ARTICLES
https://doi.org/10.1038/s41587-021-00949-w
nature biotechnology

Check for updates

Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore

Ploy N. Pratanwanich[1,2,3] ✉, Fei Yao[1,11], Ying Chen[1,11], Casslynn W. Q. Koh[1,11], Yuk Kei Wan[1,11], Christopher Hendra[1,4], Polly Poon[1], Yeek Teck Goh[1], Phoebe M. L. Yap[1], Jing Yuan Chooi[5], Wee Joo Chng[5,6,7], Sarah B. Ng[1], Alexandre Thiery[8], W. S. Sho Goh[1,9] ✉ and Jonathan Göke[1,10] ✉

RNA modifications, such as *N*^{6}-methyladenosine (m^{6}A), modulate functions of cellular RNA species. However, quantifying differences in RNA modifications has been challenging. Here we develop a computational method, xPore, to identify differential RNA modifications from nanopore direct RNA sequencing (RNA-seq) data. We evaluate our method on transcriptome-wide m^{6}A profiling data, demonstrating that xPore identifies positions of m^{6}A sites at single-base resolution, estimates the fraction of modified RNA species in the cell and quantifies the differential modification rate across conditions. We apply xPore to direct RNA-seq data from six cell lines and multiple myeloma patient samples without a matched control sample and find that many m^{6}A sites are preserved across cell types, whereas a subset exhibit significant differences in their modification rates. Our results show that RNA modifications can be identified from direct RNA-seq data with high accuracy, enabling analysis of differential modifications and expression from a single high-throughput experiment.

The molecular profile of a cell includes not only the set of genes it expresses but also the >100 chemical modifications of its RNA species. RNA modifications are essential during early development[1–6], and aberrant modifications have been associated with disease[7–10]. RNA modifications affect several post-transcriptional processes, including mRNA decay, mRNA translation, pre-mRNA splicing, RNA localization and primary microRNA processing[11–17]. Therapeutics that target RNA-modifying enzymes are under development for cancer treatment[18], highlighting the importance of RNA modifications for precision medicine. Comprehensive profiling of the transcriptome therefore requires quantification of both transcript levels and modification rates.

Although quantification of transcript levels is readily achieved by sequencing cDNA with RNA-seq, identifying and quantifying RNA modifications remains a major challenge. Certain mRNA modifications (m^{6}A, 5-methylcytosine, 5-hydroxymethylcytosine) can be mapped using short-read cDNA-sequencing-based methods[19]. For m^{6}A, one of the most abundant known RNA modifications, cross-linking-immunoprecipitation (CLIP)-based methods[20] and deamination adjacent to RNA-modification target sequencing[21] map these modifications transcriptome wide, and m^{6}A-cross-linking-exonuclease sequencing (m^{6}ACE)-seq[15] additionally enables quantification of the modification rate of individual m^{6}A sites. However, cDNA-based methods use reverse transcription, which might introduce biases, they rely on available antibodies or known enzymes that limit profiling of most modifications, and the requirement for specialized protocols prevents their large-scale application.

Third-generation sequencing using Oxford Nanopore technology promises to overcome these limitations through direct sequencing of native RNA (direct RNA-seq)[22]. Direct RNA-seq applies a fundamentally unique principle for base identification: as RNA passes through the pore, magnitudes of electric intensity across the nanopore surface are recorded and used to identify the corresponding nucleotide sequence. RNA modifications cause shifts in intensity levels that are used to computationally identify modified bases[23,24]. Identification of m^{6}A modifications can be achieved using an approach that relies on basecalling accuracy[25] or training data from synthetic sequences[26]. An alternative approach is the detection of modifications by comparison to a matched unmodified control sample, thereby removing the requirements of training data and potentially enabling identification of non-m^{6}A modifications[23,24]. However, such samples are difficult to generate, representing a major barrier to using direct RNA-seq for profiling RNA modifications.

One of the main applications for transcriptome profiling is analysis of differential expression across conditions. Similarly, analysis of differential modifications would enable the study of RNA modifications in all possible scenarios. However, current computational methods only identify modified positions but do not quantify the modification rate, and the requirement of an unmodified sample severely limits the possible design for any comparative analysis. To address this, we developed xPore, a computational method and statistical framework that defines measures of significance and effect size for differential RNA modifications from direct RNA-seq data. xPore removes the requirement for an unmodified control sample and enables simultaneous profiling of differential transcript expression and modification across conditions without any additional experiments. To determine the proportion of RNA modifications across multiple samples, we fitted a multi-sample two-Gaussian

[1]Genome Institute of Singapore, A*STAR, Singapore, Singapore. [2]Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, Thailand. [3]Chula Intelligent and Complex Systems Research Unit, Chulalongkorn University, Bangkok, Thailand. [4]Institute of Data Science, National University of Singapore, Singapore, Singapore. [5]Cancer Science Institute of Singapore, National University of Singapore, Singapore, Singapore. [6]NUS Center for Cancer Research and Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. [7]Department of Haematology–Oncology, National University Cancer Institute, National University Health System, Singapore, Singapore. [8]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore. [9]Institute of Molecular Physiology, Shenzhen Bay Laboratory, Shenzhen, China. [10]National Cancer Centre Singapore, Singapore, Singapore. [11]These authors contributed equally: Fei Yao, Ying Chen, Casslynn W. Q. Koh, Yuk Kei Wan. ✉e-mail: naruemon.p@chula.ac.th; shogoh@szbl.ac.cn; gokej@gis.a-star.edu.sg

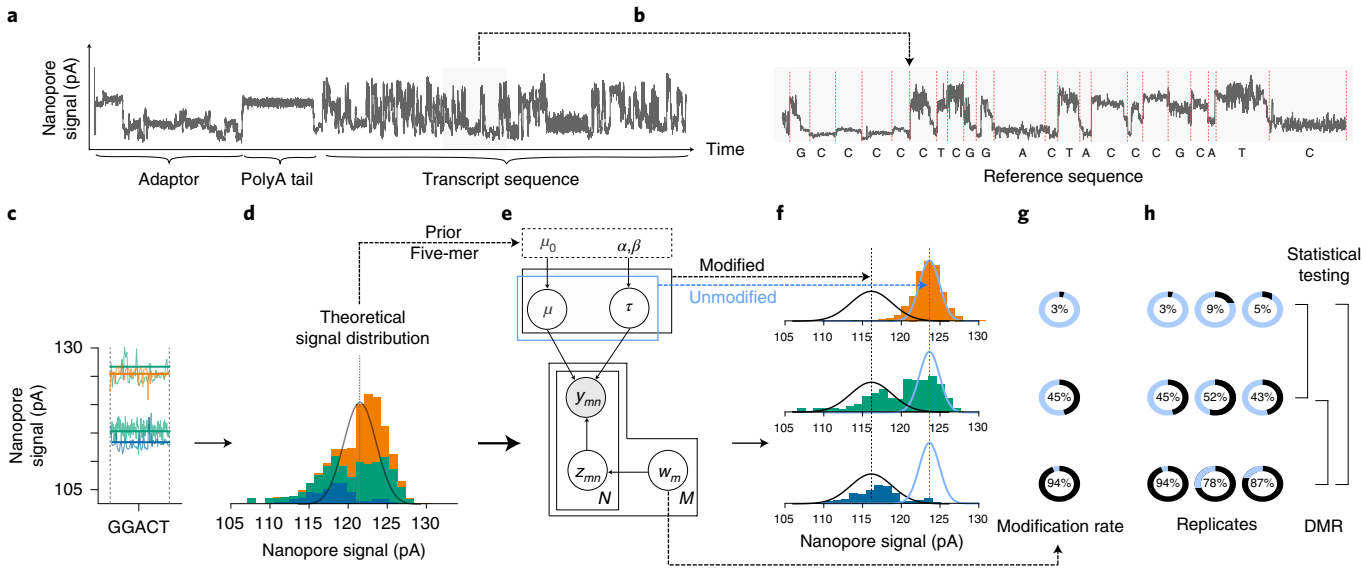

Fig. 1 | Schematic workflow: quantification of RNA modifications from direct RNA-seq data using xPore. a, Example of raw signal data from a direct RNA-seq read. **b**, A close-up view of the raw signal with the corresponding transcript sequence obtained from basecalling, sequence alignment and signal segmentation. **c**, Signal of multiple reads aligned at a GGACT site from different samples (orange, green and blue). **d**, Shown is a histogram of the mean signal from all reads covering a position for three different samples (orange, green and blue). The gray line indicates the expected distribution for unmodified RNA; samples that contain modified RNA species will show a bimodal distribution. **e**, Graphical representation of the model used by xPore to quantify the modification rate at each position. The gray circle indicates observed variables (data); white circles indicate unobserved variables that are estimated by xPore. **f**, xPore estimates the parameters for two Gaussian distributions corresponding to modified (black) and unmodified RNA species (blue). **g**, xPore summarizes the modification rate for each sample. **h**, xPore models the modification rate jointly for all samples. xPore then identifies differentially modified positions by testing for significant deviation in the DMR across replicates from the conditions of interest.

mixture model and inferred directionality of modification-rate differences by using information across all tested positions. We evaluated our method transcriptome wide in a loss-of-m^{6}A system after knockout of *METTL3*. We then applied our method to direct RNA-seq data from six of the most commonly used human cell lines including cancer tissues, providing insights into the dynamics of m^{6}A. Our study introduces a computational method that enables profiling of differential RNA modifications transcriptome wide and provides a systematic resource of direct RNA-seq data, which will be valuable as a benchmark dataset for modification detection.

Results

xPore: identification of differential RNA modifications. Nanopore direct RNA-seq generates a raw ionic current signal for each individual read (Fig. 1a). During the sequencing process, nanopores measure a signal corresponding to an RNA sequence five bases in length that resides in the pore, and a signal shift is observed when the next base enters the pore (Fig. 1b). Here we use the normalized mean signal information at each five-mer ('event') obtained after transcriptome alignment with Minimap2 (ref. [27]) and signal segmentation using Nanopolish Eventalign[28,29].

To detect a differential RNA-modification rate for each genomic position across samples (Fig. 1c), we developed xPore, a computational method that analyses direct RNA-seq data at the signal level. xPore models a mixture of two Gaussian distributions, corresponding to unmodified and modified RNA species. It uses prior information regarding the theoretical signal distribution of unmodified RNA species to guide the model estimation of Gaussian parameters (Fig. 1d and Methods). Means and variances of these two distributions are modeled to be shared across samples (Fig. 1e and Methods). After a few iterations of variational Bayesian inference[30], inferred means and variances are obtained (Fig. 1f). We then assign the inferred distribution that is closer to the theoretical mean as unmodified and assign the other distribution as modified. xPore also learns the modification probability of each read, which allows us to compute the fraction of modified reads as an estimate of the modification rate per sample (Fig. 1g).

To increase precision and control the number of false positives, we implemented two filtering steps after model fitting. First, we exclude positions where distributions for unmodified and modified signals are nearly identical, thereby reducing the number of tests and increasing power. Second, RNA modifications induce a systematic shift in the signal for each *k*-mer, as the same RNA modification will either increase or decrease the signal but not both. Although xPore does not identify the type of modification at each position, we can nevertheless restrict the analysis to a single modification at each *k*-mer by only considering one-directional signal shifts (Supplementary Fig. 1). This filter removes outliers and enables transcriptome-wide comparison of modification patterns. On the remaining sites, we perform a statistical test on differential modification rates (DMR) between samples and prioritize differentially modified sites accordingly (Fig. 1h).

The method is implemented in Python and is available as part of the open source package xPore on GitHub (https://github.com/GoekeLab/xpore). Using FAST5 files as input, xPore returns a table summarizing means and variances of unmodified and modified distributions, assignment confidence levels, modification rates for each sample and test statistics for individual positions.

xPore identifies m^{6}A sites at single-base resolution. One of the most abundant and best studied RNA modifications is m^{6}A[19]. To evaluate the ability of our method to detect differentially modified sites in the human transcriptome, we compared wild-type (WT) HEK293T cells ('WT cells') with cells in which expression of the main m^{6}A writer METTL3 was deleted via CRISPR–Cas9 ('knockout (KO) cells'). We generated three replicates for both WT and KO cells, resulting in nearly 8 million reads in total. After filtering out low-coverage positions, we had over 9 million sites to be modeled,

939,902 of which were passed through the post-modeling filter and were tested for differential modifications (see Supplementary Data 1 for the top significant differentially modified positions). For evaluation, we used single-base-resolution m^{6}ACE-seq and the presence of the DRACH motif to estimate the number of correctly predicted sites (Fig. 2a).

First, we evaluated the ability to identify differentially m^{6}A-modified sites at positions with the base A (NNANN). Among all tested positions, xPore achieves an overall area under the curve (AUC) of 0.86 when m^{6}ACE-seq is used as a reference (Fig. 2b). As the number of unmodified positions is much larger than the number of modified positions, we particularly investigated the precision of our predictions at different levels of sensitivity (Fig. 2c). The model had a precision of 0.60 at the top predictions when m^{6}ACE-seq was used as an m^{6}A reference. Compared to existing supervised[26,31], unsupervised[23] or comparative approaches[23], we found that xPore had a higher precision and recall while still being computationally more efficient (Supplementary Fig. 2, Methods and Supplementary Text). Strikingly, many of the top positions ranked by xPore that were not identified as modified by m^{6}ACE-seq still had the DRACH motif (Fig. 2d). A similar result was obtained when mapping m^{6}A at individual-nucleotide resolution using CLIP (miCLIP) or RNA digestion via m^{6}A-sensitive RNase (MAZTER)-seq data were used as reference data, suggesting that nanopore direct RNA-seq might help to identify a different set of modified sites that had been otherwise missed by antibody-based detection methods (Supplementary Fig. 3). One of the hallmarks of m^{6}A sites is the distribution along the transcript that is specifically enriched close to the stop codon[15,20]. The top ranked positions (P value < 0.001) clustered as expected of m^{6}A sites, further suggesting that m^{6}A positions identified by xPore contain only a small number of false positives (Fig. 2e). Indeed, when we combined m^{6}ACE-seq data and motif occurrences, xPore achieved an accuracy of >90% among the top significant 1,452 positions (P value < 0.001), indicating that our method can successfully identify differential m^{6}A modifications even in a search space that covers hundreds of thousands of positions (Fig. 2f).

Next, we tested the accuracy to identify differentially modified sites when the sequence context is unknown. For this evaluation, we ranked transcriptome-wide differentially modified sites in KO and WT cells, including A and non-A nucleotide positions (all k-mers) and again evaluated performance using m^{6}ACE-seq and motif content. Overall, xPore achieved an AUC of 0.86 (Fig. 2g). Compared to the A nucleotide analysis, we observed a lower precision at the same level of sensitivity (Fig. 2h). A number of m^{6}A sites that were labeled as false positives were neighboring positions next to modified m^{6}A sites, likely caused by a signal shift due to the modification (Fig. 2i). Among the top 1,500 positions, 90% were within a single-base distance of a DRACH motif or a validated m^{6}A site, demonstrating that an unbiased search for differential modifications still provides results with high precision among the top ranking positions (Fig. 2j).

Replicates increase precision. Here we analyzed data using biological replicates, which might not always be available. To evaluate the performance of xPore in the absence of replicates, we tested every pair of HEK293T WT and KO cells, resulting in nine pairwise comparisons. Single-replicate results were generally less accurate and showed higher variation compared to those from the multi-replicate analysis (Fig. 2k). These results suggest that, even in the absence of replicates, xPore can prioritize differentially modified sites; however, replicates, which naturally account for biological variation, are recommended to obtain more precise results.

Pooling data increases sensitivity. The number of positions that can be tested is limited by the sequencing depth of each sample, with lowly expressed genes being potentially excluded. To maximize the number of genes that are modeled by xPore, we therefore combined reads across replicates within the same condition. Using pooled data, we evaluated the performance for different read-coverage thresholds. As expected, we observed that a lower read-coverage threshold reduces precision (Fig. 2l,m). However, a threshold of 30 reads per position increased the number of genes being tested to more than 5,000, and a threshold of 15 enabled analysis of more than 7,000 genes (Fig. 2n). While analysis using individual replicates with a higher threshold has higher precision, pooling data increases the number of genes that are tested for differential modifications, enabling detection of RNA modifications at even genes that are lowly expressed.

xPore identifies modified positions with low stoichiometry. The proportion of RNA molecules that are modified (the stoichiometry or modification rate) can show high levels of variation across positions and samples. While some sites are modified in all RNA molecules, others seem to be modified in only a minor fraction[32]. To investigate the sensitivity of xPore to detect m^{6}A-modified positions for different levels of m^{6}A, we generated RNA mixtures with known expected average modification rates. To achieve this, we combined various proportions of WT and *METTL3*-KO cellular RNA before profiling them in multiple replicates using direct RNA-seq (Fig. 3a,b). We then compared these mixtures to KO cells to evaluate the ability to detect m^{6}A positions when the expected average modification rate was 25%, 50%, 75% or 100% of the levels observed in WT. Overall, we found that xPore achieved a similar level of precision for the most highly ranked positions for all mixtures, indicating that xPore can still detect m^{6}A at an expected modification rate of 25% (Fig. 3c). However, as expected, we observed that a higher expected modification rate resulted in a better recall (Fig. 3c). This particularly affects samples with an expected modification rate of 25%, in which a larger fraction of RNA species are likely to fall below the detection threshold. While we cannot determine the precise detection threshold (as 25% is the lowest mixture fraction), these data indicate that xPore can still accurately detect RNA modifications with modification rates higher than 25%. Of note, xPore can still identify significant differences of 25% and lower across conditions if the modification rate is sufficiently high (Supplementary Fig. 4).

To further illustrate the relevance of this observation, we generated a partial loss of m^{6}A using small interfering (si)RNA-mediated knockdown (KD) of *METTL3*. We then compared the ability to detect m^{6}A using either KD or KO samples (Fig. 3d). Results indicate that xPore can detect modified positions using partial loss of m^{6}A but with a lower recall, as expected. Interestingly, xPore also detects m^{6}A positions when KD and KO samples are compared, demonstrating the ability to detect m^{6}A positions that only show quantitative differences (Fig. 3d).

Quantitative estimation of RNA-modification rates. Quantitatively estimating these modification rates remains one of the major challenges[19]. xPore was designed to intrinsically model the probability for each read of being modified. This property enables us to calculate the fraction of reads that are assigned to the modified signal distribution, directly providing an estimate of cellular RNA-modification rates.

To evaluate whether estimated modification rates from xPore approximate the true proportion of modified reads, we analyzed the mixtures with known expected modification rates. Because not all positions are modified at 100% in WT cells, we investigated the expected relative modification rates, using WT cells and KO cells as reference points (see Supplementary Data 2 for differentially modified sites across RNA-mixture samples). Estimated global modification rates from xPore were, on average, within 7% of the expected modification rate from the RNA mixtures (Fig. 3e). Estimated modification rates across positions for each k-mer showed expected

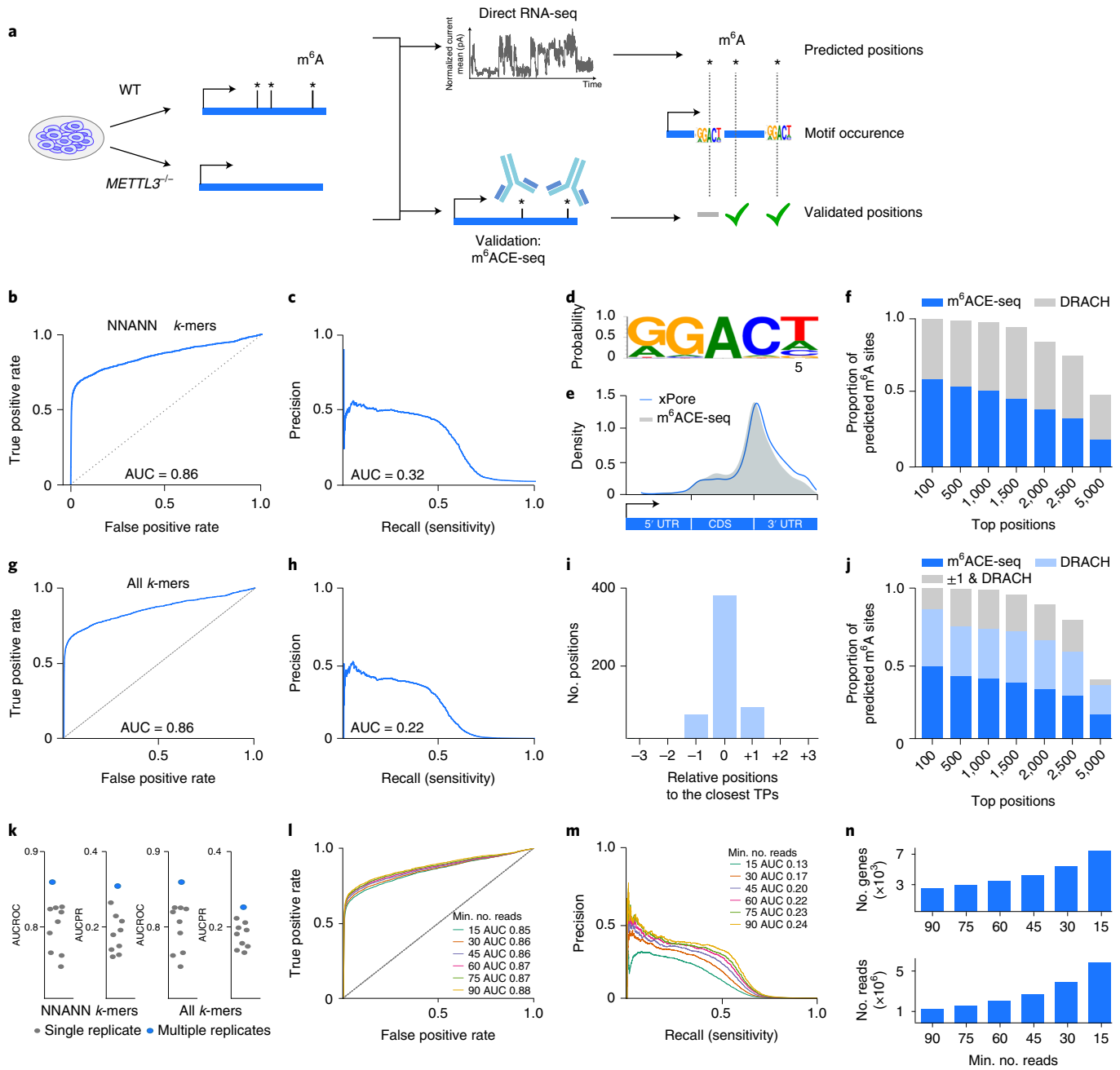

Fig. 2 | Detection of m⁶A sites in the human transcriptome. a, Experimental design: RNA from WT and *METTL3*-KO HEK293T cells is sequenced by direct RNA-seq, and differential m⁶A modification rates between both cell types are estimated using xPore. Results are validated against those from m⁶ACE-seq and tested for occurrence of the m⁶A motif (DRACH). Receiver operating characteristic (ROC) curve (**b**) and precision–recall curve (**c**) for candidate m⁶A sites identified by xPore using the set of m⁶ACE-seq sites as ground truth. Only *k*-mers with an A at the center are used in this analysis. **d**, *k*-mers from the top 1,452 differentially modified sites ($P < 0.001$) identified by xPore resemble the expected m⁶A motif. The *P* value was calculated from a two-tailed, unpooled *z*-test on the modification-rate difference, adjusted for multiple comparisons using the Benjamini–Hochberg procedure. **e**, These 1,452 sites are enriched at the 3′ end of the coding sequence (CDS), resembling the expected distribution shown for m⁶ACE-seq results. **f**, Proportion of predicted m⁶A sites that overlap with m⁶ACE-seq sites (dark blue) and resemble the DRACH motif but do not overlap with a site identified by m⁶ACE-seq (gray). ROC curve (**g**) and precision–recall curve (**h**) for candidate m⁶A sites identified by xPore using the set of m⁶ACE-seq sites as ground truth. All *k*-mers are used in this analysis. **i**, Relative positions of predicted m⁶A sites to the closest site validated by m⁶ACE-seq. TP, true positive. **j**, Proportion of predicted m⁶A sites that overlap with m⁶ACE-seq sites (dark blue), resemble the DRACH motif but do not overlap with an m⁶ACE-seq site (light blue) and resemble the DRACH motif within a distance of 1 base (gray). **k**, Area under the ROC curve and precision–recall curve (AUCROC and AUCPR) when three replicates are used (blue) compared to a single-replicate analysis (gray). Left, *k*-mers with A in the center. Right, all *k*-mers. ROC curve (**l**) and precision–recall curve (**m**) when the direct RNA-seq data from multiple replicates are pooled. Using different thresholds for the minimum (min.) number of reads influences sensitivity and precision. **n**, Read coverage of genes (bottom) and number of genes (top) that can be analyzed at these thresholds. xPore can analyze more than 7,000 genes at a minimum read coverage of 15.

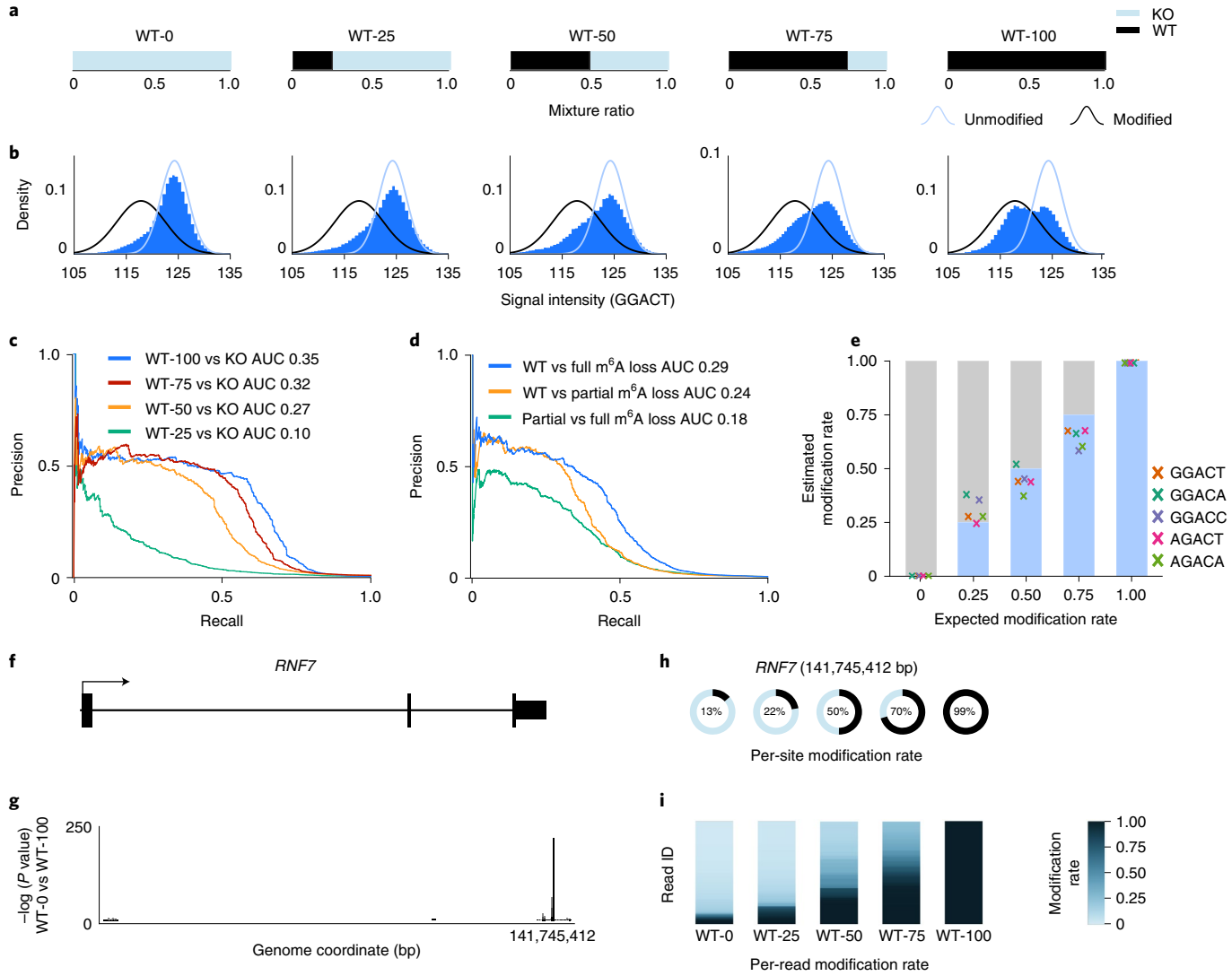

Fig. 3 | xPore modification-rate estimates correspond to the fraction of modified RNA species in the cell. a, Percentage of WT and *METTL3*-KO HEK293T cells in each mixture sample. **b**, Current-intensity-level means across all GGACT positions. The black line shows the estimated signal distribution for modified reads from xPore; the blue line shows the estimated signal distribution for unmodified reads. **c**, Precision–recall curves for candidate m^{6}A sites identified by xPore at different levels of the expected DMR using the set of m^{6}ACE-seq sites as ground truth. **d**, Precision–recall curves for candidate m^{6}A sites identified by xPore for each pairwise comparison between HEK293T WT, HEK293T KD (partial loss of m^{6}A) and HEK293T KO (full loss of m^{6}A) cells using the set of m^{6}ACE-seq sites as ground truth. Only *k*-mers with an A at the center are used in this analysis. **e**, Estimated relative modification rates of the most frequently modified m^{6}A motifs across all modified positions identified by xPore, shown for the different mixture samples. Example of a protein-coding gene, *RNF7*, (**f**) and the −log (*P* value) obtained from the pairwise comparison of 0% WT and 100% WT mixtures showing the most significant differential m^{6}A site (141,745,412 bp) (**g**). Per-site modification-rate estimates (**h**) and per-read modification rates of each mixture sample at the identified GGm^{6}ACT site (141,745,412 bp) of *RNF7* (**i**).

variation, as the RNA-mixing procedure did not control the modification rate at each individual site (Supplementary Fig. 5). A comparison of MAZTER-seq or m^{6}ACE-seq data with that from xPore showed a positive correlation of quantification estimates in HEK293T cells (Supplementary Fig. 3d,e). Together, these results suggest that, transcriptome wide, xPore accurately approximates modification rates from direct RNA-seq data at single-base resolution (Fig. 3f–i).

DMRs as estimates of effect size. The ability to estimate modification rates allows us to not only identify positions that are modified but also to quantify DMRs across conditions. Here we define the DMR as the difference between modification rates observed within each condition. Using the expected signal of unmodified RNA species as a reference, we then infer the directionality of the change, allowing us to discriminate positions that show gain and loss of RNA modifications across conditions. As modification rates correspond to the fraction of modified RNA species in the cell, the DMR can be directly interpreted as a quantitative estimate of effect size. This property enables us to study the effect of any experimental design on the stoichiometry of RNA modifications.

To demonstrate how the DMR can be interpreted, we first compared HEK293T KO and WT cells. We identified 1,923 significantly differentially modified positions at NNANN ($P < 0.001$, DMR > 0.5) (Fig. 4a). Among these, more than 90% were m^{6}A DRACH motifs (Fig. 4b). A transcriptome-wide comparison of modification rates for the most frequently changed *k*-mers demonstrates the ability to quantitatively identify RNA modifications from direct RNA-seq

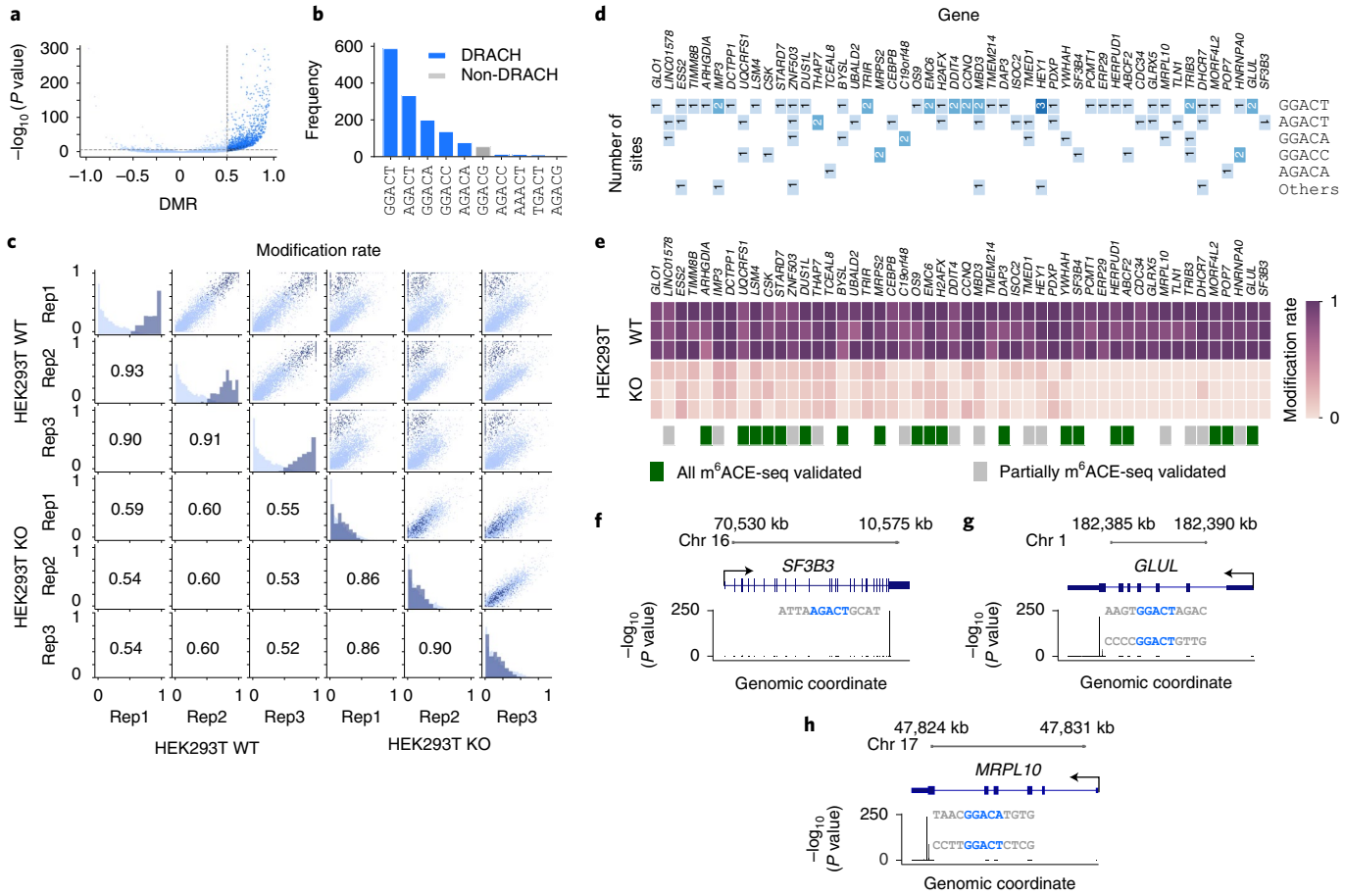

Fig. 4 | Transcriptome-wide identification of differentially modified positions. a, Shown are *P* values and DMRs for the comparison of HEK293T WT and HEK293T KO cells at A-centered *k*-mers. **b**, Frequency of the top ten *k*-mers at significantly differentially modified positions. **c**, Scatterplot comparing modification-rate estimates for HEK293T WT and KO samples, histogram of the distribution of modification rates and pairwise correlation coefficients. These non-significant and significant positions are colored in light blue and dark blue, respectively. Rep, replicate. The number of modified A sites of the top significant genes ranked by DMRs (**d**), along with the corresponding modification rates estimated by xPore across HEK23T samples (**e**). Identified differentially modified sites were all confirmed by m^{6}ACE-seq in some genes (green) and partially confirmed with newly identified A-modified sites in others (gray). **f**–**h**, Examples of the top ranked genes with corresponding *P* values and transcript sequences for the identified m^{6}A sites. The *P* values in **a**,**f**–**h** were calculated from two-tailed, unpooled *z*-tests on modification-rate differences and adjusted for multiple comparisons using the Benjamini–Hochberg procedure. Chr, chromosome.

data: modification rates within replicates are highly similar (correlation coefficient >0.9 for WT cells), whereas a systematic increase in modified sites can be observed across conditions (Fig. 4c).

Next, we identified the set of genes that show the largest change in RNA modifications after knockout of *METTL3*. We ranked genes by the highest DMR that was consistently found across replicates (Fig. 4d). Many genes appeared to contain multiple m^{6}A modifications, frequently involving different *k*-mers (Fig. 4d). Many of these positions were confirmed by m^{6}ACE-seq; however, xPore identified a number of new positions (Fig. 4d,e). Among the genes that are heavily modified by m^{6}A are genes related to RNA processing such as *SF3B3* (Fig. 4f), genes in the glutamine synthetase family such as *GLUL* (Fig. 4g) and those encoding ribosomal proteins such as *MRPL10* (Fig. 4h).

A comparison with *METTL3*-KD samples further illustrates the importance of having estimates for significance (*P* value) and effect size (DMR). While the positions that were identified as significant in KD and KO samples were largely similar (Supplementary Data 3, 90% for the top 100 positions, 81% for the top 1,000 positions), DMRs were substantially different (Supplementary Fig. 6). By providing estimates for the DMR, xPore enables not only identification of modified positions but also quantification of differences that can be essential in biology.

Identification of m^{6}A across genetically diverse cell lines. RNA modifications can be detected from direct RNA-seq as they induce a systematic shift in the signal. However, genetic variants will similarly induce a change in the signal, possibly confounding results. This effect can be avoided by comparing cell lines that are genetically similar. While this is often the case in perturbation experiments, having such a requirement would restrict analysis to mostly in vitro applications. To test whether xPore can identify differential RNA modifications in samples with a genetically different background, we compared HEK293T KO cells with five WT cell lines from the Singapore Nanopore-Expression project (SG-NEx project)[33], covering liver cancer cells (HEPG2), colon cancer cells (HCT116), breast cancer cells (MCF7), lung adenocarcinoma cells (A549) and leukemia cells (K562) (Fig. 5a, 'modified vs unmatched unmodified control'). To identify changes in RNA modification, we focused on the set of RNA species that were expressed across conditions. When we looked into modification rates at significantly differentially modified positions, loss of *METTL3* was clearly visible (Fig. 5b).

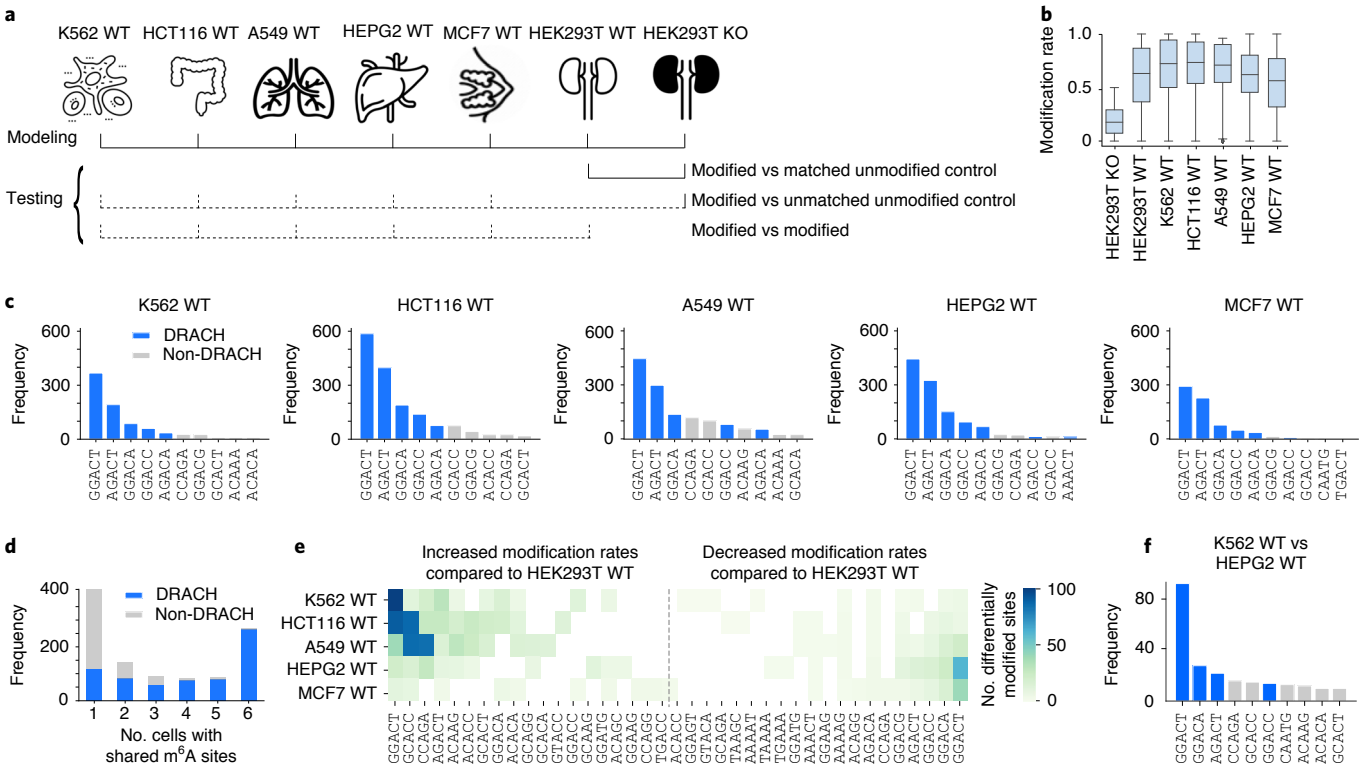

Fig. 5 | Identification of m⁶A sites across different tissues and cell lines. a, Experimental design illustrating the modeling step in which all samples across diverse cell lines are modeled simultaneously while statistical tests can be performed on any pairwise comparison including HEK293T WT versus HEK293T KO (modified and matched unmodified control) cells, comparison between one of the cancer cell lines and HEK293T KO cells (modified and unmatched unmodified control) and comparison between any WT cell lines (modified and modified). **b**, Average modification rates of significantly differentially modified positions compared with HEK293T KO cells. Boxplots display distributions of modification rates estimated for each cell line, where each box indicates 1.5 interquartile range (IQR) below the first quartile (Q1), Q1, median, the third quartile (Q3) and 1.5 IQR above Q3. **c**, Frequency of *k*-mers at significantly differentially modified sites for each cancer cell line compared with that of HEK293T KO cells; DRACH motifs are shown in blue, and non-DRACH motifs are shown in gray. **d**, Number of m⁶A sites identified in a single sample and over different numbers of samples. Most DRACH m⁶A sites are shared across multiple samples. **e**, Heatmap showing the number of significantly differentially modified *k*-mers for each cancer cell line compared with HEK293T WT cells: increased modification rates (left) and decreased modification rates (right). **f**, Frequency of the *k*-mers at significantly differentially modified sites between K562 WT and HEPG2 WT cells when jointly modeled with HEK293T KO samples in the modeling; DRACH motifs are shown in blue, and non-DRACH motifs are shown in gray.

For all cell lines, we identified between 800 and 2,000 differentially modified sites compared with control samples (Supplementary Data 4), the vast majority (71% on average) representing the m⁶A DRACH motif (Fig. 5c). These results demonstrate that xPore can identify RNA modifications even when samples have a distinct genetic background.

Variation of m⁶A across different cell lines. It was shown that m⁶A modifications differ across tissues and developmental stages, yet quantifying such changes has been challenging[34]. xPore enables comparison of such samples (Fig. 5a, 'modified vs modified'). Here we used the ability of xPore to include a variety of conditions for more precise estimation of model parameters while testing only specific conditions of interest for differential modifications. To illustrate this, we investigated the dynamics of m⁶A across the different tissues represented by SG-NEx cell lines. Globally, we found that m⁶A was stable across cell lines, with most positions being shared, possibly due to similarities of these cell lines, with primary tissues expected to show higher variation (Fig. 5d). While complete loss or gain of m⁶A was rare, we observed that a number of m⁶A sites showed quantitative differences between cells. Interestingly, the global modification rate for m⁶A appeared to differ between cells, with K562 cells showing the highest number of modified m⁶A sites compared with other cell lines (Fig. 5e). Here we used HEK293T WT cells as a reference to reduce the number of pairwise comparisons. However, a comparison of other cell lines that did not have any matched unmodified control sample in the modeling step confirmed these results (Fig. 5f and Supplementary Fig. 7 without the presence of HEK293T KO samples in the modeling), demonstrating that xPore can quantify differential RNA modifications in unperturbed samples.

Identification of m⁶A in clinical cancer samples. Clinical samples, in addition to having higher genetic diversity, are often limited by the amount of RNA that can be extracted. For high-quality direct RNA-seq, 500 ng polyA RNA is recommended, which can require more than 50 µg total RNA. To test whether RNA modifications can be identified in genetically diverse clinical samples with a low amount of RNA, we generated direct RNA-seq data from three multiple myeloma patient samples using only 5% of the recommended amount of RNA (2.5 µg) (Fig. 6a). In total, we obtained more than 1.8 million reads. When we compared data from these clinical samples to data from the *METTL3*-KO cell line, we observed that a lower number of positions were identified as significantly different relative to those in cell line samples (Supplementary Data 5). However, the top positions were

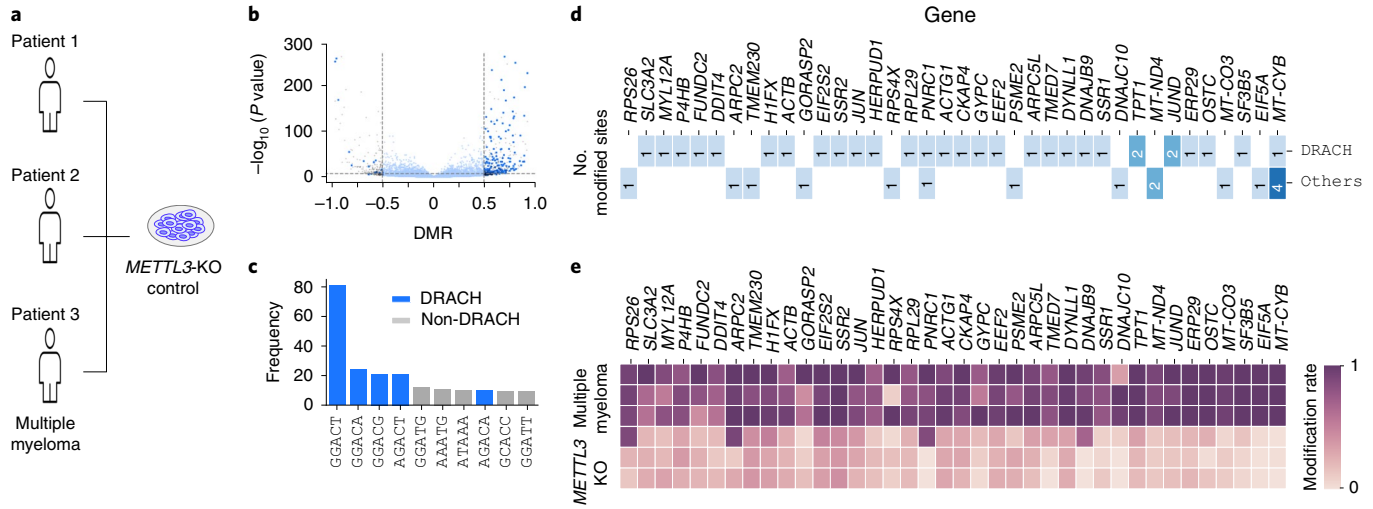

Fig. 6 | Identification of m⁶A in clinical samples using direct RNA-seq. a, RNA reads of clinical myeloma samples from three patients and those of *METTL3*-KO control samples from HEK293T cells are modeled using xPore. **b**, Shown are *P* values and DMRs for the comparison of *METTL3*-KO and multiple myeloma samples at A-centered *k*-mers. *P* values were calculated from two-tailed, unpooled *z*-tests on modification-rate differences and adjusted for multiple comparisons using the Benjamini–Hochberg procedure. **c**, Frequency of the top ten *k*-mers at significantly differentially modified positions. Number of modified A sites of the top significant genes ranked by DMRs (**d**), along with corresponding modification rates estimated by xPore across *METTL3*-KO and multiple myeloma samples (**e**).

similarly enriched in the DRACH motif (Fig. 6b,c), enabling analysis of m⁶A-modified genes in multiple myeloma samples (Fig. 6d,e). These data suggest that direct RNA-seq data can be used to identify differential RNA modifications even with genetically diverse conditions and limited RNA, opening opportunities to analyze clinical samples on a larger scale.

Discussion

Differential analysis of transcription is frequently used to understand cellular states, the impact of perturbation experiments and alterations due to diseases. Here we introduce xPore, a computational method that enables analysis of differential RNA modifications from direct RNA-seq data, opening a new layer of information that complements transcript expression profiles.

Experimental methods to detect RNA modifications such as m⁶A have been developed using different approaches[15,20,21,35–39]. Some of the most recent methods enable identification of thousands of sites at base resolution and allow quantification of modification rates[15,38]. The main limitation of these approaches is the requirement for sometimes extensive experimental procedures. By contrast, direct RNA-seq promises to enable analysis of RNA modifications from a single sequencing experiment[22]. Methods using direct RNA-seq have demonstrated the ability to identify m⁶A and other modifications[23,24,26,31,40–42]. However, they often have specific requirements, such as a control sample that lacks the modification of interest, and they do not quantify the stoichiometry of RNA modifications[23,24]. By contrast, xPore achieves base resolution for identification of m⁶A while estimating the modification rate, enabling quantitative comparison of samples across conditions even in the absence of a control sample.

Unlike other methods designed to identify RNA modification using strict case–control comparisons, xPore is not limited to such paired designs. xPore can use any combination of samples in the modeling step, while allowing any combination of sample groups to be compared in the (post-modeling) testing step. For example, here we jointly model RNA-modification rates across 18 samples from six cell lines and an unmatched loss-of-m⁶A control cell line. This approach enables xPore to use a large amount of data, it facilitates the use of an unmatched control, which is essential for patient samples and primary tissues, and it provides complete flexibility to test any comparison of interest.

The positions that are identified by xPore show a strong enrichment in the m⁶A DRACH motif and high validation rates with independent protocols, indicating high precision of predicted positions. Globally, xPore predicts a smaller number of m⁶A sites compared to other experimental approaches, partially due to stringent filtering that avoids false positives. With increased sequencing throughput and additional replicates, the number of predicted sites can likely be increased further while maintaining high levels of accuracy. Despite the smaller number of m⁶A sites identified by xPore, many have not been reported in studies using other protocols. Therefore direct RNA-seq not only provides a simplified approach to profiling RNA modifications but also identifies new sites that may be missed by complementary approaches.

Direct RNA-seq has been used to analyze m⁶A in yeast[22,26], *Arabidopsis*[25], RNA virus genomes[43] and human cells[24,31,44]. Here we showed that differential RNA modifications can be identified across a larger number of genetically diverse human cancer cell lines and patient samples. Even for m⁶A sites that were significantly different across cell lines, we found that a certain level of m⁶A was still preserved. While these differences are likely to increase when more diverse samples or primary tissues are analyzed, our results suggest that estimation of the modification rate will be key to understanding the dynamics of m⁶A. With direct RNA-seq becoming widely available, we propose that differential modification analysis can complement differential expression analysis and provide insights into the complex landscape of RNA modifications and their roles in diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-00949-w.

Received: 10 July 2020; Accepted: 10 May 2021;
Published online: 19 July 2021

References

1. Zheng, G. et al. ALKBH5 is a mammalian RNA demethylase that impacts RNA metabolism and mouse fertility. *Mol. Cell* **49**, 18–29 (2013).
2. Wang, Y. et al. N^6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nat. Cell Biol.* **16**, 191–198 (2014).
3. Zhao, X. et al. FTO-dependent demethylation of N^6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Res.* **24**, 1403–1419 (2014).
4. Geula, S. et al. Stem cells. m^{6}A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science* **347**, 1002–1006 (2015).
5. Chen, T. et al. m^{6}A RNA methylation is regulated by microRNAs and promotes reprogramming to pluripotency. *Cell Stem Cell* **16**, 289–301 (2015).
6. Xu, K. et al. Mettl3-mediated m^{6}A regulates spermatogonial differentiation and meiosis initiation. *Cell Res.* **27**, 1100–1114 (2017).
7. Mathiyalagan, P. et al. FTO-dependent N^6-methyladenosine regulates cardiac function during remodeling and repair. *Circulation* **139**, 518–532 (2019).
8. Li, Z. et al. FTO plays an oncogenic role in acute myeloid leukemia as a N^6-methyladenosine RNA demethylase. *Cancer Cell* **31**, 127–141 (2017).
9. Su, R. et al. R-2HG exhibits anti-tumor activity by targeting FTO/mA/MYC/CEBPA signaling. *Cell* **172**, 90–105 (2018).
10. Deng, X. et al. RNA N^6-methyladenosine modification in cancers: current status and perspectives. *Cell Res.* **28**, 507–517 (2018).
11. Wang, X. et al. N^6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117–120 (2014).
12. Meyer, K. D. et al. 5′ UTR m^{6}A promotes cap-independent translation. *Cell* **163**, 999–1010 (2015).
13. Alarcón, C. R., Lee, H., Goodarzi, H., Halberg, N. & Tavazoie, S. F. N^6-methyladenosine marks primary microRNAs for processing. *Nature* **519**, 482–485 (2015).
14. Alarcón, C. R. et al. HNRNPA2B1 is a mediator of m^{6}A-dependent nuclear RNA processing events. *Cell* **162**, 1299–1308 (2015).
15. Koh, C. W. Q., Goh, Y. T. & Goh, W. S. S. Atlas of quantitative single-base-resolution N^6-methyl-adenine methylomes. *Nat. Commun.* **10**, 5636 (2019).
16. Zaccara, S. & Jaffrey, S. R. A unified model for the function of YTHDF proteins in regulating m^{6}A-modified mRNA. *Cell* **181**, 1582–1595 (2020).
17. Goh, Y. T., Koh, C. W. Q., Sim, D. Y., Roca, X. & Goh, W. S. S. METTL4 catalyzes m^{6}Am methylation in U2 snRNA to regulate pre-mRNA splicing. *Nucleic Acids Res.* **48**, 9250–9261 (2020).
18. Yankova, E. et al. Small-molecule inhibition of METTL3 as a strategy against myeloid leukaemia. *Nature* **593**, 597–601 (2021).
19. Zaccara, S., Ries, R. J. & Jaffrey, S. R. Reading, writing and erasing mRNA methylation. *Nat. Rev. Mol. Cell Biol.* **20**, 608–624 (2019).
20. Linder, B. et al. Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome. *Nat. Methods* **12**, 767–772 (2015).
21. Meyer, K. D. DART-seq: an antibody-free method for global m6A detection. *Nat. Methods* **16**, 1275–1280 (2019).
22. Garalde, D. R. et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat. Methods* **15**, 201–206 (2018).
23. Stoiber, M. et al. De novo identification of DNA modifications enabled by genome-guided nanopore signal processing. Preprint at *bioRxiv* https://doi.org/10.1101/094672 (2017).
24. Leger, A. et al. RNA modifications detection by comparative Nanopore direct RNA sequencing. Preprint at *bioRxiv* https://doi.org/10.1101/843136 (2019).
25. Parker, M. T. et al. Nanopore direct RNA sequencing maps the complexity of *Arabidopsis* mRNA processing and m^{6}A modification. *eLife* **9**, e49658 (2020).
26. Liu, H. et al. Accurate detection of m^{6}A RNA modifications in native RNA sequences. *Nat. Commun.* **10**, 4079 (2019).
27. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
28. Loman, N. J., Quick, J. & Simpson, J. T. A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat. Methods* **12**, 733–735 (2015).
29. Simpson, J. T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat. Methods* **14**, 407–410 (2017).
30. Corduneanu, A. & Bishop, C. M. Variational Bayesian model selection for mixture distributions. in *Proceedings of the Eighth International Conference on Artificial Intelligence and Statistics* 27–34 (Morgan Kaufmann, 2001).
31. Lorenz, D. A., Sathe, S., Einstein, J. M. & Yeo, G. W. Direct RNA sequencing enables m^{6}A detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**, 19–28 (2020).
32. Liu, N. et al. Probing N^6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA* **19**, 1848–1856 (2013).
33. Chen, Y. et al. A systematic benchmark of Nanopore long read RNA sequencing for transcript level analysis in human cell lines. Preprint at *bioRxiv* https://doi.org/10.1101/2021.04.21.440736 (2021).
34. McIntyre, A. B. R. et al. Limits in the detection of m^{6}A changes using MeRIP/m^{6}A-seq. *Sci. Rep.* **10**, 6590 (2020).
35. Dominissini, D. et al. Topology of the human and mouse m^{6}A RNA methylomes revealed by m^{6}A-seq. *Nature* **485**, 201–206 (2012).
36. Meyer, K. D. et al. Comprehensive analysis of mRNA methylation reveals enrichment in 3′ UTRs and near stop codons. *Cell* **149**, 1635–1646 (2012).
37. Ke, S. et al. A majority of m^{6}A residues are in the last exons, allowing the potential for 3′ UTR regulation. *Genes Dev.* **29**, 2037–2053 (2015).
38. Garcia-Campos, M. A. et al. Deciphering the 'm^{6}A code' via antibody-independent quantitative profiling. *Cell* **178**, 731–747 (2019).
39. Shu, X. et al. A metabolic labeling method detects m^{6}A transcriptome-wide at single base resolution. *Nat. Chem. Biol.* **16**, 887–895 (2020).
40. Ueda, H. nanoDoc: RNA modification detection using Nanopore raw reads with Deep One-Class Classification. Preprint at *bioRxiv* https://doi.org/10.1101/2020.09.13.295089 (2020).
41. Ding, H., Bailey, A. D., Jain, M., Olsen, H. & Paten, B. Gaussian mixture model-based unsupervised nucleotide modification number detection using nanopore-sequencing readouts. *Bioinformatics* **36**, 4928–4934 (2020).
42. Price, A. M. et al. Direct RNA sequencing reveals m^{6}A modifications on adenovirus RNA are necessary for efficient splicing. *Nat. Commun.* **11**, 6016 (2020).
43. Kim, D. et al. The architecture of SARS-CoV-2 transcriptome. *Cell* **181**, 914–921 (2020).
44. Workman, R. E. et al. Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nat. Methods* **16**, 1297–1305 (2019).

Methods

xPore: a multi-sample two-Gaussian mixture model. When a modified *k*-mer passes through the pore, it causes the current intensity to differ from its canonical counterpart, enabling detection of modification at the signal level from direct RNA-seq data. Based solely on current-intensity levels, we aim to identify differentially modified positions between samples quantitatively.

Modeling a collection of transcripts at a single genomic site, we assume two distributions to correspond to the unmodified and modified RNA species that are shared across samples and allow individual reads to fit both distributions with different degrees. With this assumption, we extend a standard two-Gaussian mixture model to support multiple-sample comparison simultaneously. The multi-sample two-Gaussian mixture model allows signal properties (that is, means and variances) denoting modified and unmodified RNA species to be shared across samples, yet accommodates sample-specific mixing weights, one of which is later used as an estimate of modification rate per sample.

Let y_{mn} be an intensity-level mean of a read n from a sample m aligned at a given position i corresponding to a five-mer k. We assume that the intensity level of a modified read is independently drawn from a normal distribution of the modified k with a mean μ_{mod} and a variance τ_{mod}^{-1}, otherwise from another normal distribution of the unmodified k with a mean μ_{unmod} and a variance τ_{unmod}^{-1}. We denote z_{mn} to be 1 if the read is modified and 0 otherwise. Therefore, the conditional likelihood of all reads $y = \{y_1, \ldots, y_{mn}\}$ at the position i given $z = \{z_1, \ldots, z_{mn}\}$, $\mu = \{\mu_{\text{unmod}}, \mu_{\text{mod}}\}$, $\tau = \{\tau_{\text{unmod}}, \tau_{\text{mod}}\}$ can be written in the form of a mixture of the two Gaussian models as follows:

$$P(y \mid z, \mu, \tau) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \mathcal{N}\left(y_{mn} \mid \mu_{\text{mod}}, \tau_{\text{mod}}^{-1}\right)^{z_{mn}} \times \mathcal{N}\left(y_{mn} \mid \mu_{\text{unmod}}, \tau_{\text{unmod}}^{-1}\right)^{1-z_{mn}}.$$

where M and N_m are the total number of samples and the total number of reads in sample m, respectively. As a theoretical signal distribution of the unmodified k is available, we allow the model to favor the unmodified k unless the data reveal otherwise by imposing a normal-gamma distribution as a prior with the hyper-parameters m_0^k, λ_0, α_0, and β_0^k, where $\mathcal{N}(\cdot)$ and $\mathcal{G}(\cdot)$ denote a normal distribution and a gamma distribution, respectively:

$$P(\mu, \tau) = \prod_{i \in \text{ mod,unmod}} \mathcal{N}\left(\mu_i \left| m_0^k, (\lambda_0 \tau_i)^{-1}\right.\right) \times \mathcal{G}\left(\tau_i \left| \alpha_0, \beta_0^k\right.\right).$$

where m_0^k is the theoretical mean of the 5-mer k, $\lambda_0 = 1$, $\alpha_0 = 0.5$, and β_0^k is 0.5 × the theoretical variance of the 5-mer k. With this prior, Gaussian parameters are regularized, which can help inference for those positions that have a low number of reads.

We assume z_{mn} to follow a Bernoulli distribution with a probability w_m, to which we refer as a modification rate of sample m:

$$P(z \mid w) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} \text{Bernoulli}(z_{mn} \mid w_m) = \prod_{m=1}^{M} \prod_{n=1}^{N_m} w_m^{z_{mn}} \times (1 - w_m)^{1-z_{mn}}.$$

We also put a symmetric, uninformative beta distribution as a prior with the hyper-parameters a_0 and b_0 on each sample modification rate as follows:

$P(w) = \prod_{m=1}^{M} \text{beta}(w_m \mid a_0, b_0) = \prod_{m=1}^{M} \frac{w_m^{a_0-1}(1-w_m)^{b_0-1}}{\text{beta}(a_0, b_0)}$, where $a_0 = b_0 = 0.001$s.

To make inference, we employ a variational Bayesian approach to update all model variables iteratively.

Post-modeling filter. At positions where only a substantial fraction of reads are modified, two Gaussian models are easily distinguishable. On the other hand, at sites where either complete modification or no modification is found, one of the distribution means is to converge into its prior with a large variance, forming an uninformative distribution. Reads from all conditions tested are modeled in the other distribution, indicating no differences among conditions. To discriminate these positions, we first tested the separation of the two distributions by calculating the probability of the overlapping area of the two clusters. We considered those positions with no more than 50% overlapping area as distinguishable. We also removed those positions where one is inside the other, that is, when more than one intersection point has a density value higher than 0.1. We allow these thresholds to be adjustable by users.

When samples are diverse in genetic background, such as samples from different tissues, different nucleotides due to variants can also cause a shift in intensity levels. To avoid this confounding effect when considering base modification, we removed variants using VarScan[45].

Statistical test for the DMR. Among the remaining sites, we assign the closer mean to the prior to be unmodified RNA and the other mean to be modified RNA. For each condition, the model yields a modification rate for each sample. To prioritize positions with differential modification in a pairwise comparison, we performed a two-tailed, unpooled z-test on the modification-rate difference of any two conditions for each position:

$$z = \frac{\hat{w}_1 - \hat{w}_2}{\sqrt{\frac{\hat{w}_1(1-\hat{w}_1)}{n_1} + \frac{\hat{w}_2(1-\hat{w}_2)}{n_2}}}.$$

where $\hat{w}_1$, $\hat{w}_2$ are the estimated modification rates from the two conditions of interest and n_1 and n_2 are the corresponding read coverages.

In cases in which replicates are available, the average of read coverages and modification rates across replicates for each condition are used to compute n_1, n_2 and $\hat{w}_1$, $\hat{w}_2$, respectively. As a result, z scores were used to rank differentially modified positions. The corresponding P values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure to control the FDR at a level of 0.05.

Preprocessing. Although both signal and sequence are generated from the sequencing machine and its proprietary basecaller software, segmenting continuous signal samples into events and assigning each to a *k*-mer is an essential prerequisite for raw signal analysis. We assumed that a single event comprises a set of samples drawn at the time when a *k*-mer of a strand resides in the pore and the consecutive event when the strand moves past the pore at another base, where *k* is five for an RNA strand.

In achieving this, we applied 'Nanopolish Eventalign' to associate signal partials with their corresponding reference nucleotides[28,29]. As a result, each read is segmented, and its properties including reference *k*-mer, model *k*-mer events and their corresponding signal segments along with observed and expected normalized means were reported. To combine those events aligned to the same position into a single event, we averaged the multiple event means weighted by their event length. Moreover, we ignored the skipped positions and discarded mismatched *k*-mers. For sufficient coverage, moreover, only positions with 30–1,000 aligned reads were considered. In total, we modeled 9,509,290 genomic positions, involving 5,621 genes.

Data generation. *Tissue culture.* ATCC HEK293T CRL-3216 cells were cultivated in a sterile 5% CO_2 incubator at 37 °C in DMEM supplemented with 10% FBS and 1% penicillin–streptomycin. Cells within passages 3–20 were used for experiments. HEK293T cells were regularly subjected to testing with the MycoAlert Plus Mycoplasma kit (Lonza, LT07) to verify that they were free of *Mycoplasma*.

RNA mixtures. RNA was isolated from adherent WT or *METTL3*-KO HEK293T cells using TRIzol LS (Ambion, 10296) according to the manufacturer's instructions and quantified using the Qubit RNA HS assay (Thermo Fisher, Q32855). Total RNA was first precipitated with ethanol again to remove residual salt. PolyA RNA was then purified using the Poly(A)Purist MAG kit (Thermo Fisher, AM1922) according to the manufacturer's instructions. WT or *METTL3*-KO polyA-selected RNA (Supplementary Fig. 8) was mixed at designated ratios before undergoing direct RNA-seq library preparation.

Small interfering RNA knockdown. HEK293T cells were first seeded in a six-well plate at a seeding density of 6×10^5 cells per well. After 24 h, HEK293T cells were then transfected with a final concentration of 22 nM (50 pmol per well) of respective siRNA species using the RNAiMAX transfection reagent (Invitrogen, 13778) according to the manufacturer's instructions. siRNA s32143 (Thermo Scientific) was used to knock down *METTL3*, while siRNA 4390843 was used as a scrambled siRNA control. Twenty-four hours after transfection, HEK293T cells were diluted to inoculate new plates at a seeding ratio of 1:9 to allow the cells to divide for an additional 48 h before being collected for RNA or protein (total 72-h knockdown) (Supplementary Fig. 9).

Total protein isolation. Trypsinized HEK293T cells were washed twice with ice-cold PBS. Washed cells were lysed in RIPA buffer (150 mM NaCl, 1% NP-40, 0.5% sodium deoxycholate, 0.1% SDS, 50 mM Tris, pH 8, 1× cOmplete Mini EDTA-free protease inhibitor) by tumbling for 30 min at 4 °C. Lysates were clarified by centrifuging at 16,000*g* for 30 min at 4 °C, and protein concentrations were quantified with the Pierce BCA protein assay kit (Thermo Scientific, 23225).

Western blotting. Whole-cell lysates were diluted in 1× Laemmli buffer (Bio-Rad, 1610747) supplemented with 2-mercaptoethanol before being denatured for 10 min at 95 °C. Approximately 30 µg lysate was separated on a 10% SDS–PAGE gel in separation buffer (25 mM Tris, pH 8.3, 192 mM glycine, 0.1% SDS) and transferred onto a nitrocellulose membrane via wet transfer in cold transfer buffer (25 mM Tris, pH 8.3, 192 mM glycine, 20% methanol). The membrane was rinsed with water and then blocked with Odyssey blocking buffer (LI-COR, 927) for 1 h at room temperature. The membrane was stained for 16 h at 4 °C overnight with antibody-dilution solution (0.1% Tween-20 in Odyssey blocking buffer) containing primary antibodies. The membrane was rinsed and then washed thrice with PBS-T (1× PBS, 0.1% Tween-20) before being stained with secondary antibodies. The membrane was rinsed and then washed thrice with PBS-T and once with PBS before being imaged on a LI-COR Odyssey CLx imaging system. The following

antibodies and dilution factors were used: 200× diluted mouse anti-HSP60 (Abcam, ab110312), 1,000× diluted rabbit anti-METTL3 (Bethyl Laboratories, A301-567A-T), 10,000× diluted IRDye 680RD goat anti-mouse IgG H + L (LI-COR, 68070) and 10,000× diluted IRDye 800CW goat anti-rabbit IgG H + L (LI-COR, 32211).

Multiple myeloma patient samples. Total RNA was extracted from CD138$^+$ cells with QIAzol Lysis Reagent (Qiagen). Briefly, cells were homogenized with QIAzol by passing them through a 20-gauge needle (0.9-mm diameter). Chloroform was then added, and samples were shaken vigorously by hand. Phase separation was achieved by centrifuging samples at 12,000*g* for 15 min at 4 °C. The upper aqueous phase containing RNA was transferred into a new tube, and isopropanol was added to precipitate RNA. The RNA pellet was collected by centrifugation and washed with 75% ethanol. The pellet was then left to air dry for 15 min and dissolved in RNase-free water. RNA concentration and purity were measured by spectrophotometry (NanoDrop, Thermo Fisher Scientific) and automated electrophoresis (2100 Bioanalyzer System, Agilent). Samples from patients with multiple myeloma were obtained at diagnosis during bone marrow aspiration after informed consent. The use of these samples for research was approved by the Domain Specific Review Board in Singapore.

Profiling m^{6}A validation sets. m^{6}A-cross-linking-exonuclease sequencing. We created a validation case–control set of transcriptome-wide m^{6}A modification at single-base resolution by applying m^{6}ACE-seq to quantify relative differences in methylation levels between WT and *METTL3*-KO HEK293T RNA samples[15]. We identified METTL3-dependent m^{6}A sites using previously determined criteria[15]. Briefly, these sites exhibited a *METTL3* WT/KO relative methylation level ratio ≥4.0 (*P* value of one-tailed *t*-test, <0.05). As a result, a comprehensive profile of 15,703 genomic positions with significantly differential m^{6}A modification was generated, covering 4,508 unique genes.

Mapping m^{6}A at individual-nucleotide resolution using cross-linking and immunoprecipitation. Modified positions from miCLIP were obtained from both CIMS and CITS miCLIP libraries from the supplementary information of ref. [46]. We combined the two files and considered a position to be modified if it appeared in one of the two libraries.

RNA digestion via m^{6}A-sensitive RNase. We downloaded Table S6 ('MAZTER-Seq Quantification in Humans') from the supplementary information of ref. [38]. MAZTER-seq reports cleavage efficiency values (ranging from 0 to 1) at m6A sites, which is inversely correlated with the methylation level at each site. We used sites with cleavage efficiencies <1 for our analyses.

Basecalling and alignment. Following standard steps, the ionic current readout for each FAST5 file was basecalled using Guppy and stored in FASTQ files (nf-core/nanoseq: https://doi.org/10.5281/zenodo.3697960). To align reads to the transcriptome, we ran minimap2.1 (minimap2 '-ax map-ont -uf–secondary=no') using the GRCh38 Ensembl annotations release version 91 and a modified FASTA file by combining coding and noncoding RNA reference annotations and retaining only transcript IDs that matched reference annotations.

Data analysis. *Prioritizing differentially modified sites between two conditions.* After applying xPore on all samples of interest, we obtained the estimated modification rate for each sample, a test statistic (*z* score) and *P* value on DMRs for each pairwise condition. Because only one modification type is considered for each *k*-mer transcriptome wide, we kept only those positions where modified distributions were assigned in the same direction (lower or higher than the unmodified counterpart) as the majority per *k*-mer. Finally, we ranked differentially modified sites based on the *z* score between any two conditions of interest.

k-mer frequency among significantly differentially modified sites. To investigate the top frequent motifs that were differentially modified between any pairwise conditions, we first selected those differentially modified sites with effect size >0.5 and *P* value <0.001. Next, occurrences of each *k*-mer were counted.

Method comparison. We compared our methods with Tombo version 1.5.1 (https://nanoporetech.github.io/tombo/index.html) and EpiNano version 1.1 (https://github.com/enovoa/EpiNano.git). We ran 'tombo detect_modifications level_sample_compare' on four pairwise comparisons that were obtained from two HEK293T WT samples and two HEK293T KO samples. For the non-comparative mode, we performed 'tombo detect_modifications de_novo' and EpiNano between HEK293T WT replicate 1 and HEK293T KO replicate 1 samples. To run EpiNano version 1.1 on the HEK293T dataset, we excluded feature generation for positions without AC center nucleotides and assumed the probability of these positions being modified to be zero. We closely followed the instructions in the repository for preprocessing and ran all four SVM models provided for the comparison. We could not obtain the results from Tombo[23] using multiple replicates, and Nanocompore[24] did not return results for our dataset on single replicates or multiple replicates (status, 4.12.2020).

Evaluation metrics. To evaluate methods for m^{6}A detection, we used an ROC curve in which the true positive rate is plotted against the false positive rate at each rank cutoff. Moreover, we used precision–recall curves. Finally, we also summarized method performance by computing areas under both ROC and precision–recall (PR) curves, resulting in AUCROC and AUCPR, respectively.

Metagene analysis. We first mapped gene coordinates to transcript coordinates based on the most abundant transcript aligned to each gene. We then discretized transcripts into functional areas, that is, 5′ UTR, coding sequence and 3′ UTR, and calculated the relative positions of differentially modified sites identified by xPore. Finally, we demonstrated the occurrences of differentially modified sites distributed within each functional area in Fig. 2e.

Comparison of estimated modification rates with m^{6}ACE-seq and MAZTER-seq. We compared modification rates estimated by xPore with those from m^{6}ACE-seq, which was previously performed on the same mixture ratios (0%, 25%, 50%, 75%, 100%)[15]. Using these m^{6}ACE-seq data, we identified all the sites that are detected both by xPore and m^{6}ACE-seq and then calculated the relative abundance estimates across the 25%, 50% and 75% samples (the 0% *METTL3*-KO samples were normalized to 0; the 100% WT samples were normalized to 1). We then compared modification rates estimated by xPore and m^{6}ACE-seq across 25%, 50% and 75% mixture samples (Supplementary Fig. 3d).

We have compared xPore with MAZTER-seq[38] using raw cleavage efficiency, as no methylation-deficient background was generated for human cells. Among the top 1,452 positions ($P < 0.001$) identified by xPore, 319 have the NNACA motif, which is detected by MAZTER-seq. Among these positions, 90 were identified as significantly modified by both MAZTER-seq and xPore. Using these positions, we generated a scatterplot to show the correlations (Supplementary Fig. 3e).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data generated in this study are publicly available through the ENA (PRJEB40872). Here we used the following samples: HEK293T *METTL3*-KO cells (three replicates); HEK293T WT cells (three replicates); HEK293T *METTL3*-KD cells (three replicates); HEK293T KD control WT cells (three replicates); HEK293T WT–KO mixture, 100% modified (three replicates); HEK293T WT–KO mixture, 75% modified (four replicates); HEK293T WT–KO mixture, 50% modified (four replicates); HEK293T WT–KO mixture, 20% modified (four replicates); HEK293T WT–KO mixture, 0% modified (three replicates); multiple myeloma patient samples (three patient samples).
In addition, we used direct RNA-seq data from the SG-NEx project[33], which are available at https://github.com/GoekeLab/sg-nex-data and https://www.ebi.ac.uk/ena/browser/view/PRJEB44348.
Preprocessed files for all samples are available at https://doi.org/10.5281/zenodo.4604945 for SG-NEx data and https://doi.org/10.5281/zenodo.4587661 for the other samples, which can be directly used to identify differential RNA modifications with xPore. The list of all samples can be found in Supplementary Data 7.
Samples from the three patients with myeloma were obtained after informed consent. The use of these samples for research was approved by the Domain Specific Review Board in Singapore.

Code availability

Our implementation in Python is available at https://github.com/GoekeLab/xpore. xPore's documentation is available at https://xpore.readthedocs.io.

References

45. Koboldt, D. C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
46. Grozhik, A. V., Linder, B., Olarerin-George, A. O. & Jaffrey, S. R. Mapping m^{6}A at individual-nucleotide resolution using crosslinking and immunoprecipitation (miCLIP). *Methods Mol. Biol.* **1562**, 55–78 (2017).

Acknowledgements

This work is funded by the Agency for Science, Technology and Research (A*STAR), Singapore and by the Singapore Ministry of Health's National Medical Research Council under its Individual Research Grant funding scheme. P.N.P. acknowledges the Thailand Research Fund under grant number RTA6080013. We thank the Lezhava laboratory for assistance with sequencing. We thank M. Shee Siok Woon for help with administrative support.

Author contributions

P.N.P. designed and implemented the computational method. J.G. and W.S.S.G. conceived the project. P.N.P., J.G. and W.S.S.G. designed the study and experiments and analyzed data. A.T. contributed to design of the computational method. F.Y., Y.C., C.W.Q.K., Y.K.W., C.H., P.P., Y.T.G., P.M.L.Y., J.Y.C., W.J.C. and S.B.N. contributed to data generation, data processing and data interpretation. Y.K.W. contributed to

implementation of the computational method. P.N.P., W.S.S.G. and J.G. organized and wrote the paper with contributions from all authors.

Competing interests

W.S.S.G. has filed a technology disclosure to the institutional technology transfer office, and the office has filed a provisional patent application in Singapore on the use of photo-crosslinking RNA-modification-specific antibodies and exoribonucleases to sequence RNA modifications at high resolution. All other authors have no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-00949-w.

Correspondence and requests for materials should be addressed to P.N.P., W.S.S.G. or J.G.

Peer review information *Nature Biotechnology* thanks Angus Wilson and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Ploy N. Pratanwanich, W.S. Sho Goh, Jonathan Goeke

Last updated by author(s): 26/03/21

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	*Provide a description of all commercial, open source and custom code used to collect the data in this study, specifying the version used OR state that no software was used.*
Data analysis	Our implementation in Python is available at https://github.com/GoekeLab/xpore as open-source software.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data used in this manuscript will be available on ENA (https://www.ebi.ac.uk/ena/browser/view/PRJEB40872). Part of the data used in this study was generated by the SG-NEx project (https://github.com/GoekeLab/sg-nex-data). The preprocessed files for all samples are available at https://doi.org/10.5281/zenodo.4604945 for the SGNEx data and https://doi.org/10.5281/zenodo.4587661 for the other samples. The list of all samples can be found in Supplementary Table S7.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We evaluated our method using 3 replicates in each condition. We also compared the results when only 1 replicate is available, which still provides good results.
Data exclusions	All the filtering steps are described in the manuscript.
Replication	We evaluated our predictions against an independent method (m6ACE-Seq).
Randomization	It is not relevant because an unbiased search for differentially modified positions was performed transcriptome-wide.
Blinding	The developers were blinded to the validation set while running the model to search for differential modifications transcriptome-wide.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	anti-m6A (Synaptic Systems 202003)
Validation	It was validated by the manufacturer.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	ATCC HEK293T CRL-3216
Authentication	ATCC HEK293T CRL-3216 was authenticated via ATCC STR profiling.
Mycoplasma contamination	Cells were regularly verified to not to exhibit mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Human research participants

Policy information about studies involving human research participants

Population characteristics	Not applicable (no claims based on the population characteristics are made as this is a proof of principle study on a technology).
Recruitment	Samples from multiple myeloma patients were obtained at diagnosis during bone marrow aspiration after informed consent.
Ethics oversight	The use of these samples for research was approved by the Domain Specific Review Board (DSRB) in Singapore.

Note that full information on the approval of the study protocol must also be provided in the manuscript.