


High-resolution mapping of function and protein binding in an RNA nuclear enrichment sequence

 Yoav Lubelsky , Binyamin Zuckerman  & Igor Ulitsky* 

Abstract

The functions of long RNAs, including mRNAs and long noncoding RNAs (lncRNAs), critically depend on their subcellular localization. The identity of the sequences that dictate subcellular localization and their high-resolution anatomy remain largely unknown. We used a suite of massively parallel RNA assays and libraries containing thousands of sequence variants to pinpoint the functional features within the SIRLOIN element, which dictates nuclear enrichment through hnRNPK recruitment. In addition, we profiled the endogenous SIRLOIN RNA-nucleoprotein complex and identified the nuclear RNA-binding proteins SLTM and SNRNP70 as novel SIRLOIN binders. Taken together, using massively parallel assays, we identified the features that dictate binding of hnRNPK, SLTM, and SNRNP70 to SIRLOIN and found that these factors are jointly required for SIRLOIN activity. Our study thus provides a roadmap for high-throughput dissection of functional sequence elements in long RNAs.

Keywords hnRNPK; massively parallel assays; nuclear retention; ribonucleoprotein complex; RNA nuclear export

Subject Categories Computational Biology; Methods & Resources; RNA Biology

DOI 10.15252/embj.2020106357 | Received 27 July 2020 | Revised 14 March 2021 | Accepted 24 March 2021

The EMBO Journal (2021) e106357

Introduction

Transcription takes place almost exclusively in the nucleus, while many RNA-dependent cellular activities, including translation, occur in the cytoplasm. The export of RNA from the nucleus to the cytoplasm is of particular interest and remains poorly understood. Export of most long RNA molecules is thought to be a fast and efficient process, while some RNAs need to be retained in the nucleus. Nuclear retention is key to allow some RNAs to carry out regulatory functions as long noncoding RNAs (lncRNAs) or as a regulatory checkpoint that delays export of some mRNAs until appropriate signaling cues are present (Wickramasinghe & Laskey, 2015). It is thought to rely, at least in part, on specific sequences or “zipcodes” that are recognized by sequence-specific RNA-binding proteins

(RBPs), but the identity and modes of action of these elements remain poorly understood (Palazzo & Lee, 2018). The core set of proteins that mediate nuclear export of endogenous and viral RNAs has been identified in various species (Carmody & Wentz, 2009), and recent studies have refined the sets of transcripts regulated by them (Lee *et al.*, 2020; Zuckerman *et al.*, 2020). Within the elements that have been shown to affect RNA localization, the contribution of specific sequence motifs or structures to function or to RBP binding is largely unknown.

We have previously used a massively parallel RNA assay (MPRNA) to measure the ability of ~ 6,000 110-nt sequence tiles to drive nuclear enrichment of an AcGFP mRNA, that is otherwise efficiently exported to the cytoplasm (Lubelsky & Ulitsky, 2018). The library of sequence tiles we used (NucLibA) was derived from nuclear human lncRNAs and 3' UTRs of nuclear-enriched mouse mRNAs. Analysis of consecutive and overlapping tiles associated with nuclear enrichment identified the SIRLOIN (SINE-derived nuclear RNA LocalizatioN) element, a specific region within Alu transposable elements integrated in an antisense orientation within transcribed units and represented in four different lncRNAs in NucLibA. Based on this 42-nt sequence, we designed an additional library (NucLibB) and used it in an MPRNA that identified a region of ~12 nts centered on a GCCUCCC element that was essential for SIRLOIN function. Computational analysis of ENCODE eCLIP data and Alu sequences predicted that hnRNPK, an abundant nuclear RNA-binding protein, binds SIRLOIN. hnRNPK recognizes C-rich motifs with a particular preference for CCC repeats (Moritz *et al.*, 2014; Dominguez *et al.*, 2018). We validated that hnRNPK preferentially binds an AcGFP bearing a SIRLOIN element in its 3'UTR and that depletion of hnRNPK abolishes nuclear enrichment of SIRLOIN-containing RNAs (Lubelsky & Ulitsky, 2018). Importantly, other studies using a similar approach identified C-rich elements as driving nuclear enrichment in other contexts (Shukla *et al.*, 2018). There is therefore evidence that the presence of the SIRLOIN element, which binds hnRNPK, dictates nuclear enrichment of the host RNA. It is unknown whether hnRNPK recruitment is sufficient for SIRLOIN activity, whether bases that do not bind hnRNPK are important, and whether other factors play a role in nuclear retention of SIRLOIN-containing RNAs.

Here, we use a suite of transcriptomic and proteomic methods to dissect the grammar of the molecular recognition between hnRNPK and the SIRLOIN element and identify additional proteins that are

required for SIRLOIN function. We find that while there is a tight correlation between the ability of SIRLOIN variants to recruit hnRNPK and their ability to drive nuclear enrichment, only hnRNPK bound in a specific position and sequence context within SIRLOIN is able to drive nuclear enrichment. SIRLOIN is also bound by at least two additional proteins, SLTM and SNRNP70, that recognize overlapping sequence elements within SIRLOIN and that are also required for nuclear enrichment of endogenous SIRLOIN-containing transcripts. Importantly, we show that systematic transcriptomic and proteomic screens are able to map sequence-binding-function axes within functional modules in mRNAs and lncRNAs.

Results

A high-throughput approach for studying the effects of sequence variation on hnRNPK binding

In order to study the sequence landscape dictating binding of hnRNPK to the SIRLOIN element, we devised a screen combining our MPRNA setup with RNA immunoprecipitation of hnRNPK (MPRNA-RIP; Fig 1A). We first used NucLibB, a library that contains tiles from several lncRNAs and mRNAs that contain the SIRLOIN element, systemic mutagenesis of two SIRLOIN-containing tiles, Jpx#9 and Pvt1#22, and several additional sequences (Lubelsky & Ulitsky, 2018). We transfected the ~2,000 plasmids encoding AcGFP mRNA with NucLibB tiles integrated into the 3'UTR into MCF-7 cells, immunoprecipitated endogenous hnRNPK using a specific antibody (Fig 1A), and extracted RNAs associated with it. We separately sequenced endogenous polyadenylated RNAs (RIP-seq) and library fragments embedded in the AcGFP 3' UTR (MPRNA-RIP). Hundreds of endogenous transcripts were significantly enriched in the IP sample (Fig 1B and Table EV1). As expected, RNAs containing hnRNPK eCLIP clusters in ENCODE data were enriched in the IP sample (Fig 1C) as were nuclear-enriched transcripts (Fig 1D). Furthermore, hnRNPK-bound transcripts preferentially lost their nuclear enrichment and were down-regulated when hnRNPK was depleted in MCF-7 cells (Fig 1E and F). Interestingly, there was no correlation between changes in localization and changes in expression for the bound genes (Spearman $R = -0.12$; $P = 0.14$), and repressed transcripts were preferentially slightly more cytoplasmic at baseline ($R = 0.16$; $P = 0.025$).

For NucLibB tiles, we used the hnRNPK IP and the input libraries to compute hnRNPK binding strength (IP/Input), which was highly concordant between biological replicates (Spearman $R = 0.56$ – 0.86). Comparison of binding strength to the previously measured Nuc/Cyto ratios of the same NucLibB sequences (Lubelsky & Ulitsky, 2018) revealed a strong association between hnRNPK binding and nuclear enrichment (Spearman $R = 0.63$ $P < 10^{-15}$; Fig 1G), suggesting that the ability to efficiently recruit hnRNPK is a central hallmark of effective SIRLOIN elements. Nevertheless, we also observed substantial variation between sequences in binding, which spanned a ~16-fold range, which motivated us to further interrogate SIRLOIN sequence-binding-function axes.

NucLibB contained sequences composed of repeats of each of the 6-mers and 10-mers found in the core ~30-nt regions of SIRLOIN in Jpx#9 and Pvt1#22 tiles, separated by AT dinucleotides. As previously reported (Lubelsky & Ulitsky, 2018), these repetitive

sequences were much less effective than full-length SIRLOINs in eliciting nuclear enrichment (Fig 1H). In contrast, some of these repetitive sequences, in particular those with stretches of four or more pyrimidines, were effective in binding hnRNPK, sometimes better than the full-length SIRLOIN-bearing tiles (Fig 1H). Repeats of k-mers containing three or more consecutive purines were less effective in hnRNPK binding, even when they also contained pyrimidine stretches. These results further supported the notion that effective binding by hnRNPK is required, but not sufficient for enriching RNA in the nucleus, which also relies on additional bases that are not captured in any of the 6-mer and 10-mers. We note that we have previously shown that repeating the whole SIRLOIN core three times results in a more effective nuclear enrichment than the WT SIRLOIN sequence (Lubelsky & Ulitsky, 2018).

An oligonucleotide library enabling detailed interrogation of the variation landscape

In order to further characterize SIRLOIN architecture with MPRNA and MPRNA-RIP assays, we designed a new library, NucLibC (Fig 2A and Table EV2). Most of the sequences in NucLibC were based on the 109 nt Jpx#9 tile, which in NucLibA and NucLibB drove ~2-fold nuclear enrichment of AcGFP (Lubelsky & Ulitsky, 2018). In NucLibC, we extended the 29-nt region that was mutated in NucLibB and systematically mutated every one of the 54 bases at the 3' of Jpx#9. Within these 54 bases, we also introduced more extensive perturbations, including the following: (i) A ↔ T and G ↔ C changes in consecutive 2-, 4-, 6-, 10-, and 20-mers; (ii) shuffling of the sequence; (iii) deletions of 1–25 consecutive bases starting from each of 26 positions; (iv) all possible double mutations within a shorter window of 29 bases which flank the GCCUCCC core; (v) insertions of a strong hnRNPK binding site CCUCCC, a mutated CCAGCC site, a strong hairpin structure, or a control structure, at each possible position (Fig 2A and Table EV2).

Insertions and deletions were performed both in the context of the WT sequence and in the context of a mutated Jpx#9 ("Jpx#9 mut"), where the GCCUCCC core found 9 bases from the end of the sequence was mutated to GCAUCCC, a change that we previously showed to be sufficient to completely abolish Jpx#9 nuclear enrichment activity (Lubelsky & Ulitsky, 2018). Altogether, NucLibC contained 3,811 sequence variants of Jpx#9, 3,749 of which (98.4%) were successfully synthesized, cloned, and expressed. Importantly, all the sequence variants were of precisely the same length of 109 nt. When inserting new elements, we omitted 5' Jpx#9 sequences which exceeded this length limit, and when deleting sequences, we inserted the deleted fragment at the 5' of the tile, keeping the overall composition of the sequence, and the alignment to the 3' of the reporter fixed, facilitating a comparison that is not confounded by the effects of transcript length or distance to the 3' end of the transcript.

We used NucLibC for an MPRNA of nuclear localization and hnRNPK RIP-MPRNA, which allowed us to compare the effects of the sequence changes on both function (change in Nuc/Cyto ratios) and hnRNPK binding strength (Fig 2B and Dataset EV1). Mutations in the 19-nt region centered at GCCUCCC (which we will refer to as "SIRLOIN core," ivory-colored in Fig 2B), affected nuclear enrichment, consistent with our previous observations (Lubelsky & Ulitsky, 2018). In contrast, mutations upstream of these 19 bases had

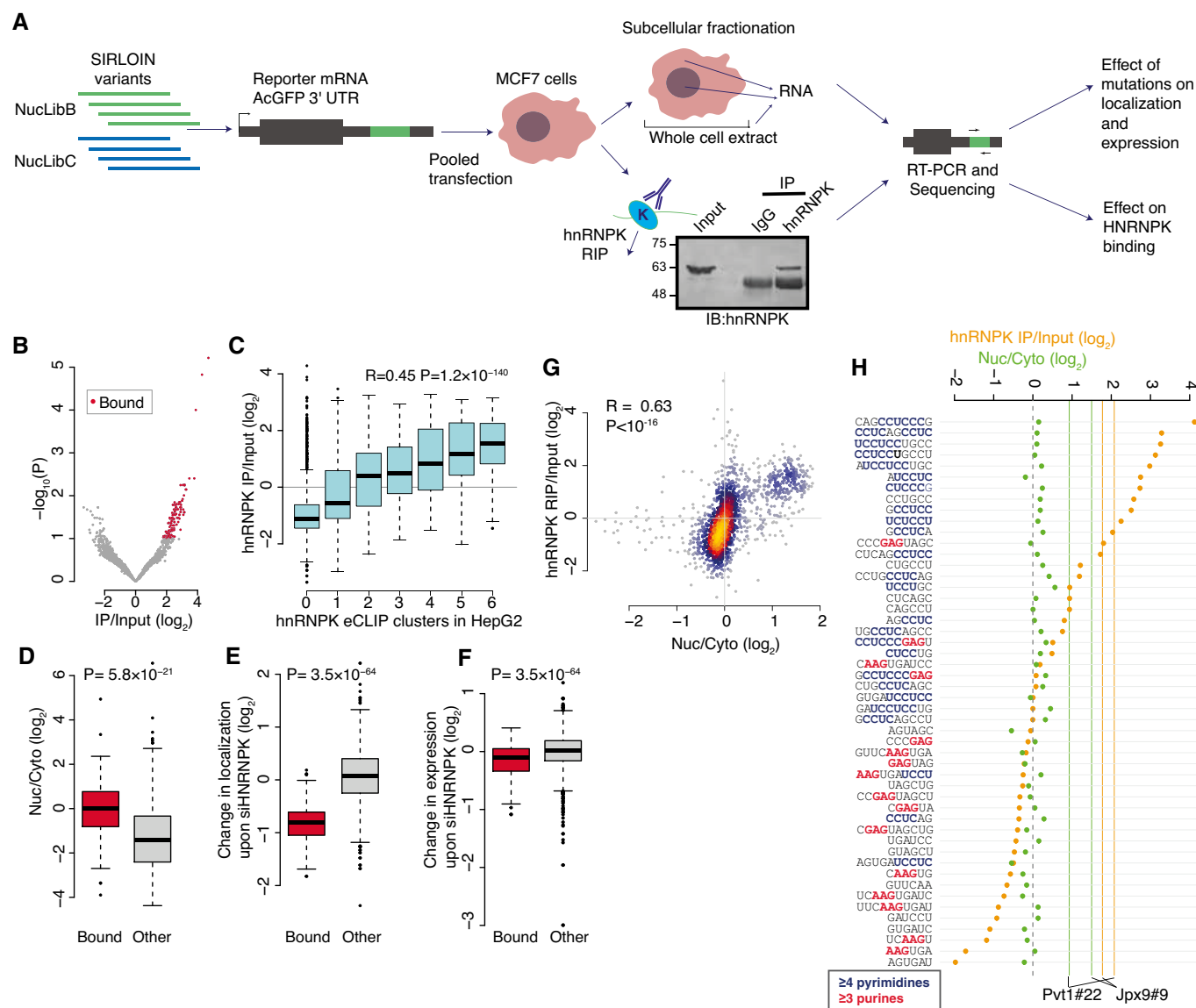


Figure 1. Identification of transcripts bound by hnRNP.

- A MPRNA scheme and hnRNP IP Western blot. The blot image is a section of the image shown in Fig EV6A.
- B Volcano plot of the RIP-seq data. Each point represents a gene. Red points correspond to the 146 genes with $\log_2(\text{IP}/\text{Input}) \geq 0.5$ and adjusted $P < 0.1$. P -values computed using Wald test as implemented in DESeq2.
- C hnRNP IP enrichment ratios for genes with the indicated number of exonic hnRNP eCLIP clusters in ENCODE data from HepG2 cells (average of the two ENCODE replicates). Enrichment ratios are \log_2 -transformed fold changes computed by DESeq2 based on four replicates. The box plots show the interquartile range (IQR) and the line indicates the median value. Whisker ends extend to 1.5 times the length of the IQR (unless the data range is smaller, in which case the whisker ends extend to the minimum or maximum value), points indicate outliers. Spearman's correlation R and P -value are indicated.
- D Nuc/Cyto ratios in MCF-7 cells (ENCODE data) for genes bound by hnRNP (from (B)) and other genes. Enrichment ratios and boxplots are as in C, P -value computed using two-sided Wilcoxon rank-sum test.
- E Change in the Nuc/Cyto ratios (\log_2) between sihnRNP- and siNT-treated cells, for the genes bound by hnRNP (from (B)) and other genes. Enrichment ratios, boxplots, and statistical test are as in D.
- F As in (E), for changes in gene expression, computed by DESeq2.
- G Correspondence between Nuc/Cyto ratios induced by each tile in NuLibB and its hnRNP IP/Input ratio. Coloring indicates local point density. Spearman's correlation R and P -value are indicated.
- H For each sequence in NuLibB containing repeats of the indicated 6-mer or 10-mer, the Nuc/Cyto (green) and hnRNP IP/Input (orange) ratios are shown. Vertical lines show the corresponding values for the Jpx9 and Pvt1#22 tiles.

no evident effect on function. When considering hnRNP binding, mutations in the CCUCC region affected hnRNP binding most strongly, consistent with its known binding preference (Moritz *et al*,

2014; Dominguez *et al*, 2018), whereas mutations in the other 13 bases had a less consistent effect (green region in Fig 2B). Conversely, mutations in an upstream region centered at another

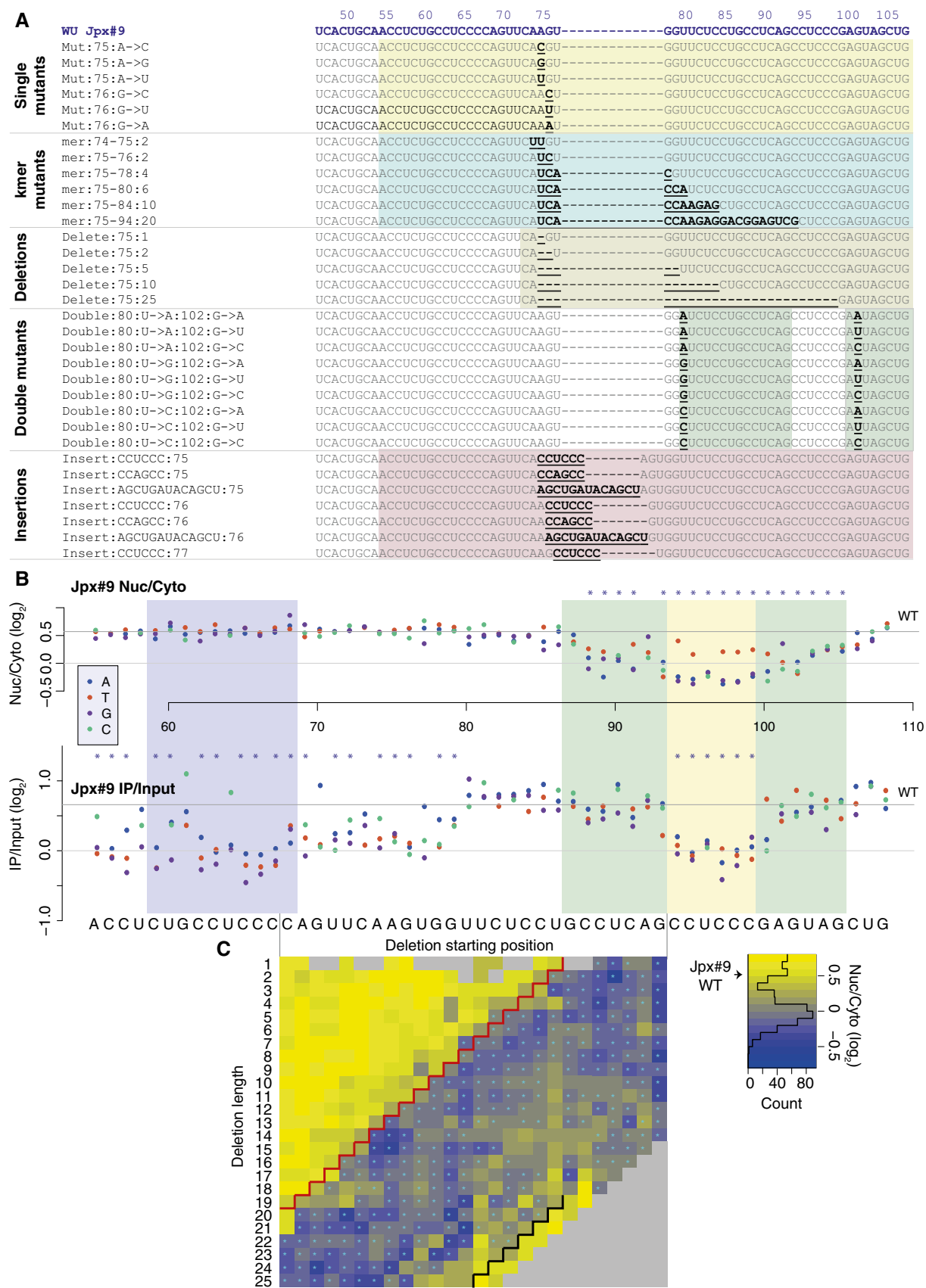


Figure 2.

Figure 2. Interrogation of SIRLOIN sequence-binding-function axes using NuLibC.

- A Groups of Jpx#9 sequence variants in NuLibC. Just the indicated bases from the 3' of the tile are shown. The WT sequence is on top and few representative variants out of the 3,810 variants in NuLibC are shown. The altered regions are in bold and underlined. For each sequence, the oligo name is shown and it includes the position(s) that were changed and the details of the change. Variants are grouped by the type of change made, and the region in which changes were introduced is shaded.
- B Nuc/Cyto (top) and IP/Input (bottom) ratios for single mutations in NuLibC. The WT sequence is shown at the bottom. Each point is a sequence variant, color-coded based on the base that was introduced. Shaded regions correspond to regions where mutations affect binding but not function (purple), affect function and binding (ivory), and affect function with small effects on binding (green). Asterisks indicate positions where the Nuc/Cyto ratios of the mutated tiles are significantly different than those of the WT tile (Wilcoxon two-sided rank-sum test $P < 0.05$).
- C Color-coded Nuc/Cyto ratios of Jpx#9 variants where the indicated number of bases were deleted (and moved to the 5' of the sequence), starting from the indicated position (aligned to the sequence in (B)). Deletions to the right of the red line affect the SIRLOIN core region. Deletions to the right of the black line include the entire SIRLOIN core region (the deleted region is relocated to the 5' of the tile). Asterisks indicate positions where the Nuc/Cyto ratios of the tile with the specific deletion are significantly different than those of the WT tile (Wilcoxon two-sided rank-sum test $P < 0.05$). The histogram on the right shows distribution of the presented Nuc/Cyto ratios.

GCCUCCC sequence found ~ 20 nt upstream of the SIRLOIN core strongly affected hnRNPK binding, but had no effect on nuclear enrichment (purple region in Fig 2B). Specifically, X → C mutations that introduced new CCC motifs increased hnRNPK binding but had no discernible effect on function. We conclude that hnRNPK binding to the 3' GCCUCCC is essential for SIRLOIN function, ~ 7 bases upstream and downstream of this motif are important for function but have a minor effect on binding, and binding of hnRNPK to other regions in Jpx#9 has a substantially lower contribution, if any, to SIRLOIN function.

We next examined the effect of deletion of 1–25 bases from different positions in the 3' part of Jpx#9. Nuc/Cyto values of these sequence variants followed a bimodal distribution, whereas deletions in the mutated Jpx#9 sequences were largely non-functional (Figs 2C and EV1). Deletions that included bases from the SIRLOIN core (right of the red line in Fig 2C) affected function, whereas deletions that were restricted to the upstream region had no evident effect. Large deletions traversing the entire SIRLOIN core (right of the black line in Fig 2C) were partially functional, presumably because they included most of the SIRLOIN core, as our NuLibC design inserted the deleted sequence at the 5' region of Jpx#9, such that the SIRLOIN core was now fully relocated to the beginning of these tiles. The reduced functionality of these sequences is consistent with the stronger SIRLOIN activity when it is found closer to the 3' end of the host transcript (Lubelsky & Ulitsky, 2018). Deletion analysis thus supports the critical importance of the SIRLOIN core for function of the Jpx#9 tile, the dispensability of the sequences upstream of the core, and the weaker yet evident effect of the position of SIRLOIN core within the transcript.

Additional hnRNPK binding sites improve binding but do not increase nuclear enrichment

A subset of NuLibC sequences included additional CCUCCC sequences, which can serve as potential hnRNPK binding sites, inserted at 54 different positions in Jpx#9. As a control, we inserted CCAGCC sequences that are not predicted to bind hnRNPK. In parallel, in order to study the effect of secondary structure, we also tested the effects of inducing a stable hairpin with 5 base-pairs AGCU-GAUACAGCU and a control region of the same length but no strong structure (ACAGCAUACAGCU). Each sequence was inserted into WT Jpx#9 as well as into the non-functional Jpx#9 mut.

When considering binding (Fig 3A, bottom), the sequences with various insertions exhibited a broad distribution, with addition of

CCUCCC motifs, as well as addition of CCAGCC in specific positions flanked by additional Cs (white arrows in Fig 3A) led to an increase in hnRNPK IP/Input ratios compared to WT and other positions. In contrast, when considering SIRLOIN functionality (Fig 3A, middle), we observed a bimodal distribution of Nuc/Cyto ratios. Addition of hnRNPK binding sites did not boost nuclear enrichment, and in contrast, insertion of CCUCCC or any of the other sequences within the SIRLOIN core abrogated SIRLOIN activity. Within the mutated Jpx#9 sequences, insertions of hnRNPK binding sites were insufficient for nuclear enrichment with the possible exception of CCAGCC insertions in specific locations containing flanking Cs (marked by yellow arrows in Fig 3A). Interestingly, insertion of a hairpin sequence at the very end of Jpx#9 increased nuclear enrichment (turquoise arrows in Fig 3A), but in a manner that appeared to be independent of a functional SIRLOIN core and of hnRNPK binding (as RIP/Input values for these sequences were not different than those of controls).

Overall, when considering together all the different variants of Jpx#9, we found that mutations that increased or decreased the number of CCC elements in the sequence had a strong corresponding effect on hnRNPK binding, but with no corresponding effect on function (Fig 3B), which relied on binding at the specific position of the GCCUCCC element, and so was typically not influenced by changes in hnRNPK binding at other positions. We conclude that a functional SIRLOIN element contains a precisely positioned and essential hnRNPK binding site, that increased binding by hnRNPK in other regions does not lead to stronger nuclear enrichment, and that functional hnRNPK binding does not appear to be influenced by drastic changes of the sequence or structure in the flanking regions.

No evident role of RNA structure within the sequences in the SIRLOIN core

In order to study the potential contribution of paired bases in the core region of SIRLOIN, we analyzed the subset of NuLibC that contained pairs of all possible double mutations in two regions—14 bases upstream and 8 bases downstream of the CCUCCC core. For each of the 2,079 combinations, we compared the “function” (Nuc/Cyto) and “binding” (IP/Input) scores to those of the corresponding single mutations. The double mutants were typically more cytoplasmic than the two corresponding single mutants, with few notable exceptions, which were almost exclusively X → C mutations (Fig 4A). When we examined the positions where a double mutation “rescued” substantially (≥ 0.3 log₂ units increase in Nuc/Cyto

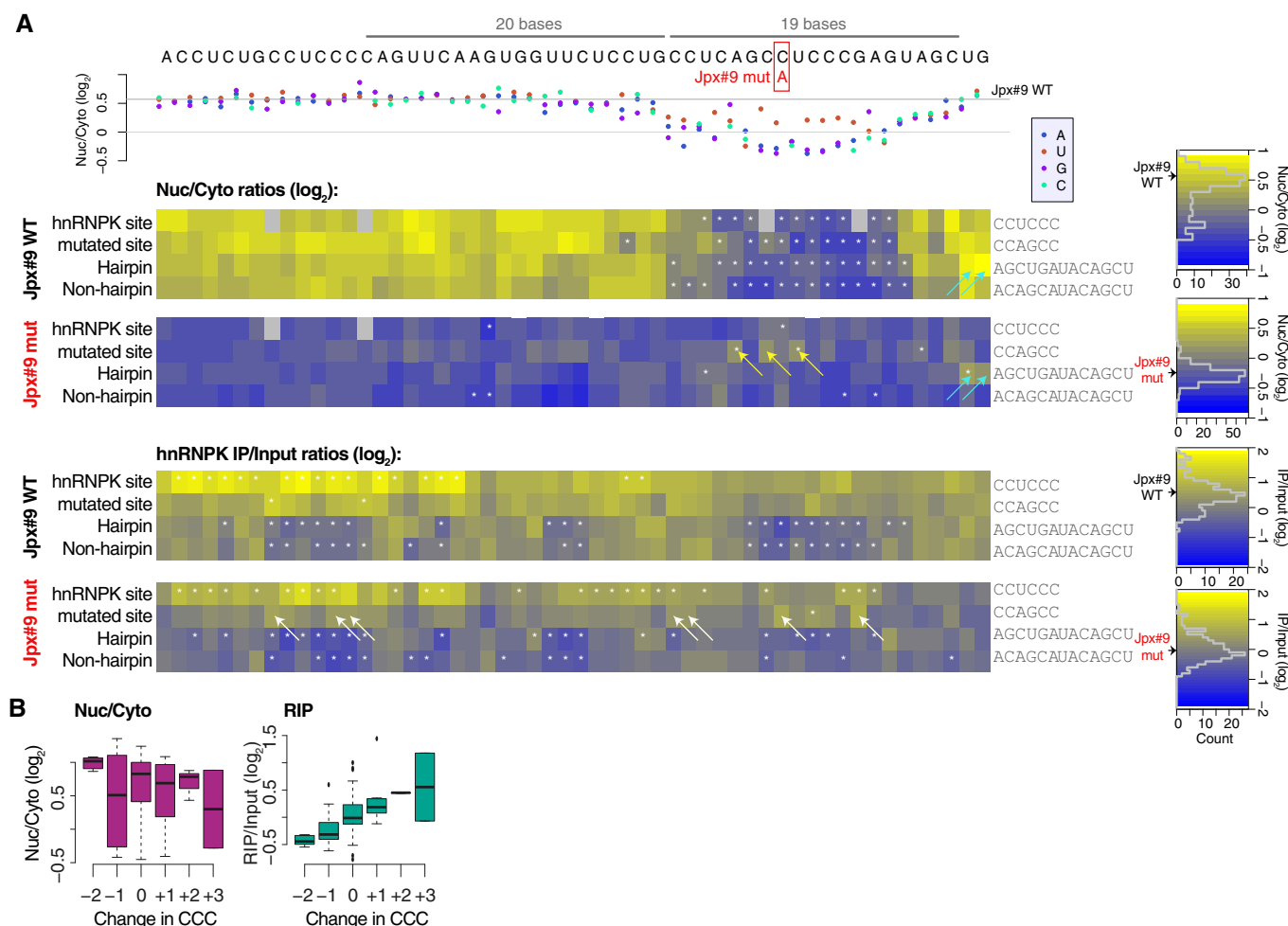


Figure 3. Effect of additional hnRNPK binding sites on SIRLOIN function.

A Top: same as the top part of Fig 2B. Middle/Bottom: Nuc/Cyto (middle) and IP/Input (bottom) ratios for variants of Jpx#9 in which the indicated sequence was introduced at the indicated position (aligned to the sequence on top) in either the WT or the mutated Jpx#9 context. Yellow, turquoise, and white arrows indicate the tiles mentioned in the text. The histograms on the right indicate the distributions of the values in each heatmap. Asterisks indicate positions where the Nuc/Cyto ratios of the mutated tiles are significantly different than those of the WT tile (Wilcoxon two-sided rank-sum test $P < 0.05$).

B Nuc/Cyto and RIP/Input ratios for sequences containing the indicated number of changes in the overall number of CCC motifs relative to the WT Jpx#9 sequence, when considering all the Jpx#9 variants in NucLibC. Ratios computed based on three replicates. Boxplots are as in Fig 1C.

ratio) over the more cytoplasmic of the two single mutants, there was a notable enrichment at position 81, where $U \rightarrow C$ mutations converted $UUUCC$ into a $UCCUCC$ element, resembling the $GCCUCC$ element in the SIRLOIN core (Figs 4B and EV2A, middle). Such mutations could functionally compensate for mutations in 6 bases upstream and 2 bases downstream of the $CCUCCC$ in the SIRLOIN core (Fig EV2A). The predicted local structure of the $GCCUCC$ core and its surrounding 60 nt on each side suggests it is a long loosely paired dsRNA (Fig EV2B). Interestingly, both the $GCCUCC$ in the SIRLOIN core and the “alternative” $UCCUCC$ element are found on the same side of the predicted structure in a somewhat similar structural context, which may contribute to their function. Nevertheless, the fact that deletions or insertions of structured or unstructured elements in the regions almost immediately flanking the main hnRNPK binding site had limited effect on function suggests that broader structural context is not very important for SIRLOIN function.

SLTM and SNRNP70 bind Jpx#9 tile in the AcGFP context

In order to identify additional factors that potentially associate with the SIRLOIN element, we used RAP-MS (McHugh & Guttman, 2018) (Fig 5A). MCF-7 cells, stably expressing AcGFP mRNA bearing either a short control 3' UTR or a 3' UTR containing the Jpx#9 tile —“AcGFP[Jpx#9]”, were used. A pool of 32 biotinylated antisense ssDNA oligos 80 nt each was then used to enrich the AcGFP mRNP (Figs 5B and EV3A). The protein constituents of the RNP were characterized using mass spectrometry (MS) and compared to proteins recovered in several control conditions (Fig 5A and Materials and Methods). Peptides originating from five proteins (SLTM, SNRNP70, EMC2, THBS1, and NOLC1) were enriched by at least 2-fold compared to a no-crosslinking control and by at least 1.4-fold compared to an empty 3'UTR (Dataset EV2). Three of these are known nuclear RNA-binding proteins (SLTM, SNRNP70, and NOLC1). We attempted to validate their binding to AcGFP[Jpx#9]

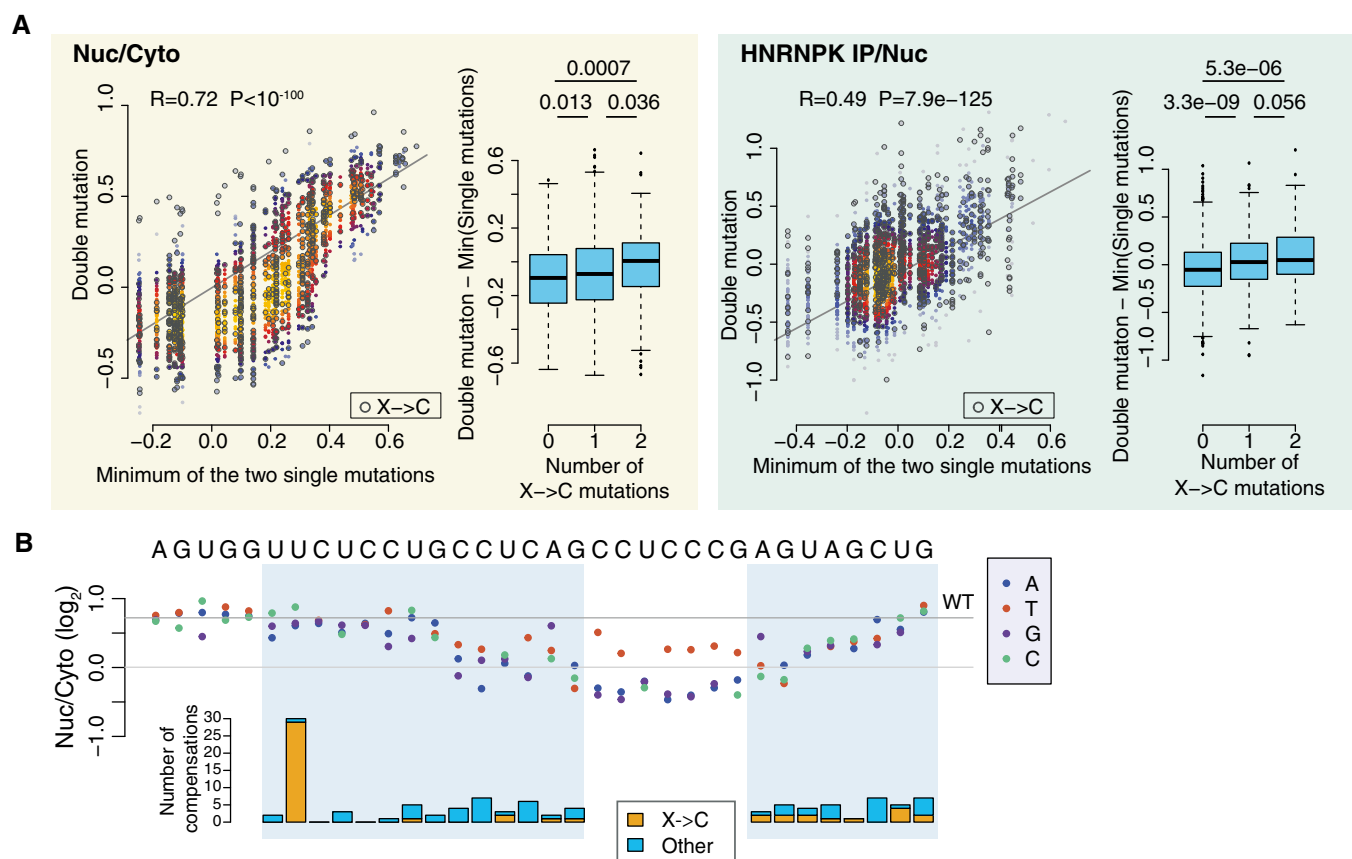


Figure 4. Effects of pairs of mutations on SIRLOIN function.

- A** Correspondence between the Nuc/Cyto (left) or hnRNPK IP/Input (right) ratios of sequences carrying a double mutation and the minimum Nuc/Cyto ratio of the two sequences which carry only one of the two mutations. Spearman's R and P -values are indicated. Color indicates point density. Boxplots compare the difference between the values on the y-axis and the x-axis for sequences with the indicated number of X → C mutations. Ratios are based on two replicates for Nuc/Cyto and three replicates for RIP. Boxplots are as in Fig 1C. Spearman's correlation R and P -value are indicated above the scatter plots. Pairwise comparison P -values shown above the boxplots were computed using two-sided Wilcoxon rank-sum test.
- B** Top: effect of single point mutations on Nuc/Cyto ratios, from Fig 2B. Bottom: The number of sequences with double mutations where the difference between the double mutant and the minimum of the two single mutants was larger than 0.3 \log_2 units, separating counting the cases when both mutations at the indicated positions are X → C.

using RIP and were successful in IP for SLTM and SNRNP70 which indeed bound AcGFP[Jpx#9] (Fig 5C and D).

SNRNP70 is a core component of the U1 snRNP, which was associated with nuclear retention in studies of individual reporters (see Discussion) and in a recent high-throughput screen (Yin *et al*, 2020). SLTM (SAFB-like, transcription modulator) belongs to the SAFB family of large and abundant nuclear RNA-binding proteins (Norman *et al*, 2016) associated with the “nuclear matrix.” SLTM co-immunoprecipitates with SNRNP70 (Huttlin *et al*, 2017; Bishof *et al*, 2018) and was recently implicated as a component of stress-induced nuclear bodies, but its function is unknown. SLTM was strongly enriched and abundant in the nucleus of MCF-7 cells as was hnRNPK (Fig EV3B). In order to seek support for the involvement of SLTM in SIRLOIN biology, we used the ENCODE eCLIP data (Van Nostrand *et al*, 2020). SLTM was ranked 11th of the 103 factors profiled by eCLIP in HepG2 for enrichment of binding to the SIRLOIN region within Alu repeats (hnRNPK is ranked first by these criteria). The number of SLTM eCLIP clusters on endogenous RNAs was significantly associated with nuclear enrichment in ENCODE

fractionation data for HepG2 cells (Spearman's $R = 0.24$; $P < 1 \times 10^{-15}$; Fig EV3C). The number of SLTM eCLIP clusters was also correlated with the number of hnRNPK clusters in nuclear-enriched transcripts in HepG2 cells ($R = 0.5$; $P < 1 \times 10^{-15}$; Fig EV3D). We note that these SLTM binding events are found almost exclusively outside of Alu elements, due to difficulties to map reads to individual Alu instances in the human genome. Existing data thus implicated SLTM and SNRNP70 as acting in nuclear enrichment and we further studied their importance for SIRLOIN function.

SLTM and SNRNP70 associate with sequences bound by hnRNPK, including SIRLOIN elements

In order to characterize in detail what dictates binding of SLTM and SNRNP70, we performed MPRNA-RIP using antibodies targeting these endogenously expressed proteins in lysates of cells transfected with NucLibB or NucLibC. The three biological replicates showed concordant enrichments (Spearman $R = 0.41$ – 0.57

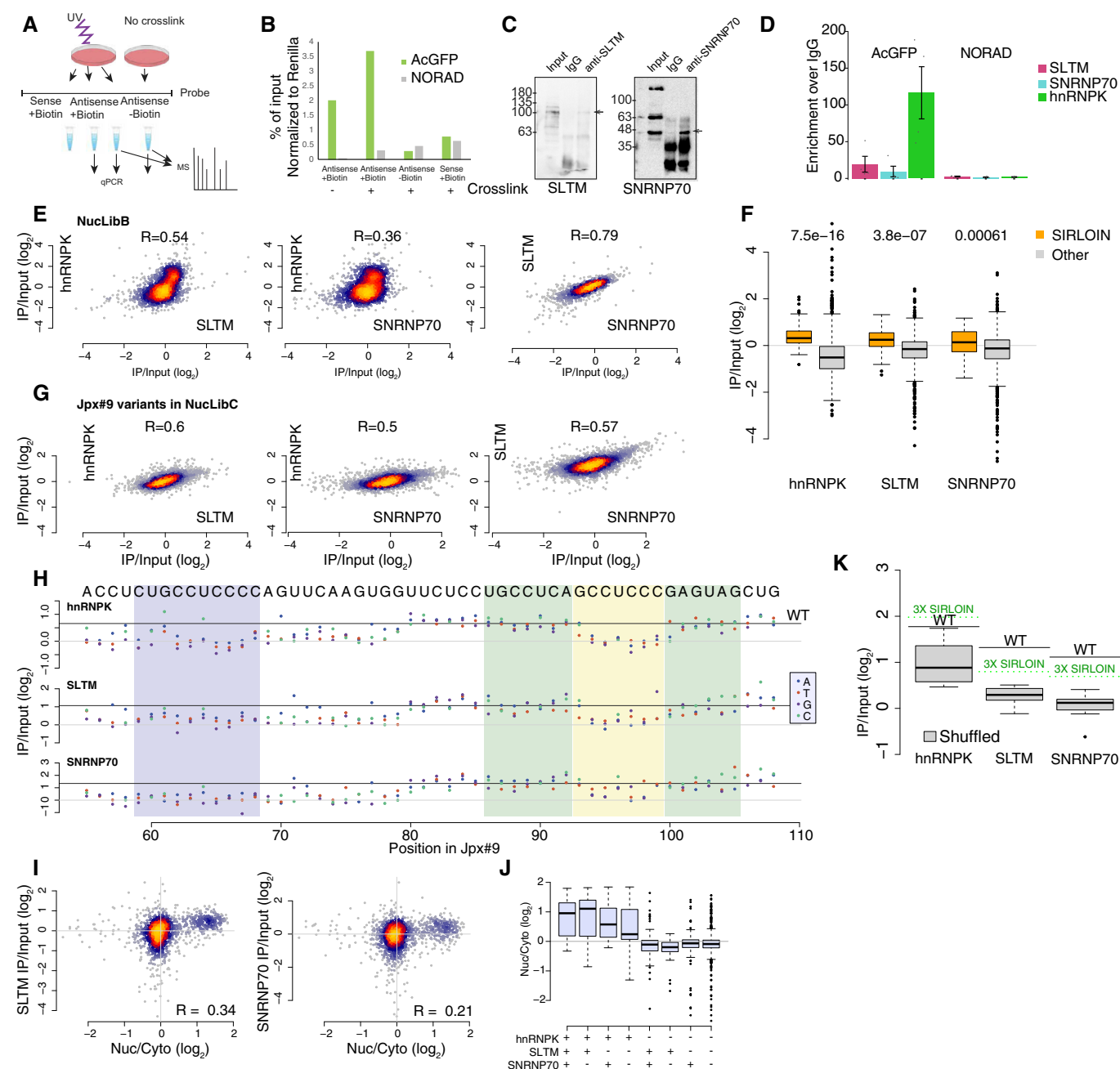


Figure 5. SLTM and SNRNP70 bind the SIRLOIN element.

- A** Outline of the RAP-MS experiment.
- B** Enrichment of the AcGFP mRNA using the antisense biotinylated probes. The NORAD lncRNA is used as a control abundant RNA.
- C** Western blot for the protein indicated below the blot, comparing input MCF-7 cells, and IP with IgG or the indicated antibody. The arrows indicate the expected size of the protein of interest.
- D** RIP-qPCR for the indicated RNA (AcGFP or NORAD) in cells transfected with the AcGFP bearing NucLibB in the 3' UTR, with RIP performed with the indicated antibody and RNA levels determined by qRT-PCR. $N = 5$. Error bars—s.e.m.
- E** Correspondence between binding of the indicated pairs of proteins to NucLibB tiles. Spearman's correlation coefficient is indicated. Color indicates point density.
- F** Comparison of IP/Input ratios for the indicated factors, for NucLibB tiles containing the SIRLOIN element (with up to 6 mismatches) and other tiles.
- G** Same as (E), for NucLibB tiles that contained sequence variants of the Jpx#9 tile.
- H** Same as 2B, for IP/Input ratios for the indicated proteins.
- I** Correspondence between binding of the indicated protein and the Nuc/Cyto ratios induced by NucLibB tiles.
- J** Nuc/Cyto ratios for NucLibB tiles, bound (+, $\log_2(\text{IP}/\text{Input}) \geq 0.5$) or not bound (–, $\log_2(\text{IP}/\text{Input}) < 0.5$) by the indicated factors. Boxplots are as in Fig 1C. P -values computed using two-sided Wilcoxon rank-sum test.
- K** IP/Input ratios for the indicated factors, for ten Jpx#9 sequences with shuffled sequences in bases 80–109, WT sequence of Jpx#9 or 3 repeats of the SIRLOIN element ("3X SIRLOIN"). Boxplots are as in Fig 1C.

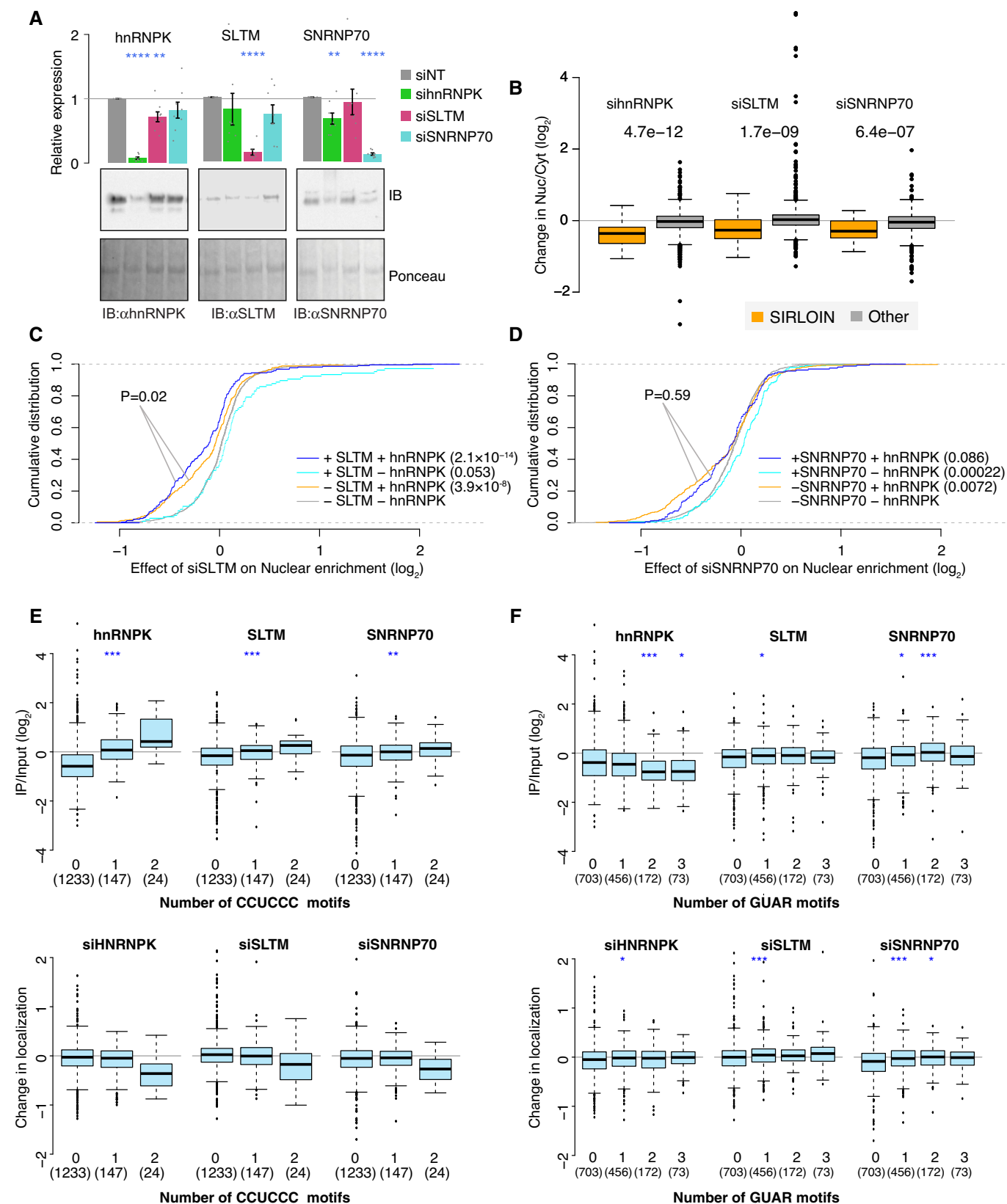


Figure 6.

Figure 6. Knockdown of SLTM or SNRNP70 affects nuclear enrichment of SIRLOIN-containing sequences.

- A mRNA (top) and protein (bottom) levels of the indicated genes in cells transfected with the indicated siRNAs. $n = 8$. Error bars are s.e.m. P -values computed two-sided t -test. $**P \leq 0.01$ and $****P \leq 0.0001$.
- B Differences in the Nuc/Cyto (\log_2) ratios for tiles containing the SIRLOIN element (up to 6 mismatches) and other NuLibB tiles (only WT sequence tiles were considered). Boxplots are as in Fig 1C. P -values computed using two-sided Wilcoxon rank-sum test.
- C, D Same differences as in (B) for tiles with the indicated protein binding pattern (+: $\log_2(\text{IP/Input}) \geq 0.5$; -: $\log_2(\text{IP/Input}) \leq -0.5$). Number in parentheses indicates the P -value when comparing the indicated group and the “– SLTM – hnRNPK” group (C) or “– SNRNP70 – hnRNPK” group (D). P -values computed using two-sided Wilcoxon rank-sum test.
- E, F IP/Input (top) and changes in Nuc/Cyto (\log_2) ratios upon siRNA targeting of the indicated factor (top) for WT NuLibB tiles with the indicated number of the indicated motifs. Number of tiles in each group is indicated in parentheses. $*P < 0.05$, $**P < 0.005$, and $***P < 0.0005$ for comparing the group with the indicated number of matches of the motif with all tiles containing a lower motif count. Boxplots are as in Fig 1C. P -values computed using two-sided Wilcoxon rank-sum test. $*P < 0.05$, $**P < 0.005$, and $***P < 0.0005$.

for SLTM, and $R = 0.42$ – 0.56 for SNRNP70), and we combined the replicates and used DESeq2 (Love *et al*, 2014) to compute IP/Input ratios. These IP/Input ratios were significantly correlated between hnRNPK, SLTM, and SNRNP70 in both NuLibB and NuLibC (Fig 5E), and all three factors preferentially bound SIRLOIN-containing tiles in NuLibB (considering only WT sequences, enriched by 1.83-, 1.33-, and 1.27-fold on average over SIRLOIN-less tiles for hnRNPK, SLTM, and SNRNP70 RIPs, respectively, Fig 5F). There was also a high correlation between the factors when considering just the 3,749 Jpx#9 variants in NuLibC (Fig 5G), showing that sequence variants that affected hnRNPK binding also typically affected binding by SLTM and SNRNP70 or vice versa. This was also evident when inspecting the effects of single mutations on the binding of the three factors to Jpx #9 (Fig 5H). As expected from these similarities, binding of SLTM and SNRNP70 was associated with nuclear enrichment of the bound tiles (Fig 5I). More importantly, when considering binding combinations, tiles co-bound by hnRNPK and SLTM were associated with stronger nuclear enrichment than those bound just by hnRNPK (Fig 5J; $P = 8.3 \times 10^{-8}$, two-sided Wilcoxon rank-sum test) and those bound by hnRNPK and SNRNP70 ($P = 2.5 \times 10^{-3}$ compared with those bound by hnRNPK and not SNRNP70, two-sided Wilcoxon rank-sum test).

The three factors thus exhibited largely similar binding preferences, but there were also some interesting differences between them. For example, NuLibB contains Jpx#9 variants where bases 80–109 encompassing the SIRLOIN core were shuffled. These sequences were still enriched by hnRNPK, although to a 1.8-fold lesser extent than WT Jpx#9 (consistent with their high C-content, and with unchanged hnRNPK binding sites in bases 1–79), whereas no enrichment was evident by SLTM and SNRNP70 (Fig 5K). Sequences containing repeats of the 80–109-nt fragment were 1.15-fold better enriched by hnRNPK than the WT Jpx#9 sequence (Fig 5K), consistent with the increased number of CCC motifs in these sequences, whereas SLTM and SNRNP70 bound 1.43- and 1.33-fold better, respectively, to the WT sequence of Jpx#9, suggesting that they might be more selective to specific features of the functional SIRLOIN element. Similarly, when considering sequences containing repeats of short k-mers, the three factors had overall similar binding preferences (Spearman $R = 0.44$ – 0.65 between pairs of factors; Fig EV4A), but hnRNPK bound poorly to repeats of the CCCGAG sequence, found downstream of the GCCUCCC in the SIRLOIN core, whereas repeats of this sequence were well-bound by SLTM and SNRNP70 (Fig EV4B).

Knockdown of SLTM and SNRNP70 affects nuclear enrichment of SIRLOIN-containing RNAs

In order to evaluate the requirement of SLTM and SNRNP70 for nuclear enrichment of SIRLOIN-containing RNAs, we separately knocked them down using siRNAs in MCF-7 cells transfected with NuLibB or control cells (Fig 6A) and studied the effects on localization of NuLibB tiles. In NuLibB, KD of SLTM or SNRNP70 led to reduction in nuclear enrichment of SIRLOIN-containing tiles compared to other tiles (Fig 6B), with an overall stronger effect for SLTM knockdown. When considering the RIP-MPRNA data, localization of tiles bound by both SLTM and hnRNPK was more affected by SLTM KD than the localization of tiles bound just by hnRNPK (Fig 6C), whereas tiles bound by SLTM and not hnRNPK did not change significantly, consistently with the general lack of nuclear enrichment of these tiles (Fig 5J). In contrast, KD of SNRNP70 affected similarly tiles bound by both hnRNPK and SNRNP70 and those bound only by hnRNPK (Fig 6D).

We next considered the number of motif occurrences within the NuLibB tiles, considering only the WT sequences to avoid the over-representation of Jpx#9 and Pvt1#22 variants. As expected, CCUCCC occurrences were associated with stronger binding by hnRNPK, and to a lesser extent by SLTM and SNRNP70 and with loss of nuclear enrichment upon their perturbation (Fig 6E). In contrast, occurrences of the U1 binding motif (a short version GUAR, or one of the three longer U1 motifs GGUAAG, GGUGAG, GUGAGU (Almada *et al*, 2013)) was associated with stronger binding by SNRNP70 and SLTM but not hnRNPK (Figs 6F and EV5A). Upon KD of SNRNP70 and SLTM, and to lesser extent hnRNPK, there was some *increase* of nuclear retention of the tiles that had these motifs (Fig EV5B), suggesting that while U1 binding to its target motif recruits SNRNP70 and SLTM to some transcripts, it is unlikely to play an important role in nuclear enrichment of SIRLOIN-bearing sequences in our specific setting. We conclude the proper expression of hnRNPK, SLTM, and SNRNP70 is required for nuclear enrichment of SIRLOIN-containing sequences, with SLTM binding specifically increasing nuclear enrichment of sequences that are bound by hnRNPK.

Discussion

We describe here a sequence-binding-function map of the SIRLOIN element (Fig 7). We show that SLTM and SNRNP70, two abundant nuclear RNA-binding proteins, bind the SIRLOIN element and are

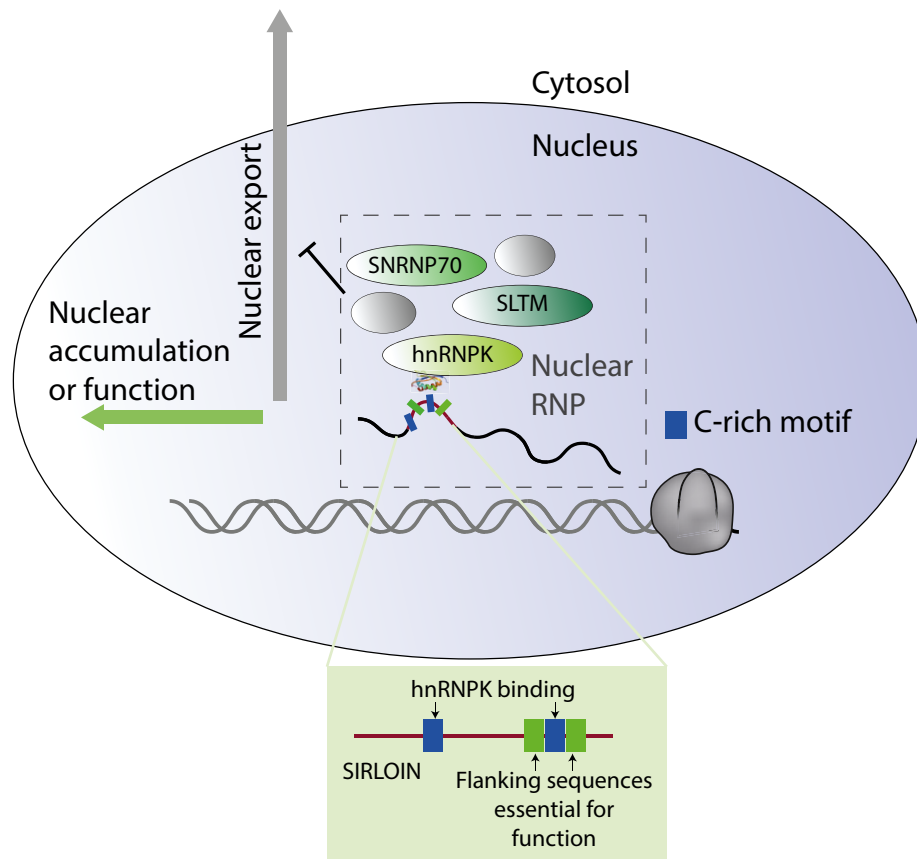


Figure 7. A model for the sequence-binding-function map of the SIRLOIN element.

SIRLOIN is required for the function of an RNP, that includes hnRNP, SNRNP70, SLTM, and potentially additional elements, which act together to enrich SIRLOIN-containing RNAs in the nucleus. Within the SIRLOIN sequence, the core part contains a single hnRNP binding site with specific flanking sequences required for function, whereas an additional region upstream of the core mediates additional hnRNP binding, but is not required for function.

required for proper nuclear enrichment of SIRLOIN-containing RNAs. One possibility that we considered is that these proteins recognize specific motifs flanking the canonical hnRNP binding site in SIRLOIN. However, the MPRNA-RIP data for SLTM and SNRNP70 rather suggest the specific bases that contribute to the binding of SLTM and SNRNP70 to SIRLOIN strongly overlap with those that facilitate hnRNP binding and that sequence changes that affect hnRNP binding typically have similar effects on SLTM and SNRNP70. One possibility is that hnRNP recruits SLTM and SNRNP70 to SIRLOIN. However, there are no known physical interactions between SLTM or SNRNP70 and hnRNP, and we were not able to detect an interaction between SLTM and hnRNP in MCF-7 cells by co-immunoprecipitation (Fig EV6A and B). Another possibility is that hnRNP binding is required for making SIRLOIN accessible to binding of the other factors, which by itself is less sequence-specific, or depends on sequences that were not mutated in NucLibC. However, knockdown of hnRNP did not substantially affect binding of SLTM to a GFP mRNA containing the Jpx#9 tile in its 3' UTR, and conversely, knockdown of SLTM did not affect hnRNP binding (Fig EV6C). SLTM and snRNP70 binding may not depend on or increase that of hnRNP but rather help enforce the hnRNP-mediated nuclear enrichment. Indeed, binding of just SLTM and/or SNRNP70 does not appear to contribute to nuclear

enrichment, whereas their binding in addition to hnRNP is associated with a more substantial nuclear presence compared to binding of just hnRNP. The effects are generally somewhat larger for SLTM than for SNRNP70, though we can not exclude the possibility that the differences are due to differences between the efficiency of the reagents we used to pulldown or deplete the two proteins.

We note that the magnitude of the effect on nuclear localization that we observe for individual SIRLOIN elements is generally modest, ~2-fold increase in nuclear presence, and that for our GFP mRNA reporter, that is typically mostly cytoplasmic, these effects do not dramatically change the overall distribution of the RNA in the cell (Fig EV6D). In the context of endogenous RNAs, the effects are likely much more substantial, as RNAs will in many cases harbor multiple SIRLOIN elements, and those are associated with a stronger effect on localization (Lubelsky & Ulitsky, 2018). Furthermore, combinations of SIRLOIN elements with other features that induce nuclear enrichment (Palazzo & Lee, 2018) will likely lead to cumulative effects that will be sufficient to cause a RNA to accumulate in the nucleus and to have a strong impact on its functionality. Indeed, when we consider endogenous RNAs with substantial binding of hnRNP and SLTM, such as *MXL1PL* discussed in Ref. Lubelsky and Ulitsky (2018) or *SRRM2* and *KMT2D* (Fig EV6E), these RNAs harbor large regions that bind hnRNP and/or SLTM, and in these

cases, the RNA is strongly enriched in the nucleus in an hnRNPK-dependent manner (Fig EV6E).

How the binding of hnRNPK, SLTM, and SNRNP70 to SIRLOIN leads to nuclear enrichment remains unknown. One possibility is that hnRNPK and/or SNRNP70, which are enriched in the nuclear speckles, help “anchor” the bound RNAs in these membraneless organelles. The binding of U1 snRNP to RNA was recently reported to be associated with nuclear enrichment of long RNA through an MPRNA (Yin *et al*, 2020), which echoed previous studies of individual genes and reporters (Chang & Sharp, 1989; Takemura *et al*, 2011; Lee *et al*, 2015; Azam *et al*, 2019). For example, nuclear retention of the *MEG3* lncRNA requires SNRNP70 but not hnRNPK (Azam *et al*, 2019). Furthermore, tethering of SNRNP70 to an RNA was shown to be sufficient for its nuclear retention (Takemura *et al*, 2011), similar to what we previously observed for hnRNPK (Lubelsky & Ulitsky, 2018). A possible model is that U1 snRNP recruitment is the functional element that yields nuclear retention and that for SIRLOIN, this recruitment is facilitated by hnRNPK binding, rather than by the presence of a specific U1 binding motif. Interestingly, the specific element that recruits U1 to the NXF1 intron and leads to nuclear retention (Yin *et al*, 2020) contains a conserved hnRNPK binding site ~100 nt upstream of the U1 binding site (Fig EV5C), suggesting that hnRNPK and U1 may also cooperate in retention of the intron-retaining NXF1 mRNA splice variant.

We describe here an integrated approach for dissection of sequence-binding-function axes in RNA elements. We focus on an element driving nuclear enrichment of a host RNA, as the functionality of such elements can be efficiently dissected by MPRNAs (Lubelsky & Ulitsky, 2018; Shukla *et al*, 2018; Yin *et al*, 2020). In addition to measuring element functionality via nuclear/cytoplasmic fractionations, we add here a dimension of binding measurements in cells, which allow us to dissect between sequence variants that are important for function, those that are important for binding specific factors, and those that affect both, which helps explain why specific positions within the sequence are functionally important. Using this approach, we show that hnRNPK binding is essential for SIRLOIN functionality but it is not sufficient as sequences that bind hnRNPK as well as the SIRLOIN element, but in a different position or sequence context, do not yield nuclear enrichment. The combination of MPRNA and RIP-MPRNA with RAP-MS allowed us to iteratively extend the regulatory circuit governing RNA element activity. In the future, additional iterations of this approach, for example, RAP-MS on SIRLOIN variants that are deficient in hnRNPK and/or SLTM binding and their comparison to each other, and identification of additional factors, are expected to further uncover the full composition of the SIRLOIN RNP, which will pave the way also for its studies using structural biology techniques, such as CryoEM. This methodology is readily applicable to other RNA activities beyond RNA localization, and we expect that it will yield substantial breakthroughs in understanding the functional anatomy of long RNAs.

Materials and Methods

Cell culture and transfection

MCF-7 cells (ATCC) were grown in DMEM (Gibco, 11-965-092) supplemented with 10% FBS and pen/strep. Cells were routinely

tested for mycoplasma contamination. Transfection of siRNA was done using DharmaFECT4 (Horizon Discovery, T-2004-03) according to the manufacturer's suggested protocol. Plasmid transfection was done using PEI (Durocher *et al*, 2002) (PEI linear, M_r 25,000, PolyScience Inc.). For generation of AcGFP-Jpx9 stable line, MCF-7 cells were transfected with linearized plasmid and selected using 500 ng/ μ l of G418 (Gibco 11811-031).

RIP

RNA immunoprecipitation (RIP) was performed as previously described (Gagliardi & Matarazzo, 2016) with slight modifications. Extracts in PLB buffer were diluted 10 \times in NET-2 buffer and incubated overnight with the primary antibody while rotating at 4°C. Magnetic beads were added to the extract-antibody mix and incubated for an additional 4 h rotating at 4°C.

Antibodies used: anti-hnRNPK (MLB, RN019P), anti-SLTM (Bethyl, A302-834A), anti-SNRNP70 (MLB, RN097PW), and Normal rabbit IgG (Millipore, 12-370).

RAP

RNA antisense purification (RAP) was modified from the methods described in (Engreitz *et al*, 2014).

Probe generation

Oligonucleotide pool tiled across the AcGFP1 sequence was ordered from Twist Bioscience (San Francisco, CA). Oligos were amplified in 96 reactions each at a volume of 50 μ l. The PCR product was concentrated using Amicon ultra tubes (0.5 ml 30 kDa millipore UFC503096) followed by purification with AMPure XP beads (Beckman-coulter, A63881) at 2:1 ratio. The product was eluted at 30 μ l of ddH₂O. For the generation of IVT templates, 2 ng of amplified oligos was used as PCR template, adding the T7 promoter sequence at the 5' end. RNA was generated by a 16-h IVT reaction using 250 ng template DNA (MEGAscript T7 Kit, Ambion AM1334M), and RNA was purified using the RNeasy kit (Qiagen) according to the manufacturer's recommended protocol. Single-strand DNA probes were generated by RT-PCR with a biotinylated primer. 1 μ g of RNA template was used for each reaction using qScript Flex cDNA synthesis kit (Quantabio, 95049-100) according to the manufacturer's recommended protocol, and 20 reactions were pulled for probe generation. Template RNA was degraded by adding NaOH to a final concentration of 100 mM and incubation at 75°C for 10 min. Acetic acid was added to a final concentration of 100 mM to stop the reaction.

Probes were purified using the RNeasy kit with the following modification to the manufacturer's protocol.

- 1 Samples were mixed with 3.5 volumes of buffer RLT.
- 2 1.5 volumes of EtOH were added to the DNA/RLT mix.

Sample preparation

Cells were washed with cold PBS and 10 ml of cold PBS were added to each 15 cm culture dish and crosslinked at 0.8 J/cm² UV (254 nm). Cells were scraped and precipitated by centrifugation (1,000 g for 5 min at 4°C). Cells were washed twice in cold PBS, and cell pellets were flash-freezed in liquid N₂.

Cell pellet was resuspended in 870 μ l cold lysis buffer (10 mM Tris–HCl pH 7.5, 500 mM LiCl, 0.5% DDM, 0.2% SDS, 0.1% DOC, supplemented with protease and RNase inhibitors (K1011, APExBio; E4210, EURx)) and incubated on ice for 10 min. During the incubation, the cells were passed through a 26G needle 5 times in order to break the pellet. The extract was sonicated (Bioraptor low setting 5 cycles 30 s ON, 30 s OFF).

DNA was degraded by adding 4.8 μ l of 200 \times DNase salt solution (500 mM MgCl₂, 100 mM CaCl₂) and 20 U of Turbo DNases (AM2238, Thermo) and incubation at 37°C for 10 min.

The samples were placed on ice and DNase reaction was stopped by adding 19.6 μ l of 500 mM EDTA (final concentration 10 mM), 9.8 μ l of 500 mM EGTA (final concentration 5 mM), and 4.9 μ l of 500 mM TCEP (final concentration 2.5 mM).

The extract was mixed with 2 volumes of 1.5 \times hybridization buffer (15 mM Tris–HCl pH 7.5, 7.5 mM EDTA, 750 mM LiCl, 0.75% DDM, 0.3% SDS, 0.15% DOC, 6 M urea, 3.75 mM TCEP) and incubated on ice for 10 min. The extract was centrifuged at 16,000 g for 10 min in a cold centrifuge. The supernatant was transferred to a new tube and flash-frozen in liquid N₂.

Pre-clearing

Streptavidin magnetic beads (NEB, S1420S) were washed 3 times in NEB wash buffer (NaCl 500 mM, 20 mM Tris–HCl pH 7.5, 1 mM EDTA) and twice in 1 \times hybridization buffer (10 mM Tris–HCl pH 7.5, 5 mM EDTA, 500 mM LiCl, 0.5% DDM, 0.2% SDS, 0.1% DOC, 4 M urea, 2.5 mM TCEP).

Cell lysate (lysate from 5×10^7 cells per pulldown) was prewarmed to 37°C and added to the beads and incubated for 30 min at 37°C in a thermomixer with 30 s ON 30 s OFF mixing at 1,100 rpm. The beads were magnetically separated, and the precleared lysate was transferred to a new tube.

Pulldown

The lysate was spiked with 1 fmol/reaction of control biotinylated RNA, and 100 μ l was taken as input RNA. 5 μ g of probe was incubated at 85°C for 3 min and placed on ice, and 1 ml of lysate was added to each probe. The reaction was incubated for 2 h at 67°C in a thermomixer with 30 s ON 30 s OFF mixing at 1,100 rpm. Magnetic beads were washed as described above and added to the lysate, and the reaction was incubated in the thermomixer for an additional 30 min. Beads were magnetically separated and washed 4 times in 1 \times hybridization buffer, each wash was incubated for 5 min at 67°C. The beads were washed 3 times in PBS in order to prepare them for on-bead digestion and mass spectrometry.

Mass spectrometry sample preparation

For the 1st experiment, samples were digested by trypsin, analyzed by LC-MS/MS on Q Exactive plus (Thermo), and identified by Discoverer software version 1.4 against the human sequence using the sequest and Mascot search engines. Semi-quantitation was done by calculating the peak area of each peptide.

For the 2nd experiment, beads were washed with 50 mM ammonium bicarbonate. Then, 50 μ l of 8 M urea was added to the beads and incubated for 30 min in room temperature. Proteins were reduced with 5 mM dithiothreitol (Sigma) for 1 h at room temperature and alkylated with 10 mM iodoacetamide

(Sigma) in the dark for 45 min at room temperature. Samples were diluted to 2 M urea with 50 mM ammonium bicarbonate. Proteins were then subjected to digestion with trypsin (Promega; Madison, WI, USA) overnight at 37°C, followed by a second trypsin digestion for 4 h. The digestions were stopped by addition of trifluoroacetic acid (1% final concentration). Following digestion, peptides were desalted using Oasis HLB, μ Elution format (Waters, Milford, MA, USA). The samples were vacuum dried and stored in –80°C until further analysis.

Liquid chromatography mass spectrometry

LC/MS was performed as previously described (Almagor *et al*, 2020), and ULC/MS grade solvents were used for all chromatographic steps. Each sample was loaded using split-less nano-ultra performance liquid chromatography (10 kpsi nanoACQUITY; Waters, Milford, MA, USA). The mobile phase was as follows: A) H₂O + 0.1% formic acid and B) acetonitrile + 0.1% formic acid. Desalting of the samples was performed online using a reversed-phase Symmetry C18 trapping column (180 μ m internal diameter, 20 mm length, 5 μ m particle size; Waters). The peptides were then separated using a T3 HSS nano-column (75 μ m internal diameter, 250 mm length, 1.8 μ m particle size; Waters) at 0.35 μ l/min. Peptides were eluted from the column into the mass spectrometer using the following gradient: 4 to 27%B in 50 min, 27 to 90%B in 5 min, maintained at 90% for 5 min and then back to initial conditions.

The nanoUPLC was coupled online through a nanoESI emitter (10 μ m tip; New Objective; Woburn, MA, USA) to Q Exactive HF mass spectrometer (Thermo Scientific). Data were acquired in data-dependent acquisition (DDA) mode, using a Top10 method. MS1 resolution was set to 120,000 (at 200 m/z), mass range of 375–1,650 m/z, AGC of 3e6 and maximum injection time was set to 60 ms. MS2 was performed by isolation with the quadrupole, width of 1.7 Th, 27 NCE, 15k resolution, AGC target of 2e3, maximum injection time of 60 ms and dynamic exclusion of 30 s.

Data processing

Raw data were analyzed using the MaxQuant software suite 1.6.6.0 (Cox & Mann, 2008) with the Andromeda search engine. The higher-energy collisional dissociation (HCD) MS/MS spectra were searched against an *in silico* tryptic digest of human proteins from the UniProt/Swiss-Prot sequence database (v. 2019_09), including common contaminant proteins. All MS/MS spectra were searched with the following MaxQuant parameters: acetyl (protein N-terminus), M oxidation; cysteine carbamidomethylation was set as fixed modification; max 2 missed cleavages; and precursors were initially matched to 4.5 ppm tolerance and 20 ppm for fragment spectra. Peptide spectrum matches and proteins were automatically filtered to a 1% false discovery rate based on Andromeda score, peptide length, and individual peptide mass errors.

Proteins were identified and quantified based on at least two unique peptides and based on the label-free quantification (LFQ) (Cox *et al*, 2014)) values reported by MaxQuant. Resulting protein groups were imported into Perseus (Tyanova *et al*, 2016). Data were filtered to include proteins identified with 2 peptides or more, and those that replicated in at least two of three replicates in at least one group. The data were transformed to log₂, and Student's *t*-test was used to identify statistically significant proteins.

NuLib data analysis

Reads were aligned to the NuLib sequences, and UMIs were counted as in Refs Lubelsky and Ulitsky (2018) and Zuckerman *et al* (2020) using a custom Java script, identifying the tile that matches the read with the minimal number of mismatches, without allowing indels. Reads in each library were normalized by the total number of alignable reads. Nuc/Cyto ratios were computed using a pseudocount of 0.5 as in (Lubelsky & Ulitsky, 2018; Zuckerman *et al*, 2020). IP/Input ratios were computed using UMI counts and DESeq2 (Love *et al*, 2014). In addition to the tiles described in the main text, NuLibC contained 690 additional tiles derived from other sequences, including NORAD lncRNA, and variants of an NICN1#53 tile from NuLibB, but these were amplified and cloned at lower frequencies (68%) and were not analyzed further in this study. Tile sequences were analyzed using the Bioconductor BioStrings package.

RNA-seq and RIP-seq data analysis

RNA-seq reads were mapped to the human genome (hg19 assembly) using STAR (Dobin *et al*, 2013) and visualized using the UCSC genome browser. Expression levels of RefSeq transcripts were quantified using RSEM (Li & Dewey, 2011), and differential expression was computed using DESeq2 (Love *et al*, 2014).

eCLIP analysis

eCLIP clusters as defined by ENCODE were obtained from <http://encodeproject.org>. Only clusters with significance < 0.01 and at least 2-fold enrichment over mock input control were considered and intersected with exons of RefSeq-annotated genes.

Single-molecule FISH and Immunofluorescence

For single-molecule FISH, AcGFP probe libraries were designed according to Stellaris guidelines and synthesized by Stellaris (Stellaris RNA FISH probes, Biosearch Technologies) as described in Ref. Raj *et al* (2008). Libraries consisted of 32 probes labeled with Quasar 570 (Table EV3). An oligo-dT(50-mer) probe labeled with FAM was used to label Poly(A)⁺ RNA for cell body segmentation. Hybridization conditions and imaging were as described previously (Lyubimova *et al*, 2013; Bahar Halpern & Itzkovitz, 2016). Hybridizations were done overnight at 30 °C with probes at a final concentration of 0.1 ng/μl. For immunofluorescence, anti-hnRNP and anti-SLTN antibodies were diluted in glucose oxidase (GLOX) buffer (1:1,000) and applied to cells for 2 h at room temperature. Secondary antibody Cy5-conjugated donkey anti-rabbit (1:500) was added to GLOX buffer for 1 h at room temperature. For nuclear staining, 1.25 μg/ml Hoechst 33342 (H3570, Thermo Fisher) was added during the washes. Images were taken with a Nikon Eclipse Ti2-E inverted fluorescence microscope equipped with a x100 oil-immersion objective and an iXon 888 EMCCD camera using NIS-Elements Advanced Research software. The image-plane pixel dimension was 0.13 μm and distance between Z stacks was 0.3 μm. Shown are 2D projections of images. Quantification was done with FishQuant V3 (Mueller *et al*, 2013). We performed automatic 2D projections as suggested in FishQuant documentation, followed by automatic cell segmentation using CellProfiler (McQuinn *et al*, 2018).

Hoechst signal was used to segment nuclei, and the oligo-dT signal was used to segment cell bodies. Following batch analysis, we manually examined segmentation and removed incorrectly segmented cells from further analysis using Fiji (ImageJ) software. Quantification of cytoplasmic and nuclear signals was performed with default parameters and recommended filters of FishQuant.

Data availability

RIP-seq, RIP-MPRNA, and MPRNA sequencing data: SRA database SRP297313 (<https://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP297313>).

Computer scripts: GitHub (<https://github.com/IgorUlitsky/MPRNA/>).

Expanded View for this article is available online.

Acknowledgements

We would like to thank Dr Amir Pri-Or from The De Botton Protein Profiling institute of the Nancy and Stephen Grand Israel National Center for Personalized Medicine for his help with mass spectrometry data analysis. Schraga Schwarz, Noam Stern-Ginossar, and members of the Ulitsky laboratory for comments on the manuscript and useful discussions. This study was funded by grants by grants to I.U. from the Israeli Science Foundation (ISF) (grant 852/19), the ISF-Natural Science from Foundation of China (NSFC) joint research program (grant 2406/18), the Germany-Israeli Foundation for Scientific Research and Development (grant I-144-417.5-2017), and the Israeli Ministry of Health as part of the ERA-NET localMND.

Author contributions

YL and IU designed the study. YL and BZ performed the experiments. IU analyzed the data. YL, BZ, and IU wrote the manuscript.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Almada AE, Wu X, Kriz AJ, Burge CB, Sharp PA (2013) Promoter directionality is controlled by U1 snRNP and polyadenylation signals. *Nature* 499: 360–363
- Almagor M, Levin Y, Halevy Amiran R, Fieldust S, Harir Y, Or Y, Shoham Z (2020) Spontaneous *in vitro* hatching of the human blastocyst: the proteomics of initially hatching cells. *Vitro Cell Dev Biol Anim* 56: 859–865
- Azam S, Hou S, Zhu B, Wang W, Hao T, Bu X, Khan M, Lei H (2019) Nuclear retention element recruits U1 snRNP components to restrain spliced lncRNAs in the nucleus. *RNA Biol* 16: 1001–1009
- Bahar Halpern K, Itzkovitz S (2016) Single molecule approaches for quantifying transcription and degradation rates in intact mammalian tissues. *Methods* 98: 134–142
- Bishof I, Dammer EB, Duong DM, Kundinger SR, Gearing M, Lah JJ, Levey AI, Seyfried NT (2018) RNA-binding proteins with basic-acidic dipeptide (BAD) domains self-assemble and aggregate in Alzheimer's disease. *J Biol Chem* 293: 11047–11066
- Carmody SR, Wente SR (2009) mRNA nuclear export at a glance. *J Cell Sci* 122: 1933–1937

- Chang DD, Sharp PA (1989) Regulation by HIV Rev depends upon recognition of splice sites. *Cell* 59: 789–795
- Cox J, Hein MY, Lubner CA, Paron I, Nagaraj N, Mann M (2014) Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol Cell Proteomics* 13: 2513–2526
- Cox J, Mann M (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* 26: 1367–1372
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29: 15–21
- Dominguez D, Freese P, Alexis MS, Su A, Hochman M, Palden T, Bazile C, Lambert NJ, Van Nostrand EL, Pratt GA et al (2018) Sequence, structure, and context preferences of human RNA binding proteins. *Mol Cell* 70: 854–867.e9
- Durocher Y, Perret S, Kamen A (2002) High-level and high-throughput recombinant protein production by transient transfection of suspension-growing human 293-EBNA1 cells. *Nucleic Acids Res* 30: e9
- Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, Grossman SR, Chow AY, Guttman M, Lander ES (2014) RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent Pre-mRNAs and chromatin sites. *Cell* 159: 188–199
- Gagliardi M, Matarazzo MR (2016) RIP: RNA immunoprecipitation. In *Polycomb Group Proteins: Methods and Protocols*, Lanzuolo C, Bodega B (eds), pp 73–86. New York, NY: Springer New York
- Hofacker IL (2003) Vienna RNA secondary structure server. *Nucleic Acids Res* 31: 3429–3431
- Huttlin EL, Bruckner RJ, Paulo JA, Cannon JR, Ting L, Baltier K, Colby G, Gebreab F, Gygi MP, Parzen H et al (2017) Architecture of the human interactome defines protein communities and disease networks. *Nature* 545: 505–509
- Lee ES, Akef A, Mahadevan K, Palazzo AF (2015) The consensus 5' splice site motif inhibits mRNA nuclear export. *PLoS One* 10: e0122743
- Lee ES, Wolf EJ, Ihn SSJ, Smith HW, Emili A, Palazzo AF (2020) TPR is required for the efficient nuclear export of mRNAs and lncRNAs from short and intron-poor genes. *Nucleic Acids Res* 48: 11645–11663
- Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12: 323
- Love M, Anders S, Huber W (2014) Differential analysis of count data—the DESeq2 package. *Genome Biol* 15: 550
- Lubelsky Y, Ulitsky I (2018) Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature* 555: 107–111
- Lyubimova A, Itzkovitz S, Junker JP, Fan ZP, Wu X, van Oudenaarden A (2013) Single-molecule mRNA detection and counting in mammalian tissue. *Nat Protoc* 8: 1743–1758
- McHugh CA, Guttman M (2018) RAP-MS: A method to identify proteins that interact directly with a specific RNA molecule in cells. In *RNA Detection: Methods and Protocols*, Gaspar I (ed), pp 473–488. New York, NY: Springer New York
- McQuin C, Goodman A, Chernyshev V, Kamensky L, Cimini BA, Karhohs KW, Doan M, Ding L, Rafelski SM, Thirstrup D et al (2018) Cell Profiler 3.0: next-generation image processing for biology. *PLoS Biol* 16: e2005970
- Moritz B, Lilie H, Naarmann-de Vries IS, Urlaub H, Wahle E, Ostareck-Lederer A, Ostareck DH (2014) Biophysical and biochemical analysis of hnRNP K: arginine methylation, reversible aggregation and combinatorial binding to nucleic acids. *Biol Chem* 395: 837–853
- Mueller F, Senecal A, Tantale K, Marie-Nelly H, Ly N, Collin O, Basyuk E, Bertrand E, Darzacq X, Zimmer C (2013) FISH-quant: automatic counting of transcripts in 3D FISH images. *Nat Methods* 10: 277–278
- Norman M, Rivers C, Lee Y-B, Idris J, Uney J (2016) The increasing diversity of functions attributed to the SAFB family of RNA-/DNA-binding proteins. *Biochem J* 473: 4271–4288
- Palazzo AF, Lee ES (2018) Sequence determinants for nuclear retention and cytoplasmic export of mRNAs and lncRNAs. *Front Genet* 9: 440
- Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S (2008) Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods* 5: 877–879
- Shukla CJ, McCorkindale AL, Gerhardinger C, Korthauer KD, Cabili MN, Shechner DM, Irizarry RA, Maass PG, Rinn JL (2018) High-throughput identification of RNA nuclear enrichment sequences. *EMBO J* 37: e98452
- So BR, Di C, Cai Z, Venters CC, Guo J, Oh J-M, Arai C, Dreyfuss G (2019) A complex of U1 snRNP with cleavage and polyadenylation factors controls telescripting, regulating mRNA transcription in human cells. *Mol Cell* 76: 590–599.e4
- Takemura R, Takeiwa T, Taniguchi I, McCloskey A, Ohno M (2011) Multiple factors in the early splicing complex are involved in the nuclear retention of pre-mRNAs in mammalian cells. *Genes Cells* 16: 1035–1049
- Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, Mann M, Cox J (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* 13: 731–740
- Van Nostrand EL, Freese P, Pratt GA, Wang X, Wei X, Xiao R, Blue SM, Chen JY, Cody NAL, Dominguez D et al (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature* 583: 711–719
- Wickramasinghe VO, Laskey RA (2015) Control of mammalian gene expression by selective mRNA export. *Nat Rev Mol Cell Biol* 16: 431–442
- Yin Y, Lu JY, Zhang X, Shao W, Xu Y, Li P, Hong Y, Cui Li, Shan Ge, Tian B et al (2020) U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature* 580: 147–150
- Zuckerman B, Ron M, Mikl M, Segal E, Ulitsky I (2020) Gene architecture and sequence composition underpin selective dependency of nuclear export of long RNAs on NXF1 and the TREX complex. *Mol Cell* 79: 251–267.e6