TECHNICAL REPORTS

nature structural & molecular biology

iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution

Julian König[1,6], Kathi Zarnack[2,6], Gregor Rot[3], Tomaž Curk[3], Melis Kayikci[1], Blaž Zupan[3], Daniel J Turner[4], Nicholas M Luscombe[2,5] & Jernej Ule[1]

In the nucleus of eukaryotic cells, nascent transcripts are associated with heterogeneous nuclear ribonucleoprotein (hnRNP) particles that are nucleated by hnRNP C. Despite their abundance, however, it remained unclear whether these particles control pre-mRNA processing. Here, we developed individual-nucleotide resolution UV cross-linking and immunoprecipitation (iCLIP) to study the role of hnRNP C in splicing regulation. iCLIP data show that hnRNP C recognizes uridine tracts with a defined long-range spacing consistent with hnRNP particle organization. hnRNP particles assemble on both introns and exons but remain generally excluded from splice sites. Integration of transcriptome-wide iCLIP data and alternative splicing profiles into an 'RNA map' indicates how the positioning of hnRNP particles determines their effect on the inclusion of alternative exons. The ability of high-resolution iCLIP data to provide insights into the mechanism of this regulation holds promise for studies of other higher-order ribonucleoprotein complexes.

A major source of proteomic diversity in multicellular eukaryotes is the production of multiple mRNA isoforms. In humans, it was recently estimated that 95–100% of all multi-exon transcripts undergo alternative splicing[1]. Splice-site selection is primarily mediated by RNA-binding proteins that bind regulatory elements within nascent transcripts[2,3]. Heterogeneous nuclear ribonucleoprotein C1/C2 (hnRNP C) was identified over 30 years ago as a core component of hnRNP particles that form on all nascent transcripts[4]. However, although hnRNP C is one of the most abundant proteins in the nucleus, its role in splicing regulation remained unresolved. Whereas some studies suggested that hnRNP particles generally facilitate splicing[5,6], individual hnRNP proteins were thought to function as splicing silencers[7,8]. Resolving these seemingly contradictory observations was hindered by the inability to locate precisely hnRNP particles on nascent transcripts *in vivo*. In particular, genome-wide mapping of hnRNP C positioning would provide critical information on how hnRNP particles control splicing. Because these highly abundant particles are likely to constitute a general platform for other splicing regulators, deciphering their function would greatly advance our understanding of splicing regulation.

UV cross-linking and immunoprecipitation (CLIP) combined with high-throughput sequencing was previously used to generate transcriptome-wide binding maps of several RNA-binding proteins[9–12]. However, because identification of binding sites relied on the analysis of overlapping sequence clusters, distances of less than 30 nucleotides were not resolved. An additional disadvantage of CLIP is the requirement of reverse transcription to pass over residual amino acids that remain covalently attached to the RNA at the cross-link site. Primer extension assays have shown that the vast majority of cDNAs prematurely truncate immediately before the 'cross-link nucleotide'[13]. Here, we exploited this apparent limitation to achieve single-nucleotide resolution by capturing these truncated cDNAs through the introduction of a second adaptor after reverse transcription via self circularization (**Fig. 1**). To quantify cDNA molecules that truncate at the same nucleotide, we added a random barcode to the DNA adaptor. This allowed us to discriminate between unique cDNA products and PCR duplicates. We successfully applied individual-nucleotide resolution CLIP (iCLIP) to study hnRNP C–dependent splicing regulation in human cells. Taken together, iCLIP enables precise mapping of protein-RNA interactions in intact cells.

RESULTS

iCLIP maps hnRNP C binding at nucleotide resolution

We used iCLIP to examine the positioning of hnRNP C on pre-mRNAs *in vivo*. We performed three replicate iCLIP experiments using an hnRNP C antibody on human HeLa cell lysates. The purified protein–RNA complex was absent when omitting UV-cross-linking or the use of hnRNP C antibody and was diminished when hnRNP C knockdown cells were used (**Supplementary Fig. 1a**). We reverse-transcribed and PCR-amplified cross-linked RNA, controlling PCR specificity with an experiment that lacked the antibody during purification (**Supplementary Fig. 1b**). High-throughput sequencing using Illumina GA2 generated a total of 6.5 million sequence reads (**Supplementary Table 1**); 4.2 million sequence reads aligned to the human genome by

[1]Medical Research Council Laboratory of Molecular Biology, Cambridge, UK. [2]European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Trust Genome Campus, Cambridge, UK. [3]Faculty of Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia. [4]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, UK. [5]European Molecular Biology Laboratory, Genome Biology Unit, Heidelberg, Germany. [6]These authors contributed equally to this work. Correspondence should be addressed to J.U. (jule@mrc-lmb.cam.ac.uk).

Received 7 December 2009; accepted 22 April 2010; published online 4 July 2010; doi:10.1038/nsmb.1838

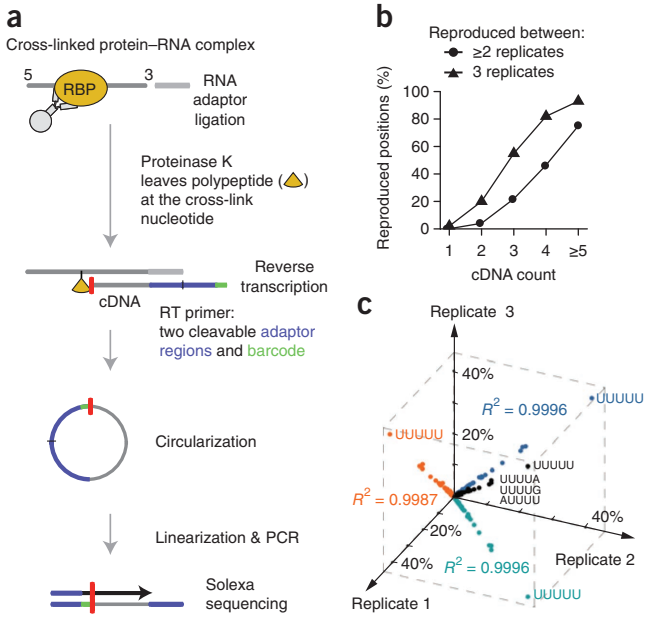

Figure 1 iCLIP identifies hnRNP C cross-link nucleotides on RNAs. (**a**) Schematic representation of the iCLIP protocol. After UV irradiation, the covalently linked RNA is co-immunoprecipitated with the RNA-binding protein (RBP) and ligated to an RNA adaptor at the 3′ end. Proteinase K digestion leaves a covalently bound polypeptide fragment on the RNA that causes premature truncation of reverse transcription (RT) at the cross-link site. Red bar, last nucleotide added during reverse transcription. Resulting cDNA molecules are circularized, linearized, PCR amplified and subjected to high-throughput sequencing. The first nucleotides of each sequence contain the barcode followed by the nucleotide where cDNAs truncated during reverse transcription. (**b**) Reproducibility of cross-link nucleotide positions. Percentage of cross-link nucleotides with a given cDNA count that were identified in at least two (circles) or all three experiments (triangles) are shown. The percentage of reproduced cross-link nucleotides increased with the incidence of hnRNP C cross-linking (cDNA count). (**c**) Reproducibility of sequence composition at cross-link nucleotides. Frequencies of pentanucleotides overlapping with cross-link nucleotides are shown for the three replicate experiments with the sequence shown for the four most highly enriched pentanucleotides. In all three replicate experiments, 42% of cross-link nucleotides overlap with UUUUU.

allowing only single genomic hits and one nucleotide mismatch. Next, we eliminated PCR amplification artifacts by removing sequences that truncated at the same nucleotide in the genome and shared the same random barcode. This identified 641,350 reads in total for the three replicate experiments, each representing a uniquely cross-linked RNA molecule. Finally, we summarized the number of sequences at each cross-link nucleotide into a 'cDNA count', representing a quantitative measure of the amount of hnRNP C cross-linking to each position (**Fig. 2a**). For the analyses of three independent no-antibody control samples, we generated a total of 18 million sequence reads. After the elimination of PCR amplification artifacts, only 1,780 unique cDNAs remained (**Supplementary Table 1**), reflecting the high quality of purification and library preparation steps.

The iCLIP data were of high positional precision. The reproducibility of iCLIP data was demonstrated by the observation that 12,790 cross-link nucleotides were identified in at least two independent experiments (**Fig. 1b**). We observed 75% of cross-link nucleotides with a cDNA count of five or more in all three experiments, showing that the strongest cross-link sites of hnRNP C are the most reproducible (**Fig. 1b**). Furthermore, there was an enrichment of cross-link nucleotides with an offset of one or two nucleotides (**Supplementary Fig. 2**). This observation may arise from protein contacts to more than one nucleotide of the RNA. In addition, the steric hindrance of the peptide fragment remaining on the RNA may cause reverse transcription to terminate more than one nucleotide upstream of the cross-link site. As an independent measure of reproducibility, we compared the occurrence of pentanucleotides overlapping the cross-link nucleotides. We found a high correlation between the three experiments (**Fig. 1c**), underscoring the high precision of iCLIP in capturing protein-RNA interactions.

iCLIP identified large-scale binding of hnRNP C across the whole transcriptome. Although only a few direct targets were known before this study, we found hnRNP C cross-linking to transcripts from 55% of all annotated protein-coding genes (**Fig. 2** and **Supplementary Fig. 3**). This places hnRNP C as a major post-transcriptional regulator of similar importance as, for example, the polypyrimidine tract–binding protein (PTB) that was shown to bind transcripts of 43% of annotated human genes[14]. Among previously described hnRNP C targets, we observed binding to the regulatory element that determines start-codon selection within the *c-myc* mRNA and to the 3′ untranslated region of the *APP* mRNA[15,16] (**Supplementary Fig. 4**). We found that 79% of cDNAs mapped in a sense orientation relative to introns, 9% relative to exons and 1% relative to noncoding RNAs; 11% mapped to intergenic regions, indicating that these harbor previously undescribed transcribed regions. Only 2% mapped in an antisense orientation relative to annotated genes, confirming that iCLIP generates strand-specific information on RNA binding (**Fig. 2d** and **Supplementary Fig. 3**). In summary, our data show that hnRNP C has a central role as a regulator of nascent transcripts.

To reduce false positive hits and to increase the resolution of the data, previous CLIP studies have applied filtering algorithms to identify clusters of CLIP cDNAs. Applying this approach to the hnRNP C dataset, we identified 33,991 clustered cross-link nucleotides (false discovery rate < 0.05)[12]. This filtering removed 94% of all cross-link nucleotides, which most likely included true binding sites. Because the iCLIP libraries prepared during this study are not fully saturated—a limitation that currently applies to all CLIP methods—many real binding sites are represented by only few cDNAs. This view was supported by the observation that 6,367 out of 12,790 reproduced cross-link nucleotides were removed during the filtering process. Therefore, we performed all the analyses described below on the complete and the filtered datasets. The results are quantitatively and qualitatively consistent (**Supplementary Fig. 5**), indicating that both sets are of high quality. To minimize the loss of information, we describe findings for the complete dataset in the remainder of this work.

hnRNP C cross-links to uridine tracts

The high resolution of iCLIP data allowed us to assess the sequence specificity of hnRNP C binding. Strikingly, uridine represented 85% of cross-link nucleotides ($P < 0.001$ by hypergeometric distribution for enrichment relative to background base frequencies; **Fig. 3a** and **Supplementary Fig. 5a**). Surrounding positions were also enriched for uridines, such that 65% of cross-link nucleotides were part of a contiguous tract of four or more uridines (**Fig. 3b** and **Supplementary Fig. 5b**). These results agree with the *in vitro* observation that the RNA recognition motif (RRM) domains of hnRNP C bind to uridine tracts[17–19], suggesting that cross-link nucleotides reflect the positions where the RRM domains contact RNA *in vivo*. In comparison, only 15–24% of cross-link nucleotides from the no-antibody control experiments were located in a tract of four or more uridines, showing a

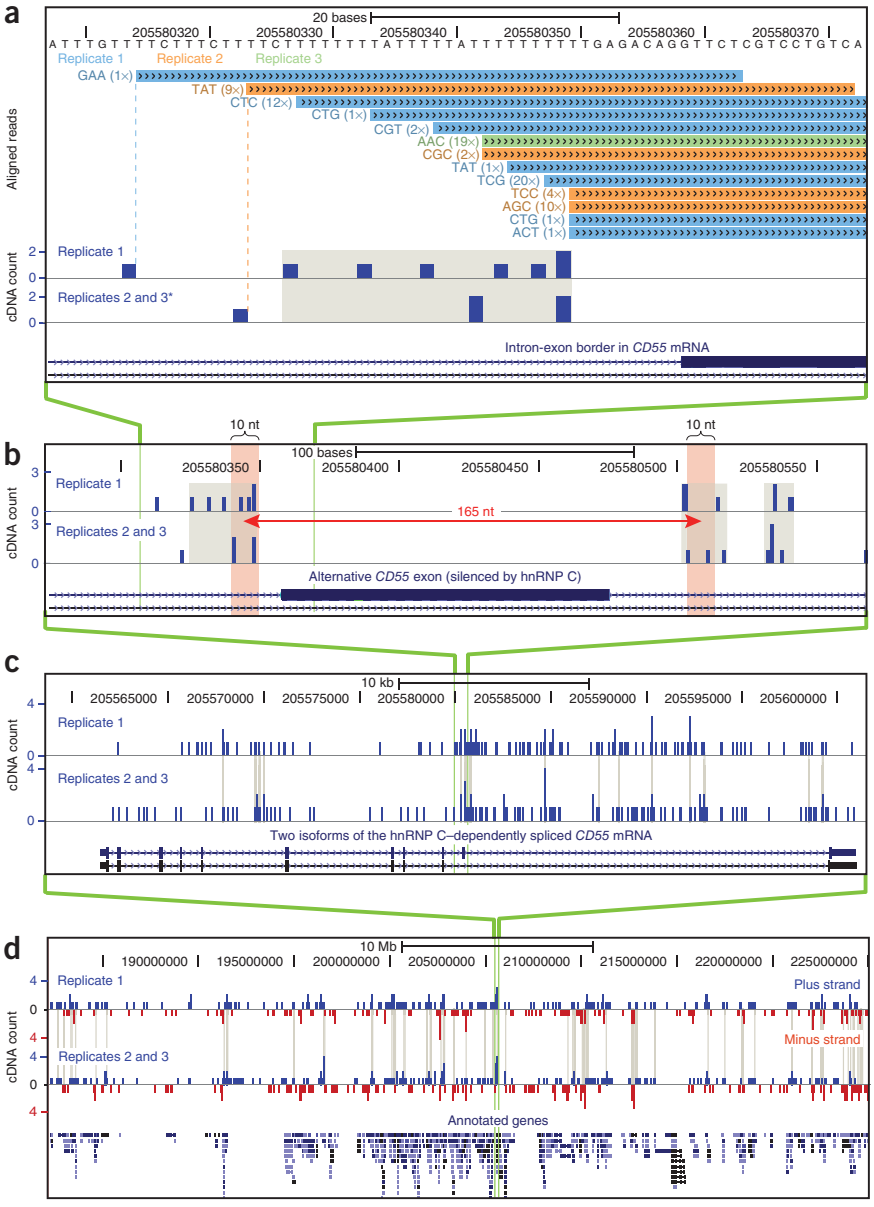

Figure 2 The genomic location of hnRNP C cross-link nucleotides. (**a**) Conversion of mapped iCLIP sequence reads into cDNA count values. Genomic sequence is shown above the color-coded positions of cDNA sequences from replicate experiments, preceded by the associated random barcode and the number of sequenced PCR duplicates (given in brackets). In the lower panel, a cDNA count was assigned to the upstream cross-link nucleotide. Cross-link nucleotides within filtered clusters are highlighted in gray. The position of an alternative exon in *CD55* mRNA is shown at the bottom. Modified image of the UCSC genome browser (human genome, version hg18, chromosome 1, nucleotides 205,580,308–205,580,373). *, due to space limitations, replicates 2 and 3 were merged into one lane. (**b**) Long-range spaced cross-link nucleotides flank the alternative exon in *CD55* pre-mRNA. A distance of 165 nucleotides is marked by a red arrow with red shaded bars on either side representing 10-nt surrounding intervals. (**c**) Cross-link nucleotides are present along the entire length of *CD55* pre-mRNA and accumulate around the alternative exon. Clustered cross-link nucleotides are indicated with gray lines. Annotation below shows the position of exons in two alternative transcripts. (**d**) Global view of cross-link nucleotides on chromosome 11 (nucleotides 182,200,000–225,000,000). cDNA counts corresponding to positions in plus- and minus-strand transcripts are shown in blue and red, respectively. Gene annotations are given below. Cross-linking to individual genes and strand specificity are reproduced between replicates.

significant enrichment of uridine tract binding in the hnRNP C iCLIP data ($P < 0.01$ by Student's *t*-test). We note that the control shows a bias to bind uridine tracts compared with the expected 5% from the background distribution in transcribed regions. However, this is in line with previous studies on single-stranded DNA–binding proteins that show preferential cross-linking to thymidine residues[20,21]. Nonetheless, the small number of sequence reads and the low cross-linking bias in the control data contrast with the strong preference for uridine by hnRNP C, indicating that the vast majority of iCLIP sequence reads reflect real hnRNP C binding events. Furthermore, the ability of iCLIP to quantify the number of cDNAs mapping to each cross-link nucleotide allowed us to analyze the affinity of hnRNP C to uridine tracts of different lengths. We found that cDNA counts increased with the number of uridines in the tract, suggesting that hnRNP C binds longer tracts with higher affinity (**Fig. 3b** and **Supplementary Figs. 5b** and **6a**).

The spacing of cross-link sites reflects hnRNP particle formation

iCLIP allowed us to resolve adjacent binding sites within uridine tracts. We found that, regardless of the length of the uridine tract, hnRNP C primarily cross-linked to the third uridine from the 3′ end (**Fig. 3c** and **Supplementary Figs. 5c** and **6b**). In addition, we identified a second peak of hnRNP C cross-linking positioned five or six nucleotides upstream on tracts longer than nine uridines. Consistently, such dual binding also occurred on shorter tracts when flanked by neighboring uridine tracts (**Fig. 3d** and **Supplementary Fig. 5d**). Because the hnRNP C tetramer binds RNA with two RRM domains positioned proximally to each other[6,22], the dual cross-linking pattern could result from adjacent binding by the two RRM domains. These results show that the high resolution of iCLIP can elucidate combinatorial binding by multiple RNA-binding domains to proximal RNA binding sites, which would otherwise remain unresolved.

In addition to the short-range spacing within uridine tracts, iCLIP also identified a pattern of long-range spacing of cross-link nucleotides. We found peaks at distances of 165 and 300 nucleotides (**Fig. 3e** and **Supplementary Fig. 5e**). Strikingly, the uridine density also peaked at the same positions (**Fig. 3e** and **Supplementary Fig. 5e**). The defined spacing between cross-link nucleotides suggests that the intervening RNA is incorporated into the hnRNP particles. This model agrees with the organization of hnRNP particles as proposed by previous studies[6,23,24]. Taken together, the precise mapping of hnRNP C cross-link sites provides insights into the structure of hnRNP particles.