Direct RNA sequencing enables m^{6}A detection in endogenous transcript isoforms at base specific resolution

Daniel A. Lorenz[1,2,3,4], Shashank Sathe[1,2,3,4], Jaclyn M. Einstein[1,2], and Gene W. Yeo[1,2,3*]

[1] Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, CA, USA

[2] Stem Cell Program, University of California San Diego, La Jolla, CA, USA

[3] Institute for Genomic Medicine, University of California San Diego, La Jolla, CA, USA

[4] These authors contributed equally

* To whom correspondence should be addressed. Email: geneyeo@ucsd.edu

Keywords: m^{6}A, Nanopore, RNA Modifications

ABSTRACT

Direct RNA sequencing holds great promise for the *de novo* identification of RNA modifications at single coordinate resolution, however interpretation of raw sequencing output to discover modified bases remains a challenge. Using Oxford Nanopore's direct RNA sequencing technology, we developed a Random Forest classifier trained using experimentally detected N6-methyladenosine (m6A) sites within DRACH motifs. Our software MINES (m6A Identification using Nanopore Sequencing) assigned m6A methylation status to over 13,000 previously unannotated DRACH sites in endogenous HEK293T transcripts and identified over 40,000 sites with isoform-level resolution in a human mammary epithelial cell line. These sites displayed sensitivity to the m6A writer, METTL3, and eraser, ALKHBH5, respectively. MINES (https://github.com/YeoLab/MINES.git) enables long-read direct RNA-seq to m6A annotation at single coordinate-level resolution.

INTRODUCTION

Since the identification of the first RNA modification over 60 years ago, over 100 different RNA modifications have been identified (Davis and Allen 1957; Jonkhout et al. 2017). These RNA modifications are capable of imparting new or altered functions in RNA and have since been collectively termed the epitranscriptome (Saletore et al. 2012). One of the most common modifications in the eukaryotic transcriptome is N^6-methyladenosine (m^6A) which is found in most classes of RNA, including mRNA, ncRNA, rRNA, and tRNAs (Deng et al. 2018; Ma et al. 2018). With the development of antibodies that recognize m^6A and coupling to high-throughput sequencing technologies, several transcriptome-wide approaches to identifying m^6A sites have been developed (Grozhik and Jaffrey 2018). These techniques have been useful in demonstrating that m^6A plays important roles in nearly every aspect of biology from yeast to mammals (Yue et al. 2015).

Biochemical studies have revealed a complex network of proteins that are involved in writing, reading, and erasing m^6A methylation. In humans, current evidence suggests that a complex, composed of proteins METTL3, METTL14 and WTAP, is responsible for installing the m^6A modification in most mRNAs (Liu et al. 2014). These sites are then recognized by several families of proteins including YTH-domain containing, IGF2BP (IMPs), and HNRNP proteins, each having uniquely characterized roles in reading m^6A, influencing processes such as splicing, transcript stability, and localization (Shi et al. 2019). M^6A modification is a dynamic process and can be removed or “erased” by demethylases, ALKBH5 and FTO. Dysregulation of any of these critical proteins results in changes to m^6A levels and has been linked to a myriad of diseases, including cancer and neurological diseases (Delaunay and Frye 2019; Chen et al. 2019).

While second-generation polymerase-based sequencing have enabled transcriptome-wide studies of RNA biology, new third-generation sequencing, are being developed to overcome limitations such as amplification biases, lack of single-molecule sensitivity and isoform ambiguity.

One of these methods, commercialized by Oxford Nanopore Technologies (ONT), uses nanopore-based sequencing to detect changes in current as a single strand of nucleic acid sequence transverses a pore protein. By deconvoluting these electrical signals, the specific nucleotide sequence can be reconstructed. This technology offers long sequencing reads (up to 2Mb) and detection of epigenetic markers (Payne et al. 2019). To illustrate, nanopore-based DNA sequencing has been able to detect the endogenous DNA modifications, m5C and m^6A (Simpson et al. 2017; McIntyre et al. 2019). Recently, Oxford Nanopore reported the first RNA sequencing method capable of directly sequencing individual RNA strands while preserving epitranscriptomic information using fully modified *in vitro* transcribed RNAs, however, single molecule detection remains problematic due to the ~90% single base accuracy (Garalde et al. 2018).

Here we evaluate the ability of nanopore-based sequencing to directly detect m^6A RNA modifications in endogenous transcripts, providing numerous benefits over traditional methodologies including single-coordinate level resolution, isoform-specific context, single experimental pipeline, and simplified bioinformatic detection. Based on changes observed in current signal from each site, MINES is able to predict known m^6A CLIP-seq sites with ~80% accuracy within certain DRACH sequences which represent ~35% of reported CLIP sites. When applied to RNA from a primary human mammary epithelial cell line, MINES identified 42,116 m^6A sites at single-coordinate and isoform level resolution. As nanopore-based sequencing becomes ubiquitous in RNA-seq studies, our approach will facilitate new discoveries regarding m^6A biology and serves as a useful framework for analyzing other RNA modifications using direct RNA sequencing.

RESULTS

DRACH Filtering is Required for *De Novo* Detection

Nanopore-based sequencing is distinct from polymerase-based sequencing in that it can preserve and detect nucleic acid modifications as a single strand of nucleic acids passes through a pore (**Figure 1A**). With the advent of commercially available direct RNA sequencing, we sought to detect one of the most abundant RNA modifications, m^6A, on cellular transcripts. A recent study suggests direct sequencing can distinguish fully modified m^6A sites in pure populations of synthetic RNAs from unmodified positions (Garalde et al. 2018). However, these recent methods are limited by the computational resources necessary to detect changes in raw current on a transcriptome-wide scale and have not yet been utilized to identify new endogenous m^6A sites (Garalde et al. 2018; Workman et al. 2018; Liu et al. 2019). Contemporaneously, software applications, such as ONT's Tombo, enable detection of RNA modifications *de novo* by calculating the difference (or error) between the observed current and a ground truth provided by the reference genome and determining a modification value. The fraction modification value is stored as site averages instead of a per read value to reduce the computational load. However, a challenge associated with all nanopore based approaches centers around a 1:13 error rate (Depledge et al. 2019). Hence, relying solely on the "error detection" of *de novo* predictions from Tombo is unreliable at this time and prevents accurate single molecule detection. This is highlighted in **Figure 1B**, with many sites exhibiting aggregate, black bars, and molecule specific, black stars, deviations from the expected current values. To overcome this limitation while simultaneously maintaining a low computational burden, we reasoned that filtering nanopore data based on the known m^6A DRACH motifs would be a pragmatic strategy for m^6A detection. By limiting our algorithm to DRACH sites, we improve the likelihood that our predictions are specific to m^6A sites and not to other mRNA modifications. Analysis of two site-specific m^6A cross-linking and immunoprecipitation sequencing (CLIP-seq) datasets from HEK293T and HeLa cells (Linder et al. 2015; Ke et al. 2017) revealed that more than 50% and 80% of sites were located within

DRACH sequences, respectively (**Figure 1C**). Deeper analysis revealed that the most common pentamers present within the DRACH motif in both datasets is GGACT, with 6 sequences (AGACT, GAACT, GGACA, GGACC, GGACT, TGACT) representing >50% of CLIP sites within DRACH sequences (**Figure 1D**). Thus, our strategy of pre-filtering nanopore reads to reduce the computational load still encompass the vast majority of m^6A sites.

Nanopore Sequencing Distinguishes m^6A Within DRACH Motifs

To evaluate the utility of our strategy, we sequenced poly-A selected RNA from HEK293T cells. Reads were aligned to the human hg19 reference genome. It should be noted that using a genomic reference in Tombo will currently only yield coverage along the 3' untranslated regions (UTRs) as Tombo aligner is not splice-aware. Hence, our initial analyses were limited to alignments within the 3'UTR but still comprises >40 percent of known m^6A sites (Yue et al. 2015; Linder et al. 2015). This limitation can be surpassed by using a cDNA reference. From Tombo's *de novo* detection algorithm we collected the fraction modification values for all genomic positions within 3' UTRs. The current pore protein used by Oxford Nanopore detects a ~5 base pair window. We therefore extended our input window to 20 base pairs centered on the "A" in the DRACH motifs to ensure detection of the site and flanking regions. Each window was labeled with a ground truth based on whether the mid-point site was found overlapping any site within the m^6A -CLIP datasets. We required that each window must have a minimum read coverage of 5 reads, due to the error rate at low coverage loci. Even with this filtering, output fraction modified values averaged around 0.5 across all windows. The aggregate modification value was obtained for each coordinate within each window and a spike in signal value was observed at positions 1 through 3 upstream of the GGACT motif compared to a randomly selected background (**Figure 2A and 2B**). A similar spike was observed for AGACT, GGACA, and GGACC motifs as seen in **Supplemental Figure 1**, along with other DRACH motifs. Encouraged by a significant difference between sites with CLIP evidence relative to non-CLIP sites, we sought to confirm that the spike in signal was

indeed due to m^{6}A. To accomplish this, we generated a HEK293T cell line stably expressing a shRNA that successfully depletes METTL3 protein (**Figure 2C-E**) and sequenced poly-A RNA with ONT. METTL3 depletion had a greater effect on m^{6}A levels in total RNA relative to the poly-A fraction (**Supplemental Figure 2**). A decrease in peak intensity was observed in the METTL3 shRNA cell line along the corresponding positions of the modified sites identified in the WT cell line. However, a similar change was not observed for randomly selected non-DRACH sites, indicating that the peak is indeed a result of the m^{6}A methylation status (**Figure 2A and B**). The METTL3 shRNA cell line also served as a validation for the sites identified in the WT cell line, independent of CLIP-based methods. Intriguingly, we found a similar decrease in peak intensity in both CLIP and non-CLIP sites, suggesting that there was a significant number of additional m^{6}A sites that were likely undetected within the previous CLIP datasets.

Random Forest Model Predicts m6A sites

After confirming that ONT is able to detect m^{6}A sites that were novel as well as ones previously found by CLIP-based methods, we elected to use a Random Forest Model (RFM) to predict methylation sites *de novo* (Pedregosa et al. 2011). The RFM was trained using 70% of the CLIP sites (positive labels) and an equal number of non-CLIP sites (n=1,515 for GGACT) as negative examples. The remaining 30% of CLIP sites were reserved as test examples. The test data also contained the remaining non-CLIP sites that were not included in the training data set. Since nanopore sequencing shows a unique sensitivity for each 5mer, we generated a separate model for each 5mer within the DRACH motif. We generated 10 models per DRACH motif based on random samples of training data and stored the model with the highest accuracy. Final accuracy values, defined as correctly predicted CLIP sites in the test data, ranged from 67-83%, while the precision values ranging from 40-92% (**Figure 3A, Supplemental Table 1**). Area under the curve (AUC) values ranged from 0.54 to 0.76; however, we believe these values were negatively affected by the presence of novel, non-CLIP m^{6}A sites (true negatives) within the test data set

(**Figure 2A, 3B,** and **Supplemental Figure 3**). Of the 18 DRACH motifs, only 4 generated models with accuracy > 0.7, precision values > 0.85, and ROC AUC values > 0.67. Combining the four top motifs, the average accuracy was 79% which represents >35% of known (CLIP-based) m^6A sites (**Figure 3C**). Interestingly, RFMs from motifs not meeting our accuracy, precision and ROC AUC standards also clearly failed to exhibit a decrease in signal in the METTL3 knockdown dataset at m^6A -CLIP sites (**Supplemental Figure 1**). This either indicates that the current pore protein is incapable of distinguishing m^6A methylation in these motif contexts or that these sites could represent off-target antibody binding or exists in such low m^6A /A ratios that we are unable to detect their change in signal.

Detection of Novel m^6A Sites in HEK293

Having generated a nanopore-enabled m^6A detection algorithm, MINES, we evaluated the non-CLIP sites and predicted their methylation status. Of the 28,925 non-CLIP sites across AGACT, GGACA, GGACC, and GGACT motifs, MINES predicted that 13,034 are likely methylated (**Figure 3C**). Surprised by the number of potentially missed m^6A sites, we analyzed the mean modification values for these predicted sites in both wild-type and METTL3 knockdown **(Figure 4A, Supplemental Figure 4)**. As expected, these sites displayed a peak in modification values that significantly decreased under METTL3 knockdown. This is in concordance with the CLIP sites correctly identified within the test data (true positives). This effect was not observed in the sites predicted to be unmodified, irrespective of whether they were previously identified from the m^6A CLIP experiments (**Figure 4A, right panels**). All other 5mers can be found in **Supplemental Figure 4**. To further characterize the wild-type peak sites, we looked at their response to METTL3 depletion on a per site basis. A METTL3-sensitive site was defined as any site with a greater mean modification value at the wild-type peak positions over METTL3 depletion. **Figures 4B and C** show the fraction of predicted m^6A and non- m^6A sites sensitive to METTL3 depletion mimics that of the CLIP data with a breakdown of each category in **Figure 4D**. Thus, this provides more

evidence that MINES is correctly predicting m^{6}A sites, as the number of sites sensitive to METTL3 increases to a similar degree as the CLIP sites.

Cell-line Independent Detection and Validation by ALKBH5 Expression

To test whether our model is able to detect m^{6}A modified sites in other cell-lines, we sequenced poly-A RNA from a primary human mammary epithelial cell line (HMEC) and a derivative cell-line that stably over-expresses the m^{6}A eraser ALKBH5. Decreased m^{6}A levels due to ALKBH5 overexpression were confirmed by western and dot blot analyses (**Supplemental Figure 5A and B**). Here, we aligned sequencing reads to a human cDNA reference to ensure full transcript coverage and evaluated the ability of MINES to predict m6a in isoform-specific level. Using Tombo's coverage data and fraction modified values, and the RFMs generated for four motifs (AGACT, GGACA, GGACC, GGACT), MINES assigned m^{6}A status to 42,116 sites. Similar to the HEK293T and METTL3 knockdown results, the mean modification values for the HMEC m^{6}A sites (true positives) were lower in the ALKBH5 overexpression cell line (**Figure 5A and Supplemental Figure 6**) compared to randomly shuffled sites **(Supplemental Figure 5C)**. Some DRACH sequences produced altered modification patterns than those found in **Supplemental Figure 1**, however, these are limited to sequences in which the accuracy and precision were poor and are not included in the final versions of MINES. The fraction of individual sites sensitive to ALKBH5 also increased in the m^{6}A predicted fraction (**Figure 5B**), similar to METTL3 knockdowns. To further assess the accuracy of MINES, we studied the distribution of predicted m^{6}A sites across all transcript isoforms to resolve the density of m^{6}A sites within different genic regions including 5'UTR, CDS and 3'UTR respectively (**Figure 5C**). This analysis revealed the characteristic density peak at the start of the 3' UTR, confirming that our model resembles results seen in traditional m^{6}A-seq approaches(Linder et al. 2015; Ke et al. 2017).

To determine differential isoform-level methylation patterns, we converted the cDNA coordinates to genomic positions. Analysis of these genomic positions identified 2,225 genes to

have isoform-specific methylation patterns out of the 6,837 m6A-containing genes (**Figure 5D**). In total there were 78,592 distinct genomic locations analyzed by MINES with 21,309 of these positions covering multiple isoforms. Comparing the methylation status of these multiple isoform sites revealed 10,415 sites that were never predicted to be methylated, 4,726 sites predicted to be consistently methylated, and 6,168 sites with isoform-specific methylation (**Figure 5D**). As an example, we looked at three ACTB isoforms that were found in our nanopore sequencing data and predicted by MINES to have isoform-specific m^6A. The three isoforms (ENST00000331789, ENST00000425660, and ENST00000462494) had seven sites which met our read depth and sequence requirements (**Figure 5E**). Two of the transcripts (ENST00000331789 and ENST00000462494) were predicted to contain one m^6A site at genomic position chr7:5527743 (hg38). The third transcript (ENST00000425660) is not methylated at this position but was instead predicted to be methylated at chr7:5528125 (hg38). Intriguingly, this third transcript is also predicted by ENSEMBL annotation to be subject to nonsense mediated decay; however, future experiments would be required to link these events. It should be noted that this isoform-level resolution is only possible if a cDNA reference was used as input to TOMBO to perform the read alignment. Thus, MINES, for the first time, enables probing of m^6A biology with isoform-specific resolution.

DISCUSSION

While effective, m^6A CLIP- and RIP-seq techniques depend on the availability of high-quality antibodies, require longer library preparation times and tailored processing pipelines for analysis. Advances in third-generation sequencing approaches have enabled direct RNA sequencing while preserving endogenous modifications with a short and straightforward library preparation and isoform-specific detection. Taking advantage of this recent technology we developed an algorithm that uses only the standard data generated from an Oxford Nanopore sequencer as input and predicts m^6A modified sites in poly-A selected mRNA.

Coupling publicly available m^{6}A datasets and Tombo's modification values, we demonstrated a largely accurate detection of m^{6}A sites at positions 1 though 3 upstream of canonical DRACH motifs. Through the sequencing of a METTL3 knockdown cell line we showed that the modification value decreases at these previously reported sites while randomly selected background sites remain unaffected. This serves as an independent validation of our results. Interestingly, we observed a decrease in modification value in several non-CLIP sites upon METTL3 knockdown, indicating potentially unannotated m^{6}A sites. We then trained an RFM using the CLIP sites as positive controls and non-CLIP sites as negative controls. Four DRACH sequences (AGACT, GGACT, GGACC, and GGACA) generated models with maximum accuracy > 70% and precision > 85%, comprising >35% of known m6A sites. Using MINES to identify methylation sites within these sequences, we predicted a total of 13,034 m^{6}A sites in HEK293T cells. These newly identified sites exhibited similar modification values and sensitivity to loss of METTL3 to those found in previous datasets. Factoring in the low individual base accuracy and high computational burden of analyzing signal deviations for each RNA molecule, we elected to use average deviations for each site and therefore cannot accurately determine the percentage of reads methylated at a given site at this time. Additionally, this averaging could result in the loss of methylated sites with low m^{6}A /A ratios as small differences could be lost to background. As improvements to the pore protein are released in the future, MINES can be easily retrained to achieve single molecule level detection.

Next, we utilized MINES to identify and annotate 42,116 m^{6}A sites in a human mammary epithelial (HMEC) cell line. As supporting validation of these sites, we generated a cell line that overexpresses ALKBH5. These newly annotated sites showed a significant increase in ALKBH5 sensitivity over non-methylated sites, consistent with our results in the METLL3 depletion in HEK293. These new sites also mimic the distribution of m^{6}A sites in other cell types with a characteristic peak at the beginning of the 3' UTR, immediately following the stop codon. Using cDNA alignments, MINES was able to predict m^{6}A methylation in an isoform-specific manner for

2,225 genes (**Figure 5D**), illustrated in **Figure 5E** with ACTB. Thus, we are confident in MINES' ability to annotate m^6A sites in any transcriptome with isoform level resolution using raw nanopore data as input. We envision this method and software to be readily adopted in the current m^6A detection field.

MATERIAL AND METHODS

Cell line Generation and Culture

HMECs expressing hTERT and tamoxifen inducible Myc-ER (Myc-ER-HMECs) were a gift from Trey Westbrook (Kessler et al. 2012). HEK293 and HMEC cell lines were cultured in DMEM supplemented with 10% FBS and Medium 171 supplemented with MEGS: S0155, respectively, following standard tissue culture practices. METTL3 shRNA plasmid (TRCN0000034717) was purchased from Sigma Aldrich. psPAX.2 and pMD2.g were a gift from Didier Trono (Addgene plasmids #12260, #12259). ALKBH5 was cloned from endogenous HMEC cDNA into doxycycline inducible pLIX403 with a C-terminal mRuby tag using gateway assembly. pLIX403 was a gift from David Root (Addgene plasmid #41395). All plasmids were confirmed with Sanger sequencing. Briefly, lentivirus was packaged in HEK293T cells by seeding 6-well plates at ~80% confluence. The following day the cells were transfected by combining 35 uL Opti-MEM, 5 uL P3000 reagent (both Thermo Fisher), 500ng psPAX.2, 50 ng pMD2.g, 500 ng shRNA/gene vector were combined. 15 uL Opti-MEM and 4 uL Lipofectamine 3000 (both Thermo Fisher) were mixed in another tube before being combined together and allowed incubate at room temperature for 20 minutes. This mixture was added to cells in a dropwise fashion. After 4-6 hours the media was replaced with fresh media. Media containing virus was harvested 48 and 72 hours post transfection. Viral particles were passed through a 0.45um sterile filter. Virus containing media was then added to HEK293T or HMEC cell lines supplemented with 8ug/mL polybrene. Media was removed after 24 hours and replaced with media containing 2ug/ml puromycin. ALKBH5

overexpression was induced with the addition of 1 ug/mL doxycycline to media for 48 hours before collecting cells.

Western Blots

Cell lysates were harvested at ~80% confluency by washing with phosphate-buffered saline (PBS), and addition of ~150uL lysis buffer (50 mM Tris-HCl pH 7.4, 100 mM NaCl, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate). Samples were sonicated, loaded on 4-12% Bis-Tris gel and transferred to PVDF membrane overnight at 30V 4 ºC. The membrane was then blocked with 5% non-fat dry milk powder in tris-buffered saline with 0.05% Tween-20 (TBST) for 1h, incubated with antibody (METTL3 - Proteintech #15073-1-AP, ALKBH5 - MBL #RN122PW, Actinin - Millipore #05-384, GAPDH- Abcam #ab8245) at 1:1000 dilution for 1h, washed 3x with TBST, incubated for 1h with HRP-conjugated anti-rabbit (Thermo Fisher #31460) or anti-mouse antibody (Thermo Fisher #31430) at 1:3000 dilution before being washed again 3x with TBST. Bands were visualized by enhanced chemiluminescence (Thermo Fisher #34096) and exposure to film.

RNA Isolation and PolyA selection

At 80% confluency in 10cm plates, cells were washed with PBS and harvested in 1mL of TRIzol reagent (Thermo Fisher) or Direct-zol kit with DNase treatment (Zymo Research). Total RNA was extracted following the manufacturer's protocol. 20ug of total RNA was poly-A selected using a poly-A magnetic resin kit (NEB E7490L). RNA was then analyzed by high-sensitivity RNA Tapestation (Agilent #5067-5579) to confirm poly-A selection and RNA quality.

m^{6}A Dot Blot

RNA was quantified prior to blotting using a Nanodrop spectrophotometer. 500 ng of RNA, unless otherwise noted, was then diluted to 100 uL in H_2O and spotted on a prewashed (100uL H_2O) nylon membrane (Hybond-XL, GE Healthcare) using a dot blot apparatus (Bio-Dot, Bio-Rad) and Lorenz et al. 12

washed with 100 uL of H_2O. RNA was then cross-linked to the membrane with a UV cross-linker fitted with 254nm bulbs at 120 mJ/cm^2. The membrane was processed and developed as described above, using an m^{6}A antibody (Synaptic Systems #202111) at 1:1000 dilution. After developing the membrane was washed 3x with TBST, methylene blue solution (0.04% methylene blue in 50 mM NaOAC, pH 5.0, Santa Cruz Biotechnology sc-215381) was added and allowed to rotate overnight. The following day the solution was removed, and membrane rinsed with 50% ethanol/water before being imaged. Dots were quantified by densitometry using ImageJ.

Nanopore Sequencing

500ng of poly-A selected RNA was used as input for the Nanopore direct RNA sequencing kit (SQK-RNA001 and 002). RNA was prepared following the manufacturer's protocol. Sequencing was carried out on an Oxford Nanopore Minion-101B using R9.4.1 flow cells for ~48 hours. Data was base called in real time using a Dell Precision 7820 Tower with either Albacore or Guppy base callers. Total reads (in millions) were HEK-WT=1.45, HEK-shMETTL3=1.1, HMEC-WT=2.14, HMEC-ALKBH5 overexpression=1.72.

Tombo Alignment and Values

Reads and modification values were aligned using the default *resquiggle* and *de novo* detection settings, respectively, in Tombo v1.4 with hg19 and GRCh38/hg38 references using either a genomic or a cDNA (transcriptomic) reference. Genomic reference (hg19) was downloaded from GENCODE and cDNA reference (GRCh38/hg38) was downloaded from ENSEMBL. WT HEK293T RNA was aligned to a custom hg19 reference containing an additional unique gene, reads mapping to this custom gene were not used. Values were obtained from the read coverage (bedgraphs) and fraction of modified reads (wiggle files) for each position within the reference.

M^{6}A site detection using Random Forest Models

Briefly, all regions within the reference containing a DRACH motif were identified and a new set of regions was generated by extending 10 bps on both sides of the "A" within the DRACH motifs. These regions were further filtered to have a minimum coverage of 5 reads. The DRACH regions were intersected with known m^6A sites to identify true positive regions obtained from GSA datasets GSM1556678 and GSM2300429 (REFs: PMID: 26121403, PMID: 28637692).

A Random Forest Classifier is Decision-Tree based classifier. The Python implementation of Random Forest (*sklearn)* was used to generate a model to predict m^6A sites from the filtered DRACH data. Since Nanopore data reflects the occurrence of a m^6A site with a change in aggregate modification values, we trained the Random Forest model on the change in corresponding modification values detected by Nanopore sequencing within each 20 bp window. Since Nanopore shows varying changes in pore current values according to the 5mer motif, we decided to build motif-specific models.

For each 5mer DRACH motif, we identified all occurrences of the motif within expressed transcripts. Using previously identified m6A sites (Linder et al. 2015; Ke et al. 2017), all occurrences of the motif were segregated into two groups of known and unknown sites. 70% of the known occurrences were used as training data, while the remaining 30% of the known occurrences were used as part of the testing data. To maintain an evenness within the training data, we added the same number of unknown occurrences to the training data. Remaining unknown occurrences were added to the testing data. A ground truth, the known m6A occurrence were considered as true m6A sites and the previously unidentified sites were considered as false m6A sites. Once the training and testing sites were identified, we extracted modification values for 10bp upstream and downstream of the 'A' within the DRACH motif. Each model was trained on these values for the given ground truth and then tested on corresponding values for the test sites.

Thus, we generated 18 RF models, each corresponding to one specific DRACH motif. Each model was trained using 10 different training datasets and the model with the highest training accuracy was selected for testing purposes. To confirm the training accuracy, each model was tested on a test dataset. To maintain the sanity of the validation, we ensured that the test datasets had not been run through the RF model in any capacity.

The purpose of the model is to identify novel m^6A sites, in addition to the known CLIP sites. We expected the accuracy of the model to be handicapped, since many of the previously unidentified DRACH sites would now be predicted as valid m^6A sites. Hence, the final accuracy of the model was determined as the accuracy of the model to detect previously known m^6A sites within the test dataset.

M6A METAGENE PLOTS

We used the metaPlotR package to plot metagene plots for m6A sites identified through MINES. MetaPlotR is a publicly available package (https://github.com/olarerin/metaPlotR) and has been previously used to perform similar analyses (Olarerin-George and Jaffrey 2017).

MINES

MINES (m6A Identification using Nanopore Sequencing) is a command line executable code that uses a compilation of the four Random Models, each corresponding to a DRACH motif, AGACT, GGACA, GGACC, and GGACT. MINES uses Tombo's fraction modified values and coverage files as inputs and outputs a bed file of predicted sites. Processing time for a full dataset is approximately 10 minutes. For more information visit [https://github.com/YeoLab/MINES.git].

AVAILABILITY

MINES source code is available at [https://github.com/YeoLab/MINES.git].

ACCESSION NUMBERS

Data files have been uploaded to GEO under accession number GSE132971.

SUPPLEMENTARY DATA

Supplementary Data are available at RNA online.

ACKNOWLEDGEMENT

The authors thank Julia Nussbacher and Kris Brannan for their help with HEK293T-shMETTL3 cell line generation and providing HEK293 RNA respectively.

Author contributions: D.A.L., S.S., J.M.E., and G.W.Y. contributed to the conception and design of the study. D.A.L. and J.M.E. performed the tissue culture, sample collection, and acquisition of data. D.A.L. and S.S. wrote the custom scripts and analyzed the data with input from G.W.Y. D.A.L., S.S. and G.W.Y. contributed to writing the manuscript.

FUNDING

This work was partially supported by grants from the National Institutes of Health [HG004659, HG009889 to G.W.Y, 2T32CA067754 to D.A.L.].

CONFLICT OF INTEREST

G.W.Y. is co-founder, member of the Board of Directors, on the SAB, equity holder, and paid consultant for Locana and Eclipse BioInnovations. G.W.Y. is a visiting professor at the National University of Singapore and receives travel reimbursement. The terms of this arrangement have

been reviewed and approved by the University of California, San Diego in accordance with its conflict of interest policies. The authors declare no other competing financial interests.

REFERENCES

Chen X, Yu C, Guo M, Zheng X, Ali S, Huang H, Zhang L, Wang S, Huang Y, Qie S, et al. 2019. Down-Regulation of m6A mRNA Methylation Is Involved in Dopaminergic Neuronal Death. *ACS Chem Neurosci* **10**: 2355–2363.

Davis FF, Allen FW. 1957. Ribonucleic acids from yeast which contain a fifth nucleotide. *Journal of Biological Chemistry* **227**: 907–915.

Delaunay S, Frye M. 2019. RNA modifications regulating cell fate in cancer. *Nature cell biology* **21**: 552–559.

Deng X, Su R, Weng H, Huang H, Li Z, Chen J. 2018. RNA N6-methyladenosine modification in cancers: current status and perspectives. *Cell research* **28**: 507–517.

Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, Mohr I, Wilson AC. 2019. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun* **10**: 754.

Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, Pantic N, Admassu T, James P, Warland A, et al. 2018. Highly parallel direct RNA sequencing on an array of nanopores. *Nature methods* **15**: 201–206.

Grozhik AV, Jaffrey SR. 2018. Distinguishing RNA modifications from noise in epitranscriptome maps. *Nature Chemical Biology* **14**: 215–225.

Jonkhout N, Tran J, Smith MA, Schonrock N, Mattick JS, Novoa EM. 2017. The RNA modification landscape in human disease. *RNA* **23**: 1754–1769.

Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vågbø CB, Geula S, Hanna JH, Black DL, Darnell JE, Darnell RB. 2017. m6A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes & Development* **31**: 990–1006.

Kessler JD, Kahle KT, Sun T, Meerbrey KL, Schlabach MR, Schmitt EM, Skinner SO, Xu Q, Li MZ, Hartman ZC, et al. 2012. A SUMOylation-Dependent Transcriptional Subprogram Is Required for Myc-Driven Tumorigenesis. *Science* **335**: 348–353.

Linder B, Grozhik AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. 2015. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nature methods* **12**: 767–772.

Liu H, Begik O, Lucas MC, Ramirez JM, Mason CE, Wiener D, Schwartz S, Mattick JS, Smith MA, Novoa EM. 2019. Accurate detection of m6A RNA modifications in native RNA sequences. *Nat Commun* **10**: 4079–9.

Liu J, Yue Y, Han D, Wang X, Fu Y, Zhang L, Jia G, Yu M, Lu Z, Deng X, et al. 2014. A METTL3-METTL14 complex mediates mammalian nuclear RNA N6-adenosine methylation. *Nature Chemical Biology* **10**: 93–95.

Ma H, Wang X, Cai J, Dai Q, Natchiar SK, Lv R, Chen K, Lu Z, Chen H, Shi YG, et al. 2018. N6-Methyladenosine methyltransferase ZCCHC4 mediates ribosomal RNA methylation. *Nature Chemical Biology* **15**: 88–94.

McIntyre ABR, Alexander N, Grigorev K, Bezdan D, Sichtig H, Chiu CY, Mason CE. 2019. Single-molecule sequencing detection of N6-methyladenine in microbial reference materials. *Nat Commun* **10**: 579.

Olarerin-George AO, Jaffrey SR. 2017. MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites ed. J. Hancock. *Bioinformatics* **33**: 1563–1564.

Payne A, Holmes N, Rakyan V, Loose M. 2019. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**: 2193–2198.

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**: 2825–2830.

Saletore Y, Meyer K, Korlach J, Vilfan ID, Jaffrey S, Mason CE. 2012. The birth of the Epitranscriptome: deciphering the function of RNA modifications. *Genome Biol* **13**: 175.

Shi H, Wei J, He C. 2019. Where, When, and How: Context-Dependent Functions of RNA Methylation Writers, Readers, and Erasers. *Molecular Cell* **74**: 640–650.

Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nature Publishing Group* **14**: 407–410.

Workman RE, Tang A, Tang PS, Jain M, Tyson JR, Zuzarte PC, Gilpatrick T, Razaghi R, Quick J, Sadowski N, et al. 2018. Nanopore native RNA sequencing of a human poly(A) transcriptome. 1–37. **PREPRINT - bioRxiv**

Yue Y, Liu J, He C. 2015. RNA N6-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes & Development* **29**: 1343–1355.

TABLE LEGENDS AND FIGURES

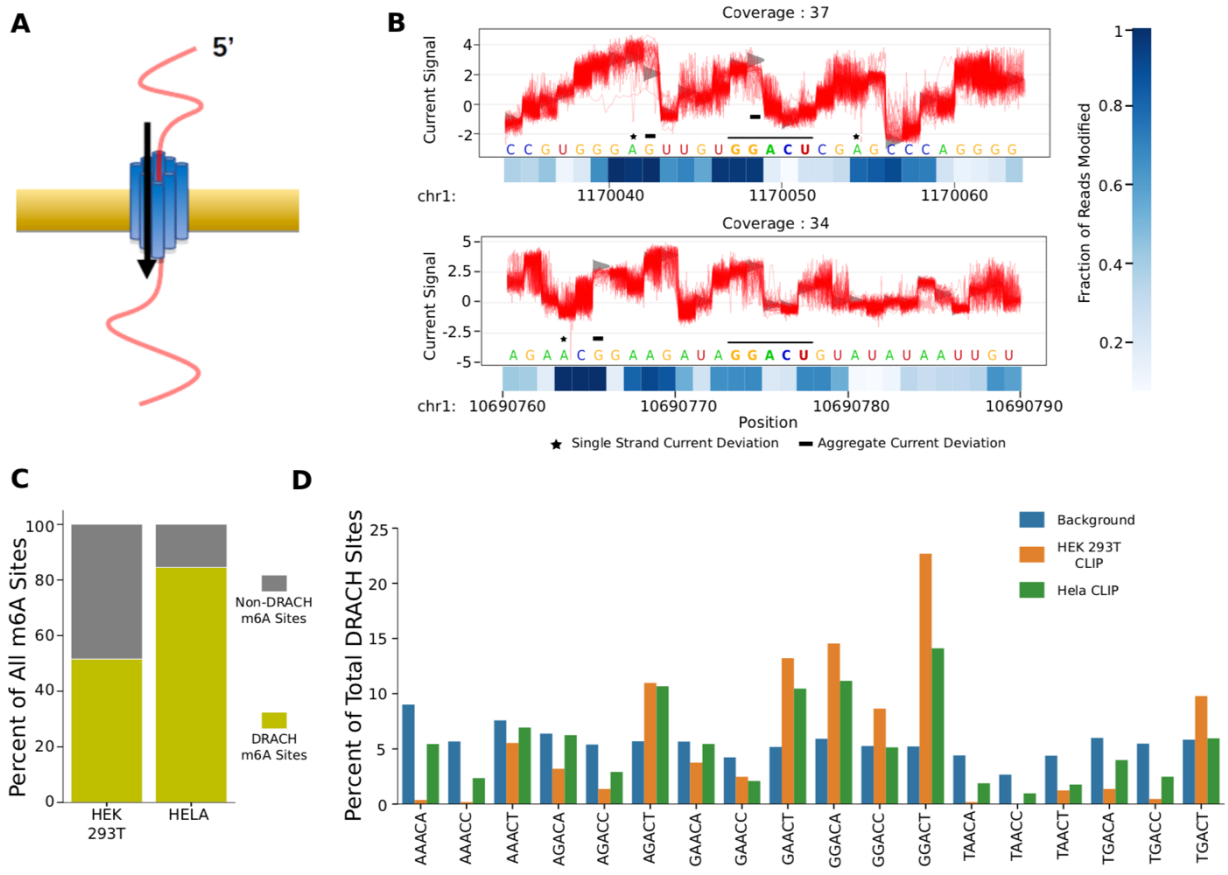


Figure 1. Filtering by DRACH motifs encompasses the majority of m^{6}A sites. (**A)** Schematic of Nanopore based sequencing. (**B**) Representative Tombo outputs depicting individual reads as red lines and expected values as grey distributions. Black bar in the middle highlights the GGACT motif. The heatmaps under each plot show Tombo's fraction modification value for each base. (**C**) Motif analysis of sites in HEK293T and HeLa cells from m^{6}A CLIP datasets. (**D**) Bars representing the percentage of each DRACH motif in m^{6}A CLIP and its relative enrichment over non-CLIP sites.

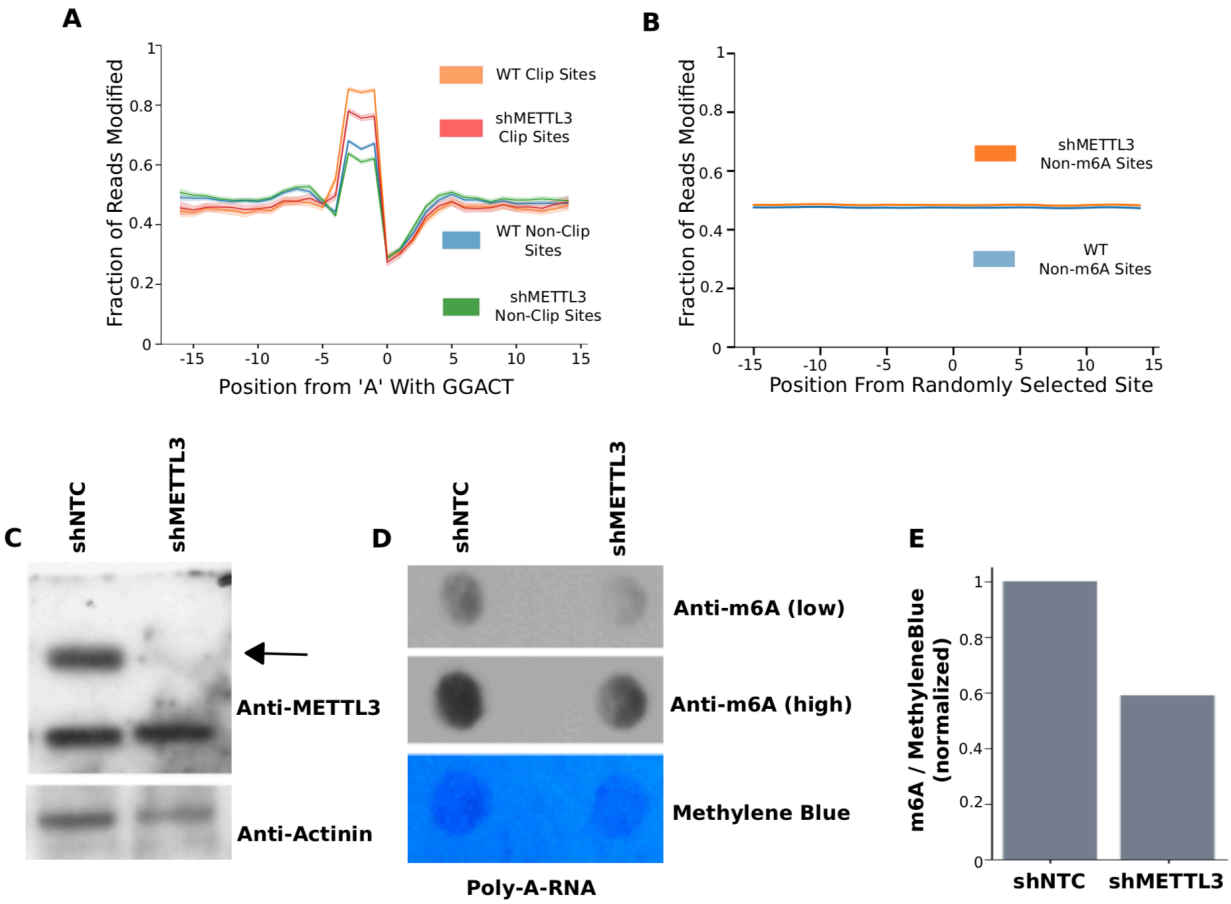


Figure 2. Nanopore sequencing can detect endogenous m^{6}A. (**A**) Line plots depicting the mean Tombo's fraction modified value across a 30-nucleotide window centered on the "A" in GGACT across all sites in RNA form HEK293T or shRNA targeting METTL3 (shMETTL3) cells. (**B**) Line plots of Tombo's fraction modified values across shuffled non-DRACH sites. (**C**) Western blot showing knockdown of METTL3 relative to non-targeting controls. Black arrow indicates expected METTL3 molecular weight. (**D**) m^{6}A dot blot of poly-A RNA from HEK293T cells treated with shNTC or shMETTL3. Methylene blue was contrast adjusted to highlight dots (**E**) ImageJ quantification and normalization of (D).

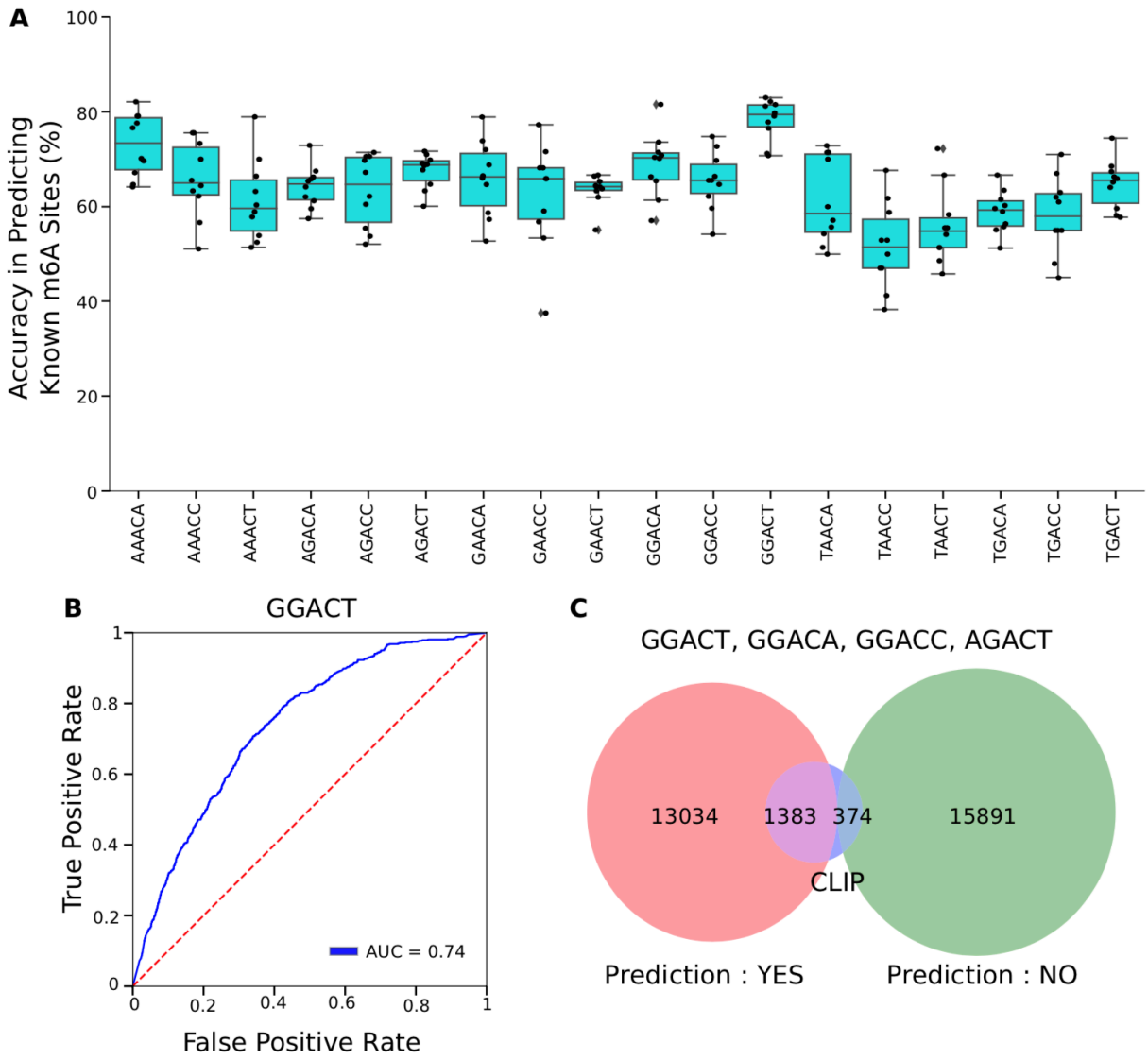


Figure 3. A trained RFM accurately predicts m^{6}A within DRACH motifs (**A**) Box plots showing the model's accuracy of predicting CLIP sites, organized by each DRACH motif across 10 training runs. (**B**) ROC curve for GGACT motif from the final model. (**C**) Venn diagram for the prediction of AGACT, GGACA, GGACC, and GGACT sites. CLIP sites represent data withheld from training for testing purposes.

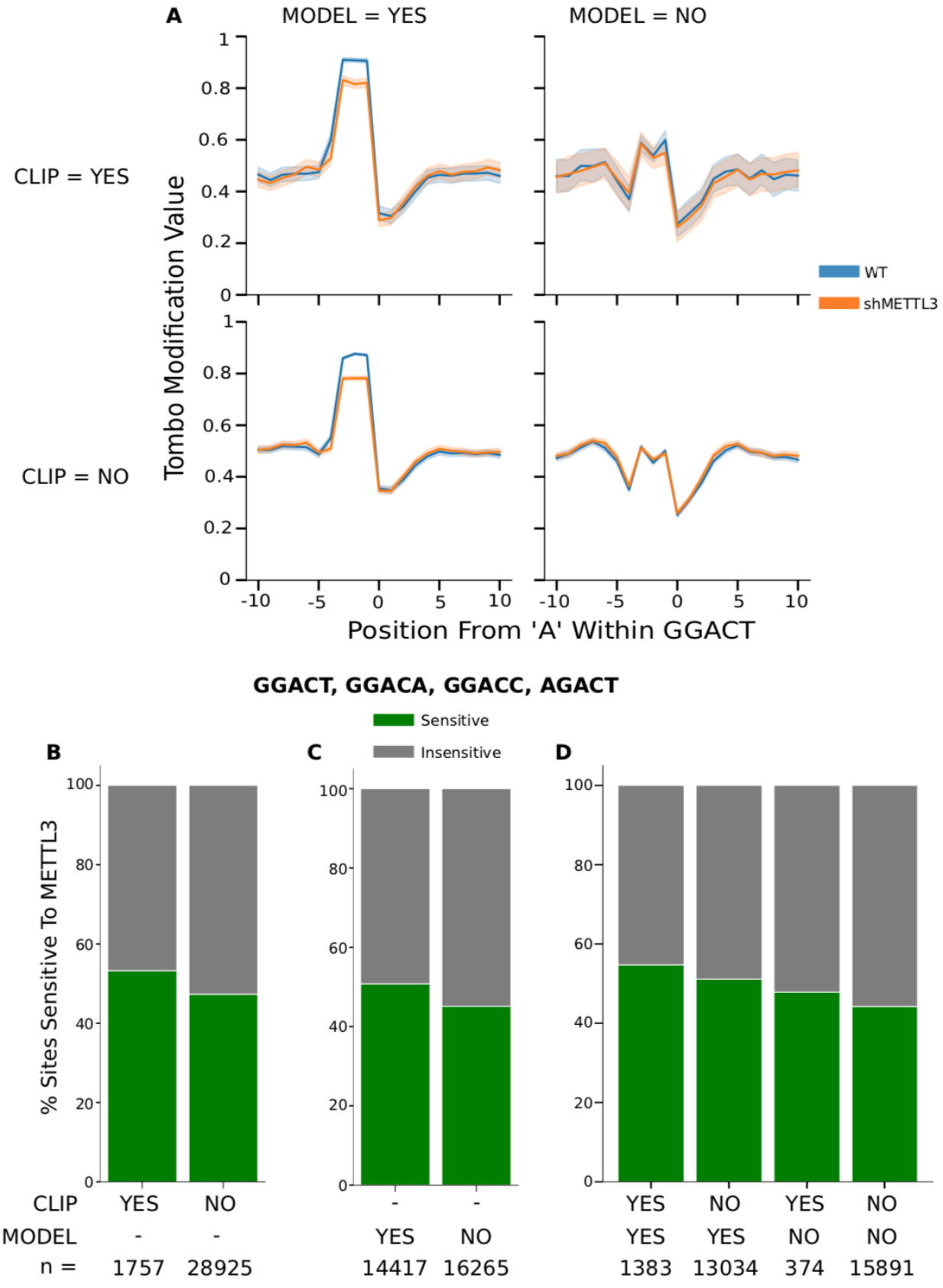


Figure 4. MINES predicted sites mimic m^6A CLIP sites. (**A**) Line plots of Tombo's fraction modified values broken down by CLIP sites and model predictions for GGACT in untreated HEK293T cells or HEK293T cells treated with shRNA targeting METTL3 (shMETTL3). (**B-D**) Percent of predicted m^6A sites sensitive to METTL3 knockdown within the AGACT, GGACA, GGACC, and GGACT motifs.

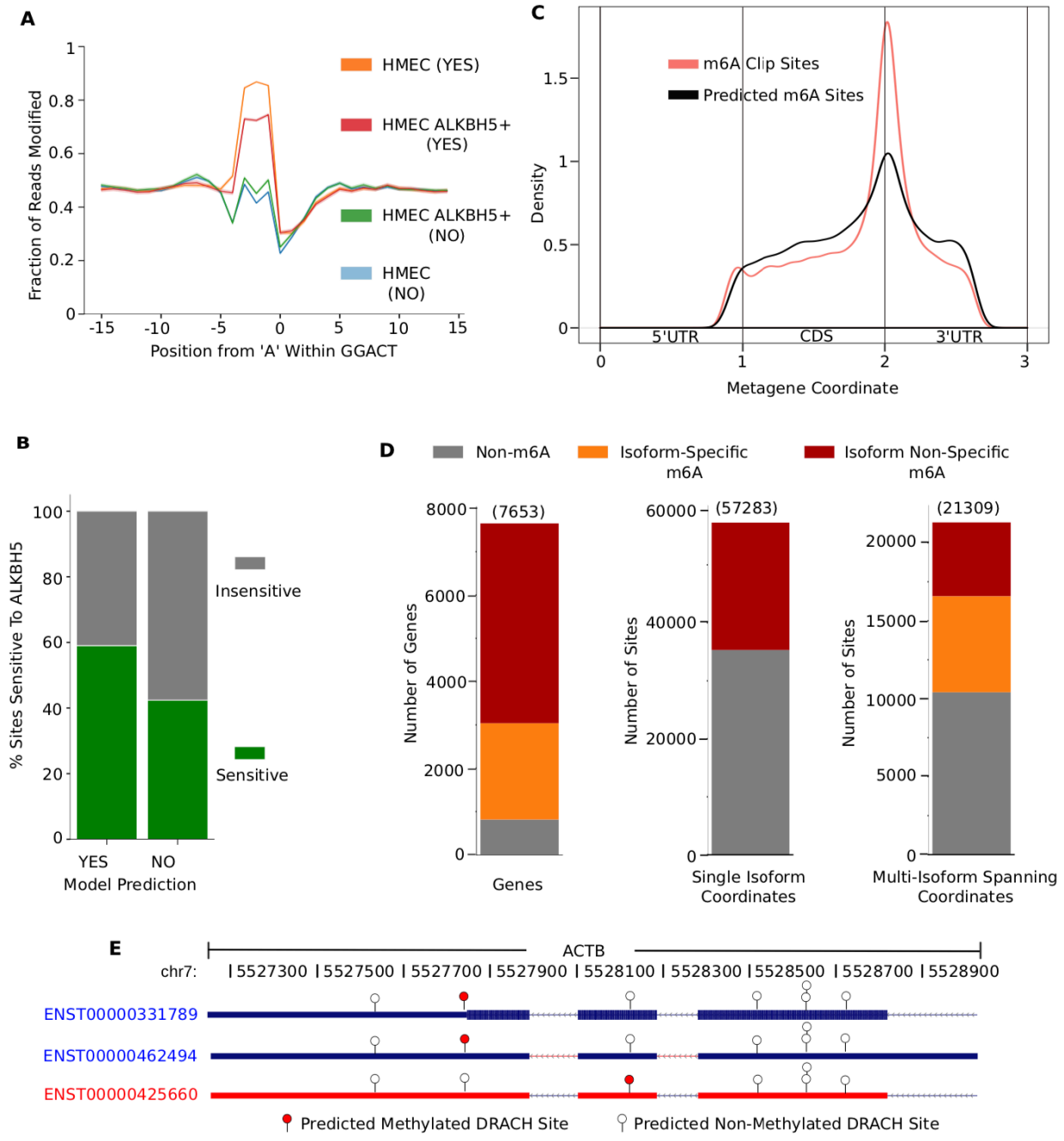


Figure 5. MINES is cell-line independent and provides isoform-level resolution. (**A**) Line plot of Tombo's fraction modified values in HMEC for GGACT and their m^{6}A prediction status. (**B**) Percent of predicted m^{6}A sites sensitive to ALKBH5 overexpression within the AGACT, GGACC, GGACT, and GGACA motifs. (**C**) Metagene analysis of m^{6}A sites in HMEC within the AGACT, GGACC, GGACT, and GGACA motifs. (**D**) Bar plots summarizing MINES' predictions

with gene- and isoform-level resolution. (**E**) MINES isoform-level prediction of ACTB. Converted to hg38 coordinates.

AAACA
AAACC
AAACT
AGACA
AGACC
AGACT
GAACA
GAACC
GAACT
GGACA
GGACC
GGACT
TAACA
TAACC
TAACT
TGACA
TGACC
TGACT
0
0.2
0.4
0.6
0.8
1
-15
-10
-5
0
5
10
15
Fraction of Reads Modified
Position from 'A' with DRACH Motif
HMEC YES
HMEC_ALKBH5+ YES
HMEC NO
HMEC_ALKBH5+ NO

MODEL | MODEL | MODEL

YES NO | YES NO | YES NO

CLIP: YES / NO

TAACA | TAACC | TAACT

TGACA | TGACC | TGACT

TOMBO Modification Value

Position From 'A' Within DRACH Motif

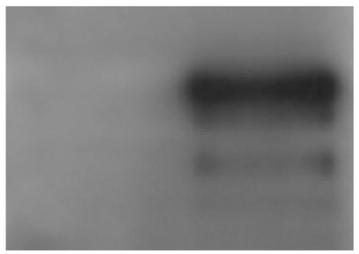
WT

shMETTL3

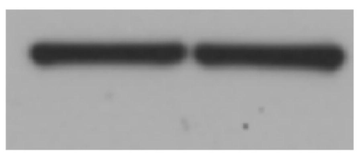
A

HMEC

Dox: - +



Anti-ALKBH5

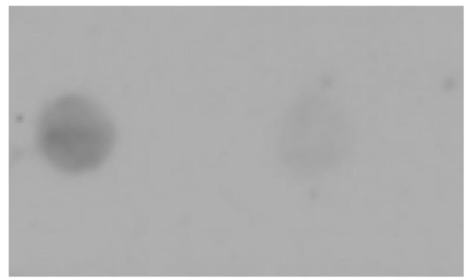


Anti-GAPDH

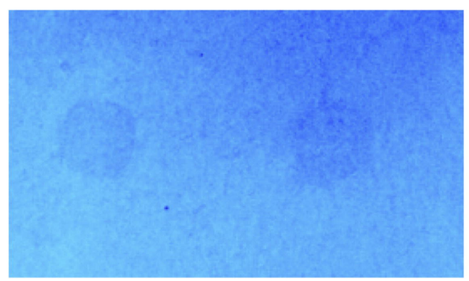
B

HMEC

Dox: - +



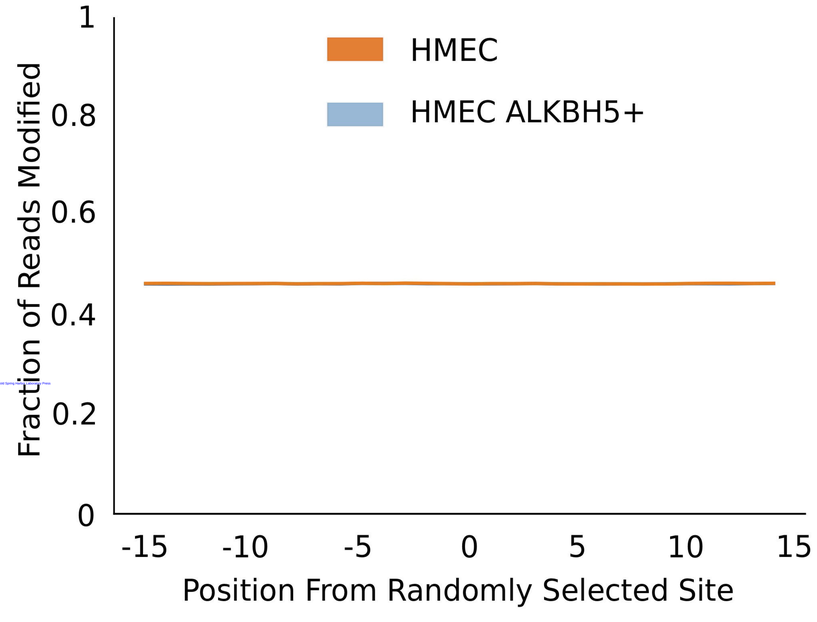
Anti-m6A

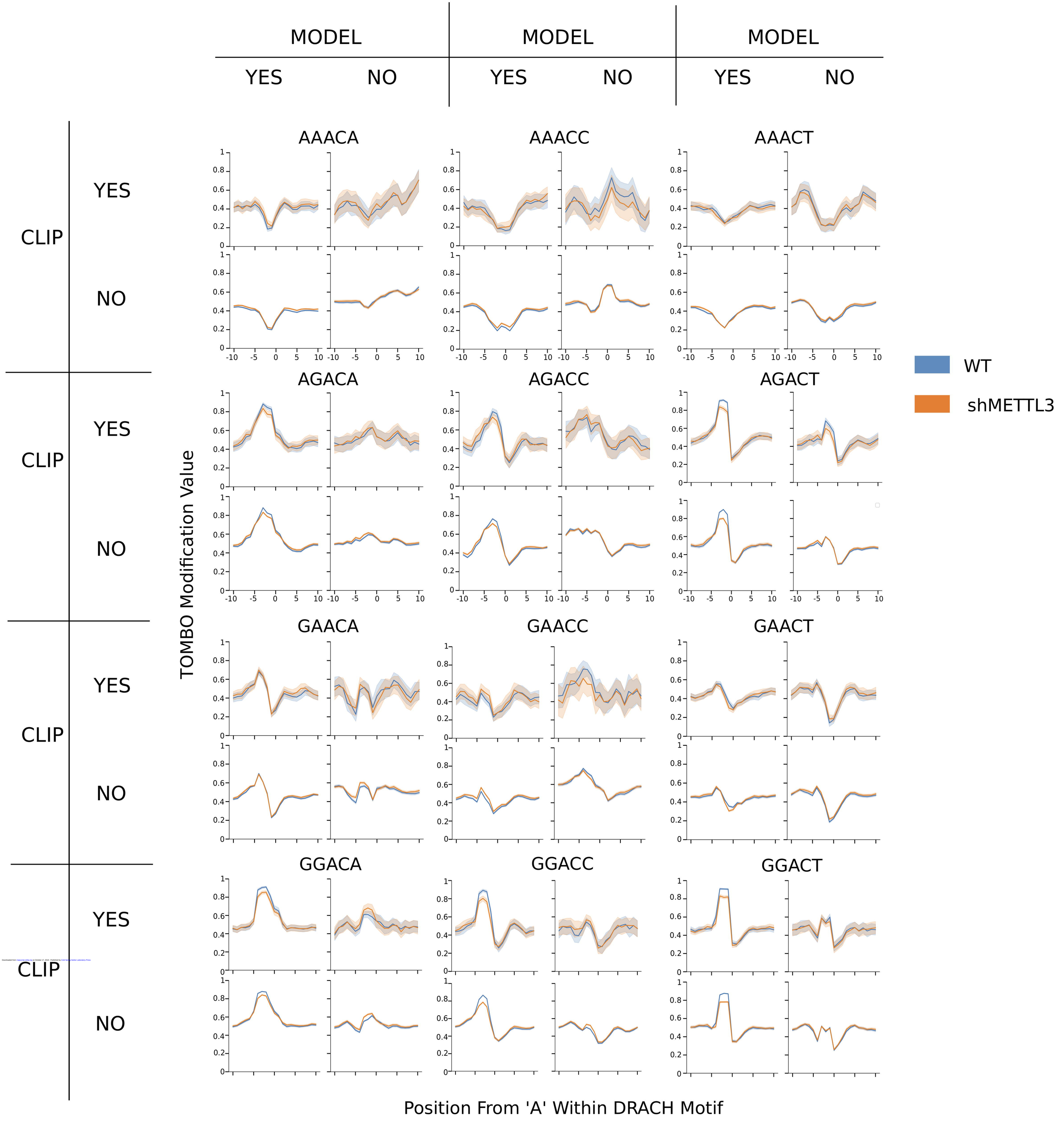


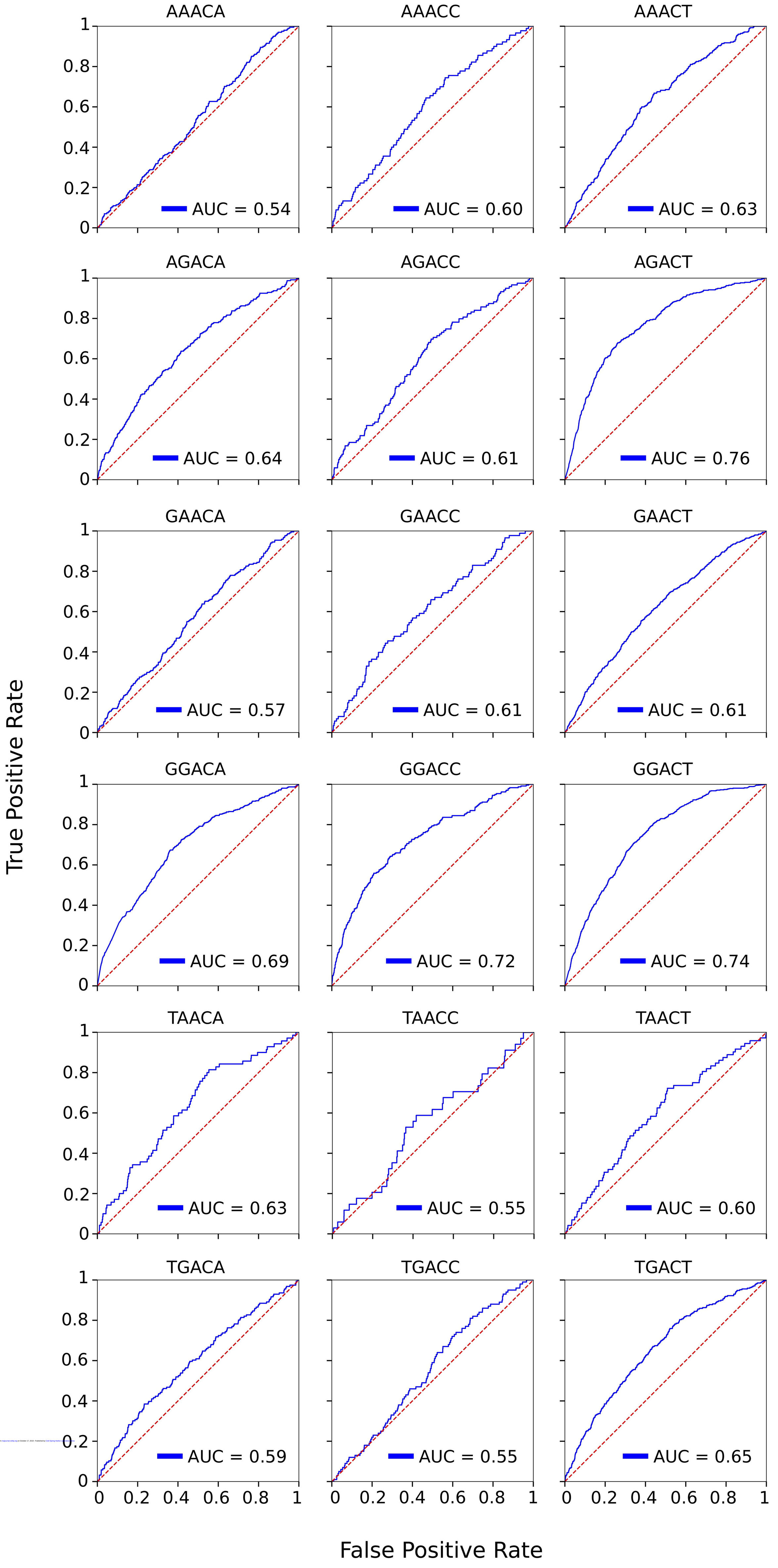
Methylene Blue

Total RNA

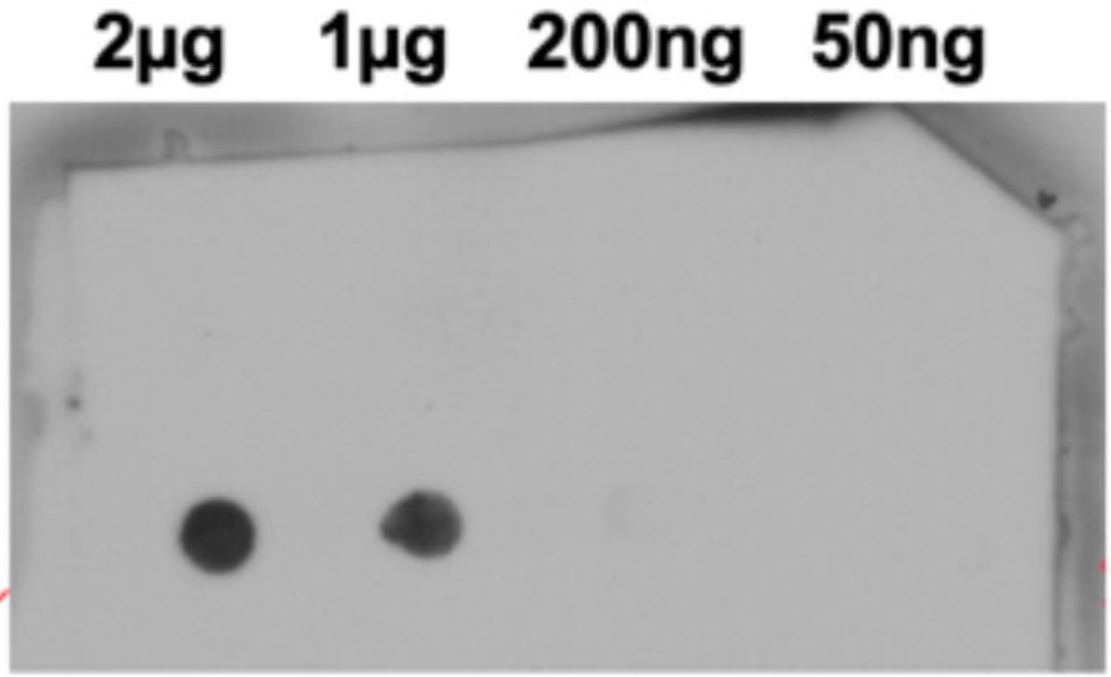
C






AAACA
AUC = 0.54
AAACC
AUC = 0.60
AAACT
AUC = 0.63
AGACA
AUC = 0.64
AGACC
AUC = 0.61
AGACT
AUC = 0.76
GAACA
AUC = 0.57
GAACC
AUC = 0.61
GAACT
AUC = 0.61
GGACA
AUC = 0.69
GGACC
AUC = 0.72
GGACT
AUC = 0.74
TAACA
AUC = 0.63
TAACC
AUC = 0.55
TAACT
AUC = 0.60
TGACA
AUC = 0.59
TGACC
AUC = 0.55
TGACT
AUC = 0.65
True Positive Rate
False Positive Rate
0
0.2
0.4
0.6
0.8
1

Motif	Max Accuracy	Mean Accuracy	Precision
AAACA	0.82	0.73	0.63
AAACC	0.76	0.66	0.63
AAACT	0.79	0.61	0.76
AGACA	0.73	0.64	0.77
AGACC	0.71	0.63	0.63
AGACT	0.72	0.68	0.89
GAACA	0.79	0.66	0.61
GAACC	0.77	0.62	0.61
GAACT	0.67	0.64	0.82
GGACA	0.82	0.69	0.92
GGACC	0.75	0.66	0.86
GGACT	0.83	0.78	0.91
TAACA	0.73	0.61	0.63
TAACC	0.68	0.52	0.4
TAACT	0.72	0.56	0.59
TGACA	0.67	0.59	0.57
TGACC	0.71	0.58	0.57
TGACT	0.74	0.65	0.81

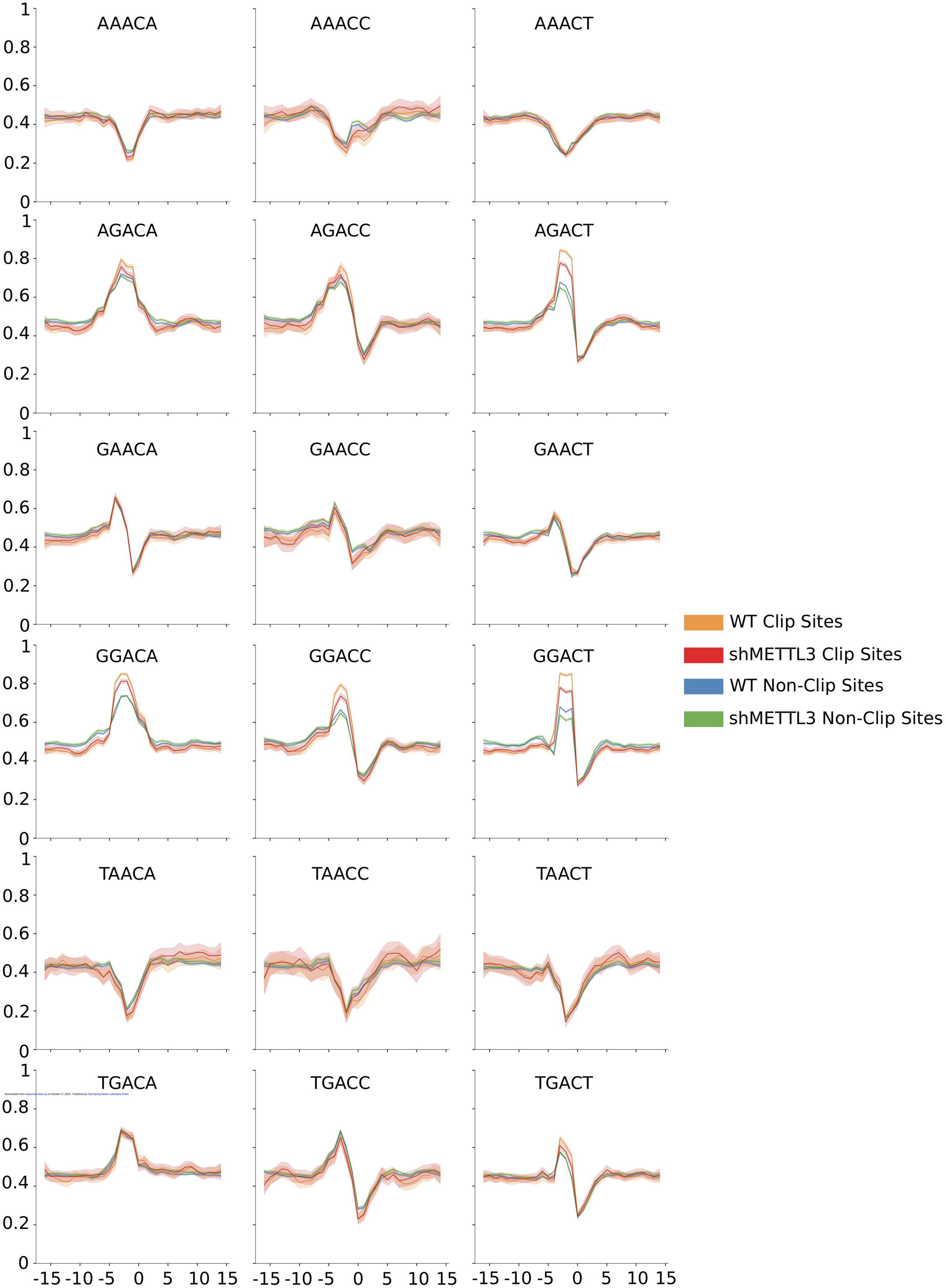


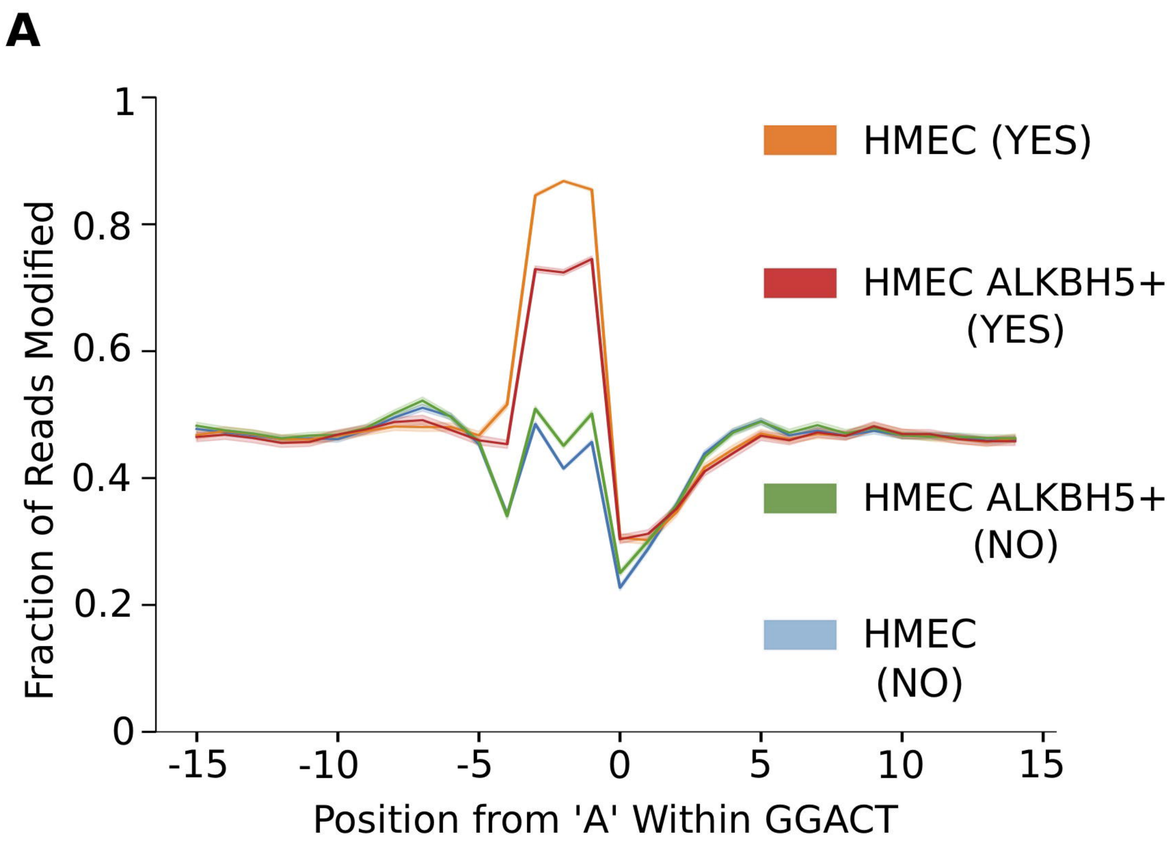
m6A antibody

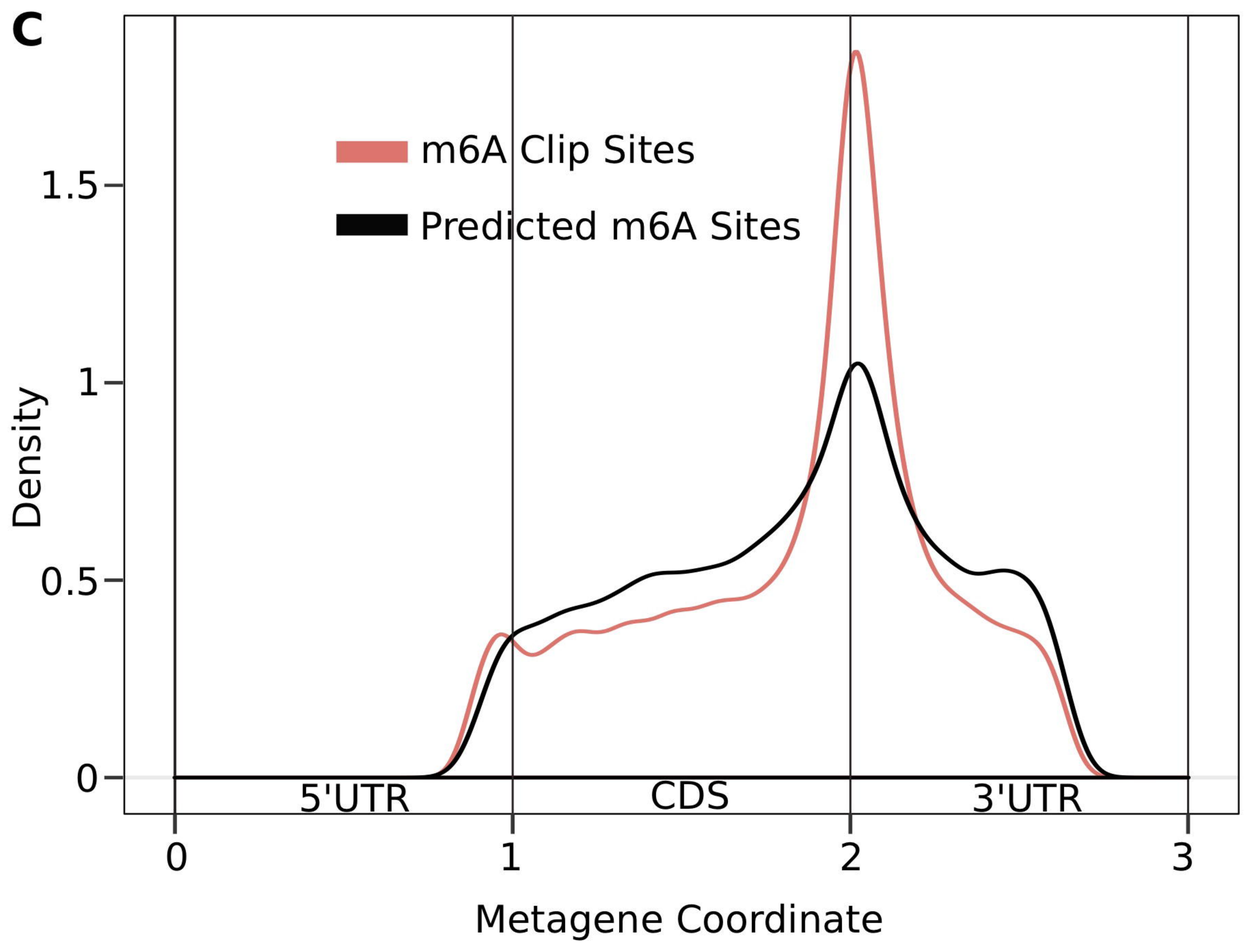
m6A antibody
(over exposed)

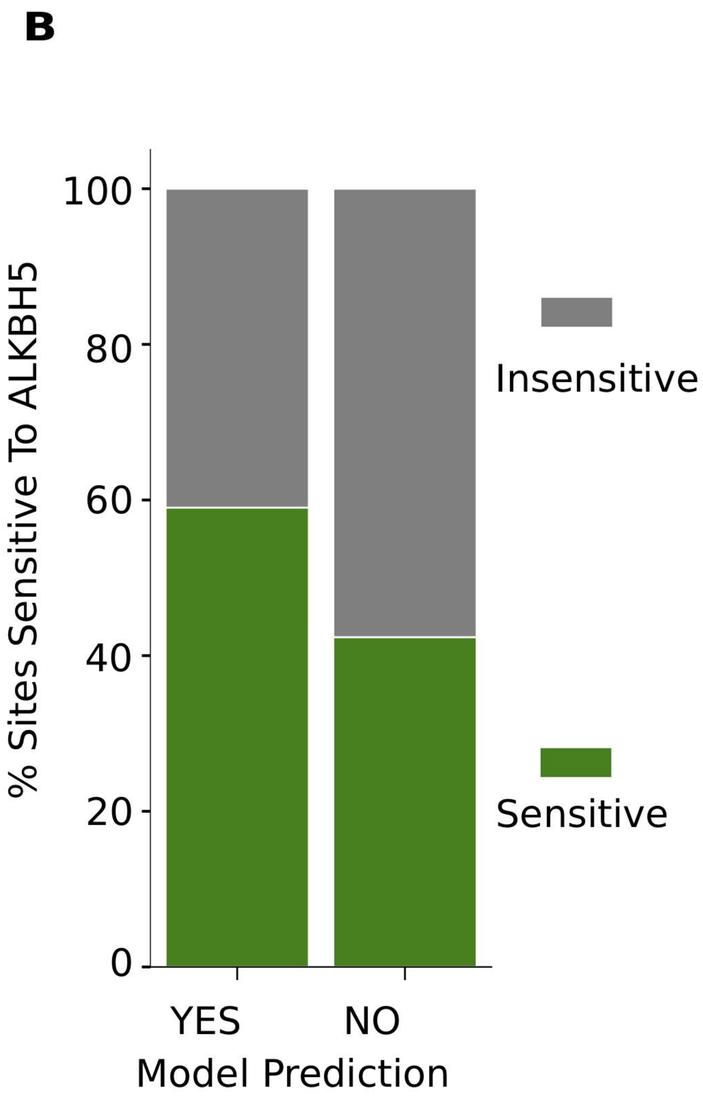
Methylene Blue

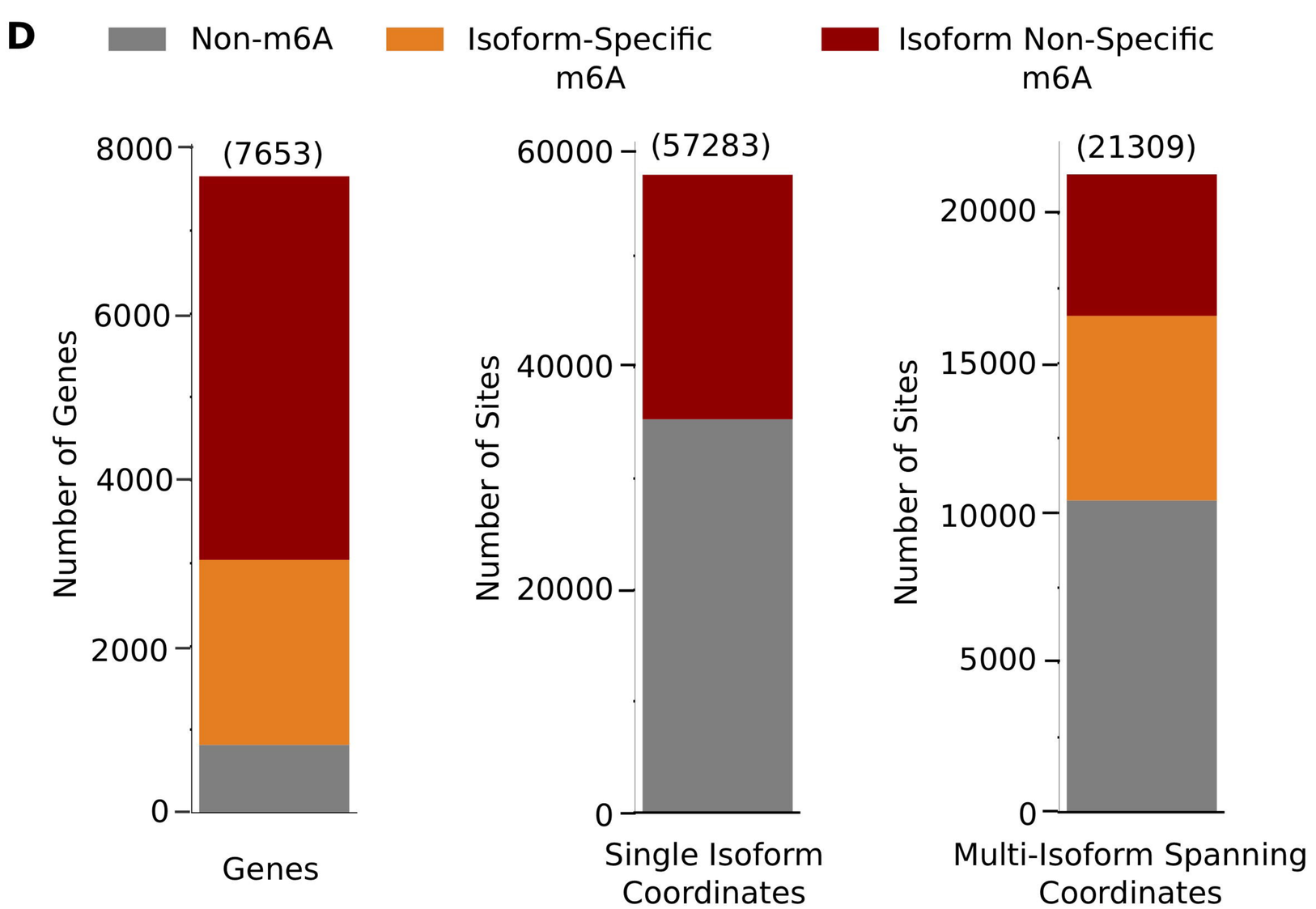
Total RNA

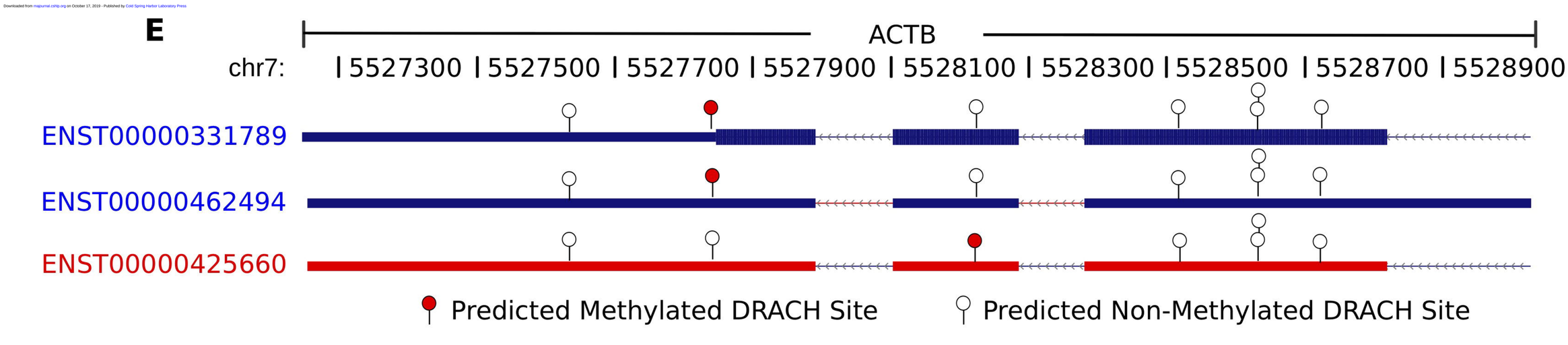


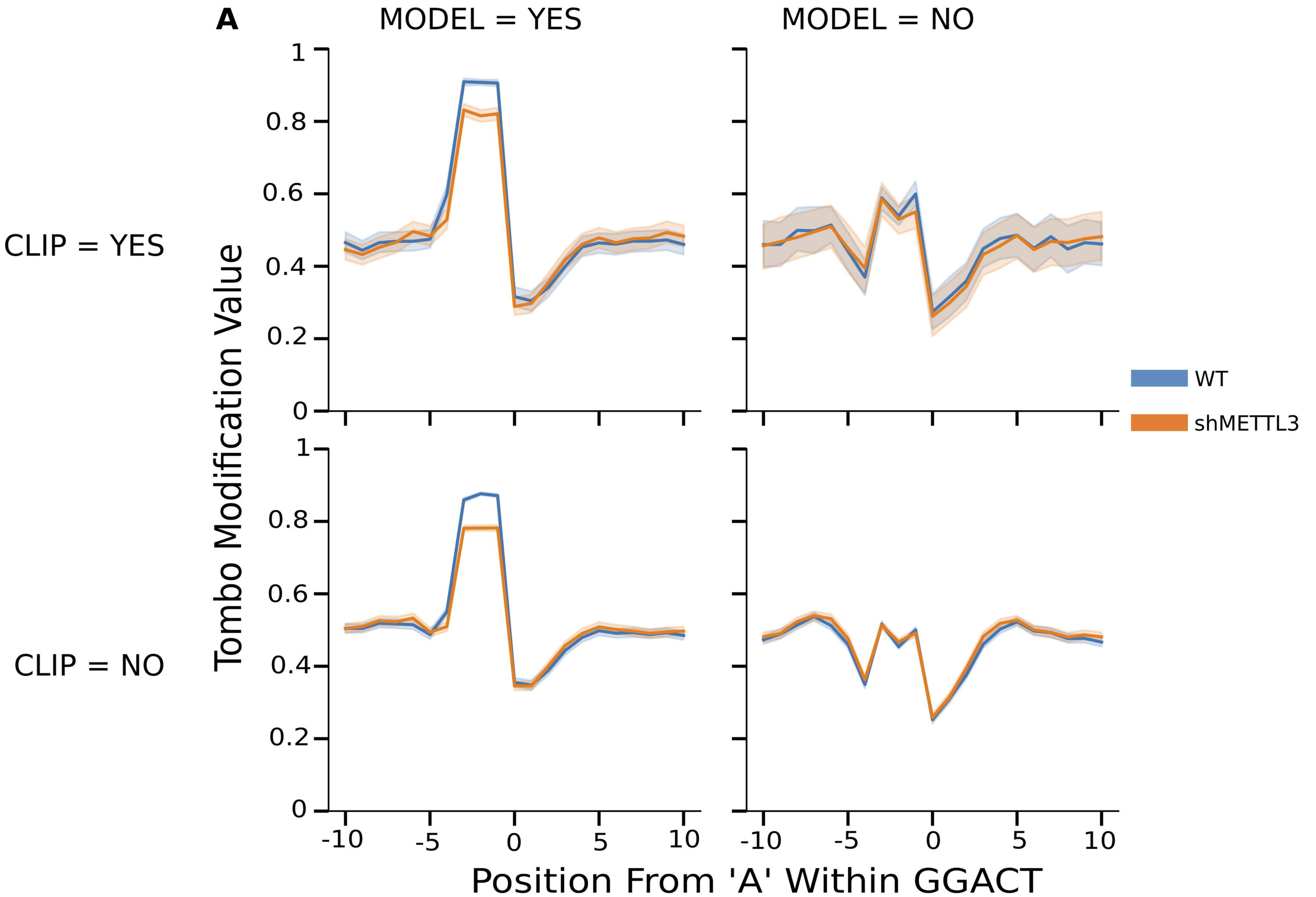


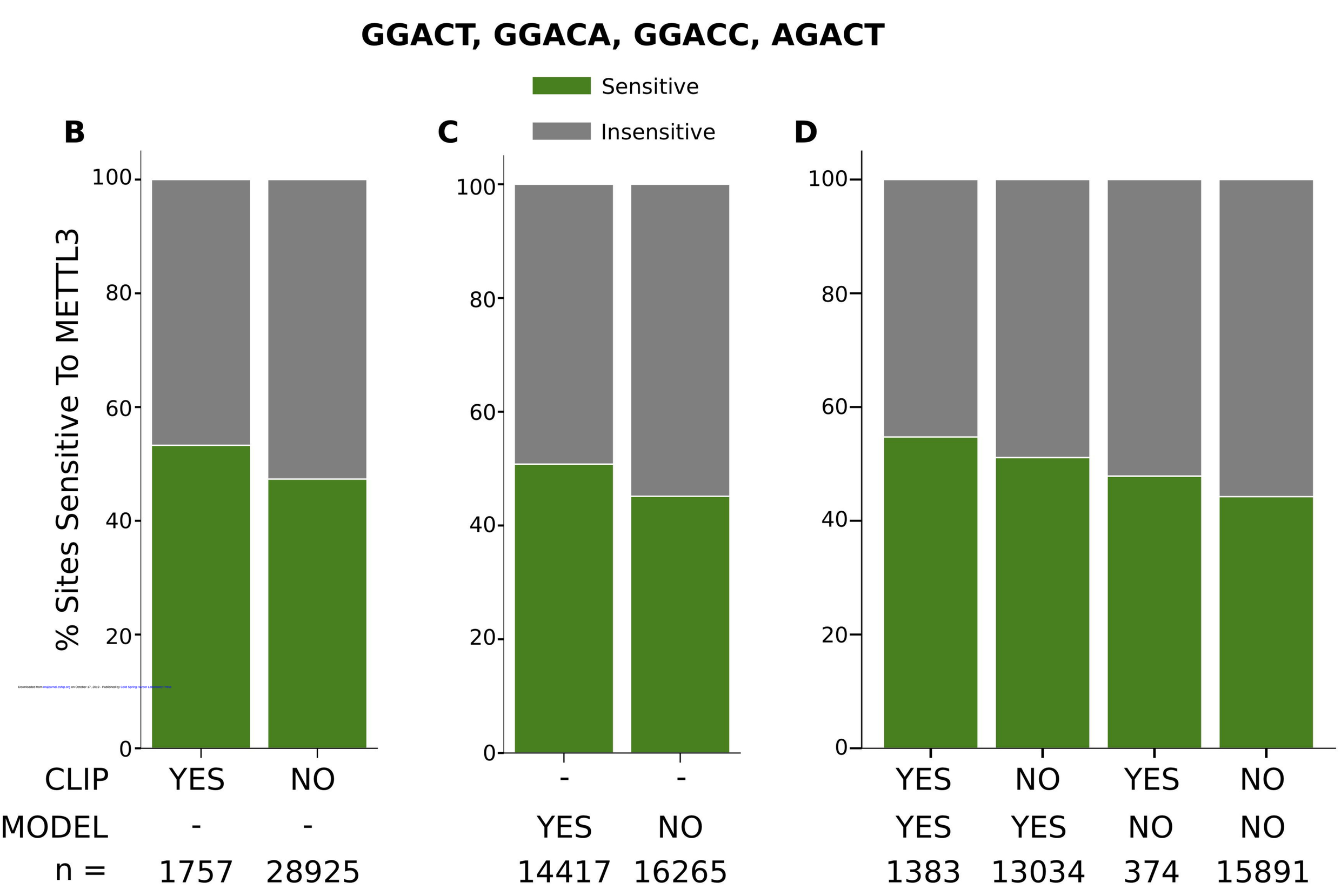


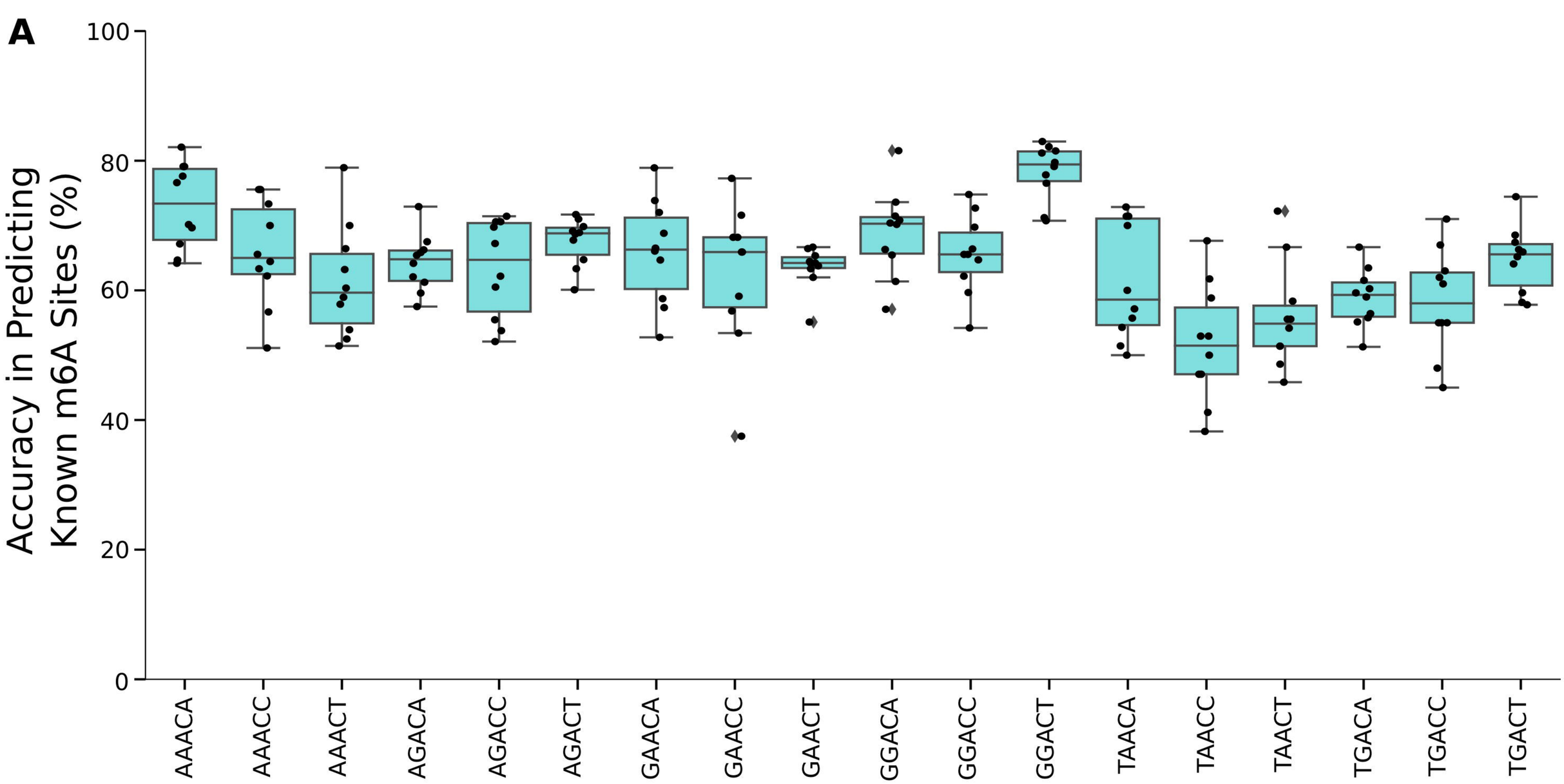


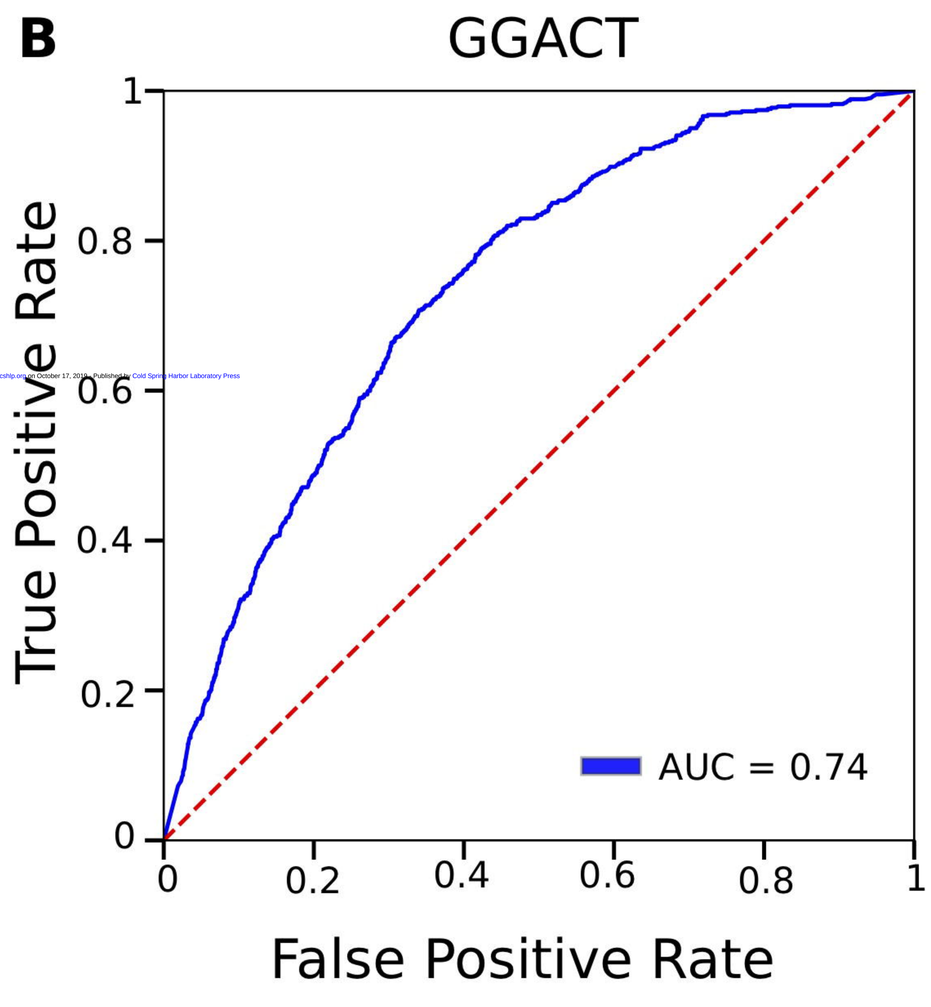


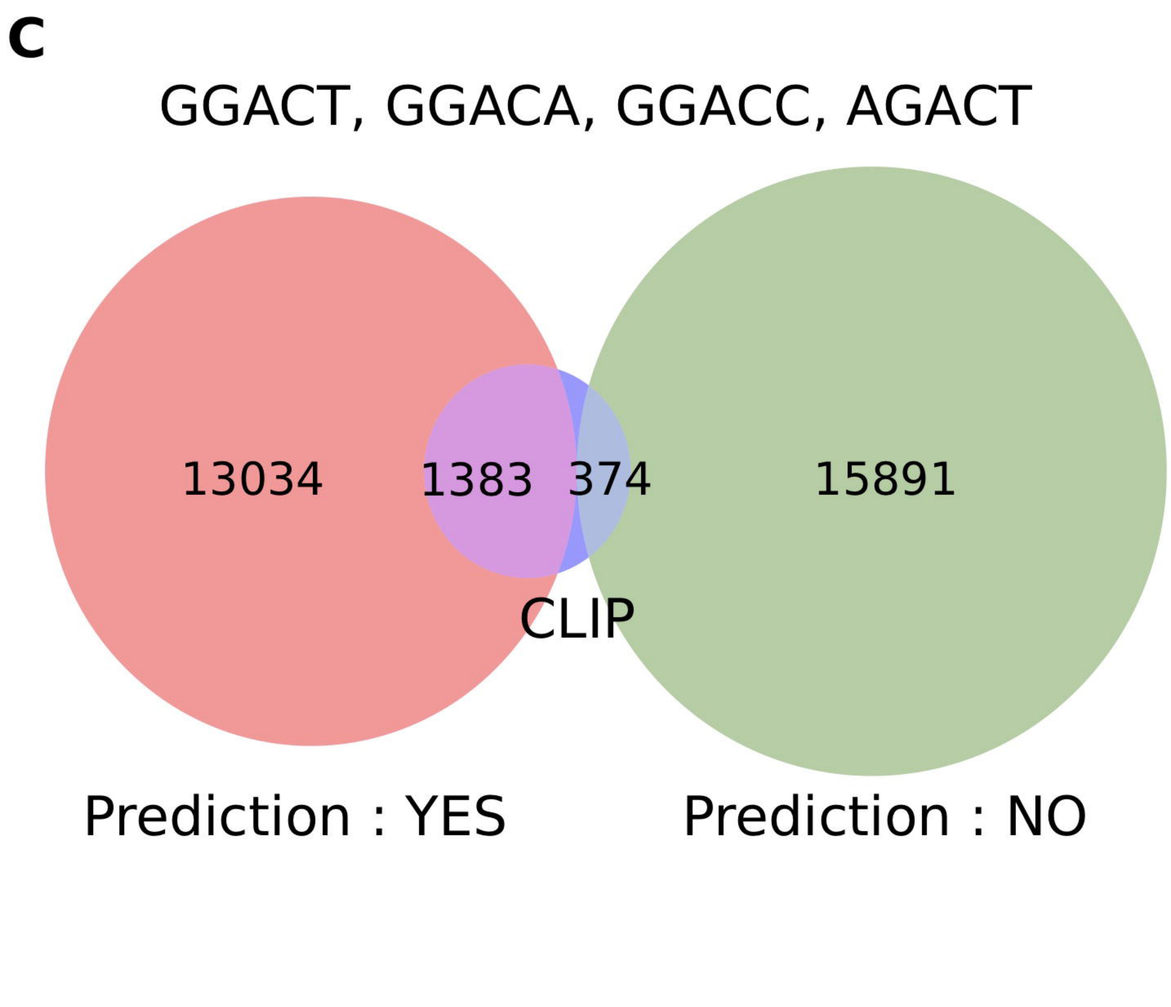




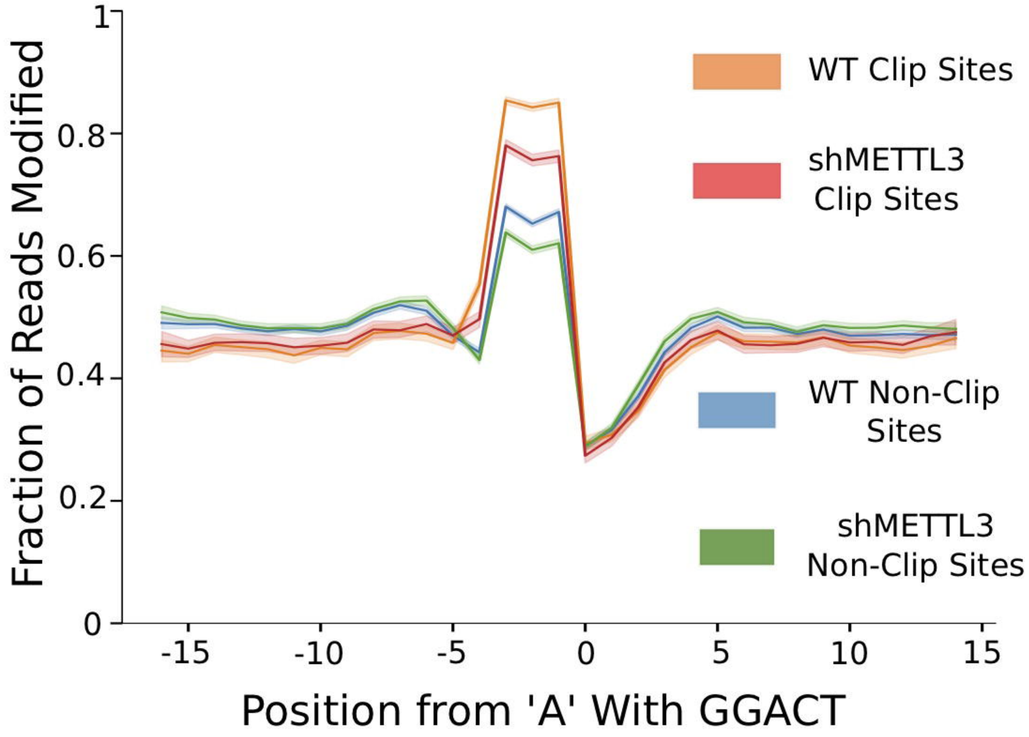




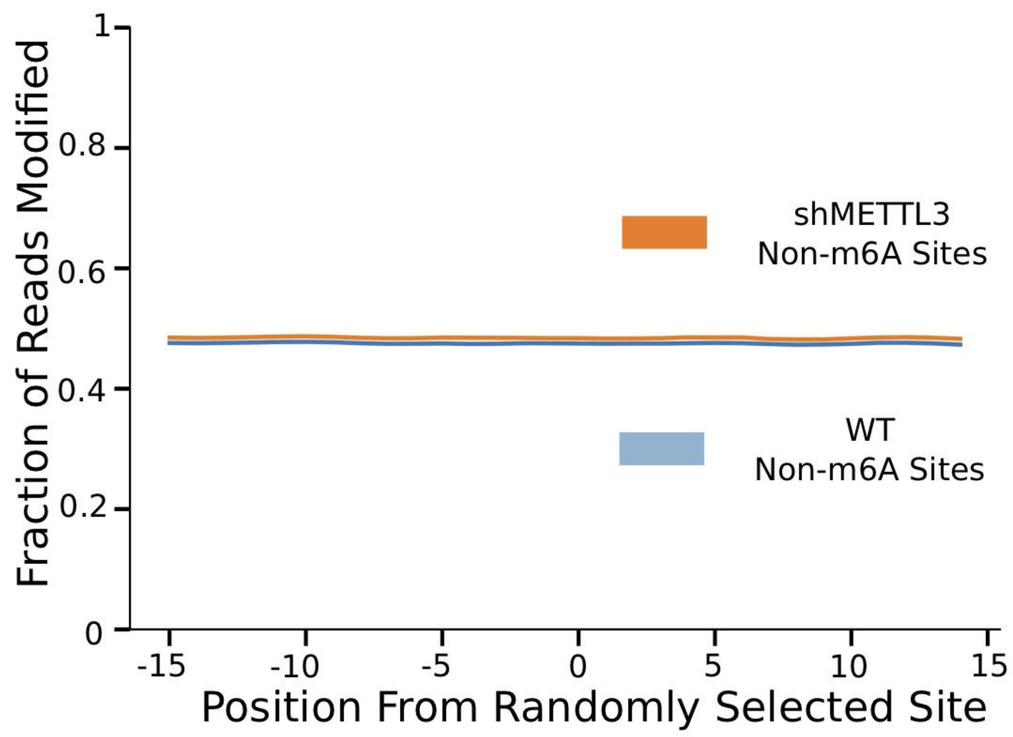




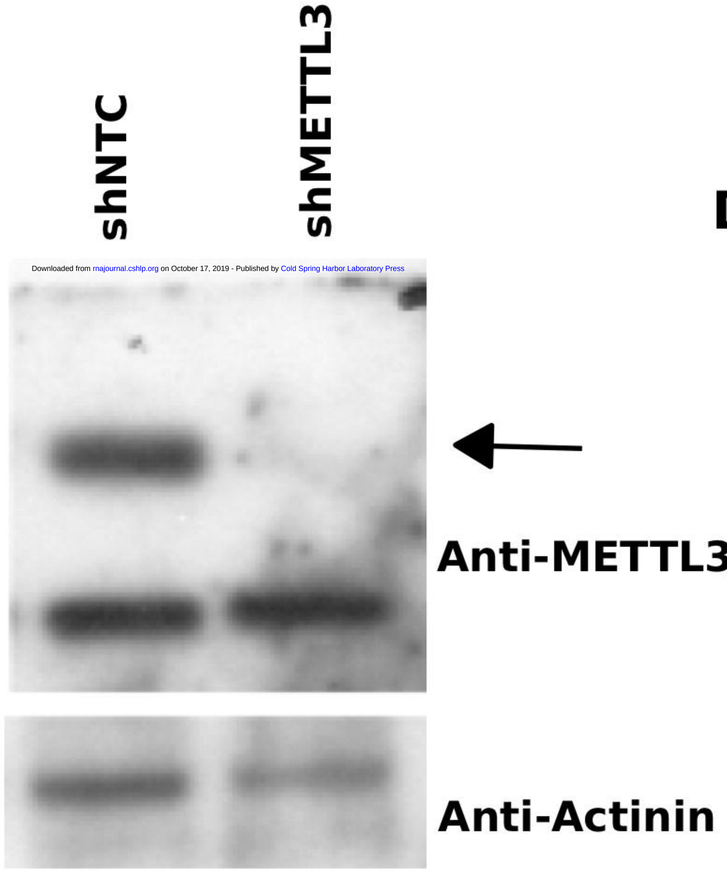
A



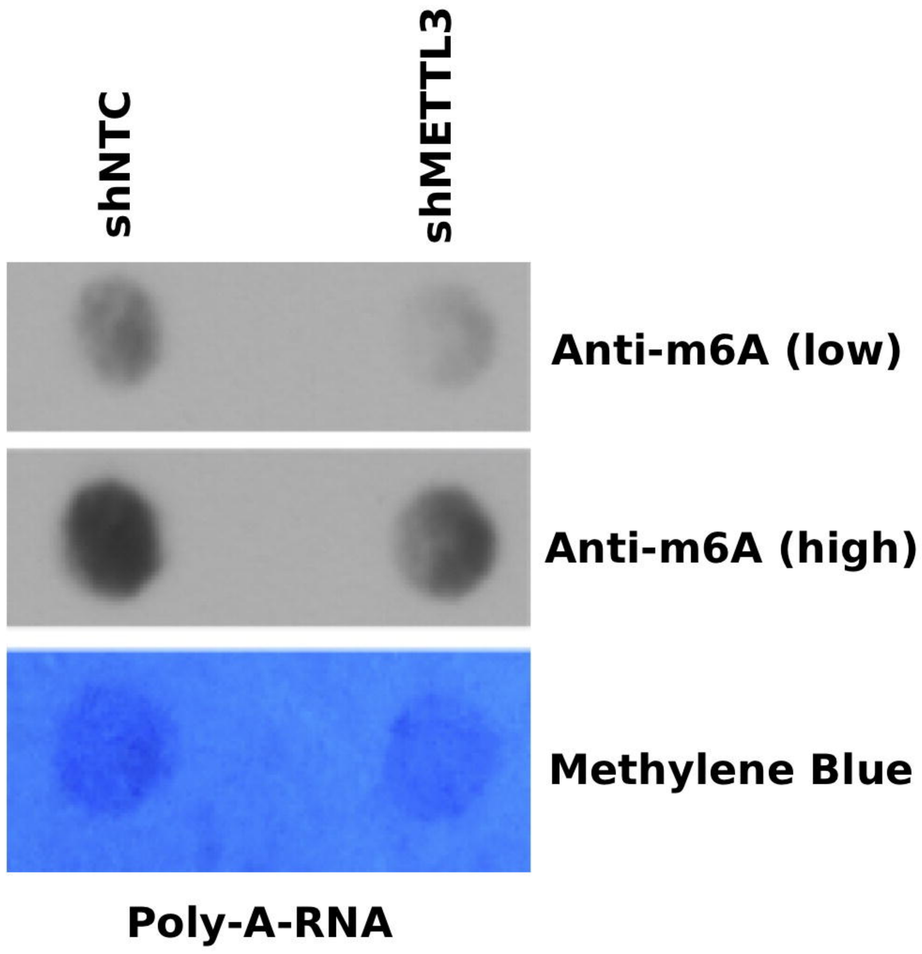
B



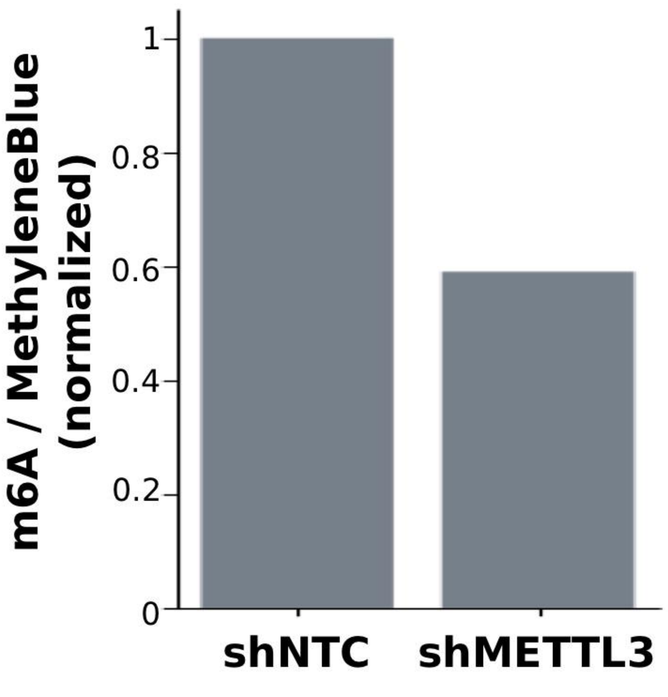
C

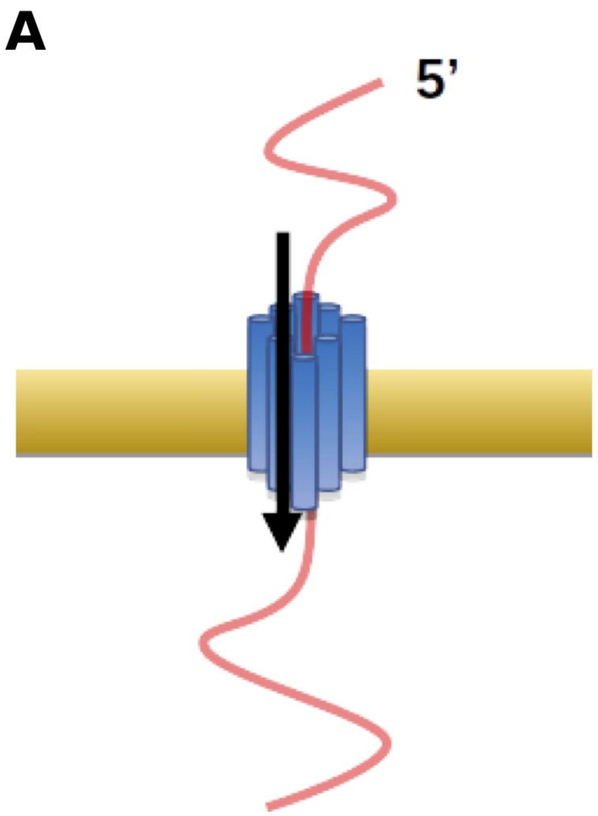


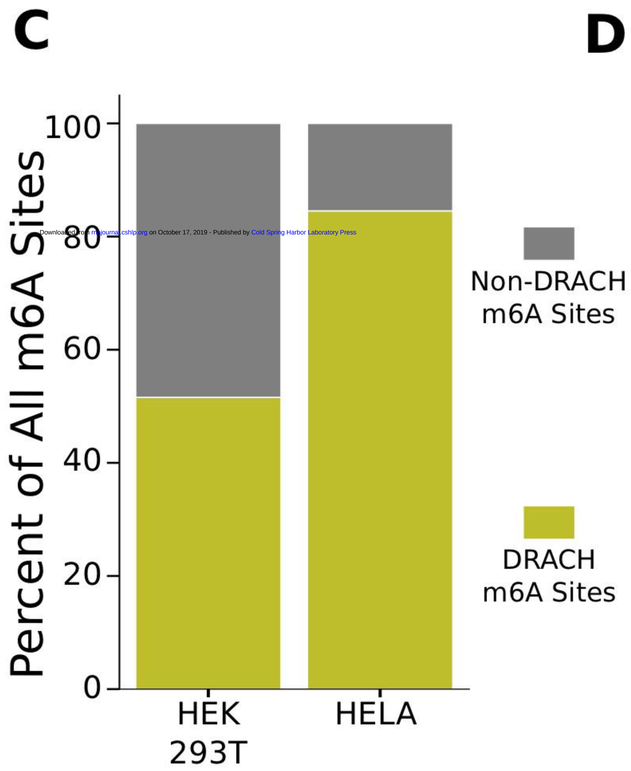
D

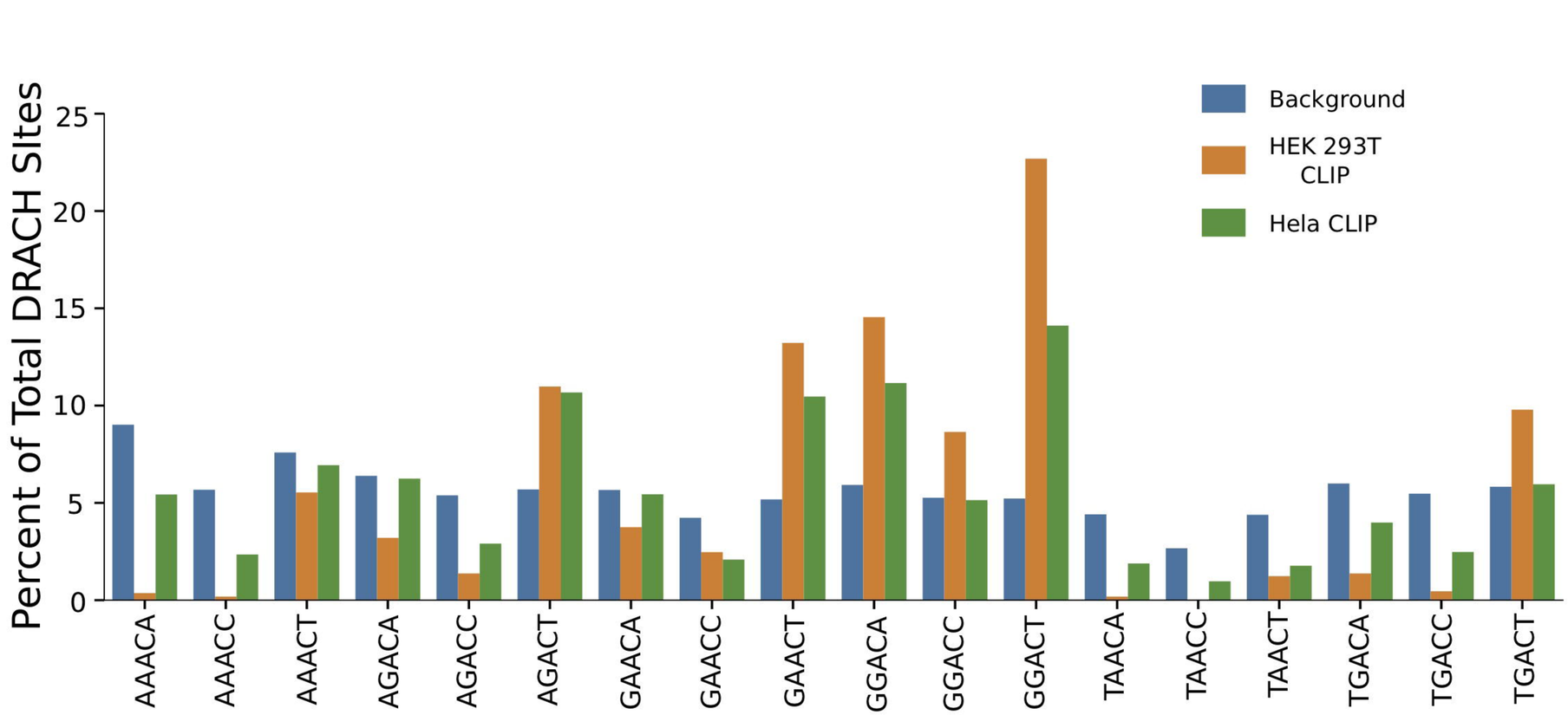


E









Direct RNA sequencing enables m6A detection in endogenous transcript isoforms at base specific resolution

Daniel A Lorenz, Shashank Sathe, Jacyln M Einstein, et al.

RNA published online October 17, 2019