

Direct RNA sequencing enables m⁶A 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}

¹Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla, California 92093, USA

²Stem Cell Program, University of California San Diego, La Jolla, California 92093, USA

³Institute for Genomic Medicine, University of California San Diego, La Jolla, California 92093, USA

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 more than 13,000 previously unannotated DRACH sites in endogenous HEK293T transcripts and identified more than 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, ALKBH5, respectively. MINES (<https://github.com/YeoLab/MINES.git>) enables m⁶A annotation at single coordinate-level resolution from direct RNA nanopore sequencing.

Keywords: m⁶A; nanopore; RNA modifications

INTRODUCTION

Since the identification of the first RNA modification more than 60 years ago, more than 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 N⁶-methyladenosine (m⁶A), 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 m⁶A and coupling to high-throughput sequencing technologies, several transcriptome-wide approaches for identifying m⁶A sites have been developed (Grozhiik and Jaffrey 2018). These techniques have been useful in demonstrating that m⁶A 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 m⁶A methylation. In humans, current evidence suggests that a complex, composed of proteins METTL3,

METTL14, and WTAP, is responsible for installing the m⁶A 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 m⁶A, influencing processes such as splicing, transcript stability, and localization (Shi et al. 2019). m⁶A 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 m⁶A levels and has been linked to a myriad of diseases, including cancer and neurological diseases (Chen et al. 2019; Delaunay and Frye 2019).

Although second-generation polymerase-based sequencing has enabled transcriptome-wide studies of RNA biology, new third-generation sequencing is 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 electric 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

⁴These authors contributed equally to this work.

Corresponding author: geneyeo@ucsd.edu

Article is online at <http://www.majournal.org/cgi/doi/10.1261/rna.072785.119>. Freely available online through the RNA Open Access option.

© 2020 Lorenz et al. This article, published in *RNA*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

offers long sequencing reads (up to 2 Mb) and detection of epigenetic markers (Payne et al. 2019). To illustrate, nanopore-based DNA sequencing has been able to detect the endogenous DNA modifications, m^5C and m^6A (Simpson et al. 2017; McIntyre et al. 2019). Recently, ONT 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 because of the one in 13 single-base error rate (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 the current signal from each site, MINES is able to predict known m^6A CLIP-seq sites with ~80% accuracy within certain DRACH sequences that represent ~35% of reported CLIP sites. When applied to RNA from a primary human mammary epithelial cell line (HMEC), 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 (Fig. 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

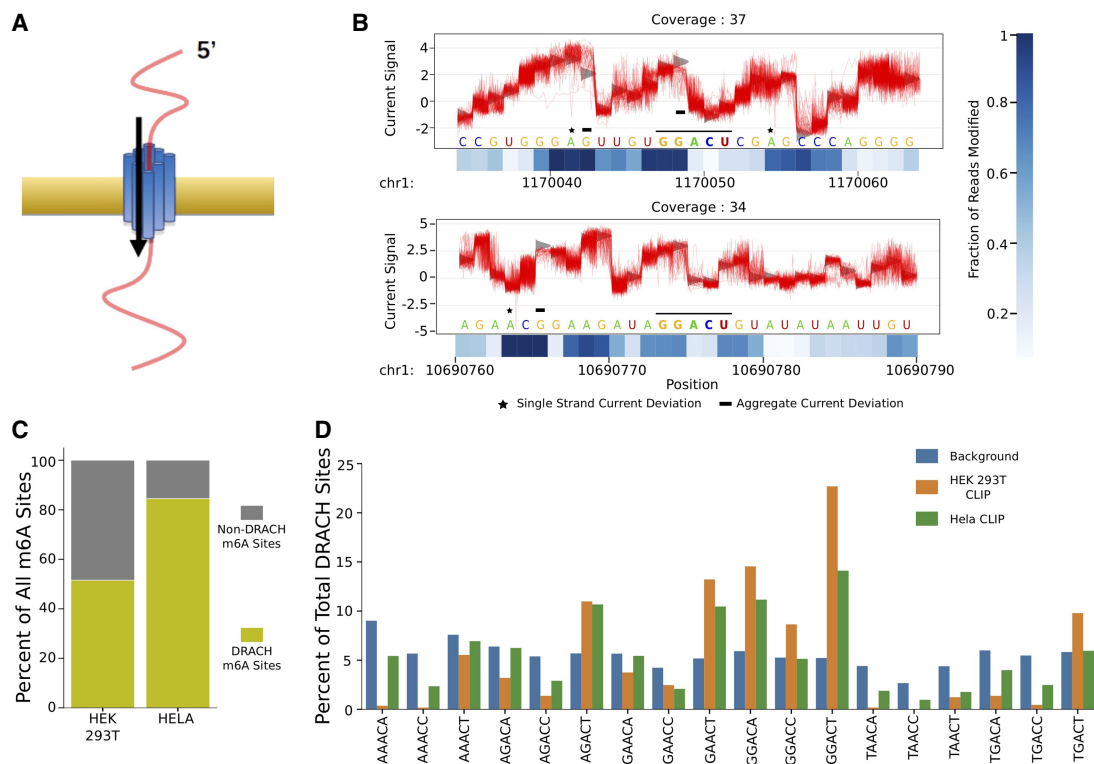


FIGURE 1. Filtering by DRACH motifs encompasses the majority of m^6A sites. (A) Schematic of Nanopore-based sequencing. (B) Representative Tombo outputs depicting individual reads as red lines and expected values as gray distributions. The black bar in the *middle* highlights the GGACT motif. The heatmaps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

applications, such as ONT's Tombo, enable detection of RNA modifications by determining a modification value from calculating the difference between the observed current and a ground truth provided by the reference genome. 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 m⁶A DRACH motifs would be a pragmatic strategy for m⁶A detection. By limiting our algorithm to DRACH sites, we improve the likelihood that our predictions are specific to m⁶A sites and not to other mRNA modifications. Analysis of two site-specific m⁶A cross-linking and immunoprecipitation sequencing (CLIP-seq) data sets from HEK293T and HeLa cells (Linder et al. 2015; Ke et al. 2017) revealed that >50% and >80% of sites were located within DRACH sequences, respectively (Fig. 1C). Deeper analysis revealed that the most common pentamers present within the DRACH motif in both data sets is GGACT, with six sequences (AGACT, GAACT, GGACA, GGACC, GGACT, TGACT) representing >50% of CLIP sites within DRACH sequences (Fig. 1D). Thus, our strategy of prefiltering nanopore reads to reduce the computational load still encompasses the vast majority of m⁶A sites.

Nanopore sequencing distinguishes m⁶A 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 comprise >40% of known m⁶A sites (Linder et al. 2015; Yue 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 an ~5-bp window. We therefore extended our input window to 30 bp 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 midpoint site was found overlapping any site within

the m⁶A CLIP-seq data sets. We required that each window must have a minimum read coverage of five reads, because of the error rate at low coverage loci. Even with this filtering, output fraction-modified values averaged ~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 (Fig. 2A,B). 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 m⁶A. To accomplish this, we generated a HEK293T cell line stably expressing a shRNA that successfully depletes METTL3 protein (Fig. 2C–E) and sequenced poly(A) RNA with ONT. METTL3 depletion had a greater effect on m⁶A levels in total RNA relative to the poly(A) fraction (Supplemental Fig. 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 m⁶A methylation status (Fig. 2A,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 m⁶A sites that were likely undetected within the previous CLIP data sets (Fig. 2A).

Random forest model predicts m⁶A sites

After confirming that ONT is able to detect m⁶A 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 ($n = 1450$ 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. Because 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% to 83%, whereas the precision values ranged from 40% to 92% (Fig. 3A; Supplemental Table 1). Area under the curve (AUC) values ranged from 0.54 to 0.76; however, we believe these values were negatively

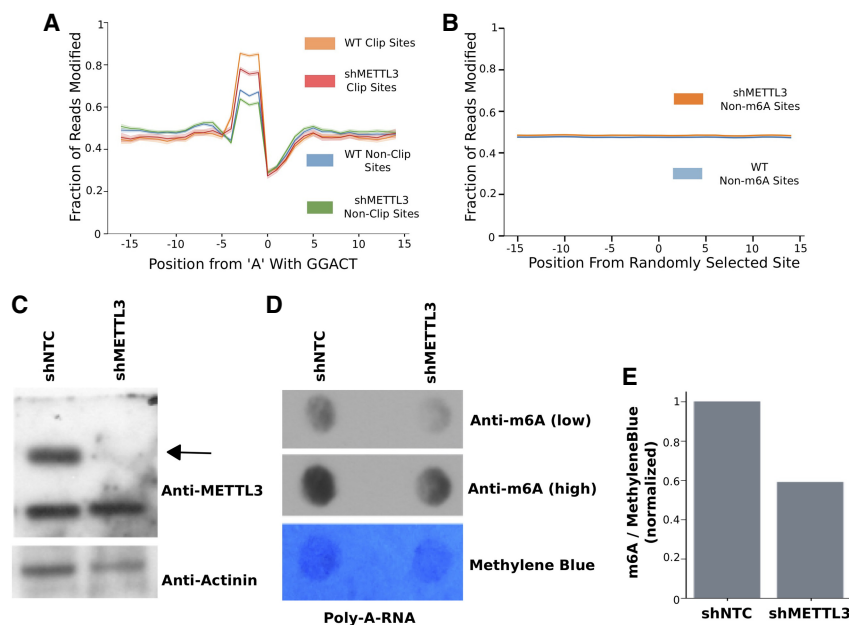


FIGURE 2. Nanopore sequencing can detect endogenous m^6A . (A) Line plots depicting the mean Tombo's fraction modified value across a 30-nt window centered on the "A" in GGACT across all sites in RNA form HEK293T or shRNA targeting METTL3 (shMETTL3) cells. (B) Line plots of Tombo's fraction modified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

affected by the presence of novel, non-CLIP m^6A sites (true negatives) within the test data set (Figs. 2A, 3B; Supplemental Fig. 3). Of the 18 DRACH motifs, only four 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 (Fig. 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 data set at m^6A CLIP sites (Supplemental Fig. 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 (Fig. 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 (Fig. 4A; Supplemental Fig. 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 (Fig. 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. Figure 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^6A 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^6A -modified sites in other cell lines, we sequenced poly(A) RNA from a primary HMEC and a derivative cell line that stably overexpresses the m^6A eraser ALKBH5. Decreased m^6A levels due to ALKBH5 overexpression were confirmed by western and dot blot analyses (Supplemental Fig. 5A,B). Here, we aligned sequencing reads to a human cDNA reference to ensure full transcript coverage and evaluated the ability of MINES to predict m^6A in isoform-specific levels. Using Tombo's coverage data and fraction modified values, and the RFMs generated for four motifs (AGACT, GGACA, GGACC, GGACT), MINES assigned m^6A status to 42,116 sites. Similar to the HEK293T and METTL3 knockdown results, the mean modification values for the HMEC m^6A sites (true positives) were lower in the ALKBH5 overexpression cell line (Fig. 5A; Supplemental Fig. 6) compared to randomly shuffled sites (Supplemental Fig. 5C). Some DRACH sequences produced altered modification patterns than those found in Supplemental Figure 1; however, these are limited to sequences in which the

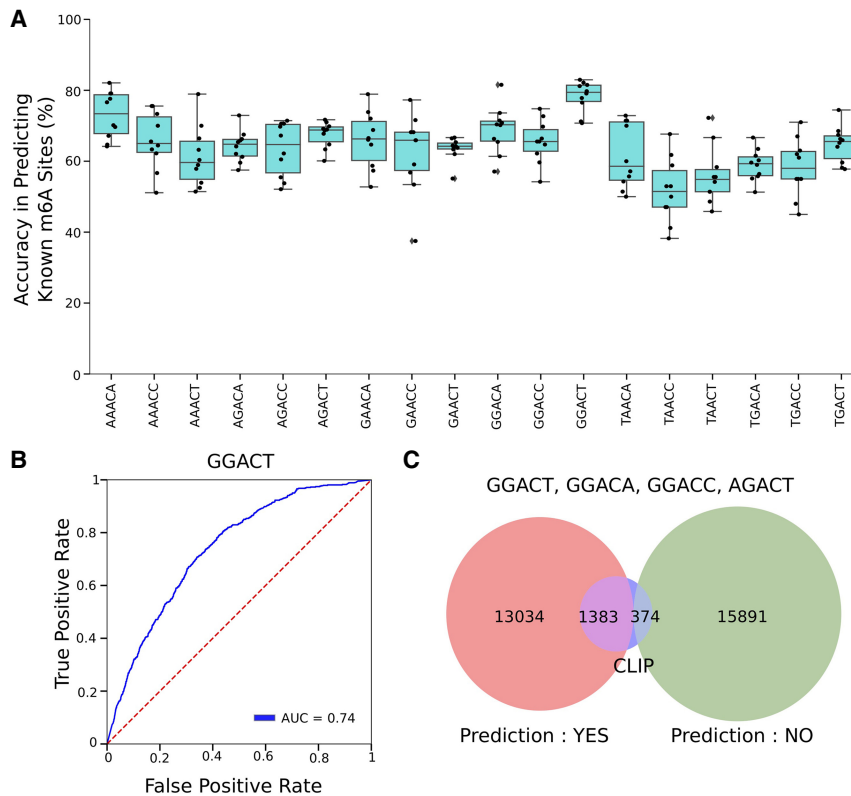


FIGURE 3. A trained RFM accurately predicts m⁶A within DRACH motifs. (A) Box plots showing the model's accuracy of predicting CLIP sites, organized by each DRACH motif across 10 training runs. (B) ROC curve for GGACT motif from the final model. (C) Venn diagram for the prediction of AGACT, GGACA, GGACC, and GGACT sites. CLIP sites represent data withheld from training for testing purposes.

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 m⁶A predicted fraction (Fig. 5B), similar to METTL3 knockdowns. To further assess the accuracy of MINES, we studied the distribution of predicted m⁶A sites across all transcript isoforms to resolve the density of m⁶A sites within different genic regions including 5' UTR, CDS, and 3' UTR, respectively (Fig. 5C). This analysis revealed the characteristic density peak at the start of the 3' UTR, confirming that our model resembles results seen in traditional m⁶A-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 2225 genes to have isoform-specific methylation patterns out of the 6837 m⁶A-containing genes (Fig. 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, 4726 sites predicted to be consistently methylated, and 6168 sites with isoform-spe-

cific methylation (Fig. 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 m⁶A. The three isoforms (ENST00000331789, ENST00000425660, and ENST00000462494) had seven sites that met our read depth and sequence requirements (Fig. 5E). Two of the transcripts (ENST00000331789 and ENST00000462494) were predicted to contain one m⁶A 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 m⁶A biology with isoform-specific resolution.

DISCUSSION

Although effective, m⁶A CLIP-seq and RIP-seq techniques depend on the availability of high-quality antibodies and 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 ONT sequencer as input and predicts m⁶A modified sites in poly(A) selected mRNA.

Coupling publicly available m⁶A data sets and Tombo's modification values, we demonstrated a largely accurate detection of m⁶A sites at positions 1 through 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, whereas randomly selected background sites remain unaffected. This serves as an independent validation of our results. Interestingly, we observed a

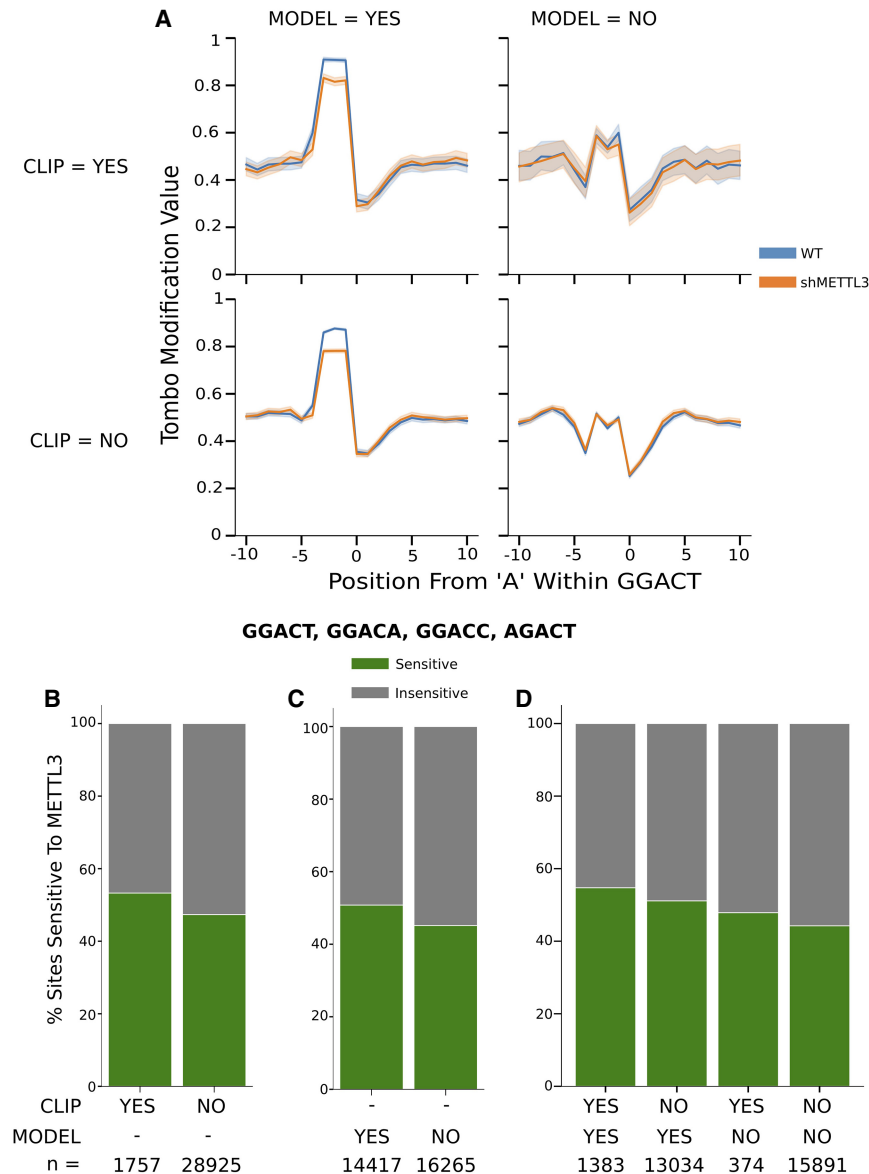


FIGURE 4. MINES-predicted sites mimic m^6A CLIP sites. (A) Line plots of Tombo's fraction modified values broken down by CLIP sites and model predictions for GGACT in untreated HEK293T cells or HEK293T cells treated with shRNA targeting METTL3 (shMETTL3). (B–D) Percent of predicted m^6A sites sensitive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

decrease in modification value in several non-CLIP sites upon METTL3 knockdown, indicating potentially unannotated m^6A 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 m^6A sites. Using MINES to identify methylation sites within these sequences, we predicted a total of 13,034 m^6A sites in HEK293T cells. These newly identified sites exhibited similar modification values and sensi-

tivity to loss of METTL3 to those found in previous data sets. 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^6A/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^6A sites in an HMEC 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 nonmethylated sites, consistent with our results in the METTL3 depletion in HEK293. These new sites also mimic the distribution of m^6A 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^6A methylation in an isoform-specific manner for 2225 genes (Fig. 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.

MATERIALS 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

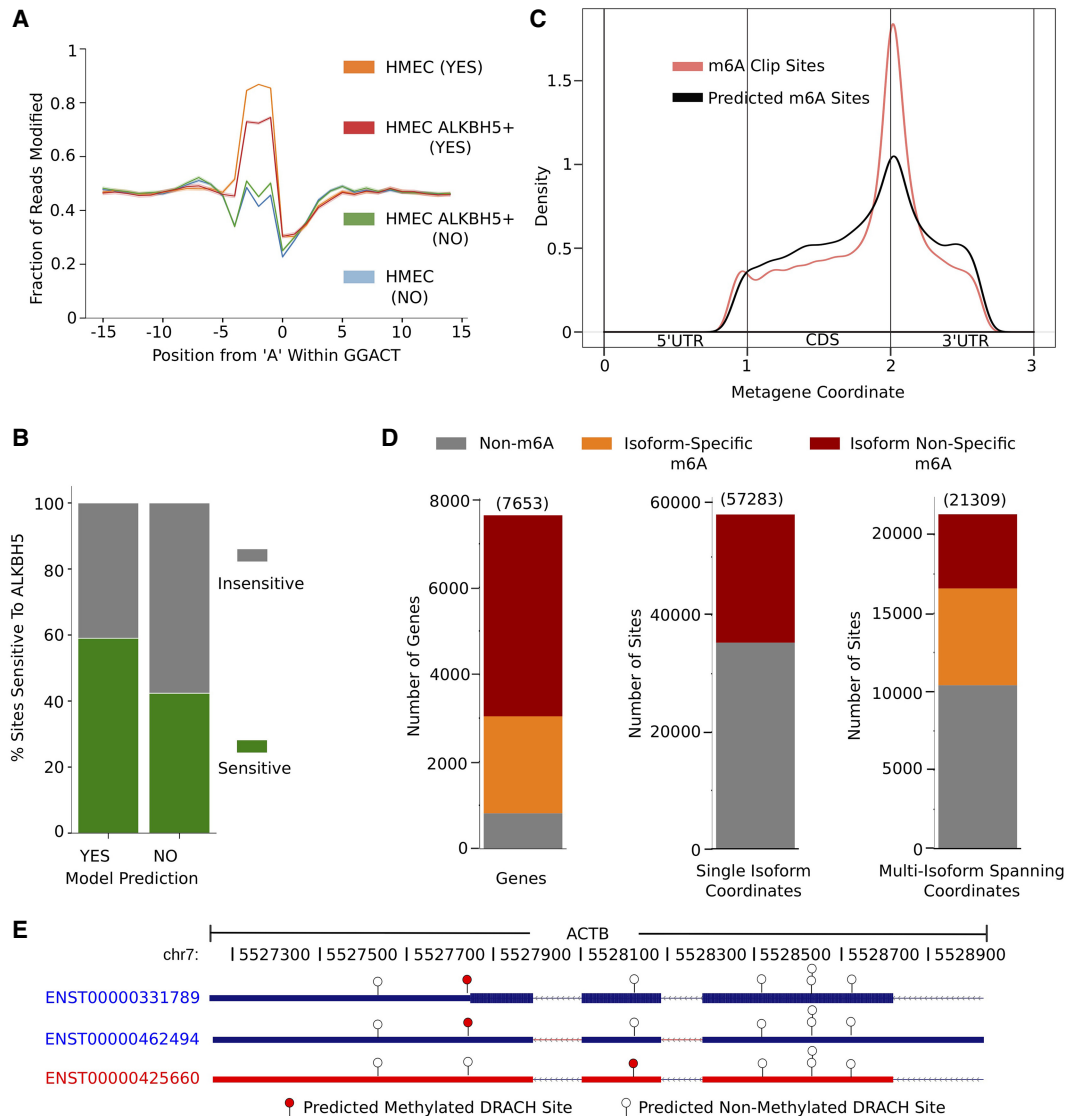


FIGURE 5. MINES is cell line-independent and provides isoform-level resolution. (A) Line plot of Tombo's fraction modified values in HMEC for GGACT and their m⁶A prediction status. (B) Percent of predicted m⁶A sites sensitive to ALKBH5 overexpression within the AGACT, GGACC, GGACT, and GGACA motifs. $n = 42,116$ (yes) and $71,365$ (no). (C) Metagene analysis of m⁶A 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.

supplemented with MEGS: S0155, respectively, following standard tissue culture practices. METTL3 shRNA plasmid (TRCN000034717) 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 carboxy-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 six-well plates at ~80% confluence. The following day the cells were transfected by combining 35 μ L Opti-MEM, 5 μ L P3000 reagent (both Thermo Fisher), 500 ng psPAX.2, 50 ng pMD2.g, and 500 ng

shRNA/gene vector. Then, 15 μ L Opti-MEM and 4 μ L Lipofectamine 3000 (both Thermo Fisher) were mixed in another tube before being combined together and allowed to incubate at room temperature for 20 min. This mixture was added to cells in a dropwise fashion. After 4–6 h, the media was replaced with fresh media. Media containing virus was harvested 48 and 72 h posttransfection. Viral particles were passed through a 0.45- μ m sterile filter. Virus containing media was then added to HEK293T or HMEC cell lines supplemented with 8 μ g/mL polybrene. Media was removed after 24 h and replaced with media containing 2 μ g/ml puromycin. ALKBH5 overexpression was induced with the addition of 1 μ g/mL doxycycline to media for 48 h before collecting cells.

Western blots

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

RNA isolation and poly(A) selection

At 80% confluency in 10-cm plates, cells were washed with PBS and harvested in 1 mL of TRIzol reagent (Thermo Fisher) or Direct-zol kit with DNase treatment (Zymo Research). Total RNA was extracted following the manufacturer's protocol. Then, 20 μ g 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.

m⁶A dot blot

RNA was quantified prior to blotting using a Nanodrop spectrophotometer. Unless otherwise noted, 500 ng of RNA was then diluted to 100 μ L in H₂O and spotted on a prewashed (100 μ L H₂O) nylon membrane (Hybond-XL, GE Healthcare) using a dot blot apparatus (Bio-Dot, Bio-Rad) and washed with 100 μ L of H₂O. RNA was then cross-linked to the membrane with a UV cross-linker fitted with 254 nm bulbs at 120 mJ/cm². The membrane was processed and developed as described above, using an m⁶A antibody (Synaptic Systems #202111) at 1:1000 dilution. After developing, the membrane was washed 3 \times with TBST, and 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 the membrane was rinsed with 50% ethanol/water before being imaged. Dots were quantified by densitometry using ImageJ.

Nanopore sequencing

Five hundred nanograms 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 h. 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 the fraction of modified reads (wiggle files) for each position within the reference.

m⁶A 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 bp on both sides of the "A" within the DRACH motifs. These regions were further filtered to have a minimum coverage of five reads. The DRACH regions were intersected with known m⁶A sites to identify true positive regions obtained from GSA data sets GSM1556678 and GSM2300429 (REFs: PMID: 26121403, PMID: 28637692).

A random forest classifier is a decision tree-based classifier. The Python implementation of random forest (*sklearn*) was used to generate a model to predict m⁶A sites from the filtered DRACH data. Since Nanopore data reflects the occurrence of a m⁶A 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.

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. About 70% of the known occurrences were used as training data, whereas 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. 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 10 bp upstream and downstream from 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 data sets, 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 data set. To maintain the sanity of the validation, we ensured that the test data sets had not been run through the RF model in any capacity.

The purpose of the model is to identify novel m⁶A sites, in addition to the known CLIP sites. We expected the accuracy of the model to be handicapped, because many of the previously unidentified DRACH sites would now be predicted as valid m⁶A sites. Hence, the final accuracy of the model was determined as the accuracy of the model to detect previously known m⁶A sites within the test data set.

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 GGAAT. 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 data set is ~10 min. For more information, visit <https://github.com/YeoLab/MINES.git>.

DATA DEPOSITION

MINES source code is available at <https://github.com/YeoLab/MINES.git>. Data files have been uploaded to GEO under accession number GSE132971.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

COMPETING INTEREST STATEMENT

G.W.Y. is cofounder, member of the Board of Directors, on the SAB, equity holder, and paid consultant for Locana and Eclipse BiolInnovations. 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.

ACKNOWLEDGMENTS

The authors thank Julia Nussbacher and Kris Brannan for help with HEK293T-shMETTL3 cell line generation and for providing HEK293 RNA, respectively. This work was partially supported by grants from the National Institutes of Health (HG004659 and HG009889 to G.W.Y., 2T32CA067754 to D.A.L.).

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.

Received August 11, 2019; accepted October 11, 2019.

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. doi:10.1021/acscchemneuro.8b00657
- Davis FF, Allen FW. 1957. Ribonucleic acids from yeast which contain a fifth nucleotide. *J Biol Chem* **227**: 907–915.
- Delaunay S, Frye M. 2019. RNA modifications regulating cell fate in cancer. *Nat Cell Biol* **21**: 552–559. doi:10.1038/s41556-019-0319-0
- Deng X, Su R, Weng H, Huang H, Li Z, Chen J. 2018. RNA N⁶-methyladenosine modification in cancers: current status and perspectives. *Cell Res* **28**: 507–517. doi:10.1038/s41422-018-0034-6
- 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. doi:10.1038/s41467-019-08734-9
- 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. *Nat Methods* **15**: 201–206. doi:10.1038/nmeth.4577
- Grozhi AV, Jaffrey SR. 2018. Distinguishing RNA modifications from noise in epitranscriptome maps. *Nat Chem Biol* **14**: 215–225. doi:10.1038/nchembio.2546
- Jonkhout N, Tran J, Smith MA, Schonrock N, Mattick JS, Novoa EM. 2017. The RNA modification landscape in human disease. *RNA* **23**: 1754–1769. doi:10.1261/ma.063503.117
- Ke S, Pandya-Jones A, Saito Y, Fak JJ, Vågbo 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 Dev* **31**: 990–1006. doi:10.1101/gad.301036.117
- 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. doi:10.1126/science.1212728
- Linder B, Grozhi AV, Olarerin-George AO, Meydan C, Mason CE, Jaffrey SR. 2015. Single-nucleotide-resolution mapping of m⁶A and m⁶Am throughout the transcriptome. *Nat Methods* **12**: 767–772. doi:10.1038/nmeth.3453
- 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. *Nat Chem Biol* **10**: 93–95. doi:10.1038/nchembio.1432
- 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. doi:10.1038/s41467-019-11713-9
- 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. *Nat Chem Biol* **15**: 88–94. doi:10.1038/s41589-018-0184-3
- McIntyre ABR, Alexander N, Grigorev K, Bezdán D, Sichtig H, Chiu CY, Mason CE. 2019. Single-molecule sequencing detection of N⁶-

- methyladenine in microbial reference materials. *Nat Commun* **10**: 579. doi:10.1038/s41467-019-08289-9
- Olarerin-George AO, Jaffrey SR. 2017. MetaPlotR: a Perl/R pipeline for plotting metagenes of nucleotide modifications and other transcriptomic sites. *Bioinformatics* **33**: 1563–1564. doi:10.1093/bioinformatics/btx002
- Payne A, Holmes N, Rakyan V, Loose M. 2019. BulkVis: a graphical viewer for Oxford nanopore bulk FAST5 files. *Bioinformatics* **35**: 2193–2198. doi:10.1093/bioinformatics/bty841
- 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. *J Mach Learn Res* **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. doi:10.1186/gb-2012-13-10-175
- Shi H, Wei J, He C. 2019. Where, when, and how: context-dependent functions of RNA methylation writers, readers, and erasers. *Mol Cell* **74**: 640–650. doi:10.1016/j.molcel.2019.04.025
- Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**: 407–410. doi:10.1038/nmeth.4184
- 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. bioRxiv doi:10.1101/459529
- Yue Y, Liu J, He C. 2015. RNA *N*⁶-methyladenosine methylation in post-transcriptional gene expression regulation. *Genes Dev* **29**: 1343–1355. doi:10.1101/gad.262766.115



RNA

A PUBLICATION OF THE RNA SOCIETY

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

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

RNA 2020 26: 19-28 originally published online October 17, 2019
Access the most recent version at doi:[10.1261/rna.072785.119](https://doi.org/10.1261/rna.072785.119)

Supplemental Material <http://rnajournal.cshlp.org/content/suppl/2019/10/17/rna.072785.119.DC1>

References This article cites 23 articles, 6 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/26/1/19.full.html#ref-list-1>

Open Access Freely available online through the *RNA* Open Access option.

Creative Commons License This article, published in *RNA*, is available under a Creative Commons License (Attribution 4.0 International), as described at <http://creativecommons.org/licenses/by/4.0/>.

Email Alerting Service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>