Nucleic Acids Research, 2023 **1**
https://doi.org/10.1093/nar/gkad466

Efficient and sensitive profiling of RNA–protein interactions using TLC-CLIP

Christina Ernst*, Julien Duc and Didier Trono*

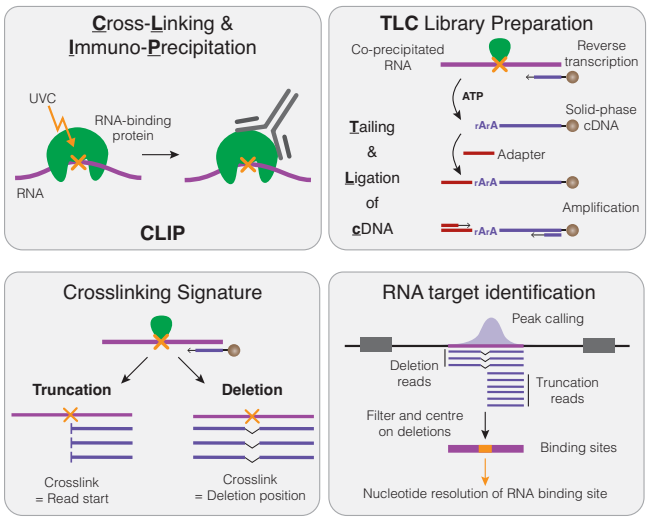
Global Health Institute, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Received August 27, 2022; Revised April 26, 2023; Editorial Decision May 10, 2023; Accepted May 15, 2023

ABSTRACT

RNA-binding proteins are instrumental for post-transcriptional gene regulation, controlling all aspects throughout the lifecycle of RNA molecules. However, transcriptome-wide methods to profile RNA–protein interactions *in vivo* remain technically challenging and require large amounts of starting material. Herein, we present an improved library preparation strategy for crosslinking and immunoprecipitation (CLIP) that is based on tailing and ligation of cDNA molecules (TLC). TLC involves the generation of solid-phase cDNA, followed by ribotailing to significantly enhance the efficiency of subsequent adapter ligation. These modifications result in a streamlined, fully bead-based library preparation strategy, which eliminates time-consuming purification procedures and drastically reduces sample loss. As a result, TLC-CLIP displays unparalleled sensitivity, enabling the profiling of RNA–protein interactions from as few as 1000 cells. To demonstrate the effectiveness of TLC-CLIP, we profiled four endogenous RNA-binding proteins, showcasing its reproducibility and improved precision resulting from a higher occurrence of crosslinking-induced deletions. These deletions serve as an intrinsic quality metric and increase both specificity and nucleotide-resolution.

GRAPHICAL ABSTRACT



INTRODUCTION

RNA-binding proteins (RBPs) are crucial in regulating gene expression by controlling all aspects of RNA metabolism. This includes co-transcriptional processes, such as splicing and editing, as well as post-transcriptional regulation, such as localisation, translation, and degradation of RNA. Being involved in such a variety of molecular processes, RBPs are essential for maintaining tissue homeostasis, and their deregulation is frequently associated with human pathologies, including neurodegenerative diseases and cancer (1,2).

The recognition and binding to RNA molecules often involves only short, relatively information-poor nucleotide sequences that occur with high frequency throughout the transcriptome, and are further influenced by secondary structures and nucleotide-independent interactions with the phosphodiester backbone (3,4). These characteristics make *in silico* predictions of RNA–protein interactions difficult, thus requiring methodologies that allow the transcriptome-wide identification of RNA targets and of their RBP-recruiting motifs.

*To whom correspondence should be addressed. Tel: +41 21 69309 40; Email: christina.ernst@epfl.ch
Correspondence may also be addressed to Didier Trono. Email: didier.trono@epfl.ch

The most commonly used protein-centric approach to study RNA–protein interactions relies on crosslinking and immunoprecipitation (CLIP) of the RBP of interest, followed by high-throughput sequencing of the co-precipitated RNA (5–9). Over the years, several variations of this technique have been developed, most prominently iCLIP (10,11) and derivations thereof such as infrared CLIP (irCLIP) (12), enhanced CLIP (eCLIP) (13–15) and more recent improvements including iCLIP2 (16) and improved iCLIP (iiCLIP) (17). These techniques enable the mapping of RNA binding sites at nucleotide resolution, and while individual steps differ between protocols, they follow the same overall strategy: cells are exposed to short wavelength ultraviolet radiation (UVC) as a crosslinking agent, lysed, and the extracts subjected to immunoprecipitation targeting the RBP of interest after partial RNA digestion. Co-purified RNA is then 3′ adapter-ligated prior to SDS polyacrylamide gel electrophoresis (SDS-PAGE) and transferred onto nitrocellulose from where RNA is liberated, purified and reverse transcribed into cDNA prior to second adapter ligation and amplification to generate sequencing-compatible libraries.

Apart from finding suitable antibodies that allow stringent purification of the RBP of interest, current CLIP protocols suffer from technical challenges during the generation of sequencing libraries from co-precipitated RNA. These challenges include low efficiency of RNA recovery after IP, time-consuming and inefficient purification procedures that are prone to sample loss, as well as suboptimal enzymatic reactions. As a consequence, current CLIP procedures require considerable amounts of starting material and are prone to experimental failure. In addition, the cost of CLIP experiments is often prohibitive due to large quantities of specialised reagents and high sequencing depth requirements due to low-complexity libraries that result from concatemerization and amplification of adapter dimers.

Here we present a streamlined library preparation strategy that allows time-efficient generation of CLIP libraries from low input material. Our approach relies on **t**ailing and **l**igation of **c**DNA molecules (TLC), which enhances enzymatic reactions and employs a fully bead-based, single-tube strategy minimising sample loss prior to amplification. This novel design makes purification of RNA–protein complexes via SDS-PAGE optional, thus providing the potential for a fully automated CLIP workflow for high-throughput settings.

MATERIALS AND METHODS

A step-by-step protocol is available on protocols.io: https://www.protocols.io/view/tlc-clip-cfmetk3e (dx.doi.org/10.17504/protocols.io.rm7vzywr4lx1/v1).

TLC-CLIP adapter generation

All adapters and oligos used throughout the protocol were ordered from Integrated DNA Technologies (IDT) and information regarding sequences, scale and purification can be found in Supplementary Table S1.

The TLC-L3 oligo for the first adapter ligation was synthesised at 250 nmol scale, carrying a 5′ phosphorylation and 3′ IRDye® 800CW (NHS Ester) (v3) modification and was purified using RNase-free HPLC with a total yield of 21.1 nmol.

Pre-adenylation was performed on 5 nmoles using the 5′ DNA Adenylation Kit (NEB, E2610L) as follows: 50 μl of 100 μM TLC-L3 adapter were set up with 25 μl 10 × 5′ DNA Adenylation Reaction Buffer, 25 μl 1 mM ATP and 50 μl Mth RNA Ligase (2.5nmol) in a total volume of 200 μl. Reaction was incubated at 65°C for 2 h followed by inactivation at 85°C for 10 min, during which it turns cloudy. Reaction was then cleaned up using the Nucleotide Removal Kit (Qiagen, Cat #28304) as follows: 200 μl were mixed with 4.8 ml of PNI buffer, distributed over 10 columns and spun down at 6000 rpm for 30 s. Columns were washed once in 750 μl PE buffer, spun for 1 min at 6000 rpm, followed by an empty spin at full speed before transferring columns to a new collection tube. 50 μl H_2O were added per column and incubated at RT for 2 min before centrifugation at 6000 rpm for 1 min. Eluates were combined with an approximate final concentration of 10 μM and 1 μM working stocks were prepared and frozen at –20°C.

Cell culture and generation of CLIP lysates

Adherent 293T cells (ATCC® CRL-1573™) were grown to ∼80% confluency in Dulbecco's modified Eagle's medium (Gibco, #41966–029) supplemented with 10% FCS (Sigma-Aldrich, #F9665-500ML, Lot #19A124) and 1% penicillin–streptomycin–L-glutamine (MED30-009-CI). Cells were rinsed in ice-cold PBS and crosslinked on ice with 254 nm UVC light at 0.3 J/cm^2 in a CL-3000 Ultraviolet Crosslinker (UVPA849-95-0615–02). Cells were collected into PBS by scraping, counted and desired cell number was aliquoted and spun down. Cell pellets were resuspended in iCLIP Lysis buffer (50 mM Tris–HCl pH 7.4, 100 mM NaCl, 1% Igepal CA-630, 0.1% SDS, 0.5% sodium deoxycholate) using 50 μl per 50 000 cells. Lysates were incubated on ice for 5 min followed by sonication for 5–10 s at 0.5 s ON and 0.5 s OFF at 10% amplitude using a tip sonicator (Branson LPe 40:0.50:4T). Protein concentration was measured using the Pierce™ Rapid Gold BCA Protein Assay Kit (Thermo Scientific, A53225) and lysates were either processed directly or stored at –80°C.

RNase treatment, immunoprecipitation and first adapter ligation

Protein-G beads (Thermo Fisher, 10004D) were washed twice in 1ml iCLIP Lysis buffer and resuspended in 100 μl iCLIP Lysis buffer per condition (100 μl of protein-G beads bind 20–30 μg of antibody and were scaled accordingly). Per IP, 1 μg of antibody against hnRNPc (Santa Cruz Biotechnology, sc-32308), RBM9 (Bethyl Laboratories, A300-864A, referred to as RBFOX2 throughout the manuscript), hnRNPA1 (4B10) (Santa Cruz Biotechnology, sc-32301), or hnRNPI (Santa Cruz Biotechnology, sc-56701) were added, and antibody-bead mixture was incubated at room temperature (RT) for 30–60 min on a rotating wheel.

Meanwhile, cell lysates at a concentration of ∼0.5 μg/μl were treated with different RNase concentrations using

0.25, 0.025 and 0.005 U of RNase I (Thermo Fisher, EN0602) for high, medium and low conditions. RNase dilution was added to cell lysates together with 2 μl Turbo DNase (Thermo Fisher, #AM2238) and lysates were incubated at 37°C for exactly 3 min at 1100 rpm, followed by 3 min on ice. Cell lysates were spun down for 10 min at 4°C at full speed and supernatant was transferred to a new tube.

Antibody-bead mixture was washed twice in iCLIP lysis buffer to remove unbound antibody and RNase-treated cell lysates were added alongside cOmplete EDTA-free Protease Inhibitor Cocktail (Merck, #11836170001) in PCR tubes and incubated for 2 h at 4°C on a rotating wheel. After IP, beads were washed twice in 200 μl High Salt Buffer (50 mM Tris–HCl pH 7.4, 1 M NaCl, 1 mM EDTA, 1% Igepal CA-630, 0.1% SDS, 0.5% sodium-deoxycholate), with the second wash at 4°C for 3 min on a rotating wheel, followed by two washes in 200 μl PNK Wash Buffer (20 mM Tris–HCl, pH 7.4, 10 mM $MgCl_2$, 0.2% Tween-20).

The volume of all washing steps has been adjusted to 200 μl to be compatible with the use of PCR tubes throughout the protocol. When using larger amounts of starting material, the volume of wash buffer should be scaled up to ensure stringency of wash steps.

Dephosphorylation of 3′ ends was performed in 20 μl of PNK reaction (70 mM Tris–HCl, pH 6.5, 10 mM $MgCl_2$, 1 mM DTT, 10 U SUPERaseIN RNase Inhibitor (ThermoFisher, #AM2696), 5 U T4 Polynucleotide Kinase (NEB, #M0201L) for 20 min at 37°C. Beads were washed once in PNK Wash Buffer and resuspended in 20 μl of ligation mix for overnight incubation at 16°C and 1200 rpm (50 mM Tris–HCl, pH 7.8, 10 mM $MgCl_2$, 1 mM DTT, 10 U SUPERaseIN RNase Inhibitor, 10 U T4 RNA Ligase (NEB, #M0204), 1 μl of 1 μM L3 adapter and 20% PEG400 (Sigma-Aldrich, #91893)).

TLC-CLIP library preparation with PAGE purification

Following the first adapter ligation, beads were washed twice in 200 μl High Salt Buffer, twice in 200 μl PNK Wash buffer and then resuspended in 20 μl 1× LDS sample buffer (Thermo Fisher, #NP0008) containing 5% beta-mercaptoethanol (Sigma-Aldrich, #M6250). Samples were denatured for 1 minute at 70°C and RNA–protein complexes were resolved on NuPAGE 4–12% Bis–Tris Gels (Thermo Fisher, #WG1402A) at 180 V for 1 h. Transfer was performed onto nitrocellulose (BioRad, #1620115) in 1× NuPAGE transfer buffer (Thermo Fisher, #NP00061) with 10% methanol at 30 V for 2 h at RT.

Nitrocellulose membranes were scanned on Odyssey® CLx Infrared Imager (LI-COR, 9141) with 169 μm resolution to visualise RNA localisation and then placed on filter paper soaked in PBS. Regions of interest were cut out from nitrocellulose membrane corresponding to ∼20–100 kDa above the molecular weight of the RBP of interest due to the ligation of TLC-L3 adapter (∼15.9 kDa) and associated RNA (with 70 nucleotides of RNA averaging ∼20 kDa). Nitrocellulose pieces were placed in LoBind Eppendorf tubes and 200 μl Proteinase K buffer (100 mM Tris–HCl, pH 7.4, 50 mM LiCl, 1 mM EDTA, 0.2% LiDS) containing 100 μg Proteinase K (Thermo Fisher, #AM2546) were added and incubated at 50°C for 45 min at 800 rpm.

Meanwhile, 10 μl of $Oligo(dT)_{25}$ Dynabeads™ (Thermo Fisher, #61005) per sample were washed in 1 ml of oligo(dT) Binding Buffer (20 mM Tris–HCl, pH 7.4, 1 M LiCl, 2 mM EDTA) and resuspended in 50 μl of oligo(dT) Binding Buffer per sample. Following Proteinase K treatment, supernatant was transferred to fresh tubes containing 50 μl of washed oligo(dT) beads and incubated for 10 min at RT on a rotating wheel. Following RNA capture, beads were washed twice in 125 μl oligo(dT) Wash Buffer (10 mM Tris–HCl, pH 7.4, 150 mM LiCl, 0.1 mM EDTA) and once in 20 μl 1× First-Strand Buffer (50 mM Tris–HCl, pH 8.3, 75 mM KCl, 3 mM $MgCl_2$). Beads were resuspended in 10 μl of Reverse Transcription Mix (1× First-Strand Buffer, 0.5 mM dNTPs, 1 mM DTT, 6 U SUPERase IN RNase Inhibitor, 20 U SuperScript™ IV Reverse Transcriptase (Thermo Fisher, #18090050)) and incubated for 15 min at 50°C followed by 10 min heating up to 96°C in an Eppendorf Thermomixer C. Samples were vortexed for 30 s at 96°C and then immediately placed on a magnet on ice. Supernatant containing adapter-ligated RNA was removed and efficiency of elution can be confirmed by dot-blotting on nitrocellulose membrane.

Solid-phase cDNA was washed once in 60 μl oligo(dT) Wash Buffer and once in 20 μl 1× T4 RNA Ligase Buffer (50 mM Tris–HCl, 10 mM $MgCl_2$, 1mM DTT, pH 7.5). Beads were resuspended in 5 μl of 5′ Adapter mix (2 μl 10× T4 RNA Ligase Buffer, 2 μl of 10 μM TLC_L## oligo (see Supplementary Table S1 – ensure balanced nucleotide composition between barcodes), 1 μl 100% DMSO), incubated at 75°C for 2 min then immediately placed on ice. 4 μl of Ligation Mix (12.5 mM ATP, 7 U Terminal Deoxynucleotidyl Transferase (TdT) (Takara, #2230B), 15 U T4 RNA Ligase High Concentration (NEB, #M0437)) were added as well as 10 μl 50% PEG8000 and reaction was mixed by slowly pipetting up- and down until beads are resuspended. Reaction was incubated at 37°C for 20 min, then cooled down to room temperature. 30 U of T4 RNA Ligase were added, the reaction mixed by pipetting and incubated at RT overnight in an Eppendorf Thermomixer C programmed to vortex the samples for 15 s at 2000 rpm every two min.

Following overnight incubation, 100 μl oligo(dT) Wash Buffer were added to ligation reaction and beads were resuspended by applying magnetic field to different sides of the tube until beads move swiftly through the solution and form a pellet when positioned on the magnet. Supernatant was then removed, and beads were washed in 100 μl oligo(dT) Wash Buffer and 20 μl 1× Phusion HF Buffer (Thermo Fisher, #F518L). Beads were resuspended in 25 μl cDNA amplification mix (1× Phusion HF PCR Master Mix (NEB, #M0531L) and 0.5 μM P5short_TLC-CLIP and P7short_TLC-CLIP primer mix (see Supplementary Table S1)) and amplification was performed with the following programme: 30 s at 98°C, 7 cycles of 10 s at 98°C, 30 s at 65°C and 30 s at 72°C followed by final extension at 72°C for 3 min. Meanwhile, 2 μl of oligo(dT) beads per sample were washed once in 1 ml oligo(dT) Binding buffer and resuspended in 5 μl per sample. After cDNA amplification, 5 μl of oligo(dT) beads were added and incubated at RT for 5 min on a rotating wheel to capture unwanted amplification by-products (see Supplementary Fig-

ure S4). Samples were placed on magnet and supernatant containing amplified cDNA was transferred to a fresh tube.

Size-selection of cDNA was performed using ProNEX® Size-Selective Purification System (Promega, #NG2002) with a ratio of 2.8X to enrich for cDNA inserts of at least 20 nucleotides in length (>80 bp). Library yield was then estimated by amplifying 1 μl of purified cDNA via qPCR using the full length P5 and P7 index primers and 2–3 cycles were subtracted from the obtained Ct value for final library amplification with P5 Universal adapter and P7 index primers (see Supplementary Table S1). Following PCR amplification, libraries were size-selected again using the ProNEX® Size-Selective Purification System, with a ratio of 1.8X to select fragments larger than 160 bp. Quality control was performed using the Agilent High Sensitivity DNA Kit (Agilent, #5067-4626) and libraries were quantified using the KAPA Library Quantification Kit (Roche, #KK4824). ProNEX size selection can be repeated with 1.8X ratio in case the Bioanalyser profiles show substantial peaks around 144 bp, which represent adapter dimers due to excess TLC-L3 adapter that was not efficiently removed prior to library preparation (see Supplementary Figure S5).

TLC-CLIP library preparation without PAGE purification

When omitting PAGE purification, the first adapter ligation was performed for 75 min at 25°C. Beads were washed as described above and either directly resuspended in Proteinase K reaction or in 20 μl of RecJ adapter removal reaction (1× NEB Buffer 2 (NEB, #B7002S), 25 U 5′ Deadenylase (NEB, #M0331S), 30 U RecJ endonuclease (NEB, #M0264S), 10 U SuperaseIN and 20% PEG400) and incubated at 37°C for 30 min prior to Proteinase K treatment. Adapter removal is achieved through incubation with the single-strand specific DNA exonuclease which results in degradation of unligated TLC-L3 adapter from the 5′ end following deadenylation. Samples were then placed on magnet, and supernatant was transferred to fresh tubes containing oligo(dT) beads, with the remaining library preparation performed as described above.

Control experiments to assess adapter self-ligation and consistent background bands

The potential for self-ligation of the TLC-L3 adapter was tested by generating control libraries obtained from the flow-through of the first adapter ligation. Following overnight incubation, the ligation mix was removed from the beads and mixed with 4X LDS and separated on NuPAGE 4–12% Bis–Tris Gels at 180V for 1 hour followed by transfer onto nitrocellulose as described above. Regions between 50–75, 75–100, 100–150 and 150–250 kDa were cut from nitrocellulose membrane and processed into sequencing libraries as described above (see Supplementary Figure S2). Sequencing reads resulting from adapter self-ligation were quantified using 'grep -c 'AGATCGGAAGAGCACACGTCTG[A]\{5,25\}'' to allow for variable length between 5 and 25 nucleotides of the polyA tail given the difficulties in obtaining accurate sequencing results for long homopolymers.

In addition, pre-adenylated TLC-L3 adapter was incubated with T4 RNA Ligase for 75 min at 37°C in the standard ligation reaction described above, but in the absence of RNA molecules functioning as acceptor molecules. The reaction was then separated on a polyacrylamide gel alongside pre-adenylated TLC-L3 adapter and infrared signal was scanned within the polyacrylamide gel at a distance of 0.5 mm as well as on the nitrocellulose membrane following transfer. For visualisation of protein following the mock ligation, polyacrylamide gels were rinsed with water and then stained with 0.01% Coomassie blue R250 (Sigma B-7920) in 50% methanol and 10% acetic acid for 10 min at RT. Gels were then rinsed with 40% methanol and 7% acetic acid, destained for 10 min in the same solution and rinsed twice in water for 5 min each. Gels were scanned at a distance of 0.5 mm on the Odyssey® CLx Infrared Imager.

Sequencing

TLC-CLIP libraries were sequenced on an Illumina NextSeq500 using the High Output Kit v.2.5 for 75 cycles, following Illumina protocol #15048776. 5% PhiX were added to final library pools for increased complexity and sequencing run was performed with custom configuration, running 86 cycles for Read 1 and 6 index cycles.

Mock ligations and denaturing polyacrylamide gel electrophoresis

Efficiency of second adapter ligation was tested in mock ligations using TLC_L01 as donor molecule and P7-3_TLC-CLIP as acceptor. 2 μl 10× T4 RNA Ligase Buffer, 1 μl 10 μM TLC-CLIP L01 oligo, 1 μl 10 μM P7-3 TLC-CLIP oligo and 1 μl DMSO were mixed and incubated at 75°C for 2 min. Reaction was placed on ice and 4 μl of Ligation mix containing 0.2 μl 0.1 M ATP, 0.5 μl TdT and 0.5 μl T4 RNA Ligase High Concentration were added followed by addition of PEG8000 to the indicated percentage. Ligation was incubated for 30 min at 37°C then cooled down to 16°C. Half the reaction was removed after 30 min at 16°C, the remaining reaction was incubated overnight.

1 μl of Ligation reaction was mixed with 1 μl Gel Loading Buffer II (Thermo Fisher, #AM8546G) and denatured at 72°C for 3 min. Samples were separated on 10% TBE-Urea gels (Thermo Fisher, #EC68752BOC) and stained with 1× SYBR® Gold (Thermo Fisher, #S11494) for 10 min in TBE buffer.

Tailing reaction in presence of NTPs or dNTPs

The processivity of two different terminal transferases were tested in the presence of adenosine triphosphate (ATP) or deoxyadenosine triphosphate (dATP), using the P5 Universal adapter as template oligo. 4 μl of 10× T4 RNA Ligase Buffer, 8 μl of 10 mM ATP or dATP, 2 μl of TdT (Takara, #2230B) or TT (NEB, #M0315L), 5 μl of 10 μM P5 Universal adapter were mixed in a total volume of 40 μl and incubated at 37°C. Aliquots were taken at 10, 20 and 30 min followed by inactivation at 75°C for 5 min. 1 μl of Tailing reaction was visualised on denaturing 10% TBE-Urea gels as described above.

Generation of total RNA-seq libraries from RBFOX2 knockout cells

RBFOX2 knockout cells were generated in a 293T background using CRISPR technology (18). In short, we designed CRISPR guides targeting upstream of the transcriptional start site of the canonical RBFOX2 isoform (guide1—chr22:35841752–35841774; guide2—chr22:35844193–35844215 (hg38)) and downstream of the transcriptional termination site (guide3—chr22:35738597–35738619; guide4—chr22:35737558–35737580 (hg38)) using CRISPOR (19). Guides against upstream and downstream region were cloned into PX459 (Addene #62988) and pgRGFP (Addgene #82695), respectively and transfected into 293T cells using Fugene (Promega #E5911). pSpCas9(BB)-2A-Puro (PX459) V2.0 was a gift from Feng Zhang (Addgene plasmid # 62988; http://n2t.net/addgene:62988; RRID:Addgene_62988) (20) and pgRGFP was a gift from Alan Mullen (Addgene plasmid # 82695; http://n2t.net/addgene:82695; RRID:Addgene_82695) (21).

Transfected cells were puromycin-selected for 3 days for the presence of Cas9 and guide targeting the 5′ region. Single cells were isolated via fluorescent-activated cell sorting into 96-well plates, gating on GFP positive cells to further select for the presence of the second guide. Clones were expanded for 3 weeks and then genotyped using PCR to identify successful knockout. Wildtype clones transfected with the same CRISPR guide combinations were used as control.

Total RNA was extracted from clonal cell populations using Trizol (ThermoFisher, #15596018) and stranded sequencing libraries were generated following ribosomal RNA depletion (NEB, #E6310L) using the NEBNext Ultra II Directional RNA Library Prep Kit for Illumina (NEB, #E7760L). Libraries were sequenced on an Illumina HiSeq4000 using the HiSeq 4000 SBS Kit for 150 cycles with a custom read configuration, running 93 cycles for Read 1 and 2 and 8 cycles for i7 and i5 indeces.

Demultiplexing and trimming with flexbar

Sequencing data was demultiplexed by i7 index reads using bcl2fastq without any read trimming. Further demultiplexing by in-read 5′ barcodes and trimming of adapter sequences was performed using Flexbar v.3.4.0 (https://github.com/seqan/flexbar) (22) in a two-step approach. In the first step, reads are demultiplexed by in-read barcodes allowing no mismatches, and UMIs are moved into the read header. Barcode sequences (see Supplementary Table S1) including the UMI designated by the wildcard character 'N' are provided in fasta format, with the arguments '-b barcodes.fasta --barcode-trim-end LTAIL --barcode-error-rate 0 --umi-tags'. In the second step, any adapter contamination at the 3′ end of the reads is removed allowing an error rate of 0.1 with the following arguments '--adapter-seq 'AGATCGGAAGAGCACACGTCTGAACTCCAGTCACNNNNNNATCTCGTATGCCGTCTTCTGCTTG' --adapter-trim-end RIGHT --adapter-error-rate 0.1 --adapter-min-overlap 1'. In addition, potential T-stretches at the 5′ end that are the result of ribotailing during ligation are removed by trimming 'T' homopolymers of 1–2 nucleotide length (see Supplementary Figure S6) using '--htrim-left T --htrim-max-length 2 --htrim-min-length 1' and reads shorter than 18 nucleotides post trimming are discarded by '--min-read-length 18'.

STAR alignment

Flexbar-trimmed reads were aligned against hg19 using STAR v.2.7.3a (https://github.com/alexdobin/STAR) (23) with the following parameters, to keep only uniquely mapping reads, removing the penalty for opening deletions and insertions and fully extending the 5′ end of reads to preserve the end of cDNA molecules: '--outFilterMultimapNmax 1 --scoreDelOpen 0 --scoreInsOpen 0 --alignEndsType Extend5pOfRead1'. To retain UMI in read header during STAR alignment, any space in header needs to be removed prior to mapping.

Bowtie2 alignment against RNA repeats and quantification

RNA repeat genome index was generated based on RepeatMasker annotation obtained from UCSC Table Browser, with the following specifications: 'clade: Mammal, genome: Human, assembly: Feb 2009 (GRCh37/hg19), group All Tables, database: hg19, table: rmsk, region: genome, filter: repFamily – does match = 'RNA' OR repClass = 'tRNA' OR repClass = 'rRNA' OR repClass = 'snRNA' OR repClass = 'scRNA' OR repClass = 'srpRNA". All regions in hg19 genome fasta file not overlapping with repetitive RNA were masked using bedtools maskfasta and genome index was built with bowtie2 v.2.3.5 using bowtie2-build (24). Reads across different repeat classes were quantified using FeautreCounts from the Subread package (25) with the following parameters: '-s 1 -t exon -g gene_id -M --fraction' and read counts were summarised at the level of ribosomal RNA (rRNA), small nuclear RNA (snRNA), transfer RNA (tRNA), small nucleolar RNA (snoRNA), 7SK small nuclear RNA, and others including 7SL signal recognition particle RNA, Y RNAs and the BC200 long-noncoding RNA.

Deduplication of reads

Aligned reads were deduplicated based on unique molecular identifiers using UMI-tools v.1.0.1 (https://github.com/CGATOxford/UMI-tools) (26). The dedup command was used with the parameters '--extract-umi-method read_id --method unique --spliced-is-unique' to group reads with the same mapping position and identical UMI, while treating reads starting at the same position as unique if one is spliced and the other is not.

Multiqc and usable reads

General quality metrics of libraries were assessed using FastQC v0.11.7 (https://github.com/s-andrews/FastQC) and QC data were collated using multiqc v.1.9 (https://github.com/ewels/MultiQC) (27) to extract information from combined log files to plot usable read fractions.

Peak calling

Enriched regions were identified using the peak calling algorithm CLIPper v.2.0.0 (https://github.com/YeoLab/clipper) (13,14) with default settings and a p-value cutoff of 0.01 '--poisson-cutoff 0.01'. Peak calling was performed on usable reads, which are defined as uniquely mapped and deduplicated, without any prior filtering for reads containing crosslinking induced deletions.

Filtering of peaks

CLIPper peaks were filtered by removing ENCODE blacklisted regions from eCLIP libraries (14) as well as peaks obtained from TLC-CLIP libraries skipping the ligation step as well as IgG controls for either Rabbit or Mouse IgG depending on RBP. An additional score filter was applied by requiring –10 log(pval) to be larger than 50 for any downstream analysis. Consensus peaks between replicates were obtained using bedtools intersect (28) requiring a minimum overlap of 25% between peaks.

Correlation plots

For correlation plots peaks or deletion positions of individual replicates were concatenated and coverage was calculated using bedtools multicov (28). Count data was normalised using the cpm function from edgeR (29) against total library size and log2 transformed. Point density plots were generated using the geom_pointdensity package available on Bioconductor and correlation coefficient was calculated using Pearson correlation.

Pairwise comparison at peak level

Fraction of overlap between filtered peaks for individual replicates of either TLC-CLIP, eCLIP or easyCLIP was calculated using the Intervene pairwise intersection module (https://intervene.readthedocs.io/en/latest/index.html) (30) requiring a minimum of 25% overlap between peaks. For comparison between TLC-CLIP and eCLIP in HepG2 cells shown in Figure 3A, peaks were restricted to genes with stable gene expression between the two cell lines, as defined by differential gene expression analysis on total RNA-seq data for 293T and HepG2 cells. Stable genes were defined as having an absolute fold change lower than 1.1 and an expression higher than 5 $\log_2$ CPM.

De novo motif discovery

De novo motif discovery was performed using Homer (31) v4.10 on peaks centred on either the apex region obtained from CLIPper. findMotifsGenome.pl was used with the parameters '-oligo -basic -rna -len5 -S10 -size given' where peak size is a 50-nucleotide window around the apex.

Extraction of crosslinking sites with htseq-clip

Individual nucleotide positions of crosslink-induced deletions within TLC-CLIP reads were extracted using the htseq-clip tools (https://github.com/EMBL-Hentze-group/htseq-clip) (32) with the following parameters: 'htseq-clip extract -e 1 -s d'.

Read start positions were extracted using 'htseq-clip extract -e 1 -s s' or 'htseq-clip extract -e 2 -s s' for public iCLIP and eCLIP data, respectively.

Density plot for deletions and motif enrichment

Motif densities were calculated using the annotatePeaks.pl function from homer on consensus motifs generated with the seq2profile.pl function allowing 0 mismatches. For density plots in Figure 3C, motif density was calculated for apex-centred peaks from TLC-CLIP and eCLIP libraries using '–size 500 -hist 5 -norevopp'. For hnRNPC '-rm 10' was specified to remove occurrences of the same motif within 10 nucleotides to avoid artificial amplification of motif enrichment through longer U-stretches, for hnRNPI the density of pyrimidine stretches was plotted, based on a 'YYYYY' stretch.

For plotting the deletion density in Figure 4A, TLC-CLIP peaks were centred onto the consensus motif, with motif files being generated using seq2profile.pl. Tag directories for deletions were generated using the homer makeTagDirectory function on the bed file obtained from htseq-count. peakSizeEstimate needs to be changed to 1 in tagInfo.txt file to avoid extension of deletion tags and preserve nucleotide resolution. Deletion enrichment was obtained using the annotatePeaks.pl with '-hist 1 -size 100' across motif-centred peaks as well as peaks shuffled across the set of target genes bound by a given RBP. For RBPs recognising palindromic sequences such as 'AGGGA' or 'CUUUC' for hnRNPA1 or hnRNPI respectively, the exact position of the crosslinking site cannot be determined during alignment if the deletion falls within the homopolymer stretch. By default, STAR will position the deletion at the first base of the ambiguous sequence based on the DNA sequence, without awareness of the strand orientation of the gene, resulting in an artificial shift of the deletion position between genes on the forward or reverse strand. To remove this artifact, deletion positions for genes on the reverse strand were shifted by two nucleotides for hnRNPA1 and hnRNPI prior to visualisation.

Deletion-centred analysis

Peaks were centred on the maximum deletion position and coverage of this nucleotide position was calculated using bedtools multicov to calculate the CID ratio, indicating the proportion of reads at a given position that carry a deletion. Motif density across peaks with different CID ratios was calculated using annotatePeaks.pl with '-size 100 -hist 5 -norevopp'.

For visualising the percentage of peaks carrying motifs according to CID ratio, findMotifsGenome.pl was used with '-find motif -size 50 -norevopp'. Peak annotation across different transcriptomic and genomic features was performed using annotatePeaks.pl.

Read counting across transcript regions

A custom simplified annotation file (SAF) was generated from the ensemble gtf file ('Homo_sapiens.GRCh37.87.gtf') to contain the following additional annotation features: promoter regions spanning 500 nucleotides (nt) upstream of the transcriptional start site, proximal introns spanning

500 nt upstream and downstream of exons, and distal introns spanning all regions further than 500 nt away from exons. Aligned reads were quantified across transcript features using FeatureCounts with the following parameters: '-F SAF -s 1 -O --fraction --read2pos 5'. Enrichment for coding sequences was quantified using the same SAF in FeatureCount with the following paramters: '-F SAF -s 1 -O --fraction --read2pos 5 -t CDS'. Hierarchical clustering of the top25 most variable coding sequences between crosslinked and non-crosslinked samples was performed using pheatmap on $\log_2$-transformed normalised counts.

Deletion visualisation

Splice site annotation from homer for hg19 was used and intersected with deletion-centred peaks for RBFOX2 from 50 000 cells, subset by CID ratios above or below 10. For hnRNPc, peaks from TLC-CLIP and publicly available irCLIP (12) were intersected with intronic antisense Alu sequences that were extracted from RepeatMasker (33). Target Alu sequences were further divided into regions shared between TLC-CLIP and irCLIP and regions specific to the individual protocols.

For visualisation, deletion positions from htseq-clip were merged across the two biological replicates per condition and converted to bam files using bedtools bedtobam. Bigwig files were then generated using deeptools function bamCoverage with a binsize of 1, normalising for total deletion count (CPM). For irCLIP, start positions were obtained using htseq-clip 'htseq-clip extract -s s -e 1' and converted to bigwig files with a binsize of 1. Heatmaps and coverage profiles were generated using the createMatrix and plotHeatmap function from deeptools (34).

Alternative splicing detection with rMATS

Total RNA-seq from RBFOX2 KO and wildtype cells were aligned against hg19 using STAR v.2.7.3a. RBFOX2 regulated exons were identified using rMATS, specifying '--libType fr-firststrand --readLength 95 --variable-readlength --gtf Homo_sapiens.GRCh37.87.gtf' (35,36). Upregulated and downregulated exons were extracted from the skipped exon table 'SE.MATS.JC' using only junction counts for splicing analysis. Exons were considered up- or downregulated if their inclusion level difference was larger than 0.1 with a p-value cut-off of 0.05 and an FDR of 0.1. Background exons were extracted from the same list with a p-value and FDR larger than 0.1 and an inclusion level difference below 0.05. Duplicate entries between the different sets were removed and exons were crossed with peaks detected in RBFOX2 TLC-CLIP libraries from 50 000 cells, searching for overlap 300 base pairs up and downstream of the splice junctions. To test for differences in the number of CIDs found at regulated and non-regulated exons, we performed two-sample Wilcoxon Mann-Whitney tests between the CID ratios of peaks found at up- or downregulated genes and the CID ratio of peaks at background exons. Coverage plots for up- and downregulated exons were generated with the deeptools plotProfile function on CPM normalised read counts with a bin size of 10 computed by the deeptools bamCoverage function.

Data visualisation

Downstream data analysis and visualisation was performed in R (v 4.1.0) using the tidyverse package (37).

UCSC browser tracks are available under the following link: http://genome-euro.ucsc.edu/s/christinaernst/TLC%2DCLIP_NAR and contain bigwig files of merged deletion positions normalised for total deletion count (CPM) with a binsize of 1 as used for visualisation in heatmaps, as well as bigwig files of CPM-normalised uniquely mapped reads with a binsize of 10.

RESULTS

TLC-L3 adapter enables easy visualisation and generation of solid-phase cDNA

The majority of current CLIP-related protocols purify RNA–protein complexes via SDS-PAGE followed by transfer onto nitrocellulose, at which point crosslinked RNA can be visualised to control IP efficiency and RNase treatment conditions using radioactive isotope labelling (10,11,16). TLC-CLIP instead employs an infrared-dye conjugated adapter that was first introduced in the irCLIP protocol (12) and has also been adopted in iiCLIP (17). This strategy avoids radioactive labelling and allows easy visualisation of adapter-ligated RNA over a wide dynamic range following transfer onto nitrocellulose, as well as during subsequent library preparation steps (Figure 1A - C and Supplementary Figure S1).

The infrared dye is positioned at the 3′ end of the TLC-L3 adapter coupled to the hydroxyl group, thus effectively functioning as a chain terminator to prevent adapter self-ligation and concatemerization (Supplementary Figure S2). Additional background bands, such as those observed around 60 kDa for hnRNPI and RBFOX2 as well as in non-crosslinked lysates, are likely the result of excess unligated TLC-L3 adapter (Supplementary Figure S1). Such molecules would not be visible using conventional radioactive labelling due to the absence of a 5′ hydroxyl group on pre-adenylated adapter molecules (38,39). The successful transfer of un-ligated adapter molecules onto nitrocellulose further suggests their association with a protein, as the DNA oligonucleotide itself does not transfer efficiently (Supplementary Figure S2). The prominent band observed around 60 kDa thus most likely represents an association with T4 RNA Ligase 1, as these bands are present in control reactions following incubation of the TLC-L3 adapter with the enzyme alone (Supplementary Figure S2).

The TLC-L3 adapter further includes the partial sequence of the Illumina Index 1 Sequencing Primer followed by a 25-nucleotide (nt) long poly(A) stretch, which confers several advantages during the library preparation. First, adapter-ligated RNA molecules can be captured using oligo(dT) beads, which are inert to high concentrations of proteinase K and denaturing agents, as first demonstrated in the easyCLIP protocol (40). This eliminates the need for time-consuming RNA precipitations that are prone to sample loss especially at low concentrations (41), and instead enables highly efficient capture of adapter-ligated RNA that occurs within minutes,

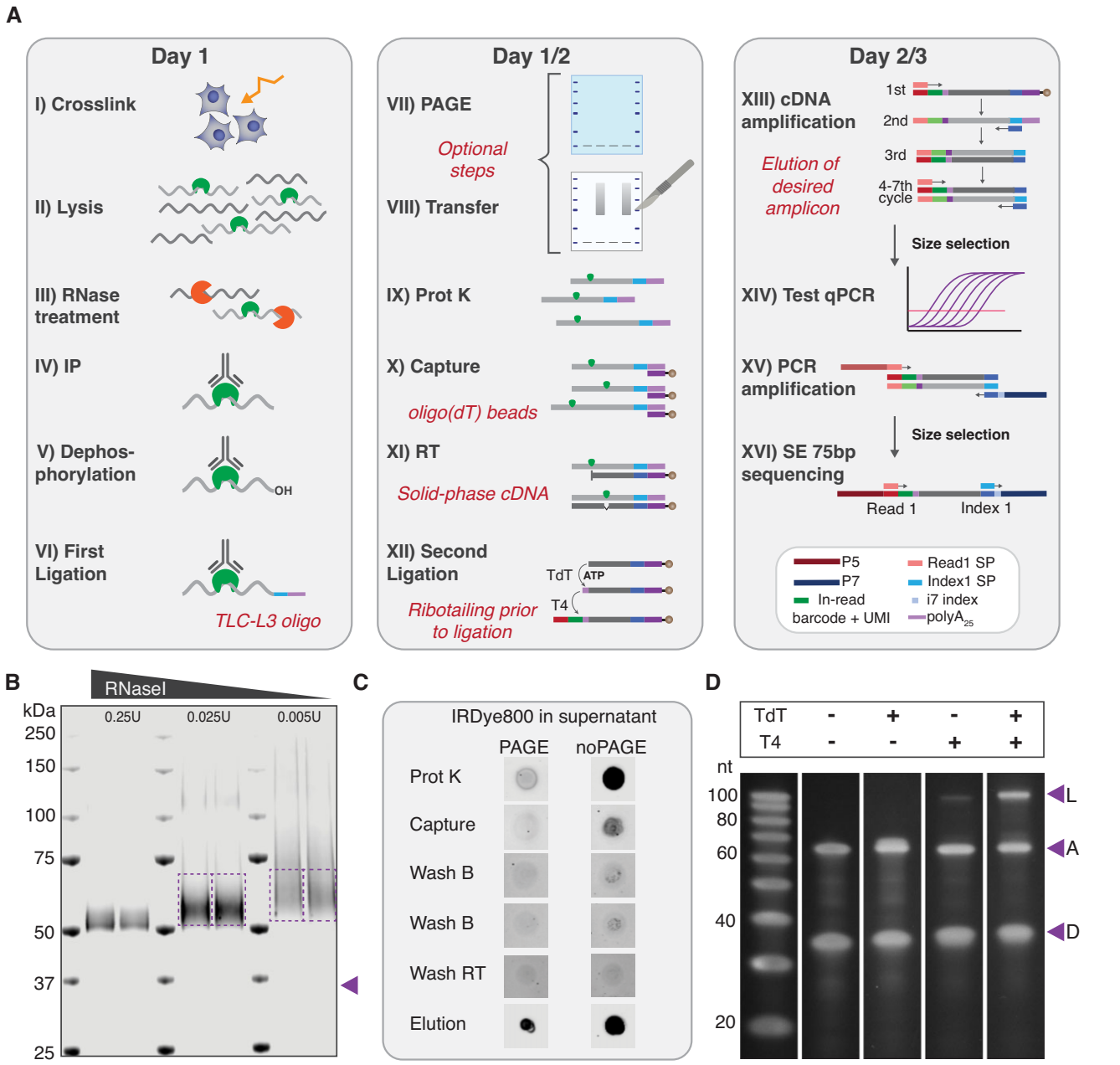


Figure 1. TLC library preparation enables efficient and time-effective generation of CLIP libraries. (**A**) Schematic overview of TLC-CLIP procedure with major changes highlighted in red italics. (**B**) Visualisation of adapter-ligated RNA on LI-COR Odyssey Clx Imager after membrane transfer for hnRNPA1 samples treated with different RNase concentrations. Molecular weight of hnRNPA1 is indicated as purple triangle and region used for library preparation is marked by a dashed rectangle. (**C**) Capture and elution of adapter-ligated RNA throughout steps IX - XI of library preparation visualised via dot blotting of supernatants on nitrocellulose. (**D**) Addition of Terminal Deoxynucleotidyl Transferase (TdT) in ligation reaction results in strongly increased ligation efficiency. D indicates donor molecule, A acceptor molecule and L ligation product.

and allows stringent washes to remove traces of proteinase K prior to subsequent enzymatic reactions without noticeable sample loss (Figure 1C and Supplementary Figure S2).

Furthermore, oligo(dT) beads are used to prime the reverse transcription (RT) reaction, resulting in first-strand cDNA molecules covalently linked to magnetic beads (42). Solid-phase cDNA can be efficiently separated from adapter-ligated RNA via heat denaturation (Figure 1C), and directly serve as acceptor molecule in the second adapter ligation. The use of oligo(dT) beads as reverse transcription primers has the additional benefit of preventing concatemerization, without the need for additional purification procedures, thus minimising sample loss and enabling a fully bead-based library preparation amenable to low input samples.

Ribotailing improves the efficiency of the second adapter ligation

The second adapter ligation presents a major bottleneck during library preparation; a problem which is not unique to CLIP, but also presents a challenge in other protocols where random priming of second strand synthesis is not favourable or feasible (43–46). Current iCLIP-related protocols tackle this step either through circularisation of cDNA molecules (10–12,17), or direct single-stranded (ss)DNA ligation (13–16), but the latter strategy is known to be enzymatically inefficient (47,48), resulting in the permanent loss of molecules that fail to ligate resulting in low-complexity libraries and large input requirements.

Our approach of **t**ailing and **l**igation of **c**DNA (TLC) greatly improves the efficiency of the ssDNA ligation by incorporating Terminal Deoxynucleotidyl Transferase (TdT) in the ATP-containing ligation mix. TdTs are highly processive in the presence of deoxynucleotide triphosphates (dNTPs), but self-terminate after incorporating only a few nucleotide triphosphates (NTPs), resulting in the addition of a short ribo-tail to the 3′ end of cDNA molecules (Supplementary Figure S3) (49). This effectively mimics the 3′ end of an RNA molecule, which is the preferred acceptor molecule of T4 RNA ligase (47,48), thus greatly increasing its affinity and ligation efficiency (Figure 1D and Supplementary Figure S3). Ribotailing of cDNA presents an efficient and cost-effective alternative to intramolecular ligation used in iCLIP, irCLIP and iiCLIP that require specialised RT primers which are not compatible with the generation of solid-phase cDNA, and improves the efficiency of the ssDNA intermolecular approach used in eCLIP and iCLIP2.

Ligated cDNA molecules remain covalently bound to the magnetic beads and can thus be easily purified through magnetic capture, eliminating the need for time-consuming procedures such as ethanol precipitation or PAGE purification that would result in further sample loss. The desired amplicons are then eluted off the beads via PCR amplification using short primers complementary to the TLC-L3 and second adapter, which removes the artificially introduced poly(A) tail (Supplementary Figure S4). Pre-amplified cDNA is then size-selected using ProNEX Size Selective Chemistry to enrich for molecules with insert sizes larger than 20 nucleotides as described in iCLIP2 (16) and sequencing-ready libraries are generated through a second PCR amplification. This step adds full-length P5 and P7 sequences as well as an i7 index to expand multiplexing capacity, followed by size-selection of fragments larger than 160bp to remove overly short inserts (Supplementary Figure S5).

The resulting TLC-CLIP libraries are compatible with single-end, two-colour chemistry sequencing protocols, with the first 15 nucleotides of the read corresponding to a 9-nt unique molecular identifier (UMI) for deduplication that is split around a 6-nt barcode for further multiplexing capacity (Materials and Methods).

TLC-CLIP libraries retain a larger fraction of usable reads

The improved library preparation strategy greatly increases sensitivity and lowers input requirements, which we demonstrated by generating high-quality libraries for four different RBPs from only 50 000 cells. We then tailored a streamlined processing pipeline (50) to the specificities of the TLC-CLIP workflow, such as the additional nucleotides added during the TLC reaction (Figure 2A). Ribotailing of the first-strand cDNA with ATP results in an over-representation of T nucleotides at the first few base positions of TLC-CLIP reads, which are removed by trimming T homopolymers of 1–2 nt length using Flexbar (22) (Supplementary Figure S6). After considering only uniquely mapping reads and removing PCR duplicates, we retain a larger fraction of reads compared to public CLIP datasets, which are termed 'usable reads', suggesting improved complexity of TLC–CLIP libraries (Figure 2B, Supplementary Figure S7 and Supplementary Table S2). Furthermore, TLC-CLIP libraries display a longer read length alongside a higher fraction of reads carrying deletions particularly in reads that show the maximum length, indicating more frequent read-through instead of truncation at the crosslinking site (Figure 2C). This is in accord with the use of Superscript IV during reverse transcription, which frequently causes crosslinking-induced mutations (CIMS) (51), resulting in deletions at similar or higher rates compared to HITS-CLIP (high-throughput sequencing of RNA isolated by CLIP) protocols (52) (Figure 2D). However, unlike HITS-CLIP protocols, which ligate adapters to both ends of the RNA molecule prior to RT and thus only amplify read-through events, truncated reads are also retained using TLC-CLIP resulting in much greater yield of usable reads, drastically reducing sequencing requirements compared to HITS-CLIP (Figure 2B).

TLC-CLIP libraries recapitulate public CLIP datasets and show increased specificity

Given these characteristics of TLC-CLIP libraries, we opted for the peak calling algorithm CLIPper (13,14,53), which does not exclusively rely on read start positions to determine cross-linking events, and can therefore be applied to different protocols, allowing a direct comparison between TLC-CLIP and public CLIP datasets. We found our TLC-CLIP libraries to display a high level of correlation at the peak-level (R^2 = 0.5–0.78, Pearson correlation), with 48–67% overlap between biological replicates (Supplementary Figure S8). Comparison between TLC-CLIP and easyCLIP libraries (40), which both profiled RBFOX2 in 293T cells, revealed up to 72% overlap at the peak-level, which is similar to the level of variation observed between biological replicates, given the stochastic nature of RNA binding and rapid turnover of intronic target sequences (Supplementary Figure S8).

Comparison with eCLIP libraries (13) showed up to 50% overlap at the peak-level, most likely due to cell-type specific differences in RNA abundance, as similar levels of overlap were observed between eCLIP libraries in HepG2 and K562 cells (Supplementary Figure S8). Accordingly, restricting the comparison between TLC-CLIP and eCLIP to genes with similar expression levels between 293T and HepG2 cells (n = 1507) increased the overlap to up to 68%, demonstrating that TLC-CLIP accurately captures RBP binding profiles (Figure 3A).

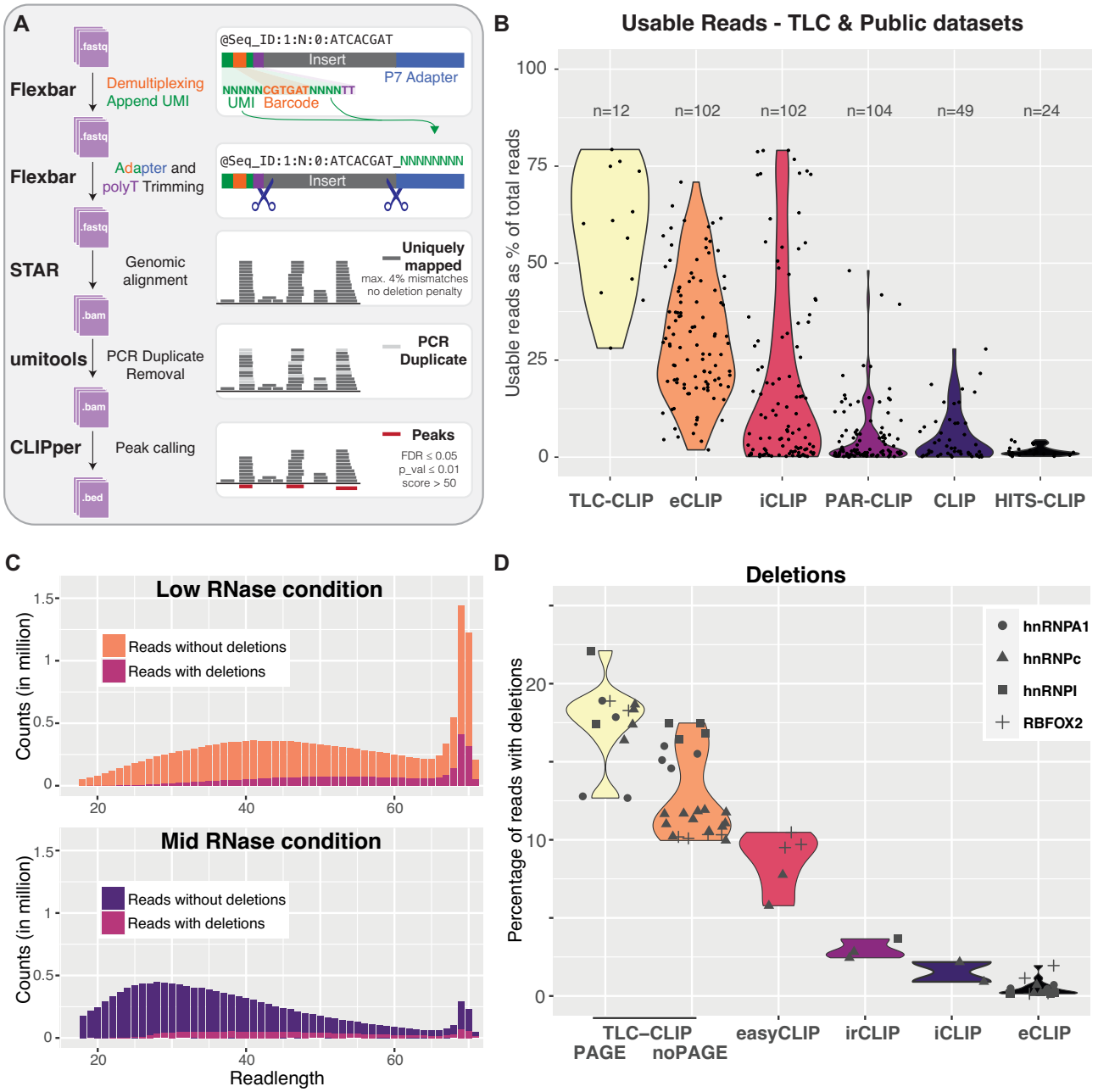


Figure 2. TLC-CLIP libraries contain more usable reads, a larger proportion of which carry crosslinking induced deletions (**A**) Schematic representation of the data processing pipeline for TLC-CLIP libraries. (**B**) Percentage of usable reads out of total read fraction is displayed for different publicly available CLIP protocols and PAGE-purified TLC-CLIP libraries. Information on usable read fraction for public datasets was obtained from van Nostrand et al. (13) and further annotated according to different experimental protocols. (**C**) Read length distributions for hnRNPA1 for different RNase conditions in PAGE purified libraries, highlighting reads with and without deletions. (**D**) Percentage of reads carrying deletions in TLC-CLIP and public CLIP libraries.

Furthermore, *de novo* motif discovery on TLC-CLIP peaks recapitulated previously reported consensus motifs for all four RBPs with high precision. Compared to eCLIP, a larger number and proportion of TLC-CLIP peaks contained a motif that closely resembles the respective consensus motif, indicating increased specificity for our protocol (Figure 3B and Supplementary Figure S8). This is further supported by stronger motif enrichment around the peak apex, which represents the region of highest read coverage defined by CLIPper, and was particularly noticeable for splicing factors with well-defined motifs such as RBFOX2 and hnRNPA1 (Figure 3C). Together, these results demonstrate that TLC-CLIP has increased specificity and resolution compared to eCLIP, despite starting with 400-times less input material.

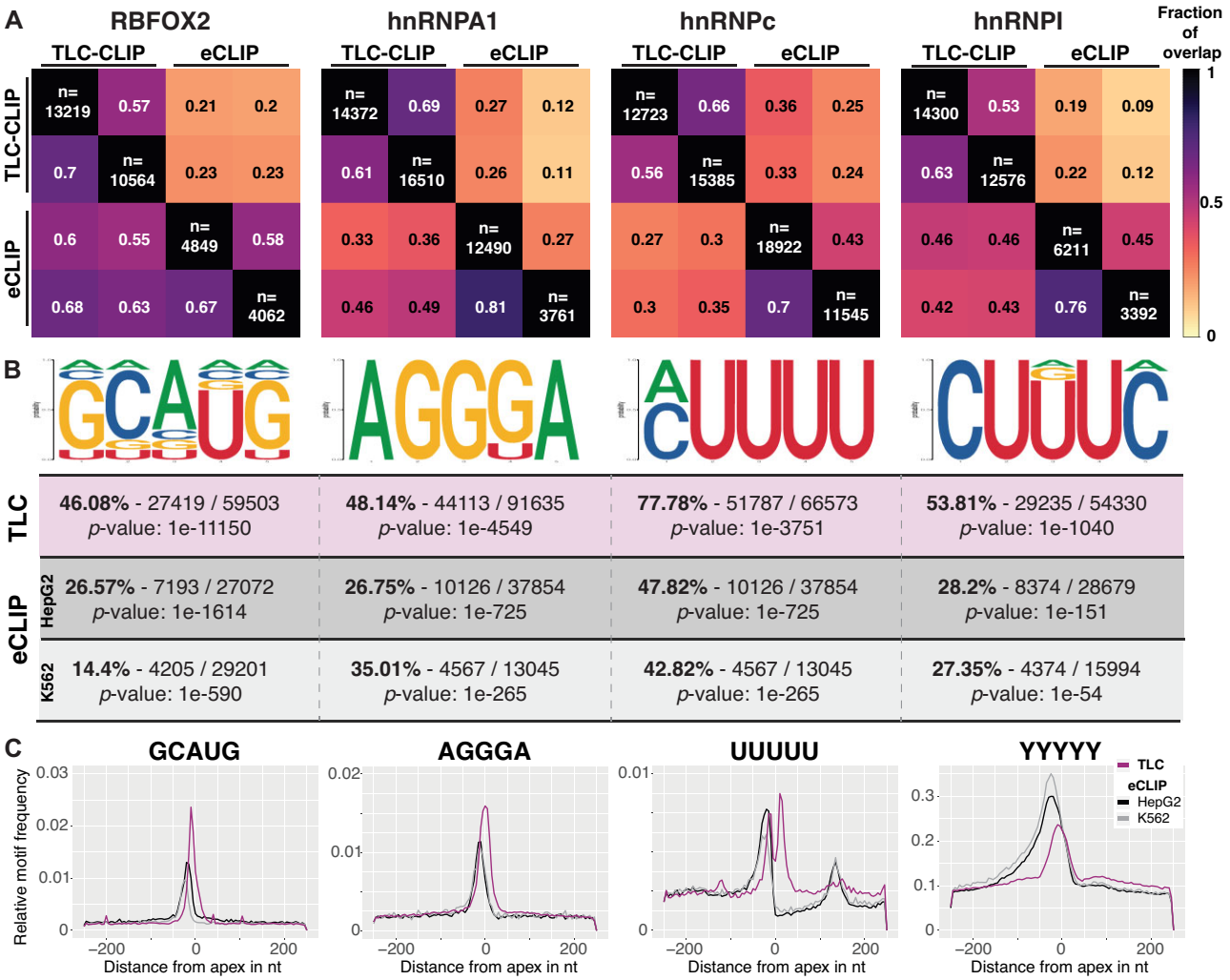


Figure 3. TLC-CLIP libraries recapitulate known binding sites and display increased precision and specificity compared to eCLIP. (**A**) Pairwise comparison showing the fraction of overlap at the peak level for individual replicates between TLC-CLIP and eCLIP libraries for four different RNA-binding proteins, requiring a minimum overlap of 25% between peaks. Peaks are restricted to genes with similar expression level between 293T and HepG2 cells (n = 1507) and individual number of peaks per replicate is annotated in the diagonal of the heatmap. (**B**) Position weight matrix of top motif from *de novo* motif discovery on TLC-CLIP peaks is displayed alongside percentage of peaks carrying motif, as well as p-value obtained from the homer motif discovery software. For eCLIP peaks, percentage of peaks carrying motif resembling the known consensus motif, as well as p-value from homer motif discovery software is displayed. (**C**) Normalised motif density plot for consensus motif across TLC-CLIP and eCLIP peaks, centred on the peak apex defined by CLIPper, which represents the region of highest read coverage. The peaks used for analysis in (B) and (C) were filtered to have a -10log(pval) score larger than 50 and to be present in both biological replicates with a minimum overlap of 25% between peaks (Materials and Methods).

Crosslinking-induced deletions improve nucleotide-resolution and specificity of TLC-CLIP

The precision of TLC-CLIP can be further enhanced by incorporating the positional information of crosslinking-induced deletions (CIDs), which are an alternative outcome to premature termination during reverse transcription, when the reverse transcriptase encounters the residual amino-acid-RNA adduct introduced by UV crosslinking (52,54). Current estimates predict that premature termination at the crosslinking site leads to truncation of cDNA molecules in about 80% of cases, while read-through occurs in the remaining 20%, with the possible introduction of crosslinking-induced deletions, insertions, or mutations (55). These proportions, and the introduction of crosslinking-induced alterations are influenced by RT reaction conditions, particularly the choice of RT enzyme, as well as crosslinking conditions and sensitivity of the profiled protein (51,54). Protocols such as HITS-CLIP rely on such events for the determination of crosslinking sites and have convincingly demonstrated that crosslinking-induced mutations (CIMS) provide excellent resolution (52,56), but they suffer from excessive sample loss and high sequencing requirements due to the exclusion of truncated reads.

TLC-CLIP libraries have a large proportion of reads with CIDs (Figure 2D), which are highly correlated between replicates at the single-nucleotide level (R^2 = 0.47–0.62, Pearson correlation) (Supplementary Figure S9). Deletions are strongly enriched at RBP binding motifs and do not exhibit the previously reported bias towards uracil (55). Instead, we observe an enrichment and high preci-

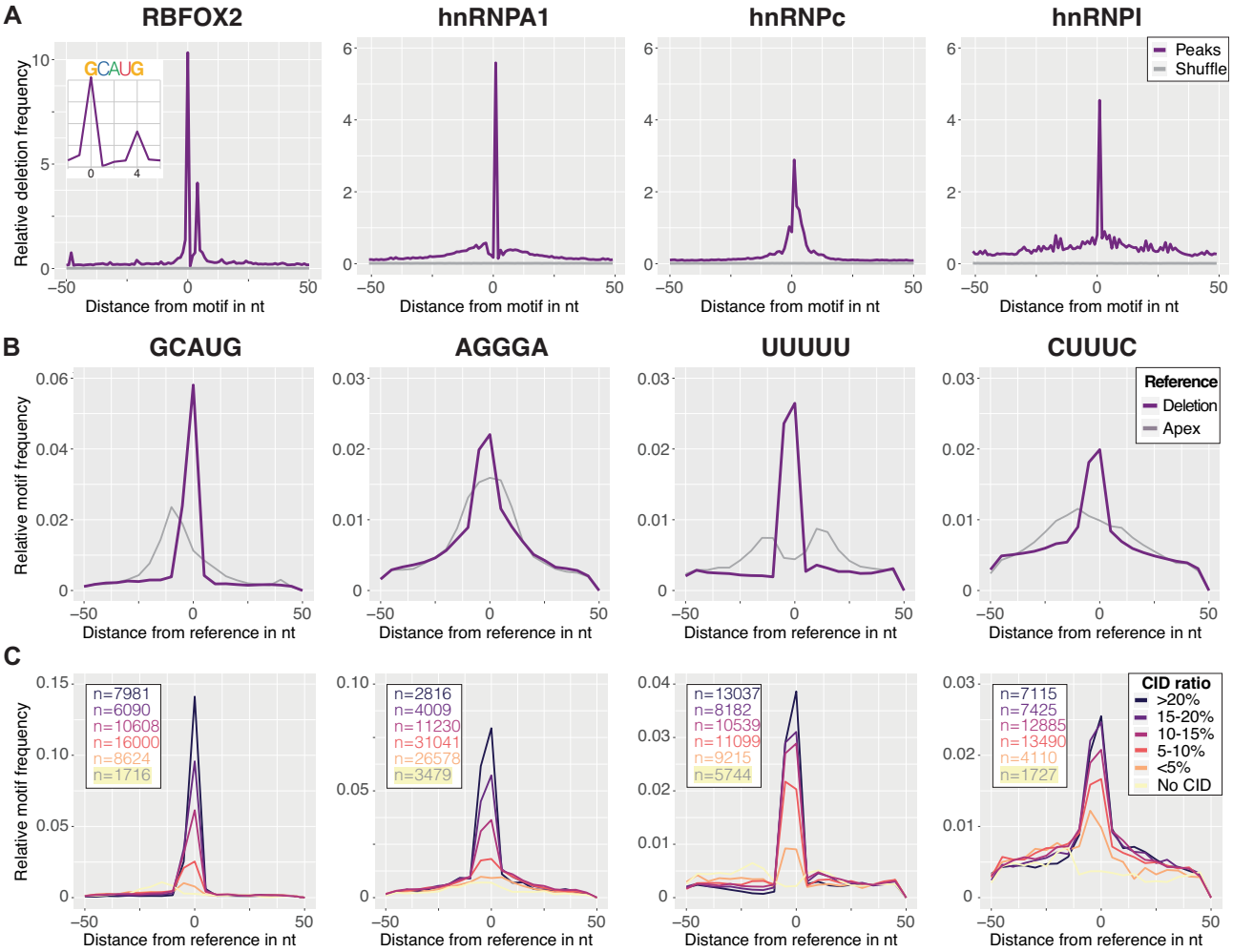


Figure 4. Crosslinking-induced deletions (CIDs) are highly enriched at RBP motifs and improve precision and specificity of TLC-CLIP data. (**A**) Density plot showing enrichment of deletions across peaks which were centred on the corresponding consensus motif for each RBP. For control regions, peaks were shuffled across target gene bodies of a given RBP. (**B**) Normalised motif density plot for consensus motif across peaks centred on either the apex region defined by CLIPper or the position with the highest deletion count. (**C**) Normalised motif density plot for consensus motif across peaks subset based on their crosslink induced deletion (CID) ratio (n_{del}/n_{reads}) at the reference point, which either presents the deletion maximum or apex region for peaks without deletions.

sion of deletions at specific guanine residues for RBFOX2 and hnRNPA1 consensus motifs, identifying the two guanines of the canonical RBFOX binding motif 'GCAUG' as crosslinking sites (57) (Figure 4A). This increased resolution of CIDs can be exploited by centring peaks on the position with the highest number of deletions to further increase the precision of TLC-CLIP data (Figure 4B).

The improvement is particularly noticeable for hnRNPc, for which the apex region defined by CLIPper does not faithfully capture the crosslinking position, resulting in an apparent bi-modal motif enrichment around the reference point (Figures 3C and 4B). Repositioning based on deletions consolidates this signal, with the observed shift away from apex regions recapitulating the bi-modal pattern for hnRNPc, which is not observed for other proteins (Figure 4B and Supplementary Figure S9).

A similar approach can be applied to both eCLIP and iCLIP datasets, by repositioning the reference point of CLIPper peaks to the highest count of start positions as both protocols rely on truncation events to identify crosslinking sites. This mainly results in an upstream shift from the apex region and improves both resolution and motif enrichment, validating our approach of recentring CLIPper peaks onto the most likely crosslinking position (Supplementary Figure S9). While iCLIP shows the strongest motif enrichment for hnRNPc when centred on the apex region, TLC-CLIP provides better resolution after repositioning peaks onto deletions, highlighting the importance of incorporating CIDs during downstream analysis of TLC-CLIP data (Supplementary Figure S9).

In addition to increased resolution, incorporating CID information also improves the specificity of TLC-CLIP data, as demonstrated by stronger motif enrichment in peaks with a higher ratio of crosslinking-induced deletions (Figure 4C). This highlights the benefit of CIDs as an intrinsic quality metric to discern true binding sites from co-purifying, non-crosslinked fragments, which increases the

specificity of TLC-CLIP data without the need to generate matched input samples.

Omission of PAGE purification enables a fully bead-based, two-day workflow amenable to automation

Using CIDs as an intrinsic quality filter is particularly important when applying TLC-CLIP without PAGE purification, which enables a 2-day fully automatable CLIP workflow amenable to high-throughput settings. We observe up to 67% overlap at the peak-level between libraries generated with or without PAGE purification, but with lower motif enrichment in peaks specific to libraries omitting PAGE (Figure 5A and Supplementary Figure S10). Lower motif enrichment is accompanied by lower CID ratios, further confirming the utility of CIDs to filter samples with higher background signal, which can result either from the omission of PAGE purification or sub-optimal RNase conditions (Figure 5B and Supplementary Figure S10).

Annotation of peaks based on different transcript regions reveals an increase in coding sequences (CDS) for peaks with lower CID ratios (Figure 5C). Out of all transcript regions, coding sequences consistently show the lowest motif enrichment for all four splicing factors that were profiled, indicating that these are likely background contamination as previously observed for hnRNPI eCLIP libraries (17) (Figure 5D and Supplementary Figure S10). Thus, while capturing the overall binding behaviour of a given RBP, performing TLC-CLIP without PAGE purification results in a larger number of contaminating background sequences. This outcome is expected as this approach lacks the stringent purification of RNA–protein complexes at the desired molecular weight under denaturing conditions and needs to be carefully considered during downstream data analysis.

Characterisation of co-purifying sequences in TLC-CLIP libraries

Given the increased background signal in samples omitting PAGE purification, we further explored the source of co-purifying fragments by generating TLC-CLIP libraries for all four proteins from non-crosslinked lysates, with and without PAGE purification (Supplementary Figure S1). As expected, libraries from non-crosslinked lysates show very low levels of CIDs confirming their introduction through UV crosslinking, whereas insertions did not show a clear UV-dependence in our data (Supplementary Figure S11).

During alignment, non-crosslinked samples show a higher proportion of reads mapping non-uniquely to the reference genome indicating a higher level of repetitive RNA sequences (Supplementary Figure S7). Indeed, alignment against a repeat index containing all classes of repetitive RNA sequences showed a much higher proportion of reads mapping to ribosomal RNA in non-crosslinked samples, irrespective of PAGE purification (Supplementary Figure S11). A similar increase in repetitive RNA sequences was observed when omitting PAGE purification for crosslinked samples profiling RBFOX2, thus identifying ribosomal RNA as a major source of non-crosslinked co-purifying fragments dependent on the protein of interest and purification method.

We further characterised the distribution of uniquely mapping reads in non-crosslinked samples across different transcript features, which showed a similar pattern for all four proteins, resulting in an increase of unassigned reads mapping to intergenic regions as well as coding sequences (Supplementary Figure S11). The increase in coding sequences was most pronounced for RBFOX2 and hnRNPA1 and included transcripts encoding for highly expressed histone and ribosomal proteins (Supplementary Figure S11). In contrast, libraries for hnRNPc and hnRNPI from non-crosslinked lysates showed fewer coding sequences overall, but mainly protein-specific patterns amongst the most variable CDS between crosslinked and non-crosslinked samples. Increased coverage at these coding sequences was also observed when omitting PAGE purification in crosslinked lysates, which highlights the importance of this purification step in removing co-purifying fragments and the need for stringent computational filtering using either CIDs or non-crosslinked controls when performing TLC-CLIP without PAGE purification (Figure 5C and Supplementary Figure S11).

TLC-CLIP maintains nucleotide resolution and high specificity at low cell numbers

Finally, we showcase the unprecedented sensitivity of our TLC-CLIP protocol, by profiling hnRNPc as well as the much less abundant splicing factor RBFOX2 from 10 000 as well as 1000 cells (Supplementary Figure S1). As expected, lowering the input material reduced the proportion of usable reads, as more amplification cycles are necessary resulting in a larger number of PCR duplicates (Supplementary Figure S7 and S12). While the overall number of peaks was decreased, a large proportion contained the expected consensus motif, which was confidently identified as the top scoring motif during *de novo* motif discovery for TLC-CLIP libraries (Supplementary Figure S12). Comparison with publicly available irCLIP data from low input material for hnRNPc (12) showed higher motif enrichment in TLC-CLIP libraries, demonstrating both increased sensitivity and specificity of TLC-CLIP (Supplementary Figure S12).

TLC-CLIP libraries from low input material maintain a strong enrichment for crosslinking induced deletions at the consensus motif, enabling the identification and visualisation of crosslinking sites with high resolution (Supplementary Figure S12). The position-dependent enrichment of hnRNPc across antisense Alu elements is clearly captured in TLC-CLIP libraries generated from 10 000 and 1000 cells and shows a stronger enrichment of crosslinking sites at the internal linker U-tract for TLC-CLIP libraries compared to irCLIP (Supplementary Figure S12). This is in line with observations by Zarnack et al. showing high affinity of hnRNPc to both internal and terminal U-tracks within Alu sequences (58), and is likely due to increased mappability of TLC-CLIP reads. Deletion-carrying reads, on average, show a longer read length (Figure 2) and span both sides of the crosslink rather than terminating within the U-tract, thus increasing the probability of unique mapping within repetitive sequences.

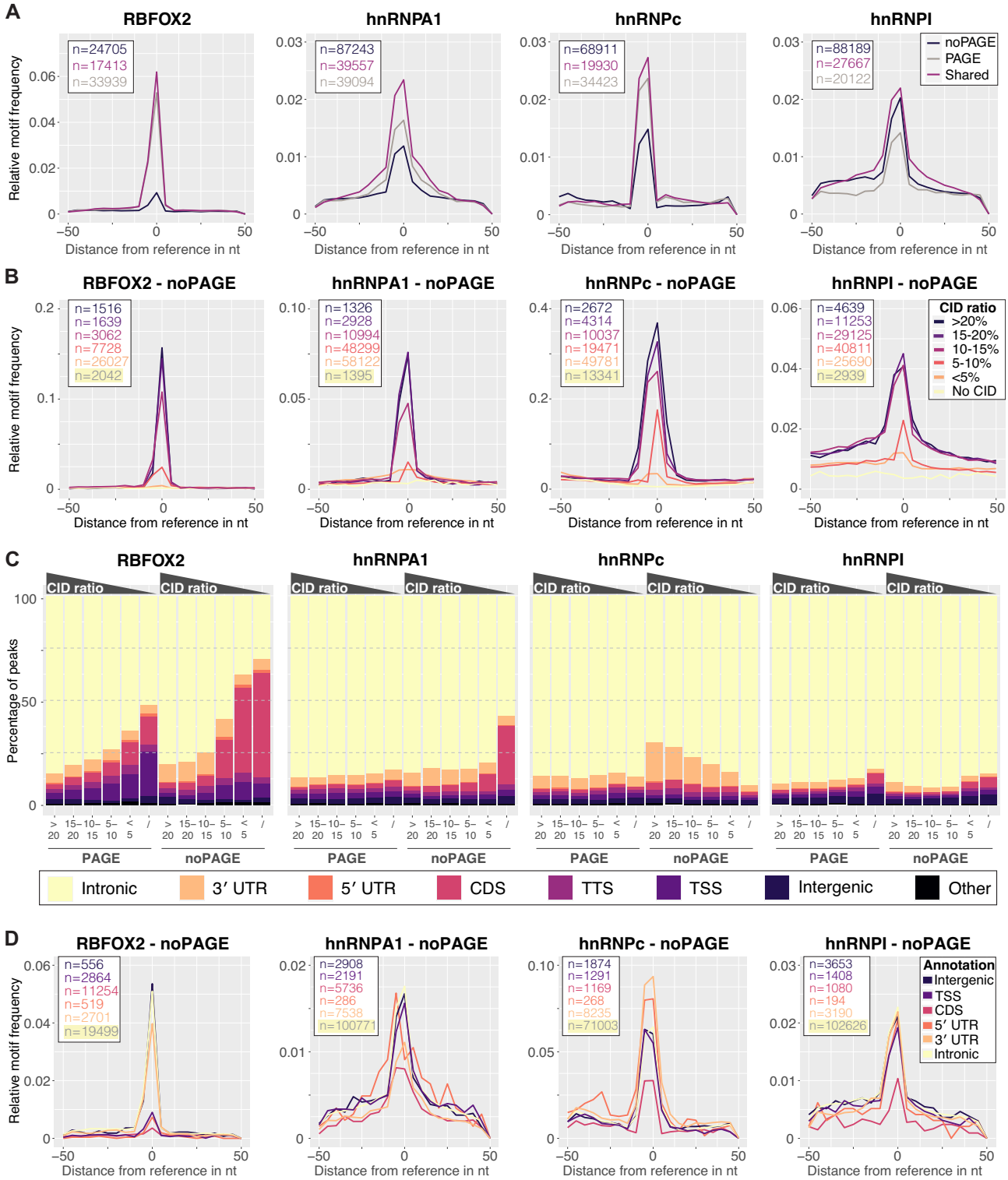


Figure 5. CIDs provide an additional quality filter to increase specificity of TLC-CLIP libraries prepared without PAGE purification. (**A**) Normalised motif density plot for consensus motif across deletion centred peaks that are shared between experimental conditions or specific for libraries generated with or without PAGE purification. (**B**) Normalised motif density plot for consensus motif across peak subsets according to their CID ratio for libraries obtained without PAGE purification. (**C**) Percentage of peaks in TLC-CLIP libraries with varying CID ratios overlapping different genomic annotation layers (Intronic, 3′ UTR = 3′ untranslated region, 5′ UTR = 5′ untranslated region, CDS = coding sequences, TTS = -100 to +1kb around Transcription Termination site, TSS = -1kb - 100bp around Transcription Start Site; Intergenic, Other = including microRNA, non-coding RNA, pseudogenes, snoRNA and scRNA). (**D**) Normalised motif density plot for consensus motif across deletion-centred peaks obtained without PAGE purification that were subset based on the different genomic annotation layers they overlap.

The high resolution of TLC-CLIP was also preserved in low input libraries generated for RBFOX2, showing a clear position-dependent enrichment around splice sites from as few as 1000 cells (Figure 6A). We further tested the enrichment of TLC-CLIP signal around alternatively spliced exons in RBFOX2 knockout (KO) cells, which showed a strong enrichment downstream of 5′ splice sites of down-regulated exons (Figure 6B and C) (59,60). This enrichment is consistent across varying input amounts, demonstrating that TLC-CLIP identifies functionally relevant binding site from extremely limited starting material. Furthermore, the CID ratios of peaks proximal to up- or down-regulated exons were significantly higher compared to control exons that remained unchanged upon RBFOX2 knockout (Figure 6D). This supports our previous observations that CIDs are a useful metric during TLC-CLIP data analysis to identify meaningful binding sites. Taken together, the libraries generated from low cell numbers demonstrate that our TLC-CLIP protocol has extremely high sensitivity and accurately captures the binding profiles even of lowly expressed RBPs while maintaining nucleotide-resolution.

DISCUSSION

The transcriptome-wide identification of RBP binding sites is fundamental to determine the effect of RNA–protein interactions on gene regulation. Protein-centric approaches rely on immunoprecipitation of the RBP of interest and can be performed in native conditions (e.g. RNA immunoprecipitation (RIP)) (61) or after covalent crosslinking of RNA–protein complexes using UV light (e.g. CLIP) (8). CLIP methods allow more stringent purification of RNA–protein complexes and generally have higher resolution compared to RIP, allowing the identification of RNA binding sites with nucleotide resolution (10).

However, CLIP-related protocols have remained technically challenging due to extended experimental procedures, making them prone to sample loss and thus requiring large amounts of starting material. TLC-CLIP presents a novel streamlined library preparation protocol for CLIP-related methods that drastically reduces both experimental time and cost of experiments, while generating high quality RBP binding profiles from low input material. A major advantage of TLC-CLIP is the fully bead-based, single-tube library preparation design that eliminates time-consuming purification procedures and thus reduces sample loss prior to amplification.

Our TLC approach addresses a crucial and often limiting step during the generation of RNA sequencing libraries, namely the generation of second-strand cDNA while preserving the original 3′ end of first-strand cDNA molecules. Current strategies such as in HITS-CLIP and PAR-CLIP circumvent this problem by ligating adapters to both ends of the original RNA molecules, which however results in a drastic loss of material and high sequencing cost as only read-through events can be amplified (5,7,62–64). Alternative approaches that introduce the second adapter sequence at the cDNA step either rely on circularisation of first-strand cDNA (10–12,17) or direct ligation of a second adapter molecule to the 3′ end of first-strand cDNA molecules (13–16). Circularisation was shown to occur with high efficiency (65), but requires specialised RT primers that are prone to concatemerisation (17) and are not compatible with the generation of solid-phase cDNA which is a crucial feature to enable our fully bead-based library preparation workflow. In contrast, the approach of ligating a single-stranded DNA oligonucleotide to first-strand cDNA is enzymatically inefficient (47,48), which could be addressed by using RNA adapters to increases the affinity of T4 RNA Ligase towards the substrate and is expected to have similar efficiency to our TLC approach, but at a higher cost.

TLC therefore presents a cost-efficient option for the second adapter ligation that greatly improves the efficiency of the ssDNA ligation reaction without the need for chimeric adapter molecules. Incorporation of a terminal transferase in the ligation mix leads to the addition of non-template ribonucleotides to the 3′ end of the cDNA in the form of a short ribotail. This greatly increases the efficiency of the ligation reaction by mimicking the 3′ end of an RNA molecule, the preferred substrate of T4 RNA Ligase (48). TLC-CLIP thus limits the requirements for specialised reagents and oligonucleotides to a minimum and generates high complexity libraries, which in turn drastically lower the sequencing depth requirements to enable affordable, large-scale profiling of RNA–protein interactions.

The larger number of crosslinking induced deletions in TLC-CLIP data further improve the precision as well as the specificity of TLC-CLIP libraries by increasing single-nucleotide resolution and distinguishing true binding sites from co-purifying, non-crosslinked fragments. Exploiting CIDs as an intrinsic quality filter during data analysis is similar to computational approaches that are frequently applied to PAR-CLIP data (66). In photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP), cells are treated with modified nucleosides prior to crosslinking which results in increased crosslinking efficiency and changes in base-pair properties causing characteristic base transitions in the obtained sequencing data (8). The enrichment of such base transitions is frequently used to identify true binding sites, as well as estimate the strength of RNA-RBP interactions (67,68). Given the strong enrichment and high resolution of CIDs at RBP binding sites, a similar approach can be applied to TLC-CLIP data to improve data quality for downstream analysis. However, as the rate of CIDs might vary between different proteins, characterisation of a wider set of RBPs with more diverse binding and crosslinking behaviour will be necessary to confirm that CIDs can be consistently used as a quality metric for TLC-CLIP data.

Additional filtering based on CIDs to identify high-confidence binding sites is particularly useful in libraries with higher background signal, which can result from sub-optimal RNase conditions, poor crosslinking efficiency, or less stringent purification of RNA–protein complexes when omitting PAGE purification. The omission of PAGE purification can be desirable in cases where excessive sample loss associated with PAGE purification and membrane transfer precludes the generation of high-quality libraries for a given RBP (69), or in high-throughput settings. For the latter it is highly recommended to perform the necessary control experiments to confirm antibody specificity and optimise RNase conditions prior to library preparation (7).

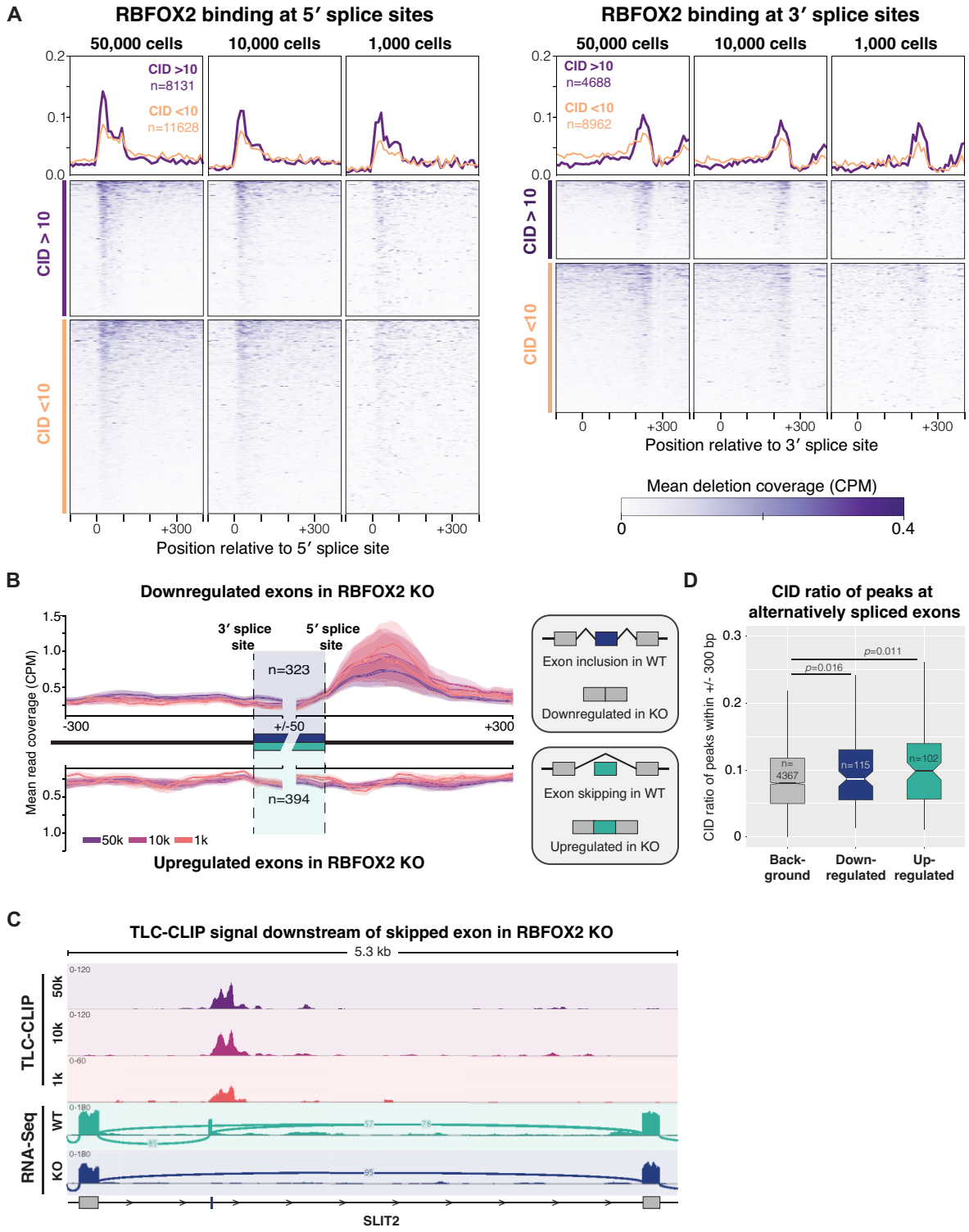


Figure 6. TLC-CLIP enables accurate detection of RBFOX2 binding sites from as few as 1000 cells. (**A**) Mean normalised deletion coverage of RBFOX2 TLC-CLIP libraries prepared from 50 000, 10 000 and 1000 cells in relation to 5′ and 3′ splice sites. Signal is shown across peaks identified in libraries from 50 000 cells subset based on their CID ratio (above or below 10). Regions are ordered based on descending signal intensity in libraries from 50 000 cells. (**B**) Mean normalised read coverage of RBFOX2 TLC-CLIP libraries prepared from 50 000, 10 000 and 1000 cells around 3′ and 5′ splice sites of RBFOX2 regulated exons. Alternatively spliced exons were identified from total RNA-seq of RBFOX2 KO cells compared to wildtype 293T cells with an inclusion level difference larger than 0.1 and a *P*-value cut-off of 0.05. (**C**) IGV genome browser screenshot displaying the enrichment of TLC-CLIP libraries downstream of the alternatively spliced exon 15 in the SLIT2 gene (73). The SLIT2-Δ15 isoform is more prominent in RBFOX2 KO samples with an inclusion level difference of 0.51. (**D**) Notched boxplots displaying the CID ratios of peaks within 300 nucleotides upstream or downstream of RBFOX2 up- or downregulated exons as well as control exons. Statistical significance is displayed in form of *P*-values using Wilcoxon-Mann-Whitney test. Boxes indicate quartiles, lines indicate mean, and whiskers show 1.5 times the inter-quartile range. Outliers are not shown.

Alternative protocols such as FLASH (Fast Ligation of RNA after some sort of Affinity Purification for High-throughput sequencing) (70) and LACE-seq (linear amplification of complementary DNA ends and sequencing) (71) have adopted strategies to profile RNA–protein interactions without PAGE purification; however, neither protocol generates data characteristics that would allow additional filtering of sequencing reads to distinguish crosslinked fragments from background contamination. In FLASH, the removal of PAGE purification is compensated by stringent affinity purification of tagged proteins and in fact, a similar increase in coding sequence contamination was observed when FLASH was performed on endogenous proteins, suggesting that CDS are a common CLIP contaminant when profiling splicing factors under less stringent conditions (70).

As highlighted by our analysis of co-purifying RNA fragments in non-crosslinked controls, we identified repetitive RNA sequences, particularly ribosomal RNA, as a major contaminant alongside increased signal across coding sequences. The latter is highly protein-dependent and further influenced by the stringency of the purification procedures throughout the protocol. This highlights that, as with other CLIP protocols, the downstream analysis of TLC-CLIP needs to be carefully adjusted for individual RBPs and the chosen purification procedure (72). Our findings show that CIDs serve as an effective indicator to distinguish genuine binding sites from non-crosslinked, co-purifying fragments. Filtering based on CIDs thus provides a straightforward computational approach to remove background signal and could provide the foundation for a more sophisticated statistical method for detecting crosslinking sites in TLC-CLIP.

The optimisation of enzymatic reactions in combination with our streamlined library preparation strategy significantly enhances the sensitivity of TLC-CLIP compared to existing protocols. This leads to a drastic reduction in input requirements and enables the generation of high-quality binding profiles from as little as 1000 cells for both hnRNPc and the less abundant RBFOX2. This makes TLC-CLIP the most sensitive protocol among CLIP-related techniques for studying endogenous proteins while preserving denaturing purification conditions.

In sum, TLC-CLIP presents a fully bead-based, single-tube library preparation strategy for CLIP protocols that generates high-quality RNA binding profiles with increased sensitivity and precision from low input material and is amenable to automation. As such, it constitutes an attractive technique in high-throughput settings such as drug or CRISPR screenings, or for studying RNA–protein complexes in lowly abundant biological samples.

DATA AVAILABILITY

Sequencing data have been deposited on Gene Expression Omnibus under accession number GSE225358, with SubSeries GSE200432 containing TLC-CLIP data and SubSeries GSE225357 containing RNA-Seq data.

UCSC browser tracks are available under the following link: http://genome-euro.ucsc.edu/s/christinaernst/TLC%2DCLIP_NAR.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Brian Zarnegar, Douglas Porter and Paul Khavari for providing reagents and guidance on initial irCLIP experiments. We thank the Gene Expression Core Facility (GECF) and Flow Cytometry Core Facility (FCCF) at EFPL for technical support.

FUNDING

European Research Council [KRABnKAP #268721 and Transpos-X #694658 to D.T.]; Swiss National Science Foundation [PZ00P3_202048 to C.E., 310030_152879 and 310030B_173337 to D.T.]; Human Frontiers Science Programme [LT000147/2019 to C.E.]; European Molecular Biology Organisation [ALTF 516-2019 to C.E.]; EPFL-Stanford Exchange programme. Funding for open access charge: EPFL.

Conflict of interest statement. C.E. and D.T. are inventors on a patent application covering specific elements of this method (i.e. construction of sequencing libraries from RNA using tailing and ligation of cDNA (TLC)) (PCT/EP2023/058731).

REFERENCES

1. Hentze,M.W., Castello,A., Schwarzl,T. and Preiss,T. (2018) A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.*, **19**, 327–341.
2. Gebauer,F., Schwarzl,T., Valcárcel,J. and Hentze,M.W. (2021) RNA-binding proteins in human genetic disease. *Nat. Rev. Genet.*, **22**, 185–198.
3. Dominguez,D., Freese,P., Alexis,M.S., Su,A., Hochman,M., Palden,T., Bazile,C., Lambert,N.J., Nostrand,E.L.V., Pratt,G.A. *et al.* (2018) Sequence, structure, and context preferences of human RNA binding proteins. *Mol. Cell*, **70**, 854–867.
4. Corley,M., Burns,M.C. and Yeo,G.W. (2020) How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol. Cell*, **78**, 9–29.
5. Ule,J., Jensen,K.B., Ruggiu,M., Mele,A., Ule,A. and Darnell,R.B. (2003) CLIP identifies Nova-regulated RNA networks in the brain. *Science*, **302**, 1212–1215.
6. Ule,J., Jensen,K., Mele,A. and Darnell,R.B. (2005) CLIP: a method for identifying protein–RNA interaction sites in living cells. *Methods*, **37**, 376–386.
7. Lee,F.C.Y. and Ule,J. (2018) Advances in CLIP technologies for studies of protein-RNA interactions. *Mol. Cell*, **69**, 354–369.
8. Hafner,M., Katsantoni,M., Köster,T., Marks,J., Mukherjee,J., Staiger,D., Ule,J. and Zavolan,M. (2021) CLIP and complementary methods. *Nat. Rev. Methods Primers*, **1**, 20.
9. Ramanathan,M., Porter,D.F. and Khavari,P.A. (2019) Methods to study RNA–protein interactions. *Nat. Methods*, **16**, 225.
10. König,J., Zarnack,K., Rot,G., Curk,T., Kayikci,M., Zupan,B., Turner,D.J., Luscombe,N.M. and Ule,J. (2010) iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat. Struct. Mol. Biol.*, **17**, 909–915.
11. Huppertz,I., Attig,J., D'Ambrogio,A., Easton,L.E., Sibley,C.R., Sugimoto,Y., Tajnik,M., König,J. and Ule,J. (2014) iCLIP: protein–RNA interactions at nucleotide resolution. *Methods*, **65**, 274–287.
12. Zarnegar,B.J., Flynn,R.A., Shen,Y., Do,B.T., Chang,H.Y. and Khavari,P.A. (2016) irCLIP platform for efficient characterization of protein–RNA interactions. *Nat. Methods*, **13**, 489–492.
13. Nostrand,E.L.V., Pratt,G.A., Shishkin,A.A., Gelboin-Burkhart,C., Fang,M.Y., Sundararaman,B., Blue,S.M., Nguyen,T.B., Surka,C.,

Elkins,K. *et al.* (2016) Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat. Methods*, **13**, 508–514.
14. Van Nostrand,E.L., Freese,P., Pratt,G.A., Wang,X., Wei,X., Xiao,R., Blue,S.M., Chen,J.-Y., Cody,N.A.L., Dominguez,D. *et al.* (2020) A large-scale binding and functional map of human RNA-binding proteins. *Nature*, **583**, 711–719.
15. Blue,S.M., Yee,B.A., Pratt,G.A., Mueller,J.R., Park,S.S., Shishkin,A.A., Starner,A.C., Van Nostrand,E.L. and Yeo,G.W. (2022) Transcriptome-wide identification of RNA-binding protein binding sites using seCLIP-seq. *Nat. Protoc.*, **17**, 1223–1265.
16. Buchbender,A., Mutter,H., Sutandy,F.X.R., Körtel,N., Hänel,H., Busch,A., Ebersberger,S. and König,J. (2020) Improved library preparation with the new iCLIP2 protocol. *Methods*, **178**, 33–48.
17. Lee,F.C.Y., Chakrabarti,A.M., Hänel,H., Monzón-Casanova,E., Hallegger,M., Militti,C., Capraro,F., Sadée,C., Toolan-Kerr,P., Wilkins,O. *et al.* (2021) An improved iCLIP protocol. bioRxiv doi: https://doi.org/10.1101/2021.08.27.457890, 27 August 2021, preprint: not peer reviewed.
18. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
19. Concordet,J.-P. and Haeussler,M. (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.*, **46**, W242–W245.
20. Ran,F.A., Hsu,P.D., Wright,J., Agarwala,V., Scott,D.A. and Zhang,F. (2013) Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.*, **8**, 2281–2308.
21. Daneshvar,K., Pondick,J.V., Kim,B.-M., Zhou,C., York,S.R., Macklin,J.A., Abualteen,A., Tan,B., Sigova,A.A., Marcho,C. *et al.* (2016) DIGIT is a conserved long noncoding RNA that regulates GSC expression to control definitive endoderm differentiation of embryonic stem cells. *Cell Rep.*, **17**, 353–365.
22. Dodt,M., Roehr,J.T., Ahmed,R. and Dieterich,C. (2012) FLEXBAR—flexible barcode and adapter processing for next-generation sequencing platforms. *Biology (Basel)*, **1**, 895–905.
23. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
24. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.
25. Liao,Y., Smyth,G.K. and Shi,W. (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
26. Smith,T.S., Heger,A. and Sudbery,I. (2017) UMI-tools: modelling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Res.*, **27**, 491–499. .
27. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
28. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
29. Robinson,M.D., McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
30. Khan,A. and Mathelier,A. (2017) Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinf.*, **18**, 287.
31. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
32. Sahadevan,S., Sekaran,T., Ashaf,N., Fritz,M., Hentze,M.W., Huber,W. and Schwarzl,T. (2023) htseq-clip: a toolset for the preprocessing of eCLIP/iCLIP datasets. *Bioinformatics*, **39**, btac747.
33. Smit,A.F.A., Hubley,R. and Green,P. (2015) RepeatMasker Open-4.0. 2013-2015.
34. Ramirez,F., Dundar,F., Diehl,S., Gruning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.
35. Shen,S., Park,J.W., Huang,J., Dittmar,K.A., Lu,Z., Zhou,Q., Carstens,R.P. and Xing,Y. (2012) MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Res.*, **40**, e61.
36. Shen,S., Park,J.W., Lu,Z., Lin,L., Henry,M.D., Wu,Y.N., Zhou,Q. and Xing,Y. (2014) rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, E5593–E5601.
37. Wickham,H., Averick,M., Bryan,J., Chang,W., McGowan,L.D., François,R., Grolemund,G., Hayes,A., Henry,L., Hester,J. *et al.* (2019) Welcome to the Tidyverse. *J Open Source Softw.*, **4**, 1686.
38. Rio,D.C. (2014) 5'-end labeling of RNA with [γ-32P]ATP and T4 polynucleotide kinase. *Cold Spring Harb. Protoc.*, **2014**, 441–443.
39. Zhelkovsky,A.M. and McReynolds,L.A. (2011) Simple and efficient synthesis of 5' pre-adenylated DNA using thermostable RNA ligase. *Nucleic Acids Res.*, **39**, e117.
40. Porter,D.F., Miao,W., Yang,X., Goda,G.A., Ji,A.L., Donohue,L.K.H., Aleman,M.M., Dominguez,D. and Khavari,P.A. (2021) easyCLIP analysis of RNA–protein interactions incorporating absolute quantification. *Nat. Commun.*, **12**, 1569.
41. Rio,D.C., Ares,M., Hannon,G.J. and Nilsen,T.W. (2010) Ethanol Precipitation of RNA and the Use of Carriers. *Cold Spring Harb. Protoc.*, **2010**, pdb.prot5440.
42. Roeder,T. (1998) Solid-phase cDNA library construction, a versatile approach. *Nucleic Acids Res.*, **26**, 3451–3452.
43. Edwards,J.B., Delort,J. and Mallet,J. (1991) Oligodeoxyribonucleotide ligation to single-stranded cDNAs: a new tool for cloning 5' ends of mRNAs and for constructing cDNA libraries by in vitro amplification. *Nucleic Acids Res.*, **19**, 5227–5232.
44. Clepet,C., Le Clainche,I. and Caboche,M. (2004) Improved full-length cDNA production based on RNA tagging by T4 DNA ligase. *Nucleic Acids Res.*, **32**, e6.
45. Scotto–Lavino,E., Du,G. and Frohman,M.A. (2006) 5′ end cDNA amplification using classic RACE. *Nat. Protoc.*, **1**, 2555–2562.
46. Matz,M., Shagin,D., Bogdanova,E., Britanova,O., Lukyanov,S., Diatchenko,L. and Chenchik,A. (1999) Amplification of cDNA ends based on template-switching effect and step-out PCR. *Nucleic Acids Res.*, **27**, 1558–1560.
47. Bullard,D.R. and Bowater,R.P. (2006) Direct comparison of nick-joining activity of the nucleic acid ligases from bacteriophage T4. *Biochem. J.*, **398**, 135–144.
48. Miura,F., Shibata,Y., Miura,M., Sangatsuda,Y., Hisano,O., Araki,H. and Ito,T. (2019) Highly efficient single-stranded DNA ligation technique improves low-input whole-genome bisulfite sequencing by post-bisulfite adaptor tagging. *Nucleic Acids Res.*, **47**, e85.
49. Schmidt,W.M. and Mueller,M.W. (1996) Controlled Ribonucleotide Tailing of cDNA ends (CRTC) by Terminal Deoxynucleotidyl Transferase: a New Approach in PCR-Mediated Analysis of mRNA Sequences. *Nucleic Acids Res.*, **24**, 1789–1791.
50. Busch,A., Brüggemann,M., Ebersberger,S. and Zarnack,K. (2020) iCLIP data analysis: a complete pipeline from sequencing reads to RBP binding sites. *Methods*, **178**, 49–62.
51. Van Nostrand,E.L., Shishkin,A., Pratt,G.A., Nguyen,T.B. and Yeo,G.W. (2017) Variation in single-nucleotide sensitivity of eCLIP derived from reverse transcription conditions. *Methods*, **126**, 29–37.
52. Zhang,C. and Darnell,R.B. (2011) Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data. *Nat. Biotechnol*, **29**, 607–614.
53. Lovci,M.T., Ghanem,D., Marr,H., Arnold,J., Gee,S., Parra,M., Liang,T.Y., Stark,T.J., Gehman,L.T., Hoon,S. *et al.* (2013) Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. *Nat. Struct Mol. Biol.*, **20**, 1434–1442.
54. Hauer,C., Curk,T., Anders,S., Schwarzl,T., Alleaume,A.-M., Sieber,J., Hollerer,I., Bhuvanagiri,M., Huber,W., Hentze,M.W. *et al.* (2015) Improved binding site assignment by high-resolution mapping of RNA–protein interactions using iCLIP. *Nat. Commun.*, **6**, 7921.
55. Sugimoto,Y., König,J., Hussain,S., Zupan,B., Curk,T., Frye,M. and Ule,J. (2012) Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biol.*, **13**, R67.
56. Moore,M.J., Zhang,C., Gantman,E.C., Mele,A., Darnell,J.C. and Darnell,R.B. (2014) Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis. *Nat. Protoc.*, **9**, 263–293.

57. Weyn-Vanhentenryck,S.M., Mele,A., Yan,Q., Sun,S., Farny,N., Zhang,Z., Xue,C., Herre,M., Silver,P.A., Zhang,M.Q. *et al.* (2014) HITS-CLIP and Integrative Modeling Define the Rbfox Splicing-Regulatory Network Linked to Brain Development and Autism. *Cell Rep.*, **6**, 1139–1152.
58. Zarnack,K., König,J., Tajnik,M., Martincorena,I., Eustermann,S., Stévant,I., Reyes,A., Anders,S., Luscombe,N.M. and Ule,J. (2013) Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, **152**, 453–466.
59. Yeo,G.W., Coufal,N.G., Liang,T.Y., Peng,G.E., Fu,X.-D. and Gage,F.H. (2009) An RNA code for the FOX2 splicing regulator revealed by mapping RNA–protein interactions in stem cells. *Nat. Struct. Mol. Biol.*, **16**, 130–137.
60. Zhou,D., Couture,S., Scott,M.S. and Abou Elela,S. (2021) RBFOX2 alters splicing outcome in distinct binding modes with multiple protein partners. *Nucleic Acids Res.*, **49**, 8370–8383.
61. Selth,L.A., Gilbert,C. and Svejstrup,J.Q. (2009) RNA immunoprecipitation to determine RNA–protein associations in vivo. *Cold Spring Harb. Protoc.*, **2009**, pdb.prot5234.
62. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M., Jungkamp,A.-C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and MicroRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
63. Garzia,A., Meyer,C., Morozov,P., Sajek,M. and Tuschl,T. (2017) Optimization of PAR-CLIP for transcriptome-wide identification of binding sites of RNA-binding proteins.*Methods*, **118–119**, 24–40.
64. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E., Wang,X. *et al.* (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
65. Heyer,E.E., Ozadam,H., Ricci,E.P., Cenik,C. and Moore,M.J. (2015) An optimized kit-free method for making strand-specific deep sequencing libraries from RNA fragments. *Nucleic Acids Res.*, **43**, e2.
66. Danan,C., Manickavel,S. and Hafner,M. (2016) PAR-CLIP: a method for transcriptome-wide identification of RNA binding protein interaction sites. *Methods Mol. Biol.*, **1358**, 153–173.
67. Corcoran,D.L., Georgiev,S., Mukherjee,N., Gottwein,E., Skalsky,R.L., Keene,J.D. and Ohler,U. (2011) PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biol.*, **12**, R79.
68. Sievers,C., Schlumpf,T., Sawarkar,R., Comoglio,F. and Paro,R. (2012) Mixture models and wavelet transforms reveal high confidence RNA–protein interaction sites in MOV10 PAR-CLIP data. *Nucleic Acids Res.*, **40**, e160.
69. Anastasakis,D.G., Jacob,A., Konstantinidou,P., Meguro,K., Claypool,D., Cekan,P., Haase,A.D. and Hafner,M. (2021) A non-radioactive, improved PAR-CLIP and small RNA cDNA library preparation protocol. *Nucleic Acids Res.*, **49**, e45.
70. Ilik,I.A., Aktas,T., Maticzka,D., Backofen,R. and Akhtar,A. (2020) FLASH: ultra-fast protocol to identify RNA–protein interactions in cells. *Nucleic Acids Res.*, **48**, e15.
71. Su,R., Fan,L.-H., Cao,C., Wang,L., Du,Z., Cai,Z., Ouyang,Y.-C., Wang,Y., Zhou,Q., Wu,L. *et al.* (2021) Global profiling of RNA-binding protein target sites by LACE-seq. *Nat. Cell Biol.*, **23**, 664–675.
72. Chakrabarti,A.M., Haberman,N., Praznik,A., Luscombe,N.M. and Ule,J. (2018) Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies. *Annu. Rev. Biomed. Data Sci.*, **1**, 235–261.
73. Lin,Y.-Y., Yang,C.-H., Sheu,G.-T., Huang,C.-Y.F., Wu,Y.-C., Chuang,S.-M., Fann,M.-J., Chang,H., Lee,H. and Chang,J.T. (2011) A novel exon 15-deleted, splicing variant of Slit2 shows potential for growth inhibition in addition to invasion inhibition in lung cancer. *Cancer*, **117**, 3404–3415.