9260–9271 *Nucleic Acids Research, 2017, Vol. 45, No. 16* *Published online 9 August 2017*
doi: 10.1093/nar/gkx646

CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome

Zijun Zhang[1] **and Yi Xing**[1,2,*]

[1]Bioinformatics Interdepartmental Graduate Program, University of California, Los Angeles, Los Angeles, CA 90095, USA and [2]Department of Microbiology, Immunology & Molecular Genetics, University of California, Los Angeles, Los Angeles, CA 90095, USA

Received January 02, 2017; Revised June 22, 2017; Editorial Decision July 13, 2017; Accepted July 19, 2017

ABSTRACT

Crosslinking or RNA immunoprecipitation followed by sequencing (CLIP-seq or RIP-seq) allows transcriptome-wide discovery of RNA regulatory sites. As CLIP-seq/RIP-seq reads are short, existing computational tools focus on uniquely mapped reads, while reads mapped to multiple loci are discarded. We present CLAM (CLIP-seq Analysis of Multi-mapped reads). CLAM uses an expectation–maximization algorithm to assign multi-mapped reads and calls peaks combining uniquely and multi-mapped reads. To demonstrate the utility of CLAM, we applied it to a wide range of public CLIP-seq/RIP-seq datasets involving numerous splicing factors, microRNAs and m^{6}A RNA methylation. CLAM recovered a large number of novel RNA regulatory sites inaccessible by uniquely mapped reads. The functional significance of these sites was demonstrated by consensus motif patterns and association with alternative splicing (splicing factors), transcript abundance (AGO2) and mRNA half-life (m^{6}A). CLAM provides a useful tool to discover novel protein–RNA interactions and RNA modification sites from CLIP-seq and RIP-seq data, and reveals the significant contribution of repetitive elements to the RNA regulatory landscape of the human transcriptome.

INTRODUCTION

Mammalian genomes encode over a thousand RNA-binding proteins (RBPs) that play important roles in RNA processing and metabolism (1,2). RBPs interact with their cognate sequences and/or structural elements within the RNA to impact diverse aspects of post-transcriptional regulation, including splicing, polyadenylation, transport, stability and translational control, as well as RNA base modifications (3). For example, many RBPs function as splicing factors through interactions with cis splicing regulatory elements within the pre-mRNA (4). In recent years, there have been intense efforts to identify and characterize RBPs using high-throughput methods. For example, technologies such as SELEX-seq (5), RNAcompete (1) and RNA Bind-n-Seq (6) have been developed to define the *in vitro* binding motifs of numerous RBPs.

A powerful strategy for transcriptome-wide mapping of RBP–RNA interactions and RNA regulatory elements is immunoprecipitation followed by high-throughput sequencing (7). Two popular approaches are CLIP-seq (crosslinking with immunoprecipitation followed by sequencing) (8–10) and RIP-seq (RNA immunoprecipitation followed by sequencing) (11). The standard protocol of CLIP-seq involves crosslinking protein–RNA interactions by UV, immunoprecipitating the RBP–RNA complexes by antibody, then sequencing cDNA library to generate short reads typically ranging between 35 and 50 bp. Three versions of CLIP-seq (HITS-CLIP, PAR-CLIP and iCLIP) deliver datasets with distinct features due to their technical differences and biases (12). RIP-seq experiments are performed in similar procedures, but RIP-seq does not include the UV-crosslinking step, resulting in reduced resolution of the binding sites and lower signal-to-noise ratios (13). Besides detecting RBP–RNA interaction sites, RIP-seq and CLIP-seq have also been utilized for detecting RNA base modifications, in particular N^6-methyladenosine (m^{6}A) (14), revealing the prevalence and dynamic landscape of reversible RNA base modifications in the human transcriptome (14,15).

Despite the increasing popularity and widespread use of CLIP-seq and RIP-seq for mapping RBP–RNA interaction and RNA modification sites, existing computational approaches for analyzing these data still have important limitations. As CLIP-seq and RIP-seq reads are short (usually <50 bp), in a conventional data analysis workflow, reads are mapped to the genome and transcriptome, uniquely mapped reads are retained and RBP binding sites are iden-

[*]To whom correspondence should be addressed. Tel: +1 310 825 6806; Fax: +1 310 206 3663; Email: yxing@ucla.edu

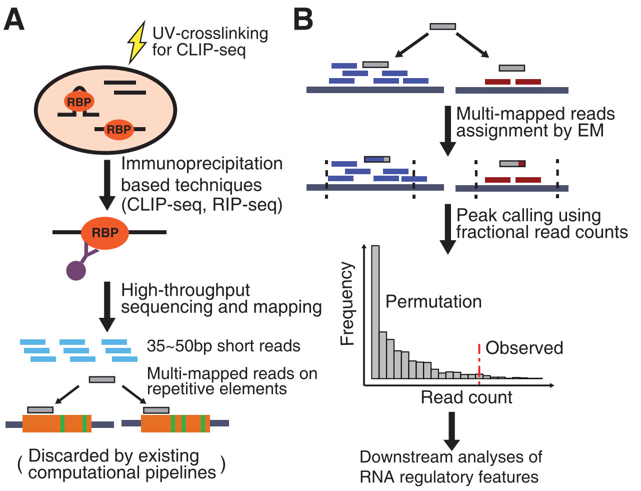
A
UV-crosslinking for CLIP-seq
RBP
RBP
Immunoprecipitation based techniques (CLIP-seq, RIP-seq)
RBP
High-throughput sequencing and mapping
35~50bp short reads
Multi-mapped reads on repetitive elements
(Discarded by existing computational pipelines)
B
Multi-mapped reads assignment by EM
Peak calling using fractional read counts
Frequency
Permutation
Observed
Read count
Downstream analyses of RNA regulatory features

Figure 1. Motivation and schematic overview of CLAM. (**A**) In immunoprecipitation-based techniques for analyzing RBP–RNA interactions (CLIP-seq, RIP-seq), RNA associated with the target RBP is subject to fragmentation after the RBP–RNA complex is immunoprecipitated by specific antibody, followed by high-throughput sequencing to generate short reads typically ranging between 35 and 50 bp. An appreciable fraction of reads, such as those originated from repetitive element derived RBP–RNA interaction sites, are mapped to multiple genomic regions and subsequently discarded by conventional data analysis pipelines. Shown here is a read mapped to two genomic copies of a repetitive element (orange boxes), which have identical sequences where the read is aligned but have mutations elsewhere between these two copies (green vertical lines). (**B**) CLAM identifies a set of genomic regions sharing multi-mapped reads. It then uses an expectation–maximization (EM) algorithm to rescue multi-mapped reads and assign them to specific genomic regions, followed by a permutation based procedure for peak calling with gene-specific FDR control. The rescued peaks are then assessed via downstream analyses of RNA regulatory features, including enrichment of consensus motifs and evaluations of RBP-specific regulatory features.

tified by appropriate statistical models for peak calling (12) (Figure 1A). However, by restricting the analysis to uniquely mapped reads and removing reads mapped to multiple genomic loci, a potentially large catalog of regulatory sites residing in duplicated and repetitive regions of the transcriptome will be under-detected or inaccessible. Given that approximately half of the human genome is comprised of transposable elements (16), and a variety of RBPs such as heterogeneous nuclear ribonucleoprotein C (hnRNPC), adenosine deaminase, RNA specific (ADAR1) and staufen double-stranded RNA binding protein 1 (STAU1) have binding sites derived from highly repetitive transposable elements (17–20), the restriction to uniquely mapped reads represents a significant source of false negatives in site identification from CLIP-seq and RIP-seq datasets.

Here, we present CLAM (CLIP-seq Analysis of Multi-mapped reads), a new computational method for CLIP/RIP-seq data analysis and peak calling utilizing multi-mapped reads. We applied CLAM to published CLIP-seq data of 18 RBPs, as well as RIP-seq data of the m^6A RNA modifications. In all datasets, CLAM recovered a large number of novel RNA regulatory sites inaccessible by conventional analyses of uniquely mapped reads. We further demonstrated the physical and functional relevance of the identified CLAM sites based on consensus motif patterns as well as correlation with relevant RNA regulatory features. Altogether, CLAM provides a useful and widely applicable computational tool to discover novel functional protein–RNA interaction sites and RNA modification events from CLIP-seq and RIP-seq data, and reveals the significant contribution of repetitive elements to the RNA regulatory landscape of the human transcriptome.

MATERIALS AND METHODS

CLIP-seq/RIP-seq read pre-processing and mapping

A typical CLIP-seq library contains 3′ adaptors due to the short length of RBP-protected fragments; and 5′ random barcodes to discriminate polymerase chain reaction (PCR) duplicates. The 3′ adaptors were first removed by fastx_clipper from fastx toolbox, available at http://hannonlab.cshl.edu/fastx_toolkit/. Low quality reads were discarded by requiring the minimum quality threshold of 30 and at least 50% of bases in a read above this quality threshold. Next, PCR duplicates were removed by collapsing the reads with the same random barcodes and identical sequences. After removal of PCR duplicates, barcodes were removed and the reads were aligned by Novoalign (available at http://www.novocraft.com/) to the human genome and transcriptome, using the hg19 version of the human genome as the genomic index and Gencode V19 (http://www.gencodegenes.org/releases/19.html) as the transcriptome annotations (21). The set of optimized Novoalign parameters for CLIP-seq data (22) was used. Specifically, the alignment cost score '–t 85' controls the mismatches as: two substitutions, two consecutive deletions or one substitution plus one deletion. The option '-l 25' requires at least 25 high-quality matches. For multi-mapped reads, reads that map to $<$100 genomic loci were retained for downstream analyses.

All mapped reads (uniquely + multi-mapped) were then merged into genomic regions. Two reads were merged if the distance between them was smaller than a threshold d. By default we set $d = 50$ for CLIP-seq and $d = 100$ for RIP-seq to match the size of RBP footprint or RNA fragment.

Expectation–maximization analysis of multi-mapped reads

Distinct genomic regions were connected through multi-mapped reads as a graph. The connected subgraphs (i.e. regions sharing multi-mapped reads) were extracted and subsequently converted to a compatibility matrix Y representing the mapping relationships between reads and genomic regions. Each genomic region corresponded to a column and each read corresponded to a row of the compatibility matrix Y. For read i uniquely mapped to genomic region k, $y_{i,\cdot} = 0$ except for $y_{i,k} = 1$. For read i mapped to multiple genomic regions $\{k_p, \ldots, k_q\}$, $y_{i,k} = 1$, for $k \in \{k_p, \ldots, k_q\}$ and 0 otherwise. Our goal was to resolve the rows with multiple 1's in the matrix Y using an EM framework (23).

In other words, our goal is to infer another indicator matrix Z to represent the true origins of mapped reads. As certain RBPs (e.g. Argonaute 2 or AGO2) could have long footprints on mRNA transcripts due to multiple overlapping binding sites, the statistical model of CLAM considers

Table 1. Three representative datasets analyzed by CLAM

Dataset	Predominant binding region	Motif	Technology	Cell line	Accession ID
hnRNPC	intronic	poly-U	iCLIP	HeLa	E-MTAB-1371
AGO2	3′-UTR	microRNA seeds	iCLIP	LCL	GSE50676
m^{6}A	3′-UTR	RRACU	RIP	H1-ESC	GSE52600

that for a potential binding site, the probability that a multi-mapped read originates from this region depends on the reads mapped to a defined local window surrounding the binding site. Hence, given the vector $\hat{\Theta}$ representing the relative abundance of multiple mapped genomic regions among RBP-bound RNAs and the compatibility matrix Y, the latent variable $\hat{z}_{i,k}$ that represents the true origin of read i from region k is computed by taking the expectation at (t+1)-th iteration as the E-step:

$$\hat{z}_{i,k}^{(t+1)} = E[z_{i,k}|\,Y,\hat{\Theta}^{(t)},c]$$
$$= \Pr(z_{i,k}=1|\,Y,\hat{\Theta}^{(t)},c) = \frac{y_{i,k}\cdot\hat{\theta}^{(t)}_{k,c_{i,k}}}{\sum_k y_{i,k}\cdot\hat{\theta}^{(t)}_{k,c_{i,k}}}$$

where $c_{i,k}$ is the center position of read i on region k, $\hat{\theta}^{(t)}_{k,c_{i,k}}$ is the relative abundance of multiple mapped genomic regions estimated at the locus $c_{i,k}$ on region k in the previous iteration. From the starting condition $t = 0$, the EM model converges to the optimal point regardless of its initial values, since the objective function to be maximized is concave (https://arxiv.org/abs/1104.3889). For simplicity, $\hat{\Theta}$ was initialized uniformly for all regions.

Next in the (t+1)-th iteration of the M-step, for any particular column y_k in Y corresponding to a specific genomic region, we estimate its relative abundance $\hat{\theta}^{(t+1)}_{k,j}$ locally at each position j among multiple mapped regions using the true origin $\hat{Z}^{(t+1)}$ within the (2w+1) window:

$$\hat{\theta}^{(t+1)}_{k,j} = \frac{\sum_i \hat{z}^{(t+1)}_{i,k}\cdot \mathbf{1}(j-w\le c_{i,k}\le j+w)}{N}$$

where N is the total number of reads (uniquely mapped and multi-mapped) in these regions sharing multi-mapped reads, w is the window size defining the local window, $c_{i,k}$ is the center position for read i on region k, $\hat{z}^{(t+1)}_{i,k}$ is the estimated true origin of read i from region k and $\mathbf{1}(\cdot)$ is the indicator function. By default we set $w = 50$ for CLIP-seq data and $w = 100$ for RIP-seq data to match the size of RBP footprint or RNA fragment.

The E-step and M-step are iterated until convergence.

Peak calling

Peak calling was performed on a gene-by-gene basis, in order to control for the expression variability among genes as in previous work (19,24,25). Briefly, CLAM was applied to genes with multi-mapped reads. For a given gene, the mapped reads could be divided into two sets: uniquely mapped reads with probability of origin $p = 1$, and multi-mapped reads with $p \in [0, 1)$. We used a random permutation procedure to obtain the background read count distribution. Specifically, uniquely mapped reads were randomly assigned a location along the gene for 1000 times. For multi-mapped reads, a uniform random variable $u \in [0, 1]$ was first drawn; if $u \le p$, this multi-mapped read was randomly assigned a location in the same manner as uniquely mapped reads; otherwise, this read was discarded in the current permutation. For position j with height h_j>0, P-value $= \frac{\sum_k 1(k\ge h_j)\cdot n_k}{\sum_k n_k}$, where n_k is the number of positions with peak height $k \in (1, 2, \ldots)$ in permutation derived null distribution, and $\mathbf{1}(\cdot)$ is the indicator function. For each gene, multiple testing was corrected by the Benjamini–Hochberg False Discovery Rate (FDR) procedure (26). Positions with gene-specific FDR < 0.001 were called as significant loci, and peaks were called as the most significant loci within 50 bp windows. If a peak was <50 bp, the peak was extended symmetrically to 50 bp. For downstream analyses, the common or rescued peaks in individual replicates were then merged by taking the union respectively.

Analysis of m^{6}A RIP-seq data

We employed a slightly different processing pipeline as well as parameters for the m^{6}A data, given the differences between RIP-seq (for m^{6}A) and CLIP-seq. We first mapped the human m^{6}A RIP-seq reads using STAR (27) v2.4.2 to the hg19 genome with the Gencode v19 transcript annotations (21), retaining reads mapped to <100 loci. Then we ran CLAM with parameters: maximum distance for collapsing reads $d = 100$; local window size $w = 100$; P-value correction using the more stringent Bonferroni correction given the lower signal-to-noise ratio of RIP-seq, and peaks were called as the most significant loci within 500 bp windows and extended to 100 bp symmetrically.

Analysis of RNA motif and regulatory features

We applied CLAM to publicly available CLIP-seq/RIP-seq datasets listed in Table 1. We analyzed two iCLIP datasets, hnRNPC iCLIP on the HeLa cell line from Zarnack *et al.* (19), and AGO2 iCLIP on the GM12878 lymphoblastoid cell line (LCL) from Wan *et al.* (28). We also analyzed one m^{6}A RIP-seq dataset on the H1 human embryonic stem cell (ESC) line from Batista *et al.* (29). For each dataset, we analyzed RNA motif and regulatory features based on the known properties of the RBP or RNA modification. Annotations of repetitive elements for the hg19 human genome were downloaded from the UCSC RepeatMasker track, available at the UCSC table browser.

Motif finding for hnRNPC peaks was performed using Zagros (30), a specialized *de novo* motif finder for CLIP-seq data. For m^{6}A sites, motif finding was performed using HOMER (31) as in our previous m^{6}A work (29) on the top 1000 peaks ranked by enrichment ratio over input control.

To assess the functional impact of hnRNPC CLAM sites on hnRNPC-dependent alternative splicing, hnRNPC shRNA knockdown followed by RNA-seq dataset in the same HeLa cell line (19) was analyzed by rMATS (32) (version 3.2.5) to detect differential alternative splicing events. The alternative exons were filtered by read counts (inclusion counts + skipping counts $\geq$20) and then ranked by $\Delta\psi$ values (control − knockdown) from the most hnRNPC repressed exons ($\Delta\psi$ =−1) to the most hnRNPC enhanced exons ($\Delta\psi = 1$). Each exon was extended symmetrically by 250 bp on both sides to include the proximal intron regions. We applied a Gene Set Enrichment Analysis (GSEA)-like analysis (33) to test if exons with CLAM sites overlapping with the extended exon regions were enriched toward the top or the bottom of the $\Delta\psi$ ranked hnRNPC-dependent differential alternative splicing events. Specifically, enrichment score (ES) was calculated as described previously (33) on the exons with CLIP-seq peaks as hits in this ranked exon list, and Kolmogorov–Smirnov test (K–S test) was performed to test for statistical significance.

To assess the effect of AGO2 CLAM sites on microRNA-mediated mRNA repression, microarray gene expression data of human cell lines upon ectopic expression or inhibition of two microRNAs were downloaded from GEO with accession number: GSE37213 (miR-21, T lymphocytes) and GSE42823 (miR-107, H4 glioneuronal cells). We selected these two microRNAs because they were both abundantly expressed in the GM12878 cell line profiled by AGO2 iCLIP, based on small RNA-seq profiling data of microRNA abundance in the original study by Wan *et al.* (28). For each AGO2 peak, we predicted the targets of these two microRNAs using TargetScan (34) (http://www.targetscan.org/vert_71/). AGO2 target genes were then separated into two categories based on whether they had common or rescued peaks. Background genes were chosen as genes without any AGO2 peaks. Affymetrix microarray probesets were matched to corresponding transcripts using BiomaRt (35).

To assess the influence of m^{6}A modification on mRNA half-life, we used transcript half-life time measured in iPS cells as in our previous m^{6}A work (36). Genes were classified similarly as in the AGO2 analysis. We performed a meta-gene analysis to obtain the m^{6}A peak distributions in 5′-UTR, coding sequence (CDS) and 3′-UTR by binning the corresponding transcript region into 10 equal-sized bins then counting in each bin the frequency of top 1000 common or rescued m^{6}A peaks respectively.

Analyses of ENCODE CLIP-seq and RNA-seq data on 17 splicing factors

We applied CLAM to 17 splicing factors with matching CLIP-seq (eCLIP) and shRNA knockdown followed by RNA-seq datasets in the HepG2 cell line from the ENCODE project. We followed the ENCODE SOP pipeline to remove adapters. We developed an in-house script for collapsing PCR duplicates based on the ENCODE SOP but preserved multi-mapped reads. Since eCLIP employed paired-end sequencing, only the second mate was extracted and fed into CLAM after mapping, following the same strategy adopted by the ENCODE consortium (37). CLAM was run using the same parameter set as in our analyses of the hnRNPC and AGO2 iCLIP data.

The CLAM sites for each splicing factor were validated in two aspects: enrichment of consensus motif (if available) and enrichment of splicing factor-dependent alternative exons upon shRNA knockdown of the splicing factor. Known consensus motifs of 12 splicing factors were retrieved from the RNAcompete database (1). Motif enrichment analysis was performed using a *Z*-score method as described previously (38), with minor modifications. Specifically, given a motif regular expression and a set of *n* CLIP-seq peaks, we first computed the number of peaks (sequences), denoted by X, containing the motif. Then we estimated the background frequency *p* of the given motif in a large collection of random genomic sequences of the same length as CLIP-seq peaks. The expected motif occurrence in the CLIP-seq peaks was hence $n \cdot p$, with the variance being $n \cdot p \cdot (1 - p)$. We applied the Z-transformation as $Z = \frac{X-np}{\sqrt{np(1-p)}}$. To account for the over-dispersion in the above *Z*-score, we computed the *Z*-scores for an additional $m = 1000$ randomers of the same length as the given motif, and estimated the sample standard deviation *s* of the randomer *Z*-scores. Hence the final *t*-statistic is $t = \frac{Z}{s}$ with degree of freedom *m*-1, and *P*-value was given by Student's *t*-distribution.

RNA-seq data of splicing factor knockdown was publicly available in the ENCODE data portal and we used our rMATS pipeline (32) (version 3.2.5) to quantify the exon inclusion level (ψ) of cassette exon skipping events. We applied a read count filter to remove exon skipping events with <20 combined (inclusion plus skipping) reads. As there were many more common peaks than rescued peaks, to account for the difference in statistical power in calculating the GSEA-like (33) K–S enrichment statistics, for each splicing factor we down-sampled the common peaks to the same number of rescued peaks and repeated the down-sampling procedure 20 times. For each exon, three non-overlapping regions were considered: upstream 250 bp flanking intron, exon body and downstream 250 bp flanking intron. The rescued peaks and each set of down-sampled common peaks were tested for enrichment in each of these three regions separately, based on a ranked list of splicing factor-dependent exons ranked by difference in exon inclusion levels ($\Delta\psi$) between control and knockdown, following the same procedure for calculating the K–S statistic as described above for the hnRNPC iCLIP data.

Code availability

The CLAM software and user manual can be downloaded from https://github.com/Xinglab/CLAM. All datasets used in this paper are publicly available in public repositories, i.e. SRA and ArrayExpress, with accession numbers listed in Table 1.

RESULTS

CLAM statistical model for multi-mapped reads in CLIP-seq and RIP-seq data

To utilize multi-mapped reads in CLIP-seq and RIP-seq data and improve peak calling in highly repetitive regions, we developed CLAM, which assigns multi-mapped reads using an EM framework, followed by peak calling with a permutation-based procedure commonly used for CLIP-seq and RIP-seq data ('Materials and Methods' section). The statistical model of CLAM was inspired by previous work on resolving multi-mapped reads in RNA-seq (23,39,40) and ChIP-seq data (41,42), while features specific for CLIP-seq and RIP-seq data were incorporated in the model. Below we briefly illustrate the CLAM algorithm, using one read mapped to two genomic regions as the example (Figure 1B). CLAM first collapses reads into genomic regions. The two genomic regions in Figure 1B have six and two uniquely mapped reads respectively, while sharing one multi-mapped read which will be resolved by CLAM. As certain RBPs (e.g. AGO2) could have long footprints on mRNA transcripts due to multiple overlapping binding sites (43), we designed the EM algorithm in CLAM to assign a multi-mapped read based on the mapping status of other reads (uniquely + multi-mapped) within a defined local window surrounding the read of interest ('Materials and Methods' section). The algorithm iterates between inferring the expected true origins of multi-mapped reads and deriving the Maximum Likelihood Estimates (MLE) for the probabilities of reads derived from specific regions, until it reaches convergence. In the hypothetical example in Figure 1B, for the multi-mapped read CLAM will assign 0.75 and 0.25 read to the left and right regions respectively to achieve the maximum likelihood. Once multi-mapped reads are reassigned, a permutation test will be performed for peak calling combining uniquely mapped reads and CLAM assignment of multi-mapped reads ('Materials and Methods' section).

To systematically evaluate the behavior of the CLAM EM framework for re-assigning multi-mapped CLIP-seq reads, we generated a benchmark dataset by truncating the hnRNPC iCLIP (19) reads by 10 bp from the 3′ end then remapping the truncated reads to the genome. This strategy enabled us to assess the algorithm performance on 'gold-standard' reads that were uniquely mapped in the full-length dataset but became multi-mapped in the truncated dataset. For comparison, we implemented and evaluated two alternative models: (i) assigning multi-mapped reads uniformly with equal weights for all mapped regions ('uniform' model) and (ii) assigning multi-mapped reads weighted by local counts of uniquely mapped reads, which corresponds to the first iteration of EM ('one-iter' model). Then we assessed the accuracy of re-assigning these reads to the known originating loci (positive loci) over the rest of multi-mapped loci (negative loci) by Area Under Receiver Operating Characteristic curve (AUROC), Area Under Precision-Recall curve (AUPR) and the median/mean weight for positive versus negative loci. As illustrated in Table 2, uniform assignment of multi-mapped reads resulted in the poorest performance. Although the CLAM model and the One-iter model achieved comparable AUROC and AUPR values, detailed analyses indicated that the CLAM model as compared to the One-iter model assigned higher weights to positive loci (0.62 versus 0.50) and lower weights to negative loci (0.02 versus 0.13), demonstrating its superior performance.

In sum, CLAM is a two-stage algorithm that first reassigns the multi-mapped reads using a statistical model (i.e. EM), followed by peak calling using the information of both uniquely mapped reads and multi-mapped reads. Compared to conventional CLIP-seq/RIP-seq peak calling procedures of using only uniquely mapped reads, CLAM can discover a large number of novel sites inaccessible by conventional methods, as demonstrated by our systematic assessments using multiple datasets below. It should also be noted that while EM and permutation test could be slow, we used computational techniques to boost the speed of CLAM. For EM-based probabilistic read assignment, we implemented Binary Indexed Tree (BIT) for faster reading and updating of weights. For permutation-based peak calling, we implemented a multi-threading framework for parallel peak-calling on a gene-by-gene basis. As a result, CLAM has reasonable running time that scales well to the total library size (Supplementary Table S1).

CLAM rescues multi-mapped reads and discovers novel sites in CLIP-seq and RIP-seq data

To assess the utility of CLAM, we first applied it to three published datasets on hnRNPC, AGO2 and m^{6}A (Table 1). We chose these three datasets because they were associated with distinct RNA regulatory processes (alternative splicing, microRNA targeting and m^{6}A methylation respectively), and included both CLIP-seq (hnRNPC, AGO2) and RIP-seq (m^{6}A) data. Each dataset had two biological replicates. After pre-processing and adapter trimming, the average read lengths were 40, 44 and 50 for the three datasets respectively. We then calculated the percentage of multi-mapped reads among all mapped reads. As shown in Figure 2A, ∼10–18% of reads were multi-mapped across the six samples. Using CLAM, we were able to rescue the vast majority (83–92%) of multi-mapped reads, representing a significant gain in read coverage especially in repetitive regions of the transcriptome (see below). A small proportion (∼10%) of multi-mapped reads were not analyzed by CLAM because they did not cluster to genomic regions (i.e. singleton reads with no other reads in vicinity), or were mapped to too many ($\geq$100) regions and therefore discarded (see details in 'Materials and Methods' section).

The rescued multi-mapped reads were significantly enriched in repetitive regions. We obtained the annotation of repetitive elements in the human genome from the UCSC RepeatMasker track then calculated the percentage of uniquely mapped and multi-mapped reads within different classes of repeats as well as non-repeat regions (Supplementary Figure S1). In all three datasets, the percentage of multi-mapped reads was much higher in repetitive regions as compared to non-repeat regions, thus creating a challenge for peak calling within repetitive regions. For example, in the hnRNPC dataset, 37% of reads mapped to antisense Alu elements were multi-mapped, as compared to only 3% for reads mapped to non-repeat regions. Overall, only 8% of multi-mapped reads in the hnRNPC dataset were mapped to non-repeat regions, while 60% of multi-mapped reads were mapped to antisense Alu elements (Supplementary Figure S1), consistent with a previous report on widespread hnRNPC binding within antisense Alu elements (19). When we ranked the repeat family by their total number of multi-mapped reads, Alu, L1 and simple re-

Table 2. Performance of CLAM and two alternative models on a synthetic benchmark dataset

Model	AUROC	AUPR	Positive loci weight (median, mean)	Negative loci weight (median, mean)
CLAM	0.88	0.79	0.62, 0.65	0.02, 0.15
One-iter	0.88	0.78	0.50, 0.54	0.13, 0.20
Uniform	0.75	0.48	0.50, 0.42	0.20, 0.25

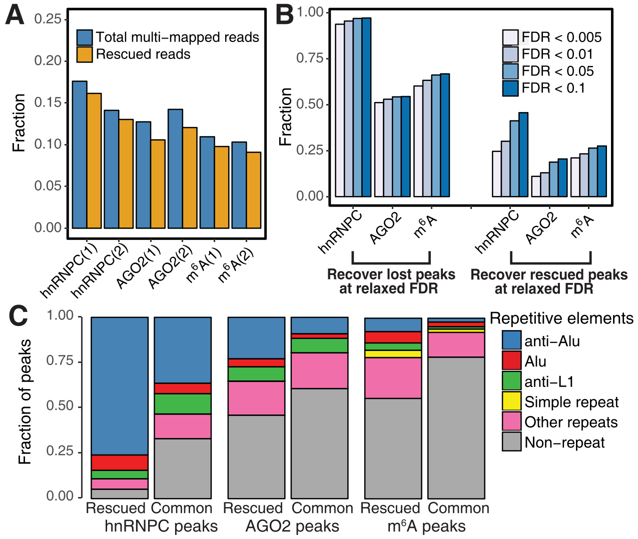

Figure 2. Summary statistics of CLAM results on three CLIP-seq/RIP-seq datasets. (**A**) Percentage of multi-mapped reads (blue) and percentage of multi-mapped reads rescued by CLAM (orange) among all mapped reads in analyzed datasets. (**B**) Sensitivity analysis at various FDR thresholds. The majority of lost peaks can be recovered using the combination of uniquely and multi-mapped reads at higher (more relaxed) FDR thresholds (bar graphs on the left), while a significantly smaller fraction of rescued peaks can be identified using only uniquely mapped reads at higher FDR thresholds (bar graphs on the right). (**C**) Fraction of rescued and common peaks derived from various types of repetitive elements. A significantly higher fraction of rescued peaks are derived from repetitive elements across all three datasets.

peat were consistently among the top families with the highest number of multi-mapped reads across the three datasets (Supplementary Figure S1).

We next assessed CLIP-seq/RIP-seq peak calling by CLAM. We adopted a commonly used permutation procedure for CLIP-seq or RIP-seq peak calling (19,24,25), and defined genomic loci with gene-specific FDR < 0.001 as peaks. We performed peak calling using: (i) uniquely mapped reads only and (ii) uniquely mapped reads plus CLAM assignments of multi-mapped reads. We classified peaks called from the above procedures into three distinct categories: 'common peaks' that were called in both procedures, 'rescued peaks' that were called only with multi-mapped reads incorporated and 'lost peaks' that were called using uniquely mapped reads but not with multi-mapped reads incorporated.

Compared to a naïve read mapping and peak calling procedure using only uniquely mapped reads, a substantial number of rescued peaks were identified from all three datasets by CLAM (Table 3). While a certain number of peaks called by the naïve peak calling procedure were lost in the CLAM results, these lost peaks were much smaller in number as compared to rescued peaks called with incorporating multi-mapped reads (Table 3). For example, in the hnRNPC dataset, we had 26 594 rescued peaks on average in the two samples, as compared to 6898 lost peaks on average. Moreover, the majority of these lost peaks can be recovered from the CLAM results of multi-mapped reads simply by using a relaxed (higher) FDR cutoff (Figure 2B), suggesting that these peaks were lost due to random statistical fluctuations. For example, by relaxing the gene-specific FDR cutoff from 0.001 to 0.005 in the hnRNPC dataset, we were able to recover 94% of lost peaks. The reverse was not true—only 25% of rescued peaks could be identified using only uniquely mapped reads at this higher FDR cutoff, demonstrating the importance of modeling multi-mapped reads in CLAM. We observed the similar trend in the AGO2 and m^{6}A datasets, in which we could recover a much higher percentage of lost peaks by relaxing the FDR cutoff, but much less so on rescued peaks if using only uniquely mapped reads (Figure 2B). We also noted that in the AGO2 and m^{6}A datasets, a number of 'lost peaks' were the only visible peaks in their respective genes when only uniquely mapped reads were considered, but could not pass the gene-specific FDR cutoff when multi-mapped reads in these genes were rescued by CLAM.

As expected, the rescued peaks were strongly enriched in repetitive elements as compared to common peaks across all three datasets (Figure 2C). For example, rescued peaks for hnRNPC were strongly enriched in antisense Alu elements, consistent with previous findings about hnRNPC binding sites within antisense Alu (19). We noted that 76% of rescued peaks for hnRNPC were located in antisense Alu elements, as compared to only 36% for common peaks. Similarly, Alu elements also showed a significant enrichment in rescued peaks for AGO2 and m^{6}A.

Rescued peaks for hnRNPC were associated with alternative splicing

Next, we assessed the functional relevance of rescued CLAM peaks, by correlating these peaks with relevant RNA regulatory features. We first analyzed the rescued CLAM peaks for hnRNPC, a splicing factor known to bind to poly-U tracts within the pre-mRNA to regulate alternative splicing. Using the Zagros *de novo* motif finder (30) for CLIP-seq data, we found a significantly enriched poly-U motif within both common peaks and rescued peaks (Figure 3A), suggesting that the rescued peaks have the same binding properties with hnRNPC as the common peaks. We then evaluated the potential functions of these rescued peaks, by investigating whether they were in the vicinity of alternative exons regulated by hnRNPC. To identify hnRNPC-dependent exons, we re-analyzed the RNA-seq data of hnRNPC knockdown in the same cell type (19) using rMATS (32), and ranked all exon-skipping cas-

Table 3. Summary of CLAM peak calling on the hnRNPC, AGO2 and m^6A datasets

Dataset	Replicate	Rescued	Common	Lost
hnRNPC	#1	24 976	99 890	6027
	#2	28 211	133 708	7769
AGO2	#1	2169	32 494	546
	#2	2243	29 774	536
m^6A	#1	3598	36 000	1790
	#2	3702	39 153	2151

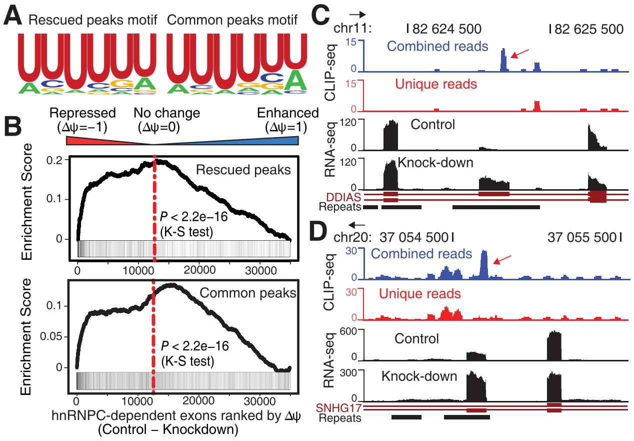

Figure 3. Functional evaluation of CLAM on the hnRNPC CLIP-seq data. (**A**) Identification of the known consensus hnRNPC motif by *de novo* motif discovery in rescued and common hnRNPC peaks. (**B**) Enrichment analysis of hnRNPC-dependent alternative exons for rescued and common hnRNPC peaks. *X*-axis represents alternative exons ranked by rMATS $\Delta\psi$ values (the difference in exon inclusion levels between control and knockdown). *Y*-axis is the enrichment score (ES) calculated via the Kolmogorov–Smirnov statistic. Both rescued and common hnRNPC peaks are strongly enriched for hnRNPC-repressed alternative exons. (**C**) Example of a rescued hnRNPC peak in *DDIAS*. (**D**) Example of a rescued hnRNPC peak in *SNHG17*.

sette exons with sufficient RNA-seq coverage for differential splicing by their rMATS $\Delta\psi$ values (i.e. the difference of exon inclusion level between hnRNPC control and knockdown; see 'Materials and Methods' section). We defined an alternative exon as being associated with a CLIP-seq peak, if the peak was located within the exon body or in intronic regions within 250 bp of the exon. We hypothesized that if rescued CLAM peaks indeed represented functional protein–RNA interaction sites, we would observe an enrichment of exons associated with rescued peaks among hnRNPC-dependent alternative exons identified by RNA-seq. Specifically, as hnRNPC is known to repress exon inclusion (19), its direct target exons should have higher splicing levels upon hnRNPC knockdown. To test this hypothesis, we performed a Kolmogorov–Smirnov statistical test similar to the GSEA algorithm (33), by comparing the rankings of exons with or without hnRNPC peaks within the $\Delta\psi$ ranked list of hnRNPC-dependent exons. Indeed, exons with rescued peaks were strongly enriched toward the left side ($\Delta\psi < 0$) of the list (P-value $< 2.2\text{e-}16$, Figure 3B, top panel), with the enrichment score (ES) peaked around RNA-seq $\Delta\psi$ of 0 then decreased gradually. We observed an almost identical trend for exons associated with common peaks (Figure 3B, bottom panel). Two representative examples of hnRNPC-dependent exons associated with rescued peaks were shown in Figure 3C and D. In Figure 3C (*DDIAS*), RNA-seq data revealed an exon with significantly elevated splicing upon hnRNPC knockdown, but no peak was observed in the vicinity of this exon using uniquely mapped CLIP-seq reads. However, this exon had a number of multi-mapped reads. These reads mapped to distinct sets of other genomic loci, while all of them mapped to this *DDIAS* exon. Therefore, CLAM rescued and assigned these multi-mapped reads to this exon, resulting in the identification of a strong hnRNPC peak. Another example was provided for *SNHG17*, in which CLAM discovered a strong hnRNPC peak within an hnRNPC-dependent alternative exon, while the coverage by uniquely mapped CLIP-seq reads was low and no peak can be identified within the exon (Figure 3D). Of note, in both genes the rescued peaks were located within a primate-specific Alu retrotransposon, indicating the creation of species-specific splicing regulatory sequences from repetitive elements.

Rescued peaks for AGO2 were associated with microRNA-mediated mRNA repression

Next, we used CLAM to analyze a CLIP-seq dataset of AGO2 in the GM12878 LCL (28). AGO2 belongs to the Argonaute (AGO) protein family and plays a critical role in RNA silencing including microRNA-mediated mRNA repression (44). CLIP-seq analysis of AGO2 allows transcriptome-wide identification of microRNA binding sites (45). CLAM rescued >2000 peaks from the AGO2 CLIP-seq data (Table 3), with over half of these rescued peaks located within repetitive elements (Figure 2C).

To assess if these rescued peaks represented functional microRNA target sites, we ran TargetScan (34) to predict the microRNA target sites within each CLIP-seq peak. We then selected two microRNAs (miR-21 and miR-107) for detailed analyses of the predicted TargetScan microRNA target sites. These two microRNAs were selected because they both were abundantly expressed in the GM12878 LCL cell line according to small RNA sequencing data, and global microarray data of mRNA expression following ectopic expression or inhibition of the microRNA were available in the published literature (see 'Materials and Methods' section for details). For each microRNA, we obtained three categories of genes: genes with common peaks containing microRNA target sites, genes with rescued peaks contain-

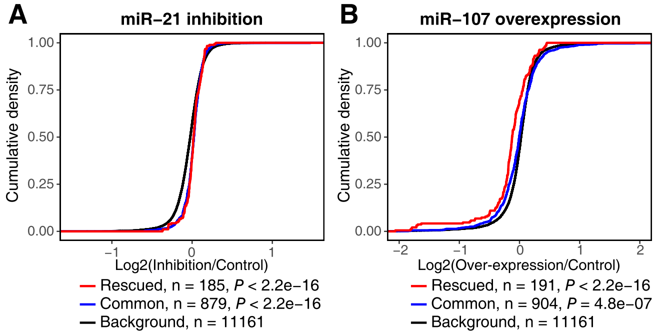

Figure 4. Functional evaluation of CLAM on the AGO2 CLIP-seq data. For each microRNA, three classes of genes are compiled: genes with common peaks containing microRNA target sites (common, blue), genes with rescued peaks containing microRNA target sites (rescued, red) and background genes without AGO2 CLIP-seq peaks (background, black). Cumulative density function is plotted for the log2 gene expression fold change upon (**A**) inhibition of miR-21 or (**B**) ectopic expression of miR-107. For both microRNAs, rescued and common target genes show the same significant shift in cumulative density function as compared to background genes.

ing microRNA target sites and background genes with no AGO2 CLIP-seq peaks. We then calculated the fold change of gene expression level upon ectopic expression or inhibition of the microRNA, then plotted the cumulative density function of the log2 fold change values for the three categories of genes (Figure 4). For miR-21, genes with commons peaks and rescued peaks both had a significant increase in expression levels as compared to background genes following microRNA inhibition (*P*-value < 2.2e-16 and *P*-value < 2.2e-16 respectively, Kolmogorov–Smirnov test), consistent with de-repression of target mRNA levels (Figure 4A). By contrast, for miR-107, genes with common peaks and rescued peaks both had a significant decrease in expression levels as compared to background genes following microRNA overexpression (*P*-value = 4.8e-7 and *P*-value < 2.2e-16 respectively, Kolmogorov–Smirnov test), consistent with repression of target mRNA levels (Figure 4B). These data are characteristic of microRNA's effects on target genes (46), suggesting that the rescued AGO2 peaks provide functional target sites for microRNA-mediated mRNA repression.

Rescued peaks for m^{6}A were associated with mRNA stability control

To test CLAM on RIP-seq data, we applied it to our published RIP-seq data of N^6-methyladenosine (m^{6}A) in the H1 human ESCs (29). The m^{6}A modification involving the addition of a methyl group to the N^6 position of adenosine is a widespread reversible RNA modification in mammalian cells. RNA immunoprecipitation by m^{6}A-specific antibody followed by sequencing is a popular strategy to identify m^{6}A sites across the transcriptome (47). CLAM rescued >3500 peaks from the m^{6}A RIP-seq data. Following an established procedure to identify the consensus m^{6}A motif from m^{6}A RIP-seq data (29), we ranked common or rescued m^{6}A peaks by the ratio of normalized read counts in the m^{6}A RIP-seq data versus the RNA-seq data of the input control, then performed *de novo* motif discovery using

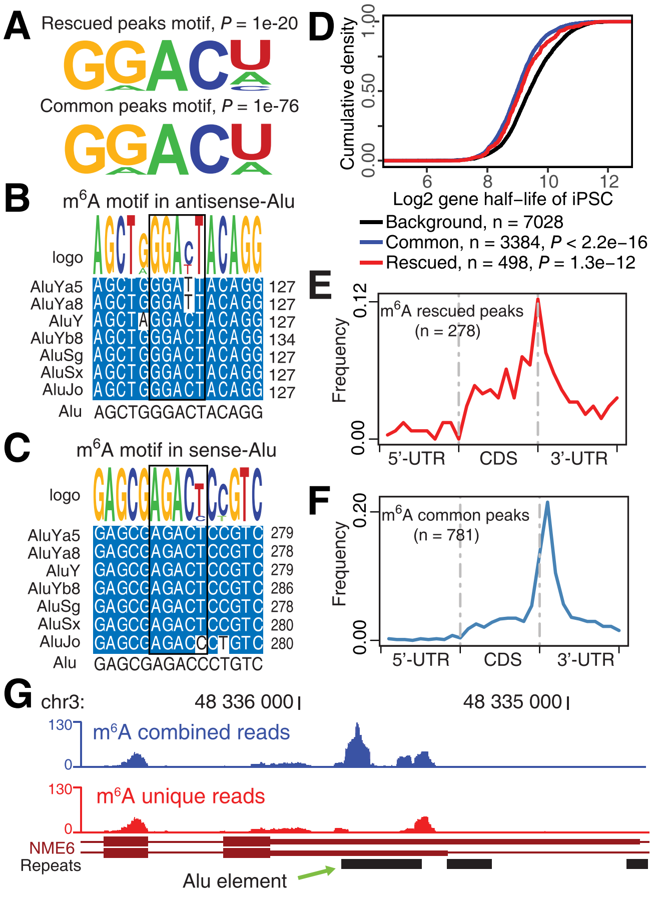

Figure 5. Functional evaluation of CLAM on the m^{6}A RIP-seq data. (**A**) Identification of the known consensus m^{6}A motif by *de novo* motif discovery in rescued and common m^{6}A peaks. The conserved m^{6}A RRACU motif in (**B**) anti-sense and (**C**) sense sequences of major Alu subfamilies. (**D**) Cumulative density function of mRNA half-life in iPSCs. Both genes with common and rescued m^{6}A peaks have significantly lower mRNA half-life as compared to background genes without m^{6}A peaks. Topological distribution of (**E**) rescued and (**F**) common m^{6}A peaks across the 5′-UTR, CDS and 3′-UTR of protein-coding genes. (**G**) Example of a rescued Alu-derived m^{6}A peak in the 3′-UTR of *NME6*.

HOMER (31) in the top 1000 common or rescued peaks. We identified a significant GGACU motif that matched the known consensus m^{6}A motif (Figure 5A). Consistent with the observation that Alu elements were enriched in the rescued m^{6}A peaks (Figure 2C), we identified the consensus RRACU m^{6}A motif in the antisense and sense sequences of Alu subfamilies (Figure 5B and C). To test if these rescued CLAM peaks contained functional m^{6}A sites, we correlated the CLAM sites of human ESCs to published data of mRNA half-life in human induced pluripotent stem cells (iPSCs) (36). As m^{6}A has a well-established role in regulating mRNA degradation and stability (48), we previously observed that genes with functional m^{6}A sites had reduced m^{6}A half-life (29). We classified genes into three categories: genes with common m^{6}A peaks, genes with rescued m^{6}A peaks and background genes without m^{6}A peaks. Genes with common or rescued m^{6}A peaks both had significantly lower mRNA half-life as compared to background genes (*P*-value < 2.2e-16 and *P*-value = 1.3e-12 respectively, Kolmogorov–Smirnov test; see Figure 5D and Supplementary Figure S2), suggesting that the rescued peaks contained functional m^{6}A sites. Furthermore, we observed significant enrichment of both common and rescued m^{6}A sites near the stop codon (Figure 5E and F), demonstrating the similar topological feature of common and rescued m^{6}A sites that matched the previously reported pattern (47). An example of a rescued m^{6}A site was shown in the 3′-UTR of *NME6*, in which a strong RIP-seq peak derived from an Alu retrotransposon was identified by CLAM combining uniquely mapped and multi-mapped reads (Figure 5G).

CLAM analysis of ENCODE CLIP-seq data of 17 splicing factors

To demonstrate the broad applicability of CLAM, we analyzed 17 splicing factors (Supplementary Table S2) with matching eCLIP (enhanced CLIP) data and shRNA knockdown followed by RNA-seq data on the HepG2 cell line from the ENCODE consortium (Figure 6). The ENCODE investigators have systematically performed eCLIP experiments on a large number of RBPs in the HepG2 cell line (37), along with RNA-seq analysis following shRNA knockdown of individual RBPs. For each of the 17 splicing factors, CLAM rescued thousands to tens of thousands of peaks (Supplementary Table S2). Twelve of the seventeen splicing factors had known consensus motifs defined previously using the RNAcompete technology (1). For these splicing factors, we calculated the enrichment *P*-values of known consensus motifs within common or rescued peaks using a *t*-statistic procedure ('Materials and Methods' section). The rescued and common peaks exhibited highly similar patterns of consensus motif enrichment in general for all 12 splicing factors (Figure 6A), despite that the *P*-value calculation could sometimes be skewed for rescued peaks due to their high content of repetitive elements and biased sequence compositions (Figure 2C).

To assess the functional relevance of rescued CLAM sites for these 17 splicing factors, we intersected the common and rescued eCLIP peaks with splicing factor-dependent alternatively spliced cassette exons, identified from RNA-seq data of the HepG2 cell line following shRNA knockdown of the splicing factor. For each exon, we defined three non-overlapping regions as the 250 bp upstream intronic region, the exon body and the 250 bp downstream intronic region. We then tested if exons containing eCLIP peaks (common or rescued) in these regions were significantly enriched toward the top or bottom of the $\Delta\psi$ ranked list of splicing factor-dependent exons using the GSEA algorithm (see 'Materials and Methods' section). As the number of common peaks was generally substantially larger than the number of rescued peaks across all splicing factors (Supplementary Table S2), in order to control for the difference in statistical power in calculating the enrichment *P*-value, we used a down-sampling strategy to randomly sample a subset of common peaks for the enrichment analysis. Our data show that across the 17 splicing factors, splicing factor-dependent alternative exons generally had similar patterns of enrichment for rescued and common peaks, and the −log10 enrichment *P*-value of rescued peaks in approximately half of the tested regions was within the mean ± standard deviation of that of common peaks generated by 20 rounds of down-sampling (marked with an asterisk next to the bar, see Figure 6B), suggesting that rescued and common peaks had similar functional effects on regulating alternative splicing. Two detailed examples were pro-

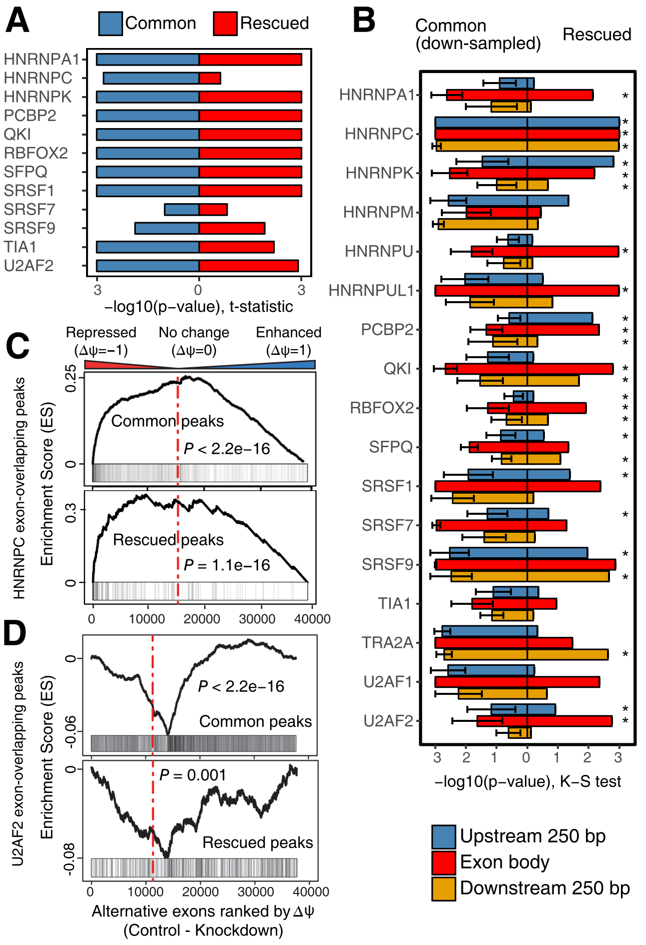

Figure 6. CLAM analysis of 17 splicing factors with ENCODE eCLIP data and matching RNA-seq data following splicing factor knockdown in the HepG2 cell line. In visualizing the negative log10 of nominal *P*-values, we added a pseudo-count of 1e-3 to all *P*-values to truncate the −log10 (*P*-value) at an upper limit of 3, while the same pattern was observed for pseudo-count of 1e-4 and 1e-5. (**A**) Negative log10 enrichment *P*-values of known splicing factor motifs within common (blue) and rescued (red) peaks. The frequency of motif occurrences were compared to randomly sampled genomic sequences and Student's *t*-distribution was fitted to measure the statistical significance of enrichment. (**B**) Barplots of negative log10 *P*-values of GSEA test on the enrichment of splicing factor-dependent alternative exons for common or rescued peaks within the upstream 250 bp intronic region (blue), the exon body (red) and the downstream 250 bp intronic region (orange). For common peaks, the −log10 *P*-value of enrichment was calculated as the average from 20 random iterations of down-sampling to the same number of rescued peaks. If the −log10 *P*-value of rescued peaks is within the mean ± standard deviation of that of common peaks, an asterisk is added next to the bar. (**C**) Enrichment analysis of hnRNPC-dependent exons for common and rescued hnRNPC exon-overlapping peaks in the ENCODE HepG2 data. Both common and rescued hnRNPC peaks are strongly enriched for hnRNPC-repressed exons. (**D**) Enrichment analysis of U2AF2-dependent exons for common and rescued U2AF2 exon-overlapping peaks. Both common and rescued peaks are strongly enriched for U2AF2-enhanced exons in the ENCODE HepG2 data.

vided for hnRNPC and U2AF2 (Figure 6C and D). For hnRNPC, we observed significant enrichment of common and rescued peaks around hnRNPC-repressed exons in the ENCODE HepG2 cells (Figure 6C), consistent with the pattern observed in the HeLa cells (Figure 3B). For U2AF2, we observed significant enrichment of common and rescued peaks around U2AF2-enhanced exons (Figure 6D), consistent with the well-established role of U2AF2 as a positive regulator of exon splicing (49).

DISCUSSION

We report CLAM, a new computational method and software program for CLIP-seq/RIP-seq peak calling incorporating multi-mapped reads. Multi-mapped reads constitute an appreciable fraction of reads in CLIP-seq/RIP-seq experiments (Figure 2A), but conventional analytic tools for CLIP-seq/RIP-seq data do not properly handle multi-mapped reads. In contrast to naïve approaches of discarding multi-mapped reads or distributing fractional counts of multi-mapped reads equally to all mapped loci (20), CLAM utilizes an EM framework to assign reads based on the local information of all mapped reads in the vicinity of multi-mapped reads. Our evaluation using a synthetic benchmark dataset demonstrates that the CLAM EM model outperforms alternative models (Table 2). It should be noted that while the EM algorithm is widely used for resolving multi-mapped RNA-seq reads (23,39,40), existing RNA-seq-based tools are not suitable for CLIP/RIP-seq data. Specifically, RNA-seq-based tools only consider reads mapped to annotated transcript regions and ignore reads in intronic regions, where a large number of CLIP/RIP-seq peaks reside. By contrast, CLAM is designed to account for unique features of CLIP/RIP-seq data. For example, CLAM assigns multi-mapped reads and calls peaks in local windows that match the size of RBP footprints. Collectively, CLAM provides a comprehensive and rigorous computational solution for CLIP/RIP-seq peak calling utilizing multi-mapped reads, and its performance is supported by comprehensive analyses of diverse datasets.

To demonstrate the utility of CLAM, we applied it to a wide range of public CLIP-seq/RIP-seq datasets involving splicing factors, microRNAs and m^{6}A RNA methylation. Consistently across all datasets, CLAM rescued the vast majority of multi-mapped reads in CLIP-seq/RIP-seq libraries, and identified a large number of novel peaks that would otherwise be missed using only uniquely mapped reads. These rescued peaks show expected patterns of consensus motif enrichment. Moreover, analyses of RNA regulatory features suggest that these rescued CLAM peaks are functional, as evidenced by association with alternative splicing (hnRNPC and other splicing factors in ENCODE), steady-state transcript abundance (AGO2) and mRNA half-life (m^{6}A).

An important application of CLAM is to comprehensively discover novel RNA regulatory sites originating from transposable elements in the genome. Extensive research in the past few decades have demonstrated that transposable elements, initially considered as 'genomic parasites' or 'junk DNAs', play important roles in essentially all aspects of gene regulation from transcription to RNA processing to protein synthesis (50). At the RNA level, transposable elements can contribute functional elements for post-transcriptional gene regulation (51). The CLIP-seq/RIP-seq technologies in principle should enable large-scale discoveries of RNA regulatory sites derived from transposable elements, but the repetitive nature of these sequences combined with the short length of CLIP-seq/RIP-seq reads creates computational challenges for peak identification. CLAM provides a statistically rigorous approach to identify CLIP-seq/RIP-seq peaks in repetitive regions of the transcriptome. Across multiple datasets, a significantly higher fraction of 'rescued peaks' identified by CLAM are derived from transposable elements, as compared to 'common peaks' that are readily identifiable using only uniquely mapped reads (Figure 2C). Of note, we identified numerous protein–RNA interaction events and RNA modification sites derived from Alu elements. As Alu elements are primate-specific retrotransposons (52), these Alu derived RNA regulatory sites have the potential to re-wire lineage-specific post-transcriptional regulatory networks, thus contributing to transcriptome diversification during primate and human evolution. For example, m^{6}A RNA methylation has recently emerged as a key player in RNA metabolism (47). While our previous m^{6}A RIP-seq analysis of human and mouse ESCs indicated significant conservation of m^{6}A patterns, we also discovered species-specific m^{6}A sites in over a thousand genes (29). However, the molecular mechanism and evolutionary source for these species-specific m^{6}A sites were unknown. In this work, using CLAM we identified 3218 Alu-derived m^{6}A sites in human genes, revealing the significant contribution of Alu elements to human-specific m^{6}A sites and potentially m^{6}A-associated regulatory effects.

A potentially very valuable feature of CLIP-seq data is the presence of diagnostic signals in CLIP-seq reads (e.g. read truncations and base substitutions) that may allow single-nucleotide-resolution mapping of protein–RNA interaction and RNA modification sites (10,22,53). For example, iCLIP was designed to have single nucleotide resolution through read truncation at the crosslinking sites (10). However, recent literature (54–56) as well as our analysis of the ENCODE data suggest that the robustness of the truncation signals in iCLIP/eCLIP data varies among datasets as well as among individual sites in a single experiment, and could depend on various experimental, technical and biological factors. One important future direction for CLAM is to model CLIP-seq diagnostic signals in a rigorous probabilistic framework to further improve read re-assignment and site identification for CLIP-seq data.

In summary, by modeling and analyzing multi-mapped reads, CLAM provides a more comprehensive solution for CLIP-seq/RIP-seq peak identification beyond commonly used existing methods that focus on uniquely mapped reads. The CLAM software and user manual can be downloaded from https://github.com/Xinglab/CLAM. With the widespread application of CLIP-seq/RIP-seq technologies as well as the rapid accumulation of datasets in the public domain (7), we expect CLAM will be of broad interest to biomedical researchers studying post-transcriptional gene regulation in diverse biological and disease processes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Emad Bahrami-Samani and Wang Xi for discussions and technical assistance. We thank the ENCODE Consortium and the ENCODE production laboratories especially Dr Gene Yeo and Dr Brenton Graveley for generating the eCLIP and RNA-seq data.

FUNDING

National Institutes of Health [R01GM088342 to Y.X.]; Alfred Sloan Research Fellowship (to Y.X.) [BR2013-117]. Funding for open access charge: National Institutes of Health [R01GM088342].
Conflict of interest statement. None declared.

REFERENCES

1. Ray,D., Kazan,H., Cook,K.B., Weirauch,M.T., Najafabadi,H.S., Li,X., Gueroussov,S., Albu,M., Zheng,H., Yang,A. *et al.* (2013) A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**, 172–177.
2. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 829–845.
3. Glisovic,T., Bachorik,J.L., Yong,J. and Dreyfuss,G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.
4. Fu,X.D. and Ares,M. (2014) Context-dependent control of alternative splicing by RNA-binding proteins. *Nat. Rev. Genet.*, **15**, 689–701.
5. Dittmar,K.A., Jiang,P., Park,J.W., Amirikian,K., Wan,J., Shen,S., Xing,Y. and Carstens,R.P. (2012) Genome-wide determination of a broad ESRP-regulated posttranscriptional network by high-throughput sequencing. *Mol. Cell. Biol.*, **32**, 1468–1482.
6. Lambert,N., Robertson,A., Jangi,M., McGeary,S., Sharp,P.A. and Burge,C.B. (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol. Cell*, **54**, 887–900.
7. Yang,Y.C.T., Di,C., Hu,B.Q., Zhou,M.F., Liu,Y.F., Song,N.X., Li,Y., Umetsu,J. and Lu,Z.J. (2015) CLIPdb: a CLIP-seq database for protein-RNA interactions. *BMC Genomics*, **16**, 51.
8. 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.
9. 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.
10. 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.
11. Zhao,J., Ohsumi,T.K., Kung,J.T., Ogawa,Y., Grau,D.J., Sarma,K., Song,J.J., Kingston,R.E., Borowsky,M. and Lee,J.T. (2010) Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol. Cell*, **40**, 939–953.
12. Wang,T., Xiao,G., Chu,Y., Zhang,M.Q., Corey,D.R. and Xie,Y. (2015) Design and bioinformatics analysis of genome-wide CLIP experiments. *Nucleic Acids Res.*, **43**, 5263–5274.
13. Bahrami-Samani,E., Vo,D.T., de Araujo,P.R., Vogel,C., Smith,A.D., Penalva,L.O. and Uren,P.J. (2015) Computational challenges, tools, and resources for analyzing co- and post-transcriptional events in high throughput. *Wiley Interdiscip. Rev. RNA*, **6**, 291–310.
14. Linder,B., Grozhik,A.V., Olarerin-George,A.O., Meydan,C., Mason,C.E. and Jaffrey,S.R. (2015) Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods*, **12**, 767–772.
15. Dominissini,D., Moshitch-Moshkovitz,S., Salmon-Divon,M., Amariglio,N. and Rechavi,G. (2013) Transcriptome-wide mapping of N6-methyladenosine by m6A-seq based on immunocapturing and massively parallel sequencing. *Nat. Protoc.*, **8**, 176–189.
16. de Koning,A.J., Gu,W., Castoe,T.A., Batzer,M.A. and Pollock,D.D. (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.*, **7**, e1002384.
17. Gong,C. and Maquat,L.E. (2011) lncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. *Nature*, **470**, 284–288.
18. Nishikura,K. (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat. Rev. Mol. Cell. Biol.*, **17**, 83–96.
19. Zarnack,K., Konig,J., Tajnik,M., Martincorena,I., Eustermann,S., Stevant,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.
20. Kelley,D.R., Hendrickson,D.G., Tenen,D. and Rinn,J.L. (2014) Transposable elements modulate human RNA abundance and splicing via specific RNA-protein interactions. *Genome Biol.*, **15**, 1–16.
21. Harrow,J., Frankish,A., Gonzalez,J.M., Tapanari,E., Diekhans,M., Kokocinski,F., Aken,B.L., Barrell,D., Zadissa,A., Searle,S. *et al.* (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.*, **22**, 1760–1774.
22. 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.
23. Xing,Y., Yu,T., Wu,Y.N., Roy,M., Kim,J. and Lee,C. (2006) An expectation-maximization algorithm for probabilistic reconstructions of full-length isoforms from splice graphs. *Nucleic Acids Res.*, **34**, 3150–3160.
24. Xue,Y., Zhou,Y., Wu,T., Zhu,T., Ji,X., Kwon,Y.S., Zhang,C., Yeo,G., Black,D.L., Sun,H. *et al.* (2009) Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping. *Mol. Cell*, **36**, 996–1006.
25. 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.
26. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methodol.*, **57**, 289–300.
27. 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.
28. Wan,Y., Qu,K., Zhang,Q.C., Flynn,R.A., Manor,O., Ouyang,Z., Zhang,J., Spitale,R.C., Snyder,M.P., Segal,E. *et al.* (2014) Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, **505**, 706–709.
29. Batista,P.J., Molinie,B., Wang,J., Qu,K., Zhang,J., Li,L., Bouley,D.M., Lujan,E., Haddad,B., Daneshvar,K. *et al.* (2014) m(6)A RNA modification controls cell fate transition in mammalian embryonic stem cells. *Cell Stem Cell*, **15**, 707–719.
30. Bahrami-Samani,E., Penalva,L.O., Smith,A.D. and Uren,P.J. (2015) Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Res.*, **43**, 95–103.
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. Shen,S., Park,J.W., Lu,Z.X., 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.
33. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a

knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.

34. Agarwal,V., Bell,G.W., Nam,J.-W. and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Elife*, **4**, e05005.
35. Durinck,S., Spellman,P.T., Birney,E. and Huber,W. (2009) Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc.*, **4**, 1184–1191.
36. Neff,A.T., Lee,J.Y., Wilusz,J., Tian,B. and Wilusz,C.J. (2012) Global analysis reveals multiple pathways for unique regulation of mRNA decay in induced pluripotent stem cells. *Genome Res.*, **22**, 1457–1467.
37. 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.
38. Pandit,S., Zhou,Y., Shiue,L., Coutinho-Mansfield,G., Li,H., Qiu,J., Huang,J., Yeo,G.W., Ares,M. and Fu,X.-D. (2013) Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol. Cell*, **50**, 223–235.
39. Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
40. Trapnell,C., Williams,B.A., Pertea,G., Mortazavi,A., Kwan,G., van Baren,M.J., Salzberg,S.L., Wold,B.J. and Pachter,L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.
41. Chung,D., Kuan,P.F., Li,B., Sanalkumar,R., Liang,K., Bresnick,E.H., Dewey,C. and Keles,S. (2011) Discovering transcription factor binding sites in highly repetitive regions of genomes with multi-read analysis of ChIP-Seq data. *PLoS Comput. Biol.*, **7**, e1002111.
42. Wang,J., Huda,A., Lunyak,V.V. and Jordan,I.K. (2010) A Gibbs sampling strategy applied to the mapping of ambiguous short-sequence tags. *Bioinformatics*, **26**, 2501–2508.
43. Boudreau,R.L., Jiang,P., Gilmore,B.L., Spengler,R.M., Tirabassi,R., Nelson,J.A., Ross,C.A., Xing,Y. and Davidson,B.L. (2014) Transcriptome-wide discovery of microRNA binding sites in human brain. *Neuron*, **81**, 294–305.
44. Hutvagner,G. and Simard,M.J. (2008) Argonaute proteins: key players in RNA silencing. *Nat. Rev. Mol. Cell Biol.*, **9**, 22–32.
45. Chi,S.W., Zang,J.B., Mele,A. and Darnell,R.B. (2009) Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, **460**, 479–486.
46. Guo,H., Ingolia,N.T., Weissman,J.S. and Bartel,D.P. (2010) Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature*, **466**, 835–840.
47. Fu,Y., Dominissini,D., Rechavi,G. and He,C. (2014) Gene expression regulation mediated through reversible m6A RNA methylation. *Nat. Rev. Genet.*, **15**, 293–306.
48. Wang,X. and He,C. (2014) Reading RNA methylation codes through methyl-specific binding proteins. *RNA Biol.*, **11**, 669–672.
49. Shao,C., Yang,B., Wu,T., Huang,J., Tang,P., Zhou,Y., Zhou,J., Qiu,J., Jiang,L., Li,H. *et al.* (2014) Mechanisms for U2AF to define 3′ splice sites and regulate alternative splicing in the human genome. *Nat. Struct. Mol. Biol.*, **21**, 997–1005.
50. Feschotte,C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, **9**, 397–405.
51. Elbarbary,R.A., Lucas,B.A. and Maquat,L.E. (2016) Retrotransposons as regulators of gene expression. *Science*, **351**, aac7247.
52. Hasler,J. and Strub,K. (2006) Alu elements as regulators of gene expression. *Nucleic Acids Res.*, **34**, 5491–5497.
53. Konig,J., Zarnack,K., Luscombe,N.M. and Ule,J. (2012) Protein-RNA interactions: new genomic technologies and perspectives. *Nat. Rev. Genet.*, **13**, 77–83.
54. 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.
55. 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.
56. 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.