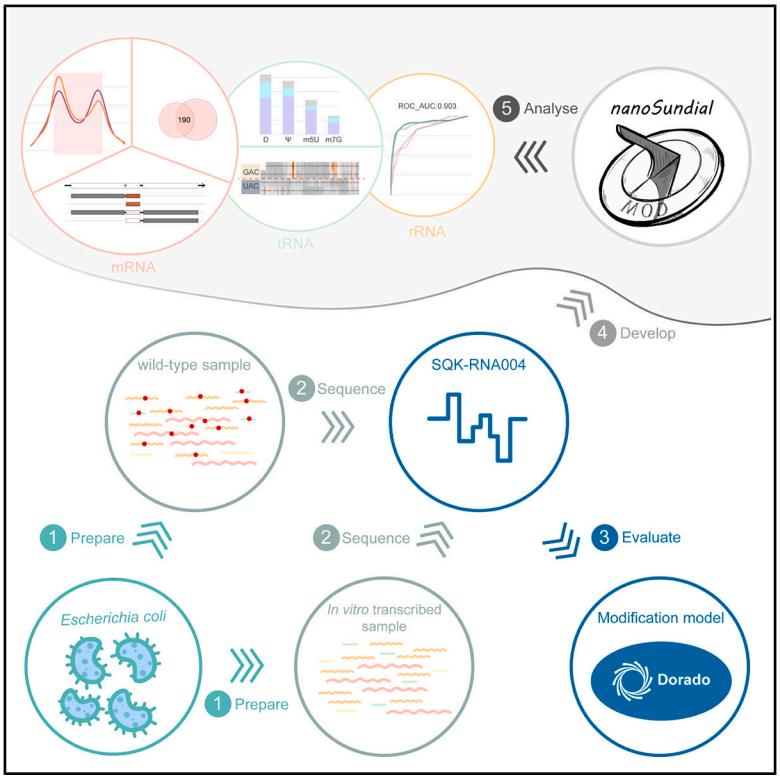
Cell Reports
Methods

Article

Enhanced detection of RNA modifications in *Escherichia coli* utilizing direct RNA sequencing

Graphical abstract



Highlights

- We evaluate sequencing yield of bacterial RNA obtained from nanopore-based approaches
- We show that Dorado RNA modification basecalling models produce false positives
- We develop nanoSundial, a modification detection tool for bacterial nanopore sequencing
- nanoSundial demonstrates strong performance in identifying bacterial rRNA modifications

Authors

Zhihao Guo, Yanwen Shao, Lu Tan, Beifang Lu, Xin Deng, Sheng Chen, Runsheng Li

Correspondence

runsheng.li@cityu.edu.hk

In brief

Guo et al. evaluate current RNA modification calling models used in nanopore sequencing on native and IVT RNA from *E. coli*. To address the false positives arising from these, they develop nanoSundial, a comparative modification detection tool for identifying prokaryotic RNA modifications.

Guo et al., 2025, Cell Reports Methods *5*, 101168
September 15, 2025
https://doi.org/10.1016/j.crmeth.2025.101168

CellPress

Cell Reports Methods

CellPress
OPEN ACCESS

Article

Enhanced detection of RNA modifications in *Escherichia coli* utilizing direct RNA sequencing

Zhihao Guo,[1,7] Yanwen Shao,[1,7] Lu Tan,[1] Beifang Lu,[2] Xin Deng,[2,3,4] Sheng Chen,[5] and Runsheng Li[1,3,4,6,8,*]

[1]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China
[2]Department of Biomedical Sciences, City University of Hong Kong, Kowloon Tong, Hong Kong, China
[3]Shenzhen Research Institute, City University of Hong Kong, Shenzhen, Guangdong 518057, China
[4]Tung Biomedical Sciences Centre, City University of Hong Kong, Hong Kong, China
[5]Department of Food Science and Nutrition, Faculty of Science, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[6]Department of Precision Diagnostic and Therapeutic Technology, City University of Hong Kong Matter Science Research Institute (Futian), Shenzhen, Guangdong, China
[7]These authors contributed equally
[8]Lead contact
*Correspondence: runsheng.li@cityu.edu.hk
https://doi.org/10.1016/j.crmeth.2025.101168

MOTIVATION Recent advancements in nanopore sequencing have significantly improved yield, accuracy, and noise reduction, opening exciting avenues for exploring epitranscriptomes. While these technologies have proven effective in various eukaryotic systems, their performance in detecting bacterial RNA modifications remains unverified. This study aims to validate and explore RNA modifications in bacterial RNA using nanopore sequencing data, addressing critical gaps in accuracy and reliability to advance our understanding of prokaryotic epitranscriptomics and its functional implications.

SUMMARY

RNA modifications play crucial roles in prokaryotic cellular processes. In this study, we found that the recent advances in direct RNA sequencing have improved yield, accuracy, and signal-to-noise ratio in bacterial samples. By evaluating four current RNA modification calling models in *Escherichia coli* transcriptome using native and *in vitro* transcribed (IVT) RNA, we found the models identified most known rRNA modifications but produced false positives. To address this, we developed nanoSundial, a comparative method leveraging raw current signals from native and IVT samples to *de novo* identify multiple RNA modifications. We optimized nanoSundial on well-studied *E. coli* rRNA modification sites and validated its effectiveness with tRNAs. It identified 190 stably modified mRNA regions, which enriched near the ends of highly expressed operons. This study highlighted the strengths and limitations of current nanopore-based modification detection methods on bacterial RNA, introduced a robust comparative tool, and elucidated previously uncharacterized mRNA modification landscapes.

INTRODUCTION

Chemical modifications of RNA occur naturally across all domains of life, with over 160 distinct types identified.[1–6] However, studying mRNA modifications in bacteria remains especially challenging. For instance, tRNAs and rRNAs are abundant and stable,[7] making their modifications relatively easy to study in both eukaryotes[8–10] and prokaryotes.[6,11–13] In contrast, bacterial mRNAs, which lack a poly(A) tail[14] and have exceptionally short half-lives,[15] are more difficult to isolate and characterize. This has made bacterial mRNA modifications particularly challenging to study, even though they likely play significant roles, leaving their modification landscape largely unexplored.[16] Recent improvements in bacterial RNA sample preparation allowed a higher mRNA yield and sequencing throughput, facilitating the analysis of bacterial transcriptome and epitranscriptome.[17,18]

Current RNA modification detection methods, such as liquid chromatography-mass spectrometry (LC-MS)[19–21] and immunoprecipitation-based techniques like methylated RNA immunoprecipitation sequencing (MeRIP-seq), have limitations in terms of single-nucleotide resolution,[22] target specificity,[23,24] and single-molecule detection.[25] Nanopore-based direct RNA

sequencing (DRS) offers a promising solution to these challenges.[26] RNA modifications generate distinct ionic current signals when RNA molecules pass through nanopores, enabling their detection via DRS.[26] Multiple computational tools are developed based on ONT RNA002 kits and can be classified into two categories based on whether their operating characteristics need a control sample. Single-mode methods include tools such as Epinano_SVM,[27] nanom6A,[28] MINES,[29] Tombo_de_novo,[30] and m6Anet.[25] And comparative methods encompass Epinano_Err,[27] nanocompore,[31] xPore,[32] ELIGOS,[33] Tombo_comp,[30] DirrErr,[34] and DRUMMER.[35] However, it is noticeable that these tools could generate highly inconsistent predictions.[17,36]

Recently, Oxford Nanopore Technologies (ONT) introduced an updated DRS kit (SQK-RNA004) and a corresponding flowcell (FLO-MIN004RA), which deliver higher sequencing yields, reduced current signal noise, and improved quality scores.[37] While this advancement opens opportunities for studying RNA modifications, existing methods optimized for RNA002 cannot be directly applied to RNA004 sequencing data. To address this, ONT officially released the Dorado rna004_130bps_sup@v5.0.0 RNA basecalling model, including several single-mode "all-context" models to detect RNA modifications with RNA004 data. These models had been trained to identify N^6-methyladenosine (m^6A),[10,38–40] pseudouridine (Ψ, pseU),[5] adenosine-to-inosine (A-to-I),[41] and 5-methylcytosine (m^5C).[42] Although these models have proven effective in various eukaryotic systems,[43] their performance in detecting bacterial RNA modifications remains unverified. Moreover, because previous nanopore-based approaches have shown limited success in detecting DNA modifications in bacteria,[44] further validation and exploration of RNA modifications in bacterial RNA004 data are crucial.

Here, we presented a high-quality bacterial transcriptome data generated via ONT DRS using the RNA004 technology, including wild-type (WT) and *in vitro* transcribed (IVT; unmodified) samples. We evaluated the performance of Dorado RNA modification models on bacterial RNA004 data. Furthermore, we introduced nanoSundial, a *de novo* comparative approach that utilized current features from WT and IVT RNA for accurate detection of prokaryotic RNA modifications based on RNA004 data. nanoSundial effectively identified multiple types of modifications and was validated on various RNA types, including tRNA, rRNA, and mRNA. Its reproducibility was confirmed through technical and biological replicates. We found enrichment of mRNA modifications at the start or end of coding sequences. In total, 190 stably modified coding sequence (CDS) regions were identified in *E. coli*, many of which cluster near the end of highly expressed transcription units (TUs) in each operon. Overall, this study not only demonstrated the feasibility of identifying bacterial RNA modifications via DRS with RNA004 data but also highlighted exciting avenues for investigating the bacterial epitranscriptome.

RESULTS

The RNA004 kit improved sequencing yield and quality for bacterial RNA

Total RNA was extracted from *E. coli* K-12 cells harvested at logarithmic growth phase. Following our previous pipeline, we performed size selection, rRNA depletion, and polyadenylation (ss&rd_RNA) before the DRS library preparation.[17] The K-12 total RNA was of high quality, with a RNA integrity number (RIN) value of 10 (Figures S1A and S1B). After size selection, most short RNA were removed (Figures S1C and S1D). After size selection and ribosome depletion (ss&rd), the 23S and 16S ribosomal RNA components were largely removed (Figures S1E and S1F). Using this ss&rd_RNA, we synthesized double-stranded cDNA and performed *in vitro* transcription to produce IVT RNA (described in STAR Methods section), which exhibited a gel profile distribution consistent with that of the ss&rd_RNA, confirming that the IVT RNA was qualified (Figures S1G and S1H). Sequencing libraries were constructed using the latest SQK-RNA004 kit and ran on the FLO-MIN004RA flowcell. The raw data were basecalled using the Dorado rna004_130bps_sup@v5.0.0 model. We refer to the native samples as ss&rd_004 and the IVT samples as IVT_neg_004.

The ss&rd_004 and IVT_neg_004 samples yielded 6.84 million reads (4201 Mb in bases) and 8.32 million reads (5659 Mb in bases), respectively, representing a more than 4-fold increase compared with previous SQK-RNA002 data (Figure 1A; Table S1). Unfiltered reads were subsequently mapped to the *E. coli* K-12 genome using minimap2. As a result, 93.65% of ss&rd_004 and 94.87% of IVT_neg_004 reads were mapped, corresponding to 92.06% and 90.78% of bases, respectively. We observed that unmapped reads are predominantly characterized by low read quality, exhibiting shorter read lengths and inferior Q scores (Figure 1B), likely attributable to technical issues during sequencing.[45] Consequently, all subsequent analyses were performed using reads filtered to exclude unmapped reads. For mapped reads, we noted higher read accuracy compared to SQK-RNA002 data. The median Q scores were 17.93 for ss&rd_004 and 19.17 for IVT_neg_004 (Figures 1B and 1C), compared to a median Q score of approximately 13 for SQK-RNA002.

All mapped bases were classified based on their annotation features (Figure 1D). In ss&rd_004, 52.3% of the total annotated bases fell within CDS, covering 3,872 genes with at least five reads each. In IVT_neg_004, 38.9% of the annotated bases were CDS regions, encompassing 3,150 genes (≥5 reads). Combined with the enhancement in sequencing yield, the number of CDS bases in both samples reached approximately 1,800 Mb (Figure 1D). In addition, rRNA bases accounted for about 2,665 Mb. This high sequencing yield provides sufficient depth for downstream analyses, including gene expression profiling as well as mRNA and rRNA modification detection.

The ss&rd_004 sample and IVT_neg_004 datasets showed a high correlation in gene expression, with a Spearman correlation of 0.95 (Figure 1E). We observed an increased ratio of ncRNA in IVT compared to WT in both RNA004 and RNA002 datasets. Specifically, in the RNA004 data, the overall proportion of ncRNA among all mapped bases rose from 1.3% in WT to 2.3% in IVT. Among these, the *ssrA* transcript showed a marked increase, with its contribution rising from 0.7% in WT to 2.0% in IVT. Previous research[16] has demonstrated that *ssrA* was highly expressed in *E. coli*, which can lead to its preferential amplification during the IVT process. This suggests that the reduced CDS overlap in IVT may reflect a relative shift in transcript composition due to biased ncRNA amplification during IVT.

CellPress
OPEN ACCESS

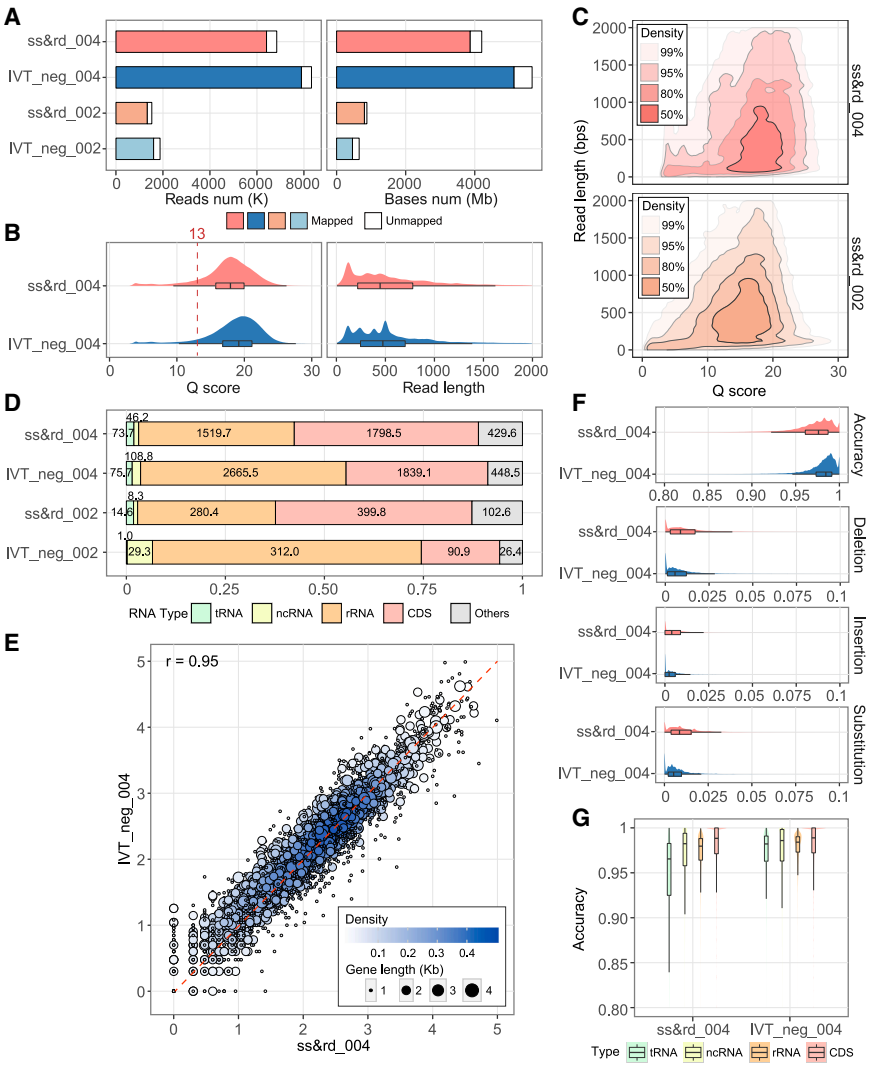


Figure 1. Raw read features and analysis of mapped reads based on Dorado basecalling results

(A) Total sequencing throughput from the samples of ss&rd and IVT_neg, sequenced with RNA002 (_002) and RNA004 (_004) technology. ss&rd means the sample was processed through size selection, rRNA depletion, and polyadenylation. IVT_neg means the *in vitro* transcribed sample is a non-modified negative control. The white blank represents the reads and bases that cannot align on the reference genome.

(B) Distribution of Q score and read length from RNA004 raw reads; 13 is the median Q score of RNA002 data.

(C) The relationship between read length and quality of raw reads was shown as a 2D density plot. Samples of ss&rd sequenced with RNA004 and RNA002 techniques were displayed. Read qualities were specified by Q scores.

(D) Proportion of mapped bases to annotation features, including tRNA, non-coding RNA (ncRNA), rRNA, CDS, and others. Numbers embedded in the bars indicate the number of bases in each category. Others covered the UTR region and intergenic region.

(E) The correlation plot shows the protein-coding gene expression levels between the ss&rd_004 and IVT_neg_004 samples. Spearman's rank correlation coefficients were calculated using CPM normalization. Undetected genes were excluded from the analyses. Each point represents one gene color coded by the density at the plot position. Gene length is indicated by the point size.

(F) Observed accuracy distribution including mapping accuracy, insertion, deletion, and substitution of ss&rd_004 and IVT_neg_004.

(G) Mapping accuracy of site level on different RNA types.

Boxplots in (B), (F), and (G) depict data distributions. The box represents the interquartile range (IQR), spanning the middle 50% of data (Q1 to Q3), with the central line indicating the median. Whiskers extend to the minimum and maximum values within 1.5 × IQR from Q1 and Q3, excluding outliers. See also Tables S1, S2, S3, and S4.

Besides, reverse transcriptase-RNA bias is known to introduce specific biases, potentially leading to the absence of certain RNAs in IVT samples.[46] In our study, we quantified the reverse transcription rates (Table S2). The similar reverse transcription rates of mRNA (4.4% for ss&rd_004 and 4.2% for IVT_neg_004) demonstrate that our samples show no notable difference in mRNA conversion efficiency.

Overall, the RNA004 kit and flowcell enabled the generation of high-yield and high-quality RNA sequencing data for both native and unmodified samples, greatly facilitating downstream transcriptome and epitranscriptome analyses.

The estimated and observed features of IVT surpassed those of the WT sample

Modifications can affect ONT raw current signals, leading to mapping errors such as insertions, deletions, and substitutions.[33,34,47] Consequently, the unmodified control, like an IVT sample, is expected to outperform the WT sample in observed accuracy across all 4 RNA types (Table S3). Consistent with this, our sequencing data revealed a median accuracy of 97.5% for the WT and 98.4% for the IVT (Figure 1F). Regarding all three types of mapping errors, WT showed significantly higher frequencies than IVT (two-tailed unpaired t test, $p < 0.001$; Table S4).

A previous study showed that the accuracy varied by basecalling models version for both WT and IVT under the RNA002 kit.[17] Here, we evaluated several features of mapped reads from WT and IVT across two recent basecalling models: rna004_130bps_sup@v3.0.1 and rna004_130bps_sup@v5.0.0 (Figure S2A). Dorado v5.0.0 model brought a significant improvement in mapping accuracy by at least 2.5% compared to v3.0.1 model. And v5.0.0 model increased the estimated Q score by approximately 7 for both WT and IVT samples, while maintaining consistent read length. Furthermore, in both versions of the basecalling models, the IVT samples demonstrated higher accuracy than the WT.

We next compared the site-level mapping accuracy among different RNA types in RNA004 dataset using Dorado v5.0.0 model. We found that the tRNA exhibited the largest accuracy gap (1.6%) between WT and IVT samples (Figure 1G). The tRNAs are known to harbor a high density of modifications. We believed these modifications account for the lower accuracy of tRNA in the WT sample compared to the other types.

Although Q scores were commonly used as filtering criteria in DNA sequencing, their applications in RNA studies remain relatively uncommon. For our RNA004 data, a Q score threshold of 7 was proved effective for pre-filtering unmapped reads in the v5.0.0 model (Figure S2B). However, this threshold may differ for other models: for example, a Q score of 5 worked better under the v3.0.1 model. To investigate the underlying significance of Q scores, we conducted a correlation analysis between the estimated Q score and the Phred-scale observed mapping accuracy. Although positive correlations were observed, the relationships were not statistically strong (Figure S2C). Therefore, it is recommended to analyze Q scores and observe accuracy independently in nanopore RNA studies.

In conclusion, our RNA004 data confirmed that IVT achieved higher read quality than WT, as reflected by both estimated and observed features across two basecalling model versions. The largest divergence occurred in tRNA, where the WT harbored numerous modifications. We also found that Q scores could be used to filter poor-quality RNA reads, but this approach should be tailored to each basecalling model version.

High false-positive rates were observed when using Dorado's modified basecalling models

To evaluate the performance of Dorado modified basecalling models in *E. coli*, a series of analyses were performed based on our WT and IVT datasets (Figure 2A). The model of rna004_130bps_sup@v5.1.0 was compatible with m^6A, Ψ, A-to-I, and m^5C all-context models. Here, we ran all the all-context RNA modified basecalling with the v5.1.0 models. The Dorado modified basecaller generates a BAM file containing modification information (Modbam). Using another official tool called Modkit, the Modbam file could be converted to a BED file (Modbed) for downstream analysis. We retained sites with adequate coverage (>20) in the Modbed file and analyzed their "fraction modified," defined as the ratio of modified nucleotides among all.

As a negative control, the "fraction modified" values of each site in IVT_neg_004 were expected to be considerably lower than that of ss&rd_004. When analyzing the result from Ψ all-context model, 92.4% of well-covered sites in IVT_neg_004 also appeared in ss&rd_004, which can be used for further comparison. However, only 40.7% of the modified fraction from well-covered sites was lower than ss&rd_004 (Figures 2B and S3A). This suggested that ss&rd_004 also may contain many false positives. When false positives are high, using a negative control can help filter them out. We calculated the difference in modified fraction by subtracting IVT from WT for each site. The average difference was close to zero, and the distribution was right-skewed (Figures 2C and S3B), suggesting certain sites displayed signal patterns indicative of true positive modifications at defined thresholds.

To verify whether the differences between WT and IVT could indicate true modifications, we initially focused on 36 well-studied modified sites in *E. coli* rRNA. These sites included ten Ψ sites, two m^6A sites, and three m^5C sites. Utilizing the Dorado v5.1.0 model for Ψ detection, we found clear differences (at least 35%) between WT and IVT (Figure 2D). Setting a threshold of 35% for the difference in "fraction modified" yielded 288 high-confidence Ψ sites, with 25% on mRNA, 39.2% on tRNA, and 30.2% on rRNA (Figure 2E). The majority of reliable sites on mRNA were enriched in the 5′ untranslated region (5′ UTR) (Figure 2F). Cross-referencing Dorado's tRNA Ψ positive sites against the 75 known tRNA sites deposited in MODOMICS database (a comprehensive public RNA modification database) identified 47 overlaps (Figure 2G). While this difference-based method captured known rRNA Ψ sites accurately, we also found five unreported rRNA sites (Figure 2H).

For the m^6A model, the differences between WT and IVT at the two known m^6A sites, A1618 and A2030, were more pronounced compared to those observed in the Ψ model, exceeding 65 (Figure S3C). Applying the same strategy, we spotted one additional unreported site on the 23S rRNA (A2448). We also identified another modified site, A1518 on 16S, which was known to be a $m^{6,6}A$ modification[48] instead of m^6A (Figure S3C). Among annotated regions, most high-confidence m^6A sites (152 in total) were in mRNA (Figure S3D). They clustered near the 5′ UTR (Figure S3E), and motif analysis with MEME revealed a strong signal in the polyadenylation A (poly(A)) region (Figure S3F).

For the m^5C and A-to-I models, a similar distribution pattern between WT and IVT "fraction modified" value difference was observed (Figures S3G and S3H). For the three known m^5C sites on rRNAs, the difference at C1962 on 23S rRNA is large (>75), but the differences at C967 and C1407 on 16S rRNA were minimal (<15) (Figure S3I). We couldn't simply use one cutoff to differentiate the true positive site using the current m^5C model. For the A-to-I, no known sites or ground truth exist in *E. coli*, so we couldn't determine a good cutoff. As a result, no further downstream analyses were done using m^5C and A-to-I Dorado models. Instead, we provided the sites whose differences between WT and IVT "fraction modified" values exceeded 50 from the two models in Table S5.

In conclusion, the Dorado modified basecalling model exhibited a high false-positive rate in *E. coli*. An IVT sample was necessary to filter out false positives. Yet, certain known Ψ and m^6A sites showed strong WT to IVT differences. This observation prompted us to leverage fraction-modified differences for high-coverage sites, thereby boosting our confidence in identifying true modification sites. Utilizing "fraction modified" differences between WT and IVT for filtering high false-positive sites could enable more detailed follow-up studies. The need to include IVT samples in Dorado's single-mode workflow has limited its direct applications for bacterial RNA modification detection. This constraint suggested that a comparative approach—leveraging WT-IVT signal differences—could offer a more effective solution.

nanoSundial, a comparative modification detection tool based on RNA004 was designed for prokaryotes

Despite the value of Dorado's modified basecalling, our analyses showed that it can produce a high false-positive rate in *E. coli*.

CellPress
OPEN ACCESS

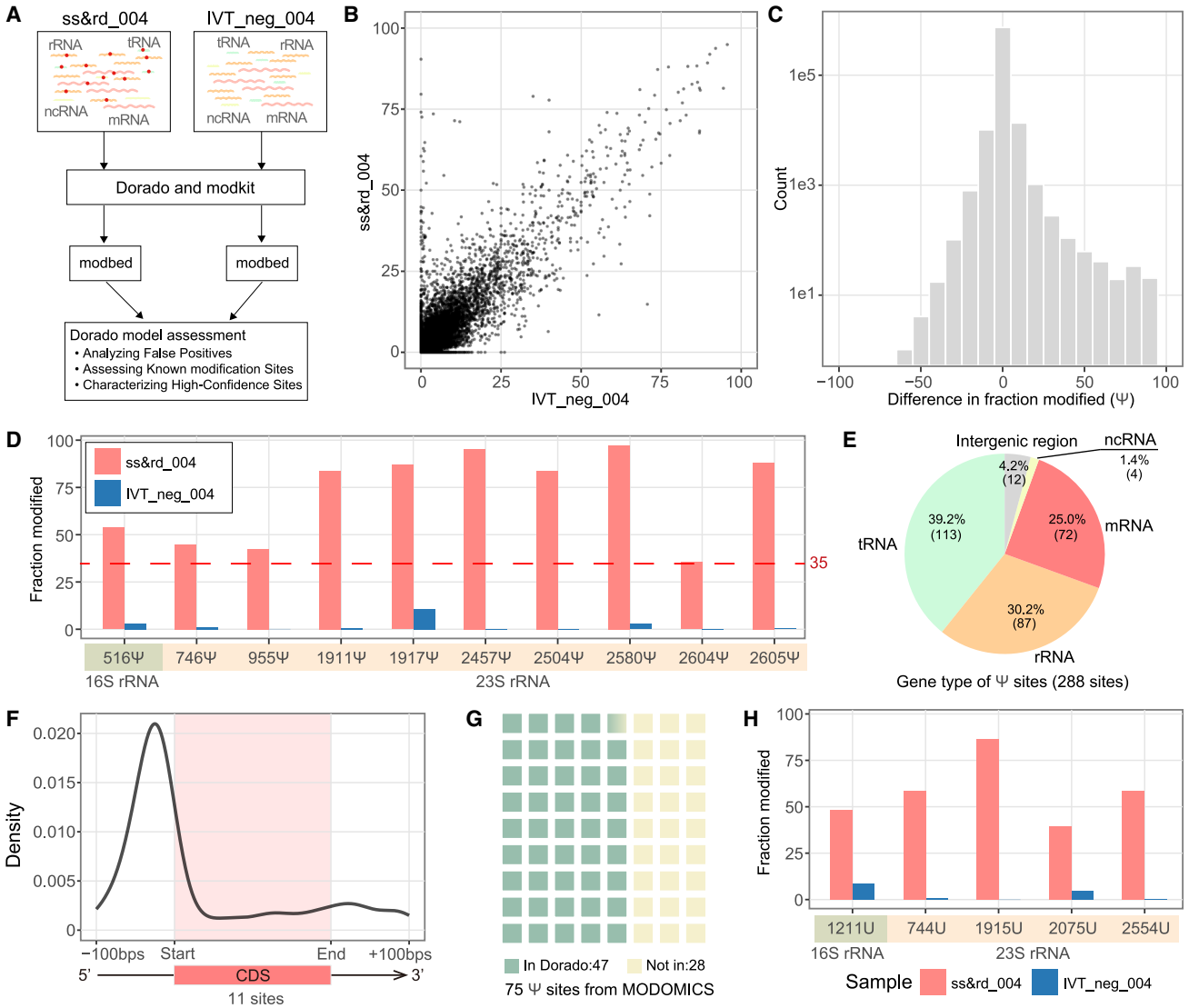


Figure 2. Evaluation and characterization of Dorado Ψ model

(A) Workflow of our analysis based on Dorado Ψ model. The raw data of ss&rd_004 and IVT_neg_004 were basecalled by Dorado v5.1.0 Ψ model. BAM files with modification information were transferred to BED files with modification information (Modbed) by the Modkit tool. Further assessments were done for the Dorado model results.

(B) The dotplot illustrated the "fraction modified" value of each position detected by the Dorado Ψ model from WT (ss&rd_004) and IVT (IVT_neg_004) samples.

(C) The right-skewed distribution of difference of "fraction modified" (WT - IVT). The *y* axis was treated with log10.

(D) "Fraction modified" value of 10 well-studied Ψ sites on *E. coli* rRNA. The lowest difference is around 35.

(E) The proportions and number of high-confidence sites within the expanded region. To investigate the characteristics of high-confidence sites located in the UTR region, we expanded the annotation region by 100 base pairs at both the start and end of each gene. This adjustment enabled the CDS to encompass nearly the entire mRNA sequence.

(F) Density distribution around the CDS region, with 11 in 72 sites located within the CDS.

(G) The waffle chart shows the number of Ψ sites validated by Dorado in MODOMICS. 47 Ψ sites on tRNA were identified in comparison with MODIMICS.

(H) "Fraction modified" value of high-confidence sites on 16S and 23S rRNA that were previously unreported.

See also Table S5.

Including an IVT (unmodified) sample helped exclude many false positives by comparing "fraction modified" between WT and IVT and indeed confirmed strong WT-to-IVT differences for certain known Ψ and m^6A sites. However, even after applying this filtering step, Dorado still detected some unreported sites, raising the possibility of algorithmic artifacts or undescribed modification types. Moreover, there were more than 160 known RNA modifications,[1–6] yet the Dorado v5.1.0 model only supports four (Ψ, m^6A, m^5C, and A-to-I). These observations underscored the need for a comparative strategy that exploits WT-IVT

signal differences and can potentially identify a wide spectrum of modifications.

To address the limitations, we developed nanoSundial, a comparative modification detection tool specifically tailored to prokaryotes using RNA004 data. nanoSundial took advantage of the improved signal-mapping accuracy in RNA004 (often referred to as resquiggle or eventalign). It then applied a statistical framework that directly compared the raw signal profiles of WT and IVT samples, rather than relying on single-mode classification from a limited subset of known modifications (Figure 3A; see STAR Methods). In doing so, nanoSundial aimed to provide a more comprehensive and reliable approach to detecting diverse RNA modifications in *E. coli*.

We used 36 well-characterized modification sites on *E. coli* rRNA (Figure 3B) as true positive sites. These included 11 sites from the 16S rRNA and 25 sites from the 23S rRNA. We generated receiver operating characteristic (ROC) curves to compare the classification performance using absolute differences in mean, dwell time, and standard deviation (STD) between ss&rd_004 and IVT_neg_004. Among these features, the absolute mean difference achieved the highest area under the curve (AUC) score of 90.3% (Figure 3C), while all four features surpassed 79.8%.

Fine-tuning the cutoff across various coverage levels has consistently posed a challenge for comparative tools.[49] To minimize false positives, we evaluated our method on two non-overlapping subsets of IVT_neg_004 (negative controls) with coverage levels from 10× to 2000× (Figure 3D). Coverage had a major impact. As coverage increased, the 99% confidence interval narrowed. A notable change in slope occurred at 50× coverage, after which the slope became more gradual. And the 99% confidence interval at higher coverage levels was close to zero. For RNA004 data, we recommend a minimum coverage of 50×, similar to the Tombo_level approach[30] used for RNA002.

Though the mean difference between IVT and WT is the key factor for RNA modification detection, analyzing the differences in dwell time and STD on the 23S rRNA demonstrated their possible contribution (Figures S4A and S4B). We would like to know if we could include one additional factor to achieve a better detection result. So, we evaluated the F1 score across various cutoff values using the mean together with dwell time or STD (Figure 3E). Combining STD with mean did not enhance the F1 score, but including dwell time did. The highest F1 score (0.34) was achieved with an absolute mean difference cutoff of 0.18 and an absolute dwell time difference cutoff of 1. These cutoffs offered an optimal balance between precision and recall. Consequently, we recommend a cutoff of 0.18 for the absolute mean difference and a cutoff of 1 for the absolute dwell time difference.

Determining an appropriate cutoff for the *p* value posed another challenge.[49] We observed that, above 200× coverage, the adjusted *p* value from MANOVA lost effectiveness (Figure S4C). However, at low coverage (50×), the adjusted *p* value remained highly informative, with the best performance at the −log10 (adjusted *p* value) = 3 (Figure 4F). Therefore, we recommend a *p* value of 1e-3 as a universal cutoff across coverage levels when using MANOVA.

Unlike tools specifically designed for m^6A modification detection, *de novo* modification detection methods lack prior knowledge of the motifs associated with each modification type. Additionally, modifications can alter the current signals in neighboring nucleotides.[31,33,50] Thus, a shift strategy is necessary to identify potential modification sites, similar to earlier pipelines.[17,33] Using our recommended cutoffs, performance plateaued at the shift of 4 bases (Figure 3F), indicating no further gains beyond that point.

After all optimizations, this approach successfully identified 34 out of 36 known rRNA sites (Figures 3G, S4D, and S4E). Aside from the detection of known positive rRNA sites on 16S and 23S rRNA, our analysis on 5S rRNA, which held no modification, found no positive hits, signifying reliable performance for true negatives. The two missing positive sites were A1618 and U1939 on 23S rRNA. A1618 may not be fully modified, and m^5U might not produce a signal change comparable to other modifications.

In summary, we developed nanoSundial—a modification detection tool optimized for bacterial data from the RNA004 kit. To optimize our approach, we employed 36 known modification sites on rRNA as ground truth. We assessed various features, including mean, median, standard deviation, and dwell time. By optimizing coverage thresholds, cutoff values, and shifting, nanoSundial demonstrated strong performance in identifying modifications on rRNA.

The performance of nanoSundial was validated in tRNA modification sites

Compared to RNA002, RNA004 data had a marked increase in sequencing yield per flowcell, accompanied by substantial improvements in site-level coverage (Figure 4A). The median coverage for tRNAs now exceeded 2,000 in RNA004. Meanwhile, tRNA expression levels in ss&rd_004 and IVT_neg_004 were highly similar (Figure 4B), with a correlation coefficient of 0.91. The enhancements in both sequencing depth and the expression level concordance on RNA004 provided a solid foundation for comprehensive tRNA modification analysis.

We drew upon 346 modification sites from MODOMICS[6] as the ground truth. By employing a 4-base shift strategy, nanoSundial successfully detected 201 of these sites (Figure 4C). Focusing exclusively on tRNAs with sufficient sequencing depth in our dataset, we achieved a 68.3% detection rate for known tRNA modifications, belonging to 34 types of tRNA modifications. We also calculated site detection ratios for the 13 most common tRNA modification types (Figure 4D). Of these 13 types, nanoSundial detected over half the sites in 9 types.

nanoSundial exhibited stable performances across tRNA paralogs. As an example, we examined mean WT to IVT differences in lysine (lys) tRNAs (Figure 4E), which include six paralogs. The distribution of mean differences was remarkably similar across all paralogs. In each case, Ψ sites exhibited the greatest mean differences compared to other modifications (Figure 4E). The example of one lys paralog (*lysW*) further indicated the dramatic change of current signals between WT and IVT around modified sites (Figure S5A). Additionally, we provided mean differences for another tRNA with multiple paralogs, aspartic acid (asp) (Figures S5B and S5C). All these showcases demonstrated nanoSundial's capability of detecting modifications on tRNAs.

nanoSundial effectively distinguished the modification profiles among tRNA paralogs that bind to different codons. Identical tRNA types may exhibit distinct modifications depending on their

CellPress
OPEN ACCESS

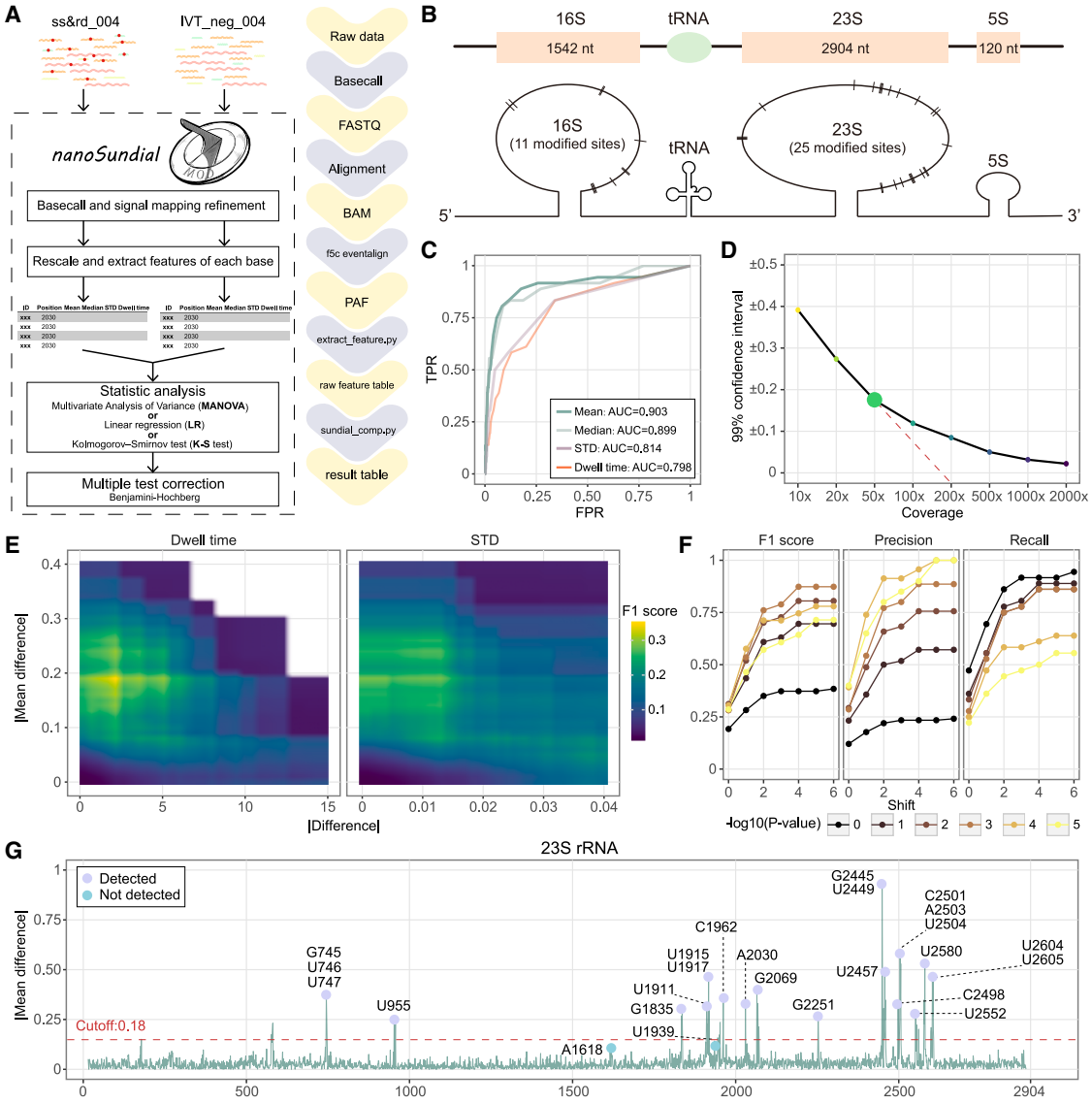


Figure 3. Workflow and optimization of the nanoSundial

(A) Raw data WT (ss&rd_004) and IVT (IVT_neg_004) were basecalled using Dorado, filtered with SAMtools, and then underwent signal mapping refinement via f5c eventalign. NanoSundial then rescales the data and extracts features including mean, median, dwell time, and STD for each site. Finally, statistical analyses, including MANOVA, LR, or the K-S test, were performed, followed by multiple testing corrections using the BH method. The right side highlights all involved programs in purple boxes and files in yellow boxes.

(B) The diagram shows the distribution of 36 known modification sites on the *E. coli* rRNA, representing the ground truth.

(C) Receiver operating characteristic (ROC) curves for the four features were generated based on the absolute differences between the two conditions.

(D) The 99% confidence interval of the absolute difference in mean when comparing two independent negative controls at different coverage levels. This value shows a sharp change in slope at a coverage of 50×.

(E) F1 score distributions when applying the absolute mean difference combined with the absolute differences in dwell time or standard deviation as co-cutoffs. The highest F1 score is achieved when the absolute difference of dwell time is around 1 and the mean is around 0.18.

(F) Dot and line plots illustrate the variation in F1 score, precision, and recall (on the *y* axis) as different cutoff values of adjusted *p* value are applied, with the *x* axis representing the shift length at a 50× coverage. The highest value is observed when −log10 adjusted *p* value reaches 3, and the results stabilize when the shift is set to 4.

(G) Plot shows the absolute mean difference on the 23S rRNA. The purple points represent the sites within a 4-base shift that can be detected by nanoSundial, whereas the blue points indicate those that were not detected.

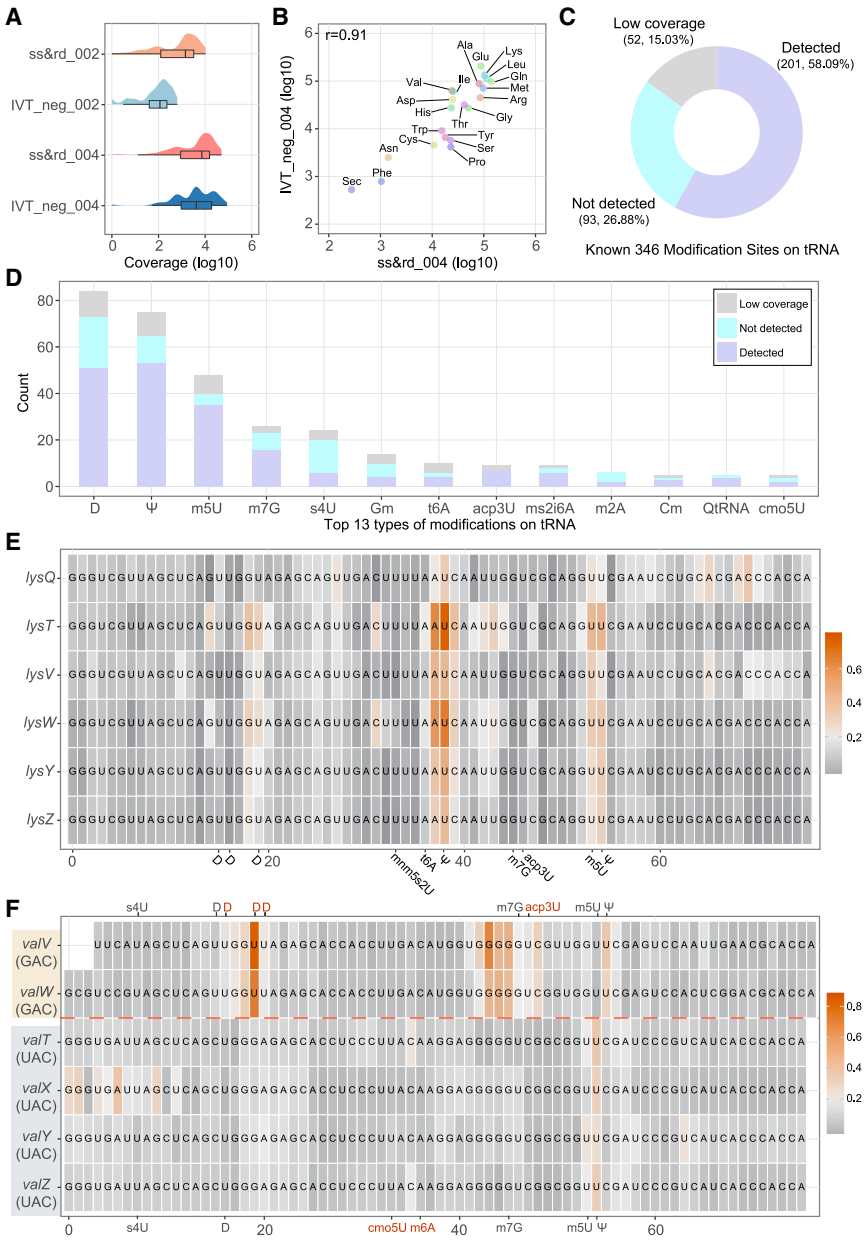


Figure 4. Characterization and validation of nanoSundial in tRNA applications

(A) Coverage (log10 scale) on tRNA for the RNA004 data compared with the RNA002 data. The box represents the interquartile range (IQR), spanning the middle 50% of data (Q1 to Q3), with the central line indicating the median. Whiskers extend to the minimum and maximum values within 1.5 × IQR from Q1 and Q3, excluding outliers.

(B) Correlation between WT (ss&rd_004) and IVT (IVT_neg_004) on tRNA expression levels.

(C) The performance of nanoSundial using MODOMICS as the ground truth, detecting 201 out of 346 sites (58%), with 52 positions (15%) having coverage below 50×.

(D) Top 13 types of modifications on tRNA.

(E and F) Heat maps illustrate the absolute mean difference for the paralogs of tRNA lysine (lys) (E) and valine (val) (F). The colors at each position represent the magnitude of these absolute differences. For valine, the paralogs can be categorized into two classes based on their binding codons (GAC and UAC), which also exhibit different modifications. These different modifications are highlighted in the red text.

where the sequencing signal ceases to change). However, tRNAs with secondary structures that are effectively unfolded during sequencing are successfully detected. Our aligned reads, which map accurately to the reference, consistently exhibit high quality and full length (Figures S5D and S5E). Furthermore, our data demonstrate that the most significant differences in tRNAs strongly correlate with known modification sites (Figures 4E and 4F). This indicates that the signals we detect are primarily driven by RNA modifications rather than secondary structure effects.

The improved sequencing data provided by RNA004 has facilitated the analysis of tRNA. Although tRNAs exhibit a larger basecalling accuracy gap due to their high modification density (Figure 1G), the pronounced signal difference between WT and IVT samples further supports the reliable detection of tRNA modifications. nanoSundial not only detected a greater number of modifications but also identified modifications across different codons within tRNAs.

codon binding. We further analyze multi-codon tRNA, such as $tRNA^{Val}$, which binds GAC or UAC (Figure 4F). Specifically, *valV* and *valW*, which pair with the GAC codon, possess three additional 5,6-dihydrouridine-5′-monophosphate (D) modifications and one additional 3-(3-amino-3-carboxypropyl) uridine (acp3U) modification compared to their paralogs that bind the UAC codon. As a result, regions adjacent to these additional modification sites displayed higher absolute mean differences in *valV* and *valW* (Figure 4F). These findings pointed out nanoSundial's ability to differentiate the modification profiles of tRNA paralogs with different binding codons.

Theoretically, tRNA secondary structures could influence nanopore sequencing by causing pore blockage or read ejection, potentially resulting in no read or low-quality reads (e.g.,

Consistent modification detection performance of nanoSundial across RNA types

Modified bases can influence the nanopore raw current signals for neighboring regions.[25] In our analysis, nanoSundial output all sites satisfied the filter of current differences, including both the modified sites and neighboring sites. To refine these outputs,

CellPress
OPEN ACCESS

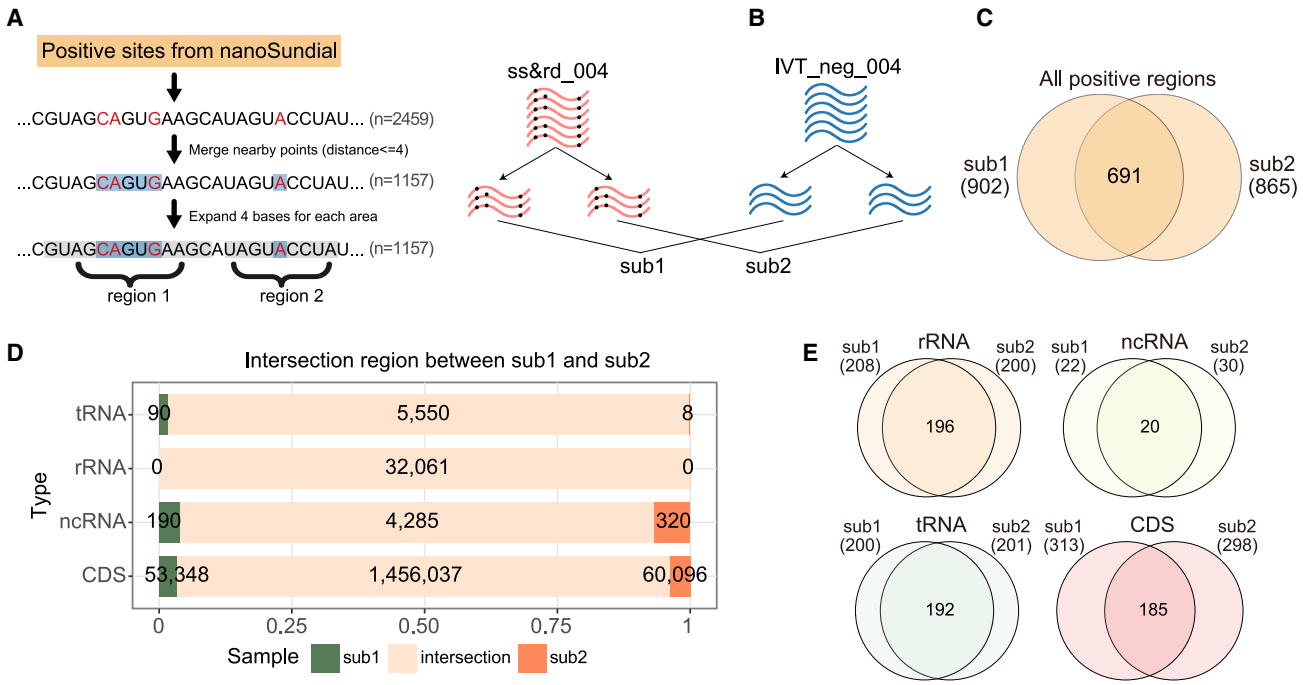


Figure 5. Reproducibility testing of technical replicates

(A) The merging strategy will connect all nearby positive sites within 4 bps, then expand by an additional 4 bps, ultimately identifying and outputting these positive regions.

(B) A workflow to subsample and analyze by nanoSundial. The datasets ss&rd_004 and IVT_neg_004 were divided into two completely non-overlapping subsets, which are then compared using nanoSundial to obtain sub1 and sub2.

(C) Venn plots show the intersection of overall and CDS between the results from two non-overlapping subsamples (technical replicates).

(D) The intersection of the detected sites in sub1 and sub2 within the annotated regions. The green areas represent sites detected only in sub1, the orange areas represent sites detected only in sub2, and the overlapping section indicates sites detected in both sub1 and sub2.

(E) The intersection of the positive regions from the nanoSundial results for sub1 and sub2.

we employed a merging strategy to consolidate adjacent sites into positive regions (Figure 5A; see STAR Methods).

Reproducibility remains a major challenge in modification detection. Previous studies suggested multiple replicates.[31] To evaluate the reproducibility of nanoSundial, we conducted both technical and biological replicate analyses. For technical replicate analysis, WT and IVT were divided into two non-overlapped halves. Each half of the WT sample was then compared with one-half of IVT, resulting in two sets of results from nanoSundial, referred to as sub1 and sub2 (Figure 5B). By applying the coverage cutoff of at least 50, approximately 4% of the sites in one subset (sub1 or sub2) were absent in the other subset and were therefore excluded (Figure 5C). Sub1 returned 902 positive regions, while sub2 returned 865 (Figure 5D). Overall, their intersection consisted of 691 regions, accounting for approximately 76% of sub1 and 78% of sub2. The intersection rate for mRNA was the lowest, with 59.1% in sub1 and 62.1% in sub2 (Figure 5E). In contrast, the intersection rates for tRNA, rRNA, and ncRNA exceeded 95% (Figure 5E). This discrepancy suggested that many modifications on mRNA may be only partially modified, resulting in differences between WT and IVT mRNA that were insufficient to be reliably detected at certain modification points. This issue became more pronounced when sequencing coverage was reduced through subsampling, making the differences even less significant.

We then explored biological reproducibility by sequencing an additional biological replicate of ss&rd_004 (WT). The size selection step before rRNA depletion and polyadenylation may result in the different proportions of small RNA, rRNA, and mRNA between batches. Nonetheless, the replicate exhibited high yield and quality, with 8,666,818 sequencing reads and a median Q score exceeding 17.93 (Table S1). The correlation of gene expression between the two WT replicates is 0.98 (Figure 6A).

Both WT biological replicates were compared to the IVT_neg_004 sample using nanoSundial. After applying the merging strategy, replicate 1 (rep1) yielded 1,157 positive regions from the initial 2,459 positive sites, and replicate 2 (rep2) yielded 1,797 positive regions from 3,512 sites (Figure 6B). Most positive regions were in the CDS, with 452 for rep1 and 856 for rep2. Overall, rep1 and rep2 shared 741 overlapping regions. The intersection rates were especially high for tRNA and rRNA. For example, 194 of ~200 rRNA regions overlapped (Figure 6C). Despite 452 regions in rep1's CDS and 856 in rep2's, only 190 overlapped, indicating a lower intersection rate in mRNA regions.

In summary, we incorporated a merging strategy and demonstrated through both technical replicate and biological replicates that nanoSundial had good reproducibility, especially on tRNA

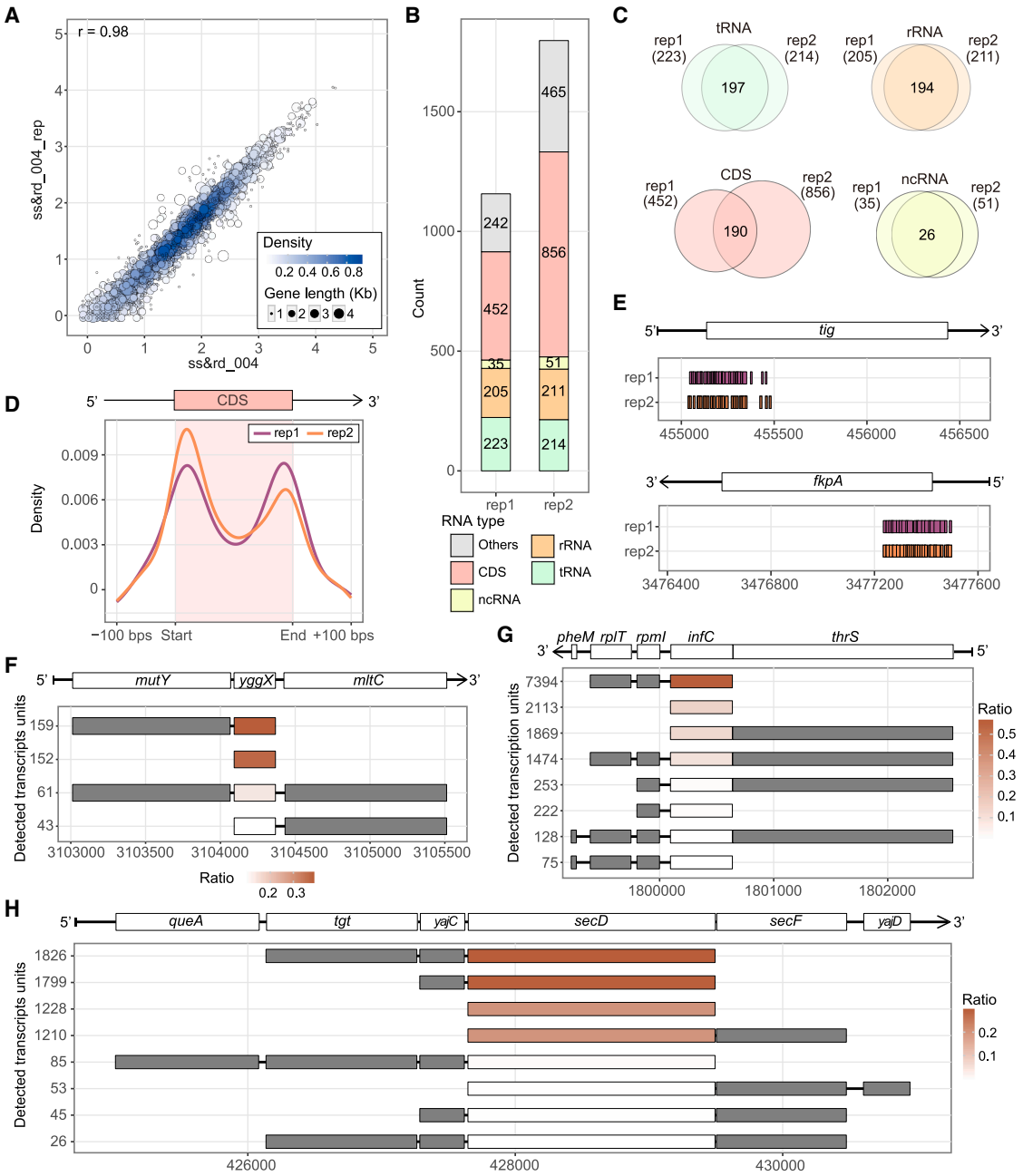


Figure 6. Exploration analysis of high-confident positive regions and transcription unit structures of stably modified genes

(A) Correlations of protein-coding gene expression levels between the ss&rd_004 with respective replicate. Each point represents a single gene, color coded by density at the plot position. The size of each point indicates gene length.

(B) The barplots demonstrate the number of positive regions within the annotation feature of the results of two biological replicates. Numbers embedded in the bars indicate the amount of positive regions in each category, including tRNA, rRNA, ncRNA, CDS, and others. Others covered the UTR region and intergenic region.

(C) The intersection of positive regions on each category of gene type.

(D) Density distribution around the CDS region of two biological replicates result.

(E) Showcase of the clustering of positive regions on the genes *tig* and *fkpA*. Positive regions were enriched in the start of mRNAs.

(F–H) The composition and proportion of the transcription units (TUs) for yggX, *infC*, and *secD*. The colors indicate the proportion of each TU relative to the total reads for one gene.

See also Table S6.

and rRNA regions. Moreover, in two biological replicates, we detected 190 intersecting positive regions within the CDS. These regions were of higher confidence compared to other positive regions in the CDS, making them prime candidates for downstream analysis.

mRNA modifications were enriched at the start and end of CDSs

By analyzing the density of all positive regions located on mRNA, we found that modifications were primarily concentrated within the gene body. The enriched peaks of modification regions were close to the CDS start and end (Figure 6D). This distribution was consistent in the two biological replicates. While observing the distribution characteristics of positive regions, we noticed a phenomenon that many nearby regions were enriched in one mRNA.

We further analyzed the location of the 190 high-confident positive regions in the CDS area. The 190 overlapping regions involve 71 genes, with operon information from RegluonDB[51] described in Table S6. We defined the 71 genes as stably modified genes. Although 55 out of 71 genes have only one or two positive regions, some mRNAs, like *tig*, *fkpA*, *cutC*, *sbmA*, and *OmpC*, exhibited a clustering of adjacent regions at around the 5′ or 3′ ends (Figure 6E).

Among the 71 genes, we found that 25 genes are part of operons containing three or more genes, and 22 in 25 genes are predominantly located at the beginning or end of the operons (see Table S6). Notably, some operons that start or end with a stably modified gene, like *xerD*, *lptD*, *thrS*, *rplQ*, and *rpsQ*, contain more than five genes. In particular, the *rpsQ* operon consists of 11 genes. Within the 71 stably modified genes, only 3 genes, including *yggX*, *infC*, and *secD*, were not located at the start or end of the containing operon. Further analysis of their TU compositions revealed that the three genes were positioned at the start or end of the most abundant TUs (Figure 6F–6H) in the operon. Each of the highly expressed TUs comprised more than 85% of the total expression for each operon.

We investigated the biological functions of operons (with more than three genes) that contained stably modified genes. Among these operons, nine had a stably modified gene located at the start of the operons, encompassing a total of 42 genes, while eleven had a stably modified gene at the end of the operons, accounting for 50 genes. Both sets were enriched in the "translation" Gene Ontology (GO) pathway (Figure S6A). Notably, the operons with stably modified genes at the ends were also associated with multiple "ribosome"-related functions. These findings suggested that operons with stably modified genes at the start play key biological roles in translation, with those with stably modified genes at the end may have broader implications for ribosome biology.

In summary, our biological replicates confirm that positive regions on mRNA tend to cluster near the CDS boundaries. Moreover, the 71 stably modified genes typically lie near the start or end of their TUs, suggesting a potential functional significance for modification sites in these regions.

DISCUSSION

In recent years, substantial progress has been made in understanding the roles and functions of RNA modifications. Modifications in eukaryotes, such as m^6A and m^5C, are well characterized, but bacterial mRNA modifications remain less understood.[52,53] The latest nanopore RNA sequencing kit (RNA004) had great potential to tackle this gap, as it provides higher yield, lower signal-to-noise ratio, and higher quality reads compared to the previous RNA002. Several computational methods, including single-mode and comparative methods, were originally designed for RNA002. To date, only the Dorado modified basecalling model has been developed for the RNA004; it represents a single-model method that does not require negative controls and can analyze modification stoichiometry at the site level.

In this case, we applied the latest RNA004 technology to detect RNA modifications in *E. coli*. We conducted an improved experimental pipeline and prepared sufficient yields of both native (WT) and unmodified (IVT) RNA samples. We noticed that the proportion of bases mapping to CDS regions was lower in the IVT sample (38.9%) compared to the WT sample (52.3%). Although we observed an increased proportion of ncRNA in IVT, including a notable rise in *ssrA* coverage from 0.7% to 2.0%, this alone cannot fully account for the overall reduction in CDS representation. One possible contributing factor is variation in ribosomal RNA depletion efficiency between samples, which can influence the relative abundance of coding versus non-coding transcripts.[54] Nevertheless, gene expression between IVT and WT samples showed a strong correlation (Spearman's ρ = 0.95), with comparable coverage of CDS bases (1798.5 Mbps for WT and 1839.1 Mbps for IVT). Furthermore, comparative analyses of CDS regions between IVT and WT samples revealed highly consistent coverage patterns (Figures S6B and S6C).

During the course of our analysis, Dorado continued to release basecalling models. Therefore, for the initial basecalling, we used v3.0.1 and v5.0.0. However, as Dorado rapidly updated its models, v5.1.0 introduced an expansion from three to five RNA modification models, which covered 4 types of modification. As a result, we adopted v5.1.0 for all analyses involving modification models. The basecalling differences between versions 5.0.0 and 5.1.0 were minimal. For version 5.1.0, a filter threshold of 7 remains optimal, preserving over 95% of mapped reads (Figure S6D).

Combined with the native and negative control (IVT) data, we evaluated the modification detection models provided by Dorado (Ψ, m^6A, m^5C, and A-to-I) on our *E. coli* data and found a large number of false positives. As a result, we concluded that the Dorado's DRS single-model modification basecalling models couldn't be directly applied to prototypes without using a negative control to filter away the false positives. By observing the "fraction modified" values of known modified sites on rRNA, we proposed a simple pipeline to identify the high-confidence sites using the difference in "fraction modified" between WT and negative control. After optimization, the Dorado pipeline could successfully identify all known Ψ and m^6A sites. However, even with the improvement with IVT, it detected additional Ψ and m^6A sites on rRNA that have not been previously reported, which are likely to be false positives. What's more, there were more than 160 types of RNA modifications, but the Dorado v5.1.0 model can only detect 4 types. These limitations pointed to the necessity of launching a *de novo* modification detection tool to detect bacteria.

Apart from the limitations of Dorado, previous RNA modification detection methods based on RNA002 were not compatible with RNA004 data. Furthermore, modification patterns can differ between prokaryotes and eukaryotes. For example, while m6A in eukaryotes primarily occurs at DRACH motifs via methyltransferase-like (METTL) proteins , prokaryotic systems may use different recognition rules. Hence, we aimed to develop a comparative method. Comparative methods required a low modification or no modification sample as a negative control. Existing RNA002-based comparative methods rely on either current features or alignment features.[36] Previous research showed that alignment features, such as mismatches, highly depended on the basecaller version.[17] Additionally, modifications like m^5C and m^4C often produce subtle or negligible current shifts, making them difficult to classify as basecalling errors.[16,33] As RNA004 improved mapping accuracy, the mismatch differences decreased between WT and IVT, making alignment features even less reliable. For example, differences at A1618 and A2030 (Figures S7A and S7B) are smaller in RNA004 than in RNA002.[27] Current features, on the other hand, remain largely unaffected by improved basecalling accuracy. So, we think that the comparative method with current features could better detect modification compared to methods using alignment features in bacteria. Therefore, we designed nanoSundial, which is a comparative modification detection tool for prokaryotes that specifically targets current features in RNA004 data.

The comparative strategy of nanoSundial is training-free, unlike machine-learning-based methods that require a large, labeled training set, iterative backpropagation optimization, and graphics processing unit (GPU) accelerated parameter tuning. Instead, it relies on direct statistical comparison with a negative-control sample, enabling broad detection of RNA modifications. Moreover, it can identify a wider range of modifications than single-mode tools. Despite the challenge in base shifting, nanoSundial detected more known modification sites in both rRNA and tRNA than a prior RNA002-based pipeline. Specifically, while the previous study[16] using ELIGOS detected 198 of 346 tRNA modifications and 31 of 36 rRNA modifications (with a broad ±10 bps window), our pipeline detected 201 of 346 tRNA and 34 of 36 rRNA modifications using a narrower ±4 bps shift window. This reduction from ±10 to ±4 bps greatly improved positional precision, enabling more confident and fine-grained detection of RNA modifications. This highlights the increased accuracy and sensitivity enabled by RNA004 and the design of nanoSundial, representing a significant advance in high-resolution *de novo* RNA modification detection.

Several methods currently exist to erase specific RNA modifications, such as METTL3 knockdown in eukaryotes.[28] This strategy could enhance the performance of comparative methods like nanoSundial, enabling more precise site-specific results. Since nanoSundial does not focus on specific modifications, we did not perform specific validation experiments. Instead, employing IVT as a negative control effectively validates known modified sites on *E. coli*.

After optimizing on rRNA and validating on tRNA, we employed a strategy to merge adjacent sites and test the reproducibility of nanoSundial. Although the reproducibility rates for tRNA, rRNA, and ncRNA exceeded 95%, the rate for mRNA was only 61%. Additionally, the replicate results showed a high intersection on rRNA and tRNA, but not on mRNA. This indicated that the modifications on mRNA are partial and unstable, consistent with observations in rRNA, tRNA, and small nuclear RNA (snRNA) sites from humans and yeast.[10]

Still, we detected 190 positive regions from the replicate intersections, spanning 71 stably modified genes on CDS, which represent high-confidence mRNA modifications. On the gene level, 68 of 71 stably modified genes were located at the start or end of the operon. For 3 genes that did not satisfy this regulation, we found these genes were located at the beginning or end of the most highly expressed TU. This suggested that mRNA modifications were closely related to operon structure, which deserved further verification and study. Besides, among these 3 genes, we observed that the *napF* operon contains 15 genes. However, in the TU research, only three genes—*napA*, *napD*, and *napF*—were identified. Notably, *napF* also conforms to the positional pattern observed at the end of the operon. This discrepancy may suggest that the operon annotations in RegulonDB[51] could be further refined, highlighting the potential application of ONT DRS in transcriptome analysis (Figure S7C).

In addition, we assessed nanoSundial's consistency with Dorado in detecting other modifications. Comparing high-confidence Ψ sites from Dorado with nanoSundial revealed a substantial overlap (177 out of 288 sites), whereas the m^6A all-context model performed poorly, intersecting at only 38 out of 234 sites (Figure S7D). This finding indicated a high level of consistency between nanoSundial and Dorado Ψ model, mainly for rRNA and tRNA (Figure S7E). By contrast, the lower overlap for m^6A likely reflects limited accuracy in Dorado's all-context m^6A model, and the presence of partial modifications in m^6A—such as A1618—which nanoSundial struggled to detect. The high Ψ overlap and low m^6A overlap between Dorado and nanoSundial suggest that Dorado may produce more false positives in bacterial m^6A detection.

In conclusion, the enhanced quality and yield of RNA004 data have substantially supported our analysis of bacterial RNA. We utilized the RNA004 data to evaluate the performance of the Dorado modified basecalling models and found that Dorado produced a high number of false positives when detecting bacterial RNA modifications. Although our method effectively identified 10 known pseU sites and 2 m6A sites on rRNA, some unreported modified sites remained. To address the limitations of the Dorado model and to establish a tool capable of precisely detecting RNA modifications in bacteria, we developed nanoSundial, a comparative method for prokaryotic RNA004 data based on current features. Following our optimizations and evaluation, nanoSundial not only demonstrated strong performances in detecting RNA modification in bacterial tRNA and rRNA but also revealed a preferential enrichment of modifications in bacterial mRNA. This work opens avenues for investigating bacterial mRNA modifications at the transcriptome scale.

Limitations of the study

Our work also has several limitations. Although mRNA and rRNA exhibit comparable reverse transcription efficiencies (~4%), we detected substantial disparities in ncRNA transcription rates,

which were 14.5% for WT versus 4.7% for IVT samples (Table S2).

For our comparative tool, first, this tool is based on a k-mer detection approach and therefore couldn't identify specific types of RNA modifications (e.g., m6A, m5C) within a given region. Second, compared to single-molecule analysis, quantifying the stoichiometry of modifications at the site level remains challenging, complicating the detection of partial modifications. Third, our method is *de novo* and focuses more on highly modified sites, so the cutoff selection was quite harsh. The ground truth for cutoff selection relies on 36 known rRNA-modified sites, most of which are highly modified. With over 160 distinct modification types, some inducing only subtle current differences between WT and IVT samples, the linear cutoff used in nanoSundial may be overly stringent for certain modifications. Future improvements to nanoSundial could involve implementing k-mer-specific cutoffs once additional ground truth data are available.

In addition, this study primarily focuses on computational metrics and provides high-quality bacterial nanopore RNA004 sequencing data. Although several modification regions in bacterial mRNA and operons were identified, further investigation and experimental validation are needed to confirm these findings.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to the lead contact, Runsheng Li (runsheng.li@cityu.edu.hk).

Materials availability

This study did not generate new reagents.

Data and code availability

- All sequencing data have been deposited in the NCBI database under BioProject NCBI: PRJNA1198641. Basecalled files has been uploaded to figshare at figshare: https://doi.org/10.6084/m9.figshare.28106741.v2, figshare: https://doi.org/10.6084/m9.figshare.28106909.v1, and figshare: https://doi.org/10.6084/m9.figshare.28106906.v1.
- The custom scripts and plot scripts used for this paper are available in Zenodo (https://zenodo.org/records/14553666) and on GitHub (https://github.com/JeremyQuo/Ecoli_004_ONT_DRS_scripts). nanoSundial is available in Zenodo (https://doi.org/10.5281/zenodo.15904129) and on GitHub (https://github.com/lrslab/nanoSundial).
- Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

Open Access made possible with partial support from the Open Access Publishing Fund of the City University of Hong Kong. We thank the High-Performance Computing Cluster at the City University of Hong Kong for providing us with computational resources. We also thank Hasindu Gamaarachchi from the Garvan Institute of Medical Research for their valuable assistance on GitHub (https://github.com/hasindu2008/f5c/issues/163). Additionally, we are grateful to Ms. Qingqiu Jiang for designing the logo for nanoSundial.

This work was supported by the Early Career Scheme (CityU 21100521) and General Research Fund (11105524) from the Research Grants Council of the Hong Kong Special Administrative Region, China; the Hong Kong Health and Medical Research Fund (project number 08194126); and new Research Initiatives support from City University of Hong Kong (project number 9610497) to R.L.

AUTHOR CONTRIBUTIONS

Z.G. and R.L. designed the research. Y.S. and L.T. conducted the experiments. Z.G. conducted data analysis, visualization, and software development. Y.S. conducted the software test. Z.G. and Y.S. wrote the initial manuscript draft; L.T., B.L., X.D., and R.L. made corrections and edits. All authors have read and approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this manuscript, the authors utilized ChatGPT to assist with language polishing. Following the use of this tool, authors carefully reviewed and edited the content as necessary and take full responsibility for the content of the publication.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Bacterial RNA preparation
- METHOD DETAILS
  - *In vitro* transcribed (IVT) RNA preparation
  - Nanopore direct RNA sequencing and data processing
  - Reference selection and annotation strategies
  - Gene function enrichment
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Gene expression analysis
  - Current feature comparison with nanoSundial
  - Pre-processing
  - Calculation of region intersection
  - Evaluation metrics calculation

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.crmeth.2025.101168.

Received: February 19, 2025
Revised: June 19, 2025
Accepted: August 14, 2025
Published: September 9, 2025

REFERENCES

1. Wang, Y., Li, Y., Toth, J.I., Petroski, M.D., Zhang, Z., and Zhao, J.C. (2014). N6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. Nat. Cell Biol. *16*, 191–198. https://doi.org/10.1038/ncb2902.
2. Geula, S., Moshitch-Moshkovitz, S., Dominissini, D., Mansour, A.A., Kol, N., Salmon-Divon, M., Hershkovitz, V., Peer, E., Mor, N., Manor, Y.S., et al. (2015). Stem cells. m6A mRNA methylation facilitates resolution of naive pluripotency toward differentiation. Science *347*, 1002–1006. https://doi.org/10.1126/science.1261417.
3. Su, R., Dong, L., Li, C., Nachtergaele, S., Wunderlich, M., Qing, Y., Deng, X., Wang, Y., Weng, X., Hu, C., et al. (2018). R-2HG Exhibits Anti-tumor Activity by Targeting FTO/m(6)A/MYC/CEBPA Signaling. Cell *172*, 90–105. e23. https://doi.org/10.1016/j.cell.2017.11.031.

4. Squires, J.E., Patel, H.R., Nousch, M., Sibbritt, T., Humphreys, D.T., Parker, B.J., Suter, C.M., and Preiss, T. (2012). Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. Nucleic Acids Res. *40*, 5023–5033. https://doi.org/10.1093/nar/gks144.

5. Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M., and Gilbert, W.V. (2014). Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. Nature *515*, 143–146. https://doi.org/10.1038/nature13802.

6. Boccaletto, P., Stefaniak, F., Ray, A., Cappannini, A., Mukherjee, S., Purta, E., Kurkowska, M., Shirvanizadeh, N., Destefanis, E., Groza, P., et al. (2022). MODOMICS: a database of RNA modification pathways. 2021 update. Nucleic Acids Res. *50*, D231–D235. https://doi.org/10.1093/nar/gkab1083.

7. Radhakrishnan, A., and Green, R. (2016). Connections Underlying Translation and mRNA Stability. J. Mol. Biol. *428*, 3558–3564. https://doi.org/10.1016/j.jmb.2016.05.025.

8. Lucas, M.C., Pryszcz, L.P., Medina, R., Milenkovic, I., Camacho, N., Marchand, V., Motorin, Y., Ribas de Pouplana, L., and Novoa, E.M. (2024). Quantitative analysis of tRNA abundance and modifications by nanopore RNA sequencing. Nat. Biotechnol. *42*, 72–86. https://doi.org/10.1038/s41587-023-01743-6.

9. Suzuki, T. (2021). The expanding world of tRNA modifications and their disease relevance. Nat. Rev. Mol. Cell Biol. *22*, 375–392. https://doi.org/10.1038/s41580-021-00342-0.

10. Roundtree, I.A., Evans, M.E., Pan, T., and He, C. (2017). Dynamic RNA Modifications in Gene Expression Regulation. Cell *169*, 1187–1200. https://doi.org/10.1016/j.cell.2017.05.045.

11. de Crécy-Lagard, V., and Jaroch, M. (2021). Functions of Bacterial tRNA Modifications: From Ubiquity to Diversity. Trends Microbiol. *29*, 41–53. https://doi.org/10.1016/j.tim.2020.06.010.

12. Stephenson, W., Razaghi, R., Busan, S., Weeks, K.M., Timp, W., and Smibert, P. (2022). Direct detection of RNA modifications and structure using single-molecule nanopore sequencing. Cell Genom. *2*, 100097. https://doi.org/10.1016/j.xgen.2022.100097.

13. White, L.K., Dobson, K., del Pozo, S., Bilodeaux, J.M., Andersen, S.E., Baldwin, A., Barrington, C., Körtel, N., Martinez-Seidel, F., Strugar, S.M., et al. (2024). Comparative analysis of 43 distinct RNA modifications by nanopore tRNA sequencing. Preprint at bioRxiv. https://doi.org/10.1101/2024.07.23.604651.

14. Cheng, M.Y., Tao, W.B., Yuan, B.F., and Feng, Y.Q. (2021). Methods for isolation of messenger RNA from biological samples. Anal. Methods *13*, 289–298. https://doi.org/10.1039/d0ay01912g.

15. Rauhut, R., and Klug, G. (1999). mRNA degradation in bacteria. FEMS Microbiol. Rev. *23*, 353–370. https://doi.org/10.1111/j.1574-6976.1999.tb00404.x.

16. Riquelme-Barrios, S., Vasquez-Camus, L., Cusack, S.A., Burdack, K., Petrov, D.P., Yesiltac-Tosun, G.N., Kaiser, S., Giehr, P., and Jung, K. (2025). Direct RNA sequencing of the Escherichia coli epitranscriptome uncovers alterations under heat stress. Nucleic Acids Res. *53*, gkaf175. https://doi.org/10.1093/nar/gkaf175.

17. Tan, L., Guo, Z., Shao, Y., Ye, L., Wang, M., Deng, X., Chen, S., and Li, R. (2024). Analysis of bacterial transcriptome and epitranscriptome using nanopore direct RNA sequencing. Nucleic Acids Res. *52*, 8746–8762. https://doi.org/10.1093/nar/gkae601.

18. Grünberger, F., Ferreira-Cerca, S., and Grohmann, D. (2022). Nanopore sequencing of RNA and cDNA molecules in Escherichia coli. RNA *28*, 400–417. https://doi.org/10.1261/rna.078937.121.

19. Abebe, J.S., Verstraten, R., and Depledge, D.P. (2022). Nanopore-Based Detection of Viral RNA Modifications. mBio *13*, e0370221. https://doi.org/10.1128/mbio.03702-21.

20. Deng, X., Chen, K., Luo, G.Z., Weng, X., Ji, Q., Zhou, T., and He, C. (2015). Widespread occurrence of N6-methyladenosine in bacterial mRNA. Nucleic Acids Res. *43*, 6557–6567. https://doi.org/10.1093/nar/gkv596.

21. Hong, A., Kim, D., Kim, V.N., and Chang, H. (2022). Analyzing viral epitranscriptomes using nanopore direct RNA sequencing. J. Microbiol. *60*, 867–876. https://doi.org/10.1007/s12275-022-2324-4.

22. Meyer, K.D., Saletore, Y., Zumbo, P., Elemento, O., Mason, C.E., and Jaffrey, S.R. (2012). Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons. Cell *149*, 1635–1646. https://doi.org/10.1016/j.cell.2012.05.003.

23. Helm, M., Lyko, F., and Motorin, Y. (2019). Limited antibody specificity compromises epitranscriptomic analyses. Nat. Commun. *10*, 5669. https://doi.org/10.1038/s41467-019-13684-3.

24. Zhang, Z., Chen, T., Chen, H.X., Xie, Y.Y., Chen, L.Q., Zhao, Y.L., Liu, B.D., Jin, L., Zhang, W., Liu, C., et al. (2021). Systematic calibration of epitranscriptomic maps using a synthetic modification-free RNA library. Nat. Methods *18*, 1213–1222. https://doi.org/10.1038/s41592-021-01280-7.

25. Hendra, C., Pratanwanich, P.N., Wan, Y.K., Goh, W.S.S., Thiery, A., and Göke, J. (2022). Detection of m6A from direct RNA sequencing using a multiple instance learning framework. Nat. Methods *19*, 1590–1598. https://doi.org/10.1038/s41592-022-01666-1.

26. Garalde, D.R., Snell, E.A., Jachimowicz, D., Sipos, B., Lloyd, J.H., Bruce, M., Pantic, N., Admassu, T., James, P., Warland, A., et al. (2018). Highly parallel direct RNA sequencing on an array of nanopores. Nat. Methods *15*, 201–206. https://doi.org/10.1038/nmeth.4577.

27. Liu, H., Begik, O., Lucas, M.C., Ramirez, J.M., Mason, C.E., Wiener, D., Schwartz, S., Mattick, J.S., Smith, M.A., and Novoa, E.M. (2019). Accurate detection of m(6)A RNA modifications in native RNA sequences. Nat. Commun. *10*, 4079. https://doi.org/10.1038/s41467-019-11713-9.

28. Gao, Y., Liu, X., Wu, B., Wang, H., Xi, F., Kohnen, M.V., Reddy, A.S.N., and Gu, L. (2021). Quantitative profiling of N6-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. Genome Biol. *22*, 22. https://doi.org/10.1186/s13059-020-02241-7.

29. Lorenz, D.A., Sathe, S., Einstein, J.M., and Yeo, G.W. (2020). Direct RNA sequencing enables m(6)A detection in endogenous transcript isoforms at base-specific resolution. RNA *26*, 19–28. https://doi.org/10.1261/rna.072785.119.

30. Stoiber, M., Quick, J., Egan, R., Eun Lee, J., Celniker, S., Neely, R.K., Loman, N., Pennacchio, L.A., and Brown, J. (2017). De novo Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. Preprint at bioRxiv. https://doi.org/10.1101/094672.

31. Leger, A., Amaral, P.P., Pandolfini, L., Capitanchik, C., Capraro, F., Miano, V., Migliori, V., Toolan-Kerr, P., Sideri, T., Enright, A.J., et al. (2021). RNA modifications detection by comparative Nanopore direct RNA sequencing. Nat. Commun. *12*, 7198. https://doi.org/10.1038/s41467-021-27393-3.

32. Pratanwanich, P.N., Yao, F., Chen, Y., Koh, C.W.Q., Wan, Y.K., Hendra, C., Poon, P., Goh, Y.T., Yap, P.M.L., Chooi, J.Y., et al. (2021). Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. Nat. Biotechnol. *39*, 1394–1402. https://doi.org/10.1038/s41587-021-00949-w.

33. Jenjaroenpun, P., Wongsurawat, T., Wadley, T.D., Wassenaar, T.M., Liu, J., Dai, Q., Wanchai, V., Akel, N.S., Jamshidi-Parsian, A., Franco, A.T., et al. (2021). Decoding the epitranscriptional landscape from native RNA sequences. Nucleic Acids Res. *49*, e7. https://doi.org/10.1093/nar/gkaa620.

34. Parker, M.T., Knop, K., Sherwood, A.V., Schurch, N.J., Mackinnon, K., Gould, P.D., Hall, A.J., Barton, G.J., and Simpson, G.G. (2020). Nanopore direct RNA sequencing maps the complexity of Arabidopsis mRNA processing and m(6)A modification. eLife *9*, e49658. https://doi.org/10.7554/eLife.49658.

35. Abebe, J.S., Price, A.M., Hayer, K.E., Mohr, I., Weitzman, M.D., Wilson, A.C., and Depledge, D.P. (2022). DRUMMER—rapid detection of RNA modifications through comparative nanopore sequencing. Bioinformatics *38*, 3113–3115. https://doi.org/10.1093/bioinformatics/btac274.

36. Zhong, Z.-D., Xie, Y.-Y., Chen, H.-X., Lan, Y.-L., Liu, X.-H., Ji, J.-Y., Wu, F., Jin, L., Chen, J., Mak, D.W., et al. (2023). Systematic comparison of tools used for m6A mapping from nanopore direct RNA sequencing. Nat. Commun. *14*, 1906. https://doi.org/10.1038/s41467-023-37596-5.

37. Mears, M.C., Read, Q.D., and Bakre, A. (2025). Comparison of direct RNA sequencing of Orthoavulavirus javaense using two different chemistries on the MinION platform. J. Virol. Methods *333*, 115103. https://doi.org/10.1016/j.jviromet.2024.115103.

38. Wang, X., Lu, Z., Gomez, A., Hon, G.C., Yue, Y., Han, D., Fu, Y., Parisien, M., Dai, Q., Jia, G., et al. (2014). N6-methyladenosine-dependent regulation of messenger RNA stability. Nature *505*, 117–120. https://doi.org/10.1038/nature12730.

39. Meyer, K.D., Patil, D.P., Zhou, J., Zinoviev, A., Skabkin, M.A., Elemento, O., Pestova, T.V., Qian, S.B., and Jaffrey, S.R. (2015). 5′ UTR m6A Promotes Cap-Independent Translation. Cell *163*, 999–1010. https://doi.org/10.1016/j.cell.2015.10.012.

40. Alarcon, C.R., Lee, H., Goodarzi, H., Halberg, N., and Tavazoie, S.F. (2015). N6-methyladenosine marks primary microRNAs for processing. Nature *519*, 482–485. https://doi.org/10.1038/nature14281.

41. Mendoza, H.G., and Beal, P.A. (2024). Structural and functional effects of inosine modification in mRNA. RNA *30*, 512–520. https://doi.org/10.1261/rna.079977.124.

42. Motorin, Y., Lyko, F., and Helm, M. (2010). 5-methylcytosine in RNA: detection, enzymatic formation and biological functions. Nucleic Acids Res. *38*, 1415–1430. https://doi.org/10.1093/nar/gkp1117.

43. Hewel, C., Hofmann, F., Dietrich, V., Wierczeiko, A., Friedrich, J., Jenson, K., Mündnich, S., Diederich, S., Sys, S., Schartel, L., et al. (2024). Direct RNA sequencing (RNA004) allows for improved transcriptome assessment and near real-time tracking of methylation for medical applications. Preprint at bioRxiv. https://doi.org/10.1101/2024.07.25.605188.

44. Kong, Y., Zhang, Y., Mead, E.A., Chen, H., Loo, C.E., Fan, Y., Ni, M., Zhang, X.-S., Kohli, R.M., and Fang, G. (2024). Critical assessment of nanopore sequencing for the detection of multiple forms of DNA modifications. Preprint at bioRxiv. https://doi.org/10.1101/2024.11.19.624260.

45. Liu-Wei, W., van der Toorn, W., Bohn, P., Hölzer, M., Smyth, R.P., and von Kleist, M. (2024). Sequencing accuracy and systematic errors of nanopore direct RNA sequencing. BMC Genom. *25*, 528. https://doi.org/10.1186/s12864-024-10440-w.

46. Verwilt, J., Mestdagh, P., and Vandesompele, J. (2023). Artifacts and biases of the reverse transcription reaction in RNA sequencing. RNA *29*, 889–897. https://doi.org/10.1261/rna.079623.123.

47. Liu, H., Begik, O., and Novoa, E.M. (2021). EpiNano: Detection of m(6)A RNA Modifications Using Oxford Nanopore Direct RNA Sequencing. Methods Mol. Biol. *2298*, 31–52. https://doi.org/10.1007/978-1-0716-1374-0_3.

48. Fellner, P. (1969). Nucleotide sequences from specific areas of the 16S and 23S ribosomal RNAs of E. coli. Eur. J. Biochem. *11*, 12–27. https://doi.org/10.1111/j.1432-1033.1969.tb00733.x.

49. Tan, L., Guo, Z., Wang, X., Kim, D.Y., and Li, R. (2024). Utilization of nanopore direct RNA sequencing to analyze viral RNA modifications. mSystems *9*, e0116323. https://doi.org/10.1128/msystems.01163-23.

50. Guo, Z., Ni, Y., Tan, L., Shao, Y., Ye, L., Chen, S., and Li, R. (2024). Nanopore Current Events Magnifier (nanoCEM): a novel tool for visualizing current events at modification sites of nanopore sequencing. NAR Genom. Bioinform. *6*, lqae052. https://doi.org/10.1093/nargab/lqae052.

51. Salgado, H., Gama-Castro, S., Lara, P., Mejia-Almonte, C., Alarcón-Carranza, G., López-Almazo, A.G., Betancourt-Figueroa, F., Peña-Loredo, P., Alquicira-Hernández, S., Ledezma-Tejeida, D., et al. (2024). RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. Nucleic Acids Res. *52*, D255–D264. https://doi.org/10.1093/nar/gkad1072.

52. Petrov, D.P., Kaiser, S., Kaiser, S., and Jung, K. (2022). Opportunities and Challenges to Profile mRNA Modifications in Escherichia coli. Chembiochem *23*, e202200270. https://doi.org/10.1002/cbic.202200270.

53. Szydlo, K., Santos, L., Christian, T.W., Maharjan, S., Dorsey, A., Masuda, I., Jia, J., Wu, Y., Tang, W., Hou, Y.-M., and Ignatova, Z. (2025). m6A modification is incorporated into bacterial mRNA without specific functional benefit. Nucleic Acids Res. *53*, gkaf425. https://doi.org/10.1093/nar/gkaf425.

54. Lahens, N.F., Kavakli, I.H., Zhang, R., Hayer, K., Black, M.B., Dueck, H., Pizarro, A., Kim, J., Irizarry, R., Thomas, R.S., et al. (2014). IVT-seq reveals extreme bias in RNA sequencing. Genome Biol. *15*, R86. https://doi.org/10.1186/gb-2014-15-6-r86.

55. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics *34*, 3094–3100. https://doi.org/10.1093/bioinformatics/bty191.

56. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*, giab008. https://doi.org/10.1093/gigascience/giab008.

57. Quinlan, A.R., and Hall, I.M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033.

58. Shen, W., Le, S., Li, Y., and Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. PLoS One *11*, e0163962. https://doi.org/10.1371/journal.pone.0163962.

59. Liu, X., Shao, Y., Guo, Z., Ni, Y., Sun, X., Leung, A.Y.H., and Li, R. (2024). Giraffe: A tool for comprehensive processing and visualization of multiple long-read sequencing data. Comput. Struct. Biotechnol. J. *23*, 3241–3246. https://doi.org/10.1016/j.csbj.2024.08.003.

60. Yu, G. (2024). Thirteen years of clusterProfiler. Innovation *5*, 100722. https://doi.org/10.1016/j.xinn.2024.100722.

61. Bayega, A., Oikonomopoulos, S., Wang, Y.C., and Ragoussis, J. (2022). Improved Nanopore full-length cDNA sequencing by PCR-suppression. Front. Genet. *13*, 1031355. https://doi.org/10.3389/fgene.2022.1031355.

62. Gamaarachchi, H., Samarakoon, H., Jenner, S.P., Ferguson, J.M., Amos, T.G., Hammond, J.M., Saadat, H., Smith, M.A., Parameswaran, S., and Deveson, I.W. (2022). Fast nanopore sequencing data analysis with SLOW5. Nat. Biotechnol. *40*, 1026–1029. https://doi.org/10.1038/s41587-021-01147-4.

63. Gamaarachchi, H., Lam, C.W., Jayatilaka, G., Samarakoon, H., Simpson, J.T., Smith, M.A., and Parameswaran, S. (2020). GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. BMC Bioinf. *21*, 343. https://doi.org/10.1186/s12859-020-03697-x.

64. Samarakoon, H., Liyanage, K., Ferguson, J.M., Parameswaran, S., Gamaarachchi, H., and Deveson, I.W. (2024). Interactive visualization of nanopore sequencing signal data with Squigualiser. Bioinformatics *40*, btae501. https://doi.org/10.1093/bioinformatics/btae501.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial and virus strains		
Escherichia coli strain K-12	American Type Culture Collection, VA, USA	BW25113
Chemicals, peptides, and recombinant proteins		
Invitrogen TRIzol reagent	Thermo Fisher Scientific	Cat#15596018
Maxima H Minus Reverse Transcriptase	Thermo Fisher Scientific	Cat#EP0752
Deoxynucleotide (dNTP) Solution Mix	BioArrow Technology Ltd.	NEB#N0447S
Thermostable RNase H	Thermo Fisher Scientific	NEB#M0523S
Poly(A) polymerase	BioArrow Technology Ltd.	NEB#M0276
Q5 Hot Start High Fidelity Master Mix	BioArrow Technology Ltd.	NEB#M0494
RNaseOut	Thermo Fisher Scientific	Cat# 10777019
SuperScript III Reverse Transcriptase	Thermo Fisher Scientific	Cat#18080044
Critical commercial assays		
RiboMinus™ Transcriptome Isolation Kit, bacteria	Thermo Fisher Scientific	K155004
MEGAscript Kit	Ambion, MA, USA	Ambion#AM1334
SQK-RNA004 Kit	Oxford Nanopore Technology, Oxford, UK	version DRS_9195_v4_revB_20Sep2023
Deposited data		
ss&rd_004 (Basecalled files)	figshare	https://doi.org/10.6084/m9.figshare.28106741.v2
ss&rd_004_rep (Basecalled files)	figshare	https://doi.org/10.6084/m9.figshare.28106909.v1
IVT_neg_004 (Basecalled files)	figshare	https://doi.org/10.6084/m9.figshare.28106906.v1
Raw sequencing data	NCBI: PRJNA1198641	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1198641
Oligonucleotides		
Template switching oligo (TSO) T7 primer: 5′-ACTCTAATACGACTCACTATAG GGAGAGGGCrGrGrG-3′	BGI Genomics, SZ	N/A
T7 extension primer: 5′-GCTCTAAT ACGACTCACTATAGG-3′	BGI Genomics, SZ	N/A
Software and algorithms		
Custom code	This study	https://zenodo.org/doi/10.5281/zenodo.14553666
Dorado v0.8.0	Oxford Nanopore Technology, Oxford, UK	https://github.com/nanoporetech/dorado
Modkit v0.3.0	Oxford Nanopore Technology, Oxford, UK	https://github.com/nanoporetech/modkit
minimap2 v2.17	Li[55]	lh3.github.io/minimap2
SAMtools v1.13	Danecek et al.[56]	https://www.htslib.org/
Epinano v1.2	Liu et al.[27]	https://github.com/novoalab/EpiNano
BEDtools v2.21.0	Quinlan and Hall[57]	https://github.com/arq5x/bedtools2
SeqKit v2.3.0	Shen et al.[58]	https://bioinf.shenwei.me/seqkit/
Giraffe-View v0.2.3	Liu et al.[59]	https://github.com/lrslab/Giraffe_View
compareCluster v4.8.3	Yu[60]	https://rdrr.io/bioc/clusterProfiler/man/compareCluster.html
nanoSundial v0.0.1	This study	https://doi.org/10.5281/zenodo.15904129
Other		
FLO-MIN004RA Flow Cell	Oxford Nanopore Technology, Oxford, UK	FLO-MIN004RA
SPRIselect Beads	Beckman Coulter, IN, USA	Product No: B23317
Agilent TapeStation System	Agilent	RNA ScreenTape 5067-5576

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
AMPure XP beads	Beckman Coulter, IN, USA	Product No: A63881
VAHTS RNA Clean Beads	Vazyme	N412-02

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Bacterial RNA preparation

E. coli strain K-12 was purchased from the American Type Culture Collection, VA, USA. The preparation steps for native RNA from *E. coli* K12 were performed as described in a previously published study 17. Total RNA was extracted from bacterial cultures at an OD600 of 0.4–0.6 using TRIzol reagent. RNA quality was assessed with the Agilent TapeStation System (RNA ScreenTape 5067–5576), ensuring a RIN score >8, confirming high-quality RNA suitable for downstream applications. Then, we purified 20 micrograms (μg) total RNA with a concentration of 200 nanograms per microliter (ng/μL), using 0.8 volume of SPRIselect Beads (Beckman Coulter, IN, USA) to remove small RNA fragments (size select). Subsequently, ribosomal RNA in size-selected RNA was depleted with RiboMinus Transcriptome Isolation Kit, bacteria (Invitrogen). The amount of size-selected RNA for ribosome-depleted was 15 μg per reaction. Around 600 ng size-selected ribosome-depleted RNA was further used for poly(A) tailing using *E. coli* Poly(A) Polymerase (New England Biolabs, NEB#M0276, MA, USA). Following the manufacturer's instructions for poly(A) tailing, we incubated the RNA at 37°C for 30 min, adding around 180-240 adenosines to the RNA primer. The native RNA of preparation was completed.

METHOD DETAILS

In vitro transcribed (IVT) RNA preparation

For the generation of *in vitro* transcribed (IVT) RNA, the poly(A)-tailed RNA underwent reverse transcription to synthesize double-stranded cDNA, which was then used for *in vitro* transcription. Primers used in reverse transcription referred to previously published protocols.[17,24] The following oligonucleotides were purchased from the BGI Genomics, SZ, PRC: template switching oligo (TSO) T7 primer, 5′-ACTCTAATACGACTCACTATAGGGAGAGGGCrGrGrG-3′, where r indicates ribonucleotide bases; T7 extension primer, 5′-GCTCTAATACGACTCACTATAGG-3′. For the reverse transcription conditions, we referred to a previously published literature.[61] In total, 300 ng RNA in 6 μL nuclease-free water was first annealed with 1 μL oligo(dT)$_{23}$VN primer (10 μM, NEB#S1327S) and 1 μL dNTP (10 mM, NEB#N0447S) for 3 min at 72°C, 10 min at 4°C, 1 min at 25°C with the lid temperature set at ≥85°C and then held at 4°C. The reverse transcription (RT) mix was assembled containing 3 μL nuclease-free water, 4.4 μL 5xRT Buffer, 1 μL RNaseOut (Invitrogen), 1 μL TSO T7 primer (75 μM), and 1 μL Maxima H Minus Reverse Transcriptase (Thermo Scientific). After adding the RT mix to the annealed RNA, the reaction mixture was incubated following the SSIV RT protocol.[61] The RNA template was subsequently hydrolyzed by the Thermostable RNase H (NEB#M0523S) according to the manufacturer's instructions. The template-switching cDNA product was purified using the VAHTS RNA Clean Beads. The second-strand cDNA synthesis reaction mixture was assembled on ice, consisting of 20 μL template-switching cDNA, 25 μL Q5 Hot Start High Fidelity Master Mix (NEB#M0494), 3.75 μL T7 extension primer (50 μM), and 1.25 μL nuclease-free water. Following the initial denaturation at 95°C for 1 min and 57°C for 30 s for annealing, the reaction mixture was incubated at 65°C for 10 min. The resulting double-strand DNA (dsDNA) was purified using 1 volume of AMPure XP beads (Beckman Coulter). The *in vitro* transcription step was performed using the MEGAscript Kit (Ambion#AM1334, MA, USA) at 37°C for 4 h. IVT RNA prepared in this article refers specifically to modification-free RNA, so the reaction mixture was composed of 2 μL each of NTPs, 2 μL Reaction Buffer, 120 ng ds-cDNA template in 8 μL nuclease-free water, and 2 μL Enzyme Mix. The IVT RNA was purified with RNA clean beads (Vazyme# N412-02) with a sample-to-bead ratio of 1:1.8.

Nanopore direct RNA sequencing and data processing

Around 500 ng of native RNA and IVT RNA were used for building libraries. The libraries were constructed following the manufacturer's instructions using the SQK-RNA004 Kit (version DRS_9195_v4_revB_20Sep2023, ONT, Oxford, UK). The optional RT step was performed using SuperScript III Reverse Transcriptase (Thermo Fisher Scientific, Cat#18080044). Nanopore direct RNA sequencing (DRS) was conducted on the MinION platform using the FLO-MIN004RA Flow Cell (ONT). The resulting POD5 files were basecalled using the Dorado workflow (v0.8.0) with the model of rna004_130bps_sup@v3.0.1 and rna004_130bps_sup@v5.0.0. The basecalling results were stored as FASTQ files and were statistically analyzed with SeqKit v2.3.0.[58] Raw read features, including read length and Q score, were extracted using Giraffe v0.2.3.[59] Subsequently, reads were aligned to the *E. coli* genome (GenBank Accession Number: NC_000913.3) using minimap2 v2.17 with parameter settings "-ax map-ont".[55] Mapping results (SAM files) were converted into BAM files and sorted and indexed using SAMtools v1.13.[56] Read alignment information was extracted from the sorted BAM files and converted into BED files using a custom Python script based on Pysam (https://github.com/pysam-developers/

pysam). And the accuracy feature in site level is used Epinano_Variants.py from Epinano v1.2.[27] The Phred-scaled mapping accuracy score is calculated as below,

$$Phred\ score = -log\,10(1 - accuracy).$$

This method is employed to convert mapping accuracy into a standardized scale that is consistent with Q scores.

For the modification model, we applied rna004_130bps_sup@v5.1.0, covering three modifications models: m5C, inosine_m6A, Ψ. After the modified basecalling, the output modbam file will be aligned with the reference genome and then sorted by SAMtools. The sorted bam files were converted into modbed files with modkit (0.3.0, https://github.com/nanoporetech/modkit). We employed "modkit pileup" to convert BAM files into BED files by using the default setting. The "Nvalid_cov" value in the BED file was used as the coverage threshold to filter out sites with lower coverage.

Reference selection and annotation strategies

When using the Dorado model, we used the genome as a reference and converted the annotation file (GFF) to a gene.bed (BED) file using a self-customized script. Then, we used Bedtools[57] with the command "bedtools intersect -a <modbed> -b gene.bed -wb -s". Since the GFF file does not include UTR information and contains only the coding sequences (CDS), we will expand the annotation region of the gene by 100 bases upstream and 100 bases downstream in our study required to focus on the UTR regions.

When we optimized nanoSundial for rRNA, we used the rRNA sequences as ref. 23S rRNA (GenBank Accession Number: NR_103073.1), 16S rRNA (GenBank Accession Number: NR_103073.1), and 5S rRNA (GenBank Accession Number: X00414.1). For the tRNA and other analyses, the reference genome was used (GenBank Accession Number: NC_000913.3).

Gene function enrichment

Gene Ontology (GO) pathways of *E.coli* K12 was obtained online (https://ftp.ebi.ac.uk/pub/databases/GO/goa/proteomes/18.E_coli_MG1655.goa). The GO annotations were obtained with GO.db v3.20.0. The GO enrichment analysis was done by compareCluster v4.8.3[60] in R v4.3.1.

QUANTIFICATION AND STATISTICAL ANALYSIS

Gene expression analysis

Gene expression correlations were subsequently calculated based on the mapping results. For DRS data, reads that did not match the strand in annotation were filtered, followed by counting the read numbers aligned to individual genes. Genes other than Protein-coding regiones were removed from the counting results. Undetected genes were also excluded from the analysis. Afterward, counts per million (CPM) were computed for each gene in different samples using the formula, followed by pairwise calculation of Spearman's rank correlation coefficients.

$$CPM_i = \frac{q_i}{\sum_j q_j} * 10^6$$

Where q_i denotes reads mapped to a gene, and $\sum_j q_j$ corresponds to the sum of mapped reads to individual gene. The above process was implemented using our custom script count_read_num_each_gene.py, which is now publicly available in our git and zenodo repository.

Current feature comparison with nanoSundial

NanoSundial is a Python package dedicated to comparative analysis of DRS nanopore sequencing raw signal to *de novo* identify RNA modification sites. As a comparison method, nanoSundial requires two samples, a wild-type sample and a low modification or no modification as negative control sample. The analysis flow is divided into three steps (Figure 3A) after pre-processing: (1) extract current features, (2) parallel processing and statistical testing, (3) Identify the positive sites and merge the positive region.

Pre-processing

To begin with, prepare the input data. Basecalling should first be run to obtain FASTQ files. We have chosen BLOW5 format,[62] which has been proven to be smaller and delivers consistent improvements across different computer architectures. Therefore, raw data needs to be converted to BLOW5 format. After processing the two samples, an appropriate reference needs to be selected, either genome or transcriptome.

1. Extract current features

The signal mapping refinement capabilities of the 004kit dataset is from f5c eventalign,[63] with parameter "–pore 004kit –rna –paf –min-mapq 0". A PAF file recording alignment information and signal index will be generated from f5c. Then, our script traverses the PAF file to extract the raw signal array from BLOW5. A thresholding normalization approach will be implemented, where the raw

signal corresponding to a mapped section of each read is normalized using median shift and MAD (median absolute deviation) scale parameters.

$$NormSignal = \frac{RawSignal - Shift}{Scale}$$

Any normalized signal values exceeding a predetermined threshold (Default: ±5) will be replaced by the threshold value. Then, for each nucleotide in each read, based on the index in the PAF file, we can obtain its window and calculate the mean, median, dwell time, and standard deviation (STD). Additionally, f5c eventalign applied a k-mer model method, there will always be some shift. Therefore, we perform the shifting operation using the shift table defined by Squigualiser.[64]

Finally, since the current changes are very unstable at the beginning and end of the sequencing read, each read will by default skip the first and last ten bases (can be changed by using option "–flip"). The current characteristics corresponding to each remaining base, including mean, median, dwell time, and standard deviation (STD), will be stored on the hard drive. To reduce memory usage and speed up subsequent analysis time, a chunking strategy is used, with the default segment size being 100k bases.

2. Parallel processing and statistical testing

For each position in the reference sequence, the four features of the two samples will be compared by statistical methods. Because statistical methods can yield significantly different results when the data sizes are unequal, the "–balance" option is used to subsample the larger dataset so that the comparison numbers are equivalent.

We employ three methods to do statistical analysis, including multivariate analysis of variance (MANOVA) for all four features, as well as logistic regression and the Kolmogorov–Smirnov (K-S) test specifically for the mean to analyze and return the *p*-values. To enhance processing speed, multiprocessing has been incorporated into our script to handle each position during statistical testing. Subsequently, multiple test corrections using the Benjamini-Hochberg algorithm will be applied to adjust the *p*-values.

Additionally, for each position in the reference, the four features of each sample will have their average values calculated, and then the differences between these average values will be determined. These differences, along with each position's coverage and the *p*-values and adjusted *p*-values, will be saved in the output file.

3. Identify the positive sites and merge the positive region

Through several optimizations on rRNA, appropriate cutoffs were determined as follows: an absolute value of 0.18 for the mean difference, 1 for the dwell time difference, and 3 for -log10(adjusted *p*-value). Our script will output all high-confidence sites using these cutoffs and apply a merging strategy to combine adjacent points. The merging strategy consists of two steps. First, all adjacent high-confidence sites (less than 4 bases apart) are connected. Second, all points, including both those that are connected and those that are not, are extended by 4 bases to form contiguous regions.

Calculation of region intersection

The intersection of regions is a more complex issue compared to the intersection at the site level. This is because, for region sets A and B, the number of elements in A intersecting B is not the same as the number of elements in B intersecting A. To describe the intersection more accurately, we use the smaller set number between $A \cap B$ and $B \cap A$, and the formula is as below,

$$|A \cap B| = \begin{cases} |A \cap B|, |A \cap B| < |B \cap A| \\ |B \cap A|, |B \cap A| < |A \cap B| \end{cases}$$

Evaluation metrics calculation

To assess the performance of the Dorado model and optimize nanoSundial, we calculate the following evaluation metrics: TP, TN, FP, FN, Recall, Precision, F1 Score, and AUC-ROC.

True Positive (**TP**): These are the instances where the model correctly predicts the positive class. True Negative (**TN**): These are the instances where the model correctly predicts the negative class. False Positive (**FP**): These are the instances where the model incorrectly predicts the positive class. This is also known as a "Type I error." False Negative (**FN**): These are the instances where the model incorrectly predicts the negative class. This is also known as a "Type II error."

Recall: Recall, also known as sensitivity, measures the proportion of actual positives that are correctly identified by the model.

$$Recall = \frac{TP}{TP+FN}$$

Precision: Precision, or positive predictive value, measures the proportion of positive predictions that are correct. It is calculated by dividing the number of true positives by the sum of true positives and false positives.

$$Precision = \frac{TP}{TP+FP}$$

F1 Score: The F1 Score is the harmonic mean of Precision and Recall, providing a single metric that balances the two. It is particularly useful when there is an uneven class distribution. It is calculated by multiplying Precision and Recall by two, then dividing the product by the sum of Precision and Recall.

$$F1\ score = 2 * \frac{Precision * Recall}{(Precision+Recall)}$$

AUC-ROC: The Area Under the Receiver Operating Characteristic (AUC-ROC) curve measures the model's ability to distinguish between classes. The ROC curve was constructed using two key performance indicators: the True Positive Rate (TPR) and the False Positive Rate (FPR). TPR, also known as sensitivity, indicates the model's ability to correctly identify positive instances. The FPR represents the proportion of negative instances that are incorrectly classified as positive.

$$TPR = \frac{TP}{TP+FN} \quad FPR = \frac{FP}{TN+FP}$$

The AUC-ROC is the area under this curve, which can be calculated by integrating the ROC curve as a function of the threshold.

$$AUC - ROC = \int_0^1 ROC(t)\, dt$$

And $ROC(t)$ is the ROC curve as a function of the threshold t. In our analysis, t represents the threshold on the absolute difference of 4 features between WT and IVT, specifically mean, median, standard deviation, and dwell time.