An informatics approach to distinguish RNA modifications in nanopore direct RNA sequencing

Ramasamy, Soundhar; Mishra, Shubham; Sharma, Surbhi; Parimalam, Sangamithirai Subramanian; Vaijayanthi, Thangavel; Fujita, Yoto; Kovi, Basavaraj; Sugiyama, Hiroshi; Pandian, Ganesh N

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

ELSEVIER

An informatics approach to distinguish RNA modifications in nanopore direct RNA sequencing

Soundhar Ramasamy [a,1], Shubham Mishra [b,1], Surbhi Sharma [a], Sangamithirai Subramanian Parimalam [a], Thangavel Vaijayanthi [a], Yoto Fujita [a], Basavaraj Kovi [c], Hiroshi Sugiyama [a,b,*], Ganesh N. Pandian [a,*]

[a] *Institute for Integrated Cell-Material Science (WPI-iCeMS), Kyoto University, Sakyo, Kyoto 606-8501, Japan*
[b] *Department of Chemistry, Graduate School of Science, Kyoto University, Sakyo, Kyoto 606-8502, Japan*
[c] *Laboratory of Crop Evolution, Graduate School of Agriculture, Kyoto University, Muko, Kyoto 617-0001, Japan*

ARTICLE INFO

Keywords:
RNA modifications
Nanopore sequencing
Epitranscriptomics
Pseudouridine
Chemical probe
Trace value

ABSTRACT

Modifications in RNA can influence their structure, function, and stability and play essential roles in gene expression and regulation. Methods to detect RNA modifications rely on biophysical techniques such as chromatography or mass spectrometry, which are low throughput, or on high throughput short-read sequencing techniques based on selectively reactive chemical probes. Recent studies have utilized nanopore-based fourth-generation sequencing methods to detect modifications by directly sequencing RNA in its native state. However, these approaches are based on modification-associated mismatch errors that are liable to be confounded by SNPs. Also, there is a need to generate matched knockout controls for reference, which is laborious. In this work, we introduce an internal comparison strategy termed “IndoC,” where features such as ‘trace’ and ‘current signal intensity’ of potentially modified sites are compared to similar sequence contexts on the same RNA molecule within the sample, alleviating the need for matched knockout controls. We first show that in an IVT model, ‘trace’ is able to distinguish between artificially generated SNPs and true pseudouridine (Ψ) modifications, both of which display highly similar mismatch profiles. We then apply IndoC on yeast and human ribosomal RNA to demonstrate that previously reported Ψ sites show marked changes in their trace and signal intensity profiles compared with their unmodified counterparts in the same dataset. Finally, we perform direct RNA sequencing of RNA containing Ψ intact with a chemical probe adduct (N-cyclohexyl-N′-β-(4-methylmorpholinium) ethylcarbodiimide [CMC]) and show that CMC reactivity also induces changes in trace and signal intensity distributions in a Ψ specific manner, allowing their separation from high mismatch sites that display SNP-like behavior.

1. Introduction

RNA in biological systems is frequently modified with chemical moieties that can cause changes in its structural or chemical properties [1,2]. Apart from being the object of focus as the intermediate carrier of genetic information between DNA and protein and as an effector molecule (tRNA and rRNA), RNA is now known to form various kinds of non-coding transcripts, which can also directly act to regulate gene expression within the cell [3,4]. RNA modifications can further add to the complexity of this regulatory network by contributing to a layer of information that exists on top of the one portrayed by the nucleotide sequence of the RNA itself, which is often studied within the realm of epigenetic. This study of the modifications themselves collectively falling under the umbrella term of “epitranscriptomics.” Of these, one of the most abundant RNA modifications is pseudouridine (commonly represented by the Greek letter Ψ), which is one of the earliest to be discovered [5]. Ψ is highly conserved across species, and artificially replacing Ψ in place of U sites has been shown to have effects such as increased protein expression [6] and recoding of nonsense to sense codons [7].

Consequently, the accurate information on the position and stoichiometry of such modifications is of increasing interest. Despite this, comprehensive profiling of modifications has remained a long-standing

* Corresponding authors at: Institute for Integrated Cell-Material Science (WPI-iCeMS), Kyoto University, Sakyo, Kyoto 606-8501, Japan.
E-mail addresses: hs@kuchem.kyoto-u.ac.jp (H. Sugiyama), namasivayam.ganeshpandian.5z@kyoto-u.ac.jp (G.N. Pandian).

[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.ygeno.2022.110372
Received 29 November 2021; Received in revised form 18 March 2022; Accepted 18 April 2022
Available online 20 April 2022

challenge as biophysical methods such as chromatography and mass spectrometry techniques which have been the gold standards for detecting modifications, can only be used on a per-case basis. Recently, the field has seen a resurgence with the advent of high-throughput sequencing technologies, which have the ability to make predictions on thousands of sites. However, these techniques require laborious sample preparation procedures and antibodies or modification-specific chemical adducts often lead to truncated products post reverse transcription [8]. Furthermore, they are error-prone to various degrees and do not easily allow simultaneous mapping of multiple modifications.

The direct RNA sequencing (dRNA-Seq) platform being developed by Oxford Nanotechnologies (ONT) is gaining increasing attention recently. ONT allows for long-read sequencing by directly passing the native RNA molecule through a modified transmembrane pore (with a constant voltage maintained across this membrane) and measuring the characteristic changes in the cross-membrane current, which happen in a sequence-dependent manner. In principle, this technology could enable the mapping of virtually any RNA modification simultaneously. However, technical and technological limitations have proven to be stumbling blocks towards this goal. Nevertheless, alternative approaches have been explored to probe RNA modifications using both ionic current and base-calling and alignment-based methods [9].Recently, *Begik* et al [10]. showed that the per-read features of trace value and signal intensity could differentiate pseudouridine-modified sites from unmodified identical uridine-containing sites.

In this work, we tested whether trace value and signal intensity can differentiate pseudouridine from SNP, with both appearing as mismatch errors in dRNA-Seq reads. In addition, we utilize the specific reactivity of (N-cyclohexyl-N′-β-(4-methylmorpholinium) ethylcarbodiimide [CMC]) towards pseudouridine in direct RNA sequencing of the CMC modified pseudouridine-containing RNA, which to our knowledge, is the first example of such an approach.

2. Results

2.1. Ψ Modified positions show significant change in trace value distributions compared with an SNP model in IVT-RNA

It has been shown that pseudouridine-modified sites, when sequenced on the ONT platform, are quite often basecalled as U/C mismatches [10,11]. This property, along with certain other base-calling features, is a relatively reliable indicator of this modification. With this approach, however, there is a potential problem of conflating true single nucleotide polymorphisms (or SNPs, sequence variations at specific single nucleotide positions present in relatively high frequencies, typically >1%, within a population) or single nucleotide variations (SNVs), with modifications.

As of the latest version, the basecalling program Guppy, which converts raw electric signal data from a sequencing experiment to the corresponding nucleic acid sequence, in addition to recording the electric current ‘signal intensity,’ also generates ‘meta-information’ on the raw signal such as “Trace” (also termed as ‘base probability’). Together with the signal intensity, ‘trace’ has been used for the measurement of the stoichiometry of pseudouridine modifications in ribosomal RNA (rRNA) of yeast by comparing the trace value distributions of modified pseudouridine sites against the same sequence context obtained from pseudouridine writer enzyme knockouts, which instead contained unmodified uridine at the same position [10].

Given the interpretation of ‘trace’ as a measure of the probability of a given base getting called correctly, it is reasonable to expect obtaining lower trace values (lesser confidence, towards zero) in the presence of a chemical modification at a given position and higher trace values in the presence of any of the canonical bases (not just uridine). This, along with other parameters such as signal intensity, in turn, could hold potential as an indicator of pseudouridine when it came to distinguishing between a true modification and an SNP at that position.

To see if a difference could indeed be observed between true pseudouridine modifications (which show up as U/C mismatches in nanopore sequenced data) and true single nucleotide variations, we generated a synthetic in vitro transcribed SNP model (by mixing U and C containing RNAs), and pseudouridine containing RNA (detailed description of the scheme and design in Methods) outlined in Fig. 1 and Fig. S1. It is important to note that while our U/C SNP model was created by mixing U and C containing RNAs in equal concentrations, the observed mismatch rate varied from the expected 50% of U/C mismatch error, which could be attributed to sequencing bias for one of the above RNAs. Predictably, the pattern of U/C mismatch errors was quite similar between the SNP model and pseudouridine modified RNA (representative IGV snapshot in Fig. 1a., a full snapshot in Fig. S2a), which indicates that mismatch profiles alone lack differentiation power between the two. We used the tool nanoRMS [10]to extract the trace value for every passed read at these U/C mismatch positions for both the sequence datasets. We then used the Kolmogorov–Smirnov statistic (KS statistic) [12] to quantify the difference between the trace value distributions for the SNP model and the Ψ containing RNA, the negative log-transformed *p*-values for which (two-sample KS test) ranked in decreasing order can be seen in Fig. 1b (the corresponding D statistic values in Fig. S2b). From the probability density plots for representative high, intermediate, and low KS statistic value positions (Fig. 1c), we observe that there is a significant density at or close to zero in the pseudouridine-modified RNA. This corresponds to those reads containing pseudouridine at these positions, whereas the density close to 1 corresponds to those with uridine. On the other hand, almost all the density is present close to 1 in the case of the SNP model since no reads, in this case, have pseudouridine at these positions (only either U or C). The distribution of the KS statistic values for all the relevant positions can be seen in Fig. 1d.

One caveat with using the KS statistic to measure the difference between two distributions is that it does not give information about the direction of the change since it only measures the most significant vertical distance between their CDFs (Fig. S3 for a graphical explanation). In the present case, since the direction of the change is relevant (i.e., the increase in density at 0 in the modified RNA compared with the unmodified RNA), we decided to also calculate the quantity ‘skewness,’ which is the third standardized moment of a probability distribution [13] and as the name suggests, is a measure of the asymmetry in a distribution function. A distribution with its long tail towards the right side will have a positive skewness, whereas one with its long tail towards the left will have a negative skewness (Fig. S4a. For a graphical explanation). Fig. 1e. shows violin plots for the skewness in the trace value distributions for all the relevant positions in the SNP model and the pseudouridine-containing RNA, where the latter has its values tightly packed around the zero mark. In contrast, the SNP model mostly encompasses a wide range of negative values expected from its right-heavy trace distributions. The distribution of the differences in skewness at every position is given in Fig. S4b.

2.2. Internal comparison strategy, IndoC, to compare Bonafide Ψ modified positions with their respective identical 5-mers

Having established the utility of trace value distributions in distinguishing between pseudouridine modifications and single nucleotide variations in RNA generated by IVT, we explored if similar observations could also be made in dRNA-seq data derived from biological RNA samples. However, unlike IVT-RNA, it is quite challenging to obtain a matched control that comes without the pseudouridine modifications at the same positions that would enable such a comparison. This typically involves generating knockouts of genes coding for enzymes that “write” these modifications onto the target RNA (Pseudouridine synthase, PUS family genes) or small nucleolar RNA (snoRNAs, that guide the modification of other RNAs) [10]. Therefore, we sought to alleviate this problem by attempting an internal comparison strategy.

In the current version of the ONT, it is known that at any given point

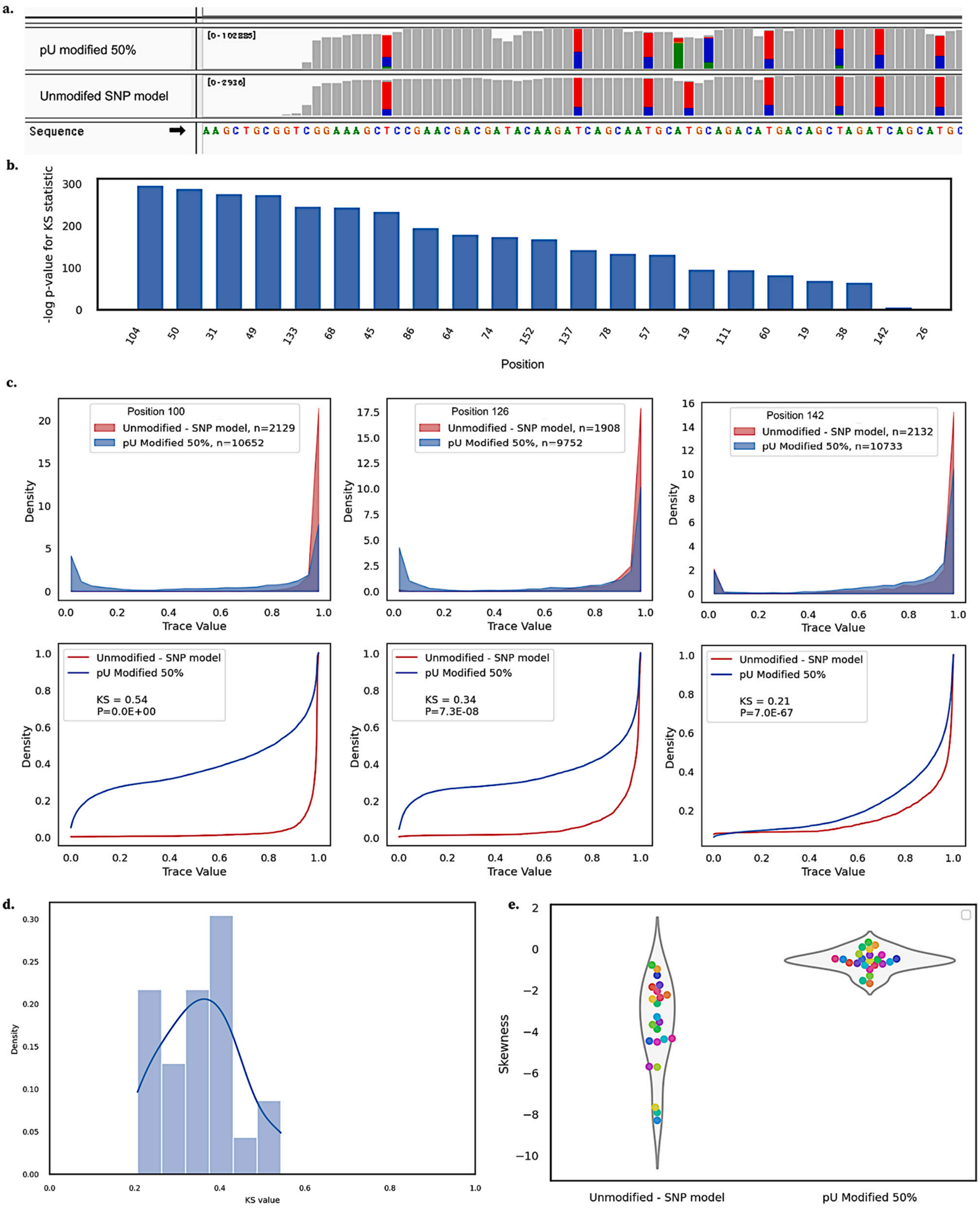

Fig. 1. ‘Trace’ based differentiation of pseudouridine and SNP model in a synthetic template. a. IGV snapshot of 50% pseudouridine modified RNA and the SNP model. b. Negative log-transformed p-values from the two-sample Kolmogorov-Smirnov tests between the pU modified and SNP model RNA’s trace value distributions at specific positions. c. Representative probability density function and cumulative density function plots for the trace values at specific positions; inset: n - number of reads, KS - KS statistic value, P - p-value obtained for the two-sample KS test performed between the two distributions (due to limitations of precision in the float32 datatype, the p-values are displayed as absolute zero in the plot, while the actual values are non-zero) d. Distribution of the KS statistics, e. Violin plot for the skewness of the trace value distributions of the pU modified RNA and SNP model RNA. Data points represent the positions compared, labeled by color.

in time, the space within the pore is occupied by a stretch of DNA or RNA five nucleotides in length. Unsurprisingly, the characteristic change in electric current across the membrane turns out to be a function of this “K-mer” (of length 5). With this in mind, we conjectured that in the absence of matched controls, in principle, it is possible to compare the basecalling parameters (signal intensity, trace, dwell time, etc.) of a given position (for, e.g., a potential pseudouridine modified site) to other positions within the dataset corresponding to the same sequencing run, which satisfy the condition of occupying the center position in the same 5-mer sequence as our site of interest. In the present case, where this is a potential pseudouridine modification, in addition to its identical 5-mers (which would have a U at the central 0-position in the 5-mer), we also compare it with 5-mers with a cytidine in the center as pseudouridine typically manifest as U/C mismatches in nanopore sequencing. We term this strategy ‘IndoC’ (*Indo*-*C*ompare), which is graphically illustrated in Scheme 1 and Fig. 2a. To see if this method of internally comparing per-read features against K-mers within the same dataset can reveal differences between known pseudouridine and unmodified nucleotides, we focused on the rRNA transcriptome of *Saccharomyces cerevisiae*, for which there have been extensive investigations into the transcriptome-wide mapping of RNA modifications, including pseudouridine [14–16].

We generated dRNA-seq data for yeast total RNA (wildtype BY4741), which we then mapped to the reference for the yeast rRNA transcriptome. For every known pseudouridine site, we identified other positions on the yeast rRNA sequence as per the IndoC method, which we call ‘identical 5-mers’ (although they also contain K-mers that differ in sequence from the site of interest, albeit only in the center 0-position, having a C instead of a U), and trimmed the list using a coverage cut-off. Then, as in the case of the IVT-RNA, we used nanoRMS to extract per-read features (signal intensity and trace value) for both the bonafide pseudouridine sites and their associated identical 5-mers.

Representative distribution plots for selected sites can be seen in Fig. 2b. In addition to the KS-statistic values of trace, we calculated the KS-statistic for signal intensity distributions. We also calculated the skewness of the trace value distributions for every bonafide pseudouridine site and its corresponding identical 5-mers (Fig. S5). The violin plot for skewness is shown in Fig. 2c. In accordance with our notion, a distinct upward shift for the skewness values can be seen in the case of the bonafide pseudouridine sites. This is also consistent with results from the synthetic pseudouridine-containing RNA (Fig. 1e), although the skewness distribution is significantly broader.

Since mismatch errors are a hallmark of pseudouridine modifications, we performed the principal component analysis with KS value of trace, KS value of signal intensity, and skewness as features to comprehend if differences in these features could provide distinguishing capacity against high mismatch sites. We then performed linear 2D-Support Vector Machine (SVM) based binary classification with bonafide Ψ sites in yeast rRNA as the positive class and high mismatch sites as the negative class. We obtained a ROC-AUC of 0.83, suggesting a reasonable degree of differentiation. We also performed the same analysis on human rRNA data (K562 leukemia cell line) to see if this pattern is preserved across different ribosomal transcriptomes, utilizing known pseudouridine positions (Fig. S6) [17]. We obtained a ROC-AUC of 0.77, which is reasonable even when it is lower than that of yeast.

2.3. dRNA-seq of CMC treatment modified Ψ containing RNA enables differentiation from high mismatch sites

Chemical probes such as N-cyclohexyl-N′-β-(4-methylmorpholinium) ethylcarbodiimide (CMC) have been previously used to detect Pseudouridine modifications by sequencing via short-read sequencing technologies [14]. These truncated cDNA products arise as a result of the CMC adduct on pseudouridines [14,16]. This has also been performed using nanopore sequencing in nanoCMC-seq [10]. While it has been shown that the presence of biological RNA modifications themselves can be identified via direct RNA sequencing on the ONT platform by detecting changes in base-calling errors and per-read features, we evaluated if direct nanopore sequencing of biological RNA (without generating cDNA from reverse transcription), after reaction with a modification-specific chemical probe could amplify these differences and act as an alternative method to detect RNA modification biological samples. (Fig. S7).

To this end, we treated yeast total RNA with CMC as per previous reports [15]. We then sequenced both the CMC treated and untreated samples and extracted per-read features for each position and four of its neighboring bases (which would correspond to the set of all the 5-mers containing the site of interest). As before, representative plots for the distributions of trace value and signal intensity for selected pseudouridine modified sites are shown in Fig. 3a. The presence of the CMC adduct appears to induce a change in these distributions, as can be seen from the KS statistics (Figs. S8 & S9).

To see if this change was specific to pseudouridine-modified sites or a globally induced effect, we compared trace, and signal intensity for CMC treated and the untreated conditions at bonafide pseudouridine sites (coverage >50). High mismatch error sites (mismatch >5%, coverage >50) are shown in violin plots in Fig. 3b. We also calculated the KS statistics for the CMC treated and untreated conditions at these high mismatch sites; the combined scatter plot is given in Fig. 3c. Visual inspection conveyed that the bonafide pseudouridine sites display higher KS-statistic values for both parameters compared to the high mismatch sites, which are by and large clustered much closer to zero. This observation is in line with the fact that these quantities are not expected to dramatically change when treated with CMC, which is specific to pseudouridine). Binary classification with an SVM, with bonafide Ψ sites in yeast rRNA as the positive class and high mismatch sites as the negative class, gave a ROC-AUC equal to 0.87. Data points shown in light pink are those high mismatch sites in Fig. 2d (IndoC method) that fell on the “modified” side of the decision boundary. We also observe that in Fig. 3c, most of these sites fall on the “high mismatch” side of the decision boundary, indicating some potential in our CMC-based method to weed out those high-mismatch sites that IndoC is unable to classify correctly. Combining both methods could thus help sieve out those positions that are likelier to contain true modifications.

3. Discussion and conclusions

Many informatics methods currently use RNA modification-induced basecalling errors, including mismatches, to identify RNA modifications [18–20]. However, the mismatch is also the hallmark of single nucleotide polymorphisms (SNPs) or single nucleotide variations (SNVs) (Fig. S11). Unsurprisingly, most of these techniques remove SNPs as a pre-processing step from the candidate list since they would otherwise act as confounding elements. Yet, multiple reports have identified RNA modifications present at or indistinguishable from such sites showing population-level variation [21–23]. We showed that trace value and signal intensity, together with the skewness of the trace value distribution, could distinguish true pseudouridine modifications from true SNPs in an IVT model to a significant degree. Although a similar pattern can be seen when we apply the same strategy to the yeast rRNA transcriptome, it is difficult to say if the high mismatch sites used for the comparison truly represent sites with a population-level variation or if the high mismatch is a manifestation of artifacts resulting from the nanopore sequencing technique itself (Fig. S11). Nevertheless, the fact that a relatively good separation can be achieved concerning the already known pseudouridine positions speaks to the merit of our approach. The future course of work will evaluate the ability of IndoC to distinguish modifications other than pseudouridine, which can also show differences in their per-read features, compared with unmodified residues.

Also, IndoC technique can be favored over matched KO controls as IndoC does not have to take cross-run variation (if any), into account. This is because all comparisons are made with data obtained from the

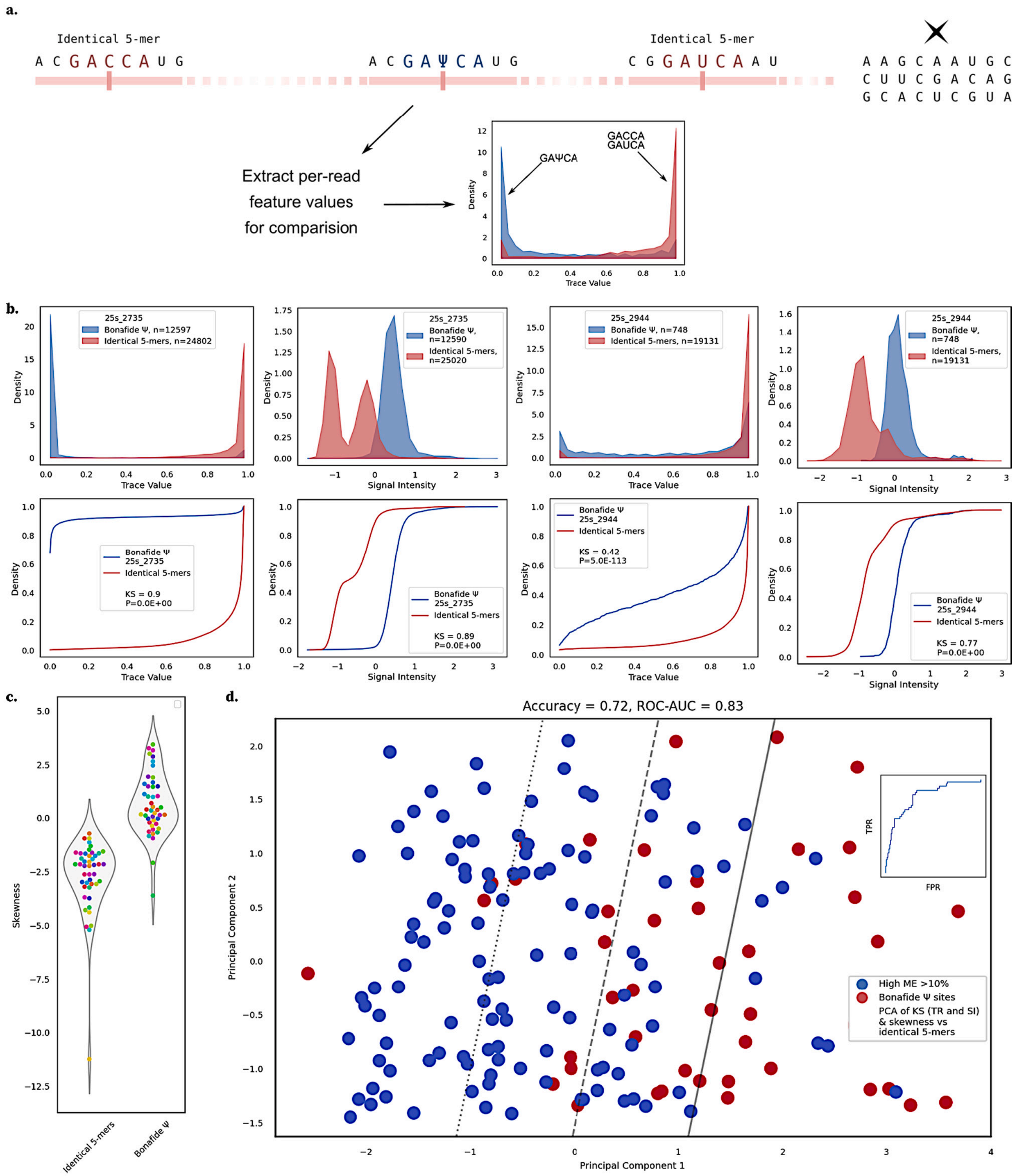

Fig. 2. Applying an internal comparison strategy, IndoC, on yeast rRNA a. IndoC scheme; locations within the entire dataset (including other RNA species) of the same sequencing run are searched for, which have the same 5-mer sequence as the site of interest, termed identical 5-mers here (in case of the pseudouridine sites, the 0-position is allowed to have a 'C') b. Representative probability density function and cumulative density function plots for trace and signal intensity at selected bonafide Ψ positions in yeast rRNA and their identical 5-mers; inset: n - number of reads, KS - KS statistic value, P - p-value obtained for the two-sample KS test performed between the two distributions. Distribution of the number of identical-5mers found for each of the bonafide Ψ positions shown in Supplementary Fig. S10 c. Violin plot for the skewness of the trace value distributions for Bonafide pseudouridine sites and their identical 5-mers d. Principal components of KS_trace, KS_signal-intensity and skewness difference of trace distributions; linear SVM binary classification with bonafide Ψ sites in yeast rRNA as the positive class and high mismatch sites as the negative class (inset: ROC curve, AUC = 0.83).

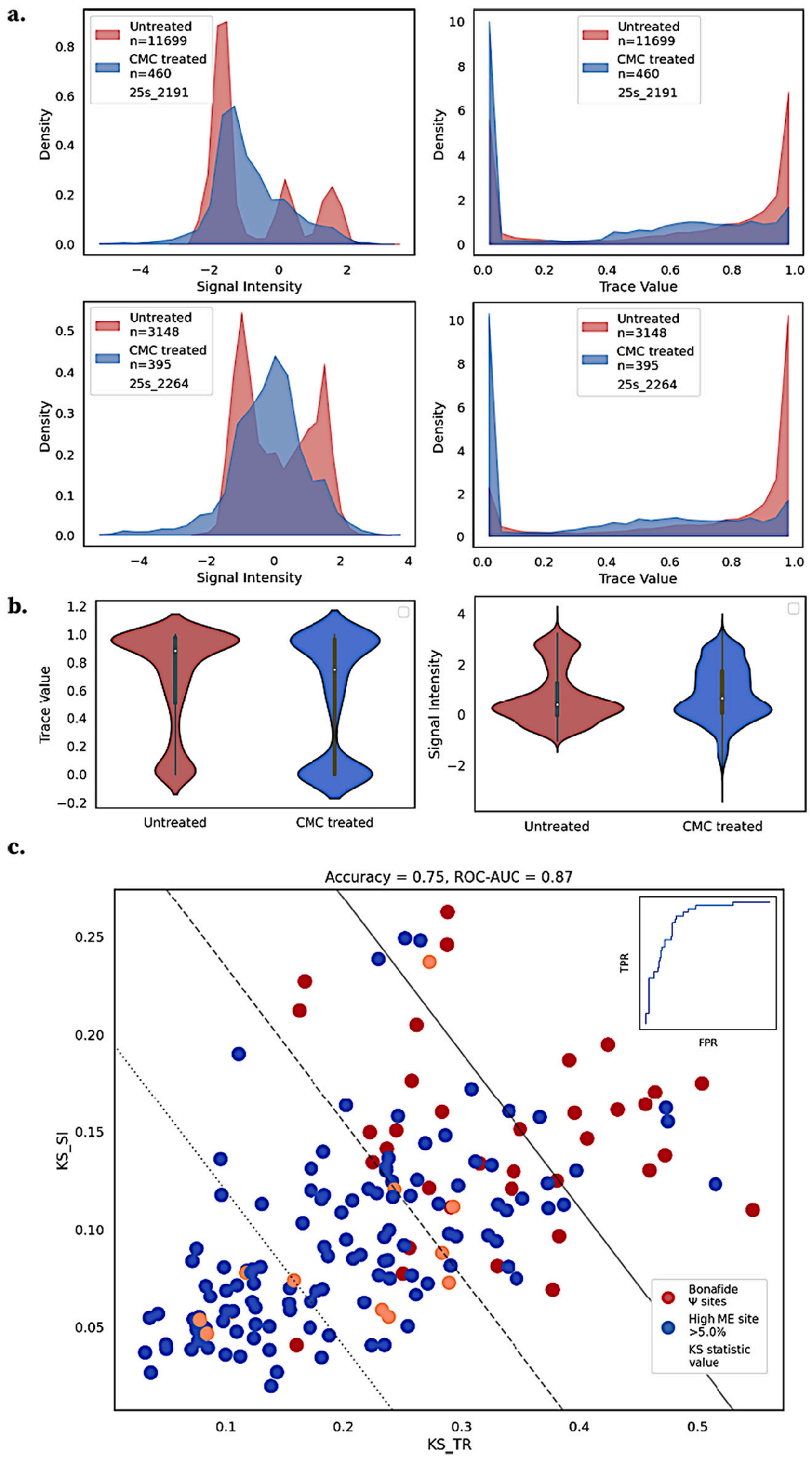

Fig. 3. CMC treatment specifically changes trace and signal intensity on Ψ sites a. Representative probability density function plots for trace and signal intensity at selected bonafide Ψ positions in yeast rRNA showing CMC treated and untreated conditions; inset: n - number of reads. b. Violin plots for trace (left) and signal intensity (right) for all the bonafide Ψ sites in yeast rRNA comparing CMC treated and untreated conditions. c. Scatter plot of KS statistics of Trace, and Signal Intensity distributions between the CMC treated and untreated conditions; linear SVM binary classification with bonafide Ψ sites in yeast rRNA as the positive class and high mismatch sites as the negative class (inset: ROC curve, AUC = 0.84); data points in light pink: high mismatch sites in Fig. 2D that fell on the "modified" side of the decision boundary. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

same sequencing experiment. Comparison with identical 5-mers within the same dataset circumvents the need for separate unmodified control. The critical assumption here is that sequence contexts larger than two neighboring bases in range (corresponding to the 5-mer around a given position) do not significantly affect signal and trace parameters. This assumption is naturally unnecessary in the case of a matched control. However, the fact that this is the same principle that nanopore basecallers utilize when identifying canonical bases (that of signal characteristics being primarily a function of a 5-mer sequence window), provides a precedent for such an approach.

Furthermore, we show that CMC treatment induces significant changes in signal intensity and trace value distributions, specifically in the bonafide pseudouridine sites in yeast rRNA. To the best of our knowledge, this is the first example of direct sequencing of CMC modified pseudouridine-containing RNA on the ONT platform to probe for modifications. One key advantage is that, compared to methods such as Pseudo-seq [14], or PSI-seq [16], which infer the positions of the modifications from the lengths of the cDNA strands that get truncated due to the CMC moiety on pseudouridine during reverse transcription, our approach requires fewer intermediate processing steps, and therefore reduces experimental time and systematic error. Our experiments also have implications for generalizing the use of chemical probes in nanopore sequencing. In principle, the direct sequencing of RNA modified with a chemical probe is possible for any biological RNA modification (though the ease of passage through the nanopore may vary), as long as a specific chemical probe exists for the same. Furthermore, in contrast to the truncation-based methods, the chemical probe-based direct RNA sequencing methods keep other modifications intact.

Over the course of our experiments, we noticed that CMC-treated samples consistently had fewer reads compared to their untreated counterparts. Since every RNA molecule passing through the pore in effect contributes to a single "measurement" of the quantities from which the sequence of its species is inferred, the number of reads obtained corresponds to the "sample size" of this "sample," which has its associated "sample error." And like any random sample drawn from a population, the difference between its sample statistics and its population statistics diminishes as the sample size increases. The reduction in the number of reads (Fig. 3a) is likely due to the large size of the bulky CMC moiety, resulting in difficulties in efficiently entering the pore. Given that nanopore technology comes with its own intrinsic noise, it is advised that care be taken in future experiments to ensure a sufficient number of reads suitable for any downstream statistical analysis. In this work, we have used per-read features from previously known pseudouridine sites in yeast rRNA to distinguish them from high-mismatch sites. However, with accuracies hovering between 70 and 80% for both indoC and CMC-treatment methods, it is not straightforward to determine unknown modified positions with high confidence. Despite this, we envision that both these approaches can aid to weed out confounding high-mismatch sites in nanopore sequencing datasets. Thus, our method assists in focusing a smaller number of sites that can be further validated by the conventional methods like LC/MS, thereby aiding the potential identification of novel RNA modification sites.

4. Methods

4.1. Generation of the SNP model

For generating the SNP model of U and C, two RNA species were synthesized separately and mixed at equal concentrations. The SNP model RNAs had identical sequences except for a matched position having U or C (NNUNN, NNCNN), i.e., they were designed such that the resulting RNAs would have sequences that differed only at certain positions, where one template would possess only uridines at these locations while the other only cytidines. The U containing template was also used to synthesize partially modified pseudouridine containing RNA with IVT performed using an equal concentration of ATP, CTP, GTP, UTP, and ΨTP (75 mM, TriLink). The RNAs were transcribed using the MEGAscript T7 transcription kit (AMB13345). 1 μg of IVT RNA was used for library preparation following the standard ONT dRNA-seq protocol (SQK-RNA002), samples were sequenced independently on a primed flow cell (FLO-MIN106D) and run on a MinION sequencer.

4.2. Yeast/HepG2 rRNA dRNA-seq library preparation and CMC treatment

The dRNA-seq libraries were prepared following the ONT dRNA-seq protocol (SQK-RNA002). Briefly, 20 μg of total RNA derived from yeast (*S. cerevisiae*, strain BY4741)/ HepG2 was polyadenylated followed by a quality check using bioanalyzer and used for library preparation (SQK-RNA002).

CMC treatment was performed as described in previous reports [15,24,25]. where 20 μg of yeast total RNA was treated with 0.34 M CMC in BEU buffer at 37 °C for 1 h followed by alkali lysis using 50 mM sodium bicarbonate, pH -10.4 at 37 °C for 3 h. The CMC labeled RNA was then purified using a quick spin column and polyadenylated. The CMC labeled and polyadenylated RNA was then purified using quick spin columns, and its quality was assessed using a bioanalyzer. The library was then loaded on a primed flow cell (FLO-MIN106D) and run on a MinION sequencer.

4.3. HPLC validation of pseudouridine incorporation and CMC reactivity

For nucleoside analysis, the IVT/CMC treated yeast RNA was digested into nucleosides using 10 μg/ml nuclease P1 (New England Biolabs, M0660S), and 0.5 U/ml Bacterial Alkaline phosphatase (Takara 2120A) at 37 °C for 1 h in 10 μl of a reaction mixture containing 20 mM HEPES–KOH pH 7.5 (Nacalai 15,639–84). The nucleosides were separated using an HPLC equipped with COSMOSIL® 5C18-MS-II packed column 4.6mmI.D.x150 mm. The analysis was performed with a mobile phase with solvent A containing 0.1% TFA (Trifluoracetic acid) in water in gradient combination with solvent B of acetonitrile. The linear gradient started with 0% solvent B at 0 min to 10% B for 30 min, at a flow rate of 1 ml min − 1, monitored at 254 nm.

4.4. DNA library preparation for nanopore sequencing

The 20 μg of total RNA was fragmented using NEB Next magnesium RNA fragmentation module at 94^{o}C for 1.5 min in a preheated thermal cycler. The fragmented RNA was purified using quick spin columns and treated with T4 polynucleotide kinase (M0201). Next, the RNA was polyadenylated followed by quick spin column purification, the quality of the RNA was assessed using a bioanalyzer. For DNA library preparation, the standard protocol mentioned in the SQK-PCS109 kit was followed. Briefly, 70 ng of RNA from the above step was taken into the library preparation protocol. Following second-strand cDNA synthesis, the PCR was performed for 14 cycles. Before adaptor ligation for sequencing, the quality of the amplified DNA was assessed using a bioanalyzer. The library was then loaded on a primed flow cell (FLO-MIN106D) and run on a MinION sequencer.

4.5. Basecalling, mapping, and extraction of nanopore parameters of dRNA-seq reads

Fast5 files were basecalled with Guppy_3.1.5, with –fast5_out, which adds the trace value to the Fast5 files for downstream analysis. The basecalled files were mapped to the reference transcriptome using Minimap2–2.17, the nanopore parameters like mismatch error, trace, and signal intensity were extracted using scripts associated with the nanoRMS package (https://github.com/novoalab/nanoRMS).

The mismatch error per base/k-mer was calculated using python3 epinano_rms.py -R reference.fa -b .sort.bam -s sam2tsv.jar.

The trace value and signal intensity per base/k-mer were calculated using get_features.py –rna -f reference.fa -t number of cores -i fast5. directory.

Data availability

All FASTQ and FAST5 files generated in this work will be publicly available at European Nucleotide Archive (ENA ENA Accession ID: PRJEB51316). We have also provided the code for the new internal comparison strategy, IndoC, on GitHub. (https://github.com/geno-verse/indoC).

Funding sources

This work was supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI [JP19H03349], Takeda Science Foundation, and Uehara Memorial Foundation to G.N.P. We thank JSPS [Grant, No. JP16H06356], NIHR01 (R01CA236350) to H·S and AMED [Grant No.

JP18am0301005] (Basic Science and Platform Technology Program for Innovative Biological Medicine), [JP20am0101101] (Platform Project for Supporting Drug Discovery and Life Science Research (BINDS)). We thank MEXT fellowship for S.M.

Authors statement

S.R. conceived the project. S.R. and S.M. contributed to the design analysis. S.R and S.S. performed HPLC and dRNA-sequencing, S.S·P and T.V. standardized the HPLC protocols, S.S.P. maintained K562 (human leukemia) cell line, Y.F. assisted in the analysis, B.K. helped in the curation of data, S.R. and S.M. wrote the draft, G.N.P. contributed to copy-editing, G.N.P. and H.S. provided critical input and discussion.

Declaration of Competing Interest

The authors declare no conflict of interest.

Acknowledgments

We would like to thank Prof. Qiu-Mei Zhang-Akiyama, for graciously providing yeast total RNA.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2022.110372.

References

[1] S.P. Preethi, P. Sharma, A. Mitra, Structural Landscape of Base Pairs Containing Post-Transcriptional Modifications in RNA, 2021, https://doi.org/10.1101/098871.

[2] E.M. Harcourt, A.M. Kietrys, E.T. Kool, Chemical and structural effects of base modifications in messenger RNA, Nature. 541 (2017) 339–346.

[3] L.F.R. Gebert, I.J. MacRae, Regulation of microRNA function in animals, Nat. Rev. Mol. Cell Biol. 20 (2019) 21–37.

[4] R.-W. Yao, Y. Wang, L.-L. Chen, Cellular functions of long noncoding RNAs, Nat. Cell Biol. 21 (2019) 542–551, https://doi.org/10.1038/s41556-019-0311-8.

[5] W.E. Cohn, E. Volkin, Nucleoside-5′-phosphates from ribonucleic acid, Nature. 167 (1951) 483–484, https://doi.org/10.1038/167483a0.

[6] K. Karikó, H. Muramatsu, J.M. Keller, D. Weissman, Increased erythropoiesis in mice injected with submicrogram quantities of pseudouridine-containing mRNA encoding erythropoietin, Mol. Ther. 20 (2012) 948–953.

[7] J. Karijolich, Y.-T. Yu, Converting nonsense codons into sense codons by targeted pseudouridylation, Nature. 474 (2011) 395–398, https://doi.org/10.1038/nature10165.

[8] I. Anreiter, Q. Mir, J.T. Simpson, S.C. Janga, M. Soller, New twists in detecting mRNA modification dynamics, Trends Biotechnol. 39 (2021) 72–89.

[9] M. Furlan, A. Delgado-Tejedor, L. Mulroney, M. Pelizzola, E.M. Novoa, T. Leonardi, Computational methods for RNA modification detection from nanopore direct RNA sequencing data, RNA Biol. 18 (2021) 31–40.

[10] O. Begik, M.C. Lucas, L.P. Pryszcz, J.M. Ramirez, R. Medina, I. Milenkovic, S. Cruciani, H. Liu, H.G.S. Vieira, A. Sas-Chen, J.S. Mattick, S. Schwartz, E. M. Novoa, Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing, Nat. Biotechnol. 39 (2021) 1278–1291.

[11] S. Ramasamy, V.J. Sahayasheela, Z. Yu, T. Hidaka, L. Cai, H. Sugiyama, G. N. Pandian, Chemical Probe-Based Nanopore Sequencing to Selectively Assess the RNA Modification, 2021, https://doi.org/10.2139/ssrn.3906935.

[12] D. Newman, The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation, Biometrika. 31 (1939) 20, https://doi.org/10.2307/2334973.

[13] The moments of the distribution for normal samples of measures of departure from normality, in: Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 130, 1930, pp. 16–28, https://doi.org/10.1098/rspa.1930.0185.

[14] T.M. Carlile, M.F. Rojas-Duran, B. Zinshteyn, H. Shin, K.M. Bartoli, W.V. Gilbert, Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells, Nature. 515 (2014) 143–146.

[15] S. Schwartz, D.A. Bernstein, M.R. Mumbach, M. Jovanovic, R.H. Herbst, B.X. León-Ricardo, J.M. Engreitz, M. Guttman, R. Satija, E.S. Lander, G. Fink, A. Regev, Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA, Cell. 159 (2014) 148–162.

[16] A.F. Lovejoy, D.P. Riordan, P.O. Brown, Transcriptome-wide mapping of Pseudouridines: Pseudouridine synthases modify specific mRNAs in *S. cerevisiae*, PLoS One 9 (2014), e110799, https://doi.org/10.1371/journal.pone.0110799.

[17] S. Huang, W. Zhang, C.D. Katanski, D. Dersh, Q. Dai, K. Lolans, J. Yewdell, A. M. Eren, T. Pan, Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling, Genome Biol. 22 (2021) 330.

[18] C. Bates, Accurate Detection of m6A RNA Modifications in Native RNA Sequences, 2021, https://doi.org/10.1242/prelights.9716.

[19] M.T. Parker, K. Knop, A.V. Sherwood, N.J. Schurch, K. Mackinnon, P.D. Gould, A.J. W. Hall, G.J. Barton, G.G. Simpson, Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m6A modification, eLife. 9 (2020), https://doi.org/10.7554/elife.49658.

[20] P. Jenjaroenpun, T. Wongsurawat, T.D. Wadley, T.M. Wassenaar, J. Liu, Q. Dai, V. Wanchai, N.S. Akel, A. Jamshidi-Parsian, A.T. Franco, G. Boysen, M.L. Jennings, D.W. Ussery, C. He, I. Nookaew, Decoding the epitranscriptional landscape from native RNA sequences, Nucleic Acids Res. 49 (2021), e7.

[21] W.M. Gommans, N.E. Tatalias, C.P. Sie, D. Dupuis, N. Vendetti, L. Smith, R. Kaushal, S. Maas, Screening of human SNP database identifies recoding sites of A-to-I RNA editing, RNA. 14 (2008) 2074–2085.

[22] T. Chai, M. Tian, X. Yang, Z. Qiu, X. Lin, L. Chen, Genome-wide identification of RNA modifications for spontaneous coronary aortic dissection, Front. Genet. 12 (2021), 696562.

[23] Y. Li, X. Yang, N. Wang, H. Wang, B. Yin, X. Yang, W. Jiang, SNPs or RNA modifications? Concerns on mutation-based evolutionary studies of SARS-CoV-2, PLoS One 15 (2020), e0238490, https://doi.org/10.1371/journal.pone.0238490.

[24] A. Bakin, J. Ofengand, Four newly located pseudouridylate residues in *Escherichia coli* 23S ribosomal RNA are all at the peptidyltransferase center: analysis by the application of a new sequencing technique, Biochemistry 32 (1993) 9754–9762.

[25] A.V. Bakin, J. Ofengand, Mapping of pseudouridine residues in RNA to nucleotide resolution, Protein Synth. 297–309 (1998).