Nature Communications

https://doi.org/10.1038/s41467-026-68419-y

Article in Press

Comprehensive mapping of RNA modification dynamics and crosstalk via deep learning and nanopore direct RNA-sequencing

Received: 5 August 2025

Accepted: 7 January 2026

[redacted]

Cite this article as: Dong, H., Gao, Y., Cai, Z. *et al.* Comprehensive mapping of RNA modification dynamics and crosstalk via deep learning and nanopore direct RNA-sequencing. *Nat Commun* (2026). https://doi.org/10.1038/s41467-026-68419-y

Han Dong, Yongsheng Gao, Zhengyi Cai, Yi Li, Xing Li, Fangqing Zhao & Jinyang Zhang

We are providing an unedited version of this manuscript to give early access to its findings. Before final publication, the manuscript will undergo further editing. Please note there may be errors present which affect the content, and all legal disclaimers apply.

If this paper is publishing under a Transparent Peer Review model then Peer Review reports will publish with the final article.

Comprehensive mapping of RNA modification dynamics and crosstalk via deep learning and nanopore direct RNA-sequencing

Han Dong[1,2], Yongsheng Gao[1,2], Zhengyi Cai[1,2], Yi Li[1], Xing Li[1,2*], Fangqing Zhao[1,2,3,*], Jinyang Zhang[1,*]

[1]State Key Laboratory of Animal Biodiversity Conservation and Integrated Pest Management, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, China
[2]University of Chinese Academy of Sciences, Beijing 100049, China
[3]Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China

[*]Corresponding authors: zhangjinyang@ioz.ac.cn (JZ); zhfq@ioz.ac.cn (FZ); li@ioz.ac.cn (XL)

Abstract

Despite the extensive studies of individual RNA modifications, the lack of methods to detect multiple modification types simultaneously has left the global epitranscriptomic landscape and its underlying crosstalk largely unexplored. Here, we present ORCA (Omni-RNA modification Characterization and Annotation), a deep learning framework that enables comprehensive mapping of RNA modification landscape using nanopore direct RNA sequencing. ORCA employs domain adversarial learning to detect and quantify a wide range of modifications by leveraging mixed stoichiometry-driven signal and sequence variability between modified and unmodified nucleotides. It also incorporates a transfer learning module for accurate annotation of modification types with minimal prior knowledge. Applying ORCA to multiple human cell lines reveals widespread, isoform-specific modification patterns, as well as intricate cooperative and competitive interactions among neighboring modification sites. This approach substantially expands the repertoire of known RNA modification sites and elucidates their spatial organization, revealing the emerging roles of RNA modifications in splicing regulation. ORCA thus provides an unbiased and generalizable framework for decoding RNA modification dynamics and their regulatory complexity across diverse biological contexts.

Introduction

RNA modifications represent a complex and dynamic layer of post-transcriptional regulation, with over 170 distinct chemical marks that regulates RNA stability[1], splicing[2, 3], translation[4, 5], and subcellular localization[6]. While the functions of individual modification such as N6-methyladenosine (m6A), pseudouridine (Ψ), and 5-methylcytosine (m5C) have been extensively studied[7, 8], the combinatorial effects and crosstalk among different RNA modifications remain largely unexplored. Recent studies have revealed the coordinated roles for m6A and Ψ in modulating translation[9], and a synergistic co-occurrence of m6A and m5C in plants under salt stress[10], highlighting the emerging role of interaction between different modifications. Deciphering these interactions is crucial for understanding the multilayered regulatory mechanisms governing RNA biogenesis and function. However, progress in this area has been hindered by the lack of transcriptome-wide tools capable of simultaneously detecting and analyzing diverse RNA modifications and their interactions, limiting our ability to decode the full regulatory potential of the epitranscriptome.

Recent Illumina-based approaches using immunoprecipitation[11-13] or chemical treatment[14-17] have enabled transcriptome-wide profiling of individual RNA modification types, but are unable to capture the global epitranscriptomic landscape simultaneously[8]. Nanopore direct RNA sequencing (DRS) overcomes this limitation by directly sequencing native RNA molecules and recording ionic current signals that reflect each nucleotide's chemical structure[18, 19]. These inherent signals produce distinct ionic signal profiles and basecalling differences between modified and unmodified bases, encoding rich information about various RNA modifications within single molecules[20]. However, most existing DRS-based tools are either trained on in vitro synthesized datasets that are restricted to a few well-characterized modifications such as m6A[9, 10, 21-25], m5C[10, 23, 26] and Ψ[9, 10, 20, 27], or rely on comparative analyses to detect condition-specific modification changes[21, 23, 28, 29]. Both strategies are unable to resolve the full spectrum of endogenous RNA modifications or reveal their complex interactions. While a few attempts have been made to simultaneously identify multiple RNA modifications and their associations[9, 10], these models remain constrained by the narrow scope of *in vitro* synthesized modification types, which limits their generalizability to unseen or uncharacterized modifications. Thus, the systematic and unbiased characterization of the full epitranscriptomic landscape and its underlying regulatory crosstalk remains a fundamental challenge.

To address these limitations, we present ORCA (Omni-RNA modification Characterization and Annotation), a deep learning framework for comprehensive profiling of RNA modifications and their interactions at isoform and single-molecule resolution. ORCA employs an adversarial learning strategy to capture both signal- and sequence-level variations arising from the mixed stoichiometry of modified and unmodified nucleotides, thereby overcoming the limited detection scope of current DRS-based approaches. Extensive benchmarking shows that ORCA serves as a powerful tool for unbiased detection and stoichiometric quantification of RNA modifications, even for modification types absent from the training data, demonstrating ORCA's broad generalizability across diverse modification types. Applying ORCA to human cell lines, we expand the known repertoire of RNA modification sites and uncover the widespread interactions among different modifications across transcript isoforms. Notably, ORCA reveal intricate cooperative and competitive relationships between neighboring modification sites, suggesting the complex crosstalk between RNA modifications and splicing regulation. Collectively, ORCA provides a robust and versatile approach for mapping the full spectrum of RNA modifications, revealing the regulatory complexity and isoform-specific crosstalk in the eukaryotic epitranscriptome.

Results

Deep-learning based detection and annotation of various RNA modifications from direct RNA-seq data

To enable generalized detection of diverse RNA modifications from nanopore direct RNA-seq data, we developed a deep-learning framework (ORCA) to systematically identify multiple RNA modification types. Briefly, ORCA first aggregates the raw current signals and basecalled sequences from all reads aligned to a given genomic region, focusing on a 9-nucleotide window centered on each candidate site (Fig.1a and Methods). Since RNA modifications exhibit mixed stoichiometry[17, 30], where not all copies of a given base are modified, modified positions should be characterized by elevated skewness in signal intensity distributions and increased basecalling errors (Supplementary Fig. 1). Thus, ORCA employes these polymorphic features to detect the presence of RNA modifications across the transcriptome. Afterwards, ORCA integrates prior knowledge from established RNA modification databases for effective annotation of a wide range of modification types.

To accurately predict the presence of RNA modifications based on signal- and base-level features, we first constructed a robust and diverse training set comprising six types of RNA modifications from the in vitro synthetic ELIGOS sequences[31] (Fig. 1b). Synthetic transcripts containing one of the six modified bases (m6A, m5C, Ψ, m1A, hm5C, and 5fC) or four canonical bases were randomly sampled and combined to simulate varying stoichiometries and sequencing depths (Methods). To mitigate k-mer bias from the limited sequence diversity of the ELIGOS sequences[10], raw sequence or absolute current levels features were deliberately excluded to ensure that these features could represent generalized modification status without sequence preference. In total, over 7,000,000 sites were generated, with positions containing >10% modified transcripts designated as the positive set[15]. To develop a generalized model capable of accurately predicting diverse RNA modifications without being restricted to specific types, we implemented a domain adversarial learning framework[32] (Fig. 1b). Here, a feature encoder comprising two LSTM layers that process the sequence in opposite directions was trained to capture contextual and sequential features and predict modification presence (modScore) and stoichiometry. Notably, a domain classifier was adversarially trained to minimize the models' ability to discriminate between different modifications using the encoder's output (Methods). This adversarial training strategy forced the encoder to learn generalized features that are shared across modifications, ensuring robust representation of modification status beyond training types while suppressing over-fitting to the modification types used for training.

Considering occurrence of the same RNA modification across different transcriptomic positions often share conserved sequence contexts or signal patterns[15, 33], we implemented a transfer-learning strategy for modification type annotation (Fig. 1c). First, an autoencoder was trained to project all predicted modification sites into a low-dimensional embedding space using both signal- and base-level features, as well as k-mer frequency profiles that capture motif preference of modification sites. These modification sites were then provisionally annotated using public RNA modification databases (RMBase v3.0[34] and DirectRMDB[35]). Subsequently, the model was fine-tuned to predict the type of annotated modification sites, with unannotated sites randomly sampled as negative controls to reject low-confidence predictions and suppress false discoveries. Finally, the classifier's predictions were transferred to all unannotated sites, enabling comprehensive and rigorous identification of unannotated RNA modification sites while minimizing dependence on pre-label training data. This framework ensures that ORCA can achieve stringent modification sites identification and annotation, with inherent flexibility to integrate new modification types with the emerging updates of RNA modification resources[36, 37].

Performance evaluation of RNA modification detection

To evaluate ORCA's performance in detecting various RNA modifications, we first assessed its sensitivity and accuracy using the synthetic ELIGOS dataset[31]. Training datasets for six *in vitro* synthesized modifications were generated as described above, and 5-fold cross-validation was applied to evaluate prediction accuracy across modification types. As shown in Fig. 2a and Supplementary Fig. 2a, ORCA achieved high recall and precision across all six modifications, with an average area under the precision-recall curve (AUPRC) of 0.95 and average area under the receiver operating characteristic curve (AUROC) of 0.94. To quantify the overall performance, we further calculated the F1-score, which balances sensitivity and false discovery rate (FDR) (Fig. 2b). Across all modification types, ORCA consistently attained high F1-scores (0.971-0.976), reflecting its reliable and accurate detection capability. Given that modified sites represent only a small fraction of the transcriptome, we further estimated the false discovery rate of ORCA using an in vitro transcribed (IVT) human mRNA transcriptome[31] devoid of endogenous modifications. As shown in Fig. 2c, ORCA exhibited a low false discovery rate of 2.25% using the default threshold (modScore > 0.9), demonstrating superior false discovery scores compared with multi-modification detection tools and comparable performance relative to several modification-specific methods (Supplementary Fig. 2b). In addition, ORCA demonstrated robust stoichiometry prediction (Supplementary Fig. 2c), showing its ability to accurately quantify diverse RNA modifications using the integrated signal- and base-level features, Together, these results suggested that ORCA provides accurate and unbiased prediction of multiple RNA modification types.

To assess the performance of ORCA in real-world transcriptomes, we benchmarked its ability to detect m6A modifications using DRS data from *Mettl3* knockout (KO) and wild-type (WT) mouse embryonic stem cells[31]. We applied ORCA to predict modification sites in individual samples and analyzed site-specific differential modification levels after *Mettl3* knockout (Methods). As shown in Fig. 2d, *Mettl3*-KO cells exhibited a significant global reduction in RNA modification, with 17.58% of modification sites showed > 0.2 stoichiometric reduction. In contrast, only 5.78% of sites retained increased modification stoichiometries, consistent with *Mettl3*'s role as a primary m6A methyltransferase. For comparison, we evaluated both typical m6A-specific models (CHEUI-solo[23], EpiNano-SVM[31], m6Anet[22] & TandemMod[10]) and comparative-based methods (CHEUI-diff[23], EpiNano-Error[21], Nanocompore[29] & xPore[28]) for detecting differential m6A sites. The m6A sites identified by miCLIP2[12] and GLORI[15] were collected as ground truth benchmarks. To ensure a fair comparison between m6A-specific and comparative-based tools, the performance of m6A prediction was evaluated at both single-base level (only adenosine within DRACH motifs was considered modified) and 5-mer level (all nucleotides within DRACH motifs were treated as modified) respectively[38]. ORCA achieved an AUPRC of 0.42 at single-base level (Supplementary Fig. 2d) and 0.43 at 5-mer level (Fig. 2e), matching the performance of state-of-the-art m6A-specific and comparative-based algorithms. Among these top differentially modified sites (ranked by change of modScore), ORCA exhibited the highest proportion of sites overlapping DRACH motifs or modified 5-mers from m6A sequencing methods (Fig. 2f), confirming its high accuracy in identifying the biologically relevant m6A modifications.

To further validate the versatility of ORCA, we applied it to detect m5C modifications in HeLa cells following *NSUN2* knockout[23], and then evaluated the performance of ORCA and other tools against m5C sites reported by BS-seq[39], bsRNA-seq[40] and RNA-BisSeq[41]. For each tool, changes in predicted stoichiometry were then calculated, and chi-squared test was employed to measure the reduction in m5C sites. Among the m5C sites that consistently detected across two biological replicates, ORCA identified a higher proportion of downregulated sites than most existing tools (Fig. 2g). While CHEUI-diff reported a marginally higher fraction of downregulated sites, ORCA uncovered a significantly larger absolute number of m5C sites with statistically significant downregulation (Fig. 2h and Supplementary Fig. 2e), indicating its strong ability to detect *NSUN2*-dependent m5C alterations. Beyond m5C, we further tested ORCA's ability to detect Ψ modifications in the ribosomal RNA dataset[27]. Compared to established tools (Tombo[42], TandemMod[10] and NanoPSU[27]), ORCA identified comparable high number of validated Ψ sites (73/89) to TandemMod (75/89, Fig. 2i) and also showed strong orthogonal overlap with each approach (Supplementary Fig. 2f), suggesting its accuracy in Ψ detection. Collectively, these results demonstrated ORCA's capacity as a effective framework for detecting a wide range of RNA modifications across different experimental conditions and modification types, including m6A, m5C and Ψ, and supported its utility for accurate and robust transcriptome-wide RNA modification discovery superior to canonical modification-specific and comparative-based tools.

ORCA enables zero-shot detection of unseen RNA modification types

A key limitation of current DRS-based RNA modification detection tools is their reliance on sophisticated training datasets, which typically derived from *in vitro*-synthesized transcripts[10]. However, many modification types are challenging to synthesize in vitro[43], limiting the development of these modification-specific models. To determine whether our strategy is applicable to modifications not included in the training set, we comprehensively benchmarked the performance of ORCA using three complementary approaches: (1) zero-shot prediction of unseen modifications using the ELIGOS dataset, (2) prediction of modifications absent in training set with reference RNA modification sequencing data, and (3) evaluation of ribosomal RNA modifications from mass spectrometry (MS)-based databases (Fig. 3a).

To systematically evaluate the performance of ORCA in detecting unseen RNA modifications, we first assessed the precision and recall in zero-shot prediction using synthetic ELIGOS sequences[31]. For each target modification, we iteratively excluded it from the training set and trained ORCA on different combinations of other modifications to evaluate ORCA's ability to generalize from arbitrary subsets to unseen targets. For zero-shot prediction of target modifications the absent in training set, ORCA achieved high prediction accuracy ~ 90% across all modification types and maintained considerable recall rates > 30% for most modifications (Fig. 3b). Both accuracy and recall increased with the inclusion of additional training modification types (Fig. 3b and Supplementary Fig. 3a), indicating that ORCA can effectively extract generalized cross-modification features. In addition, we further evaluated ORCA's ability to estimate stoichiometry for previously unseen modifications. Strikingly, our adversarial learning framework accurately encapsulated stoichiometry in the zero-shot prediction of all six modification types, achieving a strong linear correlation (average Pearson's correlation coefficient = 0.76) between predicted and ground-truth stoichiometries (Supplementary Fig. 3b). These results confirmed that ORCA can be effectively adapted to detect and quantify unseen RNA modifications without requiring prior training data for these modification types.

To further evaluate ORCA's ability to predict unseen RNA modifications transcriptome-wide, we employed three different high-throughput RNA modification sequencing datasets to validate its applicability. First, we performed ONT direct RNA-seq and m6A-SAC-seq[17] on the same mouse brain sample. To test ORCA's capacity for zero-shot prediction, m6A was excluded from the training set to construct an m6A-absent model, which was then used for transcriptome-wide modification prediction. In total, 1,000 m6A sites detected by m6A-SAC-seq were covered in the nanopore DRS data, of which 70.5% were confidently predicted as modified (modScore > 0.9) by the ORCA model trained without m6A-specific data. Moreover, these identified m6A sites showed significantly higher modification probabilities than randomly sampled DRACH motifs ($p < 1\times 10^{-308}$, Wilcoxon rank-sum test; Fig. 3c), confirming ORCA's capacity to detect transcriptome-wide m6A without prior training on this modification.

Similarly, we further evaluated ORCA's ability to detect 2'-O-methylation (Nm) and inosine (I) modifications which were not included in the training set. For Nm detection, we analyzed public DRS data from mESCs cells[31] and benchmarked ORCA predictions against Nm sites identified by 2'-OMe-seq[44]. As shown in Fig. 3d, ORCA accurately predicted Nm modifications, with 74.4% of reference Nm sites successfully identified and an overall significant enrichment of reference Nm and random control sites was also observed among ORCA's prediction ($p = 1.30\times 10^{-24}$, Wilcoxon rank-sum test). For inosine prediction, we employed public DRS data from wild-type and FY-ADAR2 yeast strains engineered to express human ADAR2, which introduce A-to-I editing in yeast without an endogenous ADAR system[45]. A-to-I editing sites was identified as reference using Illumina RNA-Seq data from the same project (Methods), and differentially modified sites between WT and hADAR2-expressing yeast were identified using ORCA. As expected, reference inosine sites showed significantly higher modification scores in hADAR2 yeast, whereas no such patterns were observed for randomly sampled background adenosines (Fig.

3e). Taken together, these results indicated the ability of ORCA to detect previously unseen RNA modifications at the transcriptome scale, even in the absence of modification-specific training datasets.

Given that ribosomal RNA harbors a diverse array of RNA modifications — many of which are undetectable by either modification-specific or comparative-based tools (Fig. 3f), we employed human and yeast ribosomal RNA sequencing datasets to evaluate ORCA's performance in predicting this broader spectrum of RNA modifications. Across human 28S and 18S rRNAs, ORCA successfully identified 78.1% of the 224 modification sites spanning 13 chemically distinct modification types (e.g. N4-acetylcytidine (ac4C) and m1acp3Ψ) supported by SILNAS mass spectrometry[46]. In parallel, a low false discovery rate (10.0%) was observed for unmodified bases, indicating high specificity in distinguishing modified from unmodified sites. Similarly, ORCA achieved comparable performance on yeast 18S and 25S rRNAs, accurately predicting 79.8% of orthogonally validated modifications. In additional, further supporting its robustness in identifying a wide range of RNA modification types across species (Supplementary Fig. 3c-d).

Furthermore, we evaluated ORCA's ability to detect non-natural 4-thiouridine (4sU) modifications. Applying ORCA to K562 4sU pulldown and DMSO control samples[47], ORCA detected significantly elevated modification scores for U-containg 5-mers in 4sU pulldown samples compared with non U-containing controls ($p = 1.38 \times 10^{-20}$, one-sided Wilcoxon rank-sum test; Supplementary Fig. 4a). The predicted 4sU levels also strongly correlated with nascent RNA expression measured by orthogonal Illumina sequencing ($p = 1.67 \times 10^{-16}$, Supplementary Fig. 4b), demonstrating ORCA's sensitivity to 4sU incorporation. Taken together, these results demonstrate that ORCA can overcome the limitations of modification-specific models, enabling transcriptome-wide discovery of unseen RNA modifications through its generalized adversarial training strategy.

ORCA uncovers the transcriptome-wide landscape of a broad range of modifications

The comprehensive identification of diverse RNA modifications is essential for understanding the post-transcriptional processing of RNAs. However, current DRS-based tools typically rely on modification-specific models that target a limited subset of modifications or utilize comparative strategies to detect changes between different condition pairs. These limitations make it challenging to profile a broad spectrum of RNA modifications in individual samples. To further demonstrate ORCA's ability to simultaneously detect multiple RNA modifications per sample, we applied it to the previously described *Mettl3*-KO mESCs dataset. All predicted modification sites were ranked by modScore and compared against known modification sites in the RMBase 3.0 database[34], as well as sites predicted by modification-specific tools. Among the top-ranked predictions, over 40% were supported by either public database or ONT-based RNA modification detection tools (Fig. 4a), indicating the high reliability of ORCA's prediction. Specifically, a comprehensive catalog of well-characterized modifications including m6A, m5C, inosine (I), Ψ, m7G, m1A and 2′-O-methylation (Nm) were detected within the top 10,000 predicted sites (Fig. 4b). For example, 41.9% and 30.9% of predicted sites in mESCs and HeLa cells were supported by m6A sites, while 16.1% and 5.2% of sites were supported by Ψ modifications. This high concordance underscored ORCA's ability to accurately resolve multiple RNA modifications in a single analysis.

For instance, ORCA identified adjacent modification sites in the 3' UTR of the *Lars2* transcript, where one Nm, two m5C and five Ψ modification sites were consistently detected in both WT and *Mettl3* KO mESCs. Notably, a previously unannotated site exhibited a dramatic reduction in modScore upon *Mettl3* knockout, suggesting it may represent an unseen m6A site (Fig. 4c). Similarly, two m6A sites in the 3' UTR of *Ets2* transcript were robustly detected in WT mESCs but were absent in *Mettl3* KO samples, consistent with substantial alterations in both sequence- and signal-level features following loss of the key m6A methyltransferase complex (Fig. 4d). Moreover, these predicted sites were also consistently supported by both modification-specific and comparative-based algorithms. We further extended this analysis to WT and *NSUN2* KO HeLa cells. For exemplary modification sites in the 3' UTR of RPL13A and in the small nuclear RNA RNA5-8SN1, ORCA revealed a selective reduction in m5C, but not in m6A or Ψ levels, consistent with the specificity of *NSUN2* as an m5C methyltransferase[41] (Supplementary Fig. 5). Taken together, these results demonstrated that ORCA enables the simultaneous detection of diverse RNA modifications while accurately resolves biologically relevant stoichiometry changes upon perturbation of specific modification writers.

Transfer learning of sequence and signal features enables accurate discovery of previously unannotated RNA modification sites.

Based on accurate modification presence prediction, we further developed a transfer learning framework to annotate modification types using both sequence- and signal-level features of high-confidence sites curated from public databases (Fig. 5a and Methods). In addition to the features used for modification presence prediction, we incorporated k-mer occurrence frequencies to capture sequence similarity specific to each modification type. Given that chemical modifications affect current signals across a 5-6 nucleotide window as the strand passes through the nanopore[48], a multi-task learning model was implemented to simultaneously predict both modification types and positional phase. The prediction results were then filtered based on the consistency between predicted modification types and the corrected nucleotide phases, and only predictions where the modification types matched the corresponding nucleotide position were retained to ensure accurate modification assignment. To avoid over-assignment of uncharacterized modifications to known categories, unannotated sites were also sampled as negative controls during training to ensure stringent and reliable identification of known modifications. Finally, the trained model was transferred to predict modification types for previously unannotated sites, enabling the discovery of RNA modification sites absent from existing databases. Overall, this approach balances the accuracy of modification type prediction with the sensitivity to uncover epitranscriptomic features.

To rigorously evaluate the performance of RNA modification annotation, ORCA was employed to present modification presence using a K562 DRS dataset from the SGNex project[49]. A total of 48,377 sites across 7 RNA modification types were then annotated with high confidence NGS-supported sites from RMBase 3.0[34] and DirectRMDB[35], and the modification annotation model was trained as described previously (Fig. 5a). In cross-validation, ORCA achieved > 90% precision and > 40% recall for most modification types (Fig. 5c-d), demonstrating robust accuracy and sensitivity. To further assess the specificity of modification type prediction, we iteratively masked each modification type during training and quantified the misclassification of masked sites into other categories. As shown in Fig. 5e and Supplementary Fig. 6a, ORCA maintained an average of > 0.83 accuracy across all modification types. Moreover, ablation of the background negative-control class markedly increased false discovery (Supplementary Fig. 7), confirming that our strategy ensure accurate and stringent modification type prediction while minimize false positive assignments of unannotated modifications.

The trained model was subsequently applied to predict modification types across all unannotated modification sites. In total, 42,449 previously unannotated RNA modification sites were identified, whose gene body distributions closely consistent with those in the established databases (Fig. 5f). Notably, ORCA largely expanded the catalog of current modification sites, annotating 29% additional m6A sites and dramatically increasing the low-abundance modifications: > 400% more m5C, Nm, Ψ and inosine, > 1,030% more m7G, and > 178% more m1A sites compared to existing annotations (Fig. 5g**)**. To validate these predictions, we then performed de novo motif analysis using XSTREME[50] on previously unannotated m5C sites. Two canonical m5C motifs CUCC (88.3% of ORCA-annotated sites) and CGGG (8.8% of ORCA-annotated sites) were identified, aligning with known *NSUN6*[51] and *NSUN2*-dependent[39] m5C sites. For example, ORCA predicted an m5C site in the *CDT1* 3' UTR that was absent in both RMBase 3.0 and DirectRMDB, but was independently validated by UBS-seq[16] in HeLa cells (Fig.5i). Besides, ORCA also demonstrated high specificity for annotating m6A sites, with 26.9% sites supported by GLORI[15], which was consistent with the validation rate of database-curated m6A sites (37.6%, Supplementary Fig. 6b). For other modifications, high validation rates were also observed for Ψ (BID-seq[14]), m7G (20% overlap with QKI CLIP-seq[11] peaks), and m1A (18.5% by m1A-seq[13]) (Supplementary Fig. 6c-e). Taken together, these results demonstrated the effectiveness of ORCA's label-transfer learning strategy in discovering and annotating RNA modification sites that were previously unannotated in existing databases with high confidence.

Furthermore, we evaluated whether database composition introduces biases associated with modification detection technologies or cell line origins. Although cross-technology overlap for the four major modification types was generally limited (Supplementary Fig. 8a-b), ORCA's annotation model trained on individual assays achieved high recall within orthogonal datasets, and performance further improved when complementary assays were incorprotatd while maintaining low FDR (Supplementary Fig. 8c-d). Consistent results were observed in cross cell-line validation, where restricting training to a single cell linereduced annotation sensitivity but did not affect precision or FDR (Supplementary Fig. 9). Together, these results demonstrate that limited overlap across technologies or cell lines does not compromise annotation accuracy, and that integrating multiple public resources effectively mitigates technology-specific biases and improves sensitivity without sacrificing precision.

Characterization of RNA modification landscape and its regulatory crosstalk in human cell lines

To demonstrate the applicability of our method, we applied ORCA to characterize the RNA modification landscape in human cell lines DRS data from the SGNex project[49]. In summary, a total of 98,586 sites were detected across all samples, with 10,954 modification sites per cell line. Notably, 70.2% of these sites were consistently detected in at least two cell lines (Fig. 6a), which is consistent with the reported stable m6A modifications shared across human cell lines[28]. To further investigate the spatial associations between different modification types, we further calculated the genomic distances between adjacent modifications. Strikingly, a substantial proportion of modifications (33%) occurred within 20-nt of each other (Supplementary Fig. 10a). We therefore clustered proximal sites using a 20-nt window, yielding 13,633 clusters with an average of 2.85 modification per cluster (Fig. 6b and Fig. 6c). As m6A was the most abundantly detected modification, most clusters were m6A-enriched, while a high degree of association between m6A and other modifications, such as m5C and m1A, was also observed, highlighting the complex spatial organization and potential crosstalk among neighboring modification types (Fig. 6d).

To further investigate the regulatory interplay between these neighboring RNA modifications, we applied an expectation-maximization (EM)-based model to estimate the single-molecule co-occurrence patterns within modification clusters (Fig. 6e and Methods). Among 443,361 modification clusters, 7,719 exhibited significant co-modification, while 39,906 were exclusively modified with competitive exclusion (Supplementary Fig. 6b). First, cooperative modified clusters were prioritized for downstream analysis. As shown in Fig. 6f, frequent co-occurrence of different modification was substantially observed, with m5C and m6A emerging as the most prevalent combinatorial pattern. To further validate these predictions, we leveraged m6A-SAC-seq and m5C-TAC-seq[52] datasets to extract short-read level co-modification evidence. For instance, a cooperative modification of a pair of m6A sites spaced 8 nucleotides apart in the 3' UTR of *DNAJB1* was identified in IM95 DRS data and independently confirmed in the HeLa m6A-SAC-seq data (Fig. 6g). Similarly, two co-occurring adjacent m5C sites in the 3'UTR of HDGF were detected in Hct116 DRS data, which was also validated by HeLa m5C-TAC-seq (Supplementary Fig. 6c). These results demonstrated the cooperative modification between different types and also validated accuracy of ORCA in resolving spatial co-existence of neighboring RNA modifications at single-molecule resolution.

To investigate the interplay between RNA modifications and splicing regulation, we then focused on exclusively modified clusters that exhibited significant isoform-specific changes in K562 cells, and integrated ENCORE eCLIP-seq data[53] of K562 cells to assess regulatory associations. The binding patterns of RNA-binding proteins, including splicing factors and RNA modification associated proteins (writers, erasers, and readers, WERs) within these modification clusters were further calculated. Notably, many isoform-specific modification clusters, particularly those associated with m6A, showed significant enrichment of splicing factors and modification-associated WERs, suggesting the widespread coupling between m6A modifications and alternative splicing events (Fig. 6h). For example, two splicing factors ELAVL1[54] and FMR1[55] were significantly enriched in isoform-specific m6A/m5C modification clusters, consistent with the FMR1's preference for binding m6A-modified RNAs[56, 57] (Fig. 6i-j). In one case, transcript-level analysis of RBIS revealed an isoform-specific exclusion pattern between neighboring m6A and m5C sites at exon 4 (Fig. 6k). m6A-modified reads were strongly associated with an upstream skipped exon, whereas the exon was consistently associated retained in m5C-modified reads. Furthermore, strong eCLIP-seq peaks for splicing factors MBNL1 and the RNA-binding protein FMR1 were also detected in the same region, consistent with their established roles in alternative splicing and m6A-mediated splicing regulation[56]. Together, these findings demonstrate that ORCA enables systematic characterization of the interactions between RNA modifications and splicing, offering a powerful platform for dissecting the multilayered regulation of eukaryotic transcriptome.

Discussion

In this study, we present a comprehensive computational framework for mapping global RNA modification landscape and regulatory crosstalk using nanopore direct RNA sequencing data. ORCA employs deep-learning algorithms for unbiased and generalized detection of RNA modification presence and enables accurate modification-type annotation by incorporating prior knowledge of validated sites. Comprehensive evaluations demonstrated that ORCA reliably detects and quantifies previously uncharacterized modification sites and revealed its applicability in uncovering complex interactions between neighboring modifications and isoform-specific RNA modification regulation.

Comprehensive detection of the full spectrum of RNA modification is essential for understanding their roles in RNA biology and epitranscriptomic regulation[58, 59]. However, current high-throughput sequencing-based approaches rely on modification-specific antibodies or chemical reactivity, substantially limiting their generalizability[11-17]. Despite rapid advancements in DRS-based algorithms for modification detection or modification-aware basecalling, existing methods remain constrained by their dependence on modification-specific training sets[9, 10, 21-27, 31]. Meanwhile, comparative profiling of nanopore direct RNA-seq data across different experimental conditions have also enabled identification of RNA modification changes without modification type limitatiion[21, 28, 29], but also overlook the unperturbed modifications thus restricting the analyses to condition-specific modification sites. Furthermore, modification types could only be inferred from experimental setup, which risk bias due to complex interaction between modification[60].

To address this limitation, ORCA leverages the mixed stoichiometry nature of RNA modifications and detects their presence based on variability in signal- and sequence-level features arising from the co-existence of modified and unmodified bases. Specifically, an adversarial learning strategy is employed to ensure unbiased detection of diverse modification by preventing modification-specific overfitting. Through comprehensive evaluation, we demonstrated that ORCA enables accurate zero-shot detection and quantification of various RNA modifications without requiring a corresponding training dataset. highlighting its broad applicability for profiling the transcriptome-wide RNA modification landscape. Furthermore, a transfer-learning based annotation assigns modification types by aligning signal- and sequence-level features of identified sites with prior knowledge from validated databases, enabling accurate co-profiling of multiple modification types. As metabolic labeling and chemical-based sequencing techniques continue to evolve, ORCA can be further extended to incorporate these reference sites, facilitating transcriptome-wide characterization of emerging RNA modifications without requiring extensive synthesis of in vitro transcription experiments.

Nanopore-based full-length RNA sequencing approaches have been widely applied to resolve transcript isoform landscape across diver RNA classes[61-63]. Beyond transcriptome-wide characterization of RNA modification sites, single-molecule RNA modification identification is critical for uncovering the underlying regulatory crosstalk between different modifications[10, 22, 23]. ORCA incorporates an expectation-maximization (EM)-based model to infer single-molecule modification states and assess competitive or cooperative interactions among neighboring modification sites. Applied to human cell lines, ORCA substantially expanded the known catalog of RNA modification sites, increasing the number of both well-characterized m6A and other low-abundance modifications. Notably, ORCA revealed the widespread interplay between different RNA modifications and uncovered the potential regulatory crosstalk between splicing factors and modification-associated RNA-binding proteins in shaping isoform-specific modification patterns. These findings highlight ORCA as a powerful platform for dissecting the complex regulatory architecture of the RNA epitranscriptome at isoform and single molecule resolution.

Recent studies have employed deep-learning model for detecting multiple RNA modifications. In particular, TandemMod employs deep-learning models to identify multiple RNA modifications (including m6A, m1A and m5C) at the single-read level, and further incorporate transfer learning to predict additional modification such as m7G, hm5C and Ψ using limited training examples[10]. While this approach enables simultaneous detection of multiple RNA modification types, it still relies on IVT-derived training sites, which restricts its ability to capture the full RNA modification landscape. In contrast, ORCA leverages a domain-adversarial learning strategy to infer modification presence based on signal polymorphism, enabling the detection of a wide range of RNA modification types without requiring corresponding IVT training sets. However, this approach requires sufficient read depth and is less effective at low-coverage sites (<10 reads). Further work integrating both strategies may enable robust de novo modification detection at single molecule level.

In addition, recent advances in RNA-004 chemistry have largely improved ionic signal quality and basecalling accuracy[64]. To assess ORCA's compatibility with the new sequencing chemistry, we trained an RNA004-specific model using the IVT curlcake dataset[65] (Supplementary Fig. 11a). Compared with Dorado, ORCA achieved similarly high performance for the three basecallable modification types, while also maintaining high accuracy on the remaining four modification types that Dorado could not detect (Supplementary Fig. 11b-c). Furthermore, ORCA exhibited reliable zero-shot prediction performance for these modification, consistent with the results obtained on RNA002 datasets (Supplementary Fig. 11d).We additionally generated a mouse brain RNA004 dataset and compared de novo m6A predictions between RNA002 and RNA004 using corresponding non-m6A models. ORCA produced highly concordant m6A signals across chemistries (Supplementary Fig. 11e-f), demonstrating stable and robust de novo detection under RNA004 chemistry (Supplementary Table 1).

Despite these advantages, ORCA also faces several limitations. First, ORCA requires sufficient sequencing depth to robustly estimate modification-induced feature variability. Although its performance becomes largely insensitive to coverage beyond a certain threshold (Supplementary Fig. 12), reliable detection remains challenging at very low read depths (<10 reads) or when attempting single-read inference. In additional, the ELIGOS training dataset exhibited limited 9-mer diversity, which might introduce sequence composition biases and affect generalization. Cross-dataset evaluation using the in vitro transcribed epitranscriptome (IVET[10]) revealed that IVET-derived model achieved superior better cross-dataset prediction performance (Supplementary Fig. 13), indicating that greater sequence diversity in the training set improves ORCA's prediction performance across diverse sequence contexts. Finally, each ELIGOS read contains only a single modification type, resulting in that no two modification types co-occur within the same read in the training dataset, which could impact the model's performance to predict co-occurring modifications in very close proximity.

In summary, ORCA comprehensively captures the full RNA modification spectrum and reveals the widespread crosstalk between different modifications and splicing regulation. This framework enables unbiased profiling of RNA modifications without requiring extensive IVT training data, providing robust identification of various RNA within individual samples and detection of biologically relevant changes across experimental conditions. By facilitating simultaneous identification, quantification and annotation of diverse RNA modifications at isoform and single-molecule resolution, ORCA uncovers the cooperative modification patterns among neighboring modification sites and highlights the potential regulatory role of adjacent RNA modifications and RBPs in isoform-specific splicing and modification dynamics. Overall, ORCA provides a powerful computational strategy towards the comprehensive elucidation of the RNAome, offering a foundation for understanding RNA biology at unprecedented resolution.

Methods

Ethics statement

All experimental procedures involving animals in this study were carried out in accordance with the guidelines for procurement and use of laboratory animals and have been approved by the Institutional Animal Ethics Committee at the Institute of Zoology, Chinese Academy of Sciences.

Animal experiments

All mice used in this study were adult C57BL/6 mice and were purchased from SiPeiFu Biotechnology. Two adult mice were used for brain tissue dissection, with one mouse used for RNA002 sequencing and the other for RNA004 sequencing. Animals were maintained under conventional specific pathogen-free conditions and a 12-h light/12-h dark cycle at 25 ° C and 40–60% humidity.

RNA isolation and nanopore direct RNA sequencing

Total RNA was extracted from two healthy adult mice brain using TRIzol (Invitrogen) according to the manufacturer's instructions. RNA integrity and quality were assessed using the Agilent 5200 Fragment Analyzer System. Nanopore direct RNA-seq library was prepared using the Direct RNA Sequencing Kit (SQK-RNA002) from Oxford Nanopore Technologies following the manufacturer's protocol and sequenced on an R9.4.1 flow cell (FLO-MIN106D) using a MinION Mk1B device for 72 hours. An adult mouse brain direct RNA-seq library was also generated using the SQK-RNA004 sequencing kit and sequenced according to the manufacturer's instructions on an FLO-MIN004RA flow cell for 72 hours.

m6A-SAC-seq and data analysis

For m6A-SAC-seq experiments, 1 μg of total RNA was subjected to ribosomal RNA depletion using the RiboErase kit (human/mouse/rat, Kapa Biosystems). The rRNA-depleted total RNA was used directly for m6A-SAC-seq library preparation following the protocol described by He et al[17]. Briefly, m6A modifications were selectively converted into allyl-labeled derivatives by *Mj*Dim1, followed by iodine-induced intramolecular cyclization. These modifications were subsequently converted into sequence mutations during by HIV-1 RT reverse transcription and detected via Illumina sequencing.

Sequencing reads were trimmed using Cutadapt[66] (v2.10) and Fastp[67] (v0.23.4) Reads aligning to rRNA sequences were removed using Bowtie2[68] (v2.3.4.3) and Samtools[69] (v1.18). Cleaned reads were then mapped to the mm10 reference genome using STAR[70] (v2.7.10b). PCR duplicates were collapsed using UMICollapse[71] (v1.0.0), and deduplicated BAM files from biological replicates were merged with Samtools. Strand-specific BAM files were generated and processed with Samtools mpileup. Somatic variants were called using VarScan[72] (2.3.9), and candidate m6A sites were identified based on mutation profiles and the presence of DRACH motifs.

Nanopore data preprocessing and feature extraction

The GRCh38 (human) and GRCm38 (mouse) reference transcriptome were obtained from the Ensembl database. For RNA002 data, raw nanopore fast5 files were basecalled using Guppy (v6.3.8) with rna_r9.4.1_70bps_hac model. For RNA004 data, raw nanopore pod5 files were basecalled using Dorado (v1.1.1+e72f1492) with rna004_130bps_hac@v5.2.0 model. POD5 files was transferred to BLOW5 format using bluecrab[73] (v0.4.0) p2s and slow5tools[74] (v0.8.0) merge commands. Basecalled reads were then aligned to the reference transcriptome using Minimap2[75] (v2.21, with the parameters '-ax splice -N 0 -uf -k14 --cs --secondary=no'). The alignment results were processed with samtools mpileup (v1.11) to generate per-base summary statistics. Ionic current signals were aligned to the reference sequence using the f5c[76] eventalign (v1.11), an accelerated implementation of Nanopolish[77], with the parameters '--min-mapq 0 --rna --signal-index --scale-events --secondary=no --collapse-events' for RNA002 and '--pore RNA004 --min-mapq 0 --rna --signal-index --scale-events --secondary=no --collapse-events' for RNA004 reads.

For modification presence prediction, both signal-level and sequence-level features were extracted within a ±2 k-mer window surrounding each candidate site. For signal-level features, raw electrical events from the eventalign output were standardized using the method defined in Nanopolish. Specifically, each event's mean signal level was normalized by subtracting the expected reference mean and then dividing by the reference standard deviation. This normalization accounts for variation in signal intensity across different sequence contexts. The standardized signal values from all reads aligned to the same genomic position were then aggregated and interpolated into a fixed-length vector of 50 values to ensure consistent input dimensions for the model. For sequence-level features, rate of insertions, deletions, and mismatches, as well as statistical metrics including the mean, median, and standard deviation of sequence quality scores of all aligned at each position were extracted based on the 'samtools mpileup' result.

For modification type annotation, three categories of features were included as input for the transfer-learning model. K-mer occurrence features were derived from the frequency of all 256 possible 4-mer motifs within an 11-nucleotide window centered on each modification site. Besides, signal-level features were computed using the event level means and standard deviations across the window surrounding each modification site. Then a Gaussian mixture model was applied to partition each feature into divide into modified and unmodified clusters, and the mean, variance, and covariance of each component were extracted as model input. Finally, sequence-level features were obtained using the same strategy as described above.

ORCA model design

ORCA comprises two neural network models designed for predicting the presence of RNA modifications and inferring their types. For modification presence prediction, ORCA adopts a domain-adversarial neural network architecture composed of an encoder and two classifier branches. The encoder utilizes a bidirectional LSTM network to capture contextual and sequential dependencies from both sequence- and signal-level features within a 11-nucleotide window surrounding each candidate site. The encoded representations are then simultaneously passed into two parallel branches: (1) a modification predictor for predicting modification presence (modScore) along with an estimate of stoichiometry; and (2) a domain classifier with a gradient reversal layer aims to distinguish between different RNA modification types. The model is trained adversarially to optimize encoder and modification predictor to accurately detect modification presence, while the encoder is simultaneously trained to learn representations that minimize the performance of domain classifier, ensuring a generalizable representation of modification presence across a diverse range of modification types without introducing modification-specific bias.

For modification type inference, ORCA employs a transfer-learning framework consisting of an autoencoder and two classifier modules. During the pretraining phase, the autoencoder learns the global low-dimensional representation of all predicted modification sites. The encoded features are subsequently passed to multi-task prediction to produce probabilities corresponding to different modification types and phase represent the exact modification position in the input window. Then, only modification type predictions that corresponding to the base at the inferred modification position are retained as valid outputs.

Model training

To train the presence prediction model, we generated a labeled dataset based on the public ELIGOS resource. Specifically, six in vitro transcribed RNA samples containing individual modifications (m1A, m6A, m5C, hm5C, 5fC, and Ψ) and a control sample composed of only canonical bases were obtained. For each modification type, raw nanopore reads were aligned to the reference sequence using Minimap2. Then, modified and unmodified reads were randomly sampled and combined to generate 3,000 training samples at varying sequencing depths (10-200) and modification rates (0-1) of individual modification site. The sequence- and signal-level features were extracted as described above, resulting in a training set that simulate the realistic features of RNA modification types and stoichiometry levels. In total, over 7 million training samples were generated. Positions with >10% modified reads were labeled as positive samples. For each modification type, the dataset was randomly split into training and testing set at 4:1 ratio. Model optimization was conducted using the AdamW optimizer with a learning rate of 0.0005. The loss functions were defined as follows: negative log-likelihood loss (NLLLoss) for modification presence prediction, NLLLoss for domain label prediction, and mean squared error (MSELoss) for stoichiometry regression.

To train the modification type annotation model, we utilized 25 direct RNA-seq datasets generated from the MinION/GridION platform, spanning nine human cell lines from the SGNex project[49]. RNA modifications were first identified using the ORCA prediction module with a modScore threshold of 0.9 to ensure high-confidence predictions. Predicted modification sites were then annotated based on the presence of NGS-supported modifications from RMBase v3.0[34] and DirectRMDB[35] using a distance threshold of ±2 nucleotides. The autoencoder was initially trained on all predicted modification sites using the AdamW optimizer with a learning rate of 0.002 and MSELoss to learn global feature representations. The model was then fine-tuned using the annotated subset to enable prediction of both modification type and phase. These downstream tasks employed cross-entropy loss and were optimized with AdamW at the same learning rate. For each modification type, the labeled dataset was split into training and testing set in an 1:4 ratio. To mitigate class imbalance, each mini-batch was constructed to contain similar numbers of samples from each modification type, preventing the loss function from being dominated by any single class. To avoid over-assignment of uncharacterized modifications to known categories, we also included 3,000 unannotated sites as negative controls during training to ensure a stringent and reliable identification of known modification types.

EM based prediction of modification interactions

ORCA outputs read-level modification predictions for neighboring modification pairs through the EM-based model, which enables assessment of read-level linkage or mutual exclusivity between modification events. After performing RNA modification presence and type prediction and across all human cell line samples, only high-confidence sites detected in at least two samples were retained for downstream analysis. To define local modification clusters, transcriptomic distances between neighboring modification sites were calculated, and sites within 20 nucleotides were iteratively merged into the same cluster. To minimize potential signal interference caused by adjacent modifications, any neighboring sites located within 4 nucleotides of each other were excluded from clustering.

To evaluate potential co-modification between modification sites within each cluster, we employed a Local Outlier Factor (LOF) score-based strategy to infer the read-level modification states. For each site, raw signal features including event level mean, standard deviation, and dwell time were extracted from a ±5 k-mer (15 nucleotide) window. LOF scores were computed for each read and normalized to the range [0, 1]. For every pair of neighboring modification sites, each read was represented as a two-dimensional coordinate (x_i, y_i), where x_i and y_i are the normalized LOF scores at the two positions, respectively.

Given the site-specific distributions of LOF scores, we employed an expectation-maximization (EM) algorithm to classify each read into one of four canonical modification states including: d_0 centered at (0,0) for unmodified reads; d_1 centered at (1,1) for dual modification; and d_2 and d_3 centered at (1,0) and (0,1) respectively for single-site modifications. Each state d_j was assigned an initial mixing weight $\theta_j = 0.25$. In the expectation step, the posterior probability that read r_i (x_i, y_i) belongs to state d_j was calculated as:

$$\gamma_{ij} = \frac{\theta_j \cdot P(r_i|d_j)}{\sum_{k=0}^{3} \theta_k \cdot P(r_i|d_k)} \tag{1}$$

where the likelihood $P(r_i|d_j) = Euclidean(r_i, d_j) \times s(r_i, d_j)$ was defined as the product of the Euclidean distance between r_i and the center of d_j, and a dispersion score $s(r_i, d_j)$, such that:

$$\begin{aligned} s(r_i, d_0) = s(r_i, d_1\} = 1 - |x_i - y_i| \\ s(r_i, d_2) = s(r_i, d_3\} = 1 - |x_i + y_i - 1| \end{aligned} \tag{2}$$

This formulation captures both the geometric proximity of the read to a canonical modification state and the consistency of LOF across the two sites.

In the maximization step, the mixing weights were updated as the mean of the posterior probability across all N reads:

$$\theta_j = \frac{1}{N} \sum_{i=1}^{N} \gamma_{ij} \tag{3}$$

The EM process was iteratively repeated until convergence, defined by the change in posterior weights falling below a predefined threshold. After convergence, each read was assigned to the modification state with the highest posterior probability γ_{ij}, enabling stratified analysis of co-modified, mutually exclusive, and unmodified read populations within each cluster.

To further quantify the interaction between each pair of modification sites, we defined a linkage score for each candidate pair as:

$$Linkage\ score = (\theta_1 - \theta_2) + (\theta_1 - \theta_3) \tag{4}$$

A modification sites was considered cooperatively modified if $\theta_1 > \max(\theta_2, \theta_3)$, indicating an enrichment of simultaneously modified reads. Conversely, modification site pairs were considered as mutually exclusive if the linkage score was less than -0.2 and $\theta_1 < \min(\theta_2, \theta_3)$, suggesting that the two modifications tend not to co-occur in the same read.

False positive rate evaluation

To evaluate the false discovery rate of ORCA, the IVT human mRNA transcriptome were downloaded from the SRA database (accession number SRP166020). ORCA was employed to predict the modification presence in the IVT transcriptome. Among 7,576,597 sites with coverage great than 10, where only 2.25% (170,708) of sites were predicted as modified using the default modScore threshold of 0.9. TandemMod, EpiNano-SVM, NanoPSU and m6ANet were also applied to the same IVT transcriptome. For each tool, the cumulative distribution function of prediction scores were calculated, and false discovery performance was measured using the area under the cumulative distribution curve.

Model benchmarking

To evaluate ORCA's performance, we benchmarked it against several representative tools for direct RNA modification detection using publicly available datasets. For m6A prediction, we included m6Anet[22] (v2.0.1), TandemMod[10] (v1.1.0), EpiNano[21] (v1.2), CHEUI[23] (v0.1), xPore[28] (v2.1), Nanocompore[29] (v1.0.4). For m5C detection CHEUI, TandemMod and xPore were also employed. For Ψ detection, Tombo (v1.5.1) and NanoPSU[27] (v1.0) were used. Benchmarking was mainly performed on three datasets: *Mettl3* knockout and wild-type (KO/WT) mouse embryonic stem cell (mESC) samples for m6A detection (SRP166020), *NSUN2* KO/WT HeLa samples for m5C (SRP393373) and a mixed rRNA sample for Ψ detection (SRP329477). All sequencing reads were aligned using Minimap2. Most tools utilized using transcriptome-based alignment, except for EpiNano, which required genome-based alignment. For signal-to-reference alignment, most tools employed the eventalign module for f5c[76] or Nanopolish[77] with tool-specific parameter configurations. However, TandemMod used 'tombo resquiggle' command to map raw ionic signals to the basecalled sequences.

m6A detection in mESCs (*Mettl3* KO/WT) samples

To benchmarking m6A detection, we evaluated several tools using the *Mettl3* knockout and wild-type mouse embryonic stem cell (mESC) sample. Experimentally detected m6A sites were obtained from the GEO database, using the union of GLORI[15] (GSE210563) and miCLIP2[12] (GSE163491) datasets as ground truth references.

For EpiNano (v1.2 SVM & Error mode), we first converted BAM files into TSV format using sam2tsv from jvarkit (v2023.09.07) and extracted basecalling error features with Epinano_Variants.py script. Then, Epinano_DiffErr.R and Epinano_Predict.py were used to identify differentiated modified sites across samples and to directly predict m6A modifications, respectively. For CHEUI (v0.1 diff & solo mode), feature extraction was performed using 'CHEUI preprocess --m6A', then followed by direct m6A prediction using CHEUI_predict_model1.py and CHEUI_predict_model2.py. Additionally, signal-level differentiated analysis between samples was carried out using CHEUI_differentialRNAMod command. For m6Anet (v2.0.1), input data were preprocessed using 'm6anet dataprep' command, and m6A sites were inferred with 'm6anet inference'. For Nanocompore (v1.0.4), event-level signal alignments were first collapsed to the site level using ‘nanocompore eventalign_collapse’ command, and sample comparisons were performed using ‘nanocompore samplcomp’. For TandemMod, raw signal features were extracted from Tombo-resquiggled fast5 files using extract_signal_from_fast5.py and extract_feature_from_signal.py, and modification predictions were performed using ‘TandemMod.py --run_mode predict’. For xPore (v2.1), signal-level features were extracted using the ‘xpore dataprep’ command, and differential modification analysis was performed with ‘xpore diffmod’. For ORCA, both wild-type and knockout samples were processed using the modification presence prediction model.

To evaluate the m6A prediction performance, the difference in ORCA’s predicted modScore between paired samples was used to rank candidate m6A sites. The overlap between top differential modified m6A sites and reference m6A dataset were further calculated. For tools that directly provide differential predictions, such as xPore and CHEUI-diff, the reported difference in modification rate was used as the ranking metric. For EpiNano-Error and Nanocompore, we used the delta_sum_err and P-values, respectively, as provided in their outputs. To reduce tool-specific biases and ensure fair comparisons, transcriptome coordinates were converted to genomic coordinates, and only genomic sites reported in both samples were retained for downstream analysis. When multiple predictions are assigned to a single genomic site, the prediction with the highest score was selected to represent that position.

m5C detection in HeLa *NSUN2* KO/WT samples

To evaluate the m5C detection performance, CHEUI, TandemMod, xPore and ORCA were applied to *NSUN2* knockout and wild-type HeLa cell samples using the same preprocessing and prediction workflow described above. Reference m5C sites identified from HeLa BS-seq[39] (GSE122260), bsRNA-seq[40] (GSE140995) and RNA-BisSeq[41] (GSE93751) were directly downloaded from the GEO database and merged to construct a unified reference set. For comparison against WT and KO samples, the same strategy was applied here to ensure consistent genomic site-level comparisons, and only sites that detected in both samples were retained for evaluation. The changes in predicted stoichiometry between WT and KO samples at m5C sites were calculated for comparison.

Ψ detection in mixed rRNA sequencing samples

For Ψ detection evaluation, we evaluated ORCA, Tombo and NanoPSU using rRNA sequencing data. Reference Ψ sites supported by SILNAS-based mass spectrometry[46] were used as the ground truth. Basecalled reads were aligned to ribosomal RNAs from four species following the procedure described in NanoPSU[27]. For downstream comparison, only reads aligned to human 18S (NR_003286.4) and 28S (NR_003287.4) rRNA sequences were retained. For Tombo, Ψ sites were identified from resquiggled fast5 files using the ‘tombo detect_modifications’ command in de_novo detection mode. For NanoPSU, the recommend pipeline was applied, where alignment, remove_intron, extract_features and prediction commands were subsequently performed for Ψ site prediction. For TandemMod, raw signal features were extracted from Tombo-resquiggled fast5 files using extract_signal_from_fast5.py and extract_feature_from_signal.py, and modification predictions were performed using 'TandemMod.py --run_mode predict' with the Ψ detection model.

Benchmarking of zero-shot modification detection

To evaluate ORCA's zero-shot detection capability, training datasets were constructed by iteratively selecting different combinations of 2 to 5 RNA modification types from the full set of six RNA modifications (m6A, m5C, Ψ, m1A, hm5C, 5fC). For each combination, the selected modifications were used to train ORCA modification presence prediction model as previously described. The trained models were then used to perform zero-shot prediction on the modification types excluded from training. Prediction accuracy and recall were calculated for each target modification type for evaluation. In addition, the Pearson correlation between predicted stoichiometry and simulated ground truth were calculated to evaluate the generalizability of modification identification and quantification.

For zero-shot m6A prediction, an m6A-absent model was trained by excluding m6A from the training dataset. This model was then applied to direct RNA-seq data obtained from mouse brain tissue, and m6A sites detected by mouse brain m6A-SAC-Seq were used as reference dataset. ORCA's modScores were computed for all adenosine (A) sites located at the center of DRACH motifs across the transcriptome. The distribution of modScores between m6A-SAC-Seq-supported m6A sites and random selected background adenosines within DRACH motifs was compared using a two-sided Wilcoxon rank-sum test.

For Nm prediction, ORCA modScores were computed for sites supported by 2'-OMe-seq experiments[44] downloaded from RMBase v3.0[34]. All other transcriptome sites with read coverage greater than 10 were used as background controls. The distributions of modScores between reference Nm sites and background transcriptomic sites were compared using a two-sided Wilcoxon rank-sum test.

For zero-shot prediction of inosine (A-to-I) RNA editing sites, a *Schizosaccharomyces pombe* direct RNA-seq dataset was downloaded from PRJEB46364. The ASM294v2 S. pombe genome and annotation were downloaded, and the recommended strategy described in DeepEdit[45] was followed to establish a high-confidence set of inosines (A-to-I) RNA editing sites. Illumina RNA-seq reads from two hADAR2$^+$ samples and two control samples were downloaded and individually aligned using HISAT2[78] and processed with bcftools[69] mpileup for single-nucleotide variant calling. Candidate sites were defined as A-to-G substitutions supported by a minimum read coverage of 50 and a variant allele frequency exceeding 10%. Sites that consistently appeared in both hADAR2$^+$ samples and were absent from any of the control sample were retained as the final ground truth set for evaluation.

For the evaluation of rRNA modification predictions, high-confidence modification sites for human 18S/28S and yeast 18S/25S rRNAs were download from a published SILNAS mass-spectrometry based study[46]. The mixed-species rRNA direct RNA-seq library was then analyzed, and ORCA-predicted modScores were benchmarked against the reference modification sites to calculate the true-positive rate and false-discovery rate.

For the evaluation of non-natural modifications, publicly available 4sU-labelled K562 direct RNA-seq data[79] and a matched DMSO control sample were downloaded and analyzed using ORCA. The change in modScore between the 4sU-labelled and control samples was compared between U-containing and non-U-containing 5-mers was calculated to measure the enrichment of modification signal at 4sU-incorporated sites. In addition, a transcript-level 4sU load was defined as the sum of ORCA-estimated modified counts across U-centered sites per transcript, and was compared with transcript abundance in matched short-read 4sU pulldown RNA-seq data using Spearman correlation to evaluate concordance between ORCA-derived 4sU signals and orthogonal measurements of 4sU incorporation.

Benchmarking of simultaneous prediction of multiple RNA modification types

For both mESCs and HeLa datasets, RNA modification sites were independently predicted from two WT replicates. Only sites commonly identified in both replicates were retained, and mean modScore across replicates was calculated and used for site ranking. To establish a reference modification set, predicted modification sites were compared against public databases, including DirectRMDB[35] and RMBase v3.0[34], as well as orthogonal long-read based prediction tools: m6Anet (m6A), EpiNano-SVM (m6A), CHEUI-solo (m6A & m5C), and NanoPSU (Ψ). To exclude SNV interference, predicted modification sites with unnormal high mutation rate of insertion, deletion or mismatch exceeding 0.5 were excluded from the prediction results. Then, the top 10,000 ranked sites were retained for downstream analysis.

Benchmarking of modification type annotation

To evaluate the performance of modification type annotation, the K562 DRS dataset (K562_replicate6_run1) from the SGNex project was first employed for modification site prediction. A total of 48,377 high-confidence modification sites were annotated using NGS-supported sites from RMBase and DirectRMDB. The modification annotation model was trained as described above using a five-fold cross-validation. The annotation precision was measured by dividing the numbers of each correct modification annotation by the number of assigning to it and other types. The recall rate was measured by dividing the number of correct modification annotation by the total number of modification sites in the validation site. The final trained model was applied to unannotated candidate sites for modification type inference, and prediction results were subsequently filtered based on the correspondence between the predicted modification type and the reference nucleotide. Specifically, an additional filtering step was applied for m6A prediction to retain only sites located within the consensus DRACH motif.

To evaluate the false discovery rate (FDR), each modification type was iteratively excluded from the training dataset. The trained model was then used to assess the fraction of excluded modification type incorrectly predicted as one of the included types. The number of such misclassified sites was used to quantify the DFR for each modification type.

For modification annotation validation, *de novo* motif analysis of ORCA-annotated m5C sites was performed using XSTREME[50]. For m7G validation, the QKI-CLIP peaks were downloaded from GSE193039, and intersected with the annotated m7G sites. For m1A validation, peaks identified by m1A-Seq (GSE70485) were downloaded, and each peak was expanded by ±150 nt around the center before intersecting with predicted m1A sites. For Ψ validation, single-nucleotide resolution Ψ sites from BID-seq (GSE179798) were downloaded and compared against predicted Ψ, non-Ψ sites, and unmodified controls. For m6A validation, m6A sites detected by GLORI were downloaded from GSE210563 and compared against public databases and ORCA-annotated m6A sites. To account for potential influence in neighboring nucleotides, coordinates from BID-seq and GLORI were converted to 5-mer regions prior to comparison.

Exclusive modification sites identification and RBP enrichment analysis

To quantify transcript-level variation in exclusively modified sites, each pair of unique genomic sites was treated as an individual unit of analysis. To focus on isoform-specific modifications, only modification pairs that were supported by reads spanning more than one transcript isoform were retained. For each modification pair, an n × 3 contingency table was constructed, where n represents the number of transcript isoforms, and the three columns correspond to the number of reads assigned to each distribution including d_1 (simultaneous modification at both sites), d_2 (modification at the upstream site only), and d_3 (modification at the downstream site only). A chi-square test was then applied to assess variation in modification patterns across different transcript isoforms. Resulting p-values were used to rank the site pairs by isoform-specific modification heterogeneity.

To explore potential regulatory mechanisms underlying these isoform-specific modification patterns, the binding of RNA-binding proteins (RBPs) near these modification sites was analyzed. Genomic binding profiles from eCLIP-seq experiments for 139 RBPs in K562 cells were obtained from the ENCODE project in bedGraph format. All isoform-specific modification pairs were ranked by statistical significance, and enrichment of RBP binding around the top-ranked site pairs was assessed using a hypergeometric test. RBPs with P-values $\leqslant 0.01$ were considered significantly enriched. To further investigate the connection between isoform-specific modifications and splicing regulation, we performed an in-depth analysis using a curated list of 21 splicing factors from the SpliceAidF[80] database and 7 known RNA modification regulators, including writers (RBM15), erasers (FTO) and readers (HNRNPA1, IGF2BP1, IFG2BP2, FMR1).

For single-read-level validation, raw reads from m5C-TAC-Seq (SRP459299) and m6A-SAC-Seq (SRP295164) raw reads were downloaded and aligned to the hg38 reference genome with STAR[70] (v2.7.10b). Modification coordinates reported by m6A-SAC-Seq (GSE162356) and the m5C-TAC-Seq (provided in their supplementary data) were used as reference loci to quantify and visualize single-molecule co-occurrence between adjacent modification sites. The same m6A and m5C loci were independently visualized using ORCA predictions derived from IM95 and HCT116 direct RNA-seq samples, respectively.

Benchmarking on RNA004 chemistry data

For RNA004 chemistry, a prediction model was trained on curlcake IVT libraries[65] containing seven synthetic RNA modification samples and an unmodified control. Training samples with defined modification fractions and read depths were generated using the same mixing strategy as for the RNA002 training sets, and the trained model was evaluated on held-out RNA004 test samples by ROC and precision-recall analysis for each modification type. For comparison with the vendor-provided state-of-art caller, Dorado RNA modification models (rna004_130bps_hac@v5.2.0 for m5C, Ψ and inosine/m6A) were applied to the same curlcake reads. Read-level modification calls from Dorado were aligned to the reference using the Dorado aligner, and per-site modification statistics were obtained with modkit pileup. To place Dorado in the same simulated evaluation framework, synthetic sites spanning a range of sequencing depths and modification fractions were generated by random sampling, and the per-site modification fractions reported by Dorado were used as continuous prediction scores. Sites with a true simulated modification fraction of at least 0.1 were treated as positives, and ROC and precision–recall curves were computed using the Dorado scores.

To examine zero-shot detection on RNA004, leave-one-modification-out models were constructed by excluding each of the seven synthetic modification types in turn from the RNA004 curlcake training set and evaluating presence-prediction performance on the held-out type. De novo m6A detection across chemistries was further assessed on RNA002- and RNA004-based mouse brain direct RNA-seq libraries using models trained without m6A. For each library, ORCA modScores were calculated at adenosines located in DRACH motifs, and score distributions were compared between m6A-SAC-seq-supported sites and background DRACH positions. Overlaps of called m6A sites between the RNA002 and RNA004 datasets were summarized at the site level.

Benchmarking on ELIGOS and IVET training data

To assess the impact of training-set composition and sequencing chemistry on the presence-prediction model, additional models were trained on alternative IVT resources. One model was trained on IVET library[81] from the TandemMod study using the same simulation pipeline and neural network architecture as the ELIGOS-derived RNA002 training set. Performance of the ELIGOS-trained and IVET-trained models was compared by computing ROC and precision-recall curves, in order to evaluate cross-dataset generalization and the influence of sequence diversity in the training data.

Benchmarking on sequencing depth dependence of ORCA prediction model

The effect of read depth on presence-prediction performance was assessed by stratifying candidate sites into coverage bins and computing evaluation metrics within each bin. In the ELIGOS IVT datasets, sites with at least 10 supporting reads were grouped into four depth ranges. For each modification type and each depth bin, AUROC and AUPRC were calculated using the simulated modification labels as reference. A depth-stratified evaluation was also carried out for endogenous m6A detection in the mESC *Mettl3* WT/KO datasets. Sites were grouped into the same four coverage bins, and precision-recall and ROC curves were computed separately for each bin.

Benchmarking on assay-specific and cell line-specific biases

Assay-specific and cell-line-specific biases were evaluated on K562 cell line by partitioning database-supported sites for m6A, m5C, m1A and Nm according to the profiling technology or cell line. For each modification type in turn, annotation models with the same architecture and training procedure were trained in a multi-class setting in which training examples for the focal modification were restricted to sites from a single technology or cell line, whereas training examples for all other modification types and the background class always included all available database-supported sites. For each training configuration, recall and 1−FDR for the focal modification were quantified on sites detected by held-out technologies or cell lines.

Statistics and reproducibility

No statistical method was used to predetermine the sample size. The in vivo mouse experiment was performed once as a small proof-of-principle, using two adult C57BL/6 mice, and is not used for formal statistical comparisons or sex-specific analyses. This sample size was chosen to obtain one high-depth RNA002 and one high-depth RNA004 direct RNA-sequencing library in order to demonstrate technical feasibility rather than to estimate variability between animals. For all computational analyses, sample sizes were determined by the size of the available public datasets, and we used all reads or samples that passed predefined quality control criteria without additional subsampling. No data were excluded from the analyses.

Where statistical tests were applied, the specific test, the exact P values and the definition of n are provided in the figure legends or Source Data. P values < 0.05 were considered statistically significant unless otherwise stated. All computational analyses were performed using Python Jupyter Notebooks with numpy, pandas and scipy for numerical/statistical calculations and matplotlib/seaborn for plotting. All codes to replicate the analysis are available as part of code availability.

Data Availability

The m6A-SAC-seq and ONT direct RNA-seq data generated in this study have been deposited in the Genome Sequence Archive[82] in National Genomics Data Center, China National Center for Bioinformation (Accession number: PRJCA040561[https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA040561]) that are publicly accessible at https://ngdc.cncb.ac.cn/gsa. Publicly available nanopore direct RNA sequencing datasets used in this study were obtained from SRP166020[https://trace.ncbi.nlm.nih.gov/Traces/?view=study&acc=SRP166020], SRP393373[https://trace.ncbi.nlm.nih.gov/Traces/?view=study&acc=SRP393373], PRJEB46364[https://www.ebi.ac.uk/ena/browser/view/PRJEB46364], SRP329477[https://trace.ncbi.nlm.nih.gov/Traces/?view=study&acc=SRP329477], PRJEB82528[https://www.ebi.ac.uk/ena/browser/view/PRJEB82528], SRP426654[https://trace.ncbi.nlm.nih.gov/Traces/?view=study&acc=SRP426654], SRP171702[https://trace.ncbi.nlm.nih.gov/Traces/?view=study&acc=SRP171702] and the SGNex project[49]. RNA modification sites detected by NGS-based sequencing technologies were collected from the RMBase v3.0[34] and DirectRMDB[35]. Additionally, individual datasets were used for detection different modification types including m6A (GSE210563[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE210563], GSE163491[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE163491]), m5C (GSE140995[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140995], GSE93751[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93751], GSE122260[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE122260]), Ψ (GSE179798[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE179798]), m1A (GSE70485[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE70485]), m7G (GSE193039[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE193039]), and inosine (PRJEB46364[https://www.ebi.ac.uk/ena/browser/view/PRJEB46364]). In particular, the m5C sites were identified as described in CHEUI[23], and inosine editing sites were *de novo* identified from RNA-seq data in the DeepEdit[45] study. Source Data are provided at Zenodo (https://doi.org/10.5281/zenodo.17960329).

Code Availability

ORCA is implemented in Python and can be freely accessed on GitHub at https://github.com/bioinfo-biols/ORCA and is archived on Zenodo under the DOI https://zenodo.org/records/17949213[83]. The software is packaged with sample datasets and has been extensively tested on Linux. The detailed software installation guide has been included in our GitHub repository. Codes for data analysis have been depositied at Zenodo (https://doi.org/10.5281/zenodo.17785932).

References

1. Wang, X. et al. N6-methyladenosine-dependent regulation of messenger RNA stability. *Nature* **505**, 117–120 (2014).
2. Haussmann, I.U. et al. m(6)A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. *Nature* **540**, 301–304 (2016).
3. Mendel, M. et al. Splice site m(6)A methylation prevents binding of U2AF35 to inhibit RNA splicing. *Cell* **184**, 3125–3142 e3125 (2021).
4. Chen, T. et al. m(6)A modification plays an integral role in mRNA stability and translation during pattern-triggered immunity. *Proc Natl Acad Sci U S A* **121**, e2411100121 (2024).
5. Wang, X. et al. N(6)-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell* **161**, 1388–1399 (2015).
6. Loedige, I. et al. mRNA stability and m(6)A are major determinants of subcellular mRNA localization in neurons. *Mol Cell* **83**, 2709–2725 e2710 (2023).
7. Zhang, Y., Lu, L. & Li, X. Detection technologies for RNA modifications. *Exp Mol Med* **54**, 1601–1616 (2022).
8. Lucas, M.C. & Novoa, E.M. Long-read sequencing in the era of epigenomics and epitranscriptomics. *Nat Methods* **20**, 25–29 (2023).
9. Huang, S., Wylder, A.C. & Pan, T. Simultaneous nanopore profiling of mRNA m(6)A and pseudouridine reveals translation coordination. *Nat Biotechnol* **42**, 1831–1835 (2024).
10. Wu, Y. et al. Transfer learning enables identification of multiple types of RNA modifications using nanopore direct RNA sequencing. *Nat Commun* **15**, 4049 (2024).
11. Zhao, Z. et al. QKI shuttles internal m(7)G-modified transcripts into stress granules and modulates mRNA metabolism. *Cell* **186**, 3208–3226 e3227 (2023).
12. Kortel, N. et al. Deep and accurate detection of m6A RNA modifications using miCLIP2 and m6Aboost machine learning. *Nucleic Acids Res* **49**, e92 (2021).
13. Dominissini, D. et al. The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA. *Nature* **530**, 441–446 (2016).
14. Dai, Q. et al. Quantitative sequencing using BID-seq uncovers abundant pseudouridines in mammalian mRNA at base resolution. *Nat Biotechnol* **41**, 344–354 (2023).
15. Liu, C. et al. Absolute quantification of single-base m(6)A methylation in the mammalian transcriptome using GLORI. *Nat Biotechnol* **41**, 355–366 (2023).

16. Dai, Q. et al. Ultrafast bisulfite sequencing detection of 5-methylcytosine in DNA and RNA. *Nat Biotechnol* **42**, 1559–1570 (2024).

17. Hu, L. et al. m(6)A RNA modifications are measured at single-base resolution across the mammalian transcriptome. *Nat Biotechnol* **40**, 1210–1219 (2022).

18. Jain, M., Abu-Shumays, R., Olsen, H.E. & Akeson, M. Advances in nanopore direct RNA sequencing. *Nat Methods* **19**, 1160–1164 (2022).

19. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K.F. Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**, 1348–1365 (2021).

20. Begik, O. et al. Quantitative profiling of pseudouridylation dynamics in native RNAs with nanopore sequencing. *Nat Biotechnol* **39**, 1278–1291 (2021).

21. Liu, H. et al. Accurate detection of m(6)A RNA modifications in native RNA sequences. *Nat Commun* **10**, 4079 (2019).

22. Hendra, C. et al. Detection of m6A from direct RNA sequencing using a multiple instance learning framework. *Nat Methods* **19**, 1590–1598 (2022).

23. Acera Mateos, P. et al. Prediction of m6A and m5C at single-molecule resolution reveals a transcriptome-wide co-occurrence of RNA modifications. *Nat Commun* **15**, 3899 (2024).

24. Gao, Y. et al. Quantitative profiling of N(6)-methyladenosine at single-base resolution in stem-differentiating xylem of Populus trichocarpa using Nanopore direct RNA sequencing. *Genome Biol* **22**, 22 (2021).

25. Lorenz, D.A., Sathe, S., Einstein, J.M. & Yeo, G.W. Direct RNA sequencing enables m(6)A detection in endogenous transcript isoforms at base-specific resolution. *RNA* **26**, 19–28 (2020).

26. Wu, Y. et al. Simultaneous profiling of ac(4)C and m(5)C modifications from nanopore direct RNA sequencing. *Int J Biol Macromol* **305**, 140863 (2025).

27. Huang, S. et al. Interferon inducible pseudouridine modification in human mRNA by quantitative nanopore profiling. *Genome Biol* **22**, 330 (2021).

28. Pratanwanich, P.N. et al. Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore. *Nat Biotechnol* **39**, 1394–1402 (2021).

29. Leger, A. et al. RNA modifications detection by comparative Nanopore direct RNA sequencing. *Nat Commun* **12**, 7198 (2021).

30. Liu, N. et al. Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA* **19**, 1848–1856 (2013).

31. Jenjaroenpun, P. et al. Decoding the epitranscriptional landscape from native RNA sequences. *Nucleic Acids Res* **49**, e7 (2021).

32. Ganin, Y. et al. in Journal of machine learning research (2015).

33. Tourancheau, A., Mead, E.A., Zhang, X.S. & Fang, G. Discovering multiple types of DNA methylation from bacteria and microbiome using nanopore sequencing. *Nat Methods* **18**, 491–498 (2021).

34. Xuan, J. et al. RMBase v3.0: decode the landscape, mechanisms and functions of RNA modifications. *Nucleic Acids Res* **52**, D273–D284 (2024).

35. Zhang, Y. et al. DirectRMDB: a database of post-transcriptional RNA modifications unveiled from direct RNA sequencing technology. *Nucleic Acids Res* **51**, D106–D116 (2023).

36. Spangenberg, J. et al. The RMaP challenge of predicting RNA modifications by nanopore sequencing. *Commun Chem* **8**, 115 (2025).

37. Cappannini, A. et al. MODOMICS: a database of RNA modifications and related information. 2023 update. *Nucleic Acids Res* **52**, D239–D244 (2024).

38. Zhong, Z.D. et al. Systematic comparison of tools used for m(6)A mapping from nanopore direct RNA sequencing. *Nat Commun* **14**, 1906 (2023).

39. Huang, T., Chen, W., Liu, J., Gu, N. & Zhang, R. Genome-wide identification of mRNA 5-methylcytosine in mammals. *Nat Struct Mol Biol* **26**, 380–388 (2019).

40. Schumann, U. et al. Multiple links between 5-methylcytosine content of mRNA and translation. *BMC Biol* **18**, 40 (2020).

41. Yang, X. et al. 5-methylcytosine promotes mRNA export - NSUN2 as the methyltransferase and ALYREF as an m(5)C reader. *Cell Res* **27**, 606–625 (2017).

42. Stoiber, M. et al. <em>De novo</em> Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. *bioRxiv*, 094672 (2017).

43. Flemmich, L., Bereiter, R. & Micura, R. Chemical Synthesis of Modified RNA. *Angew Chem Int Ed Engl* **63**, e202403063 (2024).

44. Incarnato, D. et al. High-throughput single-base resolution mapping of RNA 2´-O-methylated residues. *Nucleic Acids Res* **45**, 1433–1441 (2017).

45. Chen, L. et al. DeepEdit: single-molecule detection and phasing of A-to-I RNA editing events using nanopore direct RNA sequencing. *Genome Biol* **24**, 75 (2023).

46. Taoka, M. et al. Landscape of the complete RNA chemical modifications in the human 80S ribosome. *Nucleic Acids Res* **46**, 9289–9298 (2018).

47. Drexler, H.L., Choquet, K. & Churchman, L.S. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol Cell* **77**, 985–998 e988 (2020).

48. Georgieva, D., Liu, Q., Wang, K. & Egli, D. Detection of base analogs incorporated during DNA replication by nanopore sequencing. *Nucleic Acids Res* **48**, e88 (2020).

49. Chen, Y. et al. A systematic benchmark of Nanopore long-read RNA sequencing for transcript-level analysis in human cell lines. *Nat Methods* **22**, 801–812 (2025).

50. Grant, C.E. & Bailey, T.L. XSTREME: Comprehensive motif analysis of biological sequence datasets. *bioRxiv*, 2021.2009.2002.458722 (2021).

51. Selmi, T. et al. Sequence- and structure-specific cytosine-5 mRNA methylation by NSUN6. *Nucleic Acids Res* **49**, 1006–1022 (2021).

52. Lu, L. et al. Base-resolution m(5)C profiling across the mammalian transcriptome by bisulfite-free enzyme-assisted chemical labeling approach. *Mol Cell* **84**, 2984–3000 e2988 (2024).

53. Consortium, E.P. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

54. Chang, S.H. et al. ELAVL1 regulates alternative splicing of eIF4E transporter to promote postnatal angiogenesis. *Proc Natl Acad Sci U S A* **111**, 18309–18314 (2014).

55. Didiot, M.C. et al. The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer. *Nucleic Acids Res* **36**, 4902–4912 (2008).

56. Edens, B.M. et al. FMRP Modulates Neural Differentiation through m(6)A-Dependent mRNA Nuclear Export. *Cell Rep* **28**, 845–854 e845 (2019).

57. Arguello, A.E., DeLiberto, A.N. & Kleiner, R.E. RNA Chemical Proteomics Reveals the N(6)-Methyladenosine (m(6)A)-Regulated Protein-RNA Interactome. *J Am Chem Soc* **139**, 17249–17252 (2017).

58. Roundtree, I.A., Evans, M.E., Pan, T. & He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell* **169**, 1187–1200 (2017).

59. Barbieri, I. & Kouzarides, T. Role of RNA modifications in cancer. *Nat Rev Cancer* **20**, 303–322 (2020).

60. Wei, J. et al. Differential m(6)A, m(6)A(m), and m(1)A Demethylation Mediated by FTO in the Cell Nucleus and Cytoplasm. *Mol Cell* **71**, 973–985 e975 (2018).

61. Zhang, J. et al. Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nat Biotechnol* **39**, 836–845 (2021).

62. Zhang, J. et al. Real-time and programmable transcriptome sequencing with PROFIT-seq. *Nat Cell Biol* **26**, 2183–2194 (2024).

63. Zhang, J. & Zhao, F. Circular RNA discovery with emerging sequencing and deep learning technologies. *Nat Genet* **57**, 1089–1102 (2025).

64. Liu-Wei, W. et al. Sequencing accuracy and systematic errors of nanopore direct RNA sequencing. *BMC Genomics* **25**, 528 (2024).

65. Cruciani, S. et al. De novo basecalling of RNA modifications at single molecule and nucleotide resolution. *Genome Biol* **26**, 38 (2025).

66. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17**, 3 (2011).

67. Chen, S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *Imeta* **2**, e107 (2023).

68. Langmead, B. & Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357–359 (2012).

69. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Gigascience* **10** (2021).

70. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

71. Liu, D. Algorithms for efficiently collapsing reads with Unique Molecular Identifiers. *PeerJ* **7**, e8275 (2019).
72. Koboldt, D.C. et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* **22**, 568–576 (2012).
73. Gamaarachchi, H. et al. Fast nanopore sequencing data analysis with SLOW5. *Nat Biotechnol* **40**, 1026–1029 (2022).
74. Samarakoon, H. et al. Flexible and efficient handling of nanopore sequencing signal data with slow5tools. *Genome Biol* **24**, 69 (2023).
75. Li, H. New strategies to improve minimap2 alignment accuracy. *Bioinformatics* **37**, 4572–4574 (2021).
76. Gamaarachchi, H. et al. GPU accelerated adaptive banded event alignment for rapid comparative nanopore signal analysis. *BMC Bioinformatics* **21**, 343 (2020).
77. Simpson, J.T. et al. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods* **14**, 407–410 (2017).
78. Kim, D., Paggi, J.M., Park, C., Bennett, C. & Salzberg, S.L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907–915 (2019).
79. Drexler, H.L., Choquet, K. & Churchman, L.S. Splicing Kinetics and Coordination Revealed by Direct Nascent RNA Sequencing through Nanopores. *Mol Cell* **77**, 985–998.e988 (2020).
80. Giulietti, M. et al. SpliceAid-F: a database of human splicing factors and their RNA-binding sites. *Nucleic Acids Res* **41**, D125–131 (2013).
81. Wu, Y. et al. Transfer learning enables identification of multiple types of RNA modifications using nanopore direct RNA sequencing. *Nature Communications* **15**, 4049 (2024).
82. Chen, T. et al. The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics Proteomics Bioinformatics* **19**, 578–583 (2021).
83. Dong, H., Zhao, F. & Zhang, J. Comprehensive mapping of RNA modification dynamics and crosstalk via deep learning and nanopore direct RNA-sequencing. *Zenodo* https://doi.org/10.5281/zenodo.17949213 (2025).

Acknowledgements

This work was supported by grants from the National Key R&D Project [2023YFC2604300, 2024YFF1206600, 2025YFF1207900, 2021YFA1300500, 2021YFA1302000], the National Natural Science Foundation of China [32422020, 32025009, 32130020] and the Zhejiang Province Vanguard Goose-Leading Initiative [2025C01114].

Author Contributions

F.Z. and J.Z. conceived the project. H.D. and J.Z. designed the method. H.D. implemented the model. H.D. and J.Z. performed data analysis. Y.G., Z.C., Y.L. and X.L. performed the experiments and generated sequencing data. H.D., J.Z. and F.Z. wrote the manuscript with the contribution of all authors.

Competing Interests

The authors declare no competing interests.

Figures

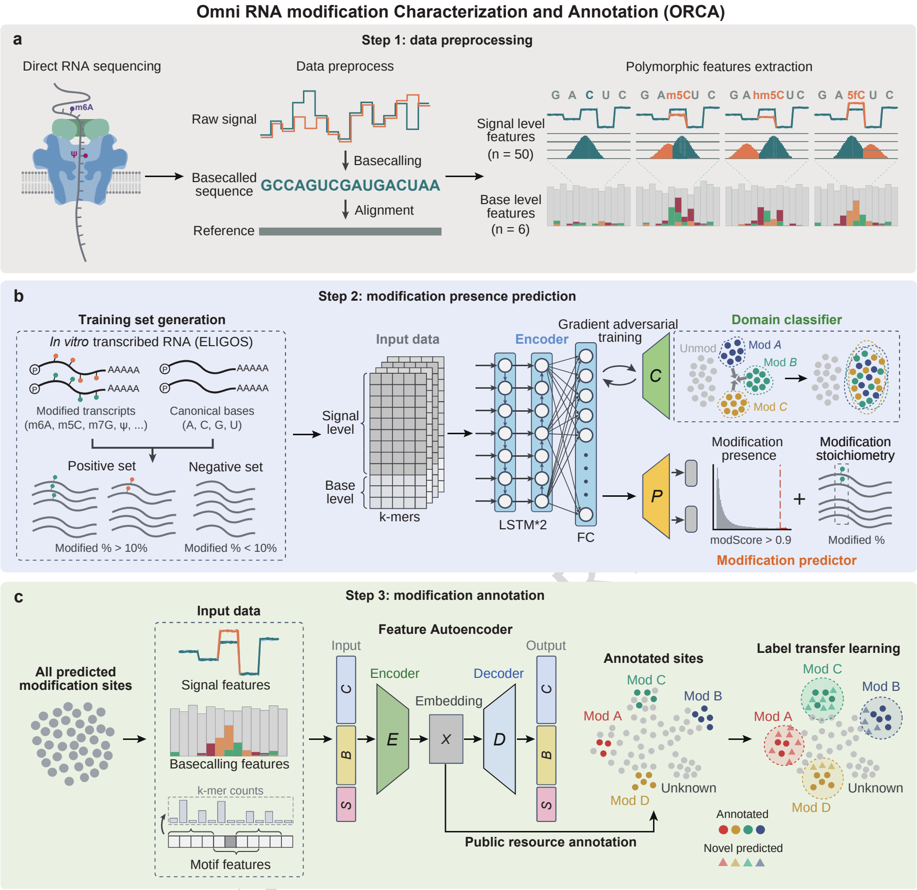

Fig. 1 | Overview of the ORCA framework for RNA modification characterization and annotation. a, Workflow of the preprocessing nanopore direct RNA sequencing data and extracting polymorphic features. Raw RNA reads were basecalled and aligned to the reference genome. Reads aligned to the same genomic position were aggregated, and both signal- and base-level features were extracted to capture the variability caused by presence of both modified and unmodified RNAs. b, Schematic of training set generation and modification presence prediction. The ELIGOS IVT dataset was sampled to simulate stoichiometries and sequencing depths of endogenous modification sites. Extracted features were used to train the ORCA modification presence prediction model, which comprising a dual-layer Bi-LSTM feature encoder, a domain classifier (C) and a modification predictor (P). The model was adversarially trained to suppress domain classifier performance while preserving accurate prediction of modification presence and stoichiometry. c, Schematic of the modification annotation model. Signal- and sequence-level features of predicted modification sites were combined with k-mer frequency features representing motif preference of RNA modifications. An autoencoder was first trained to embed all modification sites into lower dimensions, then the encoder was fine-tuned to predict modification types based on public reference annotation. The final model was then used for label-transfer prediction of previously unannotated modification sites.

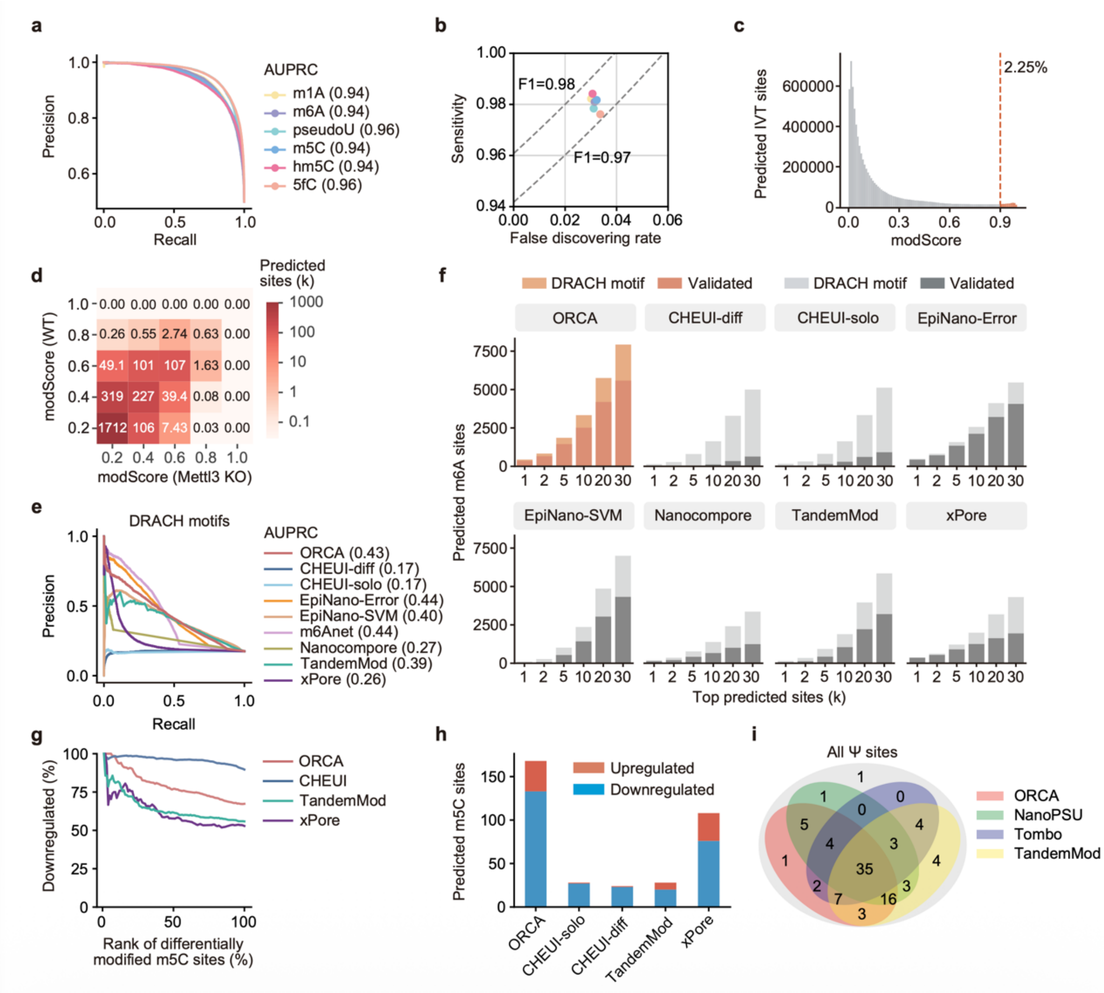

Fig. 2 | Evaluation of modification prediction performance of ORCA. a, Precision-recall (PR) curves for six RNA modifications by cross-validation using the IVT training set. b, Sensitivity and false discovery rate (FDR) for each modification type at the default modScore threshold of 0.9. Colors indicate different modification types. c, Distribution of modScore in IVT human mRNA transcriptome. Red dashed line represents the modScore threshold of 0.9. d, Heatmap showing the modScore of sites detected in both *Mettl3* KO and WT mESCs. e, PR curves comparing ORCA with other m6A-specific and comparative-based tools for detecting m6A sites at motif level. All sites within DRACH motifs were ranked based on change of modScore between WT and KO samples. f, Number of top-ranked differential modified sites supported by m6A sequencing (GLORI or miCLIP2) or DRACH motif for each method. g, Proportion of downregulated m5C sites among top differentially modified sites after *NSUN2* knockout (KO). h, Number of significantly up- and down-regulated m5C sites detected by each tool. Differentially modified sites were identified using a chi-squared test comparing WT and KO predictions (p threshold: 0.05, exact per-site p values are provided in the Source Data file). Red: upregulated; Blue:

downregulated. i, Venn diagram comparing Ψ sites prediction of on human rRNA among ORCA, NanoPSU, TandemMod and Tombo. Source data are provided as a Source Data file.

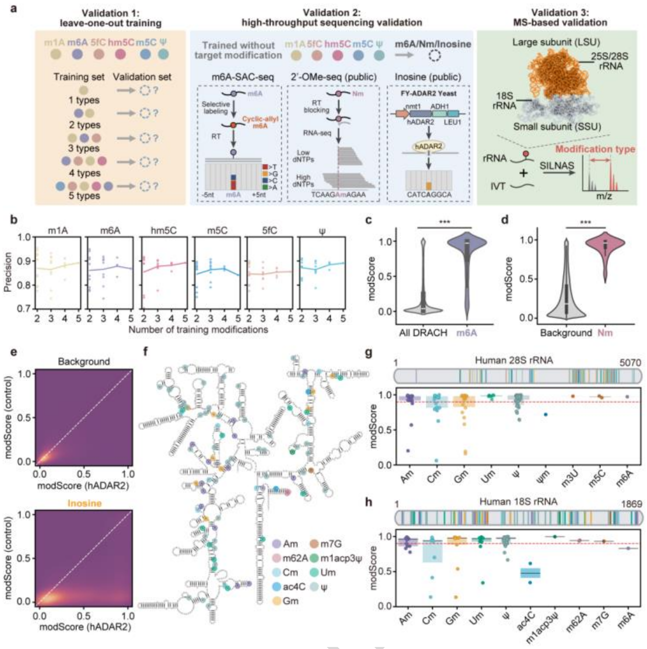

Fig. 3 | Evaluation of zero-shot detection capability for unseen RNA modification types. a, Overview of three evaluation strategies used to assess ORCA’s ability to detect unseen RNA modification types. Left: leave-one-out training using the IVT dataset to evaluate zero-shot prediction performance. Middle: evaluation of ORCA’s prediction of m6A, 2’-O-methylation (Nm) and inosine (I) sites detected using orthogonal NGS-based modification sequencing datasets. Right: prediction of ribosomal RNA medication sites validated by SILNAS mass spectrometry. b, Prediction accuracy of ORCA trained using different combination of modification types to predict modification types not included in training set. Lines represent the mean accuracy using different number of training modification types. c, modScore of m6A sites detected by m6-SAC-seq and background DRACH motifs. m6A sites, n = 2,067; background DRACH motifs, n = 2,466,196. p = 1×10^{-31}, two-sided Wilcoxon rank-sum test. d, modScore of Nm sites detected by 2’-OMe-seq. Nm sites, n =39; background, n = 5,921,554, p = 1.30×10^{-24}, two-sided Wilcoxon rank-sum test. For both violin plot panels, violin plots show the distribution of per-site values; the inner boxes indicate the interquartile range (25th – 75th percentile) with the central line marking the median, and the whiskers extend to the most extreme data points within 1.5× the interquartile range. e, Density plot of modScores across RNA-seq detected inosines (bottom) and background transcriptome positions (top) in *hADAR2*$^+$ and WT *S. pombe*. Color scale indicates site density. f, Schematic of the human 18S rRNA secondary structure and RNA modification sites detected by SILNAS-MS. Colors denote different RNA modification types. g-h, modScore distributions of different RNA modification types in human 28S (g, n = 91) and 18S (h, n = 133) rRNA. Exact site counts for each modification type are provided in the Source Data file. Each point represents one transcriptomic site. Red dashed lines indicate the default modScore threshold of 0.9. Source data are provided as a Source Data file.

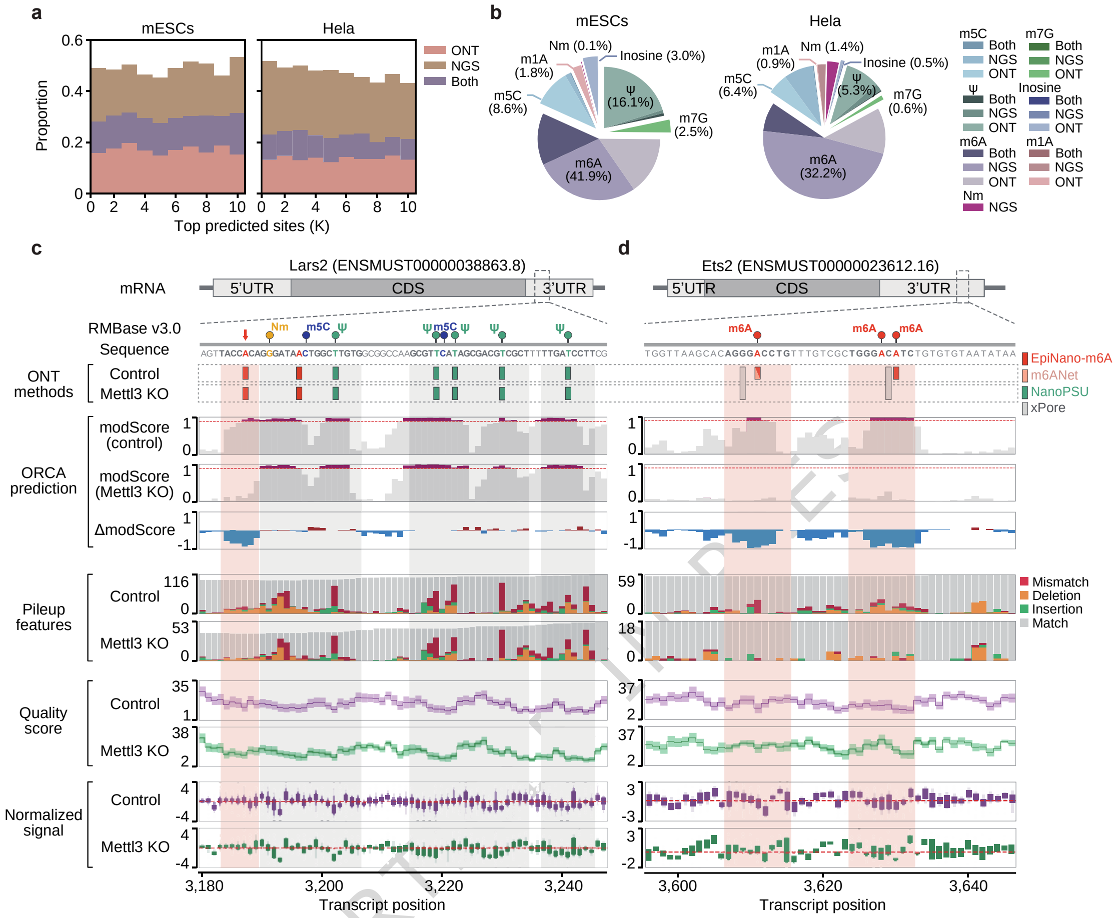

Fig. 4 | Simultaneous detection of different RNA modification types in HeLa and mESC transcriptome. a, Stacked bar plots of the fraction of predicted modification sites supported by NGS-based methods or other nanopore (ONT)-based prediction tools. Modification sites were ranked by ORCA's modScores. b, Pie charts illustrating the composition of modification types and supporting platforms among the top 10,000 predicted sites in mESCs and HeLa transcriptome. For each modification type, colors indicate dual support of b oth NGS and ONT platforms (dark), NGS-only support (medium) or ONT-only support (light). c-d, Genome track views of predicted modification sites in mouse *Lars2* (c) and *Ets2* (d) transcripts. Differentially modified m6A sites are highlighted in red, and other modification types are shown in grey above the panels. In the normalized signal tracks, the central line indicates the median, and the nested boxes represent progressively more extreme quantile ranges of the distribution (with the innermost box approximately spanning the interquartile range, 25th – 75th percentile), and n corresponds to the read depth at each transcript position; exact per-position n values are provided in source data. Source data are provided as a Source Data file.

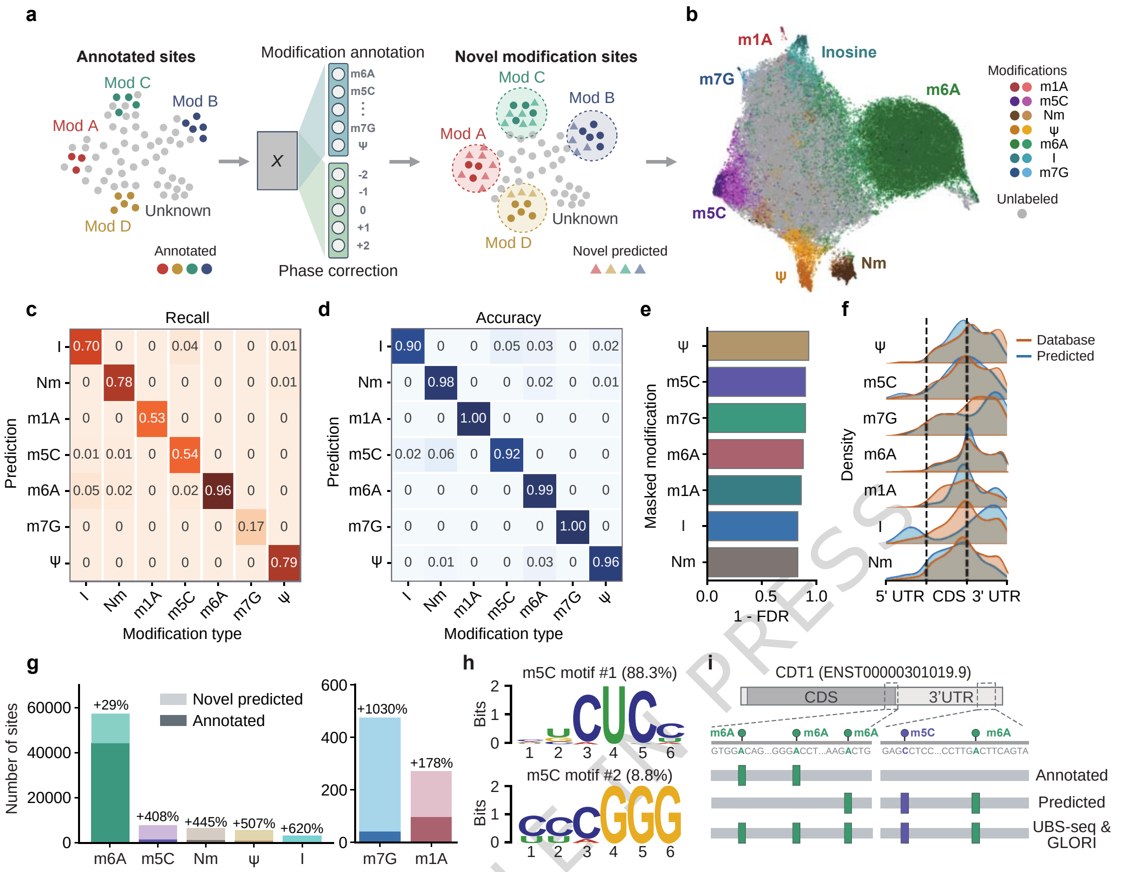

Fig. 5 | Annotation of modification sites with a transfer-learning strategy. a, Schematic overview of the modification sites annotation pipeline. Predicted sites were first used to train an autoencoder, annotated modification types from public databases were then incorporated for supervised fine-tuning. The encoder was fine-tuned to jointly predict modification type and phase shift, enabling accurate label transfer for previously unannotated modification sites. b, UMAP visualization of model-predicted logit scores assigning each site to each modification type. Colors indicate database-annotated modification sites (dark), ORCA-predicted modification sites (light), and unknown modification sites (grey). c-d, Cross-validation heatmaps showing the recall (c) and accuracy (d) of the modification type annotation model. e, Bar plot showing the specificity of the annotation model: 1 minus the false assignment rate for each modification type when that type is masked during training. f, Gene body distribution of database-annotated (orange) and ORCA-predicted (blue) modification sites. g, Number of ORCA-predicted modification sites across seven modification types. Dark bars indicate database-annotated modification sites, and light bars represent ORCA-predicted sites. h, Two representative sequence motifs enriched among ORCA-predicted m5C sites. i, Annotation of ORCA-predicted m6A and m5C sites on the *CDT1* transcript. Tracks represent database-annotated m6A sites, ORCA-predicted m6A and m5C sites, and supporting evidence from UBS-Seq and GLORI. Source data are provided as a Source Data file.

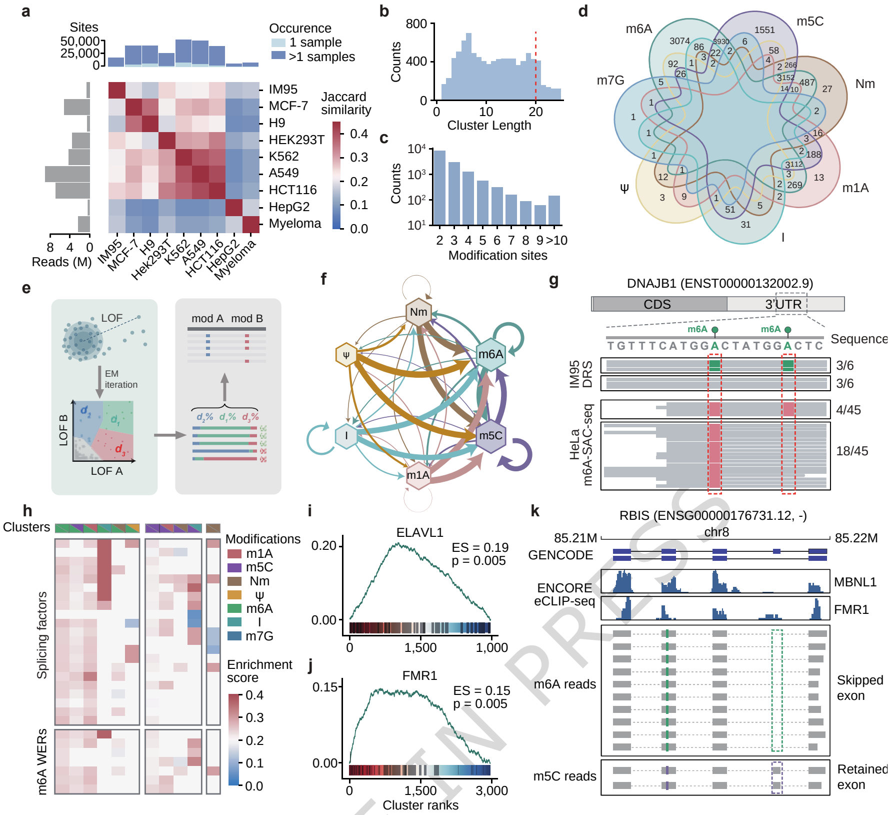

Fig. 6 | RNA modification landscape and interplay of neighboring RNA modification sites across human cell lines. a, Summary of reads counts, number of detected modification sites, and Jaccard similarity of modification sites across nine human cell lines. b, Length distribution of modification clusters. C, Number of modification clusters stratified by the number of modification sites per cluster. d, Venn diagram showing the overlap of modification types within all identified clusters. e, Schematic of the expectation-maximization (EM)-based model used to estimate co-occurrence of mutually exclusive interactions between neighboring modification sites. Local outlier factor (LOF) is used to assess the probability of each read being modified at each site, followed by statistical testing to evaluate the significance of modification interactions. f, Proportion of co-occurrence events between different modification type pairs. Arrow thickness reflects the relative frequency of source-target modification interactions. g, Example of two co-occurring m6A sites in the *DNAJB1* transcript detected in IM95 DRS data and validated by HeLa m6A-SAC-seq. h, Heatmap of enrichment score for splicing regulators and RNA modification-associated proteins (writers, erasers, readers; WERs) at isoform-specific, exclusively modified clusters. i-j, Enrichment plots for ELAVL1 (i) and FMR1 (j) at isoform-specific m6A-m5C modification clusters. P values were calculated using a two-sided permutation test. k, Genome browser view of the RBIS gene illustrating the relationship between RNA modifications and alternative splicing. Colors indicate predicted m6A (green) and m5C (purple) sites within exon 4, and dashed lines demonstrate the upstream alternative spliced exon. Source data are provided as a Source Data file.

Editor's Summary
RNA modifications influence gene regulation, but global mapping was limited. Here, the authors introduce ORCA, a deep learning framework using nanopore RNA sequencing to detect multiple modification types, revealing isoform-specific patterns and regulatory interactions.

Peer Review Information: *Nature Communications* thanks Jia Meng and Mattia Pelizzola for their contribution to the peer review of this work. A peer review file is available.