H_2O**; mix

- PCR conditions (cycle # depending on library):
  - 98 C for 30 s
  - 98 C for 15 sec -> 68 C for 30 sec -> 72 C for 40 sec (x6 cycles)
  - 98 C for 15 sec -> 72 C for 60 sec (x ? cycles)
  - 72 C 1 min
  - 4 C hold
- Typical: Input 9 total cycles (6 + 3), CLIP 16 (6 + 10) total cycles
  - Note that 18 cycles will yield ~30-50% PCR duplicated libraries (further increasing above 18 cycles), which can be ok for RBPs with few specific targets but will be challenging for broad binders.
- **Cycle # for final PCR: 3 cycles less than the qPCR Ct of the 1:10 diluted sample**

SPRI cleanup library

- Resuspend **AmpureXP beads** well
- Add 90 µl bead suspension (do not separate) per 50 µl PCR reaction, pipette mix well, incubate room temp 10 min (pipette mix 2-3x during incubation)

Nature Methods: doi:10.1038/nmeth.3810

- Magnetically separate, wash beads 2x with **75% EtOH**
- Dry beads for 5 min on magnet
- Move from magnet, resuspend in **20 µl H_2O**, let sit for 5 min
- Magnetically separate, transfer **18 µl** to new tubes

Gel-purify library

- Prepare **3% low melting temp agarose gel (**Seakem GTG LMP) in 1% TBE
    - 120 ml for larger gel tray
    - mix often while microwaving (low melting temp gel tends to boil over rapidly)
    - cool down, add **1:10,000 SybrSafe**, mix, pour

- **Prepare samples and run gel:**
    - Add 6 µl **6x OrangeG** buffer to each sample (18 µl of sample), mix
    - Prepare two **50 bp ladder** samples in Orange G buffer (Per well: 0.5 µl ladder + 2 uL Orange G + 7.5 uL H2O)
    - Load on gel, leave 1 empty well between samples, ladder on both sides of the gel
    - Run ~95V for 50 mins (longer gives better resolution but larger cut sizes)

- **Gel-extract library from gel:**
    - Under blue light illumination, cut gel slices 175-350 bp and place into 15 mL conical tubes, using fresh razor blades for each sample; keep cross-contamination to minimum
        - Keep in mind: adapter-dimer (including RNA adapter) is 142 bp, so anything below 175 will cluster & create reads on the HiSeq, but is too short to map and will be wasted

- **Cut & elute gel** using Qiagen MinElute gel extraction kit:
    - Weigh 15 mL conical with gel slice (blank with empty conical tube)
    - Calculate gel weight, add 6x volumes of **Buffer QG** to melt gel (e.g. for 100 mg gel, add 600 µL QG)
    - Melt gel at room temp (do <u>not</u> heat) on benchtop (can shake to help melt, but don't vortex)
    - After gel is melted, add 1x volume of original gel of **isopropanol** & <u>mix well</u> (100 mg gel = 100 µl isopropanol)
    - Load on column (750 µl per spin, can do multiple spins, all spins max speed 1 min)
        - **NOTE:** if gel weight is >400 mg, wash 1x with 500 uL Buffer QG after every 4 spins)
    - After all sample has been spun through, wash 1x with 500 µl **Buffer QG**
    - Add 1X with 750 µl **Buffer PE**, spin 1 min, pour out flow-through, spin again 2 min max speed
    - Carefully move column to new 1.5 mL tube (avoid any carryover of PE – if any liquid is visible on the outside of the column redo 2 min max speed spin)
    - Using a fine tip, pipette all remaining PE buffer off of the plastic purple rim of the MinElute column
    - Air dry 2 mins
    - Carefully add 12.5 µl **Buffer EB** directly to the center of the column, incubate 2 min room temp, spin max speed
    - For improved yield – repeat the elution (take the flow-through and add it to the column again)

Quantitate library (D1000 tapestation)

- 3 µl D1000 loading buffer, 1 µl sample
- Vortex to mix, spin down in microfuge

Nature Methods: doi:10.1038/nmeth.3810

Yeo Lab, UCSD - Contact geneyeo@ucsd.edu , gpratt@ucsd.edu , elvannostrand@ucsd.edu

Supplementary Protocol 2: eCLIP-seq Processing Pipeline

Programs Used & Version Information

(For all custom scripts: https://github.com/gpratt/gatk/releases/tag/2.3.2)

Yeo Lab Custom Script Versions:
Barcode_collapse_pe.py: https://github.com/YeoLab/gscripts/releases/tag/1.1
Make_bigwig_files.py: https://github.com/YeoLab/gscripts/releases/tag/1.1
Clipper: https://github.com/YeoLab/clipper/releases/tag/1.1
Clip_analysis: https://github.com/YeoLab/clipper/releases/tag/1.1
negBedGraph.py: https://github.com/YeoLab/gscripts/releases/tag/1.1
demux_paired_end.py: https://github.com/YeoLab/gscripts/releases/tag/1.1
fastq-sort: http://homes.cs.washington.edu/~dcjones/fastq-tools/fastq-tools-0.8.tar.gz
Peak_input_normalization_wrapper.pl: https://github.com/YeoLab/gscripts/tree/1.1/perl_scripts
overlap_peakfi_with_bam_PE.pl: https://github.com/YeoLab/gscripts/tree/1.1/perl_scripts
compress_l2foldenrpeakfi.pl: https://github.com/YeoLab/gscripts/tree/1.1/perl_scripts

Other programs used:
FastQC: v. 0.10.1
Cutadapt: v. 1.9.dev1
STAR: v. STAR_2.4.0i
Samtools: v. 0.1.19-96b5f2294a
bedToBigBed: v. 2.6
Bedtools: v. 2.25.0
R: v. 3.0.2

Python and Python Package Versions:
Python 2.7.11 :: Anaconda 2.1.0 (64-bit)
Pysam 0.8.3
Bx 0.5.0
HTSeq 0.6.1p1
Numpy 1.10.2
Pandas 0.17.0
Pybedtools 0.7.0
Sklearn 0.15.2
Scipy 0.16.1
Matplotlib 1.4.3
Gffutils 0.8.2
Seaborn 0.6.0
Statsmodels 0.5.0

Perl Packages used:
Statistics-Distributions-1.02

Yeo Lab, UCSD - Contact geneyeo@ucsd.edu , gpratt@ucsd.edu , elvannostrand@ucsd.edu

Script Details

Our entire processing pipeline is performed by two commands: (1) Demultiplexing of fastq files based on inline barcodes, and (2) A scala command that procedurally performs all subsequent processing steps in order. See the next section for detailed description of processing steps performed by the scala pipeline.

Demultiplexing:

Script:

```
demux_paired_end.py --fastq_1 <fastq_read_1> --fastq_2 <fastq_read_2> -b
<barcode_file.txt> --out_file_1 <fastq_read_1_out> --out_file_2
<fastq_read_2_out> --length <randomer_length> -m <metrics_file>
```

Input file Documentation:

The input file is a tab separated file that describes the barcodes to demultiplex.

Column 1: Barcode to demultiplex
Column 2: Human readable label to append to the demultiplexed file.

Example Manifest:

```
ACAAGTT    /full/path/to/files/file_R1.C01
```

Pipeline:

Script:

```
java -Xms512m -Xmx512m -jar /path/to/gatk/dist/Queue.jar -S
/path/to/qscripts/analyze_clip_seq_encode.scala --input manifest.txt --barcoded
--adapter AATGATACGGCGACCACCGAGATCTCTCTTTCCCTACACGACGCTCTTCCGATCT --adapter
CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT --adapter
AGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
ATTGCTTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
ACAAGCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
AACTTGTAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
AGGACCAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
ANNNNGGTCATAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
ANNNNACAGGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
ANNNNAAGCTGAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --adapter
ANNNNGTATCCAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT --g_adapter CTACACGACGCTCTTCCGATCT
-qsub -jobQueue home-yeo -jobNative "-W group_list=yeo-group" -runDir
/path/to/output/directory -log result.log  -keepIntermediates  --job_limit 400
-run
```

Input manifest.txt documentation:

This is a **tab separated file** that is 7 columns long.

Column 1: read 1 and read 2 input fastq files separated by a semi-colon.
Column 2: Species, either hg19 or mm9
Column 3: Biological Replicate ID. If two columns have the same ID they will be merged post mapping and duplicate removal.

Column 4: 3' adapters to be removed from the second read in the pair.
Column 5: minimum length of overlap between adapter and barcode for cutadapt. (Used with variable length barcode/random-mer structures).
Column 6: 5' adapters to be removed from the first read in the pair.
Column 7: length of random-mers to be trimmed from the 3' end of read 1

Example Manifest:

```
/full/path/to/files/file_R1.C01.fastq.gz;/full/path/to/files/file_R2.C01.fastq.
gz hg19    Merged_ID
AACTTGTAGATCGGA;AGGACCAAGATCGGA;ACTTGTAGATCGGAA;GGACCAAGATCGGAA;CTTGTAGATCGGAAG
;GACCAAGATCGGAAG;TTGTAGATCGGAAGA;ACCAAGATCGGAAGA;TGTAGATCGGAAGAG;CCAAGATCGGAAGA
G;GTAGATCGGAAGAGC;CAAGATCGGAAGAGC;TAGATCGGAAGAGCG;AAGATCGGAAGAGCG;AGATCGGAAGAGC
GT;GATCGGAAGAGCGTC;ATCGGAAGAGCGTCG;TCGGAAGAGCGTCGT;CGGAAGAGCGTCGTG;GGAAGAGCGTCG
TGT 5       CTTCCGATCTACAAGTT;CTTCCGATCTTGGTCCT     5
```

Inline barcode description:
Each inline barcode is ligated to the 5' end of Read1 and its id and sequence are listed below:

```
A01     ATTGCTTAGATCGGAAGAGCGTCGTGT
B06     ACAAGCCAGATCGGAAGAGCGTCGTGT
C01     AACTTGTAGATCGGAAGAGCGTCGTGT
D08     AGGACCAAGATCGGAAGAGCGTCGTGT
A03     ANNNNGGTCATAGATCGGAAGAGCGTCGTGT
G07     ANNNNACAGGAAGATCGGAAGAGCGTCGTGT
A04     ANNNNAAGCTGAGATCGGAAGAGCGTCGTGT
F05     ANNNNGTATCCAGATCGGAAGAGCGTCGTGT
RiL19/none    AGATCGGAAGAGCGTCGTGT
```

(see eCLIP protocol document for full description of these oligos)

We have observed occasional double ligation events on the 5' end of Read1, and we have found that to fix this requires we run cutadapt twice. Additionally, because two adapters are used for each library (to ensure proper balancing on the Illumina sequencer), two separate barcodes may be ligated to the same Read1 5' end (often with 5' truncations). To fix this we split the barcodes up into 15bp chunks so that cutadapt is able to deconvolute barcode adapters properly (as by default it will not find adapters missing the first N bases of the adapter sequence)

Column 6 is made by appending one of the barcodes below (these are the same barcode sequences used to demultiplex):
AAGCAAT A01
GGCTTGT B06
ACAAGTT C01
TGGTCCT D08
ATGACCNNNNT A03
TCCTGTNNNNT G07
CAGCTTNNNNT A04
GGATACNNNNT F05

To the 5' adapter
CTTCCGATCT

Yeo Lab, UCSD - Contact geneyeo@ucsd.edu , gpratt@ucsd.edu , elvannostrand@ucsd.edu

Human Readable Description of Steps

Note: Until the merging step each script is run twice, one once for each barcode used

Fastqc round 1: Run and examined by eye to make sure libraries look alright

```
fastqc /full/path/to/files/file_R1.C01.fastq.gz -o /full/path/to/files/ >
/full/path/to/files/file_R1.C01.fastq.gz.dummy_fastqc

fastqc /full/path/to/files/file_R2.C01.fastq.gz -o /full/path/to/files/ >
/full/path/to/files/file_R2.C01.fastq.gz.dummy_fastqc
```

Cutadapt round 1: Takes output from demultiplexed files. Run to trim off both 5' and 3' adapters on both reads

```
cutadapt -f fastq --match-read-wildcards --times 1 -e 0.1 -O 1 --
quality-cutoff 6 -m 18 -a NNNNNAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC -g
CTTCCGATCTACAAGTT -g CTTCCGATCTTGGTCCT -A AACTTGTAGATCGGA -A
AGGACCAAGATCGGA -A ACTTGTAGATCGGAA -A GGACCAAGATCGGAA -A CTTGT
AGATCGGAAG -A GACCAAGATCGGAAG -A TTGTAGATCGGAAGA -A ACCAAGATCGGAAGA -A
TGTAGATCGGAAGAG -A CCAAGATCGGAAGAG -A GTAGATCGGAAGAGC -A CAAGATCGGAAGAGC
-A TAGATCGGAAGAGCG -A AAGATCGGAAGAGCG -A AGATCGGAAGAGCGT -A
GATCGGAAGAGCGTC -A ATCGGAAGAGCGTCG -A TCGGAAGAGCGTCGT -A CGGAAGAGCGTCGTG
-A GGAAGAGCGTCGTGT -o
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.fastq.gz -p
/full/path/to/files/file_R2.C01.fastq.gz.adapterTrim.fastq.gz
/full/path/to/files/file_R1.C01.fastq.gz
/full/path/to/files/file_R2.C01.fastq.gz >
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.metrics
```

Cutadapt round 2: Takes output from cutadapt round 1. Run to trim off the 3' adapters on read 2, to control for double ligation events.

```
cutadapt -f fastq --match-read-wildcards --times 1 -e 0.1 -O 5 --
quality-cutoff 6 -m 18 -A AACTTGTAGATCGGA -A AGGACCAAGATCGGA -A
ACTTGTAGATCGGAA -A GGACCAAGATCGGAA -A CTTGTAGATCGGAAG -A GACCAAGATCGGAAG
-A TTGTAGATCGGAAGA -A ACCAAGATCGGAAGA -A TGTAGATCGGAAGAG -A
CCAAGATCGGAAGAG -A GTAGATCGGAAGAGC -A CAAGATCGGAAGAGC -A TAGATCGGAAGAGCG
-A AAGATCGGAAGAGCG -A AGATCGGAAGAGCGT -A GATCGGAAGAGCGTC -A
ATCGGAAGAGCGTCG -A TCGGAAGAGCGTCGT -A CGGAAGAGCGTCGTG -A GGAAGAGCGTCGTGT
-o /full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.fastq.gz
-p /full/path/to/files/file_R2.C01.fastq.gz.adapterTrim.round2.fastq.gz
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.fastq.gz
/full/path/to/files/file_R2.C01.fastq.gz.adapterTrim.fastq.gz >
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.metrics
```

STAR rmRep: Takes output from cutadapt round 2. Maps to human specific version of RepBase used to remove repetitive elements, helps control for spurious artifacts from rRNA (& other) repetitive reads.

```
STAR --runMode alignReads --runThreadN 16 --genomeDir
/path/to/RepBase_human_database_file --genomeLoad LoadAndRemove --
readFilesIn
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.fastq.gz
/full/path/to/files/file_R2.C01.fastq.gz.adapterTrim.round2.fastq.gz --
outSAMunmapped Within --outFilterMultimapNmax 30 --
```

```
outFilterMultimapScoreRange 1  --outFileNamePrefix
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bam --
outSAMattributes All  --readFilesCommand zcat  --outStd BAM_Unsorted  --
outSAMtype BAM Unsorted  --outFilterType BySJout  --outReadsUnmapped
Fastx  --outFilterScoreMin 10  --outSAMattrRGline ID:foo  --alignEndsType
EndToEnd >
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bam
```

Samtools view and count_aligned_from_sam: Takes output from STAR rmRep. Counts the number of reads mapping to each repetitive element.

```
samtools view
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bam |
count_aligned_from_sam.py >
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.metrics
```

Fastqc round 2: Takes output from STAR rmRep. Runs a second round of fastqc to verify that after read grooming the data still is usable.

```
fastqc
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmapp
ed.out.mate1  -o /full/path/to/files/ >
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmapp
ed.out.mate1.dummy_fastqc
```

Fastq-sort: Takes unmapped output from STAR rmRep and sorts it to account for issues with STAR not outputting first and second mate pairs in order

```
fastq-sort  --id
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmappe
d.out.mate1  >  /full/path/to/files
file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmapped.out.sorted.mate1
&& fastq-sort  --id
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmappe
d.out.mate2 >
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmappe
d.out.sorted.mate2
```

STAR genome mapping: Takes output from STAR rmRep. Maps unique reads to the human genome

```
STAR  --runMode alignReads  --runThreadN 16  --genomeDir
/path/to/STAR_database_file --genomeLoad LoadAndRemove  --readFilesIn
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmapp
ed.out.mate1
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rep.bamUnmapp
ed.out.mate2  --outSAMunmapped Within  --outFilterMultimapNmax 1  --
outFilterMultimapScoreRange 1  --outFileNamePrefix
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.bam  --
outSAMattributes All  --outStd BAM_Unsorted  --outSAMtype BAM Unsorted  -
-outFilterType BySJout  --outReadsUnmapped Fastx  --outFilterScoreMin 10
--outSAMattrRGline ID:foo  --alignEndsType EndToEnd >
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.bam
```

Barcode_collapse_pe: takes output from STAR genome mapping. Custom random-mer-aware script for PCR duplicate removal.

```
barcode_collapse_pe.py  --bam
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.bam  --
out_file
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.rmDup.b
```

```
am  --metrics_file
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.rmDup.m
etrics
```

sortSam: Takes output from barcode collapse PE. Sorts resulting bam file for use downstream.

```
java  -Xmx2048m  -XX:+UseParallelOldGC  -XX:ParallelGCThreads=4  -
XX:GCTimeLimit=50  -XX:GCHeapFreeLimit=10  -
Djava.io.tmpdir=/full/path/to/files/.queue/tmp  -cp
/path/to/gatk/dist/Queue.jar  net.sf.picard.sam.SortSam
INPUT=/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.r
mDup.bam  TMP_DIR=/full/path/to/files/.queue/tmp
OUTPUT=/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.
rmDup.sorted.bam  VALIDATION_STRINGENCY=SILENT  SO=coordinate
CREATE_INDEX=true
```

samtools index: Takes output from sortSam, makes bam index for use downstream.

```
samtools index
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.rmDup.s
orted.bam
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.rmDup.s
orted.bam.bai
```

samtools merge: Takes inputs from multiple final bam files. Merges the two technical replicates for further downstream analysis.

```
samtools  merge  /full/path/to/files/CombinedID.merged.bam
/full/path/to/files/file_R1.C01.fastq.gz.adapterTrim.round2.rmRep.rmDup.s
orted.bam
/full/path/to/files/file_R1.D08.fastq.gz.adapterTrim.round2.rmRep.rmDup.s
orted.bam
```

samtools index: Takes output from sortSam, makes bam index for use downstream.

```
samtools  index  /full/path/to/files/CombinedID.merged.bam
/full/path/to/files/CombinedID.merged.bam.bai
```

****samtools view**: Takes output from sortSam. Only outputs the second read in each pair for use with single stranded peak caller. This is the final bam file to perform analysis on.

```
samtools view -hb -f 128  /full/path/to/files/CombinedID.merged.bam  >
/full/path/to/files/CombinedID.merged.r2.bam
```

make_bigwig_files.py: Takes input from samtools view. Makes bw files to be uploaded to the genome browser or for other visualization.

```
make_bigwig_files.py  --bam /full/path/to/files/CombinedID.merged.r2.bam
--genome /path/to/hg19.chrom.sizes  --bw_pos
/full/path/to/files/CombinedID.merged.r2.norm.pos.bw  --bw_neg
/full/path/to/files/CombinedID.merged.r2.norm.neg.bw
```

Clipper: Takes results from samtools view. Calls peaks on those files.

```
clipper  -b /full/path/to/files/CombinedID.merged.r2.bam  -s hg19  -o
/full/path/to/files/CombinedID.merged.r2.peaks.bed  --bonferroni  --
superlocal  --threshold-method binomial  --save-pickle
```

fix_scores.py: Takes input from clipper: Fixes p-values to be bed compatible

```
python ~/gscripts/gscripts/clipseq/fix_scores.py  --bed
/full/path/to/files/CombinedID.merged.r2.peaks.bed  --out_file
/full/path/to/files/CombinedID.merged.r2.peaks.fixed.bed
```

bedToBigBed: Converts bed file to bigBed file for uploading to the genomeBrowser .

```
bedToBigBed /full/path/to/files/CombinedID.merged.r2.peaks.fixed.bed
/path/to/hg19.chrom.sizes
/full/path/to/files/CombinedID.merged.r2.peaks.fixed.bb -type=bed6+4
```

Peak normalization vs SMInput

Peak normalization vs paired SMInput datasets is run as a second processing pipeline (*Peak_input_normalization_wrapper.pl*). Input files for normalization pipeline include .bam and .peak.bed files (generated through the pipeline above), as well as a manifest file pairing eCLIP datasets with their paired SMInput datasets as follows:

```
uID  \t   RBP  \t   Cell line   \t    CLIP_rep1    \t   CLIP_rep2   \t    INPUT
001     RBP1    HepG2   /full/path/to/files/CombinedID_rep1.merged.r2.bam
        /full/path/to/files/CombinedID_rep2.merged.r2.bam
        /full/path/to/files/CombinedID_INPUT.merged.r2.bam
002     RBP2    K562    /full/path/to/files/CombinedID2_rep1.merged.r2.bam
        /full/path/to/files/CombinedID2_rep2.merged.r2.bam
        /full/path/to/files/CombinedID2_INPUT.merged.r2.bam
```

uID = a unique identifier for each experiment
RBP, Cell line = dataset descriptors (for labeling purposes, not used in pipeline itself)
CLIP_rep1 = full path to CLIP (replicate 1) bam file (output from ****samtools view** above)
CLIP_rep2 = full path to CLIP (replicate 2) bam file (output from ****samtools view** above) (this field is OPTIONAL – it is used for for ENCODE-style experiments with 2 eCLIP paired with 1 SMInput. For experiments where each CLIP has a paired SMInput, simply remove this column and use a 5 column manifest file).
INPUT = full path to paired SMInput bam file (output from ****samtools view** above)

- Note that this pipeline expects the .peak.bed files to be in the same folder as .bam files (as is standard output from the processing pipeline)

Final output for this pipeline is an input normalized peak file (in bed format):

```
/full/path/to/desired_output_directory/uID_Rep.basedon_uID_Rep.peaks.l2inputnor
mnew.bed.compressed.bed
```

Formatted as follows:

```
Chr \t start \t stop \t log10(p-value eCLIP vs SMInput) \t log2(fold-enrichment
in eCLIP vs SMInput) \t strand
```

Command line example:

```
>perl Peak_input_normalization_wrapper.pl /full/path/to/manfest_file.txt
/full/path/to/desired_output_directory/
```

The wrapper performs the following steps:

1) Create output directory (if not existing)
2) Create soft-links in output directory to .bam and .peak.bed files listed in manifest

3) Count usable read numbers for each .bam file (`samtools view -c -F 4`) and write to a tab-delimited text file (`/full/path/to/manfest_file.txt.mapped_read_num`)
4) Runs overlap_peakfi_with_bam_PE.pl:
   Takes in two bam files (eCLIP and SMInput) and a bed file of peak regions. For each peak, counts the # of overlapping reads in eCLIP and SMInput bam files, and performs Yates' Chi-square (in Perl) or Fisher's Exact Test (using the R statistics package) to determine enrichment significance in eCLIP relative to SMInput.

```
perl overlap_peakfi_with_bam_PE.pl
/full/path/to/desired_output_directory/CombinedID_rep1.merged.r2.bam
/full/path/to/desired_output_directory/CombinedID_INPUT.merged.r2.bam
/full/path/to/desired_output_directory/CombinedID_rep1.merged.r2.peaks.bed
/full/path/to/manfest_file.txt.mapped_read_num
/full/path/to/desired_output_directory/uID_Rep.basedon_uID_Rep.peaks.l2inputnor
mnew.bed
```

Output file has bed format:

```
Chr \t start \t stop \t log10(p-value eCLIP vs SMInput) \t log2(fold-enrichment
in eCLIP vs SMInput) \t strand
```

5) Runs compress_l2foldenrpeakfi_for_replicate_overlapping_bedformat.pl:
   As CLIPper is run at the transcript level, occasionally multiple clusters can overlap the same genomic positions. This script steps through all peaks, and resolves overlapping peaks by only keeping the peak with the greater enrichment in eCLIP over SMInput.

```
perl compress_l2foldenrpeakfi.pl
/full/path/to/desired_output_directory/uID_Rep.basedon_uID_Rep.peaks.l2inputnor
mnew.bed
```

Writes output to bed format file (same columns as above):

```
/full/path/to/desired_output_directory/uID_Rep.basedon_uID_Rep.peaks.l2inputnor
mnew.bed.compressed.bed
```