Deep and accurate detection of m^{6}A RNA modifications using miCLIP2 and m6Aboost machine learning

Authors:

Nadine Körtel[1,#], Cornelia Rücklé[1,#], You Zhou[2,#], Anke Busch[1], FX Reymond Sutandy[1,3], Jacob Haase[4], Mihika Pradhan[1], Michael Musheev[1], Dirk Ostareck[5], Antje Ostareck-Lederer[5], Christoph Dieterich[6,7], Stefan Hüttelmaier[4], Christof Niehrs[1,8], Dan Dominissini[9], Julian König[1,*], and Kathi Zarnack[2,*]

Affiliation:

[1] Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz, Germany
[2] Buchmann Institute for Molecular Life Sciences (BMLS) & Faculty of Biological Sciences, Goethe University Frankfurt, Max-von-Laue-Str. 15, 60438 Frankfurt, Germany
[3] Institute of Biochemistry II, Goethe University Frankfurt, Theodor-Stern-Kai 7, 60590, Frankfurt, Germany
[4] Institute of Molecular Medicine, Sect. Molecular Cell Biology, Martin Luther University Halle-Wittenberg, Charles Tanford Protein Center, Kurt-Mothes-Str. 3, 06120 Halle, Germany
[5] Department of Intensive Care Medicine, University Hospital RWTH Aachen, Pauwelsstr. 30, 52074 Aachen
[6] Klaus Tschira Institute for Integrative Computational Cardiology, University Hospital Heidelberg, Im Neuenheimer Feld 669, 69120 Heidelberg, Germany
[7] German Centre for Cardiovascular Research (DZHK) - Partner Site Heidelberg/Mannheim, 69120 Heidelberg, Germany
[8] Division of Molecular Embryology, DKFZ-ZMBH Alliance, Heidelberg, Germany
[9] Cancer Research Center and Wohl Institute for Translational Medicine, Chaim Sheba Medical Center, Tel HaShomer, and Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel

[#] These authors contributed equally. Order of first authors was determined by lottery.

* Corresponding authors: Kathi Zarnack (kathi.zarnack@bmls.de), Julian König (j.koenig@imb-mainz.de)

Abstract

N6-methyladenosine (m^6A) is the most abundant internal RNA modification in eukaryotic mRNAs and influences many aspects of RNA processing, such as RNA stability and translation. miCLIP (m^6A individual-nucleotide resolution UV crosslinking and immunoprecipitation) is an antibody-based approach to map m^6A sites in the transcriptome with single-nucleotide resolution. However, due to broad antibody reactivity, reliable identification of m^6A sites from miCLIP data remains challenging. Here, we present several experimental and computational innovations, that significantly improve transcriptome-wide detection of m^6A sites. Based on the recently developed iCLIP2 protocol, the optimised miCLIP2 results in high-complexity libraries from less input material, which yields a more comprehensive representation of m^6A sites. Next, we established a robust computational pipeline to identify true m^6A sites from our miCLIP2 data. The analyses are calibrated with data from *Mettl3* knockout cells to learn the characteristics of m^6A deposition, including a significant number of m^6A sites outside of DRACH motifs. In order to make these results universally applicable, we trained a machine learning model, m6Aboost, based on the experimental and RNA sequence features. Importantly, m6Aboost allows prediction of genuine m^6A sites in miCLIP data without filtering for DRACH motifs or the need for Mettl3 depletion. Using m6Aboost, we identify thousands of high-confidence m^6A sites in different murine and human cell lines, which provide a rich resource for future analysis. Collectively, our combined experimental and computational methodology greatly improves m^6A identification.

Highlights:

- miCLIP2 produces complex libraries to map m^6A RNA modifications
- *Mettl3* KO miCLIP2 allows to identify Mettl3-dependent RNA modification sites
- Machine learning predicts genuine m^6A sites from human and mouse miCLIP2 data without *Mettl3* KO
- m^6A modifications frequently occur outside of DRACH motifs and associates with alternative splicing

Introduction

The epitranscriptome collectively describes modifications in RNA and has emerged as a crucial and complex mechanism for the post-transcriptional regulation of gene expression. Pervasively occurring in all three kingdoms of life, N6-methyladenosine (m^6A) is the most prevalent internal modification on mRNA [1, 2]. The emerging interest in RNA modifications revealed m^6A as an essential regulator in almost all aspects of mRNA metabolism and uncovered diverse physiological functions [3-8].

m^6A is a dynamic modification. It is deposited by writers, recognized by readers and removed by erasers. The writing of m^6A in mRNA is mainly carried out by a highly conserved, multicomponent methyltransferase complex that catalyses the conversion of adenosine to m^6A. The methyltransferase like 3 (METTL3) acts as the catalytic active subunit, possessing an S-adenosylmethionine (SAM) binding domain (MTA-70 like domain) with the conserved catalytic DPPW motif (Asp-Pro-Pro-Trp) [9]. It installs m^6A by transferring a methyl group of a SAM donor to targeted adenosines [10]. While methyltransferase like 14 (METTL14) is catalytically inactive, it forms a stable heterodimer with METTL3; simultaneously facilitating RNA interaction and increasing the catalytic activity of METTL3 [5, 9, 11]. Additionally, different methyltransferases were identified as m^6A writers which mainly add m^6A to U2 and U6 snRNAs, lncRNA or pre-mRNA [12-14]. In mRNA, m^6A enriches in a DRACH ([G/A/U][G>A]m^6AC[U>A>C]) consensus sequence and occurs in thousands of transcripts, with an average of one to three m^6A sites per mRNA transcript [15-17]. However, only a fraction of DRACH motifs contain an m^6A modification. Furthermore, m^6A was found to cluster predominantly within the coding sequence in long internal exons, nearby stop codons and in the 3' UTR [15, 16].

In order to fully capture and understand the cellular impact of m^6A, it is essential to precisely locate the modification. Although m^6A had been identified over four decades ago, only recent fundamental technological breakthroughs allowed transcriptome-wide mapping of m^6A [15, 16, 18, 19]. Antibody-based immunoprecipitation followed by high-throughput sequencing (m^6A-seq, m^6A-MeRIP) enabled mapping of m^6A within a ~100 nucleotide (nt) window and paved the way to further understand and dissect the cellular and physiological functions of m^6A [15, 16]. Further improvements in 2015 led to an individual-nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP)-based method, called m^6A iCLIP (miCLIP), which allows the transcriptome-wide mapping of individual m^6A residues at single-nucleotide resolution [17].

Despite the novel and important insights these epitranscriptomic sequencing methods uncovered, they also suffered several limitations. A critical disadvantage is the required high amount of input material, which makes transcriptome-wide m^6A detection exclusionary for samples with limited input material. Hence, sequencing low input samples using the aforementioned techniques may lead to over-amplified

libraries with a high PCR duplication rate and low complexity. Moreover, it is broadly observed that miCLIP data comprise a lot of background signal due to limited antibody specificity, which makes computational analysis for m^6A-site identification challenging [20-23].

Here, we present the optimised miCLIP2 protocol, along with the machine-learning-based analysis tool m6Aboost to overcome these limitations. Based on our recently developed iCLIP2 protocol, miCLIP2 improvements comprise two separately ligated adapters, two independent cDNA amplification steps and a bead-based size selection [24]. These advances result in high-complexity miCLIP2 libraries using less input material at less effort. We performed miCLIP2 in murine embryonic stem cells (mESC), using wild-type (WT) and *Mettl3* knockout (KO) cells to identify peaks that are significantly depleted upon *Mettl3* KO and validated selected m^6A sites by an orthogonal method. The resulting high-confidence m^6A sites within DRACH and non-DRACH motifs were used to train a machine learning model, named m6Aboost, which recognises the specific characteristics of m^6A sites in miCLIP2 data. We applied m6Aboost to multiple miCLIP2 datasets from human and mouse. Thus, our new miCLIP2 protocol in combination with our m6Aboost machine learning model allow to globally predict m^6A sites in miCLIP2 datasets independently of a *Mettl3* KO.

Results

The miCLIP2 protocol allows profiling of m^{6}A RNA modifications

In order to allow for deep m^{6}A profiling, we combined the miCLIP procedure with our recently optimised iCLIP2 protocol, termed miCLIP2 (**Figure 1A**) [17, 24]. Experiments were performed with poly(A)+ RNA from mouse embryonic stem cells (mESCs). We first performed two consecutive rounds of poly(A)+ RNA enrichment for total RNA samples (**Figure S1A**) and optimised the RNA fragmentation time required for each sample (**Figure S1B**). The RNA was then incubated with an m^{6}A-specific antibody (Synaptic Systems), which was previously shown to yield highest truncation efficiency in miCLIP experiments (**Figure 1A**) [17]. After optimising UV irradiation (254 nm twice with 150 mJ/cm^2 strength; **Figure S1C**), crosslinked antibody-RNA complexes were immunoprecipitated using protein A beads. Co-purified RNAs were dephosphorylated with T4 polynucleotide kinase (PNK) prior to first adapter ligation (L3-APP) and radioactive labelling. After SDS-PAGE gel and transfer, the respective nitrocellulose membrane fragment was excised (**Figure S1D**). Transferred RNA was recovered by proteinase K treatment, leaving a polypeptide at the crosslinking site. Reverse transcription generally truncates at this polypeptide thus encoding the positional information about m^{6}A sites within resulting cDNA fragments [17, 25]. The residual readthrough events usually incorporate C-to-T transitions [17], which provide additional confidence for truncation-identified crosslink sites (see below). After bead-based clean-up and second linker ligation, a pre-amplification PCR (6 cycles) was employed to minimise loss of information by potential material loss in the following steps. This was followed by size selection to remove primer dimers and a second PCR which was optimized for a minimal number of PCR cycles to obtain sufficient material for sequencing (here 11 cycles). After a second size selection to remove remaining primers, the library was subjected to high-throughput sequencing (**Figure S1E**).

The majority of miCLIP2 peaks are not sensitive to *Mettl3* KO

In order to test whether miCLIP2 peaks are dependent on Mettl3, we performed miCLIP2 experiments (n=3 replicates) from wild-type (WT) as well as *Mettl3* knockout (KO) mESCs. The latter lack the primary m^{6}A methyltransferases Mettl3 and hence, lost most of m^{6}A mRNA methylation (**Figure 1B**) [26, 27]. Reads with C-to-T transitions (6%) were removed for later usage (**Table S1**). The remaining reads corresponded to a total of 261 million putative truncation events (**Table S1**). Peak calling on the data from WT mESC cells identified more than 500,000 peaks that exceeded the local background signal (peaks on all samples are reported in **Table S1**). The number of truncation events in called peaks were highly reproducible between replicates (**Figure 1C, S2A**). To allow for quantitative comparisons between transcripts, we calculated the relative signal strength of all peaks, which was independent of the underlying transcript abundance (see Methods; **Figure S2B**).

Analysis of the underlying sequence showed that most peaks resided on thymidine rather than adenosine and only 25% of these adenosines were part of a DRACH motif (**Figure 1D-F**), reflecting UV crosslinking biases and limited antibody specificity as reported previously [20, 21]. Nevertheless, the strongest peaks frequently coincided with AC and were located precisely on the A nucleotide (**Figure S2C**). We noted an additional enrichment of AC downstream of the peaks. However, these particular peaks did not harbour a DRACH motif and their signal was not reduced in the *Mettl3* KO, indicating that they are part of the unspecific background signal of the employed antibody or m^6A sites independent of Mettl3 (**Figure S2C**). Importantly, peaks at AC and DRACH motifs were specifically lost in the *Mettl3* KO, supporting that miCLIP2 detects Mettl3-dependent m^6A modifications (**Figure 1E, F, S2C, D**). The high amount of nonspecific background in the miCLIP2 data required more precise measures to define true Mettl3-dependent m^6A sites.

Differential methylation analysis detects Mettl3-dependent m^6A sites at DRACH and non-DRACH motifs

In order to learn about the features of genuine m^6A sites in the miCLIP2 data, we sought to extract all miCLIP2 peaks that significantly changed in the *Mettl3* KO mESCs. However, changes at individual peaks were overshadowed by massive shifts in gene expression in *Mettl3* KO cells, with more than 2,809 genes altered at least 2-fold in comparison to WT mESCs (false discovery rate [FDR] ≤ 0.01; **Figure 2A**). These massive shifts in the underlying transcript abundances meant that miCLIP2 read counts at individual peaks could not be compared directly. In order to overcome this shortcoming, we tested several strategies for differential methylation analysis to account for the substantial gene expression changes in the *Mettl3* KO cells (see **Supplementary Material**, Section A). Best performance was achieved with the *bin-based* approach, in which genes were stratified according to their expression change upon *Mettl3* KO (**Figure 2B, S3A-C**). All miCLIP2 peaks within the genes of same bin, i.e., with a similar change in gene expression, were then tested collectively using DESeq2 [28] (see **Supplementary Material**, Section A). As expected, the changing peaks almost exclusively showed a loss of miCLIP2 signal in the *Mettl3* KO (**Figure 2C**), and 85.3% of these downregulated peaks were located at A (**Figure 2D**), supporting that our differential methylation analysis enriched for m^6A sites. From these, we compiled a stringent set of 11,707 sites at A with reduced signal in the *Mettl3* KO (log_2-transformed fold change [log2FC] < 0, FDR ≤ 0.01), which served as 'positive set' of true m^6A sites in the following analyses (see **Supplementary Material**, Section A). As previously described, the positive sites accumulated nearby stop codons and in 3' UTRs, and the underlying sequences resembled the DRACH motif [16, 29] (**Figure 2E, F**), supporting that they indeed represented Mettl3-dependent m^6A sites. For comparison, we selected a 'negative set' of 42,090 peaks that were also located at A

but unchanged or even mildly increased upon *Mettl3* KO (log2FC ≥ 0, FDR > 0.5) and hence represented the nonspecific background in the data.

Among the DRACH motifs identified in the positive set, the most frequent pentamer was GGACT, followed by GAACT and AGACT [17] (**Figure 2G**). Surprisingly, however, we also detected 741 m^6A sites (6.3%) at non-DRACH motifs (non-DRACH m^6A). While most of these non-DRACH motifs still contained the AC dinucleotide [30], some also diverged from this, such as GGATT (**Figure 2G**). We used SELECT (single-base elongation- and ligation-based qPCR amplification) [31] as an orthogonal antibody-independent m^6A detection method to test the reliability of our approach. To this end, we compared SELECT qPCR amplification curves from WT versus *Mettl3* KO samples for an exemplary non-DRACH m^6A site from the positive set, located in the last exon of the *Trim27* gene (A at position chr13:21192298:+, GGATT). Indeed, we detected Mettl3-dependent methylation at A in the GGATT motif, reflected in a reduced efficiency of the qPCR amplification when the m^6A mark is present (**Figure 2H**). As a control, we tested an adjacent A in the same gene (position chr13:21192294:+), which remained unchanged upon *Mettl3* KO (**Figure S3D**). We similarly validated two out of two additional non-DRACH m^6A sites in the genes *Palm3* (chr8:84029842:+, GTACT) and *Hic2* (chr16:17257755:+, GGACG) (**Figure 2H, S3D**). For comparison, we also confirmed three out of three m^6A sites at *bona fide* DRACH motifs in the genes *Eif4ebp1* (chr8:27275332:+, TGACT), *Ccnt2* (chr1:127802764:+, GAACA) and *Phb2* (chr6:124716745:+, GAACT) (**Figure 2I, S3D**).

DRACH motifs were also present at 1,043 peaks (2.5%) in the negative set. The miCLIP2 signal at these peaks did not go down in the *Mettl3* KO, indicating that the antibody may show a residual background activity against the DRACH motif itself. SELECT experiments for two out of two selected sites in the genes *Nanog* (chr6:122711605:+) and *Zfp710* (chr7:8008671:+) confirmed that the respective A indeed did not carry an N6-methyl modification (**Figure 2J**).

Altogether, we defined a positive set of more than 10,000 m^6A sites, that are modified in a Mettl3-dependent manner. In addition to canonical DRACH motifs, we identified a substantial number of m^6A modifications at non-DRACH motifs.

Machine learning allows to reliably predict m^6A sites from miCLIP2 data

To allow for m^6A detection independently of an accompanying KO dataset, we built a predictive machine learning model to discriminate true m^6A sites from background signal in the miCLIP2 data (**Figure 3A**). For model training, we combined the positive (n=11,707) and negative (n=42,090) set identified in the differential methylation analysis upon *Mettl3* KO. The unbalanced setup was chosen to reflect the predominance of nonspecific background in the miCLIP2 data (**Figure 1D, F**). We randomly split the data into a training set (80%) and an independent test set (20%). The input variables for training included 10-nt flanking nucleotide sequence to either

side of A, the transcript region, and the relative signal strength. We further added, as orthogonal information, the number of coinciding C-to-T transitions in the read-through reads, which we initially removed from the data (**Figure 3B**, see **Supplementary Material**, Section B).

We tested three different machine learning algorithms, which consistently reached high predictive accuracy (support vector machine, random forest, and adaptive boosting [AdaBoost]; **Figure S4A-D**, see **Supplementary Material**, Section B). Following a series of benchmarks, we chose the AdaBoost-based predictor, which we named m6Aboost. AdaBoost is a boosting ensemble algorithm that weights the input for each iteration by the misclassification errors from previous iterations, and thereby improves the accuracy of the final predictions [32]. The error rate of m6Aboost on the independent test set reached 0.99%, with >99% area under the curve (AUC) in a precision-recall curve (**Figure 3C, S4A, D**). Evaluation on an independent test set showed that 99% of sites were correctly classified (**Figure 3D**). The performance was confirmed by five-fold crossvalidation (**Figure S4C**). The highest informative content was attributed to the immediate sequence around the modified A nucleotide, the relative signal intensity of peaks, and orthogonal information on C-to-T transitions (**Figure 3B**). Baseline models trained only on sequence information (position -10 to +10; 'sequence-only') or experimental features (relative signal strength, C-to-T transitions, and transcript region; 'feature-only') achieved worse classification results (**Figure 3C**), supporting that both types of features are required for optimal performance. Consistently, our m6Aboost outperformed a simple filter for DRACH motifs (**Figure 3C**, blue dot).

m6Aboost predicts m^{6}A sites also in lowly expressed transcripts

To test the algorithm on a complete miCLIP2 dataset, we applied m6Aboost to all peaks on A nucleotides in the mESC WT miCLIP2 data (n=117,142). In total, m6Aboost extracted 25,456 putative m^{6}A sites in 9,363 genes (**Figure 4A**). These included 11,548 sites from our initial positive set (98.6% of positive set) plus 13,908 additional m^{6}A sites. The latter were enriched in lowly expressed genes and most likely failed to reach significance in the differential methylation analysis due to low read counts (**Figure S4F**). The miCLIP2 signal in these sites coherently went down in the *Mettl3* KO (94% with log2FC < -1; **Figure 4B**), supporting that they are indeed true m^{6}A sites.

Of note, 1,813 out of 25,456 (7.1%) predicted m^{6}A sites resided at non-DRACH motifs (**Figure 4A**). These non-DRACH m^{6}A sites showed an enrichment nearby stop codons similar to the positive set and the vast majority were depleted in the *Mettl3* KO (**Figure 4C, D**), supporting that predicted non-DRACH sites are indeed true m^{6}A sites. On the other hand, m6Aboost predicted that not all peaks at DRACH motifs corresponded to true m^{6}A sites. Indeed, about half of these sites did not respond to *Mettl3* KO and distributed similarly to the negative set (**Figure 4D, S4G**), suggesting that the m^{6}A-

specific antibody shows a residual activity towards the unmodified DRACH motif. The other half had low read counts and preferentially resided in lowly expressed genes (**Figure S4G**), possibly leading to their misclassification. Importantly, m6Aboost associates a prediction score with each site that allows to minimise the number of false positives, at the expense of false negatives, by tightening the prediction score threshold (**Figure S4H**). Altogether, we conclude that m6Aboost efficiently discriminates relevant signal from nonspecific background, offering a reliable prediction of genuine m^6A sites from miCLIP2 data.

For comparison, we also performed miCLIP2 experiments on poly(A)+ RNA from RAW 264.7 cells, a mouse macrophage cell line (three biological replicates, 29.8 million truncation events on average). Out of 462,073 miCLIP2 peaks, m6Aboost identified a total of 18,885 m^6A sites (**Table S1**). Overlay with the mESC data showed that a third of the predicted m^6A sites were shared between both cell lines, rising to about 50% when focussing on genes that were highly expressed in both cell lines (TPM ≥ 20 or more; **Figure 4E, F, S5C**).

m^6A depletion triggers efficient splicing of retained introns

Since our miCLIP2 data was generated for poly(A)-selected RNA, most identified m^6A sites were located in exons. However, we also detected a number of m^6A sites in retained introns. Interestingly, the intronic m^6A sites showed a strong accumulation towards the 5' splice sites (**Figure 4G**), suggesting that they might impact intron splicing. Indeed, using IRFinder [33], we could identify 401 significantly changed intron retention (IR) events in the RNA-seq data of *Mettl3* KO mESCs (change in IR [|ΔIR|] > 3%, FDR < 0.05; **Figure 4H**). 384 out of 401 significantly changed introns showed reduced coverage in the *Mettl3* KO, as seen for intron 5 in the gene *Mif4gd* (**Figure 4J, K**), indicating increased splicing efficiency. Of note, this was reflected in a global reduction in IR across the transcriptome, as 4,563 out of 4,925 measured IR events (92.7%) showed a ΔIR < 0 (**Figure 4H**). Generally, introns harbouring m^6A modifications showed a significant trend towards more IR reduction compared to unmodified introns (**Figure 4I**), indicating that modifications on retained introns may directly influence splicing efficiency.

m^6A sites in HEK293T cells are bound by the cytoplasmic reader protein YTHDF1

To test m6Aboost on miCLIP2 data from a different species, we performed miCLIP2 experiments with poly(A)+ RNA from human HEK293T cells (n=4 replicates with 30 million truncation events on average, **Figure S1F, G**). Starting from more than 788,758 miCLIP2 peaks, m6Aboost identified 36,556 m^6A sites in 7,552 genes, corresponding to 21% of all peaks at A (**Table S1**). The m^6A sites occurred with a median of 3 sites per gene and accumulated around stop codons (**Figure 5A, S5A)**, mirroring the distribution in the mouse cells. As an independent line of evidence, we overlapped the m6Aboost-predicted m^6A sites with binding sites of the cytoplasmic m^6A reader protein YTHDF1

from published iCLIP data [34]. Metaprofiles showed a sharp peak in YTHDF1 binding precisely at the predicted m^6A sites at DRACH motifs (**Figure 5B, C, S5B**). Although less pronounced, we detected considerable YTHDF1 binding also at predicted m^6A sites at non-DRACH motifs, further supporting that these indeed represent genuine m^6A sites.

To further support the predicted m^6A sites, we compared our miCLIP2 data with published miCLIP and m6ACE-seq data for the same cell line [29, 35]. m6A-Crosslinking-Exonuclease-sequencing (m6ACE-seq)-seq is a recently developed tool which incorporates 5' to 3' exoribonuclease treatment after m^6A-antibody crosslinking to increase the resolution and omit radioactive gel electrophoresis [35]. We found that almost half of our m^6A sites overlapped at single-nucleotide level with at least one further dataset (**Figure 5D**). The remaining sites occurred on lowly expressed genes, but still showed an m^6A-typical distribution along transcripts and overlapped with YTHDF1 binding (**Figure 5E, F, S5D**). This suggests that these m^6A sites were missed in other studies due to experimental variability and technical limitations rather than lack of modification.

As a second human cell line, we performed miCLIP2 experiments on poly(A)+ RNA from C643 cells, a human thyroid cancer cell line (three biological replicates, **Table S1**). Here, m6Aboost predicted a total of 19,406 m^6A sites. Comparison with HEK293T showed that similar to mouse, 49.5% of all m^6A sites on highly expressed genes were shared between mESC and C643 cells (TPM ≥ 20 or higher; **Figure 5G, H, S5C**), an estimate that is stable with increasing expression. We therefore conclude that about half of all m^6A modifications are constitutively present in different cell types in human and mouse.

miCLIP2 allows to map m^6A sites from low input material

Most current protocols for antibody-based m^6A detection start from 5-10 µg of poly(A)+ mRNA [36, 37]. In our standard setup, we use just 1 µg, from which we obtain more than 30 million unique miCLIP2 reads on average with low PCR duplication rates (**Table S1**). However, when working with scarce material such as tissue samples, the amount of extractable RNA is often limited. We therefore tested whether miCLIP2 can be applied with even lower RNA input. To this end, we used poly(A)+ mRNA from mouse heart tissue samples and titrated the amount of input RNA down to 50 ng. The resulting miCLIP2 libraries contained 2-50 million truncation events (**Table S1**).

We found that even with these small amounts of input RNA, the miCLIP2 signals were still reproducible at nucleotide level (**Figure 6A, S5E**). As expected, the sensitivity of miCLIP2 progressively decreased with lower input material. The precision, however, was hardly compromised, since the identified sites were highly overlapping at all concentrations (**Figure 6B**). Moreover, m^6A sites from all RNA input concentrations were consistently enriched at DRACH motifs and nearby stop codons (**Figure 6C, D**).

Together, these results suggest that our approach can be used to identify m^6A modifications even from a limited amount of input RNA.

Discussion

Knowledge on the precise location of m^6A sites is essential to unravel the molecular effects and biological functions of this universal RNA modification. With the advent of next-generation sequencing, new experimental protocols allow for a systematic mapping of m^6A sites, often with single-nucleotide resolution [38]. Although alternative methods recently became available [21, 22, 39, 40], the most widely used approaches rely on a set of available antibodies against the modified nucleotide [38]. These methods suffer from the broad reactivity of these antibodies, which crossreact with unmodified adenosines or related modifications such as m^6Am, thereby generating excess false positives [17]. Moreover, many protocols require high amounts of starting material, or target only a restricted subset of m^6A sites that occur for instance in a specific sequence context [21, 22, 36, 37]. In this study, we tackle these limitations by combining the optimised miCLIP2 and the machine learning model m6Aboost to reliably map m^6A modifications at high resolution and depth. Our approach builds on three major experimental and computational innovations that are critical for its efficiency and accuracy.

First, we boosted the efficiency of the experimental miCLIP2 protocol by incorporating the recently published iCLIP2 library preparation [24], including separately ligated adapters, two rounds of PCR amplification and a bead-based clean-up strategy. This reduces the processing time to just four days and provides high-complexity datasets without PCR duplicates. With this setup, we now routinely obtain more than 30 million unique miCLIP2 reads from 1 µg input RNA - twenty times less than in the original protocol [17]. The moderate duplication rate (**Table S1**) indicates the miCLIP2 libraries in this study were not sequenced to saturation, suggesting that many more m^6A sites could still be identified from the same libraries. Moreover, it is possible to obtain reproducible data down to 100 ng and less of input RNA. The reduced input requirement will be particularly useful for studies on nascent RNA or clinical samples and *in vivo* disease models where starting material is limiting.

Second, we tackled the high false positive rate from the m^6A-specific antibodies, which is inherent to antibody-based approaches, through the direct comparison with *Mettl3* KO cells. Using a custom-tailored differential methylation analysis strategy, we identify more than 10,000 Mettl3-dependent m^6A sites in the WT mESC miCLIP2 data that constitute the positive set of high-confidence m^6A sites for subsequent model training (see below). Of note, we find that m^6A modifications frequently occur outside of DRACH motifs (6.3%) and validate selected m^6A sites at non-DRACH motifs using an orthogonal antibody-independent method. Similar motifs were previously reported and recently confirmed in direct RNA sequencing data (Oxford Nanopore Technologies) [17, 41]. Importantly, since the m^6A sites at non-DRACH motifs were included in the m6Aboost model training, similar sites can be readily identified in future miCLIP2 experiments. In addition, we propose that the sequence composition

of the high-confidence m^6A sites from the differential methylation analysis (**Figure 2E**), captured for instance in a position weight matrix, could be used to filter other datasets in a more effective way. Moreover, our strategies to account for changes in transcript abundance in order to identify differentially methylated sites will be applicable for other RNA modifications, such as 5-methylcytosine (m^5C) in m^5C-miCLIP [38, 42].

Third, we trained a machine learning model, termed m6Aboost, to accurately extract Mettl3-dependent m^6A sites from any miCLIP2 dataset. Several machine learning approaches have been developed to predict m^6A sites from the primary RNA sequences [43-45]. However, most existing models were trained on data of limited resolution and size, and consequently perform poorly for single-nucleotide predictions. Here, we apply machine learning to predict m^6A sites in miCLIP2 data based on a high-confidence positive set of Mettl3-dependent m^6A sites. We therefore tackle the inherent problem of false positives that impair most antibody-based m^6A detection protocols [38]. The resulting m6Aboost model allows one to transfer our gained knowledge to other miCLIP datasets without the need for an accompanying *Mettl3* KO, which is not feasible in many biological settings [27]. Because m6Aboost allows for m^6A sites at non-DRACH motifs and sorts out false positive miCLIP2 signals, even at DRACH, it outperforms the commonly used DRACH motif filter [29, 36, 40]. The stringency against false positives can be tuned according to the requirements of the user by adjusting the prediction score of m6Aboost.

We note that our model was trained on miCLIP2 data that was obtained with a specific m^6A antibody (Synaptic Systems). It is known that certain biochemical features such as the truncation rate at the crosslinked antibody and the distribution of C-to-T transitions varies with each antibody [17, 38]. We envision that our approach can be retrained on data for other antibodies against m^6A and other RNA modifications that can be mapped via miCLIP2, if an accompanying depletion dataset is available. The same holds true for the related RNA modification m^6Am, which is present in the miCLIP2 data due to crossreactivity of the m^6A antibody, but currently not extracted by m6Aboost.

In this study, we generated m^6A profiles for four human and mouse cell lines that will serve as a resource for future studies. Comparing the methylation profiles revealed that about half of all m^6A sites are shared between cell lines in either species. Moreover, we confirm that m^6A is mainly deposited around stop codons and within the 3' UTR [15, 16]. Interestingly, we also observe an accumulation near the 5' splice sites of retained introns. Further, our data indicates that m^6A can promote intron retention. This adds a new angle to the controversy surrounding the impact of m^6A modifications on alternative splicing. While some studies reported on extensive splicing alterations upon *Mettl3* depletion, others rebutted a strong connection between m^6A and splicing [46-49]. Consistent with the latter view, we generally observe very few changes in cassette exon splicing in the *Mettl3* KO mESCs. Intron

retention, however seemed to be systemically affected, with retained introns being spliced more efficiently throughout the transcriptome of *Mettl3* KO cells.

In essence, the combination of miCLIP2 and m6Aboost allows for a deep and accurate detection of m^6A sites. Our study illustrates how artificial intelligence helps to eliminate background signals in order to decode high-throughput data.

Materials and methods

LC-MS/MS analysis of m^6A levels

The experiments were performed as described in [50]. Ribonucleoside (A, m^6A) standards, ammonium acetate, and LC/MS grade acetonitrile were purchased from Sigma-Aldrich. $^{13}C_9$-A was purchased from Silantes, GmbH (Munich, Germany). $^2H_3$-m^6A was obtained from TRC, Inc (Toronto, Canada). All solutions were prepared using ultrapure water (Barnstead GenPure xCAD Plus, Thermo Scientific). 0.1–1 µg of poly(A)+ RNA was degraded to nucleosides with 0.003 U nuclease P1 (Roche), 0.01 U snake venom phosphodiesterase (Worthington), and 0.1 U alkaline phosphatase (Fermentas). Separation of the nucleosides from the digested RNA samples was performed with an Agilent 1290 UHPLC system equipped with RRHD Eclipse Plus C18 (95Å, 2.1 × 50 mm, 1.8 µm, Zorbax, USA) with a gradient of 5 mM ammonium acetate (pH 7, solvent A) and acetonitrile (solvent B). Separations started at a flow rate of 0.4 ml/min and linearly increased to 0.5 ml/min during first 7 min. Then, washing and re-conditioning was done at 0.5 ml/min for an additional 3 min and linearly decrease to 0.4 ml/min during the last minute. The gradients were as follows: solvent B linear increase from 0 to 7% for first 3 min, followed by isocratic elution at 7% solvent B for another 4 min; then switching to 0% solvent B for last 4 min, to recondition the column. Quantitative MS/MS analysis was performed with an Agilent 6490 triple quadrupole mass spectrometer in positive ion mode. Details of the method and instrument settings are described in [51]. MRM transitions used in this study were 269.2→137.2 (A), 278.2→171.2 ($^{13}C_9$-A), 282.1→150.1 (m^6A), and 285.1→153.1 ($^2H_3$-N6-mrA). Quantification of all samples utilised biological triplicates, and averaged values of m^6A normalised to A, with the respective standard deviation are shown.

Cell culture and RNA samples

The HEK293T cell line was cultured in DMEM (Life Technologies) containing 10% FBS (Life Technologies), 1% L-glutamine (Life Technologies) and 1% penicillin-streptomycin (Life Technologies) at 37°C with 5% CO_2. All cell lines were monitored for mycoplasma contamination. Anaplastic thyroid carcinoma-derived C643 cells (CLS, RRID:CVCL_5969) were cultured on 15 cm dishes in Dulbecco's Modified Eagle Medium (DMEM; Thermo Fisher Scientific) supplemented with GlutaMAX (Thermo Fisher Scientific) and 10% fetal bovine serum (FBS) at 37°C and 5% CO_2.

Mouse embryonic stem cells (mESC) with wild-type and *Mettl3* KO genotype were taken from a previous publication [27] and cultured as described therein. RAW 264.7 cells (ATCC, Wesel, Germany, TIB-71) were cultured in DMEM (Thermo Fisher Scientific, 12430054) supplemented with 10% heat inactivated FBS (Biochrom, Berlin, Germany, S0613) and 1x penicillin/streptomycin (Thermo Fisher Scientific, 15140-122).

RNA isolation and poly(A) selection

For RNA extraction from HEK293T cells and mESCs, cells were washed in ice-cold PBS and collected on ice for the isolation of total RNA using the RNeasy Plus Mini Kit (Qiagen) following the manufacturer's recommended protocol. For C643 and RAW 264.7 cells, cells were washed with PBS, and total RNA was extracted using TRIzol reagent (Thermo Fisher Scientific) according to the manufacturer's instructions. Prior to isolation of poly(A) RNA, total RNA samples were treated with DNase I (New England Biolabs) according to the manufacturer's protocol, and subsequently cleaned up again by using TRIzol LS (Thermo Fisher Scientific).

For HEK293T and C643 cells, poly(A)+ RNA was extracted using Oligo $d(T)_{25}$ Magnetic Beads using the manufacturer's recommended protocol (Thermo Fisher Scientific, 61002). Poly(A)+ concentration was measured using Qubit™ RNA HS Assay Kit (Thermo Fisher Scientific). For RAW 264.7 cells, poly(A)+ RNA was extracted by incubating 100 µg total RNA with 200 µl Dynabeads solution (Dynabeads mRNA Direct Purification Kit, Thermo Fisher Scientific, 61012) and purified following manufacturer's protocols.

The quality of poly(A)+ RNA was ensured using High sensitivity RNA screen tapes for the 2200 Tape station system (Agilent). If a predominant peak for ribosomal RNA was still detectable, an additional round of poly(A) selection was performed, resulting in one round of selection for mESC and RAW 264.7 cells, and two rounds for HEK293T and C643 cells.

RNA fragmentation

Poly(A)+ RNA was fragmented using RNA fragmentation reagents from Thermo Fisher Scientific. 1 µg of poly(A)+ RNA was filled up to 22 µl with H_2O for each condition. 1 µl of 0.1-0.4x diluted fragmentation buffer was added (always prepared freshly). The mixture was incubated for 7-12 min at 70°C in thermomixer at 1,100 rpm and put immediately on ice. 1 µl of 0.1 - 0.4x diluted STOP solution was added. The solution was mixed and placed back on ice until use. Time of fragmentation and dilution of fragmentation reaction solutions were optimised prior to miCLIP2 experiments for each new batch of RNA.

miCLIP2 experiments

All miCLIP2 experiments were performed with rabbit anti-m^6A antibody purchased from Synaptic Systems (order number 202 003).

<u>UV crosslinking and immunoprecipitation</u>

50 µl of protein A Dynabeads (Dynal, 100.02) were magnetically separated, washed 2 times in 900 µl IP buffer (50 mM Tris, pH 7.4, 100 mM NaCl, 0.05% NP-40) and then resuspended in 50 µl IP buffer and put at 4°C until use. 6 µg of m^6A antibody was added to the 24 µl of fragmented RNA and rotated for 2 h at 4°C. The IP mixture was placed on a parafilm coated dish and UV irradiated with 2x150 mJ/cm² of UV 254 nm. The mixture was placed back into the tube, another 500 µl of IP buffer and 50 µl of

washed protein A beads were added. The mixture was rotated at 4°C for 1 h. The beads were magnetically separated and the supernatant was discarded. The beads were washed 2 time with high-salt wash (50 mM Tris-HCl, pH 7.4, 1 M NaCl, 1 mM EDTA, 1% Igepal CA-630 (Sigma I8896), 0.1% SDS, 0.5% sodium deoxycholate). The second wash was rotated for at least 1 min at 4°C. Subsequently, the beads were washed 2 times with PNK buffer (20 mM Tris-HCl, pH 7.4, 10 mM $MgCl_2$, 0.2% Tween-20) and resuspended in 1 ml PNK buffer (the samples can be left at 4°C until ready to proceed).

3' end RNA dephosphorylation

The beads were magnetically separated and resuspended in 20 µl of 3' end RNA dephosphorylation mixture (4 µl 5x PNK pH 6.5 buffer, 0.5 µl PNK [New England Biolabs; with 3' phosphatase activity], 0.5 µl RNasin, 15 µl water). The mixture was incubated for 20 min at 37°C in a thermomixer at 1100 rpm. The beads were washed 1x with PNK buffer, 1x with high-salt wash (rotate wash for at least 1 min at 4°C) and again washed 2x with PNK buffer.

L3 linker ligation

The supernatant was magnetically removed and the beads were resuspended in 20 µl of L3 linker ligation mixture (8 µl water, 5 µl 4X ligation buffer, 1 µl RNA ligase [New England Biolabs], 0.5 µl RNasin [N2615, Promega GmbH], 1.5 µl pre-adenylated linker L3-App [20 µM; 5'-/rApp/AGATCGGAAGAGCGGTTCAG/ddC/-3'], 4 µl PEG400 [202398, Sigma]). The mixture was incubated overnight at 16°C at 1,100 rpm in a thermomixer. Subsequently, 500 µl of PNK buffer was added. The beads were washed 2 times with 1 ml high-salt buffer and 2x with 1 ml PNK buffer. After the first wash, the mixture was transferred to a new tube.

5' end labelling

The beads were magnetically separated and 4 µl of hot PNK mix (0.2 µl PNK [New England Biolabs], 0.4 µl ^{32}P-γ-ATP, 0.4 µl 10x PNK buffer [New England Biolabs], 3 µl H_2O) was added and incubated for 5 min at 37°C in a thermomixer at 1,100 rpm. Next, the supernatant was removed and 20 µl of 1x NuPAGE loading buffer (4x stock was mixed with water and reducing agent and antioxidant was used to avoid potential interference of antibodies) was added to the beads and incubated at 70°C for 5 min. The beads were magnetically separated and the eluate was loaded on a gel.

SDS-PAGE and nitrocellulose transfer

The beads were magnetically separated and the eluate was loaded on a 4-12% NuPAGE Bis-Tris gel (Invitrogen). 0.5 l of 1x MOPS running buffer (Invitrogen) was used. Additionally, 5 µl of a pre-stained protein size marker was loaded. The gel was run for 50 min at 180 V. The dye front was cut and discarded as solid radioactive waste. For transferring the protein-RNA complexes to a Protan BA85 Nitrocellulose

Membrane, a Novex wet transfer apparatus was used according to the manufacturer's instructions. The transfer was performed for 1 h at 30 V in 1x transfer buffer with 10% methanol. After the transfer, the membrane was rinsed in 1x PBS buffer. Afterwards it was wrapped in saran wrap and exposed to a Fuji film at 4°C for 30 min, 1 h, or overnight. The film was exposed to a Typhoon phosphoimager.

RNA isolation

The protein-RNA complexes were isolated by using the autoradiograph as a mask by cutting the respective regions out of the nitrocellulose membrane. The fragments were placed in a 1.5 ml tube and 10 µl proteinase K (Roche, 03115828001) in 200 µl PK buffer (100 mM Tris-HCl, pH 7.4, 50 mM NaCl, 10 mM EDTA) was added and incubated at 37°C for 20 min at 1100 rpm. 200 µl of PK buffer + 7 M urea (100 mM Tris-HCl pH 7.4, 50 mM NaCl) was added and incubated at 37°C for 20 min at 1,100 rpm. The solution was collected and added together with 400 µl phenol/chloroform (Sigma P3803) to a 2 ml Phase Lock Gel Heavy tube (713-2536, VWR). The mixture was incubated for 5 min at 30°C at 1100 rpm. The phases were separated by spinning for 5 min at 13000 rpm at room temperature. Next, the aqueous layer was transferred into a new tube. Precipitation was performed by addition of 0.75 µl glycoblue (Ambion, 9510), 40 µl 3 M sodium acetate pH 5.5 and addition of 1 ml 100% ethanol. After mixing, the mixture was placed at -20°C over night. The mixture was spinned for 20 min at 15000 rpm at 4°C. After removing the supernatant, the pellet was washed with 0.9 ml 80% ethanol and spinned again for 5 min. After removing the supernatant, the pellet was resuspended in 5 µl H_2O and transferred to a PCR tube.

Reverse transcription

RT primers and dNTPs (1 µl primer Rtclip2.0 [5´-GGATCCTGAACCGCT-3'], 0.5 pmol/µl and 1 µl dNTP mix, 10 mM) were added to the resuspended pellet and incubated in a thermocycler (70°C, 5 min, 25 °C hold until RT mix is added). After adding the RT mix (7 µl H_2O, 4 µl 5x RT buffer [Invitrogen], 1 µl 0.1 M DTT, 0.5 µl RNasin, 0.5 µl Superscript III) the mixture was incubated in a thermocycler (25°C, 5 min; 42°C, 20 min; 50°C, 40 min; 80°C, 5 min; 4°C, hold). 1.65 µl of 1 M NaOH was added and incubated at 98°C for 20 min. Subsequently, 20 µl of 1 M HEPES-NaOH pH 7.3 was added. This will eliminate radioactivity from strongly labelled samples after the next step and prevent RNA from interfering with subsequent reactions.

Silane clean-up

For bead preparation: 10 µl MyONE Silane beads were magnetically separated per sample and the supernatant was removed. The beads were washed with 500 µl RLT buffer and resuspended in 93 µl RLT buffer. For cDNA binding the beads in 93 µl were added to each sample. After mixing, 111.6 µl of 100% ethanol was added. The mixture was carefully mixed and incubated for 5 min at RT. After incubation, the mixture was again mixed and incubated for 5 min further. After magnetically separating the beads

and removing the supernatant, 1 ml of 80% ethanol was added and the mixture was transferred to a new tube. The beads were washed twice in 80% ethanol. The beads were magnetically separated and the supernatant was removed. The tube was briefly mixed in a picoFuge and the remaining supernatant was removed. The beads were air-dried for 5 min at RT. The beads were resuspended in 5 µl H_2O and incubated for 5 min at RT before performing the on-bead ligation. Radioactivity should be removed. If radioactivity is still detected, continue in hot-lab area.

Ligation of 5´ linker to cDNA (on-bead)

The linker was prepared by heating the linker mix (2 µl L##clip2.0 (10 µM stock) 1 µl 100% DMSO) for 2 min at 75°C and keeping it on ice afterwards for >1 min. The linker L##clip2.0 has the sequence 5´-/5Phos/NNNNXXXXXXNNNNNAGATCGGAAGAGCGTCGTG/3ddC/-3´, where N's are the 4-nt and 5-nt random nucleotides from the unique molecular identifier (UMI) and X's are the 6-nt the sample-specific experimental barcode given in **Table S1**. After adding the linker mix to the bead containing sample, the ligation mixture (2.0 µl 10x RNA Ligase Buffer [with DTT; New England Biolabs], 0.2 µl 0.1 M ATP, 9.0 µl 50% PEG 8000, 0.3 µl H_2O, 0.5 µl high conc. RNA Ligase [New England Biolabs]) was pipetted on ice. To ensure homogeneity, the ligation-master-mix was mixed by flicking and spinning it down and was subsequently added with the linker-sample-mix. After vigorous stirring, another 1 µl RNA ligase was added to each sample and mixed by stirring. The mixture was incubated at RT at 1,100 rpm over night.

Silane cleanup of linker-ligated cDNA

Per sample, 5 µl MyONE Silane beads were prepared. The MyONE Silane clean-up was performed as described in the previous Silane clean-up step with following modification: After washing the beads in 500 µl RLT, the beads were resuspended in 60 µl RLT buffer and added to the already bead-containing sample. After the precipitation was performed as previously described, the dried beads were resuspended in 22.5 µl H_2O.

First PCR amplification

The PCR mixture (2.5 µl primer mix 1st PCR [P5Solexa_s, 5´-ACACGACGCTCTTCCGATCT-3´, and P3Solexa_s, 5´-CTGAACCGCTCTTCCGATCT-3], 10 µM each, 25 µl Phusion High Fidelity PCR Master Mix [New England Biolabs, M0531S] was prepared and added to the 22.5 µ of sample from the previous step. A 6-cycle PCR was performed in a thermocycler (98°C, 30 s; 6x [98°C, 10 s; 65°C, 30 s; 72°C, 30 s]; 72°C, 3 min; 16°C, hold).

First ProNex size selection

In order to remove primer and primer-dimers, a bead-based size selection was performed prior to preparative PCRs. In addition to the samples, 50 µl of 'Ultra Low

Range Ladder' (ULR, Thermo Fisher Scientific) will be size selected in parallel to monitor ProNex size selection efficiency. ProNex chemistry was adjusted to RT by keeping it for 30 min at RT. 50 µl of ULR-Phusion mix (1.2 µl ULR Ladder, 28.8 µl H_2O, 30 µl Phusion PCR mastermix [New England Biolabs] and the samples were mixed with 147.5 µl ProNex chemistry. This is a 1:2.95 v/v ratio of sample:beads. This was optimised in previous experiments [24]. The mixture was mixed 10x by pipetting and incubated for 10 min at RT. The sample-bead mixture was placed on a magnetic stand for 2 min and the supernatant was removed. While leaving the bead on the magnetic stand, 200 µl ProNex wash buffer was added to the sample. The buffer was incubated for 60 s before removal. The washes were repeated for a total of 2 washes. After removal of the supernatant, the beads were air-dried for 8-10 min (<60 min) until cracking starts. The beads were eluted in 23 µl H_2O. After 5 min of incubation, the mixture was returned to the magnetic stand for 1 min and the supernatant was carefully transferred to a new tube. The size selection efficiency was monitored for the ULR sample on a High Sensitivity D1000 TapeStation Kit. For comparison, the selected and unselected ULR Phusion mix was analysed. The 75-nt/50-nt ladder fragment ratio was compared which should be around 2.5.

Optimise PCR amplification

In order to prevent over amplification of the library, the PCR cycle has to be optimised to a minimum. Therefore, optimise PCR amplification reactions have to be performed for each sample with each 6 and 10 cycles. The PCR mixture (0.5 µl primer mix P5Solexa [5´-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3´] / P3Solexa [5´-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3´], 10 µM each, 5 µl Phusion High Fidelity PCR Master Mix [New England Biolabs, M0531S], 3.5 µl water) was added to 1 µl of the pre-amplified library. The PCR reaction was performed in a thermocycler (98°C, 30 s; 6 or 10x (98°C, 10 s; 65°C, 30 s; 72°C, 30 s); 72°C, 3 min; 16°C, hold). 2 µl of the amplified library was run on a High Sensitivity D1000 Kit in a TapeStation system. Repeat this step until libraries are seen without over-amplification.

Preparative PCR

From previous results of the PCR cycle optimisation, the minimum of PCR cycles was used to amplify ½ of the library. Here, 2.5 times more concentrated cDNA is used, therefore 1 cycle less is needed than in the preliminary PCR. The PCR mix (8 µl H_2O, 2 µl primer mix P5Solexa/P3Solexa, 10 µM each, 20 µl Phusion HF Mix [New England Biolabs]) was added to 10 µl cDNA. The PCR was performed in a thermocycler using the same program as in the optimization PCR with the optimised cycle number. 2 µl of the amplified library was run on a High Sensitivity D1000 Kit in a TapeStation system. If the results looked fine, the second half of the library was also amplified and

combined with the first half. Finally, the concentration under the peak was determined using TapeStation software, and replicates were combined either in equal molarities or equal volumes.

Second size selection by ProNex

Before submitting the samples for sequencing, another round of bead-based size selection was performed to remove residual primers. This ProNex size selection was performed as described above with the following modifications: After ULR preparation, the samples and beads were mixed in a 1:2.4 v/v ratio of sample:beads. This was optimised in previous experiments in [24]. After the incubation and washing steps, the dried beads were eluted in 20 µl H_2O. Again, for comparison the selected and unselected ULR Phusion mix was analysed as described previously. The 100-nt/75-nt ladder fragment ratio should be around 4.5.

SELECT experiments to validate m^6A modifications

We used the elongation and ligation-based qPCR amplification method SELECT [31] to independently test for m^6A modifications at several putative m^6A sites identified from our miCLIP2 data (**Figure 2H-J, S3D**). All experiments were performed with RNA from WT cell lines and compared to RNA from *Mettl3* KO cell lines (mESC).

Normalisation of input RNA

For *Mettl3* KO cell lines, the amount of m^6A is greatly reduced (**Figure 1B**). Due to m^6A-mediated RNA degradation or stabilisation processes, absence of m^6A may influence the abundance of specific transcripts. To ensure usage of same RNA amounts, Qubit (Thermo Fisher Scientific) with Qubit™ RNA HS Assay Kit (Thermo Fisher Scientific) was used to precisely measure RNA concentrations. To ensure usage of equal amounts of transcripts, qPCR experiments were performed for normalisation of input RNA amounts in WT versus *Mettl3* KO cell lines.

Elongation and ligation-based qPCR amplification

For the quantitative real-time PCR (qPCR)-based validation of a presumed m^6A site (termed X site), two primers (Up and Down primer) were designed flanking the site of interest. To precisely measure RNA concentrations before each experiment, Qubit™ RNA HS Assay Kit (Thermo Fisher Scientific) was used. An influence of m^6A on transcript stability may lead to a difference in transcript abundance upon *Mettl3* KO. Therefore, qPCR for the respective transcript was performed and the amount of total RNA for each SELECT experiment was normalised. To further monitor usage of equal amounts of input material, an Up and Down primer were designed flanking an adjacent nucleotide (termed N site). N sites at X-2 and X+2 were used as input control. According to the previously published SELECT method, 20 ng of total RNA or 10 ng of poly(A)+ RNA was used per experiment. The RNA was mixed in a total volume of 17 µl in 1xCutSmart buffer containing 40 nM Up primer, 40 nM Down primer and 5 µM

dNTPs. The RNA and primers were annealed by incubation in a thermocycler (90°C to 40 °C with a decrease of -10°C after 1 min, then left at 40°C for 6 min). 0.02 U Bst 2.0 DNA polymerase, 0.5 U SplintR ligase and 10 nmol ATP in a volume of 3 µl in 1xCutSmart buffer was added and incubated at 40°C for 20 min. After denaturation at 80°C for 20 min, the mixture was kept at 4°C. Using the Applied Biosystems ViiATM7 Real-Time PCR system, qPCR was performed. The 20 µl qPCR reaction mixture contained 2 µl of the final reaction mixture after denaturation, 0.2 nM per qPCR primer, 2x Luminaris HiGreen Lox Rox (Thermo Fisher Scientific) and ddH_2O. The quantitative qPCR reaction condition was run as follows: 95°C, 5 min; (95°C, 10 s; 60°C, 35 s) x 40 cycles; 95°C, 15 s; 60°C, 1 min; 95°C, 15 s (collect fluorescence at a ramping rate of 0.05°C/s); 4°C hold. qPCR data analysis was performed using QuantStudioTM Real-Time PCR Software v1.3. All experiments were performed in three technical replicates (separate SELECT reactions). Oligonucleotides used for SELECT are listed in **Table S2**.

miCLIP2 read processing

Multiplexed miCLIP2 libraries were sequenced as 91-nt or 92-nt single-end reads on an Illumina NextSeq500 sequencing system including a 6-nt sample barcode as well as 5-nt+4-nt unique molecular identifiers (UMIs).

Initial data processing was done as described in Chapters 3 and 4.1 of [52] for iCLIP data. In short, after checking the sequencing qualities with FastQC (v0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and filtering reads based on sequencing qualities (Phred score) of the barcode region (FASTX-Toolkit v0.0.14) (http://hannonlab.cshl.edu/fastx_toolkit/), seqtk v1.3 (https://github.com/lh3/seqtk/), reads were de-multiplexed based on the experimental barcode (positions 6 to 11 of the reads) and adapter sequences were removed from the read ends (Flexbar v3.4.0) [53]. UMIs were trimmed as well and added to the read names. Reads shorter than 15 nt were removed from further analysis. Individual samples were then mapped to the respective genome (assembly version GRCh38.p12 for all human samples, GRCm38.p6 for all mouse samples) and its annotation (GENCODE release 31 for all human samples, GENCODE release M23 for all mouse samples) [54] using STAR (v2.7.3a) [55]. When running STAR (with parameter --outSAMattributes All), up to 4% mismatches were allowed per read, soft-clipping was prohibited on the 5' end of reads and only uniquely mapping reads were kept for further analysis. Following mapping, sorted BAM files were indexed (SAMtools v1.9) [56] and duplicate reads were removed (UMI-tools v1.0.0) [57]. Reads were defined duplicates if their 5' ends map to the same position and strand in the genome and they have identical UMIs.

After removing duplicates, all mutations found in reads were extracted using the Perl script parseAlignment.pl of the CLIP Tool Kit (CTK, v1.1.3) [58]. The list of all found mutations specifies the mutations, their locations in the genome as well as the names

of the reads in which they were found. The list was filtered for C-to-T mutations using basic Bash commands and kept in BED file format as described in [36]. Based on the filtered list of C-to-T mutations, de-duplicated reads were separated into two BAM files holding reads with and without C-to-T mutation, respectively, using SAMtools and basic Bash commands. The BAM file of reads without C-to-T mutation was transformed to a BED file using bedtools bamtobed (BEDTools v2.27.1) [59] and considering only the 5' mapping position of each read. Afterwards, the BED file was sorted and summarized to strand-specific bedGraph files which were shifted by one base pair upstream (since this nucleotide is considered as the cross-linked nucleotide) using bedtools genomecov (BEDTools v2.27.1). Similarly, the BED files of C-to-T mutations were also sorted and summarized to strand-specific bedGraph files using bedtools genomecov. Finally, all bedGraph files were transformed to bigWig track files using bedGraphToBigWig of the UCSC tool suite (v365) [60]. Reads containing C-to-T transitions were filtered out using parseAlignment.pl from the CTK package [58] as described in [36].

The code for miCLIP2 data processing as described here is available from two recent data analysis publications [36, 52].

Peak calling, transcript assignment and relative signal strength

Bam files with reads without C-to-T mutation were used for peak calling with PureCLIP (v1.3.1) [61] individually on each replicate for each condition. PureCLIP significant sites per replicate were then filtered for presence in at least two out of three replicates for a given condition (PureCLIP peaks in **Table S1**). For assigning a host gene to each PureCLIP peak, transcript annotations were taken from GENCODE (release 31, GRCh38.p12 for human and release M23, mm10 for mouse), and filtered for a transcript support level ≤ 3 and support level ≤ 2. For overlapping transcripts, the longest annotation was chosen. We next assigned the miCLIP2 peaks to the transcripts.

In order to calculate the relative signal strengths of all peaks within a transcript, we calculated the mean number of truncation events for all peaks in the same transcript. Then, we divided the individual truncation read number of each peak by the mean of the peak strength in the corresponding transcript, leading to a value representing the relative peak strength.

Differential methylation analysis to identify Mettl3-dependent m^6A sites

Similar to iCLIP, the miCLIP2 signal is strongly influenced by the underlying transcript abundance [62, 63] (**Figure S2B**). Therefore, when applying DESeq2 [28] collectively to all peaks (*one-run*), any change of transcript abundance will lead to incorrect fold change and FDR estimations, resulting in false positive calls in down-regulated genes (**Figure S3A**). We tested three different approaches to overcome this, namely separately running DESeq2 on peaks of individual genes (*gene-wise*) or groups of

genes with similar abundance change (*bin-based*) as well as by using DEXSeq (*dexseq-run*) [64] instead of DESeq2 (**Figure S3A**). The different approaches are explained in more detail in the **Supplementary Material**. The best performance was seen for the *bin-based* approach, which was used for all following analyses.

Training and evaluation of the machine learning model m6Aboost

Based on the log2-transformed fold change (log2FC) and the false discovery rate (FDR) from the *bin-based* differential methylation analysis between WT and *Mettl3* KO cells, we used peaks at A to compile a positive (log2FC < 0, FDR ≤ 0.01; n=11,707) and negative (log2FC ≥ 0, FDR > 0.5; n=42,090) set. Both were combined and then randomly split into a training set (80%) and an independent test set (20%). We then extracted 27 features, including the nucleotide sequence in a 21-nt window around the central A, the transcript region as well as the relative signal strength ($\log_2$) and the number of associated C-to-T transitions ($\log_2$). We initially tested three different machine learning algorithms (AdaBoost, support vector machine [SVM], random forest) and evaluated their performance based on precision-recall curves and area under the curve (AUC) as well as by comparing F1-score, Matthews correlation coefficient (MCC), precision, accuracy, sensitivity and specificity on the independent test set (**Figure 3C, S4**). Based on these measures, we selected the AdaBoost-based predictor, which we named m6Aboost (see **Supplementary Material**, Section B for details).

RNA-seq read processing

RNA sequencing (RNA-seq) libraries were sequenced on an Illumina NextSeq500 as 84-nt single-end reads, yielding 31-35 million reads per sample. Basic sequencing quality checks were applied to all reads using FastQC (v0.11.8) (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were mapped to the mouse genome (assembly version GRCm38.p6) and its annotation based on GENCODE release M23 using STAR (v2.6.1b) [55]. When running STAR, up to 4% mismatches were allowed per read and only one location was kept for multi-mapping reads. Coverage tracks for visualisation were obtained by merging bam files for each condition using samtools (v1.11). Coverage was calculated with bamCoverage (v3.5.0) from the deepTools suite [65] using RPGC normalisation and –effectiveGenomeSize calculated by ucsc-facount (v377).

For differential gene expression analysis, mapped reads were counted with htseq-count (v0.12.4, -s reverse) [66] into gene annotation based on GENCODE release M23. Differential expression analysis was performed with DESeq2 (v1.30.0) [28] using the method "apeglm" for shrinkage of $\log_2$-transformed fold changes.

Intron retention (IR) analysis was done with IRFinder (v1.3.0) [33] using built-in script analysisWithLowReplicates.pl for differential analysis [67]. We adapted some built-in filtering steps by overwriting line 179 of analysisWithLowReplicates.pl into:

my $ok = ($pA[8] > 0 || $pB[8] > 0) && ($pA[19] > 0 || $pB[19] > 0) && separatedAB(\@repsIR, $repsA, $repsB);

and line 186 into:

if (($pA[8] > 0 || max($pA[16],$pA[17]) > 0) && ($pB[8] > 0 || max($pB[16],$pB[17]) > 0)) {

For downstream analysis, IR events were filtered for IRratio ≥ 0.03 in at least one condition and mean IntronDepth ≥ 3. *P* values were corrected using Benjamini-Hochberg adjustment.

YTHDF1 iCLIP processing and overlap with predicted m^{6}A sites

YTHDF1 iCLIP reads were quality filtered and processed as in Busch et al, 2020 {Busch, 2020 #58}, used tools versions are as described above for miCLIP. For peak calling with PureCLIP [61] reads from the four replicates were merged. Resulting peaks were filtered to be present in at least two out of four replicates. To generate binding sites, peaks closer than 4 nt were merged, allowing no overlapping binding sites. Finally, binding sites were centred at the position with the highest truncation read number as described in [52]. All predicted m^{6}A sites were aligned and spanned with a 21-nt window to count the presence of YTHDF1 binding sites in that area.

Data availability

All miCLIP2 and RNA-seq data sets generated in this study were submitted to the Gene Expression Omnibus (GEO) under the SuperSeries accession GSE163500 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?&acc=GSE163500). Code is available at GitHub (https://github.com/ZarnackGroup/m6Aboost).

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) SPP1935 in a joint project by JK (KO4566/3-2) and KZ (ZA881/5-2). AOL received DFG funding OS290/6-1 within the SPP1935. Support by the IMB Genomics Core Facility and the use of its NextSeq500 (INST 247/870-1FUGG) are gratefully acknowledged.

Acknowledgements

The authors gratefully acknowledge Tina Lence and Jean-Yves Roignant for help with miCLIP, Lina Worpenberg with initial help with SELECT, and Yannic Schumacher for preparing poly(A)$^+$ RNA from RAW 264.7 macrophages. We are grateful to Marcel Schulz for advice on machine learning, and Peter Hoch-Kraft for advice on mESC cells. We kindly thank members of the IMB Genomics and Bioinformatics Core Facilities for technical assistance and reagents. We would like to thank all members of the König and Zarnack groups for lively discussions.

References

1. Wei CM, Gershowitz A, Moss B: **Methylated nucleotides block 5' terminus of HeLa cell messenger RNA.** *Cell* 1975, **4:**379-386.
2. Yue Y, Liu J, He C: **RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation.** *Genes Dev* 2015, **29:**1343-1355.
3. Lin S, Gregory RI: **Methyltransferases modulate RNA stability in embryonic stem cells.** *Nat Cell Biol* 2014, **16:**129-131.
4. Lence T, Akhtar J, Bayer M, Schmid K, Spindler L, Ho CH, Kreim N, Andrade-Navarro MA, Poeck B, Helm M, Roignant JY: **m(6)A modulates neuronal functions and sex determination in *Drosophila*.** *Nature* 2016, **540:**242-247.
5. Zheng G, Dahl JA, Niu Y, Fedorcsak P, Huang CM, Li CJ, Vagbo CB, Shi Y, Wang WL, Song SH, et al: **ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility.** *Mol Cell* 2013, **49:**18-29.
6. Li A, Chen YS, Ping XL, Yang X, Xiao W, Yang Y, Sun HY, Zhu Q, Baidya P, Wang X, et al: **Cytoplasmic m(6)A reader YTHDF3 promotes mRNA translation.** *Cell Res* 2017, **27:**444-447.
7. Wang X, Lu Z, Gomez A, Hon GC, Yue Y, Han D, Fu Y, Parisien M, Dai Q, Jia G, et al: **N6-methyladenosine-dependent regulation of messenger RNA stability.** *Nature* 2014, **505:**117-120.
8. Zaccara S, Ries RJ, Jaffrey SR: **Reading, writing and erasing mRNA methylation.** *Nat Rev Mol Cell Biol* 2019, **20:**608-624.
9. Sledz P, Jinek M: **Structural insights into the molecular mechanism of the m(6)A writer complex.** *Elife* 2016, **5**.
10. Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, et al: **A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation.** *Nat Chem Biol* 2014, **10:**93-95.
11. Wang P, Doxtader KA, Nam Y: **Structural Basis for Cooperative Function of Mettl3 and Mettl14 Methyltransferases.** *Mol Cell* 2016, **63:**306-317.
12. Gu L, Wang L, Chen H, Hong J, Shen Z, Dhall A, Lao T, Liu C, Wang Z, Xu Y, et al: **CG14906 (mettl4) mediates m(6)A methylation of U2 snRNA in Drosophila.** *Cell Discov* 2020, **6:**44.
13. van Tran N, Ernst FGM, Hawley BR, Zorbas C, Ulryck N, Hackert P, Bohnsack KE, Bohnsack MT, Jaffrey SR, Graille M, Lafontaine DLJ: **The human 18S rRNA m6A methyltransferase METTL5 is stabilized by TRMT112.** *Nucleic Acids Res* 2019, **47:**7719-7733.
14. Warda AS, Kretschmer J, Hackert P, Lenz C, Urlaub H, Höbartner C, Sloan KE, Bohnsack MT: **Human METTL16 is a N(6)-methyladenosine (m(6)A) methyltransferase that targets pre-mRNAs and various non-coding RNAs.** *EMBO Rep* 2017, **18:**2004-2014.
15. Dominissini D, Moshitch-Moshkovitz S, Schwartz S, Salmon-Divon M, Ungar L, Osenberg S, Cesarkas K, Jacob-Hirsch J, Amariglio N, Kupiec M, et al: **Topology of**

the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature* 2012, **485:**201-206.

16. Meyer KD, Saletore Y, Zumbo P, Elemento O, Mason CE, Jaffrey SR: **Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons.** *Cell* 2012, **149:**1635-1646.
17. Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR: **Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome.** *Nat Methods* 2015, **12:**767-772.
18. Desrosiers R, Friderici K, Rottman F: **Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells.** *Proc Natl Acad Sci U S A* 1974, **71:**3971-3975.
19. Schäfer KP: **RNA synthesis and processing reactions in a subcellular system from mouse L cells.** *Hoppe Seylers Z Physiol Chem* 1982, **363:**33-43.
20. Mauer J, Jaffrey SR: **FTO, m(6) Am , and the hypothesis of reversible epitranscriptomic mRNA modifications.** *FEBS Lett* 2018, **592:**2012-2022.
21. Garcia-Campos MA, Edelheit S, Toth U, Safra M, Shachar R, Viukov S, Winkler R, Nir R, Lasman L, Brandis A, et al: **Deciphering the "m(6)A Code" via Antibody-Independent Quantitative Profiling.** *Cell* 2019, **178:**731-747 e716.
22. Meyer KD: **DART-seq: an antibody-free method for global m(6)A detection.** *Nat Methods* 2019, **16:**1275-1280.
23. McIntyre ABR, Gokhale NS, Cerchietti L, Jaffrey SR, Horner SM, Mason CE: **Limits in the detection of m(6)A changes using MeRIP/m(6)A-seq.** *Sci Rep* 2020, **10:**6590.
24. Buchbender A, Mutter H, Sutandy FXR, Körtel N, Hänel H, Busch A, Ebersberger S, König J: **Improved library preparation with the new iCLIP2 protocol.** *Methods* 2020, **178:**33-48.
25. König J, Zarnack K, Rot G, Curk T, Kayikci M, Zupan B, Turner DJ, Luscombe NM, Ule J: **iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.** *Nat Struct Mol Biol* 2010, **17:**909-915.
26. Bokar JA, Shambaugh ME, Polayes D, Matera AG, Rottman FM: **Purification and cDNA cloning of the AdoMet-binding subunit of the human mRNA (N6-adenosine)-methyltransferase.** *RNA* 1997, **3:**1233-1247.
27. Geula S, Moshitch-Moshkovitz S, Dominissini D, Mansour AA, Kol N, Salmon-Divon M, Hershkovitz V, Peer E, Mor N, Manor YS, et al: **Stem cells. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation.** *Science* 2015, **347:**1002-1006.
28. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol* 2014, **15:**550.
29. Boulias K, Toczydlowska-Socha D, Hawley BR, Liberman N, Takashima K, Zaccara S, Guez T, Vasseur JJ, Debart F, Aravind L, et al: **Identification of the m(6)Am**

Methyltransferase PCIF1 Reveals the Location and Functions of m(6)Am in the Transcriptome. *Mol Cell* 2019, **75:**631-643 e638.

30. Liu N, Parisien M, Dai Q, Zheng G, He C, Pan T: **Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA.** *RNA* 2013, **19:**1848-1856.
31. Xiao Y, Wang Y, Tang Q, Wei L, Zhang X, Jia G: **An Elongation- and Ligation-Based qPCR Amplification Method for the Radiolabeling-Free Detection of Locus-Specific N(6) -Methyladenosine Modification.** *Angew Chem Int Ed Engl* 2018, **57:**15995-16000.
32. Freund Y, Schapire RE: **A desicion-theoretic generalization of on-line learning and an application to boosting.** In *Computational Learning Theory EuroCOLT 1995 Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence). Volume* 904. Edited by Vitányi P: Springer, Berlin, Heidelberg; 1995
33. Middleton R, Gao D, Thomas A, Singh B, Au A, Wong JJ, Bomane A, Cosson B, Eyras E, Rasko JE, Ritchie W: **IRFinder: assessing the impact of intron retention on mammalian gene expression.** *Genome Biol* 2017, **18:**51.
34. Patil DP, Chen CK, Pickering BF, Chow A, Jackson C, Guttman M, Jaffrey SR: **m(6)A RNA methylation promotes XIST-mediated transcriptional repression.** *Nature* 2016, **537:**369-373.
35. Koh CWQ, Goh YT, Goh WSS: **Atlas of quantitative single-base-resolution N(6)-methyl-adenine methylomes.** *Nat Commun* 2019, **10:**5636.
36. Hawley BR, Jaffrey SR: **Transcriptome-Wide Mapping of m(6) A and m(6) Am at Single-Nucleotide Resolution Using miCLIP.** *Curr Protoc Mol Biol* 2019, **126:**e88.
37. Grozhik AV, Linder B, Olarerin-George AO, Jaffrey SR: **Mapping m(6)A at Individual-Nucleotide Resolution Using Crosslinking and Immunoprecipitation (miCLIP).** *Methods Mol Biol* 2017, **1562:**55-78.
38. Anreiter I, Mir Q, Simpson JT, Janga SC, Soller M: **New Twists in Detecting mRNA Modification Dynamics.** *Trends Biotechnol* 2020.
39. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al: **Highly parallel direct RNA sequencing on an array of nanopores.** *Nat Methods* 2018, **15:**201-206.
40. Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS, Smith MA, Novoa EM: **Accurate detection of m(6)A RNA modifications in native RNA sequences.** *Nat Commun* 2019, **10:**4079.
41. Pratanwanich PN, Zao F, Chen Y, Koh CWQ, Hendra C, Poon P, Goh YT, Yap PML, Yuan CJ, Chng WJ, et al: **Detection of differential RNA modifications from direct RNA sequencing of human cell lines.** *bioRxiv* 2020**:**2020.2006.2018.160010.
42. Hussain S, Sajini AA, Blanco S, Dietmann S, Lombard P, Sugimoto Y, Paramor M, Gleeson JG, Odom DT, Ule J, Frye M: **NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs.** *Cell Rep* 2013, **4:**255-261.

43. Chen Z, Zhao P, Li F, Wang Y, Smith AI, Webb GI, Akutsu T, Baggag A, Bensmail H, Song J: **Comprehensive review and assessment of computational methods for predicting RNA post-transcriptional modification sites from RNA sequences.** *Brief Bioinform* 2020, **21:**1676-1696.

44. Zhao Z, Peng H, Lan C, Zheng Y, Fang L, Li J: **Imbalance learning for the prediction of N(6)-Methylation sites in mRNAs.** *BMC Genomics* 2018, **19:**574.

45. Zhou Y, Zeng P, Li YH, Zhang Z, Cui Q: **SRAMP: prediction of mammalian N6-methyladenosine (m6A) sites based on sequence-derived features.** *Nucleic Acids Res* 2016, **44:**e91.

46. Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vagbo CB, Geula S, Hanna JH, Black DL, Darnell JE, Jr., Darnell RB: **m(6)A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover.** *Genes Dev* 2017, **31:**990-1006.

47. Zhao X, Yang Y, Sun BF, Shi Y, Yang X, Xiao W, Hao YJ, Ping XL, Chen YS, Wang WJ, et al: **FTO-dependent demethylation of N6-methyladenosine regulates mRNA splicing and is required for adipogenesis.** *Cell Res* 2014, **24:**1403-1419.

48. Xiao W, Adhikari S, Dahal U, Chen YS, Hao YJ, Sun BF, Sun HY, Li A, Ping XL, Lai WY, et al: **Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing.** *Mol Cell* 2016, **61:**507-519.

49. Price AM, Hayer KE, McIntyre ABR, Gokhale NS, Abebe JS, Della Fera AN, Mason CE, Horner SM, Wilson AC, Depledge DP, Weitzman MD: **Direct RNA sequencing reveals m(6)A modifications on adenovirus RNA are necessary for efficient splicing.** *Nat Commun* 2020, **11:**6016.

50. Leismann J, Spagnuolo M, Pradhan M, Wacheul L, Vu MA, Musheev M, Mier P, Andrade-Navarro MA, Graille M, Niehrs C, et al: **The 18S ribosomal RNA m(6) A methyltransferase Mettl5 is required for normal walking behavior in Drosophila.** *EMBO Rep* 2020, **21:**e49443.

51. Schomacher L, Han D, Musheev MU, Arab K, Kienhöfer S, von Seggern A, Niehrs C: **Neil DNA glycosylases promote substrate turnover by Tdg during DNA demethylation.** *Nat Struct Mol Biol* 2016, **23:**116-124.

52. Busch A, Brüggemann M, Ebersberger S, Zarnack K: **iCLIP data analysis: A complete pipeline from sequencing reads to RBP binding sites.** *Methods* 2020, **178:**49-62.

53. Roehr JT, Dieterich C, Reinert K: **Flexbar 3.0 - SIMD and multicore parallelization.** *Bioinformatics* 2017, **33:**2941-2942.

54. Frankish A, Diekhans M, Ferreira AM, Johnson R, Jungreis I, Loveland J, Mudge JM, Sisu C, Wright J, Armstrong J, et al: **GENCODE reference annotation for the human and mouse genomes.** *Nucleic Acids Res* 2019, **47:**D766-D773.

55. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics* 2013, **29:**15-21.

56. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

57. Smith T, Heger A, Sudbery I: **UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy.** *Genome Res* 2017, **27:**491-499.

58. Shah A, Qian Y, Weyn-Vanhentenryck SM, Zhang C: **CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data.** *Bioinformatics* 2017, **33:**566-567.

59. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

60. Kent WJ, Zweig AS, Barber G, Hinrichs AS, Karolchik D: **BigWig and BigBed: enabling browsing of large distributed datasets.** *Bioinformatics* 2010, **26:**2204-2207.

61. Krakau S, Richard H, Marsico A: **PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data.** *Genome Biol* 2017, **18:**240.

62. König J, Zarnack K, Luscombe NM, Ule J: **Protein-RNA interactions: new genomic technologies and perspectives.** *Nat Rev Genet* 2012, **13:**77-83.

63. Chakrabarti AM, Haberman N, Praznik A, Luscombe NM, Ule J: **Data Science Issues in Studying Protein–RNA Interactions with CLIP Technologies.** *Annual Review of Biomedical Data Science* 2018, **1:**235–261.

64. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from RNA-seq data.** *Genome Res* 2012, **22:**2008-2017.

65. Ramirez F, Ryan DP, Gruning B, Bhardwaj V, Kilpert F, Richter AS, Heyne S, Dundar F, Manke T: **deepTools2: a next generation web server for deep-sequencing data analysis.** *Nucleic Acids Res* 2016, **44:**W160-165.

66. Anders S, Pyl PT, Huber W: **HTSeq--a Python framework to work with high-throughput sequencing data.** *Bioinformatics* 2015, **31:**166-169.

67. Audic S, Claverie JM: **The significance of digital gene expression profiles.** *Genome Res* 1997, **7:**986-995.

Figure legends

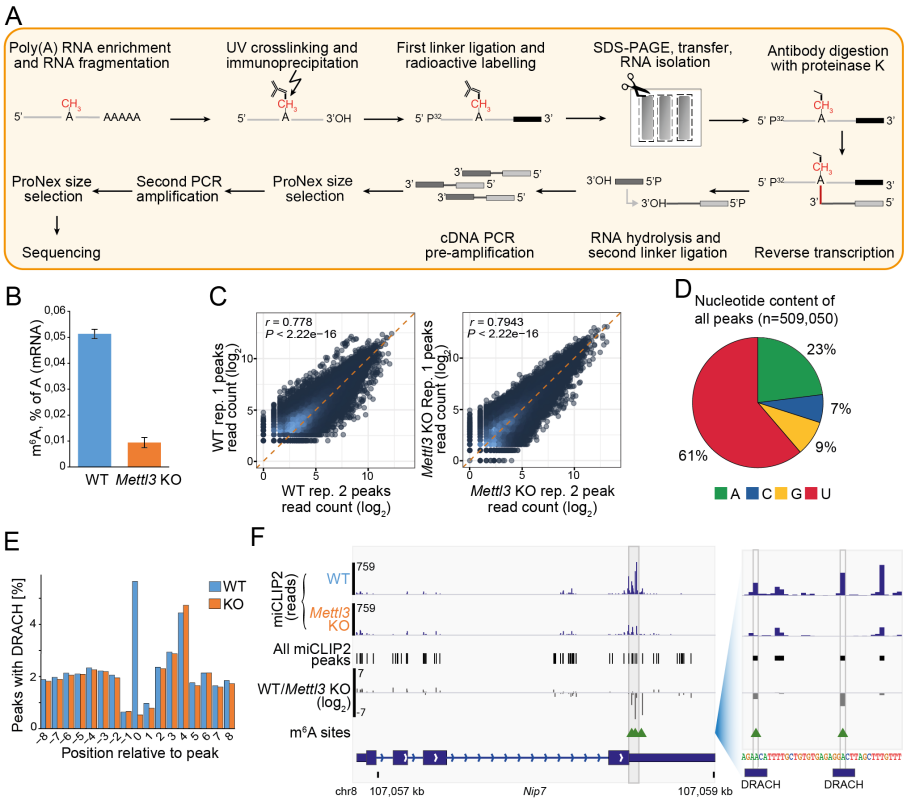

Figure 1. The optimised miCLIP2 protocol produces high complexity libraries with high reproducibility. *A.* An overview of the miCLIP2 protocol. ***B.*** mESC *Mettl3* KO cells show a significant depletion of m^{6}A on mRNAs. m^{6}A levels measured by liquid chromatography-tandem mass spectrometry (LC-MS/MS) for poly(A)+ RNA from WT and *Mettl3* KO mESCs. Quantification of m^{6}A as percent of A in RNA. Error bars indicate standard deviation of mean (s.d.m.), n = 3. ***C.*** miCLIP2 data are highly reproducible between replicates. Pairwise comparison of the miCLIP2 truncation reads within peaks from two miCLIP2 replicates from WT and *Mettl3* KO mESCs. Pearson correlation coefficients (*r*) and associated *P* values are given. Additional replicates are shown in **Figure S2A**. ***D.*** Most peaks are located at uridines and adenines. Pie chart representing the nucleotide distribution of all miCLIP2 WT peaks. ***E.*** DRACH motifs are enriched at miCLIP2 WT peaks. Metaprofile of DRACH motifs around aligned miCLIP2 peaks (position 0). Percentage of DRACH motifs (counted at position of A within DRACH) around the miCLIP2 peaks of WT and *Mettl3* KO mESCs are shown. ***F.*** *Mettl3* KO miCLIP2 signal is reduced at specific positions in the *Nip7* 3' UTR. Genome browser view of miCLIP2 data (blue) from WT and *Mettl3* KO mESCs and fold change between conditions (grey). Identified miCLIP2 peaks (black bars) and m6Aboost-predicted m^{6}A sites (green arrowheads) are given. A zoom-in (right) shows a more detailed view of two m^{6}A sites.

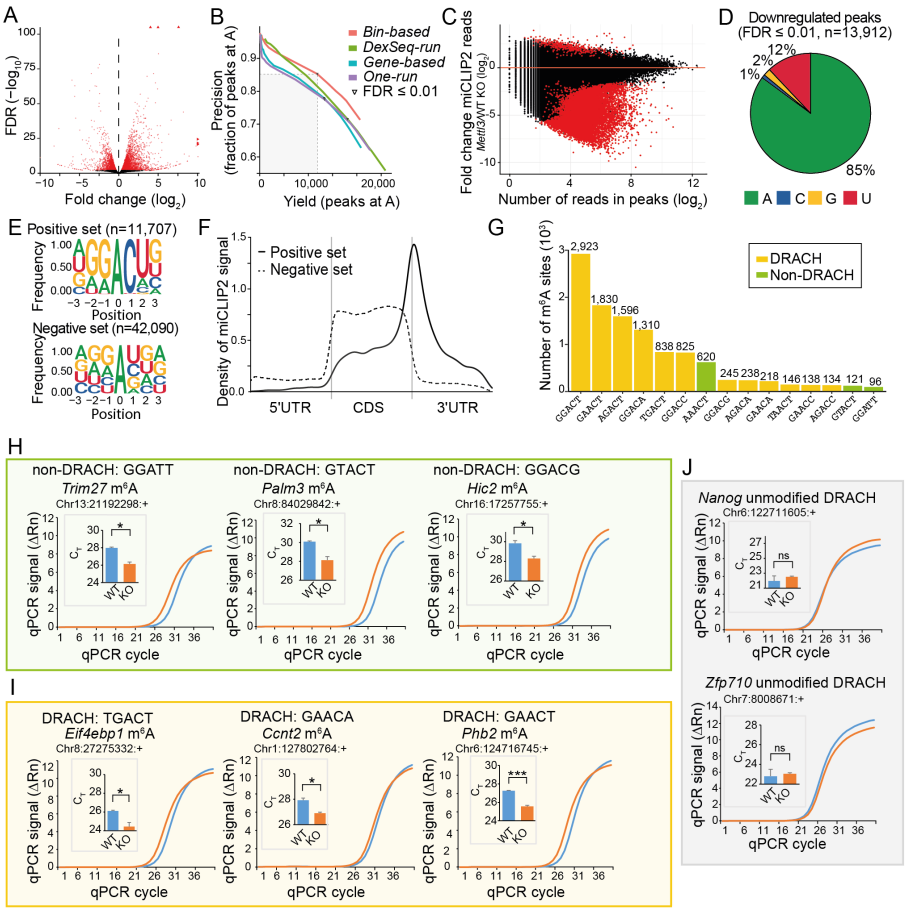

Figure 2. Differential peak analysis allows to identify true m^{6}A sites from miCLIP2 data. ***A.*** *Mettl3* KO causes drastic changes in gene expression. Volcano plot shows $\log_2$-transformed fold change (log2FC) of gene expression between WT and *Mettl3* KO against $\log_{10}$-transformed false discovery rate (FDR). Significantly changing genes are highlighted in red (FDR ≤ 0.01). ***B.*** The *bin-based* approach for differential methylation analysis outperforms other tested strategies. Number of identified peaks at A (x-axis) and fraction of peaks at A (y-axis) are given for different approaches (see **Supplemental Material**, Section A). Curves were generated by step-wise increases in stringency (FDR). FDR ≤ 0.01 is marked for each approach. ***C.*** Most changing peaks go down upon *Mettl3* KO. Comparison of log2FC in miCLIP2 signal per peak between WT and *Mettl3* KO (y-axis) against reads per peak ($\log_2$-transformed, x-axis). Significantly regulated peaks are highlighted in red (|log2FC| > 1, FDR ≤ 0.01). ***D.*** Most significantly downregulated peaks are located at adenosines. Pie chart represents nucleotide distribution of downregulated peaks. ***E.*** Sequence motifs of peaks in the positive (top) and negative (bottom) set. Logos show relative frequency of nucleotides at positions -3 to +3 around central A. ***F.*** Peaks in the positive set accumulate around stop codons. Density plot shows distribution of peaks in scaled transcript regions. UTR, untranslated region, CDS, coding sequence. ***G.*** The most frequent pentamers include non-DRACH motifs. Number of peaks (positive set) located at specific pentamer at DRACH (orange) and non-DRACH (olive) motifs. ***H-J.*** Selected m^{6}A sites were validated

by SELECT experiments. Exemplary real-time fluorescence amplification curves (normalised reporter value, ΔRn) and quantifications of threshold cycle (C_T) values (technical replicates) for mESC WT versus *Mettl3* KO samples are shown for m^{6}A sites at non-DRACH (H) and DRACH (I) motifs as well as unmodified DRACH motifs with a miCLIP2 peak (J). Neighbouring unmodified A nucleotides as control for each tested site are given in **Figure S3D**. *** *P* value < 0.001, ** *P* < 0.01, * P < 0.05, ns, not significant, two-sided Student's t-test, n=3.

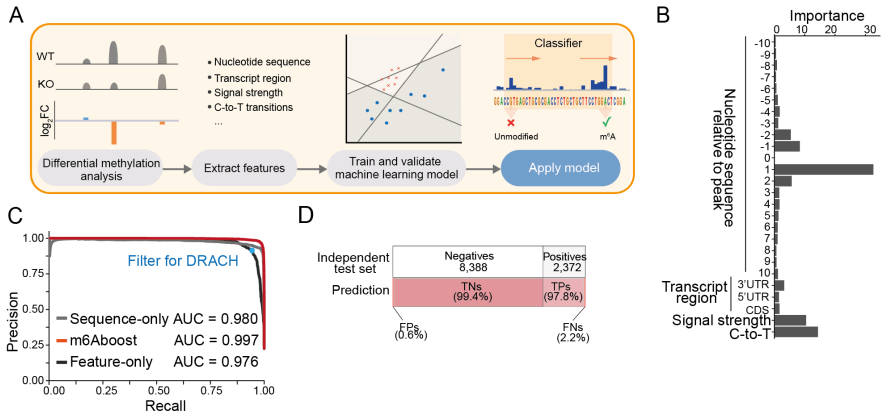

Figure 3. The machine learning classifier m6Aboost reliably predicts m^{6}A sites from miCLIP2 data. *A.* Overview of building the machine learning classifier. First, miCLIP2 WT and *Mettl3* KO datasets are analysed for differential methylation to identify Mettl3-dependent m^{6}A sites. The resulting positive and negative sets are used to extract features and train a machine learning model. The model is validated on an independent test set. Finally, the model can be applied to new miCLIP2 datasets to classify the miCLIP2 peaks as modified m^{6}A sites versus unmodified background signal. ***B.*** Highest informative content lies in the nucleotide sequence, the relative signal strength of the peak and the number of C-to-T transitions. Bar plot shows the features used for m6Aboost prediction and their associated importance ranking. UTR, untranslated region, CDS, coding sequence. ***C.*** m6Aboost outperforms baseline models trained only on sequence (sequence-only) or experimental features (feature-only). Precision-recall curve shows performance of m6Aboost compared to baseline models with the corresponding area under the curve (AUC). Precision and recall when solely filtering for DRACH motifs is shown for comparison (blue dot). ***D.*** m6Aboost achieves 99% accuracy on an independent test set. Bars visualise composition of independent test set (n=10,760) from positive (22%) and negative (78%) peaks (top) and the resulting m6Aboost predictions (bottom). In total, 10,658 peaks (99%) were correctly predicted, while 102 peaks were misclassified. TNs, true negatives, TPs, true positives, FNs, false negatives, FPs, false positives.

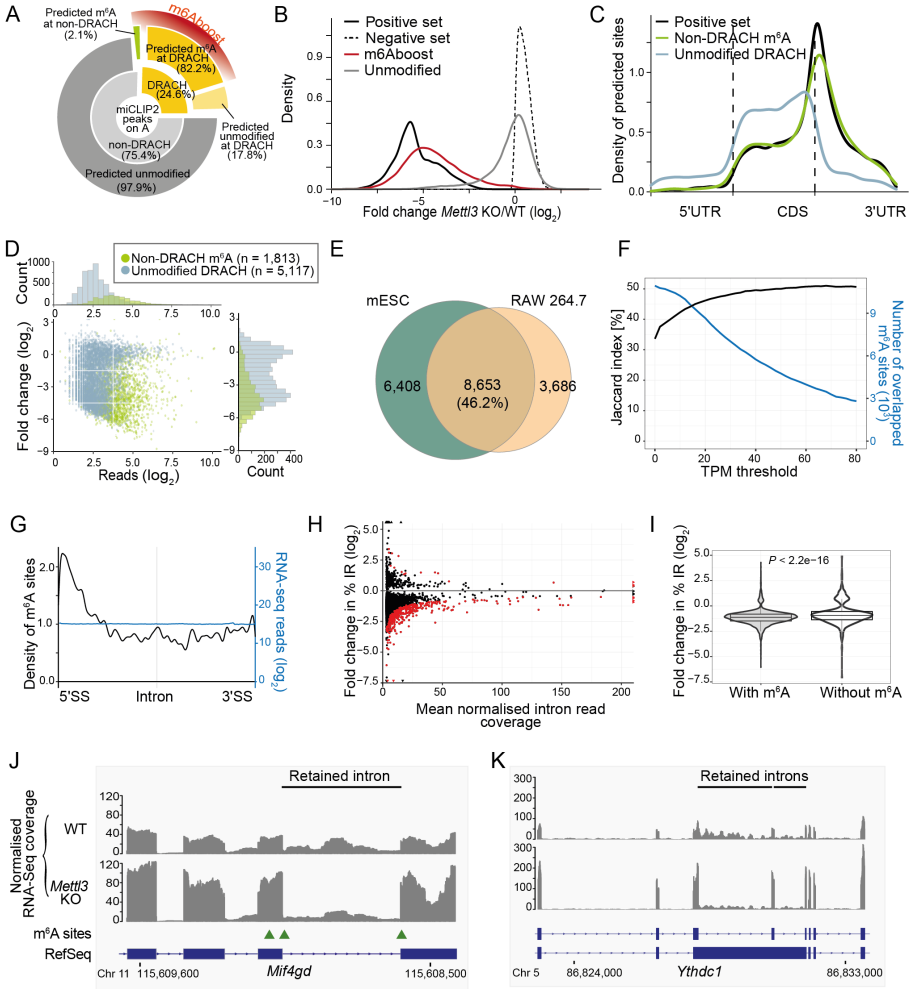

Figure 4. m^{6}A sites occur at non-DRACH motifs and accumulate in retained introns. ***A.*** m6Aboost predicts m^{6}A sites at DRACH and non-DRACH motifs in mESC WT miCLIP2 data. Inner circle of donut chart shows occurrence of DRACH (n=28,760, 24.6%) and non-DRACH (n=88,382, 75.4%) motifs for all miCLIP2 peaks at A. Outer circle shows m6Aboost prediction results (marked in red) with 23,643 m^{6}A sites and 5,117 unmodified sites at DRACH (82.2% and 17.8%, respectively, of all peaks at DRACH) as well as 1,813 m^{6}A sites and 86,569 unmodified sites at non-DRACH (2.1% and 97.9%, respectively, of all peaks at non-DRACH). ***B.*** Predicted m^{6}A sites (n=25,456) go down upon *Mettl3* KO, whereas predicted unmodified sites (n=91,686) remain unchanged. Density plot shows distribution of $\log_2$-transformed fold changes in miCLIP2 signal between *Mettl3* KO and WT samples. Positive and negative set are shown for comparison. ***C.*** m^{6}A sites at non-DRACH motifs (n=1,813) show a similar accumulation at stop codons as the positive set. Visualisation as in Figure 2F. ***D.*** m6Aboost predicts that not all peaks at DRACH motifs are m^{6}A sites. Scatter plot and associated histograms show fold change in miCLIP2 signal ($\log_2$-transformed, y-axis) against number of reads per peak ($\log_2$-transformed, x-axis) for 5,117 peaks at DRACH motifs (light blue) that are predicted to be unmodified by m6Aboost. m^{6}A sites at non-DRACH

motifs (olive) are shown for comparison. ***E.*** Most m^{6}A sites are shared between two different mouse cell lines. Venn diagram shows overlap of predicted m^{6}A sites in expressed genes (TPM ≥ 20, n=4,490) from mESC WT and RAW 264.7 cells. A Venn diagram without expression filter is shown in **Figure S5C**. ***F.*** The overlap of m^{6}A sites between two different mouse cell lines increases in higher expressed genes. TPM threshold representing the gene expression (x-axis) against the Jaccard index (y-axis). Number of overlapping m^{6}A sites are shown as comparison (blue). ***G.*** m^{6}A sites accumulate towards the 5' splice sites of introns. Metaprofile shows density of m^{6}A sites along scaled introns (n=3,509 m^{6}A sites on 1,465 different Introns). Coverage of RNA-seq reads on the same introns is shown for comparison (blue). SS, splice site. ***H.*** Intron retention (IR) is globally reduced in the *Mettl3* KO cells. Scatter plot shows fold change in relative IR (%IR, y-axis) against mean normalised RNA-seq reads on the introns across all samples (x-axis) for 4,925 measured IR events. 401 significantly changed IR events are highlighted in red (FDR ≤ 0.05). ***I.*** Introns harbouring m^{6}A sites show a significant trend towards more IR reduction. Violin plot compares distribution of fold changes in %IR for retained introns with (n=4,098) and without (n=827) m^{6}A sites. *P* value < 2.22e-16, Wilcoxon rank-sum test. ***J,K.*** IR is reduced in the genes *Mif4gd* (*J*) and Yrhdc1 (*K*). Genome browser view of RPGC-normalised RNA-seq read coverage tracks are shown for merged replicates from WT and *Mettl3* KO mESCs. Predicted m^{6}A sites are indicated with green arrowheads.

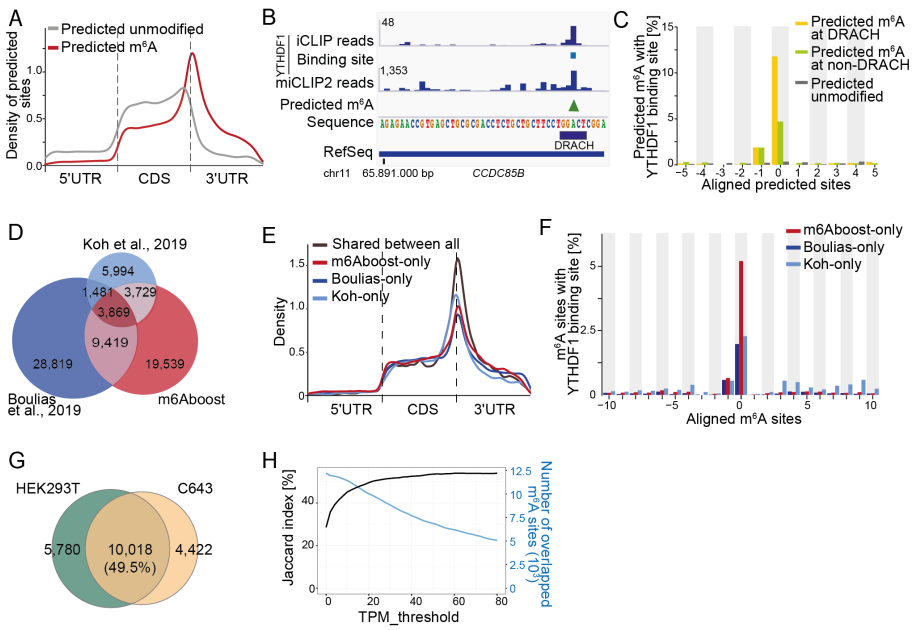

Figure 5. m6Aboost predicts 36,556 m^{6}A sites from HEK293T miCLIP2 data. *A.* Predicted m^{6}A sites are located around the stop codon. Visualisation as in Figure 2F. ***B.*** Predicted m^{6}A sites overlap with YTHDF1 binding sites. Genome browser view of a predicted m^{6}A site in the gene *CCDC85B* shows crosslink events from published YTHDF1 iCLIP data for HEK293T together with miCLIP2 signal (merge of four replicates) and m6Aboost-predicted m^{6}A site (green arrowhead) from our HEK293T miCLIP2 data.

C. YTHDF1 precisely binds at the predicted m^{6}A sites. Percentage of m^{6}A sites (position 0) with YTHDF1 binding sites (y-axis) in a 21-nt window are given for predicted m^{6}A sites at DRACH (yellow) and non-DRACH (green), as well as predicted unmodified sites (grey). ***D.*** Predicted m^{6}A sites from HEK293T miCLIP2 overlap with published m^{6}A data. Venn diagram shows single-nucleotide overlap with miCLIP and m6ACE-seq data (m^{6}A antibody by Synaptic Systems and Abcam, respectively). Note that m^{6}A sites in Boulias et al, 2019 had been filtered for DRACH motifs. ***E,F.*** Analysis of m^{6}A sites that are unique to one of the three datasets compared in (D). ***E.*** Unique m^{6}A sites accumulate around stop codons. Visualisation as in Figure 2F. ***F.*** Unique m^{6}A sites are enriched in YTHDF1 binding sites. Visualisation as in (C). ***G.*** Most m^{6}A sites are shared between two different human cell lines. Venn diagram shows overlap of predicted m^{6}A sites in expressed genes (TPM ≥ 20, n=3,298) from HEK293T and C643 cells. A Venn diagram without expression filter is shown in **Figure S5C**. ***H.*** More m^{6}A sites are shared between two different mouse cell lines in higher expressed genes. TPM threshold representing the gene expression (x-axis) against the Jaccard index (y-axis). Number of overlapping m^{6}A sites are shown as comparison (blue).

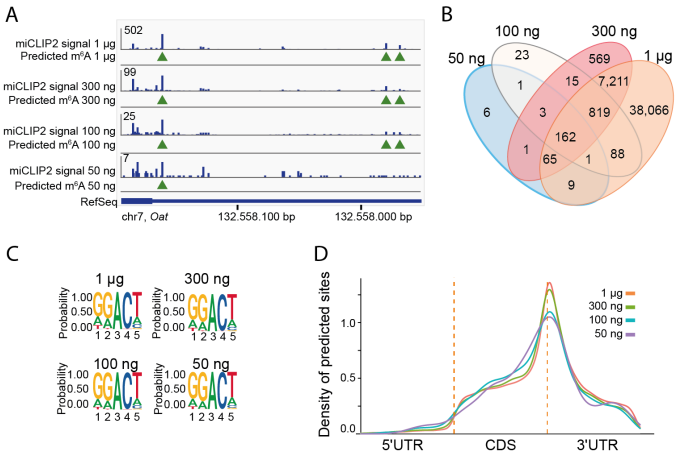

Figure 6. miCLIP2 allows to map m^{6}A sites from low input material *A.* m6Aboost predicts overlapping m^{6}A sites from miCLIP2 data for different RNA input concentrations. Example genome browser view of the *Oat* gene shows miCLIP2 signals and corresponding m6Aboost predictions (green arrowheads) for 1 µg, 300 ng, 100 ng and 50 ng of input RNA. ***B.*** The majority of m^{6}A sites predicted from low-input libraries overlap with 1 µg input library. Overview of the overlap of predicted m^{6}A sites from different concentrations. ***C.*** All predicted sites from different concentrations resemble a DRACH motif. Sequence logo of the predicted m^{6}A sites from miCLIP2 from different RNA input concentrations including the surrounding four nucleotides. ***D.*** Predicted m^{6}A sites from miCLIP2 with different RNA input concentrations cluster around the stop codon. Visualisation as in Figure 2F.