doi: 10.1093/nar/gky858

Monitored eCLIP: high accuracy mapping of RNA-protein interactions

Rémi Hocq[†], Janio Paternina[†], Quentin Alasseur, Auguste Genovesio[*] and Hervé Le Hir[*]

Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS UMR8197, INSERM U1024, PSL Research University, 75005 Paris, France

Received May 03, 2018; Revised August 31, 2018; Editorial Decision September 11, 2018; Accepted September 13, 2018

ABSTRACT

CLIP-seq methods provide transcriptome-wide snapshots of RNA-protein interactions in live cells. Reverse transcriptases stopping at cross-linked nucleotides sign for RNA-protein binding sites. Reading through cross-linked positions results in false binding site assignments. In the 'monitored enhanced CLIP' (meCLIP) method, a barcoded biotinylated linker is ligated at the 5′ end of cross-linked RNA fragments to purify RNA prior to the reverse transcription. cDNAs keeping the barcode sequence correspond to reverse transcription read-throughs. Read through occurs in unpredictable proportions, representing up to one fourth of total reads. Filtering out those reads strongly improves reliability and precision in protein binding site assignment.

INTRODUCTION

Post-transcriptional gene regulation is governed by hundreds of RNA binding proteins (RBPs). RBPs form ribonucleoprotein complexes with all kind of RNAs to function as genetic information support, structural scaffold, interaction guide, or enzyme. The repertoire of eukaryotic RBPs comprises over 1500 different RBPs in human (1). In the case of human messenger RNAs (mRNAs), literally covered by proteins, RNA is in direct contact with >800 different RBPs (2,3), which modulate transcript processing and destiny (4). Despite the physiological importance of RBPs evidenced by their implication in diverse pathologies (5), the precise function of most RBPs remains obscure. The development of the cross-linking and immunoprecipitation (CLIP) method represented a pioneering step in the quest of RBP mapping (6). The basic principle of this strategy is the covalent binding of RBPs with their direct RNA targets by ultraviolet (UV) light irradiation. Once cross-linked, RNA digestion separates RNA-protein complexes before immunoprecipitation under stringent washing conditions. Coupled to high throughput sequencing, CLIP offers a transcriptome-wide snapshot of RNA-protein interactions in live cells as covalent links are formed before any disturbing purification step (7). The importance of CLIP methods prompted the community to further improve their efficiency, specificity, and accuracy, as reviewed recently by Lee and Ule (8).

A major caveat of CLIP methods is the poor efficiency of UV-C crosslinking, which is estimated not to exceed a few percent (9). The crosslinking efficiency *per se* can be strongly improved by using photoactivatable ribonucleosides combined with UV-A irradiation (PAR-CLIP) (10). However, incorporation in living cells of nucleoside analogs into RNA is likely to introduce a bias in the RNA sequences that interact with RBPs.

In addition to cross link, cDNA library preparation further decreases CLIP efficiency. After purification and protein digestion, cross-linked peptides remain attached to RNA fragments. This cross-linking mark partially blocks reverse transcriptase (RTase) progression during cDNA synthesis (11). This issue is circumvented by CLIP strategies in different ways. In the HITS-CLIP protocol, cDNA library preparations are based on adaptors ligated at both RNA extremities. Hence, cDNA fragments terminated at the cross-linking site do not harbor the 5′ adaptor and cannot be amplified by PCR. Thus, only cDNA fragments resulting from RTase bypassing the cross-linking site (read-through) are sequenced (Figure 1A). It was then suggested that the center of these read-through reads corresponds on average to the binding site (12), and thus shorter RNA fragments provide higher binding site accuracy. This is limited, however, by the minimal read length (of around 20 nt) required for an unambiguous mapping (13).

The individual-nucleotide resolution CLIP (iCLIP) protocol was conceived to recover truncated cDNA, which may constitute a large fraction of the total cDNA fragments (14). With this approach, a single adaptor is ligated to the 3′-end of RNA fragments before reverse transcription. After circularization and relinearization, cDNAs are amplified by PCR independently of cDNA termination (Figure 1B). The

[*]To whom correspondence should be addressed. Tel: +33 1 44 32 39 45; Fax: +33 1 44 32 39 45; Email: lehir@ens.fr
Correspondence may also be addressed to Auguste Genovesio. Email: auguste.genovesio@ens.fr
[†]The authors wish the first two authors to be regarded as joint First Authors.
Present address: Rémi Hocq, IFP Energies nouvelles, Département Biotechnologie, Rueil-Malmaison 92852, France.

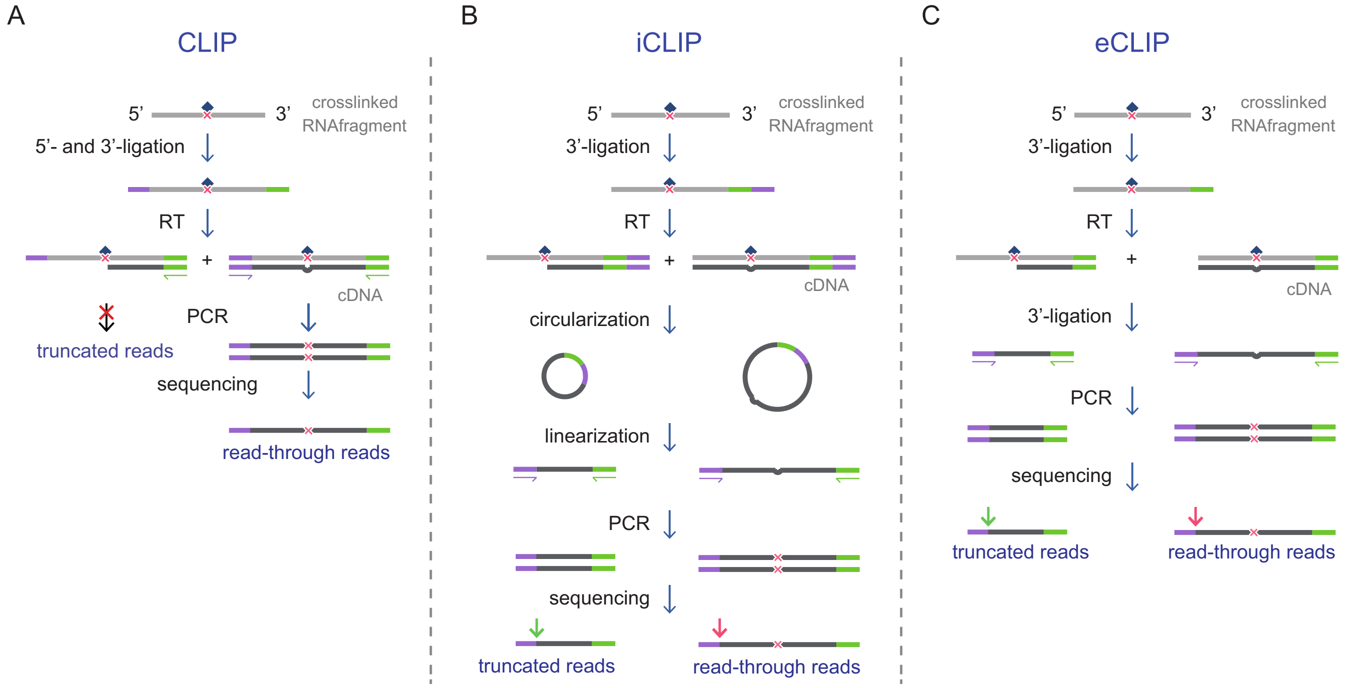

Figure 1. Comparison of CLIP, iCLIP, and eCLIP procedures. Scheme of the CLIP protocol. Immunoprecipitated RNA fragments are coupled to a peptide (blue square) at the crosslinking site (red cross). Reverse-transcription (RT) either stops or reads through the crosslinking site. (**A**) For CLIP, adaptors (purple and green) are ligated at both extremities of the crosslinked RNA fragments. Only read-through cDNAs can be amplified by PCR using primers complementary to adaptors generating read-through reads. (**B**) For iCLIP, a single bipartite adaptor is ligated at the 3′ extremity of the crosslinked RNA fragments. Full-length or truncated cDNAs are circularized and then linearized leading to the presence of adaptors at both extremities. PCR amplifies both truncated and read-through reads. (**C**) For eCLIP, a single adaptor (green) is ligated at the 3′ extremity of the crosslinked RNA fragments. After RT, a second adaptor (purple) is ligated to the 3′ extremity of the cDNAs. PCR amplifies both truncated and read-through reads. The green arrows point towards the position of the crosslinking site. The red arrows point towards the 3′ extremity of the cDNA, upstream the crosslinking site.

first sequenced nucleotide of truncated reads, after 5′ adaptor removal, corresponds to the nucleotide where the reverse transcriptase stopped, one nucleotide downstream of the cross-linking site (14). More recently, the enhanced CLIP (eCLIP) (15), infrared-CLIP (irCLIP (16)) and bromodeoxyuridine CLIP (BrdU-CLIP) (17) methods also suggested improvements of the library construction in order to capture all cDNAs (18). In the case of eCLIP notably, adaptors are ligated first at the 3′-end of RNA and next at the 3′-end of the cDNA, hence bypassing a relatively low-yield circularization step (Figure 1C). In addition, eCLIP includes a parallel analysis of the size-matched input (SM-input) control to identify the most abundant non-specific RNA fragments contributing to background signal (15).

iCLIP- and eCLIP-related methods provide single nucleotide resolution of the cross-linked site as reverse transcriptase tends to stop one nucleotide downstream of the cross-linking site (14) (Figure 1B, C). While in theory the truncation site is independent of the read length, considering various cDNA lengths helps RBP binding site assignment (19). This may result partly from a non-negligible population of read-through reads, whose mapping precision is affected by read length. This population can nonetheless be computationally estimated. Indeed, when passing through the cross-linked nucleotide, reverse transcriptases generate mutations (20). These cross-linking induced mutation sites (CIMS) are valuable both for CLIP-related methods to localize the binding sites and for iCLIP/eCLIP derivatives to estimate the proportion of read-through reads (21). However, CIMS occurrence is variable between RBPs and often remains low among read-through reads, thus preventing a precise binding site mapping (12,21). Furthermore, when long RNA fragments are purified, many sequenced reads are too short to reach the CIMS. In the case of RBPs recognizing specific motifs, adding a motif search can help to map binding sites (12,19,21). However, using motif search is limited by RBPs generally targeting low complexity sequences (22).

For both iCLIP and eCLIP, mapping accuracy depends upon the proportion of truncated reads versus read-through reads, as the latter correspond to spurious cross-linking sites. Consequently, a major hurdle hit by CLIP-related methods is the unpredictable behavior of reverse transcriptases: they either stop or read through the crosslinking site on RNA fragments in a stochastic way. Read-through reads may represent a large percentage of total reads, hence may distort binding site assignment. To discriminate truncated reads from read-through reads, we modified the eCLIP pipeline to establish a 'monitored eCLIP' (meCLIP) protocol. The major modification consisted in adding a biotinylated oligonucleotide ligation to the 5′-end of the RNA fragments to discriminate read-through from truncated cDNA fragments. The ability of the reverse transcriptase to read through the RBPs cross-linking site is monitored to systematically discriminate and filter out read-through reads that generate imprecise peaks. Here, we applied meCLIP to the human RNA helicases, eIF4A3 (eukaryotic initiation factor 4A3), a core component of the exon junction complex (EJC) (23), and UPF1 (up-frameshift factor 1), an essential factor for the nonsense-mediated RNA decay (NMD) (24).

MATERIALS AND METHODS

Plasmids and molecular cloning

Genome edition of endogenous *eIF4A3* was achieved through expression of Cas9 (*Streptococcus pyogenes*) nickase from pX335 (Addgene) and sgRNAs from sgRNA expression vectors (kind gift from Edouard Bertrand). The sgRNA expression vector displays an optimized sequence for the improved expression of the sgRNA as previously described (25). Briefly, the sgRNA scaffold has been engineered to remove an RNA polymerase III stop motif and to stabilize the hairpin structure recognized by Cas9. sgRNA protospacers sequences were designed using eCRISP (http://www.e-crisp.org/E-CRISP/). Insertion of those sequences in the sgRNA expression vectors was done by Golden Gate assembly into Bbs1 restriction sites. For UPF1 genome edition, sgRNA sequence from pX335 was replaced with the one from the sgRNA expression vector and a second one was added with the Bsa1 restriction site by Gibson assembly (26). Corresponding sgRNAs targeting the C-terminal region of UPF1 were then cloned by Golden Gate assembly in Bbs1 and Bsa1 restriction sites.

eIF4A3 homology regions (800 bp upstream and downstream the stop codon, with modification of PAM sequences to prevent re-cutting) were chemically synthesized (Genewiz) in pUC19. UPF1 homology regions (1000 bp upstream and downstream the stop codon) were amplified by PCR on HeLa genomic DNA. pUC57 vectors comprising sequences coding for the 3× HA affinity tag, an IRES2, the puromycin resistance gene and the SV40 polyadenylation signal were a gift from E. Bertrand's laboratory. For UPF1 edition, the tagging cassette was modified by replacement of the TEV cleavage site by a 3C proteolytic site repositioned after the HA affinity tag and addition of a 3xFLAG tag. These modifications were ordered as a gBlock DNA fragment (IDT). Final repair plasmids were obtained by assembly of the homology regions and the tagging cassettes by Gibson assembly (26).

Cell culture

Human HeLa cells were grown in DMEM supplemented with GlutaMAX, 4.5 g/l glucose, 110 mg/l sodium pyruvate, 10% fetal bovine serum, 100 U/mL penicillin and 100 μg/ml streptomycin (Life Technologies). Cells were passaged every 3–4 days following standard procedure and cultivated in a humidified incubator at 37°C with 5% CO_2. Five million cells at 50% confluency were co-transfected using homemade JetPEI reagent with 0.5 μg pX335, 1.5 μg of each of the sgRNAs expression vectors and 1.5 μg of repair plasmid (eIF4A3) or 0.5 μg of modified pX335 and 4.5 μg of repair plasmid (UPF1). Cells were split in five 15 cm dishes 24h post-transfection and puromycin (InvivoGen) at various concentrations (0, 250, 500, 1000 and 2000 ng/ml) was added 24 h later. Medium was replaced every 3–4 days with fresh antibiotic for 10–15 days. Clones obtained at the highest puromycin concentration were picked in 96-well plates and expanded for an additional 15 days.

Transgene integration and expression

For genomic DNA extraction, cells were lysed in PXL lysis buffer containing proteinase K and RNase A at 37°C. Lysates were centrifuged and the supernatants were then precipitated with 2 volumes of ethanol following phenol–chloroform–alcohol isoamyl (25:24:1) extraction. Transgene integration at the correct locus was verified by PCR with primers annealing upstream the targeted region and in the insertion. Homozygosis was then investigated by PCR with primers annealing upstream and downstream the homology regions. Expression was tested by western blot on the soluble fraction of a cell lysate with antibodies directed against the endogenous proteins and/or the affinity tag itself. Correct integration of eIF4A3 into the native EJC was tested on cell line eIF4A3-HA (clone B) by co-immunoprecipitation and western blot.

Antibodies for immunoprecipitation

Anti-eIF4A3 were previously described (27). HA & Flag tagged proteins were respectively immunoprecipitated with Pierce anti-HA magnetic beads (Life Technologies) and M2 anti-FLAG magnetic beads (Sigma).

Oligonucleotides design and sequences

RNA and DNA linker sequences from the published eCLIP procedure (15) were modified in order to allow sequencing of the library in single-end mode and to be compatible with the P3/P5 PCR primers from Solexa used in standard iCLIP. Random and multiplex barcodes were placed on the second ligation primer. All the sequences are available in Supplementary Text S1. All the oligonucleotides were purchased from IDT and Eurofins Genomics and were ordered desalted, except for the P3/P5 primers that were ordered PAGE purified.

eCLIP and meCLIP library preparation

eCLIP procedure is similar to the one developed by Van Nostrand and colleagues with a few modifications notably towards oligonucleotides sequences (cf. Results section). A few kit-based manipulations were also replaced by conventional biology methods, such as ethanol precipitation. eCLIP and meCLIP step-by-step protocols are available in Supplementary Text S1. Briefly, 20 million cells per sample were crosslinked at 150 mJ/cm^2. Sample underwent partial RNase 1 digestion. The soluble fraction was precleared (wild-type eIF4A3) with unconjugated protein A beads before IP or directly immunoprecipitated (eIF4A3-HA and UPF1-FLAG) on pre-coupled corresponding magnetic beads. Two percent of RNase-treated lysate was kept at 4°C to be used as SM-input negative control. RNP complexes were washed stringently with a buffer containing 1M NaCl and 2M Urea. Cross-linked RNAs were subsequently 5′ and 3′ dephosphorylated, followed by 3′ RNA linker ligation as previously reported (15). In meCLIP experiments, a 5′ phosphorylation event was added with T4 PNK to allow the subsequent 5′ ligation of the biotinylated linker. Resulting RNPs and SM-input control were purified by SDS-PAGE and transferred onto a nitrocellulose membrane. Size selection was performed by comparison to a radiolabeled control (5% of the beads can be radiolabeled with γ-^{32}P ATP and T4 PNK) and elution of RNAs was achieved by proteinase K treatment, acid phenol–chloroform extraction and ethanol precipitation. SM-input samples were 5′ and 3′ dephosphorylated. 3′ RNA linker was then ligated and the resulting RNAs, as well as the eCLIP samples, were reverse transcribed. cDNAs were purified by Exo1 treatment to remove unused RT primers and alkaline treatment to remove RNAs. A second 3′ ligation step was then performed in conditions optimized by Van Nostrand *et al.* Ligation products were then purified with Agencourt AMPure XP beads modified with a cutoff set at 50-mer (28). Final quantities of the libraries were estimated using qPCR and samples with close Cp were multiplexed prior to final PCR amplification. PCR product were size-selected (175–300 bp) by PAGE and eluted by diffusion. Samples were then precipitated and submitted to single-end sequencing on a NextSeq 500 sequencer (Illumina) in two separate runs (t1 and t2).

Reverse transcription assays

As indicated, SuperScript-IV (Invitrogen) was used on extracted RNAs according to the manufacturer's protocol. Following denaturation, reverse transcription was performed at a high temperature (55°C) to decrease RNA secondary structures. SuperScript-III (Invitrogen) was used in a similar manner, but at a lower temperature (42°C, then 50°C) since this enzyme is less thermostable than SSIV. AffinityScript (Agilent Technologies) was used at 55°C, as previously described (15). TGIRT-III (InGex) was used at 60°C, as previously described (16). Detailed protocols are available in the Supplementary Text S1.

Read pre-processing and mapping

We performed de-multiplexing of raw reads using a custom script that identifies sample 5′ end barcodes (four nucleotides within a 9 bp randomer). We applied PCR duplicate removal on the de-multiplexed data, and once again after merging the reads from the same sample originating from different lanes; reads with the exact same sequence, including the 9 bp randomer were considered as PCR duplicates. After barcode trimming, we used cutadapt (version 1.10) to trim the 13 bp 5′ end linker (CAGTCCGACGATC) of read-through reads and simultaneously separate them from truncated (untrimmed) reads. Finally, we trimmed the 3′-Illumina adaptor and poor-quality bases with trimmomatic (v.0.36), discarding reads that were <20 bp long after trimming (options ILLUMINACLIP:/path/to/Trimmomatic-0.36/adapters/TruSeq2-PE.fa:2:30:1 SLIDINGWINDOW:5:25 LEADING:25 TRAILING:25 MINLEN:20). To sort the reads into the categories of short and long cDNA fragments, we used cutadapt after PCR duplicate removal with the following options:

```
cutadapt -a AGATCGGAAGAGCGGTTCAGCA
 GGAATGCCGAGACCGATCTCGTATGCCG
 TCTTCTGCTTG –m 20 –untrimmed-output =
 my_sample_long_fragments.fastq my_sample.fastq >
 my_sample_short_fragments.fastq
```

Untrimmed reads correspond to long fragments that were not fully sequenced and thus lack the Illumina 3′ end adapter. All sorted reads were mapped separately against the reference genome.

For genome visualization, datasets were mapped to the human genome (hg38, Ensembl 85, with processed transcripts and pseudo genes masked), using STAR (version 2.5.1b) with the following parameters:

```
STAR –readFilesIn raw_reads.fastq.bz2\
–outFileNamePrefix/path/to/output/mapped_reads.\
–readFilesCommand bunzip2-c\
–outReadsUnmapped Fastx\
–genomeDir/path/to/genome/index/–sjdbOverhang 100–
 sjdbGTFfile/path/to/annotation/hg38.gtf\
–outFilterType BySJout –alignSJoverhangMin 8–align
 SJDBoverhangMin 1\
–outFilterMatchNminOverLread 0.4–outFilterScore Mi-
 nOverLread 0.4–outFilterMultimap Nmax 20\
–outFilterMismatchNmax 999–outFilterMismatch
 NoverLmax 0.06–alignIntronMin 20\
–alignIntronMax 1000000–outSAMattributes All\
–outSAMtype BAM SortedByCoordinate\
–outWigStrand Stranded –quantMode GeneCounts
```

For the meta-exon profiles, we downloaded spliced transcript sequences from Ensembl (hg38, Ensemble 85), and mapped the reads using bowtie2 (version 2.3.2) with its default parameters. We used spliced transcriptome sequences as reference rather than the genome sequence because we systematically obtained a depletion of reads in the 6nt region upstream of the exon junction. In both cases, we mapped reads to one representative transcript per gene, selecting the isoform with the maximum number of exons, using the longest exonic size as a tiebreaker. To compute the number of uniquely mapped reads, we used the number of uniquely mapped reads reported in the STAR final log file. We computed the read-through percentage of each meCLIP library as the number of uniquely mapped read-through reads over the sum of uniquely mapped truncated and read-through reads.

Peak detection and intersection

We used the CTK suite (Shah *et al.*) to detect cross-linking induced truncation sites (CITS) (https://zhanglab.c2b2.columbia.edu/index.php/CTK_Documentation) using the following commands:

```
perl /path/to/ctk/parseAlignment.pl -v –map-qual 255 \
–min-len 18 –mutation-file mutations.txt - \
parsedReads.bed
perl /path/to/ctk/getMutationType.pl
-t del mutations.txt \
parsedReads.deletions.bed
perl /path/to/ctk/CITS.pl -big -gap 10 -p 0.001 \
parsedReads.bed \
```

parsedReads.deletions.bed \
cits.output.bed

The CITS detection tool finds significant truncation sites when the number of truncations per position are compared to a shuffled read-start distribution. It is therefore suitable for single-nucleotide resolution binding site assignment in iCLIP and eCLIP experiments. Next, we computed 2-fold SM-input enrichment of the CITS reported by CTK and of the same set of CITS with shuffled genomic coordinates. For downstream analyses, we selected CITS with a 2-fold enrichment higher than the 95th percentile of the shuffled distribution (which corresponds to a P-value <0.05). Multitest correction was not used due to a major loss of power (high number of false negatives).

Prior to CITS intersection, we increased the CITS region with bedtools slop (version 2.27.1), using the option -b 5. We carried out all read and truncated read CITS intersection with BedTools' intersect (version 2.27.1) with options –c –s –a truncated.cits.bed –b all.reads.cits.bed to obtain the number of common peaks; to retrieve the peaks only found in either dataset, options –v and –s were used. We used matplotlib-venn (version 0.11.5) to plot the Venn diagrams. The percentage of common CITS corresponds to the Jaccard index multiplied by 100; the percentage of CITS found only on either all reads or truncated reads was computed by dividing the respective number of peaks by the total number of peaks detected on both datasets.

regplot function from the Seaborn python library (version 0.8.0) was used to plot both the percentage of read-through reads and the number of uniquely mapped reads against the fraction of peaks detected in the all read datasets.

Peak correlation scatterplots

To assess the correlation between replicate peaks, we followed the ENCODE procedure to find significantly SM-input enriched peaks (Ref. Yeo). We calculated SM-input fold-enrichment of all CITS detected by CTK and applied False Discovery Rate as multiple test correction. Peaks with fold-enrichment higher or equal to 8 and P-value under 10^{-5} were considered significantly enriched.

Next, we obtained the CITS in common between replicates by intersecting all detected CITS (independently of their SM-input enrichment), using the same parameters as the intersection described above. We plotted the fold-enrichment value of each replicate for the common peaks, coloring the fraction of peaks that were significantly enriched in the first replicate. We computed the squared Pearson's correlation coefficient (R^2) using the SciPy python library (version 0.19.1), both on fold-enrichment values of all CITS and of significantly enriched CITS of the first replicate.

Distribution of 5′ ends relative to the exon junction (meta-exon plot)

We used BedTools intersect (version 2.27.1) to intersect uniquely mapped reads to Ensembl85 exon annotations of the hg38 assembly of the human genome, which was generated using the header of the transcript sequences downloaded from Ensembl; the genomic coordinates of exons were converted into transcript coordinates to be consistent with the mapping output from bowtie2. We only considered reads mapped to protein coding genes, mapped to exons longer than 30 bp, and whose 5′ end mapped inside the boundaries of the exon; the distance of the 5′ end of each read was plotted to either the start or the end of the exon, correcting the exon distribution and library size by dividing the counts at each relative position by the number of exons covered at that position and the total number of mapped reads. For genome visualization, BAM files (STAR aligner output) were converted to bedGraph files using BedTools genomecov function. To find 'canonical exons', we automated the retrieval of individual exon coordinates by selecting exons with a high proportion of 5′ ends inside the approximate canonical binding region (between 29 and 19 nucleotides upstream of the exon junction); similarly, we identified exons with no read-through signal by selecting exons which intersect with truncated reads but do not intersect with read-through reads.

RESULTS

Immunoprecipitation strategy for CLIP

CLIP efficiency suffers from the caveats inherent to IP such as epitope accessibility, affinity and antibody specificity. This is particularly critical when using large-scale proteomic or transcriptomic approaches to characterize protein complexes. Indeed, the depth of such strategies determines signal discrimination form noise. Unfortunately, suitable commercial antibodies for IP are not always available, especially for newly discovered RBPs. Furthermore, dedicated antibody production is long and uncertain. As an alternative option, exogenous proteins fused to well-characterized affinity tags can be used. However, expression conditions of recombinant proteins (strong synthetic promoters, optimized codons, high gene copy number/cell ratio) may generate artefactual interactions that differ from the endogenous cellular context. Moreover, the competition between endogenous and recombinant RBPs may provoke biases in the outcome.

Recently, Van Nostrand and colleagues addressed some of the aforementioned immunoprecipitation issues by using a version of eCLIP (TAG-eCLIP) (30), in which CRISPR-Cas9 mediated gene editing was used to generate endogenous RBPs fused with affinity tags. However, when gene modifications are heterozygous, only a portion of the protein of interest is concerned by the immunoprecipitation of the affinity tag, thus impacting IP yield and RNA-protein complexes recovery. Here, we used CRISPR/Cas9 (31) (Supplementary Figure S1) to knock-in an affinity tag and a selection cassette to select positive insertions (Figure 2A, Supplementary Figures S2 and S3). To reduce the number of clones resulting from random plasmid integration and increase the yield of homozygous insertions, a selection marker was inserted at the C-terminal region of the RBP loci as an independent open reading frame driven by an internal ribosomal entry sequence (IRES) to avoid its fusion to the tagged protein. We successfully obtained homozygous HeLa cell lines making eIF4A3 and UPF1 proteins

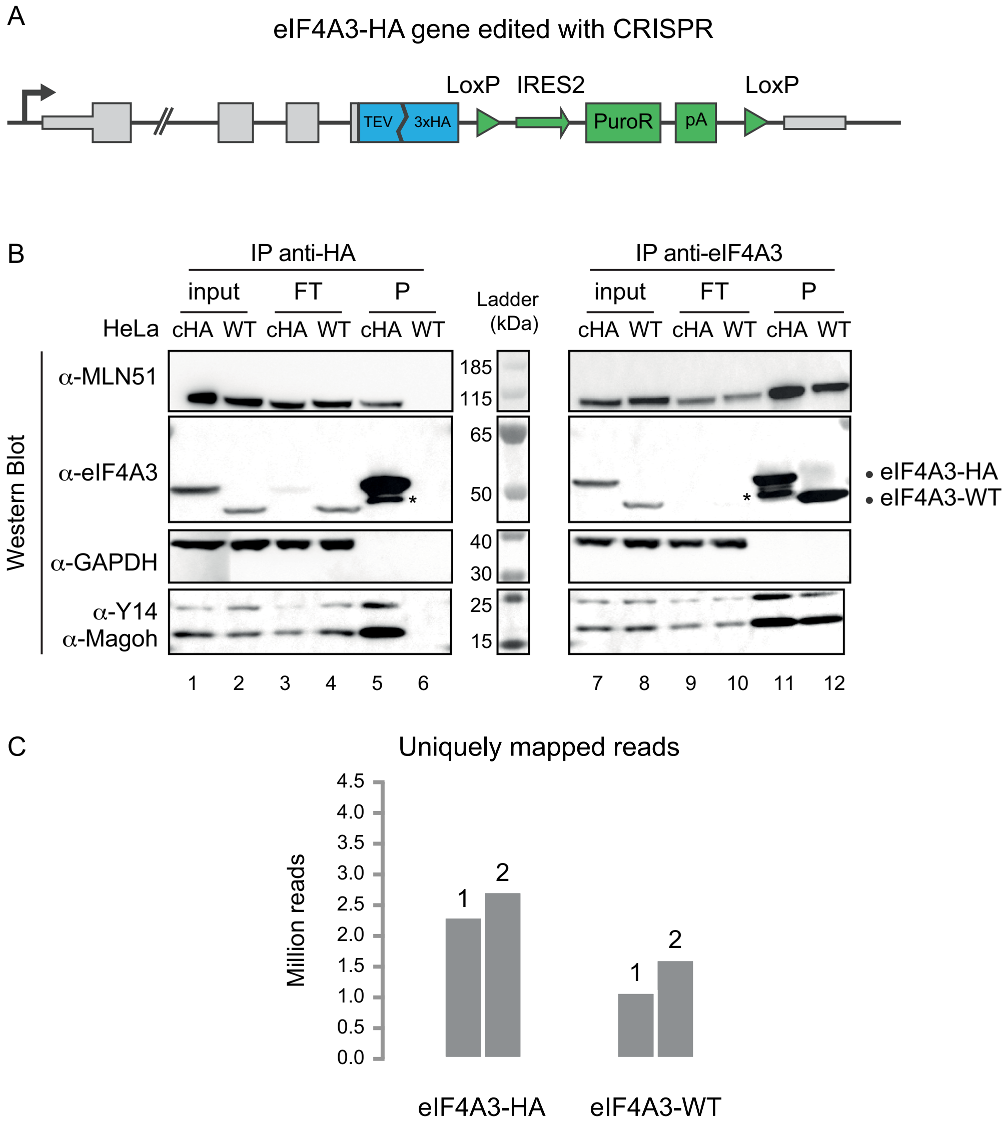

Figure 2. CRISPR/Cas9 editing of eIF4A3. (**A**) Schematic representation of the edited eIF4A3 gene. C-terminal insertion harbors a TEV proteolytic cleavage site and a 3xHA affinity tag, fused to a Internal Ribosomal Entry Site (IRES2)-controlled puromycin (PuroR) selection cassette encompassed by LoxP recombination sites. pA: poly A signal. (**B**) Lysates from wild type (WT) or eIF4A3-edited (cHA) Hela cells were immunoprecipitated with anti-HA or anti-eIF4A3 antibodies and probed for EJC core subunits by Western Blot. FT: flow-through; P: precipitate. The star indicates a C-terminal truncated form of eIF4A3-HA. **C.** Number of uniquely mapped reads from eIF4A3 eCLIP libraries obtained either with endogenous (eIF4A3-WT) or with anti-HA (eIF4A3-HA).

fused to either 3xHA or 3xFLAG affinity tags. Sequencing of the corresponding genomic regions showed that both gene alleles were correctly edited (Supplementary Figures S2 and S3). eIF4A3 expression levels were compared in both WT and edited HeLa cells using anti-eIF4A3 or anti-HA antibodies. Single bands showed that both *eIF4A3* alleles had been successfully modified and the HA tag was confirmed not to affect protein expression levels (Figure 2B, lanes 1, 2, 7, 8). Moreover, immunoprecipitation with anti-HA or anti-eIF4A3 antibodies confirmed that both forms of the protein co-precipitated as efficiently their core EJC partners MAGOH, Y14 and MLN51, demonstrating that editing neither altered eIF4A3 expression, nor its incorporation into EJCs.

We next employed the eCLIP pipeline (15) to compare eCLIP efficiency performed with anti-eIF4A3 and anti-HA antibodies. After sequencing, read pre-processing and mapping against the human genome, we found that a higher number of uniquely mapped reads was obtained using the anti-HA antibody for the immunoprecipitation step (Figure 2C). Thus, the high affinity anti-HA antibody against CRISPR-tagged eIF4A3 improves eCLIP library preparation efficiency.

Sorting out truncated cDNA reads using 'monitored eCLIP' or meCLIP.

A bottleneck of current CLIP procedures resides in their inability to determine the proportion and variability of reverse transcriptase that read through cross-linked nucleotides, and the consequences on the accuracy of RBP binding site assignment. To alleviate this hurdle, we modified the standard eCLIP pipeline, to establish 'monitored eCLIP' or meCLIP. The major modification of meCLIP compared to eCLIP consists in ligating an oligonucleotide containing a barcode at the 5′ end of RNA fragments before RT (Figure 3). The 5′ linker is reverse transcribed if, and only if, the RT manages to pass the RNA-peptide cross-linking site. It is then possible to quantify the ratio between the numbers of reads that harbor the 5′ linker and those that do not. However, the ligation yield significantly varies across experiments (10,32) and unligated RNAs may significantly bias such an approach. To circumvent this obstacle, the 5′ linker was biotinylated. Purification of biotinylated RNA fragments eliminates unligated RNAs before reverse transcription. After sequencing, reverse transcription termination events are then easily monitored by detecting the biotinylated linker sequence at the beginning of the read (Figure 3). The steps of eCLIP and meCLIP methods shown in Figure 3 are described in detail in the Methods section. In addition to the 5′ linker ligation, we replaced all adaptors for single-end sequencing compatibility. Since both the cross-linking site and the biotinylated linker are at the 5′-end of the read, pair-end sequencing is not necessary. Additionally, random and multiplexing barcodes were placed on the 3′ DNA linker, avoiding the use of costly RNA multiplexing linkers. In summary, with the biotinylated 5′ RNA linker enables it is possible to distinguish truncated reads from read-through reads.

meCLIP reveals a highly variable proportion of read-through reads

In order to estimate the impact of the extra steps added to the eCLIP protocol onto library preparation efficiency, we first performed four eCLIP and meCLIP experiments in parallel using anti-HA to target eIF4A3-HA. Quantitation of cDNA libraries by RT-qPCR revealed that meCLIP preparation is on average only 3.5 times less efficient than eCLIP. Moreover, we verified that fragments that are not ligated to the biotinylated primer are not retained on streptavidin beads, which indicated the high specificity of biotinylated fragment purification. Then, we performed meCLIP using anti-eIF4A3 antibodies, anti-HA (to target eIF4A3-HA) or anti-FLAG (to target UPF1-FLAG) using *SuperScript IV* RTase. We observed that the quantity of read-through reads reached up to one-fourth of the total number of reads (Figure 4A). In addition, there were great variations in read-through proportions between replicates or between targeted RBPs. It is important to note that read-through read percentage reflects the ability of a reverse transcriptase to bypass a cross-linked nucleotide. As this ability may differ from one RTase to another, we repeated eIF4A3-HA meCLIP with three additional RTases: *AffinityScript*, *SuperScript III* and *TGIRT III*. These enzymes have different biological origins and have also been employed for various CLIP experiments (14–16). Each reaction was carried out at the optimal conditions of each enzyme. In addition, to test the variability of this feature and to simulate laboratory-to-laboratory variations, meCLIP experiments

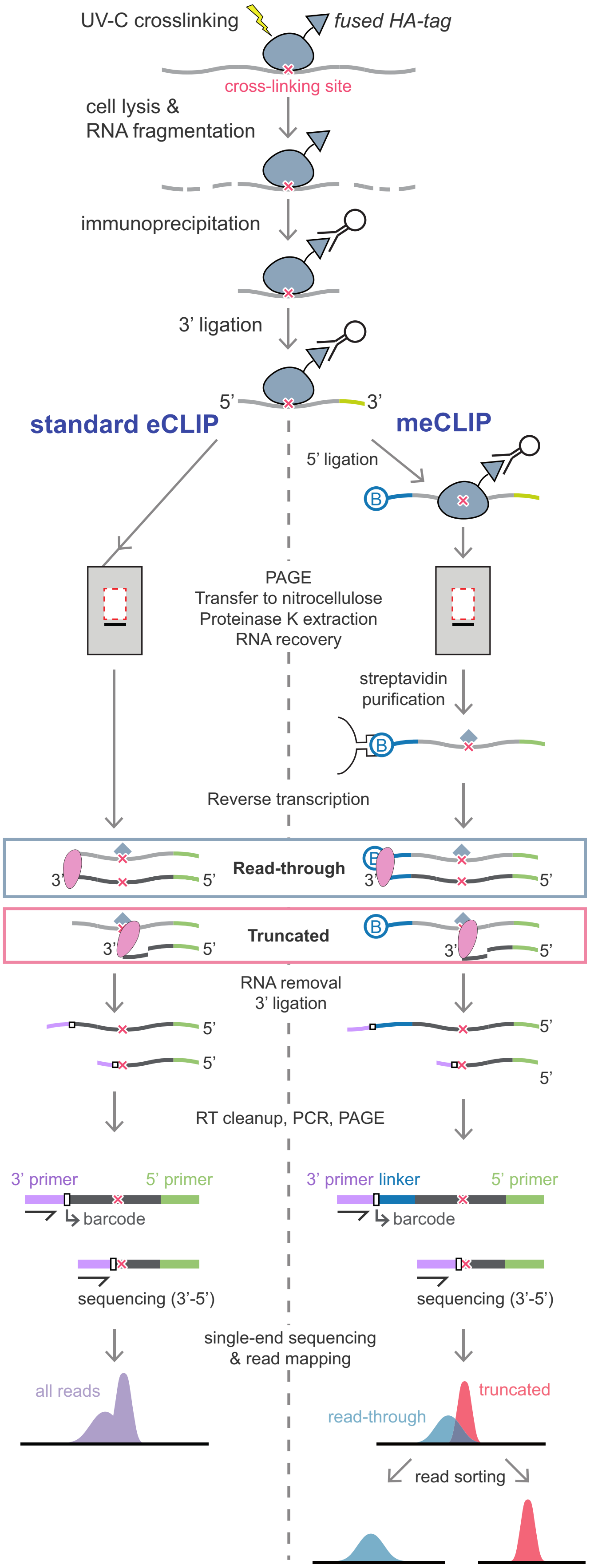

Figure 3. Detailed comparison of the eCLIP and meCLIP protocols. Presentation of the different steps involved in eCLIP and meCLIP procedures.

were done in duplicate by two different experimenters and in independent sequencing runs. For the various eIF4A3-HA meCLIP libraries, the percentage of read-through reads was highly variable (from 2 to 24%) and depended on the RTase used (Figure 4B). The lowest proportion of read-through reads was observed with *AffinityScript* while the three other enzymes generated more variable and greater proportion of read-through reads (at least 8–25%). qPCR quantification of cDNA obtained after RT showed variable efficiency measures among different enzymes (Supplementary Figure S4A). However, different cDNA yield did not correlate directly with PCR duplicate rate in sequenced libraries (Supplementary Figure 4B, C). These results illustrate how the meCLIP protocol can sort out the highly variable and significant amount of read-through reads from the final dataset independently of experimental conditions.

meCLIP removes noise from binding site detection

Detection of significant peaks among mapped reads is a critical step in CLIP-seq downstream analysis to identify potential RBP binding sites. To assess the impact of read-through reads on peak detection, we used the cross-link induced truncation site (CITS) detection software from the CTK suite (29) on meCLIP datasets corresponding either to all reads (equivalent to an eCLIP dataset), or to truncated reads only. As detailed in the Methods section, we selected CITS with a significant SM-input enrichment ($P < 0.05$). The comparison is illustrated first by a few examples that show how peaks are distributed on some annotated exons (Figure 5). On exon 47 of the *LAMA1* gene, two of the four CITS (underlined by a black line) are detected on the all read signal. When applying CITS detection on truncated reads only, these truncation sites are no longer detected, and the right-most CITS is shifted upstream toward a stronger truncation signal. (Figure 5A). The binding site detected in all reads on exon 7 of *FBLX6* corresponds mainly to read-through read signal and thus is not likely to be a truncation site (Figure 5B). In contrast, the binding site on exon 11 of *INPPL1* corresponds exclusively to truncated read signal, and most likely corresponds to a cross-linked RNA region (Figure 5C). These examples illustrate that, within a given eCLIP experiment, the proportion of read-through versus truncated reads is highly variable and unpredictable from one transcriptome region to another (Figure 5A–C). CITS detected in the whole eCLIP dataset but absent in the truncated reads dataset clearly constitute incorrect cross-linking sites.

To determine the impact of read-through reads on CITS detection at the genome scale, we intersected the sets of CITS detected on all reads and on truncated reads separately. In the case of *SuperScript III*, replicate 1, most peaks are common (∼82%), a small proportion (0.34%) are detected only in the truncated read dataset, and 17% of the peaks are detected only in all reads (Figure 5D). Since the only difference between all read and truncated read data sets is the presence of read-through reads, these CITS were considered as originating from the read-through signal. By comparing eIF4A3 datasets obtained with different RTases, as well as the UPF1 replicates, we found a direct relationship between the percentage of 'read-through CITS' and the proportion of read-through reads in each library (Figure 5E), but not with the total number of mapped reads (Figure 5F); we did not observe this relationship when comparing with the CITS detected exclusively on truncated reads (Supplementary Figure S5). Taken together, these results show that read-through reads generate imprecise binding sites that can be sorted out with the meCLIP procedure.

Next, we followed the ENCODE eCLIP pipeline (15) to assess the reproducibility of meCLIP peaks detected on all reads and on truncated reads (Supplementary Figure S6). Using this approach, comparison of meCLIP datasets corresponding to the unrelated proteins eIF4A3 and UPF1 showed very poor correlation coefficients. In contrast, comparison of meCLIP replicates showed higher correlation coefficients, varying between 0.19 and 0.58. Overall, removing read-through reads had little to no effect on replicate reproducibility, as shown when comparing truncated reads to all reads.

meCLIP improves precision of cross-linking site positioning

We next investigated the influence of read-through reads on the localization of the binding sites of a given protein. We used the data of eIF4A3 which has been shown to bind to a precise position upstream of the exon junction (23) and not the data of UPF1 that has been shown to be widely dispersed over mRNA 3′UTR regions (33,34). We first compared the transcriptome-wide distribution of 5′ ends using the addition of truncated and read-through datasets and positioned them relative to the exon junction. A sharp peak was observed centered on the 27th nucleotide upstream the exon junction (Figure 6A). Most of the reads around this position belong to the truncated read category. In contrast, the 5′ ends of read-through reads are distributed upstream of the 27th nucleotide peak, as expected for reads bypassing the cross-linking site. We next examined the 5′ end position of truncated and read-through reads on individual exons. In agreement with the transcriptome-wide distribution, read-through signals often appear upstream of a cluster of truncated reads 5′ ends (Supplementary Figure S7A and B). Upstream read-through signal varies in positions and intensities from one exon junction to another. They are sometimes absent despite a read-through percentage of over 10% in the *SuperScript III*, replicate 1 library (Supplementary Figure S7C). Additionally, we found examples of CITS far away from the canonical EJC deposition site that may correspond to non-canonical EJC binding sites (27,33) (Supplementary Figure S6D). These examples illustrate how the meCLIP

←

Following UV crosslinking, RNase treatment, and RBP purification, an RNA adaptor (green) is ligated at the 3′ end. For meCLIP, a biotinylated RNA linker (blue) is incorporated at the 5′ end. RNAs are fractionated by electrophoresis and eluted from gels. For meCLIP, biotinylated RNAs are purified on Streptavidin beads using stringent conditions. Reverse transcription (RT) is then performed, which leads to two distinct cDNA populations. One of them bears the 5′ linker if the reverse transcriptase reads through the crosslinked peptide (read-through cDNAs). The other one lacks the 5′ linker due to a stop of RT at the crosslinked peptide. A second adaptor (purple) is ligated at the 3′ end of the cDNAs which are next amplified by PCR and submitted to high-throughput sequencing. For meCLIP, two populations of reads are easily sorted out based on the presence or absence of the biotinylated linker sequence.

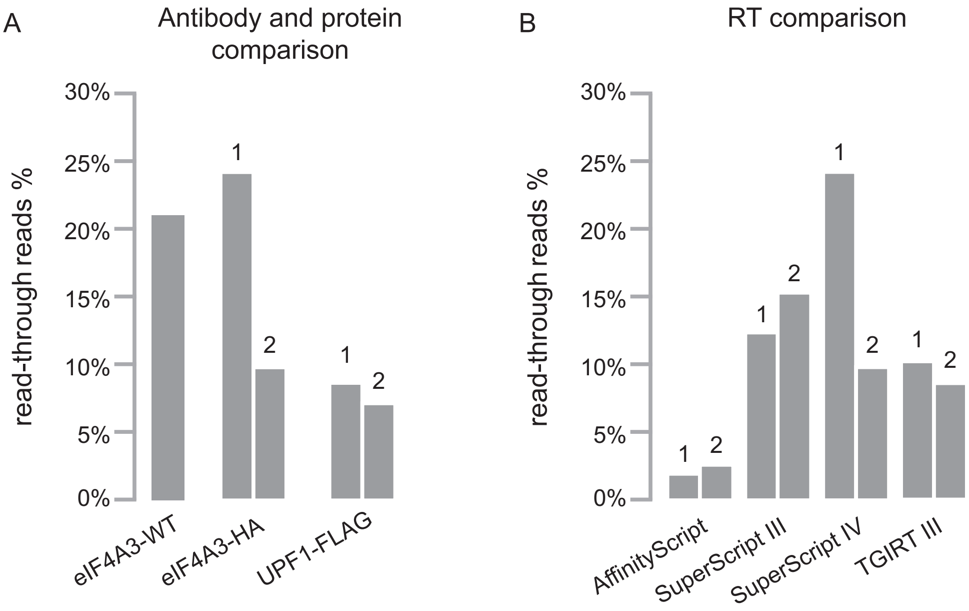

Figure 4. Read-through reads percentage for meCLIP depends on experimental conditions. Libraries generated using (**A**) different cross-linked proteins and (**B**) different reverse transcriptases (RTase) and HA-tagged eIF4A3.

protocol reduces the noise of RBP signals and helps improving the precision of binding site localization.

Read length was recently debated (13,19) as having an impact on the precision of 5′ end of reads obtained from iCLIP, and iCLIP-derived experiments. Hauer *et al.* pointed out that using the center of short fragments instead of the 5′ end of iCLIP reads increased the precision of RBP mapping. To verify this feature, we sorted truncated and read-through reads based on the presence (short fragments) or absence (long fragments) of the Illumina 3′-adapter sequence. The same sharp peak centered on the 27th nucleotide upstream the exon junction was observed for both short and long truncated reads (Figure 6B). However, the signal upstream of the main peak was weaker for short truncated reads compared to long truncated reads. Notably, the signal shifts upstream for both short and long read-through reads. Thus, the use of short truncated reads further increases the precision of binding site assignment.

Another strategy for binding site assignment consists in detecting cross-linking induced mutation sites (CIMS), which stem from errors during reverse transcription. This strategy is used in the case of HITS-CLIP and PAR-CLIP data analysis, where only read-through reads are exploited. We used the short read-through fraction of reads, which is more likely to harbor most CIMS, to quantify the number of mutations in our libraries. We found that deletions occur in <0.5% of short fragments (with the exception of one library that reaches over 1.2%); for libraries obtained with AffinityScript, they are not detectable at all; insertions are consistently negligible (Supplementary Figure S8). Considering that short fragments are about 50% of all uniquely mapped reads, a sensitive CIMS detection would require high sequencing depth, which can be especially costly in the case of multiplexed experiments. In contrast, meCLIP, as well as other iCLIP-derived methods, identifies the crosslinking-induced truncation sites by using directly the 5′-ends of the truncated reads, which on average make up approximately 90% of uniquely mapped reads.

The sharp enrichment of truncated reads observed in the meta-exon plot for the eIF4A3 datasets (Figure 6A) prompted us to compare the distribution of reads relative to the exon junction of four successive CLIP protocols: the original CLIP or HITS-CLIP (27), iCLIP (19), eCLIP and meCLIP. By applying exactly the same data analysis pipeline to all datasets, we observed both a sharpening and an increase of the meCLIP signal at the 27th nucleotide upstream of the exon junction relative to the other protocols (Figure 6C), indicating that a higher proportion of meCLIP truncated reads map to this precise position. Altogether, we demonstrated that the coupling of a nucleotide-resolution CLIP method to the identification of read-through reads and to CRISPR/Cas9-mediated genome editing for affinity tagging, improves the accuracy of the cross-linking site localization.

DISCUSSION

Although extremely informative on transcriptome-wide RBP binding sites, performing a CLIP experiment is a difficult task. Following UV irradiation, a small amount of cross-linked RNA fragments must be isolated by immunoprecipitation from a tremendous excess (possibly $>10^6$ fold) of undesired RNA fragments. Despite this challenge, the analysis of CLIP reads generated by deep sequencing are expected to narrow down the assignment of the RBP binding sites to one nucleotide. In this study, we show that ligation of a biotinylated barcode linker to the 5′ end of RNA fragments markedly improves the CLIP cDNA library preparation to identify and discard misassigned binding sites. Furthermore, genome edition brings a convenient strategy to

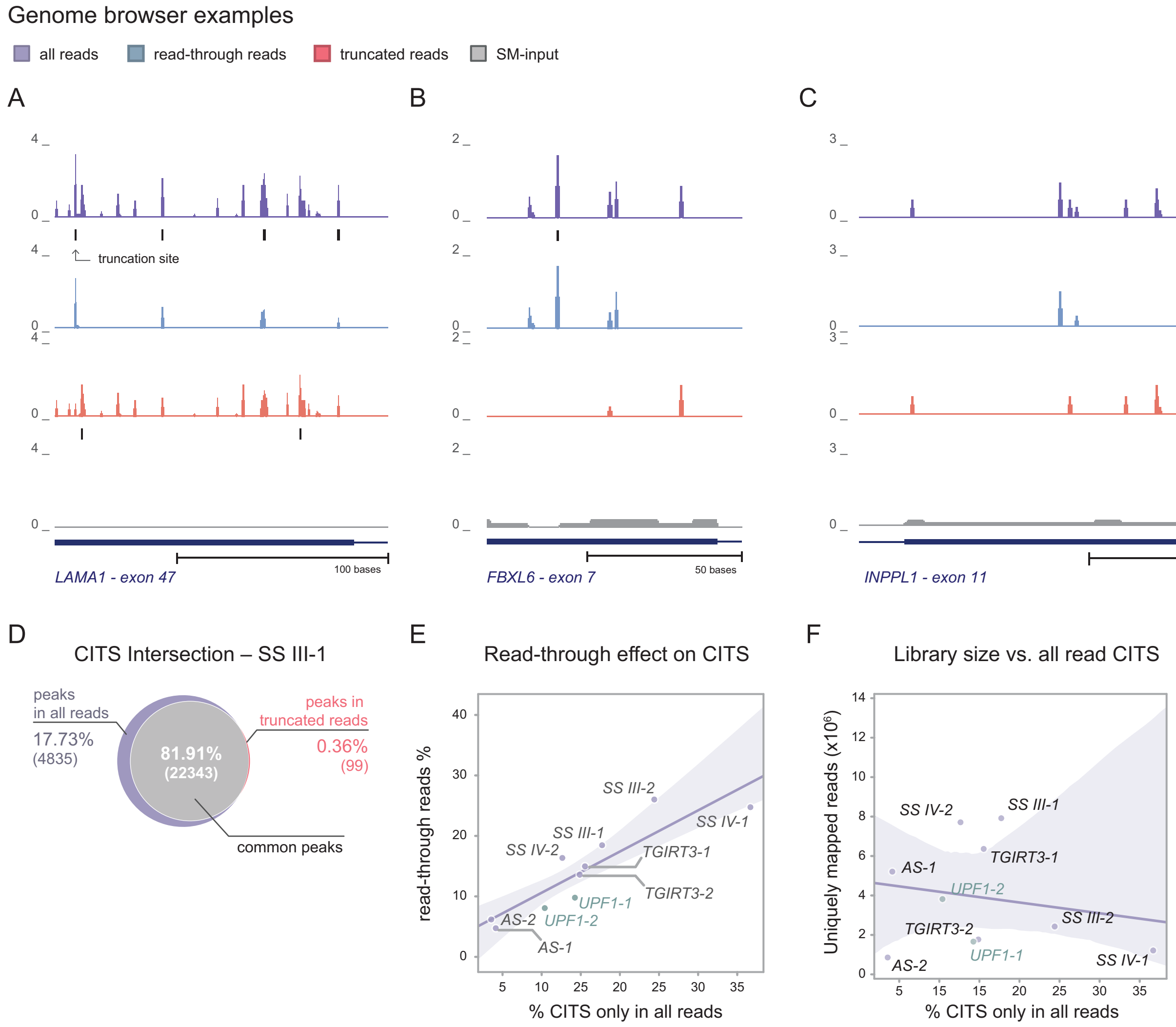

Figure 5. CITS detection is biased by read-through reads. (**A**–**C**) meCLIP reads mapped on three examples of exons. Each black underline corresponds to a CITS detected with CTK. Read coverage is in Reads Per Million (RPM). (**D**) Venn diagram representing the intersection of peaks detected in the unfiltered (all reads) and filtered (truncated reads) data sets from *SuperScript-IV* replicate 1. (**E**) Read-through reads percentages are plotted against the percentage of truncation sites (CITS) detected exclusively in all reads (purple in D). (**F**) Number of uniquely mapped reads versus the percentage of CITS detected exclusively in all reads; the shade around the line indicates the confidence interval (95%) of the linear regression. *AS*: *AffinityScript*, *SSIII*: *SuperScript III*, *SSIV*: *SuperScript IV*.

bypass the need of specific antibodies for immunoprecipitation.

The tagging of endogenous proteins by genome editing has recently proved to be an alternative to antibodies for CLIP analysis (15). Consistently with the work of van Nostrand *et al.*, we successfully obtained cell lines expressing tagged RBPs. In the case of eIF4A3, this incorporated modification did not alter its expression nor its capacity to correctly assemble the EJC. We optimized the genomic modification strategy to obtain a high yield of homozygous clones (Supplementary Figures S1–S3). Homozygosis prevents competition between tagged and untagged versions of the protein within the same cell, while also ensuring that a higher quantity of edited protein is available. The comparison of the expression of native or HA-tagged eIF4A3 in HeLa cells, and the comparison of the CLIP data in the two cell lines (Figure 2) confirmed that the CRISPR-Cas9 mediated fusion of affinity tags to human RBPs constitutes a compelling alternative to specific antibodies (15). Furthermore, this strategy offers the opportunity to target proteins for which no antibodies suitable for CLIP are available, broadening perspectives for the vast number of poorly characterized RBPs. Simultaneous CRISPR–Cas9 protein tagging of several proteins with different tags offers the possibility of co-purifying proteins of interest—a strategy certainly appropriate to shed light on the dynamics of RBPs, which often function in several different RNP complexes.

After obtaining cell lines expressing tagged RBPs, the eCLIP protocol was engineered to distinguish CLIP reads resulting from either reverse transcriptase termination or read-through at the cross-linking site. Until now, the frequency of RT stalling at the peptide–RNA cross-linking site has never been strictly assessed and indirect estimations suggested that it could be variable (12,21). Our strategy al-

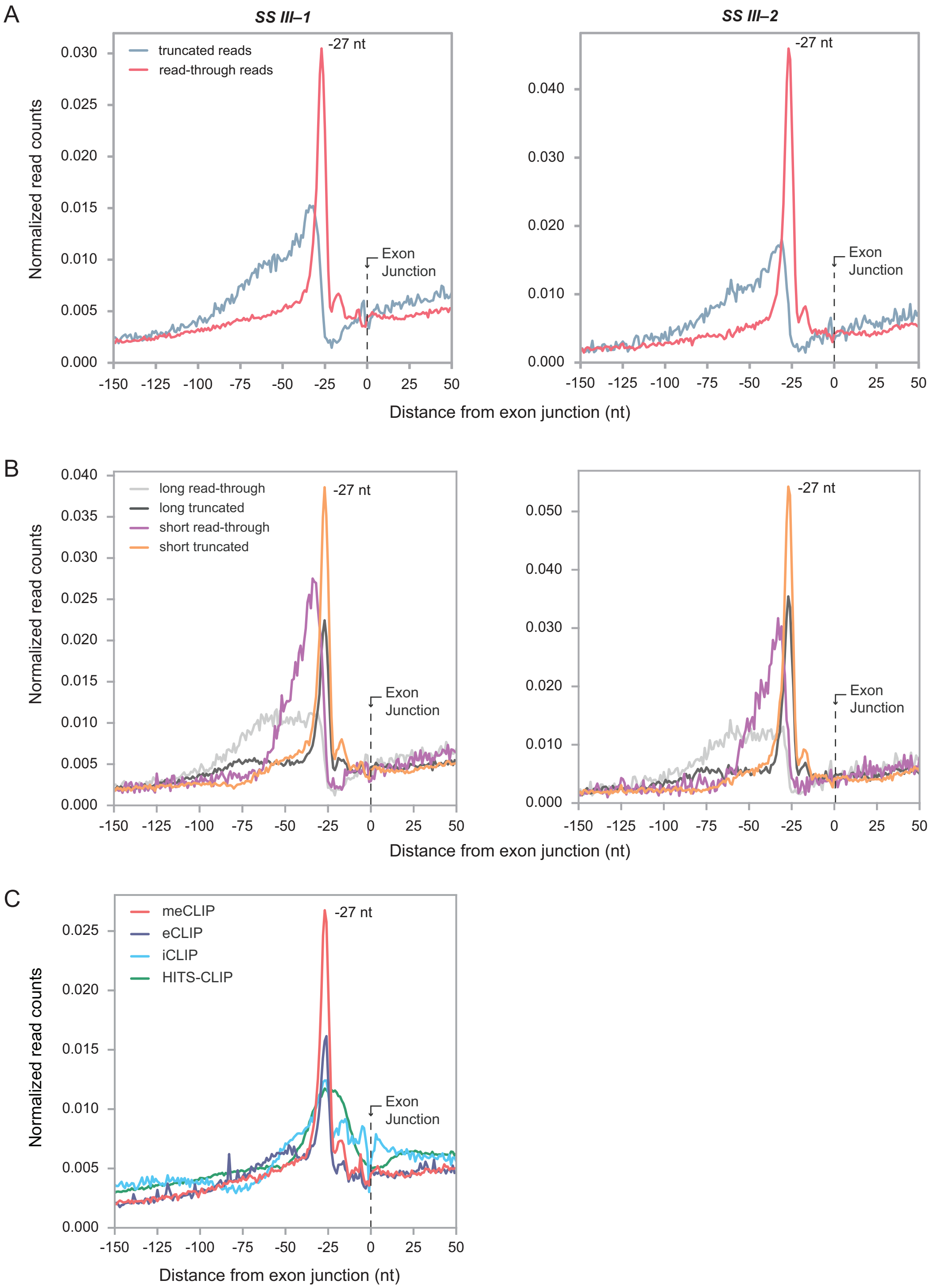

Figure 6. Increased accuracy of cross-linking site positioning. Positioning of 5′ ends of meCLIP reads relative to the exon junction. (**A**) Distribution of eIF4A3-HA meCLIP replicates: truncated reads (red) and read-through reads (blue). (**B**) Distribution of eIF4A3-HA meCLIP replicates: short truncated reads (orange), short read-through reads (purple), long truncated reads (dark grey), and long read-through reads (light gray). (**C**) Distributions of eIF4A3 reads obtained with the meCLIP, eCLIP, iCLIP and HITS-CLIP procedures. meCLIP signal corresponds to truncated reads, and is normalized using the number of uniquely mapped truncated reads.

lowed to precisely measure for the first time the proportion of RT from the truncated ones. The meCLIP data set analysis shows that the percentage of read-through reads reaches up to 24% of all reads and is also highly variable and unpredictable (Figure 5 and Supplementary Figure S6). Among the 10 meCLIP experiments performed for this study, the percentage of read-through reads varied from 2 to 24%. Variations from 7 to 24% exist for identical meCLIP libraries prepared by two different experimenters, or by the same experimenter at different times, which clearly demonstrates that read-through frequency cannot be predicted even for a single RBP. Importantly, we show that meCLIP distinguishes read-through reads from truncated reads despite variable proportions of read-through, different RT experimental conditions, and different efficiency levels among reverse transcriptases. For two different RBPs, meCLIP datasets contained a relatively high percentage of read-through reads. This underlines the importance of using meCLIP to eliminate the unpredictable proportion of spurious reads.

Moreover, we found that mapped read-through reads generate a signal that leads to incorrect RBP binding site assignment, as seen by the direct relationship between read-through CITS and read-through read percentage (Figure 5E). Thus, discarding the more imprecise read-through signal is necessary for precise single-nucleotide binding site assignment. The benefit of read-through reads elimination for RBP mapping was clearly visible in the case of eIF4A3, which is known to be deposited upstream of mRNA splice junctions. Visualization of individual exons shows that meCLIP signal corresponding to read-through reads is in most cases located upstream of signal of truncated reads. This trend is even more visible on the transcriptome-wide distribution of the relative distance of the 5′-end of meCLIP reads to the exon junction. Indeed, read-through read signal is enriched upstream of the canonical position of eIF4A3, which is centered 27 nucleotides upstream of spliced junctions. Although this meta-exon analysis is highly reproducible between replicates (Figure 6), we observed similar correlation coefficients between eCLIP and meCLIP when comparing replicates at the peak level (Supplementary Figure S6). However, the moderate correlation of significantly enriched peaks shows that more efforts are necessary to improve single nucleotide CLIP reproducibility.

Importantly, the proportion of eIF4A3 read-through signal greatly varies from one exon to another revealing that even within the same experiment the frequency of RT stalling is highly variable and unpredictable. As previously noticed in the case of eIF4A3 mapping (13,19), the consideration of short reads further increases the precision of binding site assignment. This is probably due to the fact that the longer the RNA fragments are prior to RT, constraints such as non-specific cross-links or secondary RNA structures are more likely to impair RT and cDNA truncation at the *bona fide* crosslinking sites. However, despite a slight enrichment upstream eIF4A3 canonical position, long truncated reads accumulate in a similar pattern as their short counterparts. As long reads represent a large fraction of total reads, discarding them risks decreasing sequencing depth and negatively affecting downstream analyses. A better option would be to select experimental conditions that maximize the proportion of short fragments (e.g. optimized ribonuclease treatment, size-selection). Finally, we demonstrated the improvement brought by the meCLIP strategy through the comparison of four successive CLIP protocols targeting eIF4A3: the original CLIP or HITS-CLIP (27), iCLIP (19), eCLIP and meCLIP (Figure 6c), under similar experimental conditions (such as cell lines and antibodies) and using the same analysis pipeline. Future meCLIP analyses of eIF4A3 will certainly help to understand the mechanisms that regulate the EJC deposition. More generally, future analyses of meCLIP shall tell whether this method adds a quantitative and localized dimension to the study of RBP dynamics and function.

In summary, our meCLIP method significantly improves the accuracy of RBP binding site mapping by unambiguously filtering out CLIP reads that do pinpoint RBP crosslinking sites and consequently translate into biased peaks. Furthermore, the combination of genome editing to fuse efficient affinity tags to RBP of interest with meCLIP paves the way to elucidate poorly characterized RBP functions and their role in post-transcriptional gene regulation.

DATA AVAILABILITY

meCLIP data sets have been deposited to the Sequence Read Archive (accession number SRP154888).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank lab members and J. Ule for fruitful discussions. We are grateful to E. Bertrand for plasmids and to L. Bastianelli for bioinformatics contribution.
Author contributions: R.H. and H.L.H. designed the genome editing and CLIP experiments. R.H. performed genome editing. R.H. and Q.A. performed CLIP experiments. J.P. did the computational analyses. J.P., A.G., R.H. and H.L.H. designed bioinformatics strategies. H.L.H., R.H., A.G. and J.P. wrote the paper.

FUNDING

Centre National de la Recherche Scientifique; Ecole Normale Supérieure; Agence Nationale de la Recherche (ANR) [2011-BLAN-01801 and ANR-13-BSV8-0023 to H.L.H.]; program ≪ Investissements d'Avenir ≫ launched by the French Government and implemented by ANR [ANR-10-LABX-54 MEMOLIFE and ANR-10-IDEX-0001-02 PSL* Research University]. Funding for open access charge: Agence Nationale de la Recherche.
Conflict of interest statement. None declared.

REFERENCES

1. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
2. Baltz,A.G., Munschauer,M., Schwanhausser,B., Vasile,A., Murakawa,Y., Schueler,M., Youngs,N., Penfold-Brown,D., Drew,K., Milek,M. *et al.* (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. *Mol. Cell*, **46**, 674–690.

3. Castello,A., Fischer,B., Eichelbaum,K., Horos,R., Beckmann,B.M., Strein,C., Davey,N.E., Humphreys,D.T., Preiss,T., Steinmetz,L.M. *et al.* (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, **149**, 1393–1406.
4. Singh,G., Pratt,G., Yeo,G.W. and Moore,M.J. (2015) The clothes make the mRNA: Past and present trends in mRNP fashions. *Annu. Rev. Biochem.*, **84**, 325–354.
5. Castello,A., Fischer,B., Hentze,M.W. and Preiss,T. (2013) RNA-binding proteins in Mendelian disease. *Trends Genet.*, **29**, 318–327.
6. 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.
7. 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.
8. Lee,F.C.Y. and Ule,J. (2018) Advances in CLIP technologies for studies of Protein-RNA interactions. *Mol. Cell*, **69**, 354–369.
9. Darnell,R.B. (2010) HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev. RNA*, **1**, 266–286.
10. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, 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.
11. Urlaub,H., Hartmuth,K. and Luhrmann,R. (2002) A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles. *Methods*, **26**, 170–181.
12. 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.
13. 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.
14. Konig,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.
15. Van Nostrand,E.L., 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.
16. 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.
17. 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.
18. Martin,G. and Zavolan,M. (2016) Redesigning CLIP for efficiency, accuracy and speed. *Nat. Methods*, **13**, 482–483.
19. Haberman,N., Huppertz,I., Attig,J., Konig,J., Wang,Z., Hauer,C., Hentze,M.W., Kulozik,A.E., Le Hir,H., Curk,T. *et al.* (2017) Insights into the design and interpretation of iCLIP experiments. *Genome Biol.*, **18**, 7.
20. Granneman,S., Kudla,G., Petfalski,E. and Tollervey,D. (2009) Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 9613–9618.
21. Sugimoto,Y., Konig,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.
22. Helder,S., Blythe,A.J., Bond,C.S. and Mackay,J.P. (2016) Determinants of affinity and specificity in RNA-binding proteins. *Curr. Opin. Struct. Biol.*, **38**, 83–91.
23. Le Hir,H., Sauliere,J. and Wang,Z. (2016) The exon junction complex as a node of post-transcriptional networks. *Nat. Rev. Mol. Cell Biol.*, **17**, 41–54.
24. Isken,O. and Maquat,L.E. (2008) The multiple lives of NMD factors: balancing roles in gene and genome regulation. *Nat. Rev. Genet.*, **9**, 699–712.
25. Chen,B., Gilbert,L.A., Cimini,B.A., Schnitzbauer,J., Zhang,W., Li,G.W., Park,J., Blackburn,E.H., Weissman,J.S., Qi,L.S. *et al.* (2013) Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell*, **155**, 1479–1491.
26. Gibson,D.G., Young,L., Chuang,R.Y., Venter,J.C., Hutchison,C.A. 3rd and Smith,H.O. (2009) Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods*, **6**, 343–345.
27. Sauliere,J., Murigneux,V., Wang,Z., Marquenet,E., Barbosa,I., Le Tonqueze,O., Audic,Y., Paillard,L., Roest Crollius,H. and Le Hir,H. (2012) CLIP-seq of eIF4AIII reveals transcriptome-wide mapping of the human exon junction complex. *Nat. Struct. Mol. Biol.*, **19**, 1124–1131.
28. Clarke,A.C., Prost,S., Stanton,J.A., White,W.T., Kaplan,M.E., Matisoo-Smith,E.A. and Genographic,C. (2014) From cheek swabs to consensus sequences: an A to Z protocol for high-throughput DNA sequencing of complete human mitochondrial genomes. *BMC Genomics*, **15**, 68.
29. Shah,A., Qian,Y., Weyn-Vanhentenryck,S.M. and Zhang,C. (2017) CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics*, **33**, 566–567.
30. Van Nostrand,E.L., Gelboin-Burkhart,C., Wang,R., Pratt,G.A., Blue,S.M. and Yeo,G.W. (2017) CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. *Methods*, **118–119**, 50–59.
31. Sander,J.D. and Joung,J.K. (2014) CRISPR-Cas systems for editing, regulating and targeting genomes. *Nat. Biotechnol.*, **32**, 347–355.
32. Zhang,Z., Lee,J.E., Riemondy,K., Anderson,E.M. and Yi,R. (2013) High-efficiency RNA cloning enables accurate quantification of miRNA expression by deep sequencing. *Genome Biol.*, **14**, R109.
33. Hurt,J.A., Robertson,A.D. and Burge,C.B. (2013) Global analyses of UPF1 binding and function reveal expanded scope of nonsense-mediated mRNA decay. *Genome Res.*, **23**, 1636–1650.
34. Zund,D., Gruber,A.R., Zavolan,M. and Muhlemann,O. (2013) Translation-dependent displacement of UPF1 from coding sequences causes its enrichment in 3' UTRs. *Nat. Struct. Mol. Biol.*, **20**, 936–943.