tRNA isodecoder analysis using Nanopore ionic current signals and deep learning

Stuart Akeson[1,*], Pooria Daneshvar Kakhaki[2,*], Neda Ghohabi Esfahani[1], Julia L. Reinsch[3,4], Margaret L. Barry[3,4], Maddy Zamecnik[5], Talia Tzadikario[1], Robin L. Abu-Shumays[6], David M. Garcia[3,4], Kristin S. Koutmou[5], Miten Jain[1,8,9]

* These authors contributed equally to this work.
Correspondence should be addressed to: mi.jain@northeastern.edu

1 Department of Bioengineering, Northeastern University, Boston, MA 02115, USA
2 Department of Electrical & Computer Engineering, Northeastern University, Boston, MA 02115, USA
3 Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA
4 Department of Biology, University of Oregon, Eugene, OR 97403, USA
5 Department of Chemistry, University of Michigan, Ann Arbor, MI 48109, USA
6 Biomolecular Engineering Department, University of California Santa Cruz, Santa Cruz, CA 95064, USA
7 Center for Molecular Biology of RNA, University of California Santa Cruz, Santa Cruz, CA 95064, USA
8 Department of Physics, Northeastern University, Boston, MA 02115, USA
9 Khoury College of Computer Sciences, Northeastern University, Boston, MA 02115, USA

SIMPLIFIED BANNER GRAPHIC

KEYWORDS
Nanopore direct tRNA sequencing
Ionic current analysis
Deep learning
Isodecoder classification
tRNA modifications

ABSTRACT
We describe application of Nanopore direct tRNA sequencing and ionic current signal for isodecoder-level analysis of bacterial and yeast tRNA. The computational advance is the application of deep learning for predicting tRNA isodecoders directly from raw nanopore ionic current signals. We then combined ionic current analysis with a pairwise sequence alignment strategy to improve tRNA sequence alignments. We applied this strategy to Nanopore data for *E. coli* and *S. cerevisiae* tRNA and aligned 2.6% and 13.1% more reads respectively when compared to present strategies for Nanopore tRNA analysis. Notably, this increase in alignments also came with a significant increase in alignment identity. The ionic current models confirmed enrichment of both an individual tRNA isotype and a tRNA fraction in separate experiments. Finally, we demonstrate that raw ionic current analysis is information-rich and has the potential to deconvolute dense chemical modifications in complex molecules like tRNA. This

is an advance for tRNA analysis in prokaryotes and lower eukaryotes with potential to elucidate the role of tRNA in discovery biology and in human health applications.

INTRODUCTION

tRNA are the most abundant type of RNA in terms of molecular concentration. They are typically ~76 nucleotides in length and include extensive chemical modifications[1,2]. Conventionally, tRNA molecules' abundance, sequence, and chemical modifications have been analyzed using Northern blots[3–5], liquid chromatography paired with tandem mass spectrometry (LC-MS)[2,6–8], and various short-read sequencing technologies that require reverse transcription and PCR for generating cDNA copies[8,9].

Due to the structured and highly modified nature of tRNA, complementary methods can help resolve aspects of tRNA biology[6,8]. For example, we previously combined Nanopore direct RNA sequencing, LC-MS, and genetic controls in order to study the relationship between modifications in the T-Loop of *Saccharomyces cerevisiae*[6]. These approaches can help advance documentation of alterations in tRNA sequence and modifications that have been linked to cancer, metabolic disorders, and neurodegenerative disease[1].

tRNA are typically categorized into three categories: i) Isotypes – different tRNA that carry the same amino acid ; ii) Isoacceptors – different tRNA that carry the same amino acid but have distinct anticodon sequences; and iii) Isodecoders – different tRNA that carry the same amino acid and have the same anticodon sequence but have small sequence differences in the tRNA "body" (any nucleotide other than the anticodon)[10]. Isodecoders are the most challenging to distinguish because they can differ by as little as a single nucleotide[11]. As a result, the majority of tRNA sequence analysis relies on isoacceptor-level discrimination, or will errantly map similar isodecoders to a single reference sequence[6,12]. This can lead to isodecoders that are removed from the reference sequence options, analyzed under the umbrella of the parent isoacceptor without isodecoder disambiguation, or randomly assigned based on alignment algorithm implementation for reads that map equally well to multiple isodecoder reference options. Disambiguation of isodecoders at the single nucleotide sequence level requires higher accuracy than has been achieved by previously used tools.

Prior work demonstrated that Nanopore DRS provides full-length readouts of native tRNA from bacterial, yeast, human, and other species (e.g. trypanosomes)[6,12–15]. During Nanopore tRNA sequencing, the tRNA passes through the protein pore, across which voltage has been applied, and the specific order of nucleotides and modified nucleotides produces a series of unique ionic current signatures[16]. Decoding these ionic signatures into a sequence of bases is challenging, particularly for modified tRNA that exists in cells. However, improvements have been achieved through a combination of updated Nanopore sequencing chemistry and basecalling software. For example, Nanopore DRS RNA004 platform, released in 2024, and Dorado 1.0.0 basecaller have improved median mRNA alignment accuracy to 98.67%[17,18]. This was an improvement from the 90.65% that was achieved using the older Nanopore DRS RNA002 platform and Guppy basecaller. Despite these advances in direct mRNA sequencing, tRNA sequence alignment accuracy remains lower than mRNA accuracy. This impact on accuracy is, in part, due to` the presence of complex and dense chemical modifications in tRNA. RNA modifications can cause perturbations in Nanopore ionic current. These perturbations affect the basecalling step, where a deep learning model predicts nucleotide sequence based on ionic current. The

presence of RNA modifications can sometimes cause systematic basecall errors. This observation has been previously exploited for RNA modification analysis[6,13,19–25].

Use of Nanopore ionic current data is a principled approach for resolving complex modifications or discriminating isodecoders. Recent efforts have demonstrated that ionic current analysis can be applied to tRNA. These include: i) measuring the aminoacylation status of tRNA[14]; ii) applying ionic current-based modification models trained for mRNA for "off-label" modification predictions in tRNA[26]; and iii) ionic current classification for demultiplexing barcoded tRNA[27]. However, conventional alignment approaches lack isodecoder-level resolution and are limited at an isoacceptor-level analysis[12,13]. Here, we achieved isodecoder-level resolution in bacterial and yeast tRNA using Nanopore direct tRNA sequencing and a bespoke ionic current analysis strategy. This strategy (tRNAZAP) includes ionic current classification and a pairwise alignment strategy that uses both sequence and ionic current information. We documented significant improvements in alignment of Nanopore tRNA strand reads compared to previous methods for *E. coli* and *S. cerevisiae*. We also demonstrated discrimination among isodecoders, with as little as a single nucleotide difference, using ionic current analysis.

RESULTS AND DISCUSSION

Ionic current modeling using *in vitro* transcribed *Escherichia coli* tRNA

We performed *in vitro* transcription for the complete set of gtRNAdb predicted *E. coli* tRNA. This set was resolved at the complete isodecoder-level, in contrast from most previous tRNA nanopore work. The IVT tRNA mirrored the predicted biological tRNA sequences exactly including the majority of the biological tRNA sequencing adapters (**Figure 1A**). To facilitate ionic current learning we augmented our previous adapter design[6,13] to include a longer 3′ component. We sequenced these IVT tRNA using Oxford Nanopore Technologies (ONT) PromethION flow cell and Nanopore DRS RNA004 chemistry. This sequencing experiment produced 18,985,579 reads. Using BWA-MEM (-x ont2d -w 13 -K 6) 15,310,260 of these were primary alignments with mapping quality over 0. The aligned reads had a median alignment identity of 97.26% and a median aligned read length of 140. To visualize the relationship between alignment identity and alignment length we plotted the per read metrics (**Figure 1B**). We documented a median deletion proportion of 0.015 and median insertion proportion of 0.0097. To dampen the impact of this bias, we only used reads that achieved an alignment identity of 75% in the tRNA region to minimize misclassification in the training set. Across the 51 unique *E. coli* tRNA isodecoders representing the complete set of gtRNAdb predicted *E. coli* sequences, we recorded a median of 108,422 reads per *E. coli* isodecoder, which was sufficient for training. Since the isodecoder assignments are based on BWA-MEM primary alignment, this introduces BWA-MEM bias into the model training. We documented over 100,000 reads per IVT tRNA isodecoder for all but three isodecoders (*E. coli* Gln-TTG-1-1 (1,086 reads), *E. coli* Leu-TAA-1-1 (256 reads), and *Saccharomyces cerevisiae* Val-AAC-1-1 (789 reads)). Additional sequencing using templates exclusively for those three isodecoders provided sufficient reads for model training. With the merged data sets, we documented a minimum of 39,067 primary reads with a mapping quality greater than 0 across all 51 isodecoders. For model training, we used the total read count of Val-GAC-2-1, our lowest documented isodecoder, to define the split for training and validation counts. We split the reads into 80% training, 10% validation, and 10% testing sets for Val-GAC-2-1. For all other tRNA isodecoders, the same number of training and validation reads were assigned to ensure balanced isodecoder representation with all excess

reads being assigned to the test set. In total, this meant that our training set consisted of 1,599,564 training reads and 199,920 validation reads split across each of the 51 *E. coli* tRNA isodecoders.

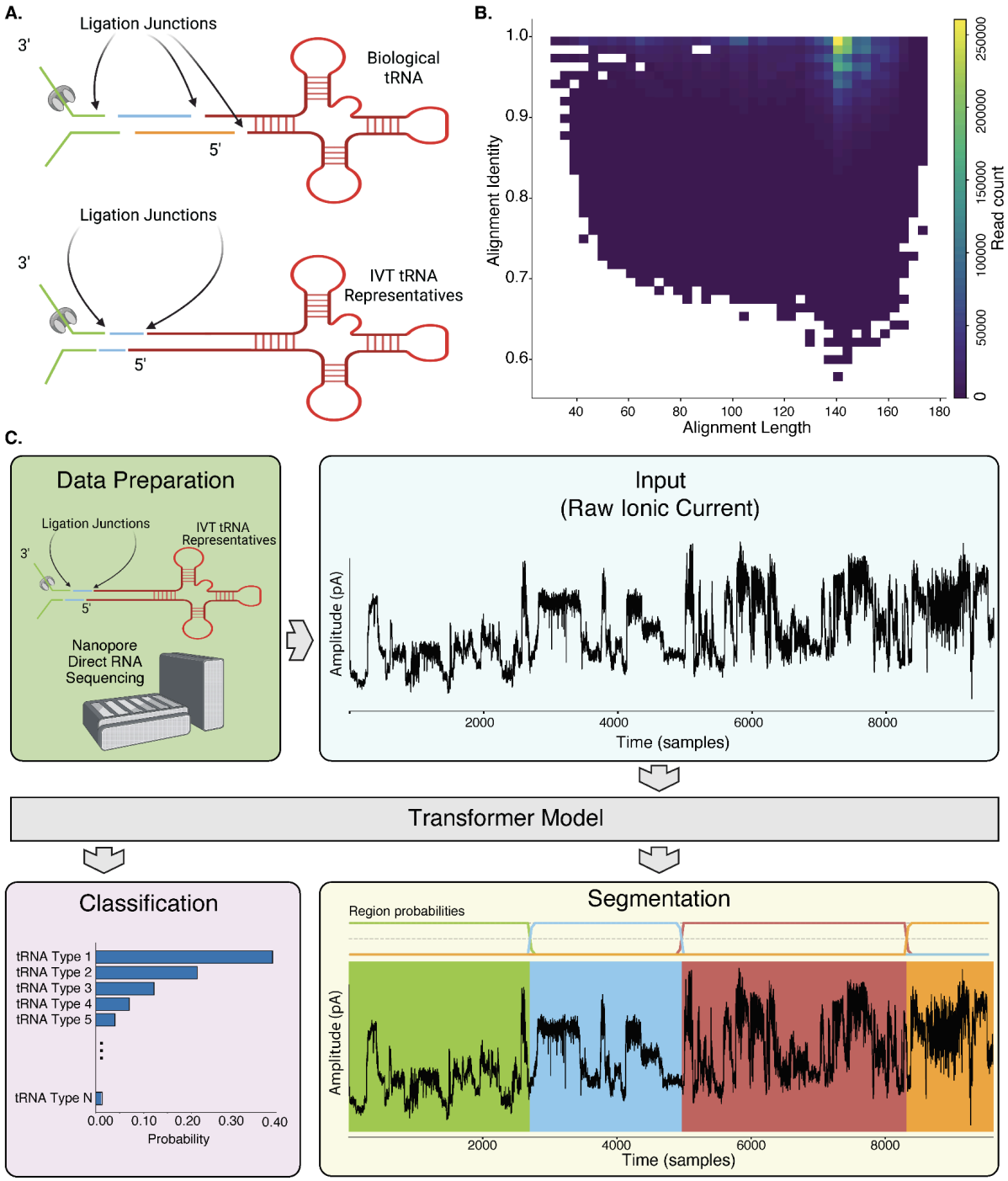

Figure 1. Nanopore adaptation, sequencing, and ionic current analysis for tRNA. **A.** Schematic detailing the relationship between the ligation strategy for biological tRNA and the production of IVT tRNA representatives and their subsequent adaptation and ligation. The final designed sequences for the fully ligated biological tRNA and the IVT tRNA representative are identical to allow for the closest ionic current approximation possible. **B.** Alignment length vs. alignment identity for the IVT tRNA representatives. This analysis only included primary alignments with a mapping quality > 0. Each cell of

the heatmap represents the number of reads that are aligned at the indicated length and the indicated identity. **C.** An overview of the transformer-based end-to-end pipeline for tRNA analysis. The model takes an input ionic current signal (in picoamperes), classifies the corresponding tRNA isodecoder, and segments the signal into biologically relevant regions.

Ionic current model training using IVT *E. coli* tRNA data

We used IVT tRNA Nanopore DRS data to train a transformer-based model. Transformers are a class of neural networks that learn complex patterns in sequential data by selectively focusing on the most informative parts of the input. The testing set for the trained model was composed of independent reads (see Methods). The model used raw ionic current data corresponding to Nanopore readouts of adapted tRNA and had two principal purposes:

i) classification of ionic current traces into tRNA isodecoders; and

ii) segmentation of ionic current traces into the various components of an adapted tRNA. These included the ONT sequencing adapter, tRNA splint 1 (ligated to tRNA 3′-end), tRNA, and tRNA splint 2 (ligated to tRNA 5′-end).

Figure 1C illustrates the overall workflow. By combining the segmentation results with the move table (generated during basecalling by ONT Dorado software), we extracted the tRNA-associated nucleotide sequence. For each tRNA readout, this provided both the predicted tRNA isodecoder (from classification) and the Dorado basecalled sequence (from segmentation and move table), which are required inputs for the downstream alignment algorithm (see below).

We first tested the transformer model on *E. coli* IVT tRNA data **(Figure 2)**. We assessed model performance for both tRNA isodecoder classification and tRNA region segmentation. We used BWA-MEM-generated labels as the reference for this assessment (see Methods). For tRNA isodecoder classification, the model's micro-average-accuracy (overall accuracy across all samples in the dataset) was 95% and macro-average-accuracy (unweighted average accuracy across isodecoders) was 96%. The confusion matrix in **Figure 2A** shows that all row-normalized diagonal values (per-class recalls) fall between 94.5% and 99.7%, indicating that the model was able to learn the ionic current signal from the training data.

For tRNA region segmentation, we measured the overlap between predicted and tRNA regions, identified using BWA-MEM alignments, using Intersection-over-Union (IoU) (see Methods). Briefly, IoU measures overlap between a prediction and corresponding ground truth. The distribution of IoU scores is shown in **Figure 2B**, with the overall mean IoU reaching 0.96. For the set of reads that showed 0 IoU we observed a pattern of reads with extended "stalling" on the nanopore characterized by extensive periods of monotonic current with exaggerated noise profiles. These regions sometimes exceeded the total length of the sequence associated with ionic current. We do not currently have a metric for measuring these stalling periods, or mathematically associating them with the 0 IoU instances. **Figure 2C** illustrates the distribution of classification scores across all isodecoders. The x-axis denotes the BWA-MEM label, and the y-axis shows the model's confidence score for that same isodecoder, computed only over the reads assigned to that isodecoder by BWA-MEM. The color intensity indicates the log-count of samples within each bin. Across all isodecoders, 98.46% of predictions fell within the [0.975–1.0] confidence range, indicating that the model assigned consistently high confidence to the same isodecoders as the labels assigned by BWA-MEM.

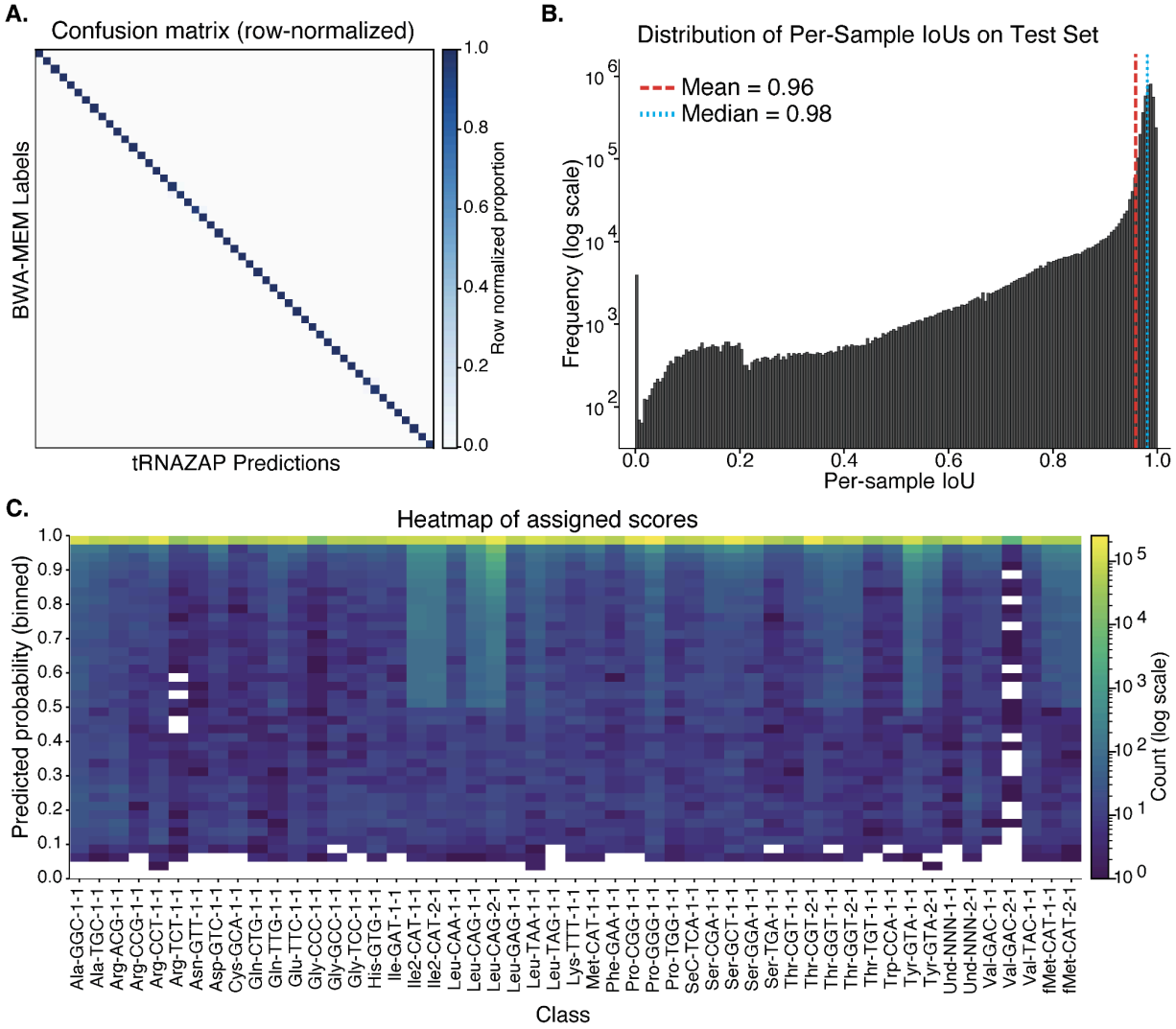

Figure 2. tRNAZAP's inference model performance on the test subset of *E. coli* IVT reads when compared to BWA-MEM based sequence level classification as well as the Dorado move table's signal to sequence mapping output. **A.** Confusion matrix illustrating the correspondence between BWA-MEM-derived reference labels and the model's final classifications. BWA-MEM reference labels were enforced to have a minimum of 75% identity. **B.** Distribution of Intersection-over-Union (IoU) scores between the model-predicted segment ranges and the BWA-MEM assigned labels. The Intersection-over-Union metric allows for the mathematical analysis of overlap between BWA-MEM and move table based segmentation of the tRNA region against the tRNAZAP inference based segmentation of the inference region. **C.** Heatmap depicting the distribution of classification confidence scores for each isodecoders. White cells indicate bins with zero observations (i.e., no reads falling within that confidence range for the corresponding isodecoder). Lower confidence scores indicate that the model was not able to separate a read from one or more alternative isodecoders. The confidence score is the final output layer of the isodecoder prediction component of the model and is normalized by the softmax function.

Model tuning using biological *E. coli* tRNA data

The tRNA ionic current model performed well on IVT tRNA test data. But the training data did not include ionic current corresponding to biological tRNA molecules that are chemically

modified. We reasoned that these biological tRNA data could be used to refine the ionic current model. To that end, we sequenced *E. coli* total tRNA and used BWA-MEM to assign isodecoder-level tRNA labels. For isodecoders with equal lengths, which is all of the *E. coli* isodecoder pairs, we assigned reads to an isodecoder if, and only if, the tRNA was aligned with the isodecoder-specific positions in the reference sequence.

Fine-tuning is the supervised refinement of a previously trained model using additional data. This is intended for a model to adapt to variations that might be absent in training data. We fine-tuned the tRNA ionic current model using data for biological *E. coli* total tRNA, which include biologically derived and chemically modified tRNA. These data were not included in the initial IVT training dataset. After fine-tuning, we observed an increased concordance between BWA-MEM-based reference selection and tRNAZAP classification (denoted as positive change/green on the diagonal of the heatmap; **Figure 3A**), we attributed this mainly to the presence of chemical modifications in biological tRNA that were absent in IVT tRNA. For the test subset of 13,098,894 total reads aligned by BWA-MEM, 665,968 reads that were previously assigned to different isodecoders by BWA-MEM and tRNAZAP, were assigned to the same isodecoder after fine-tuning. In contrast, 18,700 reads that were originally assigned to the same isodecoder by BWA-MEM and tRNAZAP no longer matched following fine-tuning. A total of 11,833,017 reads that were assigned to the same isodecoder by BWA-MEM and tRNAZAP remained assigned to the same isodecoder, and 581,209 reads assigned to different isodecoders remained assigned to different isodecoders after fine-tuning. Overall, these shifts improved all evaluated metrics, raising accuracy by 0.0494, macro-average F1 score (the classwise average of F1 scores) by 0.0974, micro-average F1 score (F1 score when all individual predictions are given equal weight) by 0.0494, and weighted-average F1 score (classwise F1 score weighted by isodecoder proportion in the dataset) by 0.0449. When aggregated by the assigned BWA-MEM label, the classifications shifted towards agreement for all isodecoders except Thr-CGT-2-1 and Und-NNN-1-1 (**Figure 3B**). Both of these isodecoders were under-represented (107 and 24 total reads respectively) compared to the median fine-tuning read count of 190,000. The reasons for poor representation of these isodecoders of tRNA include the tRNA having either low or no expression in the *E. coli* strain being assessed, or the tRNA being a computational prediction but not documented in *E. coli*.

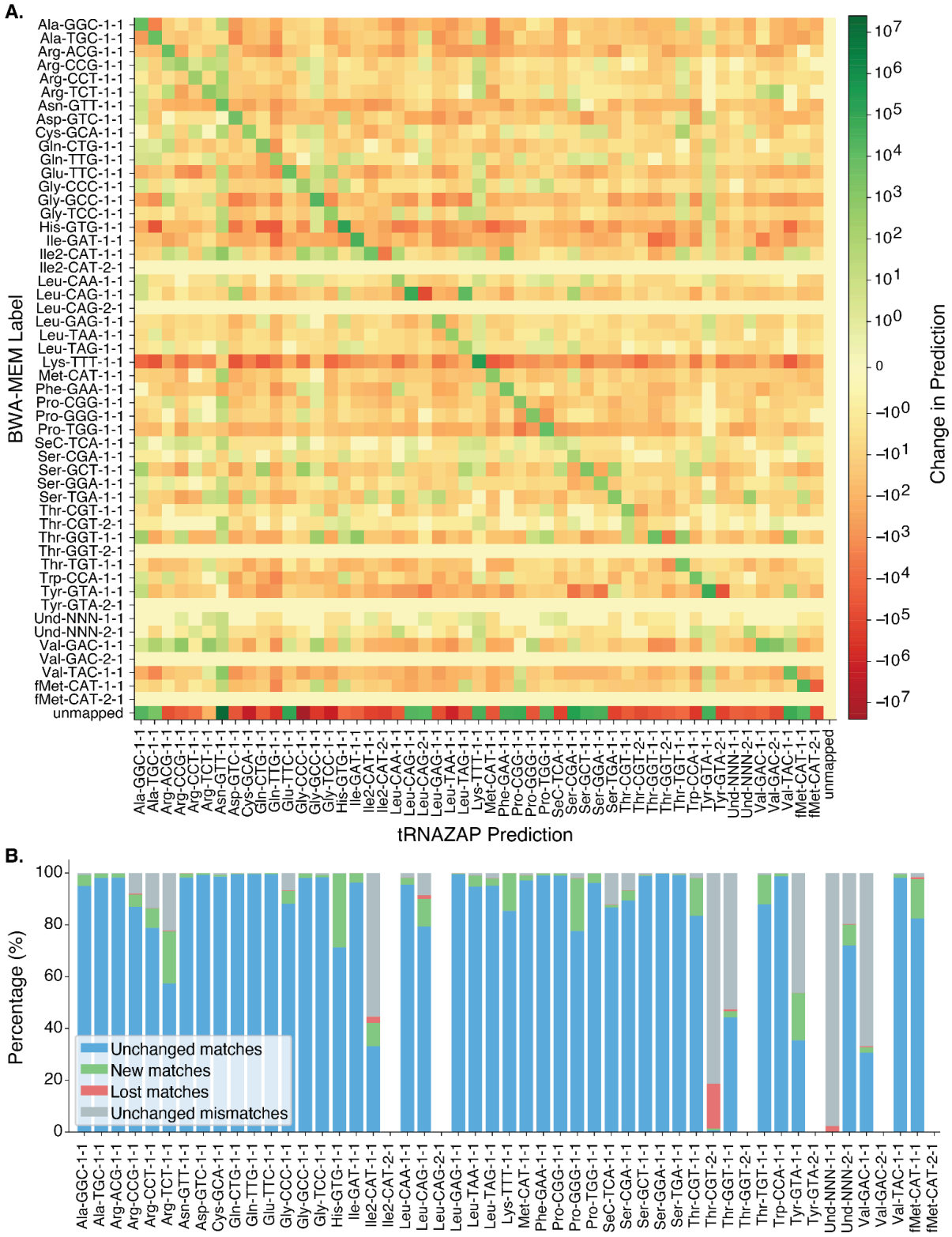

Figure 3. Effect of model fine-tuning on performance for biological data. **A.** Heatmap illustrating changes in tRNA isodecoder assignments following biological fine-tuning. Rows correspond to BWA-MEM assigned labels, while columns represent tRNAZAP model predictions. The color scale indicates the log-count of samples per cell. Increased cell values along the diagonal demonstrates

improved agreement between tRNAZAP predictions and BWA-MEM labels after fine-tuning. The rows composed entirely of 0s come from tRNA isodecoders that are withheld from the reference to accommodate the BWA-MEM alignment scoring system. The only exception is the Isoacceptor Thr-CGT-2-1 which has sufficient sequence level differences from Thr-CGT-1-1 to minimize alignment impact. **B.** Class-wise redistribution of tRNAZAP prediction outcomes following fine-tuning. For each isodecoder defined by the BWA-MEM reference alignment, the stacked bars show the percentage of reads grouped according to how tRNAZAP's predicted isodecoder changes with fine-tuning. Blue indicates reads for which the predicted isodecoder matches the BWA-MEM label both before and after fine-tuning (unchanged matches). Green denotes reads whose predicted isodecoder differs before fine-tuning but matches the BWA-MEM label after fine-tuning (new matches). Red represents reads whose predicted isodecoder matches the BWA-MEM label before fine-tuning but differs after fine-tuning (lost matches). Gray corresponds to reads whose predicted isodecoder differs from the BWA-MEM label both before and after fine-tuning (unchanged mismatches). Percentages are computed independently within each BWA-MEM-defined isodecoder and normalized to 100% for each bar.

An ionic current guided pairwise sequence alignment strategy

Nanopore tRNA sequence analysis has historically relied on using BWA-MEM for alignment, a seed and chain strategy. Seed and chain is an alignment strategy that first identifies short sequences of exact matches between a read and a set of candidate reference sequences, then extends outward while chaining together short lengths of sequence until the read is either fully aligned or the quality of continued chaining falls below early termination thresholds. One limitation of this approach is that alignments are performed at isoacceptor level instead of isodecoder level. This is because BWA-MEM is an aligner designed for short reads with high sequencing accuracy (99.9%+).

In addition to the isodecoder-level classification for a given tRNA, the tRNAZAP inference model also predicted which section of ionic current signal corresponds to the tRNA. The emit-moves option of Dorado basecalling can then be used to map the predicted signal to a section of basecalled sequence. We developed a custom strategy to align this section of the basecalled sequence with the corresponding reference for the predicted tRNA isodecoder. We used a custom pairwise alignment implementation strategy for this purpose. In our implementation, the alignments were forced to account for the entire length of tRNA reference, but the read sequence is not penalized for bases that extend beyond the tRNA reference. This is similar to soft clipping in other alignment strategies, an acknowledgement that sequencing technologies can produce inaccuracies at the start and end of predicted sequences. tRNAZAP aligned each read to the predicted reference isodecoders sequence. Because tRNAZAP predicted the tRNA isodecoder, including isodecoder level discrimination, at the ionic current level only one pairwise alignment was required.

While this alignment strategy worked well for full-length Nanopore tRNA readouts, we reasoned if it could also "rescue" truncated tRNA strand reads. These truncations could be caused due to incomplete tRNA ligation, software issues; or they could represent bonafide tRNA fragments. To that end, we applied a final alignment modality where the optimal local alignment was calculated between the read sequence and the top three predicted tRNA isodecoder sequences using a Smith-Waterman-based algorithm. This modality was only used if the primary isodecoder did not yield a successful alignment. For biological applications we selected an identity cut off threshold of 75% and we required a minimum of 25 matching bases in the tRNA region of the reference.

The length of 25 was selected to accommodate the potential for tRNA fragments split at the anticodon loop[28] which accounts for the characteristic loss of the last ~10-12 nucleotides of 5' sequence when using the nanopore sequencing[29].

Assessing tRNAZAP performance for biological *E. coli* tRNA

We next compared tRNAZAP performance using total tRNA from wild-type *E. coli* MRE600 strain. Nanopore DRS produced a total of 51,034,636 reads. We first filtered these data to exclude 280,747 reads that had undergone read splitting by Dorado (denoted by the 'pi' tag). Read splitting happens when the ONT software (MinKNOW) artificially fuses two reads into one, and records them as a single read. Dorado has the ability to correctly split them as two independent molecules at the read sequence level. Because this does not happen at the ionic current level, we exclude such molecules from our analysis. This filtering reduced this down to 50,753,889 reads. Among these, BWA-MEM aligned 19,079,732 (37.59%) reads while tRNAZAP alignment strategy aligned 19,590,521 reads (38.60%). This small difference of 510,879 reads was still a statistically significant increase (2.68% increase when comparing tRNAZAP to BWA-MEM; χ^2=24,124.58, $p<1x10^{-308}$, **Figure 4A**).

A comparison of identity for all aligned reads between the two alignment methods showed a mean alignment identity of 88.97% and 89.15% for BWA-MEM and tRNAZAP respectively: a mean difference of 0.18% (Mann-Whitney U = $1.84x10^{14}$, $pp<1x10^{-308}$). For the set of reads where both aligners had a successful alignment BWA-MEM had a mean alignment identity of 89.01% and tRNAZAP had a mean alignment identity of 89.37%. This 0.36% mean identity increase for tRNAZAP in comparison to BWA-MEM was statistically significant (paired Wilcoxon Rank Sum test; W = $9.05x10^{12}$, $p<1x10^{-308}$, **Figure 4B**). Such minor increases in alignment identity likely offer minimal practical advantage for aligning *E. coli* tRNA.

However, for the set of reads where the two aligners selected different primary tRNA isodecoders BWA-MEM had a mean alignment identity of 84.48% and tRNAZAP had a mean alignment identity of 88.45%. This difference of 3.97% alignment identity in favor of tRNAZAP. This difference was statistically significant (paired Wilcoxon Rank Sum test; W = $2.43x10^{11}$, $p<1x10^{-308}$, **Figure 4C**). The confusion matrix of final alignments between BWA-MEM and tRNAZAP shows the two aligners agreeing on the best isodecoder for 18,167,400 reads and disagreeing on 780,090 reads (4.12% of reads where both aligners found a satisfactory alignment). tRNAZAP and BWA-MEM agreed on the primary tRNA assignment for 94.88% of reads that were aligned by both aligners (**Figure 4D**).

These results indicate that tRNAZAP aligned more reads at an identity threshold at least as high as BWA-MEM and for the reads where the two aligners disagreed on the primary tRNA isodecoder, tRNAZAP produced a significantly higher alignment identity. This means that despite the inclusion of isodecoders in the possible reference set tRNAZAP aligned more reads at a higher alignment identity than BWA-MEM. This level of isodecoder resolution had not been previously achieved with BWA-MEM and Nanopore DRS for *E. coli* tRNA.

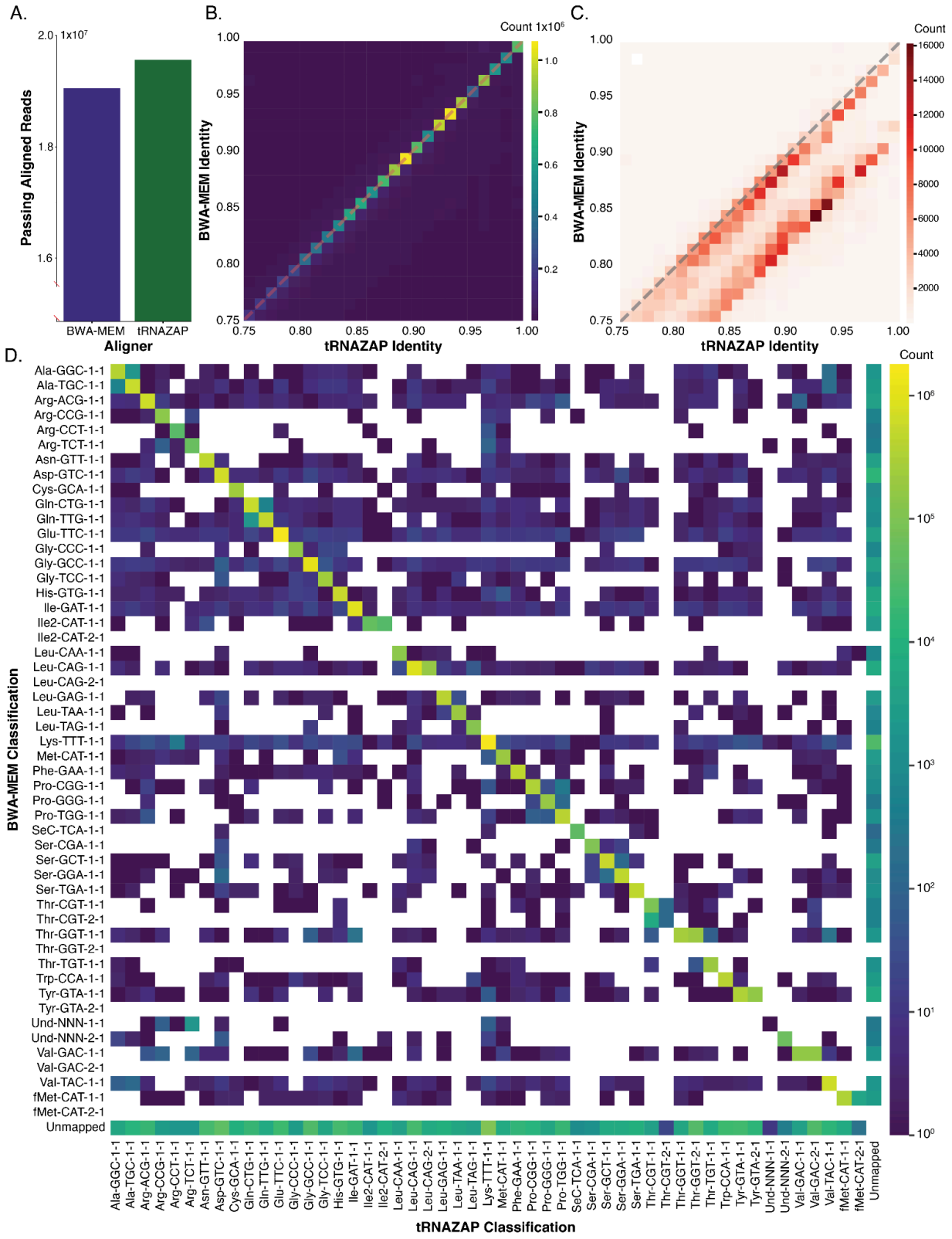

Figure 4. tRNAZAP performance for *E. coli* biological tRNA compared with BWA-MEM **A.** Count of reads (y-axis) that achieved a minimum of 25 matching bases in the tRNA region and a minimum of 75% alignment identity for tRNAZAP and BWA-MEM (x-axis). **B.** Per read alignment identity for the tRNA portion of the reference compared at a paired read level between tRNAZAP and BWA-MEM.

Colors represent the raw count of reads per bin. **C.** Per read alignment identity compared between tRNAZAP and BWA-MEM for reads where the read aligned to different primary tRNA. Colors represent the raw number of reads per bin. **D.** Read assignments based on the tRNAZAP alignment algorithm (x-axis) compared to the BWA-MEM alignment algorithm (y-axis). Colors represent a log transform of the count of reads per bin. The cell representing the condition where both aligners failed to align a read was artificially set to 0 to increase the granularity of the color scale.

Applying tRNAZAP on a biological *E. coli* tRNA sample enriched for $tRNA^{Phe}$

It is possible for deep learning models to memorize the distribution of training classes, which we anticipated might be reflected in Nanopore tRNA data from wild-type *E. coli* strains cultured under similar growth conditions. However, this static distribution of tRNA isodecoders does not reflect the extent of tRNA experimental conditions. There are several enrichment methods for enriching a specific set of populations or even individual tRNA populations[30,31]. Assessing these enrichments is critical for a broadly applicable Nanopore-based tRNA analysis strategy. We tested if tRNAZAP was robust for samples that had different underlying distribution of tRNA compared to their conventional wild-type counterparts. To that end, we performed an HPLC purification designed to enrich *E. coli* tRNA Phe-GAA-1-1 and sequenced the associated fraction. We documented a Fold Change (FC) of 17.01 transcripts per million (TPM) for Phe-GAA-1-1 between the total tRNA condition and the purified condition (**Figure 5A**, **Supplementary Table 1**). When we examined the normalized read counts between the two conditions we observed the expected enrichment of Phe-GAA-1-1, but also a co-enrichment of Glu-TTC-1-1 (2.59 FC of TPMs), Ser-GCT-1-1 (4.78 FC of TPMs), and Ser-GGA-1-1 (3.47 FC of TPMs). Visual inspection of the enriched alignments showed a decrease in positional base mismatch profiles we have previously associated with a pseudouridine modification at tRNA-Phe-GAA-1-1 position 55 (**Figure 5B.**).

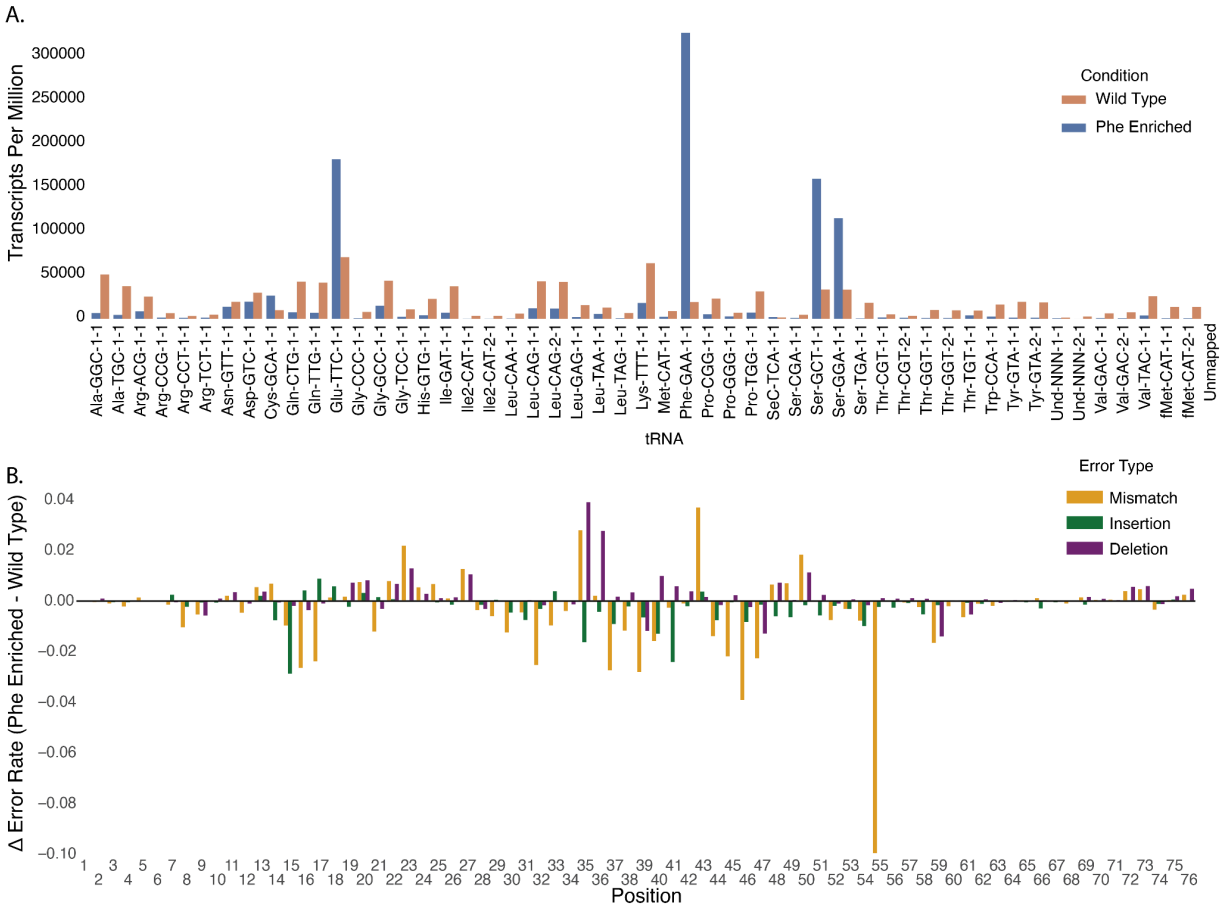

Figure 5. tRNAZAP performance on wild-type and tRNA^{Phe} enriched *E. coli* tRNA samples. **A.** Transcripts per million (TPM) for *E. coli* tRNAs for wild-type and for tRNA-Phe-GAA-1-1 enriched samples. **B.** Changes in positional error profiles for tRNA-Phe-GAA-1-1. Each error type is colored separately and insertions are attributed to the 5′ position relative to the insertion.

Training tRNAZAP for a model eukaryote

Eukaryotic tRNAs contain more chemical modifications than *E. coli,* as profiled in previous studies[32]. To test our tRNAZAP strategy for eukaryotic tRNA, we trained a different model for the budding yeast *Saccharomyces cerevisiae* (*S. cerevisiae*). The increased complexity in modifications impacted basecalling quality and alignment identity. For the biological *E. coli* tRNA dataset, the BWA-MEM median alignment identity was 89.47%. In contrast, for biological *S. cerevisiae* tRNA, the BWA-MEM median alignment identity was 82.35%. Additionally while the *E. coli* model had to account for isodecoders arising from 7 isoacceptors, the *S. cerevisiae* reference had isodecoders arising from 12 isoacceptors. One of these isoacceptors had three isodecoders. We first generated an IVT tRNA-based model for *S. cerevisiae* using the same strategy that we had employed for *E. coli*.

The IVT-trained model for *S. cerevisiae* showed similar performance (98.83% macro-average-accuracy (unweighted average accuracy across isodecoders); 0.98 median IoU). The heatmap of classification scores per tRNA showed reduced confidence for sets of isodecoders that share an isoacceptor, which was expected considering as low as a single nucleotide difference between some of the isodecoders. To resolve this, we fine-tuned the

model using data from three separate *S. cerevisiae* biological tRNA sequencing datasets. *S. cerevisiae* has two pairs of isodecoders (Asp-GTC-1-1 vs. Asp GTC-2-1 and Ser-AGA-1-1 vs. Ser-AGA-2-1) with different sequence lengths making direct nucleotide comparison uninformative. For these sets of isodecoders we assigned fine tuning labels based on an edit distance metric between the read sequence and the two isodecoder references.

We tested the fine-tuned *S. cerevisiae* model on data from a fully withheld sequencing experiment, this means that the model had not trained on any of the data from the withheld sequencing experiment. This allowed us to test both the traditional performance of the model and the impact of technical variation of ionic current on the model. We used the same filtering thresholds for our comparison of BWA-MEM and tRNAZAP for our *S. cerevisiae* model as we did for the *E. coli* comparison. The *S. cerevisiae* sequencing run produced a total of 34,133,760 reads of which we filtered to 34,012,875 after removing reads which underwent read splitting. Of these reads BWA-MEM produced 8,994,290 alignments that passed filtering thresholds (26.44%) and tRNAZAP aligned 10,173,310 reads under the same filtering thresholds (30.19%). This amounts to a total increase of 1,179,020 aligned reads, resulting in a 13.11% increase in total aligned reads (**Figure 7A**). This is a statistically significant increase in aligned reads (χ^2 = 119551.89, $p<1x10^{-308}$).

To compare the global alignment identity between tRNAZAP and BWA-MEM we first considered the set of all reads that were aligned by both aligners. BWA-MEM had a mean alignment identity of 82.94% and tRNAZAP had a mean alignment identity of 83.14%. This 0.20% higher mean alignment identity for tRNAZAP was statistically significant (Mann-Whitney $U=4.66x10^{13}$, $p<1x10^{-308}$). However, to understand the comparison at a per read level we explained only reads that were aligned by both aligners. For reads where both aligners had a successful alignment we documented a mean alignment identity of 82.95% and 83.44% for BWA-MEM and tRNAZAP respectively. We performed a paired Wilcoxon signed rank test that confirmed that the observed 0.49% median identity increase for the shared aligned reads was significant ($W=6.31x10^{12}$, $p<1x10^{-308}$, **Figure 7B)**. This led to the question of how well reads that were assigned to different isodecoders compared between the two isodecoders. For this set of reads BWA-MEM had a mean alignment identity of 80.92% and tRNAZAP had a mean alignment identity of 82.40%. This difference in mean identity of 1.48% in favor of tRNAZAP was significantly higher ($W=7.25x10^{10}$, $p<1x10^{-308}$, **Figure 7C**). tRNAZAP and BWA-MEM agreed on the primary tRNA assignment for 90.96% of reads that were aligned by both aligners (**Figure 7D**).

Interestingly, higher complexity increased the comparative performance of tRNAZAP showing a 2.68% increase in aligned reads for *E. coli* and a 13.11% increase in aligned reads for *S. cerevisiae* when compared to BWA-MEM. This demonstrates that the ionic current informed pairwise alignment strategy of tRNAZAP is superior to the seed and chain strategy of BWA-MEM, especially for eukaryotes. It is also important to note that this increase further facilitates analyses of more tRNA strand reads per experiment, thus improving usable throughput for Nanopore direct tRNA sequencing. An added benefit of this higher identity alignment strategy is that it can also be used for alignment-guided modification analysis strategies that are being employed by the community for tRNA.

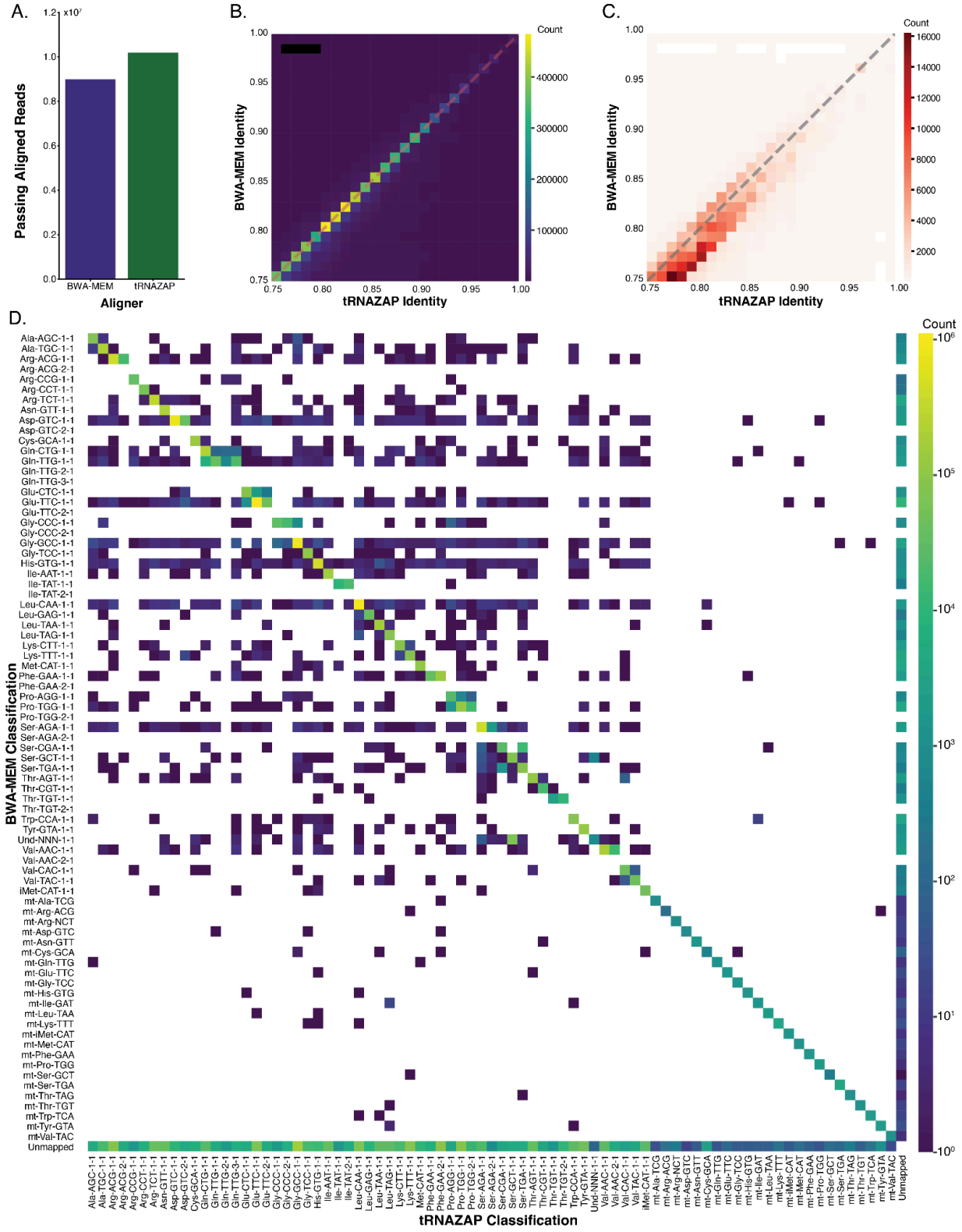

Figure 6. tRNAZAP performance for *S. cerevisiae* tRNA nanopore sequencing data **A.** Count of reads (y-axis) that achieved a minimum of 25 matching bases in the tRNA region and a minimum of 75% alignment identity for tRNAZAP and BWA-MEM (x-axis) for *S. cerevisiae* biological tRNAs. **B.** Per read alignment identity for the tRNA portion of the reference compared at a paired read level between

tRNAZAP and BWA-MEM. Colors represent the raw count of reads per bin. **C.** Per read alignment identity compared between tRNAZAP and BWA-MEM for reads where the read aligned to different primary tRNA. Colors represent the raw number of reads per bin. **D.** Read assignments based on the tRNAZAP alignment algorithm (x-axis) compared to the BWA-MEM alignment algorithm (y-axis). Colors represent a log transform of the count of reads per bin. The cell representing the condition where both aligners failed to align a read was artificially set to 0 to increase the granularity of the color scale.

Applying tRNAZAP to biological *S. cerevisiae* tRNA

Nearly all eukaryotic organisms harbor tRNA genes in two different genomes: the nuclear genome and the mitochondrial genome. When *S. cerevisiae* are grown in glucose-based media fermentation permits growth in the absence of mitochondrial respiration and so they tend to have many fewer mitochondria and thus much less abundant mitochondrial tRNA than nuclear tRNAs[33]. The abundance of mitochondrial tRNA can be increased in a sample by using different growth media and an enrichment of the mitochondrial fraction[30,34–36]. We wanted to test if tRNAZAP reflected the enrichment of an entire subpopulation of tRNA simultaneously. To that end, we sequenced tRNA from a mitochondria-enriched tRNA sample (MT+, see methods). We compared the MT+ condition with the glucose growth condition (MT-) and documented a 4.1-to-8.7 Log_2 FC for the 24 *S. cerevisiae* mitochondrial tRNA isodecoders, demonstrating tRNAZAP's ability to capture subpopulation enrichment (**Figure 7A**). We noted that the median alignment identity for mitochondrial tRNA from the enriched condition was equal to or greater than mitochondrial tRNA from the yeast grown in glucose across all isodecoders (**Supplementary Figure 1, Supplementary Table 2**).

tRNAZAP resolves isodecoders in *E. coli* and *S. cerevisiae* data

Isodecoder-level resolution of aligned tRNA read sequences has been challenging for Nanopore DRS[12,13,37]. This is due to the limited number of positional differences between isodecoders and impact of modifications on sequencing accuracy. Prior nanopore tRNA analysis approaches, including our work, focused on isoacceptor-level analyses using curated reference sequences[6,11–13]. This loss of information oversimplifies tRNA biology and obscures disease relevant discoveries. For instance, it has been shown that a single base mutation in the T-loop of mouse tRNA-Arg-UCU isodecoder encoded by *n-Tr-20* can cause fatal neurodegeneration in mice[38]. Analysis strategies that resolve isodecoders at a single-molecule level will advance tRNA research. We reasoned that tRNAZAP could help achieve this because it does not need to predict the entire sequence context, rather it only needs to identify regions of differential ionic current between isodecoders. We first constructed a training set for isodecoder-level fine-tuning by aligning reads to a reference set of isoacceptors. For each isoacceptor we identified all of the positional mismatches between associated isodecoders and assigned reads to the appropriate isodecoder if, and only if, the read had the expected nucleotide for a given isodecoder at each relevant position. We acknowledge that this technique has some risk of introducing a BWA-MEM based alignment bias into model training. An alternative approach, beyond the scope of the current study, would be purification of isodecoders for the same isoacceptor followed by model training, which would be experimentally challenging[31].

To investigate model performance, we compared ionic current profiles between isodecoder pairs. In some cases, ionic current visualization showed a separation of populations (**Figure 7B.i**). However, this is not always expected due to the stochastic nature of ionic current data. We aggregated ionic current traces and estimated the probability density function for both the

median and standard deviation for ionic current at each position. These kernel density estimates show a similar trend to the raw ionic current, but at an aggregate level (**Figure 7B.ii**) The combination of these two metrics allowed us to model the ionic current, showing a separation (0.59 Jensen-Shannon (JS) divergence, **Figure 7B.iii**). A two-dimensional interpretation of ionic current data is an oversimplification, failing to account for the temporal interactions between data points. The inference model considers not only these baseline statistics (visualized here) but also more complex interactions between datapoints. When we compared the Gaussian mixture models across all isoacceptors with isodecoders the JS divergence ranged from 0.36 to 0.98 (**Supplementary Figures 2 and 3**, **Supplementary Table 3**). The isodecoder pair Ser-AGA-1-1 and Ser-AGA-2-1 had the lowest JS divergence for *S. cerevisiae* at 0.42. This can possibly be attributed to the nature of the sequence level divergence between the two isodecoders: Ser-AGA-2-1 has a dinucleotide GG spanning consecutive Sprinzl coordinates V4,V11 where Ser-AGA-1-1 lacks the V4 position entirely, meaning it only has a G mononucleotide at V11. This means that the entire sequence difference between the two isodecoders is GG for Ser-AGA-2-1 and G for Ser-AGA-1-1[37,39]. For *E. coli* the fMet-CAT-1-1 and fMet-CAT-2-1 had the lowest JSD at 0.36. fMet-CAT-1-1 and fMet-CAT-2-1 have a Guanosine and Adenosine at position 46 respectively. According to MODOMICS fMet-CAT-1-1 can have a m^7G at position 46, potentially further complicating differentiating the two isodecoders.

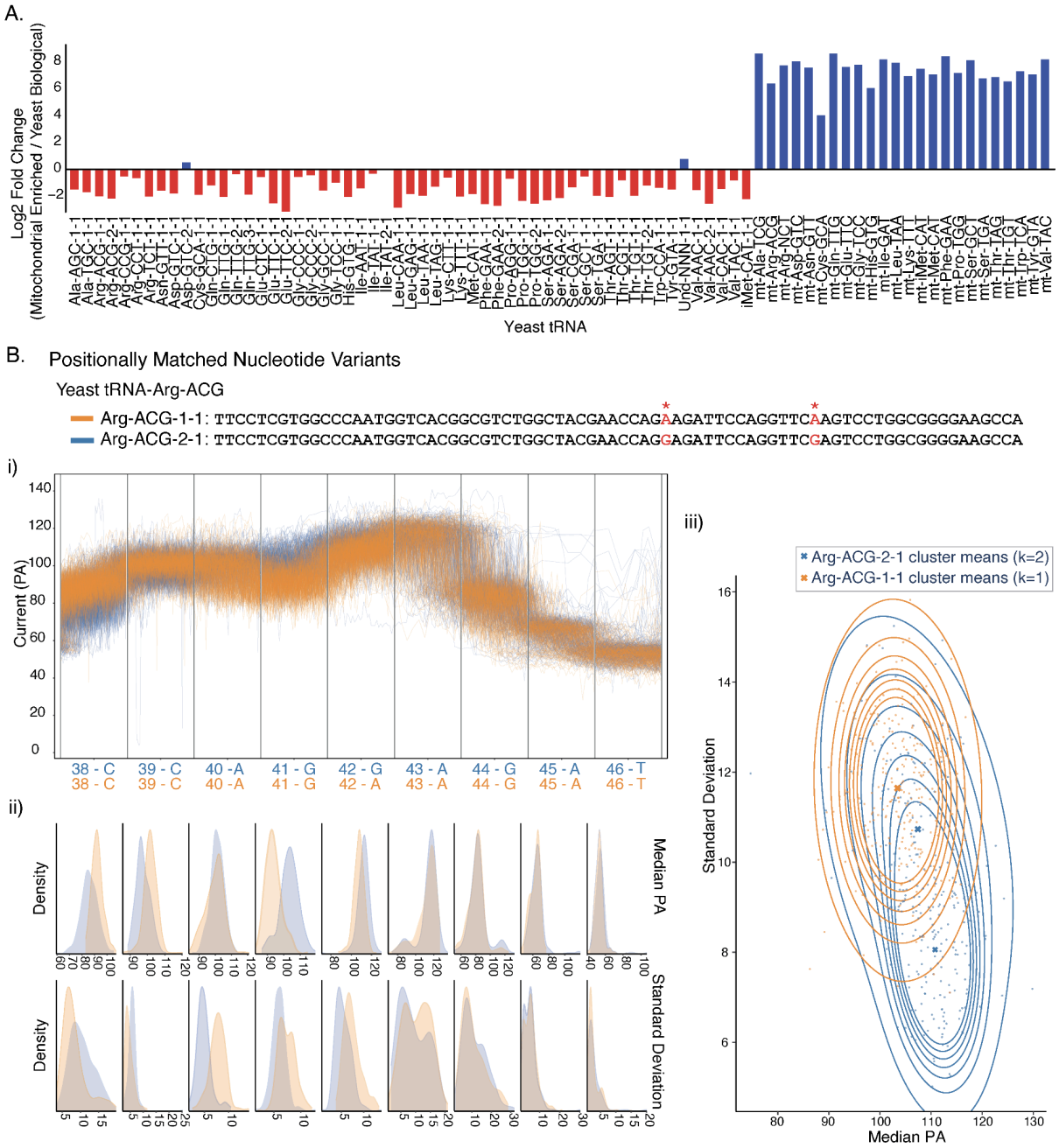

Figure 7. A. Log_2 Fold Change between the TPMs of the mitochondrial-enriched condition and the non-enriched condition. The mitochondrial enrichment condition is the numerator and wild wild-type condition is the denominator when calculating Log_2 Fold Change. **B. i.)** Raw Ionic current traces for two isodecoders of *S. cerevisiae* tRNA Arg-ACG (250 reads each) identified by tRNAzap inference and aligned using tRNAzap align. Position 42 is one of two single-nucleotide variants that define the two isodecoder populations. **ii.)** Kernel Density Estimates of the Probability Density Functions for the median and standard deviation of ionic current picoamperes for each position in the 9 nucleotides spanning one of the single nucleotide variants. **iii.)** Gaussian Mixture Model of median and standard deviation for 3-mer window (position 42 in middle) visualizing the compounding separating power of multiple summary statistics for isodecoder discrimination. The number of components was empirically calculated by selecting the number of components, Isomass centers denoted that minimized the

Bayesian information criterion (BIC).

tRNAZAP for nucleotide modification analysis

In previous work, we and others investigated the modification circuit of the T-loop in *S. cerevisiae* tRNA[6,12,40]. We combined mass-spectrometry, genetic controls, and direct RNA sequencing to analyze the inter-dependence of T-loop modifications. Because the tRNAZAP model is trained using wild-type *S. cerevisiae* Nanopore tRNA data, we were curious to see if we saw a decrease in accuracy on Nanopore tRNA data from a genetically altered strain of *S. cerevisiae* that affected a widespread chemical modification. Thus we sequenced total tRNA using RNA004 from a *pus*4Δ strain, lacking the pseudouridine synthase Pus4 that modifies U55 on nearly all *S. cerevisiae* tRNAs[6] As for wild-type *S. cerevisiae* tRNA, tRNAZAP aligned more reads than did BWA-MEM alignment for the *pus*4Δ strain. And there was a significantly higher identity amongst reads aligned by both aligners; for reads where the tRNA isodecoder disagreed, tRNAZAP had a higher identity (**Supplementary Figure 4**).

This increase in data coupled with higher alignment identity lends itself to investigations of ionic current perturbations caused by putative modifications. While miscalls have provided suggestive evidence in tRNA modification analysis using Nanopore DRS, it is important to adopt principled approaches using raw ionic current. To qualitatively recapitulate the miscall-based findings of our previous work, we investigated the ionic current profiles for the T-loop in *S. cerevisiae* tRNA. We examined Cys-CGA-1-1 because it has been shown to bear a 5-methyluridine (m^5U) at position 54, pseudouridine (Ψ) at position 55, and a 1-methyladenosine (m^1A) at position 58 in our previous work. This cluster of modifications produced a distinct miscall pattern. When we performed an enzymatic knockout of Pus4 previously, we observed that the miscall signal for both Ψ55 and m^1A58 reduced significantly[6]. Nucleoside MS has previously confirmed that all three modifications in the Cys-CGA-1-1 T-loop are reduced when Pus4 is knocked out. This work was performed with RNA002, an older Nanopore DRS platform that had lower throughput and lower basecalling accuracy. Using the RNA004 DRS platform, we examined raw ionic current traces for Cys-CGA-1-1 tRNA T-loop region from the wild type condition, an IVT condition, and the Pus4 knockout condition (**Figure 8A**). We noted two distinct populations in the ionic current in the vicinity of the modification-bearing positions in the T-loop. We then aggregated the ionic current data and estimated the probability density functions (PDF) for the median picoamperes (pA), standard deviation, and duration of each event (**Figure 8B**). These PDFs show two distinct ionic current level populations. The IVT and *pus*4Δ ionic current traces show high levels of concordance throughout the region, this follows the documented modification circuit that would suggest without Ψ55, m^1A58 and m^5U54 are absent or reduced in abundance. The wild type condition has some reads that follow the same ionic current pattern observed in the IVT and *pus*4Δ populations suggesting a similar modification profile. But additionally the wild type condition has a second population of reads that produce both a unique ionic current pattern observed in **Figure 8A** positions 53-57 and distinct lower peaks in median picoamperes in **Figure 8B** positions 53-57. A possible explanation for this is the wild type specific patterns of ionic current are for transcripts that have the complete or partial set of tRNA Cys-GCA-1-1 T-loop modifications: m^5U54, Ψ55, and m^1A58. It is important to note that the nanopore sensor measures approximately 5 consecutive nucleotides in an RNA strand at a time, moving in a one nucleotide stepping pattern. As a result, the influence of a nucleotide

modification can extend beyond the nucleotide position. In the wild type condition we observed approximately 5 positions of impacted signal, a likely explanation for this is the combination of ionic current impacts from the trio of modifications m^5U54, Ψ55, and m^1A58 spread across the observed window of positions. The separation of populations within the wild-type condition suggests that with proper training, controls, and orthogonal validation ionic current analysis may provide a sensitive medium for documenting the stoichiometry of modifications in tRNA.

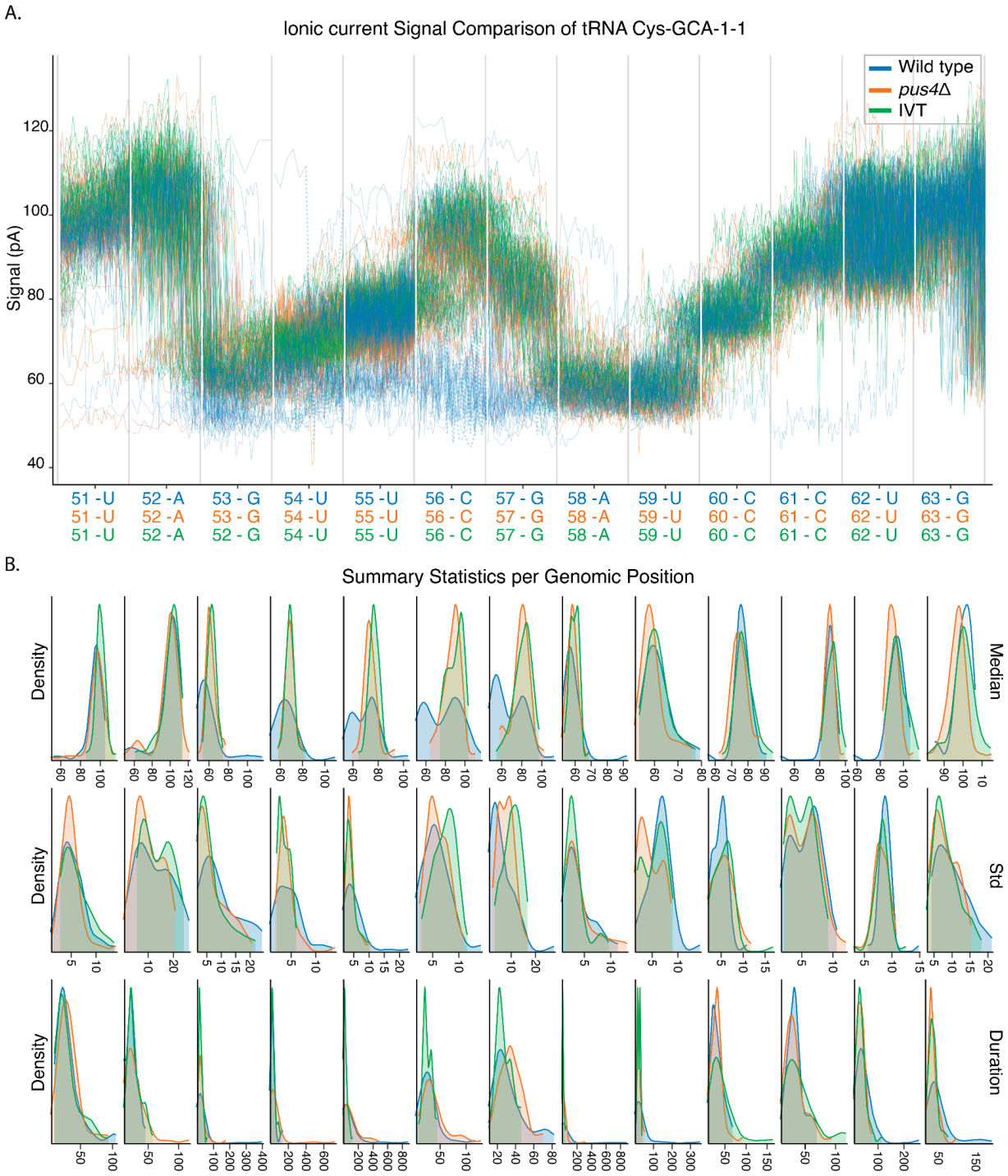

Figure 8. Ionic current signal comparison of tRNA Cys-GCA across conditions with *pus4*Δ. A.

Representative ionic current traces for tRNA Cys-GCA across three experimental conditions: Wild type (blue), *pus4*Δ (orange), and IVT (green), each with ~50 reads. Signals are displayed in picoamperes (pA) and tRNA sequence positions (51-63). Traces are visualized based on the Dorado move table (a signal-to-read alignment that maps blocks of raw signal timepoints to basecalled nucleotides. The x-axis indicates genomic position and corresponding nucleotide identity). **B.** Summary statistics per position showing distributions of signal characteristics across the three conditions. Each column represents a nucleotide position with three statistical measures displayed as density plots: median signal intensity (Median, top row), standard deviation of signal intensity (Std, second row), and signal dwell time in timepoints (Duration, bottom row). Density curves are color-coded as in panel A, with shaded areas representing the distribution of per-read values for each condition at each position.

CONCLUSION

In this study, we analyzed all tRNA isodecoders from *E. coli* and *S. cerevisiae* using Nanopore direct tRNA sequencing. Our strategy included training ionic current models for tRNA isodecoder classification. Additionally, these models performed ionic current segmentation of tRNA to identify associated basecall sequence. We then implemented a custom pairwise alignment strategy that aligned this tRNA isodecoder sequence to the expected reference sequence. We applied these models to demonstrate successful classification of isodecoders for *E. coli* and *S. cerevisiae*. We successfully applied the model to enriched tRNA samples ($tRNA^{Phe}$ in *E. coli* and mitochondrial tRNA in *S. cerevisiae*). Additionally, we also documented that the model could also identify isodecoders for a modification knockout strain (e.g. *pus*4Δ where the Ψ55 is knocked out) of *S. cerevisiae*.

The present implementation of Nanopore direct tRNA sequencing requires ligation of splint adapters that anneal to the conserved NCCA tRNA overhang. It is important to note that we designed new splint adapters for the tRNAZAP strategy. This was done to aid sequencing performance and ionic current analysis. As others implement the tRNAZAP ionic current classification strategy, they should be aware that the newer splint adapters must be used. This is because transformer-based deep learning architectures, especially those with attention mechanisms, consider the entire ionic current readout (including the adapters). Nanopore tRNA data generated using other splint adapters could behave unpredictably for ionic current classification and for segmentation. However, we anticipate that the benefits of the tRNAZAP strategy will encourage the community to adopt these new adapter strategies for tRNA ionic current analysis.

Going forward, we anticipate that raw ionic current will be a key component of the principled approaches to study tRNA isodecoders and modifications. Nanopore DRS is poised to play an essential role for studying the landscape of tRNA modifications (e.g., in tRNA loops and stems). The success of this strategy for *E. coli* and *S. cerevisiae* is predictive of its potential for more complex tRNA (e.g., human). It is also important to validate DRS-derived results regarding modifications using orthogonal techniques such as LC-MS and other molecular methods, in addition to the continued use of genetic controls (e.g., knockouts or knockdowns). The Nanopore-based single-molecule, ionic-current readouts of tRNA are information rich. tRNAZAP, which is available for use by the community, leverages these rich ionic current readouts to achieve more aligned reads than BWA-MEM, and with higher identity. Additionally, tRNAZAP provides isodecoder-level resolution for tRNA as demonstrated for a model prokaryote and eukaryote in this study. Nanopore ionic current-based analysis tools will play a foundational role

for single-molecule, native tRNA strand analysis. Combined with orthogonal methods, this will reveal unprecedented insights into tRNA biology and their role in human health and disease.

METHODS

All of the adapter oligonucleotides, IVT oligonucleotides and DNA ultramer pools used in these studies were manufactured by IDT. The adapter oligonucleotides were HPLC purified,

RNA samples The biological *E. coli* tRNA sample sequenced was from MRE600 total tRNA purchased from Roche Pharmaceuticals. The *S. cerevisiae* samples were from strain BY4741 collected during log phase growth.

Enrichment of tRNA-Phe from *E. coli* total RNA: Wild-type E. coli K-12 BW25113 were transformed with a pUC57 plasmid containing a lac-inducable tRNA-Phe-GAA-1-1 gene for overexpression, then struck onto LB-amp plates and grown overnight at 37°C. Individual colonies were selected and grown to saturation in LB with ampicillin. An expression culture was inoculated at a 5:1000 ratio in enriched TB with ampicillin, 0.05% glucose, and 0.2% lactose for autoinduction, and then grown for 16 hours before harvesting in a centrifuge. Pseudopurified tRNA phe was generated as previously described[41]. Briefly, cells were lysed and total RNA was isolated using a phenol chloroform extraction. Tris-acetate pH 8 was used to deacylate the tRNA in the sample and then total tRNA was isolated using a Cytiva Resource Q anion exchange column on an FPLC. This total tRNA was ethanol precipitated and a Waters XBridge BEH C18 OBD prep HPLC column was used for pseudopurification. Fractions containing high tRNA phe concentrations were identified using a test charging with radiolabeled phenylalanine and pooled.

***S. cerevisiae* wild type growth conditions:** Wild-type BY4741 *S. cerevisiae* were struck out from glycerol stocks onto YPD plates and grown for three days at 30°C. Single colonies were inoculated in 3mL liquid YPD and grown overnight at 30°C, after which 400uL of overnight culture were diluted into 5.2mL liquid YPD and grown at 30°C to an OD of 0.5-1.0. Cell pellets were washed with 1X PBS and flash frozen in liquid nitrogen. Total RNA was extracted using the hot phenol method as previously described[6].

***S. cerevisiae* mitochondrial growth conditions and mitochondrial enrichment:** Wild-type BY4741 *S. cerevisiae* were grown in 1 L YPE (Yeast extract, Peptone, 2% Ethanol) and flash frozen in liquid nitrogen at an OD600 of 1.0-1.4. Differential centrifugation was used to isolate the crude mitochondrial fraction from *S. cerevisiae* cells as previously described[30]. RNA was isolated from the crude mitochondrial fraction using the hot phenol-chloroform method.

Biological tRNA adapter augmentation for deep learning: Previous work by our group and others share the common theme of adapting the tRNA by ligating 5' and 3' adapter strands to the conserved NCCA tRNA 3' overhangs[12,14]. To facilitate model learning we amended our original adapter lengths to have approximately equal lengths of sequence bearing constructive ionic current information. Our final constructs had a 5' adapter length of 36 and 3' length of 60, of which 10 are poly(A) and another 18 are the DNA nucleotides required for the customized RTA ligation. Here we consider the 32 nucleotides before the poly(A) to approximately equal in length to the 5' adapter. The 5' adapter was designed to complement each of the 5 known 3' tRNA overhangs (ACCA, UCCA, GCCA, CCCA, CCA). The 3' adapter included 18 DNA

nucleotides for the addition of the ONT SQK-RNA004 sequencing adapter and a poly(A) homopolymer for demarcation of the start of the 3' RNA sequence. The sequences for these oligos can be found in **Supplementary Table 4**.

Candidate tRNA Reference Curation: For *Saccharomyces cerevisiae* (https://gtrnadb.org/genomes/eukaryota/Scere3/sacCer3-mature-tRNAs.fa) and *Escherichia coli* (https://gtrnadb.org/genomes/bacteria/Esch_coli_K_12_MG1655/eschColi_K_12_MG1655-mature-tRNAs.fa) we downloaded the set of genomically predicted tRNA sequences from gtRNA-db. The set of unique sequences were curated for each species and the sequences were attributed to the first predicted tRNA of that sequence based on tRNA naming conventions [37]. This set of tRNAs represents a comprehensive set of possible tRNA sequences based on genomic predictions. In our previous work a more restrictive iso-acceptor level reference was curated to account for the difficulties of resolving minimal sequence differences using a seed and chain algorithm strategy such as BWA-MEM. These sequences were used as references for alignment using tRNAZAP (**Supplementary Table 5, Supplementary Table 6).**

Design of tRNA DNA Templates for *in vitro* transcription: We appended the 3' adapter and the appropriate 5' adapter to the corresponding side of each target tRNA sequence. The IVT template is the reverse complement of this unified sequence with an additional T7 binding site sequence (TATAGTGAGTCGTATTAAATGATG) included on the 3' most end of each template. From the complete set of template sequences the interspecies and intraspecies pairwise sequence similarity was calculated using levenshtein distance (edit distance) as the score. We pooled the *E. coli* and *S. cerevisiae* tRNA sequences and split them into 8 groups that allowed a minimum similarity score of 16, where lower scores represented more similar sequences. Each group of sequences was ordered as an Oligo pool from Integrated DNA Technologies. We prepared a custom reference for the merged *E. coli* and *S. cerevisiae* IVT products (**Supplementary Table 7**)

***In Vitro* Transcription:** The 130 IVT templates were pooled in 10ul (162.5pmol total) and annealed with a two fold molar amount (325pmol) of the T7 oligonucleotide in 1X TNE by heating to 75C for 1 min followed by slow cooling**.** One microgram of the annealed oligo/template mix, was used to produce our tRNA representative RNA using the NEB HiScribe T7 Quick High Yield RNA Synthesis Kit (NEB Catalog #E2050) following manufacturer's recommendations for standard RNA synthesis of short RNA (<300nt). The reactions went for 16 hours, then were DNAse I treated for 30 minutes and cleaned up using 1.5X RNACleanXP beads (Beckman-Coulter) according to manufacturer's protocol and eluted in NF H20.

***In Vitro* Transcription Library Preparation and Sequencing:** For each experiment an augmented SQK-RNA004 (Oxford Nanopore technology)protocol was followed that used a custom splint adapter in the place of the RTA adapter. The first ligation was composed of 2 ug of the purified in vitro transcription reaction, 6µl of 5x quick ligation buffer (NEB Catalog #B6058S), 34 pmol of the custom adapter and 3 µl of T4 DNA ligase (2,000,000 units/mL, NEB Catalog #M0239S) in a total volume of 30ul. The protocol was followed from this point forward. Sequencing took place on two separate devices; a PromethION48, and PromethION2.

Biological Library preparation and Sequencing: RNA Ligase II (NEB) was used to ligate tRNA to the five splint adapters RNA. The reactions typically had 8-10 pmol of deacylated tRNA

and a 1.1 to 4 fold excess of adapters. Further detail of the ligation conditions for different runs are shown in **Supplementary Table 8**. The ligation reactions went for 45-60min at RT and were cleaned up with 1.3X Ampure RNACleanXP Spri beads and eluted in 23ul of NFH20. The eluate was then ligated to 4.5ul RLA motor associated adapter using T4 DNA ligase (2,000,000 U/mL) for 30 min and cleaned up using 0.8Xbeads, washed with WSB and eluted in 33ul REB. PromethION flow cell priming and sequencing followed the RNASQK004(version X) and were run with basecalling off.

Basecalling and BWA-MEM Alignment: All experiments were basecalled with Dorado 1.0.0 model rna004_130bps_sup@v5.2.0. For standard alignment each experiment was aligned to a custom isoacceptor level reference that included the sequencing adapters and using BWA-MEM (-W 13 -k 6 -x ont2d). Tags bearing ionic current and training relevant information were transferred from the unmapped reads to aligned reads using pysam. [42]

Data Labeling for IVT training: Data were aligned to our complete custom reference sequence set (**Supplementary Table 7**) representing the expected isoacceptor level IVT product using BWA-MEM.

```
BWA-MEM -x ont2d -W 13 -k 6
```

Reads with a mapping quality of 0 were removed and reads were labeled according to their primary tRNA alignment. An additional round of filtering removed:

1. Reads with less than 15 positions of aligned tRNA sequence,
2. Reads with lower than 75% identity across the entire reference span including splints

In addition to labeling the isodecoder of each read, sections of signal were labeled as the DNA adapter, 3' Adapter, tRNA of interest, and 5' Adapter by assigning bases based on the corresponding Dorado move table signal to base assignments.

The produced data and their corresponding labels were fed into the model training pipeline.

tRNAZAP Inference Pipeline

Our pipeline processes nanopore ionic current signals through two integrated components: (1) the classification and segmentation model and (2) the Aligner.

Classification and Segmentation model:

Our deep learning model processes raw nanopore ionic current signals to perform two key functions for tRNA analysis:

- **tRNA Classification:** Determines the specific tRNA isodecoder present in the signal.
- **Ionic Current Segmentation:** Isolates the tRNA-containing region by identifying and removing ONT adapters and 3'/5' splints.

Model Architecture: We implement these functions using a multitask Transformer neural network that divides the continuous ionic current into discrete, non-overlapping time windows (chunks). This chunking approach enables the model to learn both local patterns within individual windows and global dependencies across the entire signal. To achieve the two aforementioned functions, the multitask transformer simultaneously produces two outputs:

- **Classification Output:** Analyzes the complete signal to assign scores across all possible tRNA isodecoders. The signal is classified as the tRNA type with the highest score.
- **Segmentation Output:** Labels each time window as one of four categories: ONT adapter, 3'/5' splints, or tRNA region. The windows labeled as tRNA are then used to extract the tRNA region from the signal.

Overall, the classification and segmentation module thus extracts both a tRNA type and a tRNA region from an ionic current signal, which are required by the subsequent Aligner module to perform sequence alignment.

Model Training: The multitask transformer was trained jointly on the classification and segmentation objectives using the labels from section *Data Labeling*. The loss function was a weighted combination of classification and segmentation objectives, with weights of 0.3 and 0.7, respectively. Models were optimized using the Adam optimizer with learning rate scheduling. Training was performed for 50 epochs and a base learning rate of 0.001. During finetuning, Model parameters were updated using a learning rate of 0.0001 over 30 epochs.

Evaluation Metrics: Performance was evaluated with specific metrics for each task by comparing predicted labels versus the labels from section *Data Labeling*.

For sequence classification, we report micro and macro accuracy. Micro accuracy reflects overall performance across all reads, while macro accuracy averages performance across isodecoders to account for class imbalance.

For segmentation, we report the mean Intersection-over-Union (IoU) of the predicted versus BWA-MEM tRNA region across the dataset.

For each sequence *i*, the IoU is computed between the predicted and BWA-assigned tRNA regions (defined by start and end positions of the tRNA region). Let P_i denote the predicted region and L_i the BWA-MEM-assigned region. The IoU is given by:

$$IoU_i = \frac{|P_i \cap L_i|}{|P_i \cup L_i|}$$

where $|P_i \cap L_i|$ is the length of the overlapping region (Intersection of regions) and $|P_i \cup L_i|$ is the length of the union of the two regions. A visual example is provided in **Figure 10**.

The overall segmentation score is then reported as the average across all N sequences:

$$mIoU = \frac{1}{N}\sum_{i=1}^{N} IoU_i$$

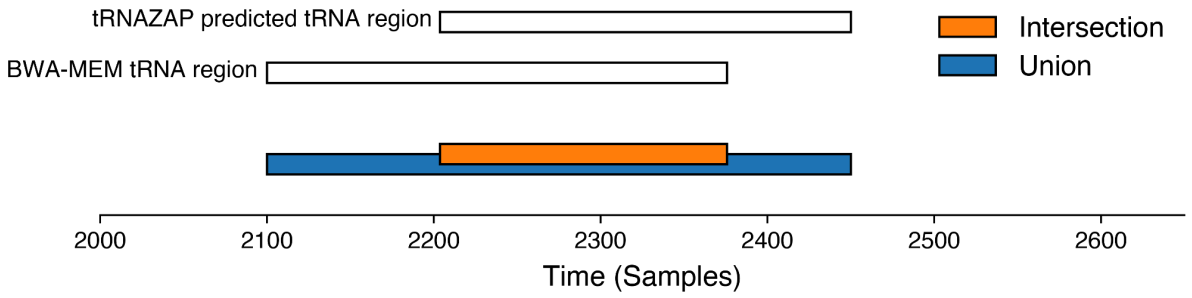

Figure 10. Intersection-over-Union (IoU) Illustration

Model Polishing with Biological Data, Data Labeling:
For both *E. coli* and *S. cerevisiae*, we used biological sequencing runs to fine-tune the model to be cognizant of tRNA-specific modification profiles in ionic current space. To disambiguate isodecoders from biological data we aligned to the isoacceptor level reference and created a lookup table for all of the isodecoder level variants. For a read to count towards an isodecoder, it had to have a minimum alignment identity of 75% and have a positive match at the correct position for every isodecoder-specific variant. In the case of *S. cerevisiae* Asp-GTC and Ser-AGA which have isodecoder with different lengths, we assigned isodecoder identity by a Needleman-Wunsch edit distance calculation. The isodecoder requiring fewer edits was selected for fine-tuning. Reads that could not fulfill the overall identity, isodecoder-specific identity, or edit distance requirements were not used in fine-tuning.

Model Polishing with Biological Data, Training:
For fine-tuning the IVT models on biological data, a collection of biological tRNA were first collected. Biological data exhibit isodecoder imbalance due to rarity of some of the tRNA types in nature. To address this imbalance, we randomly sampled 200,000 examples per isodecoder for fine-tuning. isodecoders with fewer than 20,000 samples retained all available examples. This procedure reduces the overrepresentation of highly abundant isodecoders while keeping the population of biologically rare tRNA types limited. The models were allowed to train on the data for 30 epochs. Performance was evaluated with the same metrics for IVT model performance.

tRNAZAP Aligner:
tRNAZAP aligner uses the predicted tRNA isodecoder, ionic current boundaries for the tRNA, and the move table produced by Dorado to align to an isodecoder-specific tRNA reference with an augmented Needleman-Wunsch algorithm. The algorithm finds the path traversing the entire tRNA reference for the predicted isodecoder that maximizes the ratio of query bases explained over edit distance. Optionally, the aligner can consider the highest probability isodecoder or the second highest probability isodecoder, ultimately selecting the option with the lower edit distance. For possibly fragmented reads a check is performed: if 15% of the reference length is accounted for by deletions an affine-gap Smith-Waterman alignment is performed to find a sub alignment between the query and the reference. Finally, an optional rescue alignment can be performed if the tRNA region alignment fails. The last shot alignment performs a Smith-Waterman alignment of the entire query sequence against the top 3 tRNA references. A

minimum of 25 matches in the tRNA sequence is required as well as a minimum identity of 75%. These alignments are output in a SAM compliant format and preserve all of the original reads SAM Tags.

Statistical analysis of isodecoder signal differences

For each isodecoder pair, we compared ionic current signal features extracted from a fixed window centered on the site of sequence difference. For each read, two summary features were computed: the median ionic current and the standard deviation of the ionic current within the window. Thus, isodecoder pairs p and q are represented with two-dimensional feature vectors $X^{(p)} = (X_1^{(p)}, X_2^{(p)})$ and $X^{(q)} = (X_1^{(q)}, X_2^{(q)})$, where X_1 is the vector of per-read medians and X_2 is the vector of per-read standard deviations of the ionic currents measured within the window.

For each feature $i \in \{1, 2\}$, we tested the marginal null hypothesis

$$H_{0,i}\colon\ F_{p,i}(x) = F_{q,i}(x) \text{ for all } x$$

where $F_{p,i}$ and $F_{q,i}$ denote the cumulative distribution functions of feature i for the two isodecoders. The corresponding alternative hypothesis was

$$H_{1,i}\colon\ F_{p,i}(x) \neq F_{q,i}(x) \text{ for all } x$$

Marginal hypotheses were evaluated using two-sample Kolmogorov–Smirnov tests. To control the family-wise error rate across the two tested features, *p*-values were adjusted using the Bonferroni correction. The family-wise significance level was set to $\alpha = 0.05$, corresponding to a per-feature significance level of $\alpha/2 = 0.025$.

The family-wise null hypothesis, that all marginal feature distributions are identical between isodecoders, was rejected if any Bonferroni-adjusted p-value was below α.

To summarize differences in the joint distribution of the two features, we additionally computed the Jensen–Shannon (JS) divergence between two-dimensional Gaussian mixture models fitted to the feature distributions of each isodecoder. JS divergence was used as a descriptive measure of joint distributional differences and was not used for formal hypothesis testing.

Data Availability

Sequence-level data will be publicly available through the European Nucleotide Archive. Signal files are available on request from the authors (mi.jain@northeastern.edu).

Competing Interests Statement

M.J. is a consultant to ONT and has received reimbursement for travel, accommodation, and conference fees to speak at events. Other authors declare no competing interests.

Funding

This work was supported by NIH grants R35 GM143125 (D.M.G), R01 HG013876 ([M.J., K.S.K. and D.M.G.), and T32GM149387 to (J.L.R.).

References

(1) Suzuki, T. The Expanding World of tRNA Modifications and Their Disease Relevance. *Nat Rev Mol Cell Biol* **2021**, *22* (6), 375–392.

(2) Suzuki, T.; Yashiro, Y.; Kikuchi, I.; Ishigami, Y.; Saito, H.; Matsuzawa, I.; Okada, S.; Mito, M.; Iwasaki, S.; Ma, D.; Zhao, X.; Asano, K.; Lin, H.; Kirino, Y.; Sakaguchi, Y.; Suzuki, T. Complete Chemical Structures of Human Mitochondrial tRNAs. *Nat Commun* **2020**, *11* (1), 4269.

(3) Khalique, A.; Mattijssen, S.; Maraia, R. J. A Versatile tRNA Modification-Sensitive Northern Blot Method with Enhanced Performance. *RNA* **2022**, *28* (3), 418–432.

(4) Nostramo, R. T.; Hopper, A. K. A Novel Assay Provides Insight into tRNAPhe Retrograde Nuclear Import and Re-Export in S. Cerevisiae. *Nucleic Acids Res* **2020**, *48* (20), 11577–11588.

(5) Gu, W.; Li, M.; Zhao, W. M.; Fang, N. X.; Bu, S.; Frazer, I. H.; Zhao, K.-N. tRNASer(CGA) Differentially Regulates Expression of Wild-Type and Codon-Modified Papillomavirus L1 Genes. *Nucleic Acids Res* **2004**, *32* (15), 4448–4461.

(6) Shaw, E. A.; Thomas, N. K.; Jones, J. D.; Abu-Shumays, R. L.; Vaaler, A. L.; Akeson, M.; Koutmou, K. S.; Jain, M.; Garcia, D. M. Combining Nanopore Direct RNA Sequencing with Genetics and Mass Spectrometry for Analysis of T-Loop Base Modifications across 42 Yeast tRNA Isoacceptors. *Nucleic Acids Res* **2024**, *52* (19), 12074–12092.

(7) Baczynskyj, L.; Biemann, K.; Hall, R. H. Sulfur-Containing Nucleoside from Yeast Transfer Ribonucleic Acid: 2-Thio-5(or 6)-Uridine Acetic Acid Methyl Ester. *Science* **1968**, *159* (3822), 1481–1483.

(8) Zheng, G.; Qin, Y.; Clark, W. C.; Dai, Q.; Yi, C.; He, C.; Lambowitz, A. M.; Pan, T. Efficient and Quantitative High-Throughput tRNA Sequencing. *Nat Methods* **2015**, *12* (9), 835–837.

(9) Behrens, A.; Rodschinka, G.; Nedialkova, D. D. High-Resolution Quantitative Profiling of tRNA Abundance and Modification Status in Eukaryotes by Mim-tRNAseq. *Mol Cell* **2021**, *81* (8), 1802–1815.e7.

(10) Holmes, A. D.; Chan, P. P.; Chen, Q.; Ivanov, P.; Drouard, L.; Polacek, N.; Kay, M. A.; Lowe, T. M. A Standardized Ontology for Naming tRNA-Derived RNAs Based on Molecular Origin. *Nat Methods* **2023**, *20* (5), 627–628.

(11) Kochavi, A.; Velds, A.; Suzuki, M.; Akichika, S.; Suzuki, T.; Beijersbergen, R. L.; Agami, R. Exploiting Nanopore Sequencing Advances for tRNA Sequencing of Human Cancer Models. *NAR Cancer* **2025**, *7* (4), zcaf044.

(12) Lucas, M. C.; Pryszcz, L. P.; Medina, R.; Milenkovic, I.; Camacho, N.; Marchand, V.; Motorin, Y.; Ribas de Pouplana, L.; Novoa, E. M. Quantitative Analysis of tRNA Abundance and Modifications by Nanopore RNA Sequencing. *Nat Biotechnol* **2024**, *42* (1), 72–86.

(13) Thomas, N. K.; Poodari, V. C.; Jain, M.; Olsen, H. E.; Akeson, M.; Abu-Shumays, R. L. Direct Nanopore Sequencing of Individual Full Length tRNA Strands. *ACS Nano* **2021**, *15* (10), 16642–16653.

(14) White, L. K.; Radakovic, A.; Sajek, M. P.; Dobson, K.; Riemondy, K. A.; Del Pozo, S.; Szostak, J. W.; Hesselberth, J. R. Nanopore Sequencing of Intact Aminoacylated tRNAs. *Nat Commun* **2025**, *16* (1), 7781.

(15) Stuart, C. J.; Hurtig, J. E.; Tzadikario, T.; Thomas, N. K.; Jain, M.; van Hoof, A. The Highly Conserved Intron of Tyrosine tRNA Is Critical for m1A58 Modification and Controls the Integrated Stress Response. *Proc. Natl. Acad. Sci. U. S. A.* **2025**, *122* (23). https://doi.org/10.1073/pnas.2502364122.

(16) Workman, R. E.; Tang, A. D.; Tang, P. S.; Jain, M.; Tyson, J. R.; Razaghi, R.; Zuzarte, P. C.; Gilpatrick, T.; Payne, A.; Quick, J.; Sadowski, N.; Holmes, N.; de Jesus, J. G.; Jones, K. L.; Soulette, C. M.; Snutch, T. P.; Loman, N.; Paten, B.; Loose, M.; Simpson, J. T.; Olsen, H. E.; Brooks, A. N.; Akeson, M.; Timp, W. Nanopore Native RNA Sequencing of a Human poly(A) Transcriptome. *Nat Methods* **2019**, *16* (12), 1297–1305.

(17) Esfahani, N. G.; Stein, A. J.; Akeson, S.; Tzadikario, T.; Jain, M. Evaluation of Nanopore Direct RNA Sequencing Updates for Modification Detection. *bioRxiv*, 2025.

https://doi.org/10.1101/2025.05.01.651717.
(18) Tzadikario, T.; Akeson, S.; Esfahani, N. G.; Stein, A.; Choudhary, U.; Amar, K.; Jain, M. Genomic in Vitro Transcription and Nanopore Direct RNA Sequencing of a Human B-Lymphocyte Cell Line. *bioRxiv*, 2025. https://doi.org/10.1101/2025.04.25.650674.
(19) Smith, A. M.; Jain, M.; Mulroney, L.; Garalde, D. R.; Akeson, M. Reading Canonical and Modified Nucleobases in 16S Ribosomal RNA Using Nanopore Native RNA Sequencing. *PLoS One* **2019**, *14* (5). https://doi.org/10.1371/journal.pone.0216709.
(20) Begik, O.; Lucas, M. C.; Pryszcz, L. P.; Ramirez, J. M.; Medina, R.; Milenkovic, I.; Cruciani, S.; Liu, H.; Vieira, H. G. S.; Sas-Chen, A.; Mattick, J. S.; Schwartz, S.; Novoa, E. M. Quantitative Profiling of Pseudouridylation Dynamics in Native RNAs with Nanopore Sequencing. *Nat. Biotechnol.* **2021**, *39* (10), 1278–1291.
(21) Jenjaroenpun, P.; Wongsurawat, T.; Wadley, T. D.; Wassenaar, T. M.; Liu, J.; Dai, Q.; Wanchai, V.; Akel, N. S.; Jamshidi-Parsian, A.; Franco, A. T.; Boysen, G.; Jennings, M. L.; Ussery, D. W.; He, C.; Nookaew, I. Decoding the Epitranscriptional Landscape from Native RNA Sequences. *Nucleic Acids Res* **2021**, *49* (2), e7.
(22) Abebe, J. S.; Price, A. M.; Hayer, K. E.; Mohr, I.; Weitzman, M. D.; Wilson, A. C.; Depledge, D. P. DRUMMER-Rapid Detection of RNA Modifications through Comparative Nanopore Sequencing. *Bioinformatics* **2022**, *38* (11), 3113–3115.
(23) Liu, H.; Begik, O.; Lucas, M. C.; Ramirez, J. M.; Mason, C. E.; Wiener, D.; Schwartz, S.; Mattick, J. S.; Smith, M. A.; Novoa, E. M. Accurate Detection of mA RNA Modifications in Native RNA Sequences. *Nat Commun* **2019**, *10* (1), 4079.
(24) Huang, S.; Zhang, W.; Katanski, C. D.; Dersh, D.; Dai, Q.; Lolans, K.; Yewdell, J.; Eren, A. M.; Pan, T. Interferon Inducible Pseudouridine Modification in Human mRNA by Quantitative Nanopore Profiling. *Genome Biol* **2021**, *22* (1), 330.
(25) Makhamreh, A.; Tavakoli, S.; Fallahi, A.; Kang, X.; Gamper, H.; Nabizadehmashhadtoroghi, M.; Jain, M.; Hou, Y.-M.; Rouhanifard, S. H.; Wanunu, M. Nanopore Signal Deviations from Pseudouridine Modifications in RNA Are Sequence-Specific: Quantification Requires Dedicated Synthetic Controls. *Sci Rep* **2024**, *14* (1), 22457.
(26) Rübsam, F. N. M.; Liu-Wei, W.; Sun, Y.; Patel, B. I.; van der Toorn, W.; Piechotta, M.; Dieterich, C.; Kleist, M. von; Ehrenhofer-Murray, A. E. MoDorado: Enhanced Detection of tRNA Modifications in Nanopore Sequencing by off-Label Use of Modification Callers. *Nucleic Acids Res* **2025**, *53* (15). https://doi.org/10.1093/nar/gkaf795.
(27) van der Toorn, W.; Naarmann-de Vries, I. S.; Liu-Wei, W.; Dieterich, C.; von Kleist, M. WarpDemuX-tRNA: Barcode Multiplexing for Nanopore tRNA Sequencing. *Nucleic Acids Res* **2025**, *53* (17). https://doi.org/10.1093/nar/gkaf873.
(28) Liu, B.; Cao, J.; Wang, X.; Guo, C.; Liu, Y.; Wang, T. Deciphering the tRNA-Derived Small RNAs: Origin, Development, and Future. *Cell Death Dis* **2021**, *13* (1), 24.
(29) Mulroney, L.; Wulf, M. G.; Schildkraut, I.; Tzertzinis, G.; Buswell, J.; Jain, M.; Olsen, H.; Diekhans, M.; Corrêa, I. R., Jr; Akeson, M.; Ettwiller, L. Identification of High-Confidence Human poly(A) RNA Isoform Scaffolds Using Nanopore Sequencing. *RNA* **2022**, *28* (2), 162–176.
(30) Reinsch, J. L.; Garcia, D. M. Concurrent Detection of Chemically Modified Bases in Yeast Mitochondrial tRNAs by Nanopore Direct RNA Sequencing. *bioRxiv*, 2025. https://doi.org/10.1101/2025.05.09.653160.
(31) Miyauchi, K.; Ohara, T.; Suzuki, T. Automated Parallel Isolation of Multiple Species of Non-Coding RNAs by the Reciprocal Circulating Chromatography Method. *Nucleic Acids Res* **2007**, *35* (4), e24.
(32) Zhang, W.; Foo, M.; Eren, A. M.; Pan, T. tRNA Modification Dynamics from Individual Organisms to Metaepitranscriptomics of Microbiomes. *Mol Cell* **2022**, *82* (5), 891–906.
(33) Mueller, D. M.; Getz, G. S. Steady State Analysis of Mitochondrial RNA after Growth of Yeast Saccharomyces Cerevisiae under Catabolite Repression and Derepression. *J. Biol. Chem.* **1986**, *261* (25), 11816–11822.
(34) Miyakawa, I.; Miyamoto, M.; Kuroiwa, T.; Sando, N. DNA Content of Individual Mitochondrial Nucleoids Varies Depending on the Culture Conditions of the Yeast

Saccharomyces Cerevisiae. *Cytologia (Tokyo)* **2004**, *69* (1), 101–107.
(35) Meisinger, C.; Pfanner, N.; Truscott, K. N. Isolation of Yeast Mitochondria. *Methods Mol Biol* **2006**, *313*, 33–39.
(36) Gregg, C.; Kyryakov, P.; Titorenko, V. I. Purification of Mitochondria from Yeast Cells. *J Vis Exp* **2009**, No. 30. https://doi.org/10.3791/1417.
(37) Chan, P. P.; Lowe, T. M. GtRNAdb 2.0: An Expanded Database of Transfer RNA Genes Identified in Complete and Draft Genomes. *Nucleic Acids Res* **2016**, *44* (D1), D184–D189.
(38) Ishimura, R.; Nagy, G.; Dotu, I.; Zhou, H.; Yang, X.-L.; Schimmel, P.; Senju, S.; Nishimura, Y.; Chuang, J. H.; Ackerman, S. L. RNA Function. Ribosome Stalling Induced by Mutation of a CNS-Specific tRNA Causes Neurodegeneration. *Science* **2014**, *345* (6195), 455–459.
(39) Lin, B. Y.; Chan, P. P.; Lowe, T. M. tRNAviz: Explore and Visualize tRNA Sequence Features. *Nucleic Acids Res* **2019**, *47* (W1), W542–W547.
(40) Barraud, P.; Gato, A.; Heiss, M.; Catala, M.; Kellner, S.; Tisné, C. Time-Resolved NMR Monitoring of tRNA Maturation. *Nat Commun* **2019**, *10* (1), 3373.
(41) Monroe, J. G.; Smith, T. J.; Koutmou, K. S. Investigating the Consequences of mRNA Modifications on Protein Synthesis Using in Vitro Translation Assays. *Methods in enzymology* **2021**, *658*, 379.
(42) *GitHub - pysam-developers/pysam: Pysam is a Python package for reading, manipulating, and writing genomics data such as SAM/BAM/CRAM and VCF/BCF files. It's a lightweight wrapper of the HTSlib API, the same one that powers samtools, bcftools, and tabix*. GitHub. https://github.com/pysam-developers/pysam (accessed 2025-08-15).