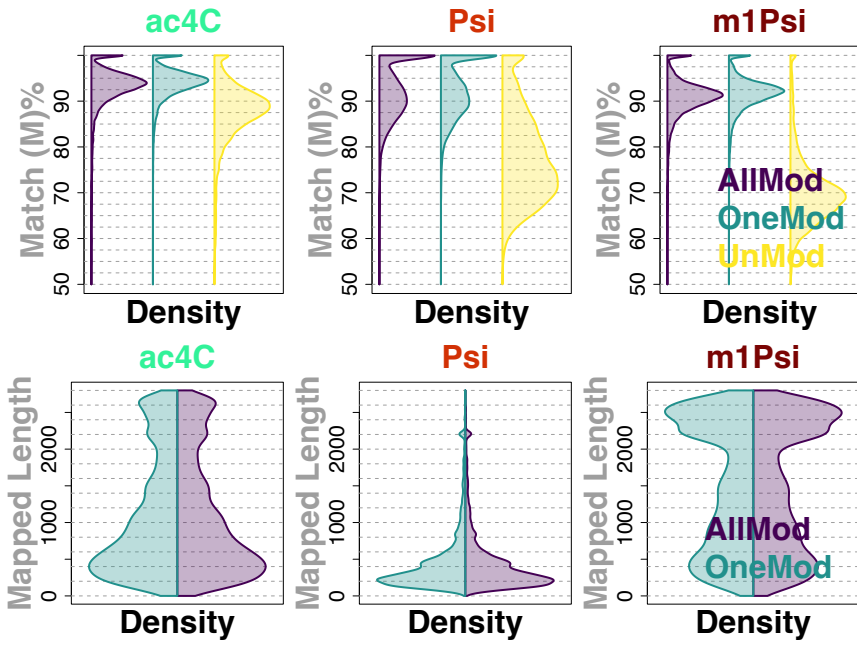
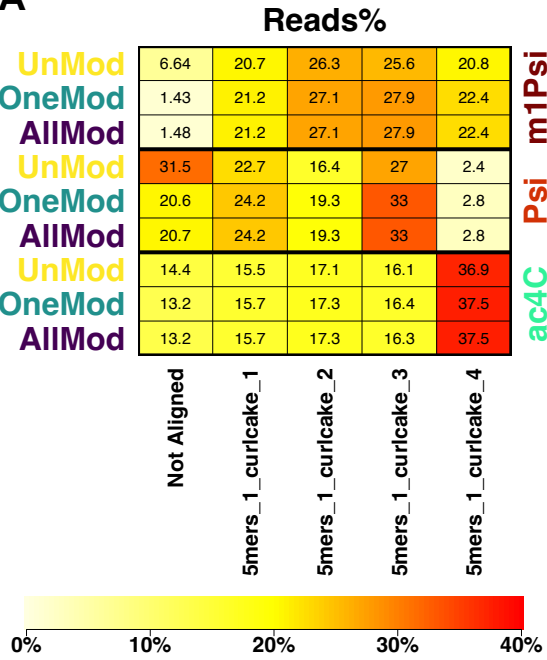


**Training Data Diversity Enhances the Basecalling of Novel RNA
Modification-Induced Nanopore Sequencing Readouts**

Supplementary Information

A



B

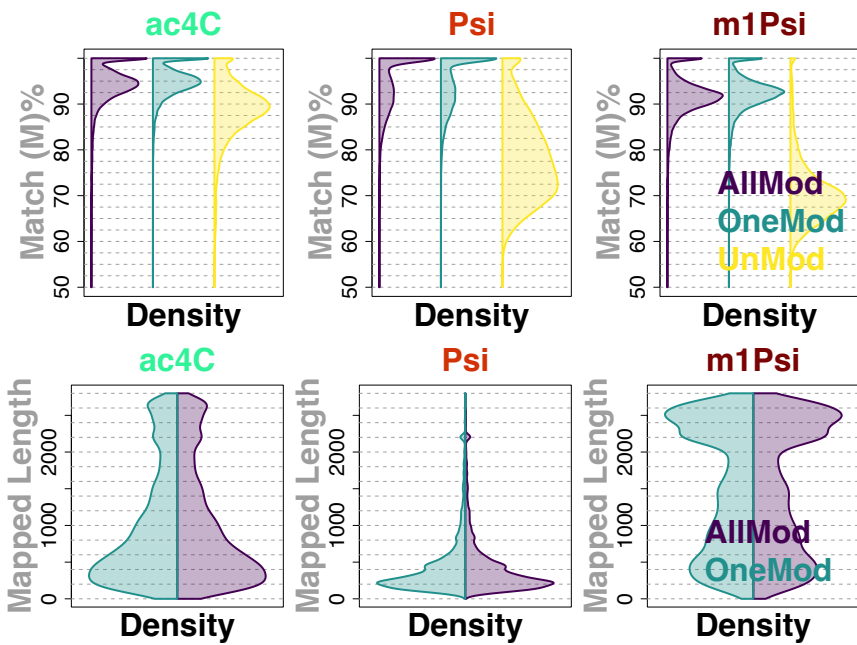
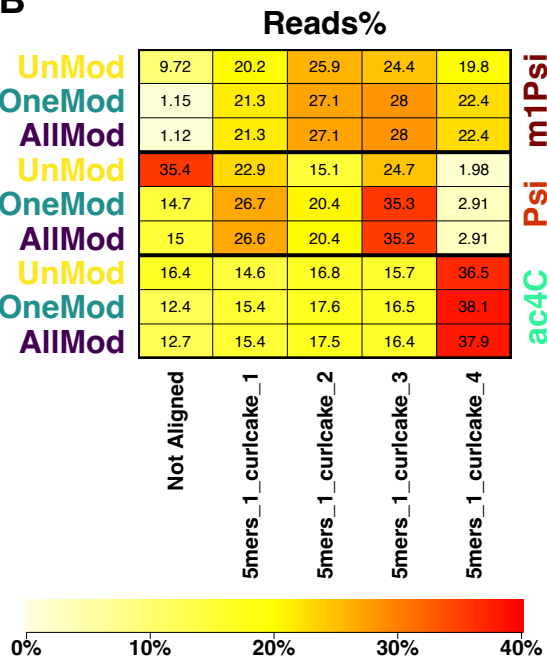
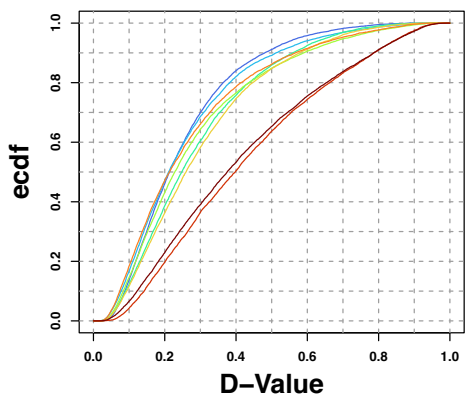
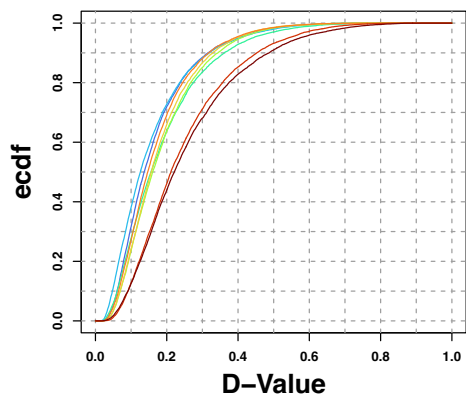


Figure S1. Bonito and Dorado trained with diverse known modifications gain the capability to basecall novel modifications. The basecalling performance of Bonito and Dorado were presented in (A) and (B), respectively. Same as in Figure 2B-D, the mappability, per-read CIGAR match fraction and mapped length distribution were used as quantification metrics. AllMod, the basecaller trained by all the modifications except for the one to be basecalled; OneMod, the basecaller trained with only the modification to be basecalled; UnMod, the basecaller trained by only unmodified reads.

Mean



SD



Dwell

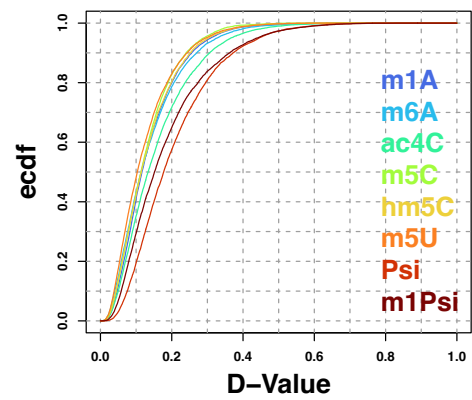


Figure S2. Modifications deviate nanopore sequencing signal features. Mean, SD and Dwell denote the mean, standard deviation and dwell time of Remora signal events. Acronyms denote different modification oligo categories. The KS-test D-values measure Mean, SD and Dwell distribution differences between modified and unmodified oligos. Ecdf denotes the empirical cumulative distribution function.

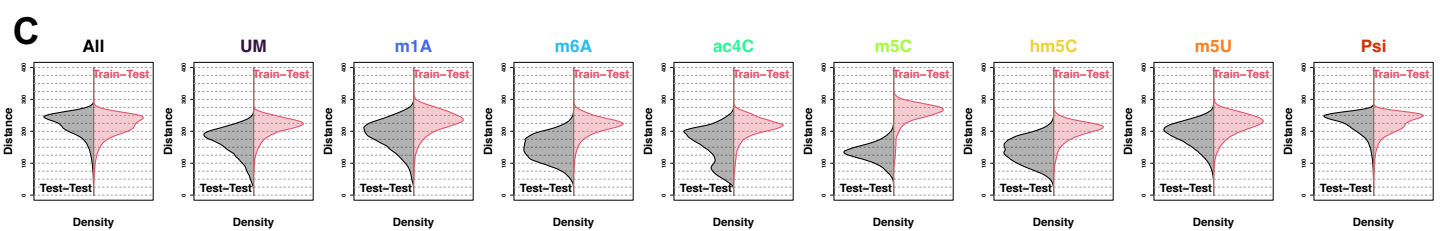
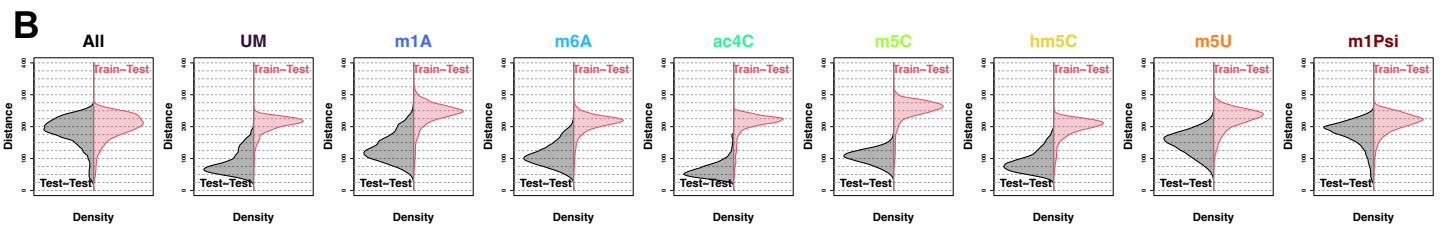
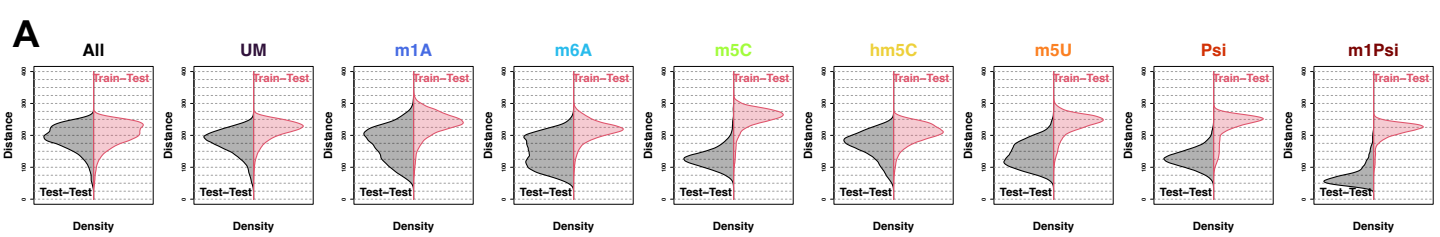


Figure S3. Quantifying the read-level similarity in the representation space. Reads were projected into the representation space and then subjected to PCA. Within the PC space, Euclidean distance was calculated to quantify the read-level similarity. Analyses for ac4C, Psi and m1Psi test reads were presented in (A), (B) and (C), respectively. All, the jointly-trained basecaller by all the oligo groups except for the test; other acronyms denote individually-trained basecallers. Train-Test and Test-Test represent the distance between training and test reads and among test reads, respectively. We took Test-Test groups as baselines to quantify the Train-Test similarity.

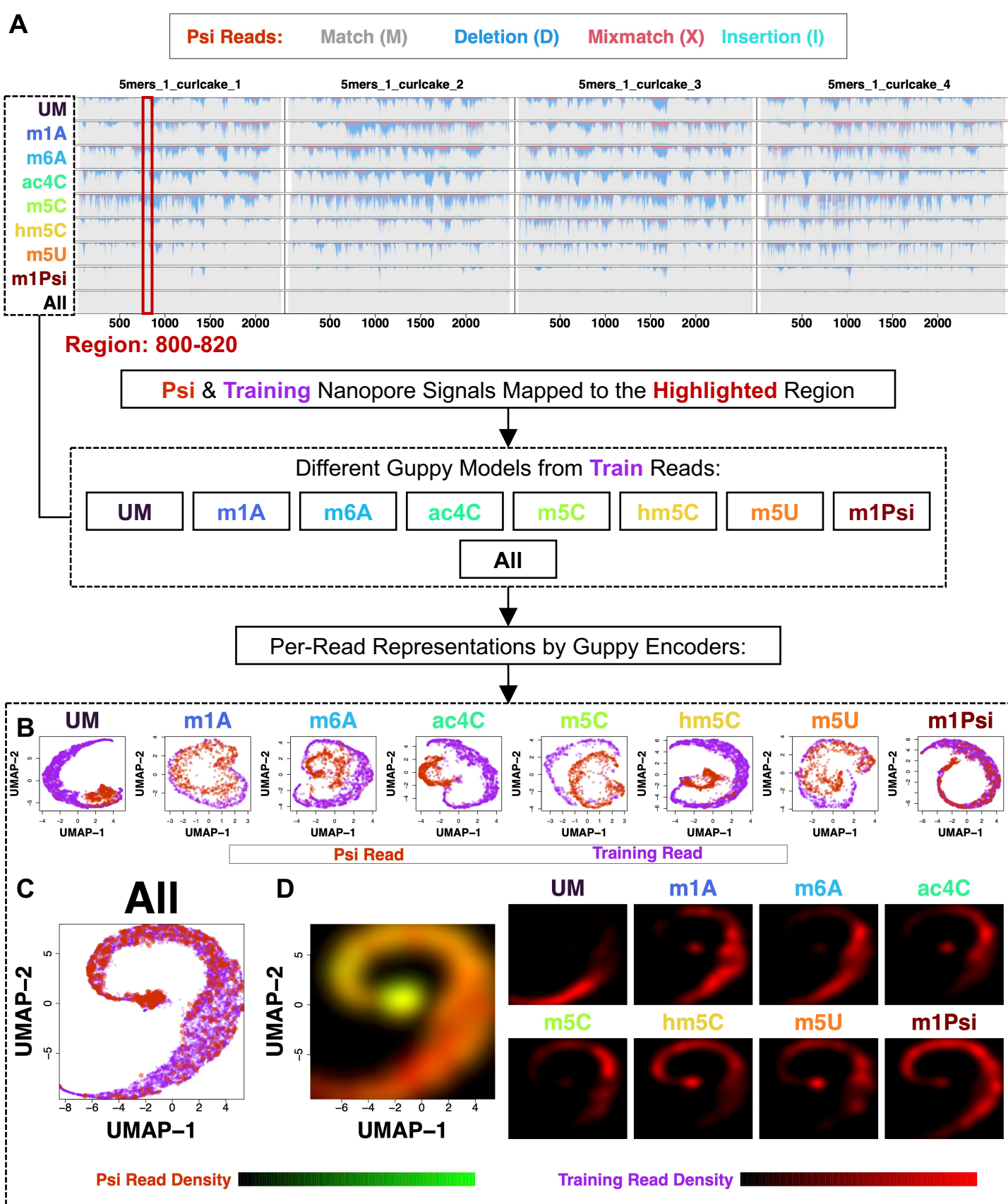
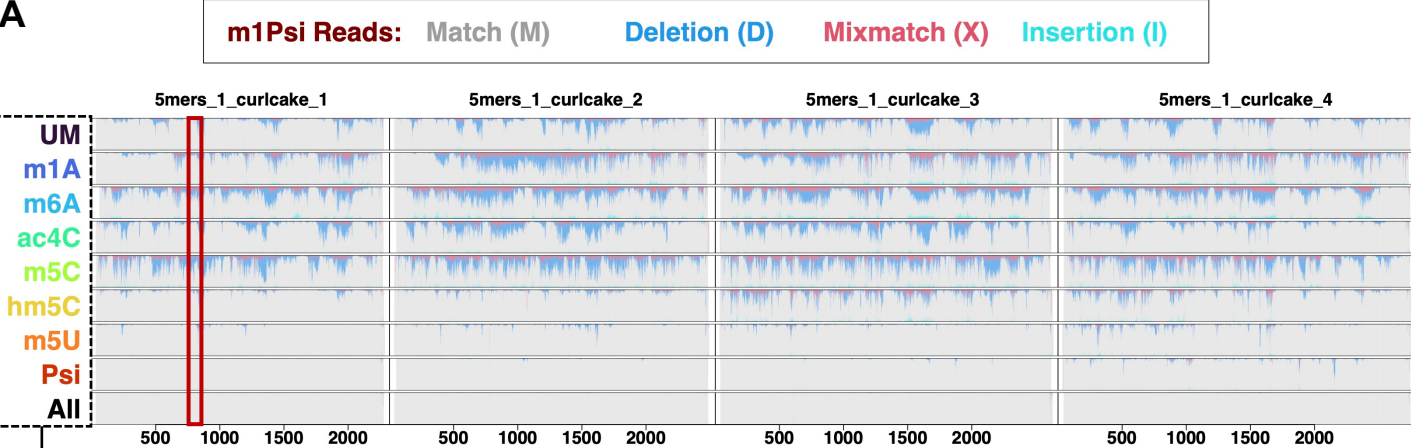


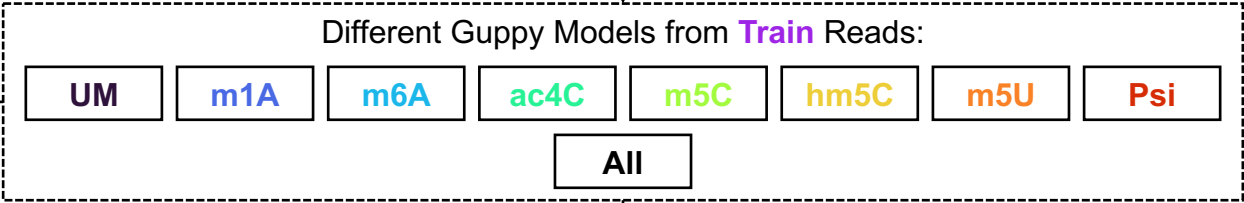
Figure S4. A generalizable representation space produced from diverse training modifications facilitates the basecalling of the out-of-sample Psi modification. (A) Performance of individually and jointly-trained basecallers on Psi reads was visualized with the genome viewer graph, which shows per-nucleotide CIGAR fractions. All, the jointly-trained basecaller by all the oligo types except for Psi; other acronyms denote individually-trained basecallers. For individually (B) and jointly-trained (C) basecallers, read fragments mapped to the boxed region were first converted as representation vectors with different basecaller encoders, then visualized by a UMAP plot. Train denotes reads used for training the corresponding basecaller. (D) Spatial distributions of different oligo types in the UMAP space as shown in (C). Black-to-green and red palette denotes Psi and training reads, respectively.

A



Region: 800-820

m1Psi & Training Nanopore Signals Mapped to the Highlighted Region



Per-Read Representations by Guppy Encoders:

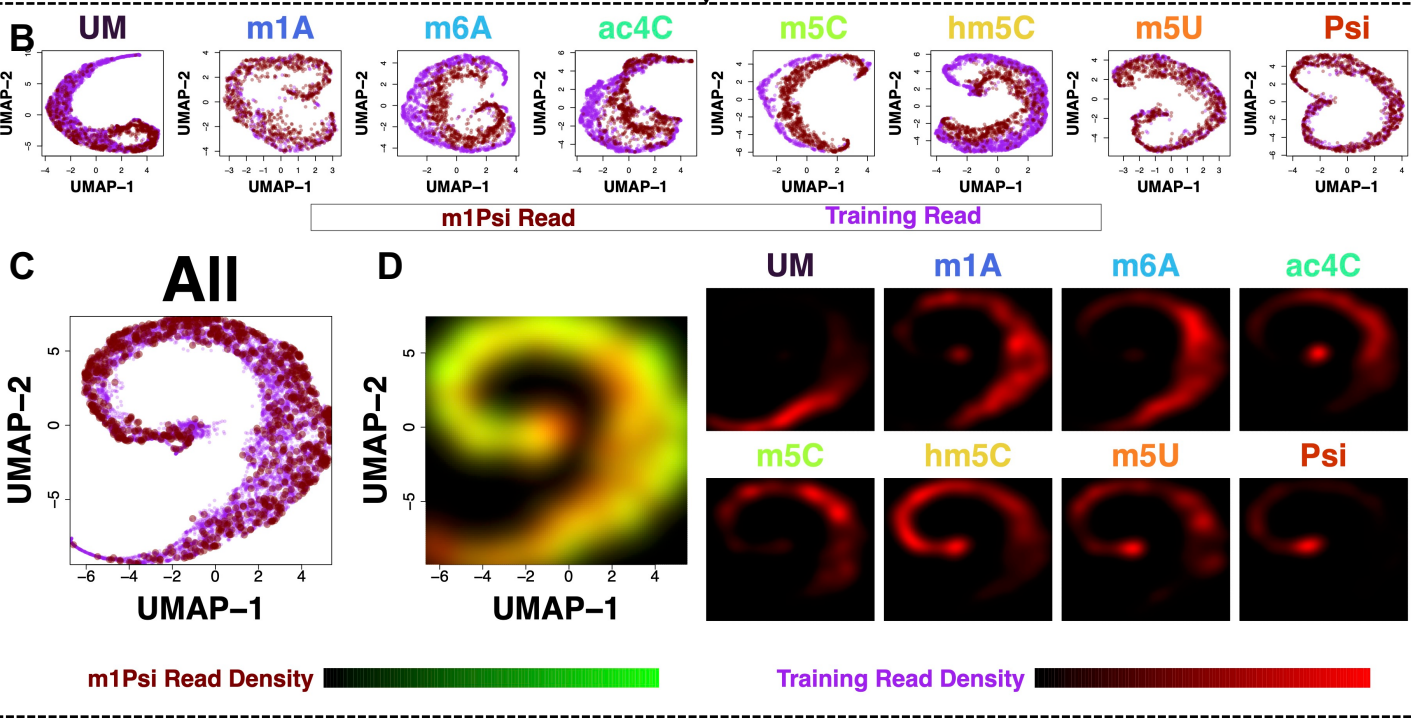


Figure S5. A generalizable representation space produced from diverse training modifications facilitates the basecalling of the out-of-sample m1Psi modification.

(A) Performance of individually and jointly-trained basecallers on m1Psi reads was visualized with the genome viewer graph, which shows per-nucleotide CIGAR fractions. All, the jointly-trained basecaller by all the oligo types except for m1Psi; other acronyms denote individually-trained basecallers. For individually (B) and jointly-trained (C) basecallers, read fragments mapped to the boxed region were first converted as representation vectors with different basecaller encoders, then visualized by a UMAP plot. Train denotes reads used for training the corresponding basecaller. (D) Spatial distributions of different oligo types in the UMAP space as shown in (C). Black-to-green and red palette denote m1Psi and training reads, respectively.

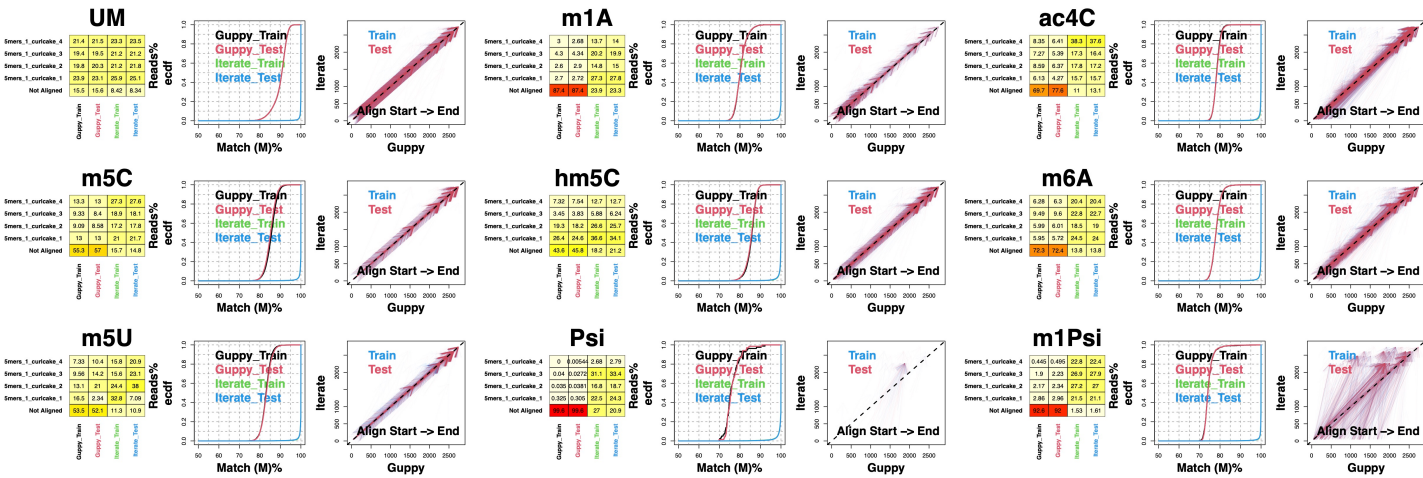


Figure S6. The iterative basecalling of RNA oligos. Guppy and iterate denote Guppy and 4th iteration (final iteration) basecallers, respectively. Train and test denote training and the independent test datasets, respectively. UM denotes unmodified reads. Match denotes the CIGAR category M. Ecdf, empirical cumulative distribution function. Start and End denote the alignment direction.