

Coral: a dual context-aware basecaller for nanopore direct RNA sequencing

Table of contents.....	1
Supplementary Note 1 Computational efficiency analysis.....	3
Table S1.....	4
Table S2.....	5
Table S3.....	6
Table S4.....	7
Table S5.....	8
Table S6.....	9
Table S7.....	10
Table S8.....	11
Table S9.....	12
Table S10.....	13
Table S11.....	14
Table S12.....	15
Table S13.....	16
Table S14.....	17
Table S15.....	18
Table S16.....	19
Table S17.....	21
Table S18.....	22
Table S19.....	23
Table S20.....	24
Fig. S1.....	25
Fig. S2.....	26
Fig. S3.....	27
Fig. S4.....	28

Fig. S5.....	29
Fig. S6.....	30
Fig. S7.....	31
Fig. S8.....	32
Fig. S9.....	33
Fig. S10.....	34
Fig. S11.....	35
Fig. S12.....	36
Fig. S13.....	37
Fig. S14.....	38
Fig. S15.....	39
Fig. S16.....	40
Fig. S17.....	41
Fig. S18.....	42
Fig. S19.....	43
Fig. S20.....	44
Fig. S21.....	45
Fig. S22.....	46
Fig. S23.....	47
Fig. S24.....	48

Supplementary Note 1 Computational efficiency analysis

The computational efficiency of different basecallers was evaluated through both theoretical complexity analysis and empirical benchmarking. A comparison of the theoretical decoding complexities reveals differences between Coral variants and other basecallers (**Table S14**). The complexity of CTC-based methods (e.g., Guppy, Dorado) scales primarily with the length of signal features T . In contrast, Coral's LSTM-based decoder (Coral-LSTM; dynamic programming and beam search for decoding) exhibits a higher complexity, scaling with factors including output sequence length L , signal feature length T , and beam size B . Coral's Transformer-based decoder (two variants, Coral and Coral-FAST) utilizes efficient cross-attention for alignment and standard beam search (default beam size is 1) for decoding, resulting in a complexity scaling primarily with output sequence length L , offering a more favorable profile compared to Coral-LSTM.

Empirical benchmarks were performed using 16,000 randomly sampled reads from the A549 RNA002 data. The results, shown in **Fig. S8**, compare the median read accuracy, runtime, and memory usage across various models. As expected, CTC-based tools such as Guppy, Dorado, and RODAN achieved the fastest runtime due to their parallel decoding nature. While slower than CTC methods, the Coral Transformer variants (e.g., Coral, Coral-FAST) offered significant speed improvements (1.44×~3.64× speedup) over the Coral-LSTM and achieved the highest median read accuracy. Furthermore, Coral-FAST (predicting k bases per auto-regressive step, $k=5$) achieves a runtime reduction to 39% of Coral (predicting a single base per auto-regressive step) while maintaining a median read accuracy loss of only ~0.5%, demonstrating the effectiveness of predicting five consecutive bases simultaneously in balancing speed and accuracy. We also compared the computational efficiency of the auto-regressive decoder with that of the CTC decoder (Coral-CTC), with a focus on their respective parallel processing capabilities. As illustrated in **Fig. S8**, Coral-CTC achieves a runtime that is 2.37× faster than Coral-FAST, but with a ~2.26% reduction in median read accuracy. This result underscores the trade-off between the speed benefits of CTC, which allows for parallelization, and the accuracy benefits of auto-regressive decoding, which leverages sequence context for higher performance. In terms of memory consumption, Coral-FAST required ~3.6 GB of RAM during inference, comparable to RODAN (~3.5 GB) and substantially more memory-efficient than Coral-LSTM (~5.3 GB). While ONT official basecallers Guppy and Dorado exhibited the lowest memory usage (~1.5GB and ~1.1GB, respectively), Coral's memory footprint remains practical.

To further evaluate computational efficiency, particularly with the new RNA004 chemistry, we also benchmarked various basecallers using 20,000 randomly sampled reads from the HEK293T RNA004 data. As shown in **Fig. S9**, CTC-based tools like Guppy and Dorado remained the fastest. Coral transformer variants, including Coral and Coral-FAST, again achieved the highest median read accuracy. Notably, the Coral-FAST model (predicting k bases per auto-regressive step, $k=6$) demonstrated a 4-fold speed improvement over the Coral model, with a negligible median accuracy reduction of only ~0.03%. This excellent speed-accuracy balance on RNA004 data may be attributed to its higher raw signal quality, which facilitates more accurate multi-base predictions per auto-regressive step. Memory consumption for all basecallers on RNA004 data was broadly comparable.

Table S1. Details of training dataset

Species	Flow cell version	Library kit	SRA/ENA accession ID or Signal dataset url	Data link
Arabidopsis	R9.4.1	RNA002	SRX14052225	RNA001/002 training dataset: https://github.com/BioinfoSZU/Coral/blob/master/README.md RNA004 training dataset: https://github.com/BioinfoSZU/Coral/blob/master/README.md
H. Sapiens ¹	R9.4	RNA001	http://s3.amazonaws.com/nanopore-human-wgs/rna/links/Bham_Run1_20171009_DirectRNA.files.txt	
Epinano synthetic	R9.4.1	RNA001	SRX5177013	
C. elegans	R9.4	RNA001	ERX3290234	
E. coli	R9.4	RNA001	SRX8525119	
Mus musculus	R9.4.1	RNA002	ERX12326526	
S. cerevisiae S288C	R9.4.1	RNA001/002	SRX8120016, SRX17942557	
Populus trichocarpa	R9.4.1	RNA001/002	SRX9290560, SRX9290561	
Danio rerio ²	R9.4.1	RNA002	ERX9536682	
H. Sapiens (Universal human reference RNA) ³	R10.4.1	RNA004	ERX12369345	

1. from the NA12878 project (BHAM Run1)
2. reads that are distinct from those of the test dataset are used for training
3. from the Garvan Institute Long Read Sequencing Benchmark Dataset (PNXRX240011)

Table S2. Details of test datasets

Species	Flow cell version	Library kit	SRA/ENA accession ID	Transcriptome reference url	Signal dataset url
Arabidopsis	R9.4.1	RNA001	ERX3766448	https://zenodo.org/records/4557005/files/transcriptomes.tgz?download=1	https://zenodo.org/records/4557005/files/arabidopsis-dataset.tgz?download=1
H. Sapiens ¹	R9.4	RNA001	N/A	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_45/gencode.v45.transcripts.fa.gz	https://zenodo.org/records/4557005/files/human-dataset.tgz?download=1
Mus musculus	R9.4.1	RNA001	ERX3444723	https://zenodo.org/records/4557005/files/transcriptomes.tgz?download=1	https://zenodo.org/records/4557005/files/mouse-dataset.tgz?download=1
S. cerevisiae S288C	R9.4.1	RNA002	SRX8120036	https://zenodo.org/records/4557005/files/transcriptomes.tgz?download=1	https://zenodo.org/records/4557005/files/yeast-dataset.tgz?download=1
Populus trichocarpa	R9.4.1	RNA001	SRX9347134	https://zenodo.org/records/4557005/files/transcriptomes.tgz?download=1	https://zenodo.org/records/4557005/files/poplar-dataset.tgz?download=1
Danio rerio	R9.4.1	RNA002	ERX9536682	https://www.ncbi.nlm.nih.gov/datasets/genome/GCF_000002035.6	https://zenodo.org/records/11632496/files/zebrafish-dataset.zip?download=1
H. Sapiens (HEK293T cell) ²	R10.4.1	RNA004	N/A	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_45/gencode.v45.transcripts.fa.gz	http://sg-nex-data.s3.amazonaws.com/data/sequencing_data_ont/fast5/SGNex_Hek293T_directRNA_replicate5_run1/SGNex_Hek293T_directRNA_replicate5_run1.pod5

1. from the NA12878 project (HOPKINS RUN1)
2. from the SG-NEx project (SGNex_Hek293T_directRNA_replicate5_run1, accessed on [2024-11-21] at registry.opendata.aws/sg-nex-data)

Table S3. Basecaller configuration

Guppy (v6.5.7) basecalling configuration	
Download url	https://hub.docker.com/r/genomicpariscentre/guppy-gpu/tags?name=6.5.7
Basecalling command	<code>guppy_basecaller -i [fast5_dir] --recursive -s [output_dir] --device cuda:0 -c rna_r9.4.1_70bps_hac.cfg -q 0 --disable_qscore_filtering</code>
RODAN (v1.0) basecalling configuration	
Download url	https://github.com/biodlab/RODAN
Basecalling command	<code>python RODAN/basecall.py -m RODAN/rna.torch -a RODAN/rnaarch [fast5_dir] > [output_dir]/rodan.fasta</code>
Coral (v2.0) basecalling configuration	
Download url	https://github.com/BioinfoSZU/Coral
Basecalling command	<code>python Coral/coral.py --prefix coral --gpu 0 --input-dir [fast5_dir/pod5 file] --output-dir [output_dir] --kit [RNA002 or RNA004]</code>
GCRTcall (v1.0) basecalling configuration	
Download url	https://github.com/liqingwen98/GCRTcall
Basecalling command	<code>python GCRTcall.py --fast5-dir [fast5_dir] --output-dir [output_dir] --device-id 0</code>
Dorado (v0.8.3) basecalling configuration	
Download url	https://github.com/nanoporetech/dorado
Basecalling command	<code>dorado basecaller --recursive --device cuda:0 --emit-sam [rna002_70bps_hac@v3 or rna004_130bps_sup@v5.1.0] [fast5_dir or pod5_dir] > output_dir/dorado.sam</code>

Table S4. Basecalling performance of leave-one-species-out cross-validation on RNA001/002 test datasets.

Dataset (Read Count)	Basecallers	Read Accuracy (%)		Mismatch(%)		Insertion(%)		Deletion(%)		Read Length (bp)		Unaligned Reads
		mean	median	mean	median	mean	median	mean	median	mean	median	
Human (100000)	Guppy*	89.88	90.6	2.56	2.32	2.21	2	5.35	4.91	964	713	N/A
	Dorado*	90.77	91.32	2.76	2.55	2.17	2	4.3	3.93	900	665	1316
	RODAN	90.45	91.11	2.32	2.09	1.84	1.63	5.39	4.85	847	681	4063
	GCRTcall	91.68	92.69	1.8	1.56	1.87	1.42	4.65	4.07	973	722	1445
	Coral	<u>94.8</u>	<u>95.55</u>	<u>1.5</u>	<u>1.28</u>	<u>1.23</u>	<u>1.01</u>	<u>2.46</u>	<u>2</u>	<u>987</u>	<u>730</u>	<u>117</u>
Mouse (100000)	Guppy*	87.17	87.66	3.88	3.68	2.66	2.52	6.29	5.96	647	524	N/A
	Dorado*	88.84	89.47	3.59	3.35	1.96	1.8	5.61	5.21	639	519	1700
	RODAN	87.78	88.12	3.19	3.05	1.62	1.44	7.4	7.11	498	450	4813
	GCRTcall	89.54	90.19	2.69	2.51	1.67	1.49	6.09	5.6	643	522	911
	Coral	<u>92.69</u>	<u>93.46</u>	<u>2.45</u>	<u>2.23</u>	<u>1.26</u>	<u>1.09</u>	<u>3.6</u>	<u>3.03</u>	<u>652</u>	<u>529</u>	<u>154</u>
Yeast (100000)	Guppy*	90.51	91.36	2.78	2.38	2.51	2.32	4.2	3.8	842	597	N/A
	Dorado*	93.21	94.42	2.21	1.69	1.8	1.54	2.78	2.25	842	598	848
	RODAN	91.15	91.9	2.51	2.21	2.08	1.89	4.25	3.81	652	553	3465
	GCRTcall	91.94	92.97	2.22	1.87	1.98	1.77	3.86	3.24	847	601	1071
	Coral	<u>94.65</u>	<u>95.83</u>	<u>1.91</u>	<u>1.46</u>	<u>1.36</u>	<u>1.05</u>	<u>2.08</u>	<u>1.57</u>	<u>853</u>	<u>604</u>	<u>273</u>
Arabidopsis (100000)	Guppy*	90.73	91.59	2.32	2.03	2.18	1.93	4.77	4.32	964	855	N/A
	Dorado*	93.36	94.36	1.86	1.51	1.69	1.39	3.09	2.63	966	854	509
	RODAN	91.6	92.29	2.01	1.81	1.85	1.59	4.54	4.13	723	667	1907
	GCRTcall	92.17	93.24	1.74	1.48	1.91	1.5	4.17	3.59	969	859	565
	Coral	<u>95.21</u>	<u>96.15</u>	<u>1.38</u>	<u>1.09</u>	<u>1.02</u>	<u>0.74</u>	<u>2.4</u>	<u>1.9</u>	<u>979</u>	<u>870</u>	<u>65</u>
Poplar (100000)	Guppy*	89.26	90.16	2.95	2.64	2.87	2.51	4.91	4.47	792	665	N/A
	Dorado*	92.51	93.66	2.26	1.86	2.17	1.74	3.07	2.57	795	668	477
	RODAN	90.87	91.64	2.49	2.24	2.25	1.95	4.38	4	617	559	3025
	GCRTcall	90.85	91.94	2.44	2.11	2.41	2.01	4.3	3.72	797	669	1168
	Coral	<u>94.48</u>	<u>95.58</u>	<u>1.75</u>	<u>1.41</u>	<u>1.56</u>	<u>1.1</u>	<u>2.2</u>	<u>1.73</u>	<u>805</u>	<u>676</u>	<u>70</u>
Zebrafish (100000)	Guppy*	86.45	86.9	3.86	3.61	2.18	1.94	7.51	7.23	950	823	N/A
	Dorado*	92.81	93.66	2.31	2	1.93	1.63	2.94	2.52	1011	880	506
	RODAN	88.7	89.27	3.16	2.95	1.75	1.59	6.4	5.92	759	696	2798
	GCRTcall	88.93	89.6	2.96	2.72	1.8	1.54	6.31	5.83	977	848	1537
	Coral	<u>93.7</u>	<u>94.55</u>	<u>2.25</u>	<u>1.95</u>	<u>1.52</u>	<u>1.2</u>	<u>2.53</u>	<u>2.07</u>	<u>1008</u>	<u>875</u>	<u>348</u>

* Since there are no training pipelines available for Guppy and Dorado, the corresponding the original pretrained official models were used. LOOCV was no applicable to these two methods. There is no guarantee that data identical to the test set, or data from the same or closely related species, was unseen during the training process.

The best results among all methods are marked as bold text.

The best LOOCV results (excluding Guppy and Dorado) are underlined.

“N/A” on unaligned reads for Guppy in the RNA001/002 test datasets, which include only the reads successfully basecalled by Guppy and aligned to the reference sequence.

Table S5. Basecalling performance of leave-one-chromosome-out evaluation on the chr22 reads of HEK293T RNA004 sample.

Dataset (Read Count)	Basecallers	Read Accuracy (%)		Mismatch(%)		Insertion(%)		Deletion(%)		Read Length (bp)		Unaligned Reads
		mean	median	mean	median	mean	median	mean	median	mean	median	
HEK293T RNA004 chr22 reads (233936)	Guppy*	86.43	87.02	3.81	3.68	2.57	2.37	7.20	6.75	1130	966	26408
	Dorado*	96.77	98.13	0.76	0.55	0.98	0.36	1.49	0.79	1122	898	1747
	RODAN	91.31	92.44	1.94	1.72	1.32	1.09	5.44	4.55	1074	875	8327
	GCRTcall	92.86	94.15	1.48	1.27	1.32	0.93	4.34	3.49	1094	<u>880</u>	3545
	Coral	<u>96.54</u>	<u>97.78</u>	<u>0.93</u>	<u>0.70</u>	<u>0.70</u>	<u>0.36</u>	<u>1.83</u>	<u>1.03</u>	<u>1099</u>	869	<u>1750</u>

* Since there are no training pipelines available for Guppy and Dorado, the corresponding the original pretrained official models were used. LOOCV was no applicable to these two methods. There is no guarantee that data identical to the test set, or data from the same or closely related species, was unseen during the training process.

The best results among all methods are marked as bold text.

The best LOOCV results (excluding Guppy and Dorado) are underlined.

Table S6. Architecture and parameter size of ablated models used in RNA001/002 test datasets.

Ablated model	Encoder type	Decoder type	Encoder layers	Decoder layers	Feature dimension	Param size(MB)
LSTM-Transformer	CNN-LSTM	Transformer	5	12	512	351
Coral-CTC	CNN-Transformer	CTC	18	1	512	292
Coral-LSTM	CNN-Transformer	LSTM	18	1	512	309
Coral-S	CNN-Transformer	Transformer	3	2	256	44
Coral-M	CNN-Transformer	Transformer	1	6	384	115
Coral-L3	CNN-Transformer	Transformer	18	3	512	356
Coral-L6	CNN-Transformer	Transformer	18	6	512	420
Coral (i.e., Coral-L12)	CNN-Transformer	Transformer	18	12	512	548
Coral-FAST	CNN-Transformer	Transformer	18	12	512	554

Table S7. Basecalling performance of Coral-L3, Coral-LSTM, and Coral-CTC on RNA001/002 test datasets.

Dataset (Read Count)	Model	Read Accuracy (%)		Mismatch(%)		Insertion(%)		Deletion(%)		Read Length (bp)		Unaligned Reads
		mean	median	mean	median	mean	median	mean	median	mean	median	
Human (100000)	Coral-L3	96.43	97.29	0.94	0.72	0.94	0.66	1.68	1.11	987	730	125
	Coral-LSTM	94.84	95.56	1.45	1.23	1.3	0.97	2.42	2.01	988	731	77
	Coral-CTC	94.28	95.03	1.28	1.09	0.88	0.72	3.55	3.03	967	717	226
Mouse (100000)	Coral-L3	92.11	93.97	1.89	1.67	1.03	0.84	4.97	2.99	619	508	469
	Coral-LSTM	92.38	93.01	2.48	2.27	1.26	1.06	3.87	3.49	657	534	115
	Coral-CTC	91.46	92.13	2.25	2.06	0.93	0.81	5.37	4.87	638	518	369
Yeast (100000)	Coral-L3	96.43	97.65	1.12	0.74	0.88	0.59	1.57	0.8	850	603	148
	Coral-LSTM	95.44	96.36	1.63	1.25	1.1	0.81	1.83	1.42	855	605	135
	Coral-CTC	94.67	95.6	1.49	1.15	0.89	0.72	2.95	2.39	842	597	355
Arabidopsis (100000)	Coral-L3	96.35	97.36	0.97	0.7	0.96	0.63	1.73	1.11	976	866	62
	Coral-LSTM	95.07	95.97	1.45	1.16	1.22	0.86	2.27	1.87	983	874	52
	Coral-CTC	94.39	95.3	1.26	1.01	0.88	0.64	3.47	2.92	961	852	166
Poplar (100000)	Coral-L3	95.44	96.68	1.35	1.01	1.37	0.88	1.85	1.17	803	674	60
	Coral-LSTM	94.28	95.34	1.81	1.47	1.66	1.13	2.25	1.83	809	680	60
	Coral-CTC	93.49	94.59	1.69	1.38	1.25	0.86	3.57	2.98	789	662	201
Zebrafish (100000)	Coral-L3	95.31	96.18	1.66	1.36	1.15	0.83	1.88	1.38	1011	878	242
	Coral-LSTM	93.98	94.78	2.05	1.77	1.51	1.08	2.46	2.06	1019	888	251
	Coral-CTC	93.18	93.96	1.96	1.69	1.03	0.77	3.82	3.35	988	858	385

The best results are marked as bold text.

Table S8. Basecalling performance of Coral-L3 and LSTM-Transformer on RNA001/002 test datasets.

Dataset (Read Count)	Model	Read Accuracy (%)		Mismatch (%)		Insertion (%)		Deletion (%)		Read Length (bp)		Unaligned Reads
		mean	median	mean	median	mean	median	mean	median	mean	median	
Human (100000)	Coral-L3	96.43	97.29	0.94	0.72	0.94	0.66	1.68	1.11	987	730	125
	LSTM-Transformer	95.38	96.49	0.97	0.78	1.19	0.94	2.46	1.48	970	714	330
Mouse (100000)	Coral-L3	92.11	93.97	1.89	1.67	1.03	0.84	4.97	2.99	619	508	469
	LSTM-Transformer	92.83	93.76	2.02	1.83	1.33	1.12	3.82	2.98	636	519	459
Yeast (100000)	Coral-L3	96.43	97.65	1.12	0.74	0.88	0.59	1.57	0.8	850	603	148
	LSTM-Transformer	95.55	96.89	1.15	0.82	1.2	0.92	2.1	1.13	836	591	500
Arabidopsis Yeast (100000)	Coral-L3	96.35	97.36	0.97	0.7	0.96	0.63	1.73	1.11	976	866	62
	LSTM-Transformer	95.53	96.59	1.03	0.8	1.23	0.92	2.21	1.45	963	853	248
Poplar (100000)	Coral-L3	95.44	96.68	1.35	1.01	1.37	0.88	1.85	1.17	803	674	60
	LSTM-Transformer	94.27	95.65	1.44	1.17	1.76	1.3	2.54	1.58	787	658	352
Zebrafish (100000)	Coral-L3	95.31	96.18	1.66	1.36	1.15	0.83	1.88	1.38	1011	878	242
	LSTM-Transformer	93.69	94.9	1.84	1.57	1.66	1.35	2.81	1.87	988	855	585

The best results are marked as bold text.

Table S9. Basecalling performance of Coral (predict single base per step) and Coral-FAST (predict 5 bases per step) on RNA001/002 test datasets.

Dataset (Read Count)	Model	Read Accuracy (%)		Mismatch (%)		Insertion (%)		Deletion (%)		Read Length (bp)		Unaligned Reads
		mean	median	mean	median	mean	median	mean	median	mean	median	
Human (100000)	Coral	96.93	97.69	0.84	0.59	0.84	0.56	1.4	0.92	991	733	93
	Coral-FAST	96.18	97.23	0.97	0.77	0.82	0.59	2.03	1.14	980	724	160
Mouse (100000)	Coral	94.14	94.97	1.89	1.64	1.02	0.83	2.94	2.34	657	533	92
	Coral-FAST	93.71	94.61	2.05	1.84	1.03	0.85	3.21	2.46	651	529	175
Yeast (100000)	Coral	96.92	98.07	1.02	0.6	0.79	0.47	1.27	0.65	855	605	108
	Coral-FAST	96.48	97.7	1.17	0.8	0.8	0.51	1.55	0.75	848	601	153
Arabidopsis (100000)	Coral	96.76	97.63	0.92	0.64	0.9	0.57	1.42	0.97	984	874	50
	Coral-FAST	96.35	97.42	1	0.74	0.89	0.58	1.76	1.05	976	866	68
Poplar (100000)	Coral	95.79	96.93	1.31	0.96	1.34	0.83	1.55	1.06	808	679	54
	Coral-FAST	95.33	96.69	1.36	1.05	1.3	0.82	2	1.13	800	671	104
Zebrafish (100000)	Coral	95.44	96.27	1.65	1.34	1.2	0.87	1.72	1.28	1013	880	200
	Coral-FAST	95.19	96.18	1.67	1.38	1.11	0.79	2.03	1.33	1009	877	241

The best results are marked as bold text.

Table S10. Basecalling performance of Coral variants with different scale on RNA001/002 test datasets.

Dataset (Read Count)	Model	Read Accuracy (%)		Mismatch (%)		Insertion (%)		Deletion (%)		Read Length (bp)		Unaligned Reads
		mean	median	mean	median	mean	median	mean	median	mean	median	
Human (100000)	Coral-S	93.59	94.25	1.42	1.21	1.29	1.03	3.7	3.13	965	715	298
	Coral-M	94.42	95.09	1.64	1.4	1.53	1.14	2.41	1.99	984	728	291
	Coral-L3	96.43	97.29	0.94	0.72	0.94	0.66	1.68	1.11	987	730	125
	Coral-L6	96.69	97.44	0.91	0.68	0.93	0.65	1.48	1.02	991	733	95
	Coral-L12	96.93	97.69	0.84	0.59	0.84	0.56	1.4	0.92	991	733	93
Mouse (100000)	Coral-S	88.44	89.13	2.55	2.36	1.36	1.18	7.65	6.93	614	500	513
	Coral-M	90.64	91.18	3.21	3.01	1.82	1.59	4.33	3.93	650	528	490
	Coral-L3	92.11	93.97	1.89	1.67	1.03	0.84	4.97	2.99	619	508	469
	Coral-L6	93.97	94.79	1.97	1.72	1.12	0.94	2.94	2.34	656	532	115
	Coral-L12	94.14	94.97	1.89	1.64	1.02	0.83	2.94	2.34	657	533	92
Yeast (100000)	Coral-S	93.99	95.08	1.5	1.14	1.24	0.94	3.27	2.44	839	596	418
	Coral-M	94.73	95.89	1.81	1.36	1.46	1.03	2	1.45	855	604	581
	Coral-L3	96.43	97.65	1.12	0.74	0.88	0.59	1.57	0.8	850	603	148
	Coral-L6	96.75	97.84	1.08	0.68	0.89	0.56	1.27	0.71	856	606	110
	Coral-L12	96.92	98.07	1.02	0.6	0.79	0.47	1.27	0.65	855	605	108
Arabidop sis (100000)	Coral-S	93.51	94.4	1.43	1.18	1.3	0.99	3.76	3.13	960	850	214
	Coral-M	94.38	95.19	1.68	1.39	1.5	1.09	2.44	2.04	977	867	228
	Coral-L3	96.35	97.36	0.97	0.7	0.96	0.63	1.73	1.11	976	866	62
	Coral-L6	96.63	97.49	0.96	0.69	0.98	0.65	1.42	1.01	984	874	48
	Coral-L12	96.76	97.63	0.92	0.64	0.9	0.57	1.42	0.97	984	874	50
Poplar (100000)	Coral-S	92.46	93.56	1.95	1.62	1.85	1.4	3.74	3.08	791	663	260
	Coral-M	93.07	94.15	2.25	1.88	2.1	1.53	2.58	2.13	803	674	280
	Coral-L3	95.44	96.68	1.35	1.01	1.37	0.88	1.85	1.17	803	674	60
	Coral-L6	95.67	96.79	1.36	1.01	1.42	0.91	1.56	1.08	809	679	50
	Coral-L12	95.79	96.93	1.31	0.96	1.34	0.83	1.55	1.06	808	679	54
Zebrafish (100000)	Coral-S	91.8	92.63	2.31	2.03	1.67	1.36	4.22	3.67	993	863	561
	Coral-M	92.19	93.02	2.75	2.44	1.86	1.49	3.2	2.77	1006	874	640
	Coral-L3	95.31	96.18	1.66	1.36	1.15	0.83	1.88	1.38	1011	878	242
	Coral-L6	95.42	96.25	1.66	1.35	1.19	0.86	1.74	1.3	1013	881	207
	Coral-L12	95.44	96.27	1.65	1.34	1.2	0.87	1.72	1.28	1013	880	200

The best results are marked as bold text.

Table S11. Architecture and parameter size of ablated models used in RNA004 test dataset.

Ablated model	Encoder type	Decoder type	Encoder layers	Decoder layers	Feature dimension	Param size(MB)
LSTM-Transformer	CNN-LSTM	Transformer	5	12	512	351
Coral-CTC	CNN-Transformer	CTC	18	1	512	292
Coral-LSTM	CNN-Transformer	LSTM	18	1	512	309
Coral-S	CNN-Transformer	Transformer	3	2	256	44
Coral-M	CNN-Transformer	Transformer	1	6	384	115
Coral-ML	CNN-Transformer	Transformer	15	3	512	308
Coral-L3	CNN-Transformer	Transformer	18	3	512	356
Coral-L6	CNN-Transformer	Transformer	18	6	512	420
Coral (i.e., Coral-L15)	CNN-Transformer	Transformer	18	15	512	612
Coral-FAST	CNN-Transformer	Transformer	18	15	512	634

Table S12. Basecalling performance of Coral ablated models on HEK293T RNA004 dataset.

Dataset (Read Count)	Model	Read Accuracy (%)		Mismatch (%)		Insertion (%)		Deletion (%)		Read Length (bp)		Unaligned Reads
		mean	median	mean	median	mean	median	mean	median	mean	median	
HEK293T RNA004 sampled reads (623781)	Coral-CTC	95.32	96.39	0.83	0.63	0.61	0.41	3.25	2.32	1025	759	176111
	Coral-LSTM	96.28	97.55	0.90	0.64	0.81	0.50	2.01	1.17	1044	772	171153
	LSTM- Transformer	96.05	98.15	0.74	0.38	1.02	0.33	2.18	0.74	1015	735	173998
	Coral-S	93.38	94.37	1.57	1.28	1.65	1.23	3.40	2.66	1026	754	181249
	Coral-M	94.44	96.05	1.47	1.02	1.26	0.79	2.84	1.77	1030	756	178516
	Coral-ML	97.33	98.71	0.63	0.31	0.46	0.13	1.58	0.61	1045	775	171036
	Coral-L3	97.35	98.80	0.65	0.33	0.60	0.19	1.41	0.49	1050	777	170695
	Coral-L6	97.45	98.84	0.59	0.29	0.50	0.14	1.46	0.48	1046	775	170323
	Coral-FAST	97.75	99.37	0.46	0.15	0.38	0.00	1.41	0.21	1045	774	170233
	Coral	97.74	99.39	0.39	0.06	0.35	0.00	1.52	0.33	1047	776	169982

The best results are marked as bold text.

Table S13. Basecalling performance of Coral-L3 and Coral-L3 without up-sampling in the encoder on sampled reads from HEK293T RNA004 data.

Dataset (Read Count)	Model (Aligned reads)	Read Accuracy (%)		Mismatch (%)		Insertion (%)		Deletion (%)		Read Length (bp)		GFLOPS	GPU Memory Usage (MB)
		mean	median	mean	median	mean	median	mean	median	mean	median		
HEK293T RNA004 sampled reads (830885)	Coral-L3 (596880)	97.40	98.81	0.64	0.32	0.58	0.18	1.37	0.47	1000	726	131	1526
	Coral-L3 (w/o up-sampling) (595795)	97.21	98.81	0.60	0.24	0.50	0.11	1.69	0.58	993	722	524	2780

The best results are marked as bold text. GFLOPS (Giga Floating-point Operations Per Second) denotes the computational cost of the matrix-multiplication operations in the Transformer encoder layers under a batch size of 64 and a feature dimension of 512. GPU memory usage is reported for the operations performed within the Transformer encoder.

Table S14. Complexity and architecture of basecaller models

Basecaller	Encoder	Decoder	Param size (MB)	Complexity
Guppy	CNN-LSTM	CTC	11	$O(T * d * V)$
RODAN	CNN	CTC	41	$O(T * d * V)$
GCRTcall	CNN-Transformer	CTC	349	$O(T * d * V)$
Dorado	CNN-LSTM (RNA002); CNN-Transformer (RNA004)	CTC	25 (RNA002); 300 (RNA004)	$O(T * d * V)$
Coral-LSTM	CNN-Transformer	LSTM (dynamic programming and beam search)	309	$O(L * T * B * d * V)$
Coral-CTC	CNN-Transformer	CTC	292	$O(T * d * V)$
Coral	CNN-Transformer	Transformer (beam search, predict single base per auto-regressive step)	548 (RNA002); 612 (RNA004)	$O(L * N * d * d)$
Coral-FAST	CNN-Transformer	Transformer (beam search, predict k bases per auto-regressive step)	554 (RNA002); 634 (RNA004)	$O(\frac{L}{k} * N * d * d)$

(T is length of signal features; d is feature dimension size; V is alphabet size; L is length of output sequence; B is beam size; N represent the number of transformer decoder layers; k=5 for RNA001/002 kits and k=6 for RNA004 kit).

Table S15. Details of human transcriptome analysis datasets

Cell line	Dataset	Library kit	Fast5/Pod5 Data link	Illumina data link
A549	SGNex_A549_directRNA_replicate6_run1	RNA002	http://sg-nex-data.s3.amazonaws.com/data/sequencing_data_ont/fast5/SGNex_A549_directRNA_replicate6_run1/SGNex_A549_directRNA_replicate6_run1.tar.gz	http://sg-nex-data.s3.amazonaws.com/data/sequencing_data_illumina/fastq/SGNex_A549_Illumina_replicate1_run1/
HCT116	SGNex_Hct116_directRNA_replicate3_run1	RNA002	http://sg-nex-data.s3.amazonaws.com/data/sequencing_data_ont/fast5/SGNex_Hct116_directRNA_replicate3_run1/SGNex_Hct116_directRNA_replicate3_run1.tar.gz	http://sg-nex-data.s3.amazonaws.com/data/sequencing_data_illumina/fastq/SGNex_Hct116_Illumina_replicate3_run1/
NA12878/ GM12878	JHU Runs 1,2,3,4,5; Notts Runs 1,2,3,4,5; OICR Runs 1,2,3,4,5; UBC Runs 1,2,3,4,5	RNA001	https://github.com/nanopore-wgs-consortium/NA12878/blob/master/nanopore-human-transcriptome/fastq_fast5_bulk.md	N/A
MCF-7	SGNex_MCF7_directRNA_replicate2_run3	RNA002	http://sg-nex-data.s3.amazonaws.com/data/sequencing_data_ont/fast5/SGNex_MCF7_directRNA_replicate2_run3/SGNex_MCF7_directRNA_replicate2_run3.tar.gz	N/A
HEK293T _SIRV	HEK293T-1_Direct_RNA, HEK293T-2_Direct_RNA, HEK293T-3_Direct_RNA	RNA002	https://sra-pub-src-2.s3.amazonaws.com/SRR17414694/HEK293T-1_Direct_RNA.tar.gz.4 https://sra-pub-src-2.s3.amazonaws.com/SRR17414696/HEK293T-2_Direct_RNA.tar.gz.3 https://sra-pub-src-2.s3.amazonaws.com/SRR17414698/HEK293T-3_Direct_RNA.tar.gz.3	N/A

Table S16. Basecalling performance of Guppy v6.5.7, RODAN v1.0, GCRTcall v1.0, Dorado v0.8.3, and Coral v2.0 on Human datasets used in downstream analysis.

Dataset	Reads Count	Tool	Read Accuracy (%)		Mismatch (%)		Insertion (%)		Deletion (%)		Read Length (bp)		Unaligned Reads
			mean	median	mean	median	mean	median	mean	median	mean	median	
A549	1254612	Guppy	89.19	89.87	2.92	2.68	2.61	2.37	5.28	4.86	1106	853	22539
		RODAN	90.28	91.26	2.58	2.31	2.97	2.13	4.17	3.71	1046	825	31110
		GCRTcall	91	92.37	2.43	2.07	2.32	1.97	4.25	3.38	1089	842	29416
		Dorado	92.86	94.01	2.07	1.67	1.95	1.64	3.12	2.57	1076	846	19013
		Coral	96.53	97.54	1.05	0.73	0.99	0.66	1.43	0.9	1129	871	14609
HCT116	987488	Guppy	91.4	91.98	2.12	1.94	2.37	2.17	4.1	3.72	1155	885	40341
		RODAN	91.98	92.76	1.94	1.78	2.89	2.09	3.18	2.85	1060	833	38244
		GCRTcall	92.32	93.56	1.94	1.64	2.18	1.88	3.56	2.71	1134	865	44816
		Dorado	93.77	94.57	1.71	1.45	1.83	1.59	2.69	2.25	1136	871	39546
		Coral	97.14	97.85	0.8	0.6	0.87	0.61	1.19	0.76	1170	897	38495
NA12878	8540683	Guppy	89.53	90.48	2.63	2.31	2.23	2	5.61	5.01	990	734	2523486
		RODAN	92.29	93.23	1.82	1.54	2.24	1.66	3.65	3.09	918	671	2108927
		GCRTcall	92.68	94.03	1.8	1.41	1.89	1.58	3.64	2.74	956	700	2309372
		Dorado	90.62	91.4	2.74	2.44	2.14	1.94	4.5	3.99	722	510	2403539
		Coral	96.27	97.57	1.02	0.61	0.95	0.58	1.77	0.96	962	702	2031373
MCF-7	7230921	Guppy	85.05	85.53	4.56	4.36	2.63	2.45	7.76	7.4	728	595	2007657
		RODAN	86.71	87.1	3.89	3.81	1.94	1.71	7.45	6.93	563	478	1672852
		GCRTcall	87.89	88.6	3.7	3.45	2.09	1.86	6.32	5.8	688	556	1573782
		Dorado	88.93	89.87	3.57	3.23	2.07	1.86	5.43	4.83	680	545	1446827
		Coral	93.45	94.78	2.11	1.63	1.12	0.8	3.32	2.44	688	552	1053644
HEK293T _SIRV (sample 1)	1258880	Guppy	89.52	90.51	2.79	2.46	2.64	2.35	5.04	4.46	1008	796	179190
		RODAN	92.27	92.81	1.82	1.77	2.13	1.93	3.77	3.48	758	683	117586
		GCRTcall	90.87	92.45	2.47	2.03	2.42	2.03	4.25	3.27	982	765	177011
		Dorado	92.78	94.22	2.1	1.6	2.03	1.67	3.1	2.4	988	774	162359
		Coral	95.99	97.43	1.2	0.74	1.15	0.71	1.65	0.92	1008	790	144146
	1999948	Guppy	90.32	91.31	2.52	2.19	2.46	2.19	4.7	4.13	1008	792	231440
		RODAN	91.17	92.34	2.24	1.92	2.99	2.1	3.6	3.08	970	761	225991

HEK293T _SIRV (sample 2)		GCRTcall	91.59	93.2	2.21	1.77	2.27	1.9	3.93	2.93	982	763	228510
		Dorado	93.57	94.95	1.81	1.35	1.8	1.46	2.82	2.14	993	775	207788
		Coral	96.56	97.84	1.01	0.6	1.02	0.62	1.41	0.76	1012	790	187493
HEK293T _SIRV (sample 3)	1449168	Guppy	90	91.06	2.65	2.28	2.59	2.32	4.76	4.17	1023	805	185874
		RODAN	90.82	92.07	2.37	2.02	3.06	2.2	3.75	3.17	1022	802	189665
		GCRTcall	91.27	92.91	2.33	1.87	2.38	2	4.02	3.02	997	777	182214
		Dorado	93	94.49	2.02	1.5	1.98	1.62	3.01	2.3	1001	784	169653
		Coral	96.13	97.58	1.16	0.69	1.14	0.7	1.56	0.84	1024	804	151299

The best results are marked as bold text.

Table S17. Details of the top-10 abundant annotated transcripts uniquely discovered by Coral from the A549 dataset.

Chromosome	Transcript ID	Transcript Name	Transcript Length (bp)	Abundance	Coral Accuracy (%)	Dorado Accuracy (%)	GCRTcall Accuracy (%)	RODAN Accuracy (%)	Guppy Accuracy (%)
chr3	ENST00000305366.8	TM4SF1-201	1555	855.99	96.44	93.44	91.47	90.59	89.35
chr17	ENST00000225740.11	ALDH3A1-201	1639	578.05	97.37	93.89	92.16	90.91	90.05
chr22	ENST00000484558.3	CHCHD10-202	700	317.99	96.35	92.5	89.55	88.37	88.4
chr15	ENST00000300289.10	PDIA3-201	3680	302.83	97.74	95.06	93.62	92.45	91.6
chr17	ENST00000494157.6	ALDH3A1-212	1572	264.12	96.9	93.09	91.29	89.81	89.38
chr12	ENST00000262056.14	EIF4B-201	3856	193.65	97.22	94.51	92.65	91.73	90.38
chr17	ENST00000457500.6	ALDH3A1-206	1925	188.03	96.22	93.14	90.17	88.74	88.76
chr2	ENST00000335508.11	SF3B1-201	6463	167	97.7	95.8	94.08	93.13	92.32
chrX	ENST00000335972.11	UBA1-201	3572	164.29	97.18	94.44	92.59	90.91	90.57
chr14	ENST00000216802.10	PSME2-201	793	153	97.33	94.92	93.09	92.17	91.02

1. The best results are marked as bold text.
2. Abundance is reported as the read count estimate of a transcript isoform. Each read contributes at most 1 count, either to a single transcript isoform or distributed as fractional counts to multiple transcript isoforms.
3. “Coral Accuracy”, “Dorado Accuracy”, “GCRTcall Accuracy”, “RODAN Accuracy” and “Guppy Accuracy” refer to the median read accuracy (by transcriptome alignment) for the reads assigned to an annotated transcript discovered only in Coral.

Table S18. Details of the top-10 abundant annotated transcripts uniquely discovered by Coral from the HCT116 dataset.

Chromosome	Transcript ID	Transcript Name	Transcript Length (bp)	Abundance	Coral Accuracy (%)	Dorado Accuracy (%)	GCRTcall Accuracy (%)	RODAN Accuracy (%)	Guppy Accuracy (%)
chr22	ENST00000484558.3	CHCHD10-202	700	486.44	96.97	92.82	91.61	90.41	90.75
chr11	ENST00000322028.5	POLR2L-201	876	462	98.94	95.43	93.54	92.75	92.68
chr15	ENST00000300289.10	PDIA3-201	3680	377.62	98.08	95.48	94.65	93.79	93.27
chr21	ENST00000290299.7	ATP5PO-201	756	328	98.28	95.96	94.63	94.01	93.47
chr7	ENST00000450169.7	LSM5-208	2214	187.9	98.03	95.82	94.21	93.41	93.1
chr3	ENST00000346169.7	EIF4G1-202	5405	163.7	97.34	94.81	93.52	92.55	92.61
chr14	ENST00000216802.10	PSME2-201	793	128.98	97.79	95.07	94.15	92.53	92.61
chr11	ENST00000379719.8	IPO7-201	6162	116	97.01	94.85	93.18	92.52	92.43
chr6	ENST00000412585.7	HLA-B-249	1536	115	96.3	91.37	91.76	91.5	89.17
chr1	ENST00000360001.12	SDF4-202	1933	94.58	97.34	94.34	93.55	92.79	91.8

1. The best results are marked as bold text.
2. Abundance is reported as the read count estimate of a transcript isoform. Each read contributes at most 1 count, either to a single transcript isoform or distributed as fractional counts to multiple transcript isoforms.
3. “Coral Accuracy”, “Dorado Accuracy”, “GCRTcall Accuracy”, “RODAN Accuracy” and “Guppy Accuracy” refer to the median read accuracy (by transcriptome alignment) for the reads assigned to an annotated transcript discovered only in Coral.

Table S19. Details of the top-10 abundant annotated transcripts uniquely discovered by Coral from the NA12878 dataset.

Chromosome	Transcript ID	Transcript Name	Transcript Length (bp)	Abundance	Coral Accuracy (%)	Dorado Accuracy (%)	GCRTcall Accuracy (%)	RODAN Accuracy (%)	Guppy Accuracy (%)
chr6	ENST00000376809.10	HLA-A-203	1535	4656.61	97.77	90.35	93.95	93.26	89.71
chr6	ENST00000376806.9	HLA-A-202	1854	2330.06	97.96	90.59	94.3	93.59	90.05
chr14	ENST00000216802.10	PSME2-201	793	1942.1	97.6	92.46	94.46	93.38	91.7
chr22	ENST00000484558.3	CHCHD10-202	700	1401.88	96.63	90.67	92.71	91.62	90.23
chr12	ENST00000262056.14	EIF4B-201	3856	1106.54	96.8	91.37	93.11	92.32	90.86
chr6	ENST00000706900.1	HLA-A-218	1472	564.65	98.02	90.27	94.03	93.49	89.76
chr15	ENST00000300030.8	CIAO2A-201	919	563.76	97.05	92.46	94.09	93.06	91.55
chr17	ENST00000613922.2	CCL3-204	780	523	96.51	91.07	92.95	92.35	90.02
chr5	ENST00000264932.11	SDHA-201	2693	409.49	97.55	92.3	93.82	93.03	91.9
chr10	ENST00000369937.5	CUEDC2-201	1100	379.81	96.53	91.33	93	92.05	90.23

1. The best results are marked as bold text.
2. Abundance is reported as the read count estimate of a transcript isoform. Each read contributes at most 1 count, either to a single transcript isoform or distributed as fractional counts to multiple transcript isoforms.
3. “Coral Accuracy”, “Dorado Accuracy”, “GCRTcall Accuracy”, “RODAN Accuracy” and “Guppy Accuracy” refer to the median read accuracy (by transcriptome alignment) for the reads assigned to an annotated transcript discovered only in Coral.

Table S20. Training configuration

Batch size	128
Epoch number	10
Signal chunk length	4096
Optimizer	AdamW
Optimizer Betas	(0.9, 0.999)
Optimizer eps	1e-08
Optimizer weight decay	0.01
Initial learning rate	2e-4
Learning rate scheduler	Linear Warmup Cosine Annealing
Warmup steps	500
Automatic Mixed Precision	Float16
Gradient Norm Clip	1.0

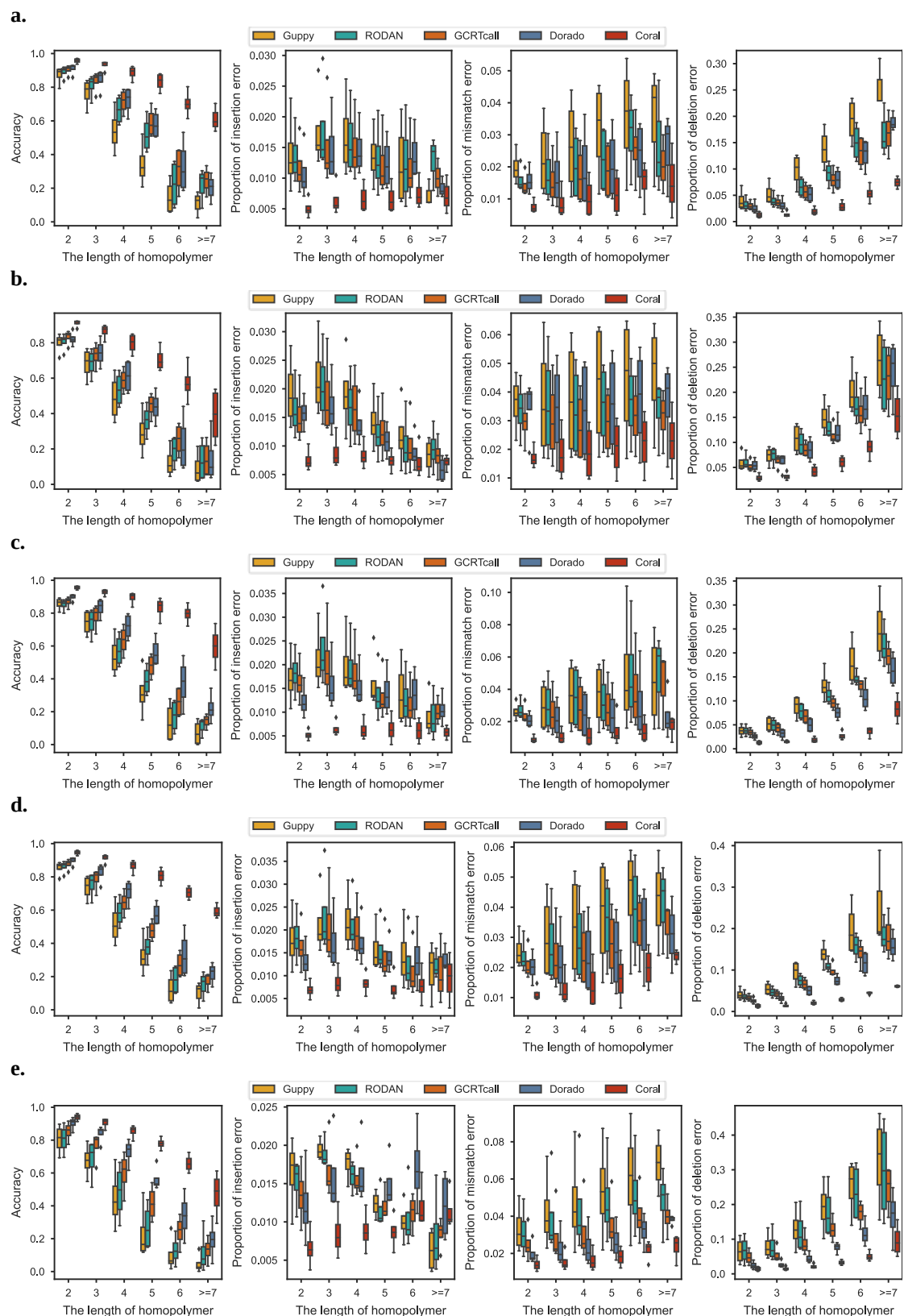


Fig. S1. Comparison of accuracy, insertion proportion, mismatch proportion, and deletion proportion for homopolymers of varying lengths across Guppy, RODAN, GCRTcall, Dorado and Coral. **a**, Results for the Arabidopsis dataset. **b**, Results for the Mouse dataset. **c**, Results for the Yeast dataset. **d**, Results for the Poplar dataset. **e**, Results for the Zebrafish dataset. Statistics were derived from the successfully aligned subset of $n = 100,000$ input reads per dataset. In the box plots, the center line represents the median, the box limits indicate the 25th and 75th percentiles, and the whiskers extend to $1.5 \times$ the interquartile range (IQR). Data points beyond the whiskers are potential outliers.

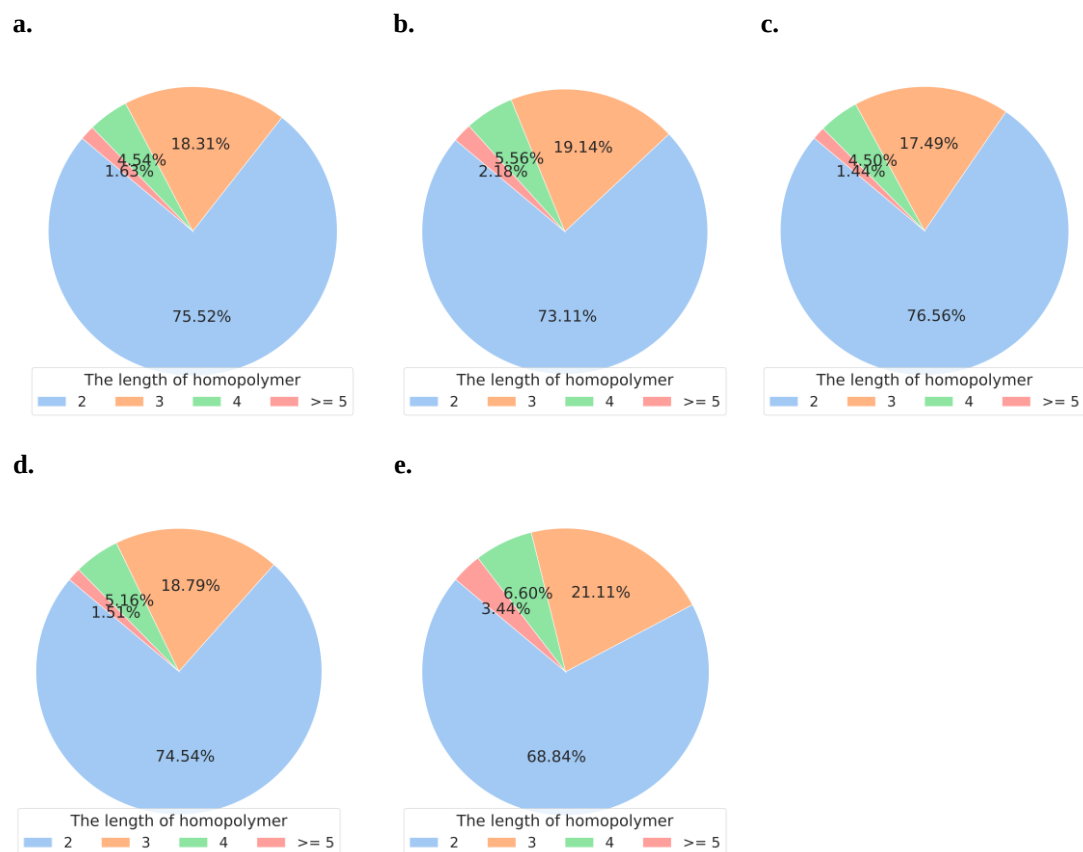


Fig. S2. Proportion of different homopolymer lengths. **a**, Results for the Arabidopsis dataset. **b**, Results for the Mouse dataset. **c**, Results for the Yeast dataset. **d**, Results for the Poplar dataset. **e**, Results for the Zebrafish dataset. Distributions were calculated from the successfully aligned subset of $n = 100,000$ input reads per dataset.

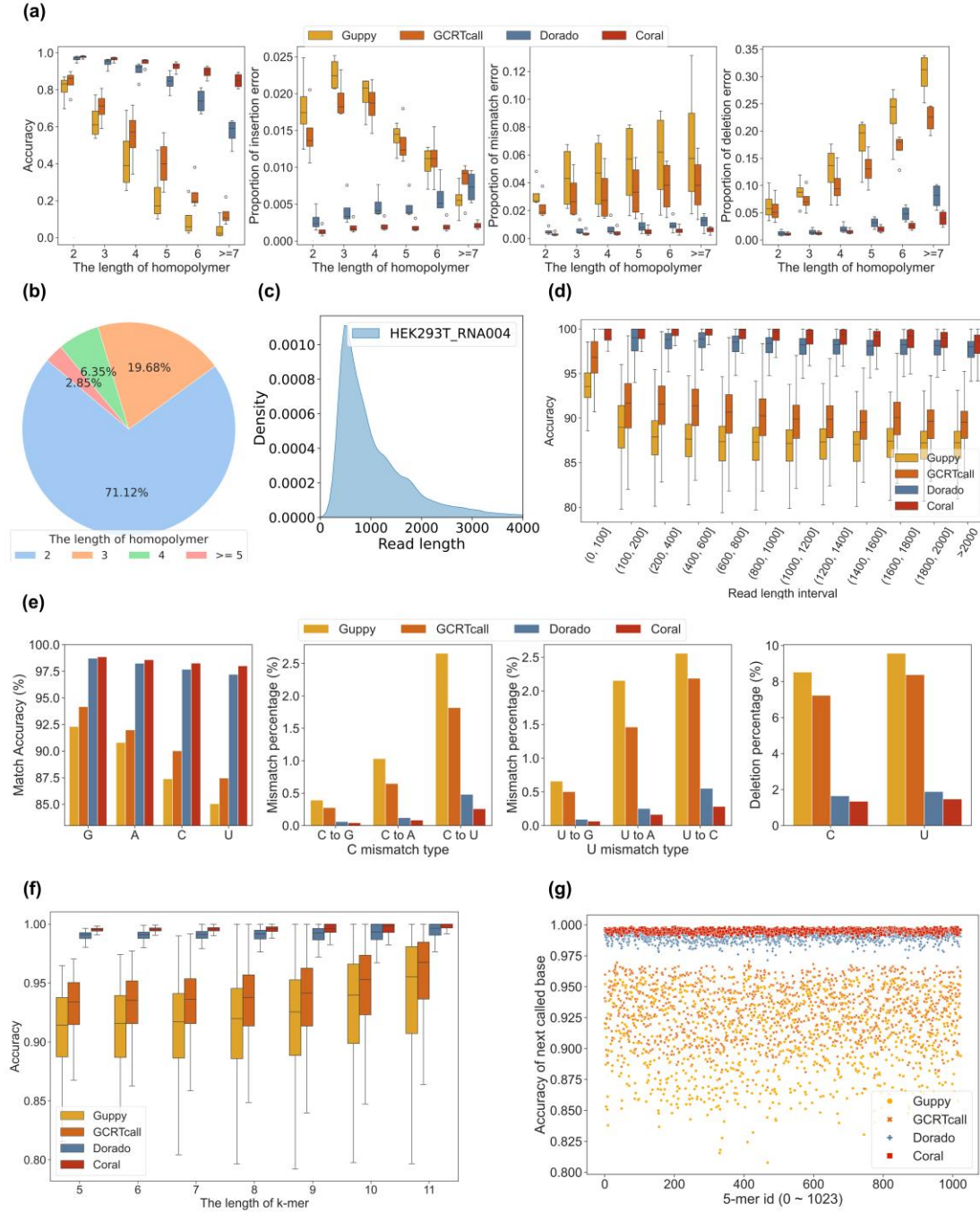


Fig. S3. Performance of the basecallers on homopolymers, various read lengths, individual bases, and k-mer contextual basecalling in the HEK293T RNA004 dataset. **a**, Accuracy and proportions of insertion, mismatch, and deletion errors for homopolymers of varying lengths. **b**, Proportion of different homopolymer lengths. **c**, Read length distribution of Coral output. **d**, Read accuracy stratified by read length intervals. **e**, Average four-base match accuracy and mismatch/deletion error of C/U at the single nucleotide level. **f**, Accuracy of k-mer predictions (lengths 5 to 11). **g**, Scatter plot of next-base prediction accuracy given a correct 5-mer context. All statistics were derived from $n = 100,000$ randomly sampled reads from the HEK293T RNA004 dataset. For all box plots in **a**, **d** and **f**, the center line represents the median, the box limits indicate the 25th and 75th percentiles, and the whiskers extend to $1.5 \times$ the interquartile range (IQR). Data points beyond the whiskers are plotted individually as potential outliers.

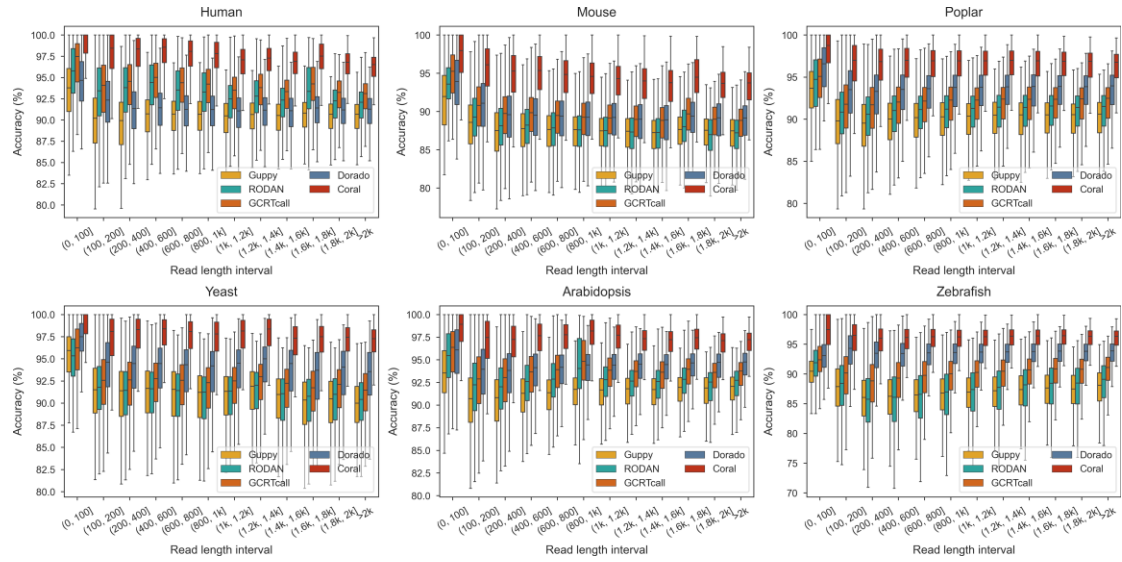


Fig. S4. The boxplots showing the basecalling accuracy as a function of read length for the six testing datasets. Statistics were derived from the successfully aligned subset of $n = 100,000$ input reads per dataset. The center line represents the median, the box limits indicate the 25th and 75th percentiles, and the whiskers extend to $1.5\times$ the interquartile range.

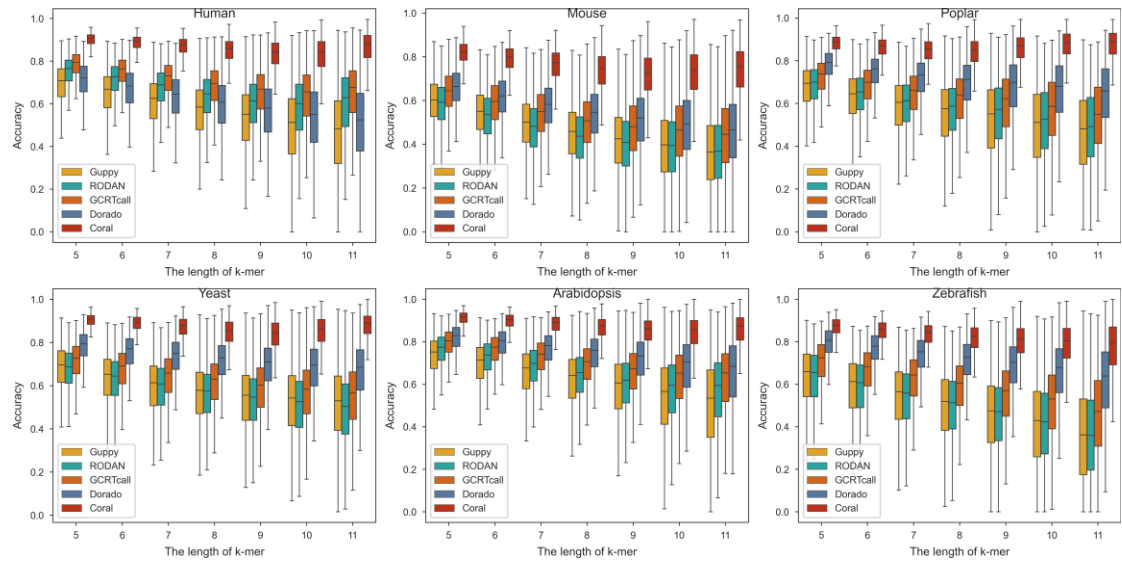


Fig. S5. The boxplots showing the accuracy of k-mer predictions (from lengths 5 to 11) for Guppy, RODAN, GCRTcall, Dorado and Coral in different testing samples. Statistics were derived from the successfully aligned subset of $n = 100,000$ input reads per dataset. The center line represents the median, the box limits indicate the 25th and 75th percentiles, and the whiskers extend to $1.5\times$ the interquartile range.

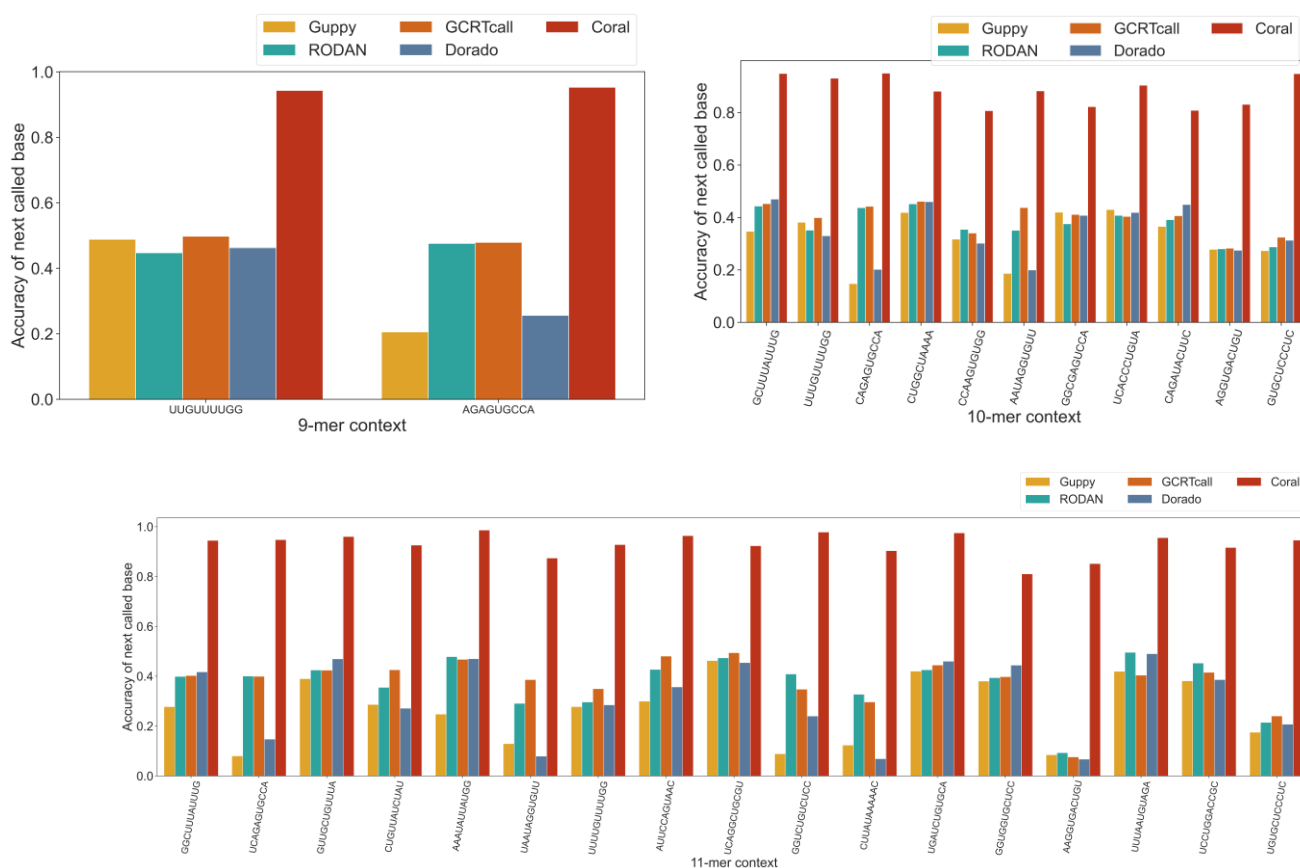


Fig. S6. Bar plots depicting the accuracy of the next called base given several specific 9-mer, 10-mer and 11-mer contexts in the human test dataset. Statistics were derived from the successfully aligned subset of $n = 100,000$ input reads.

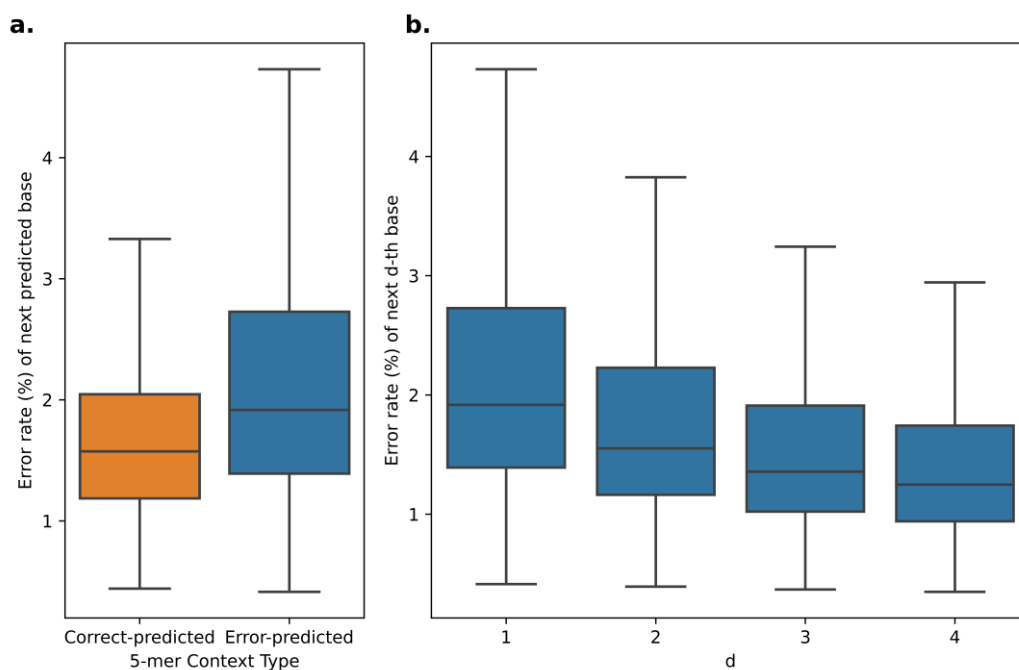


Fig. S7. Box plots depicting the impact of error-predicted 5-mer context on the prediction of the subsequent bases. **a**, Comparison of the error rate for the next predicted base given a correct or erroneous predicted 5-mer context. **b**, Error rate of d -th subsequent base given an incorrectly predicted 5-mer context. Statistics were derived across all six RNA001/002 test datasets. In the box plots, the center line represents the median, the box limits indicate the 25th and 75th percentiles, and the whiskers extend to $1.5\times$ the interquartile range.

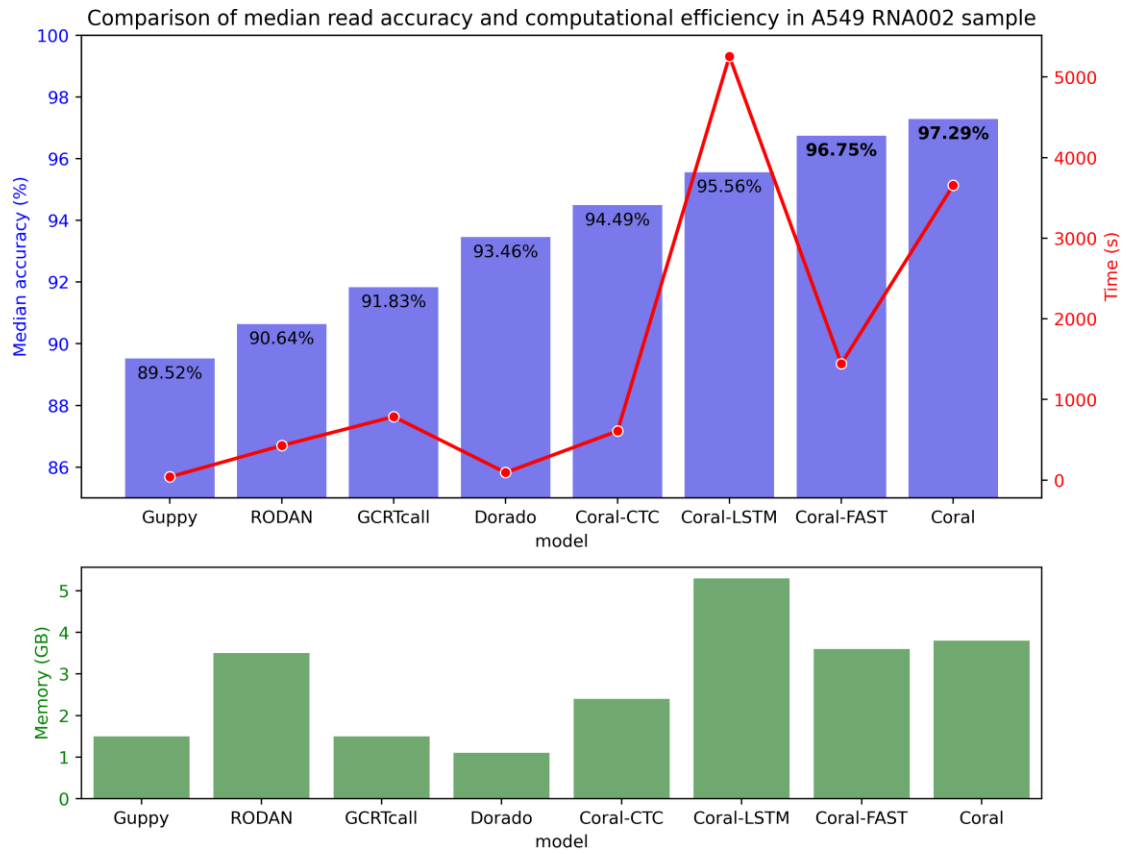


Fig. S8. Median basecalling accuracy (blue bars, left axis) and decoding time (red points, right axis) for each model on $n=16,000$ sampled reads from A549 RNA002 data. Memory usage is shown below (green bars).

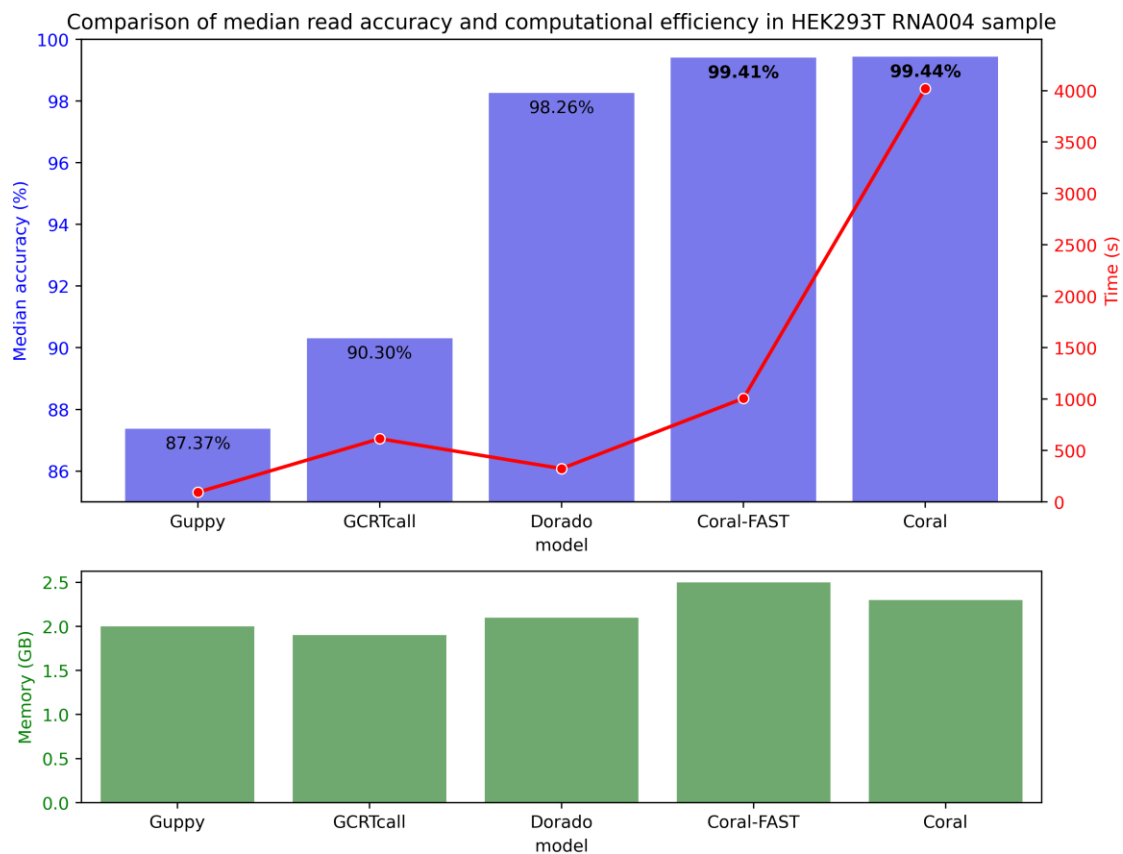


Fig. S9. Basecalling accuracy and runtime on $n=20,000$ sampled reads from HEK293T RNA004 data across different methods. Memory usage is shown below (green bars).

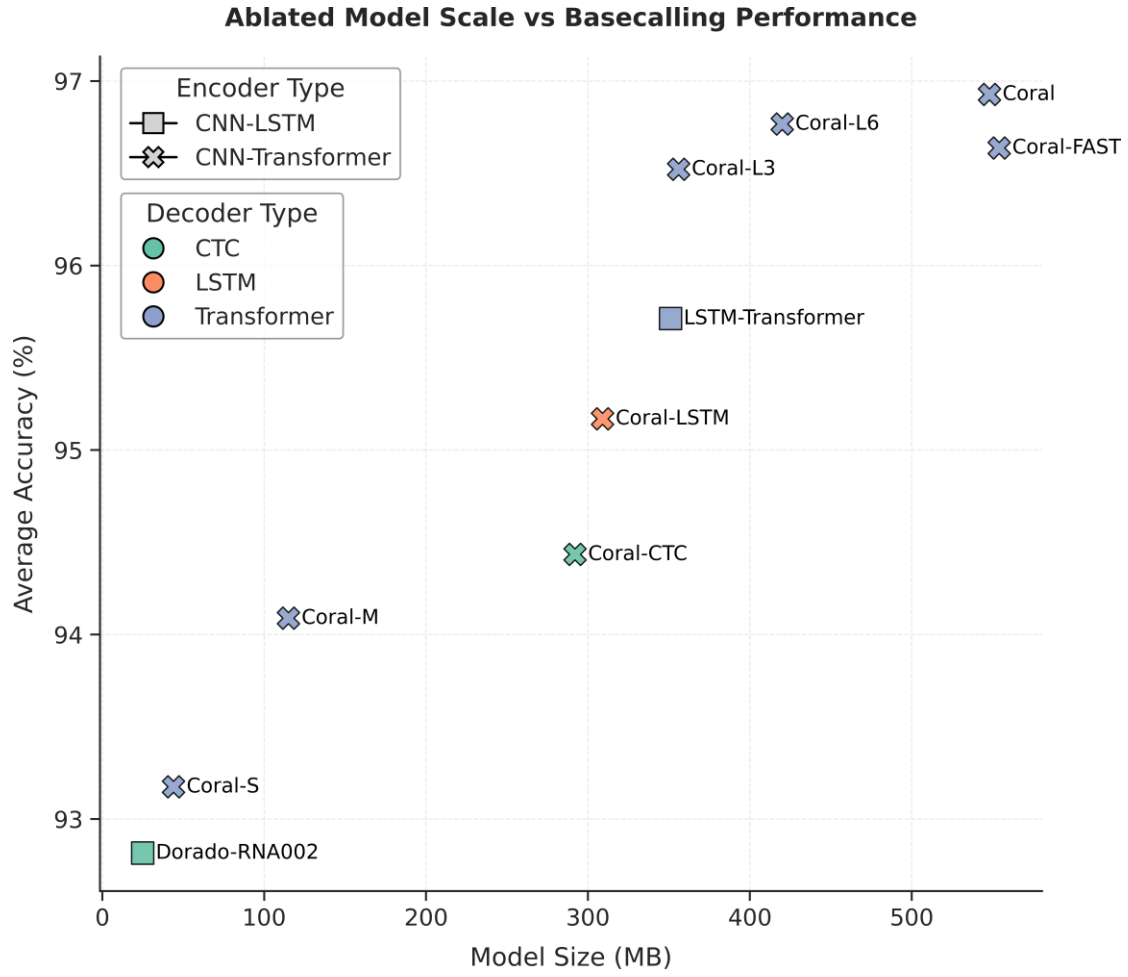


Fig. S10. Basecalling performance of Coral ablated models with different size on RNA001/002 test datasets. ‘Coral-S’ represents a small-scale variant of the Coral model. ‘Coral-M’ represents a medium-scale variant of the Coral model. ‘Coral-L3’, ‘Coral-L6’ and ‘Coral’ represent large-scale variants of the Coral model with 3/6/12 decoder layers. ‘Coral-CTC’ and ‘Coral-LSTM’ are two Coral decoder variants that replace the Transformer decoder with ‘CTC’ and ‘LSTM’. ‘LSTM-Transformer’ is Coral encoder variant that replace the CNN-Transformer encoder with CNN-LSTM encoder. The basecalling performance is measured using average median accuracy across six RNA001/002 test datasets.

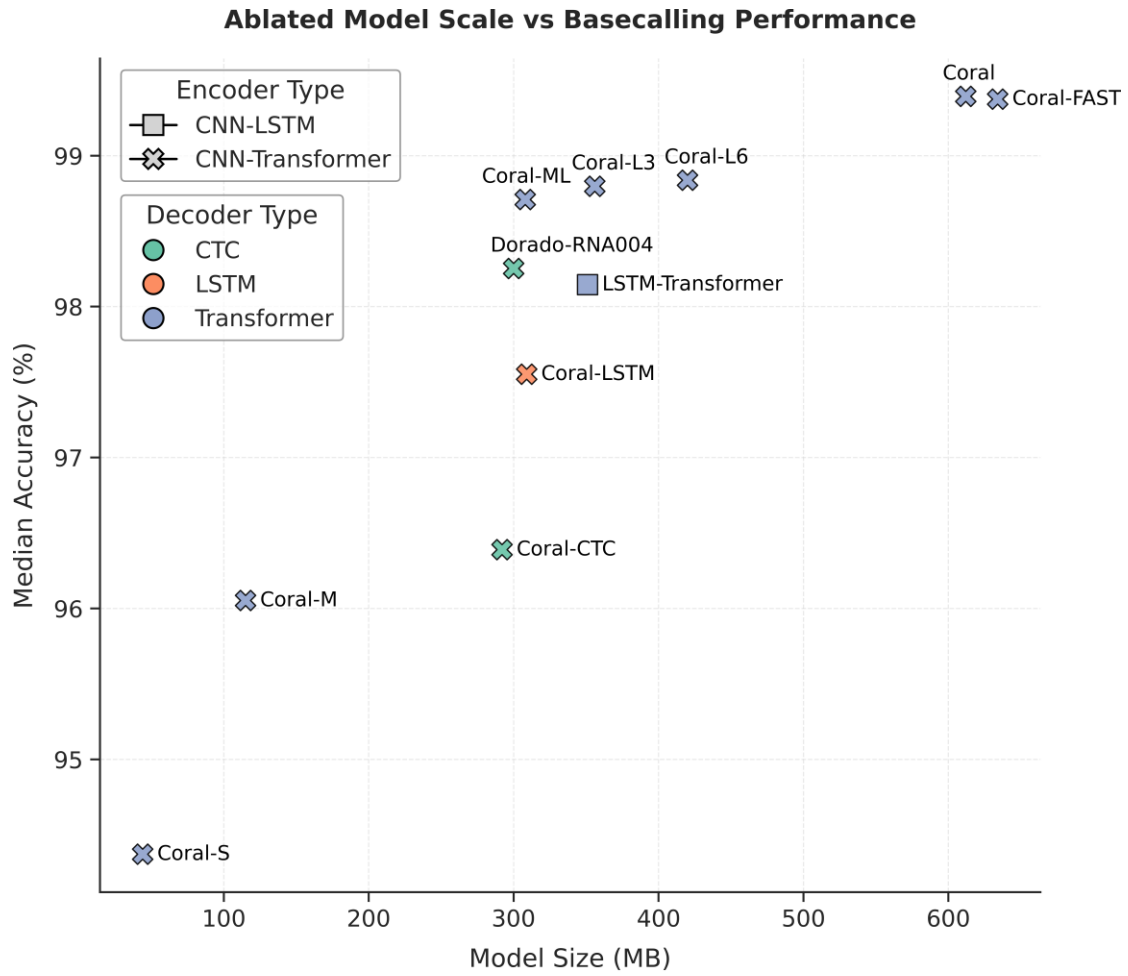


Fig. S11. Basecalling performance of Coral ablated models with different size on HEK293T RNA004 dataset. ‘Coral-S’ represents a small-scale variant of the Coral model. ‘Coral-M’ represents a medium-scale variant of the Coral model. ‘Coral-ML’ represents a medium-large-scale variant of the Coral model. ‘Coral-L3’, ‘Coral-L6’ and ‘Coral’ represent large-scale variants of the Coral model with 3/6/15 decoder layers. ‘Coral-CTC’ and ‘Coral-LSTM’ are two Coral decoder variants that replace the Transformer decoder with ‘CTC’ and ‘LSTM’. ‘LSTM-Transformer’ is Coral encoder variant that replace the CNN-Transformer encoder with CNN-LSTM encoder. The basecalling performance is measured using median accuracy on 623k sampled reads from HEK293T RNA004 data.

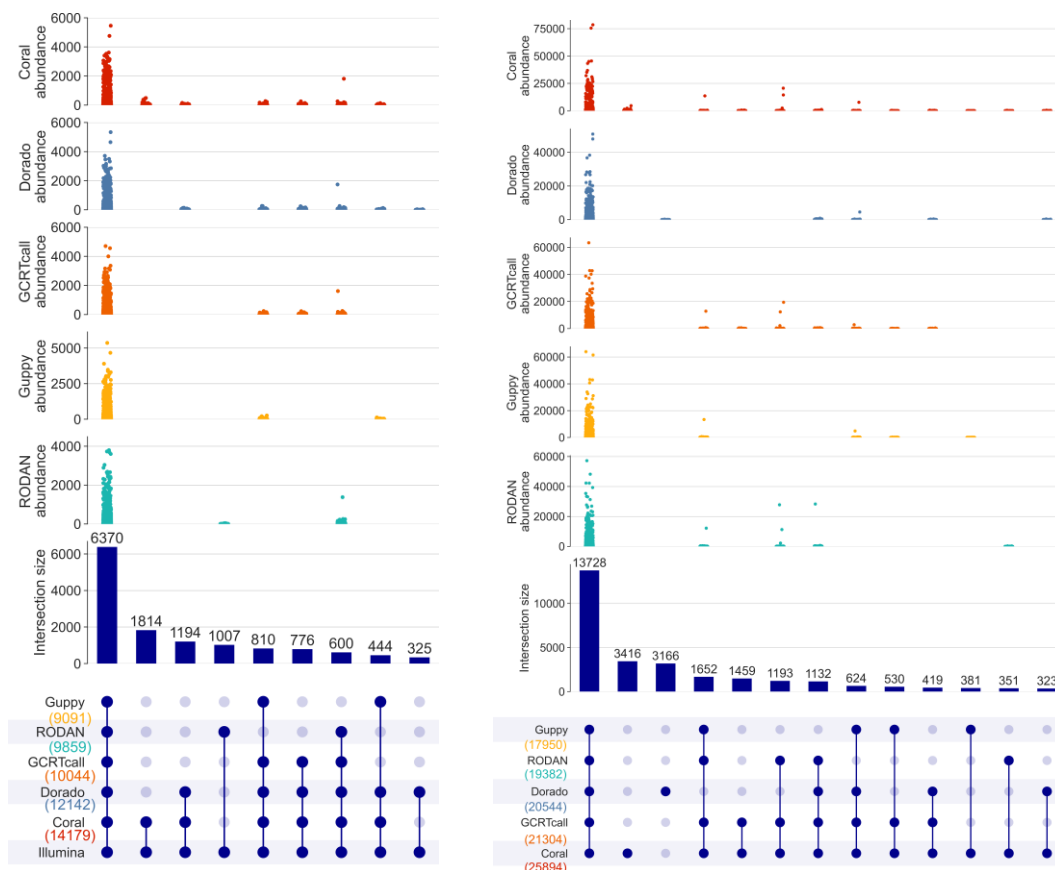
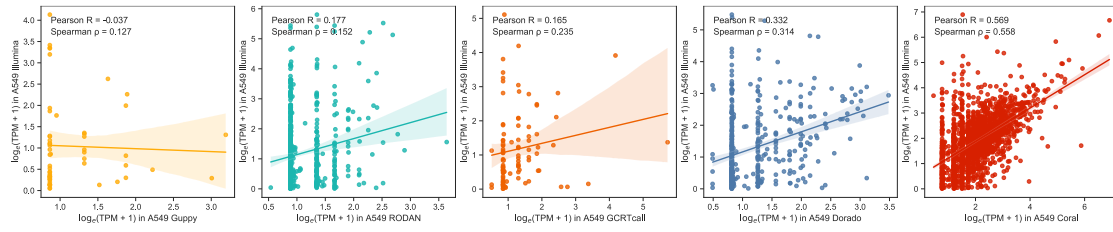


Fig. S12. Overlap analysis of detected isoforms across basecallers in the HCT116 (left, overlapped with Illumina-discovered) and NA12878 (right) datasets. Upset plot (bottom) displays the intersection size of annotated transcripts discovered by Illumina (only for HCT116 dataset) and the specific basecallers indicated by linked filled dots below each bar. Intersection sets smaller than 300 elements are not displayed. The total number of annotated transcripts identified by different basecallers is shown in parentheses, with each method distinguished by a unique color. Strip plot (upper) shows the abundance distribution of each basecaller across different annotated transcript sets.

a.



b.

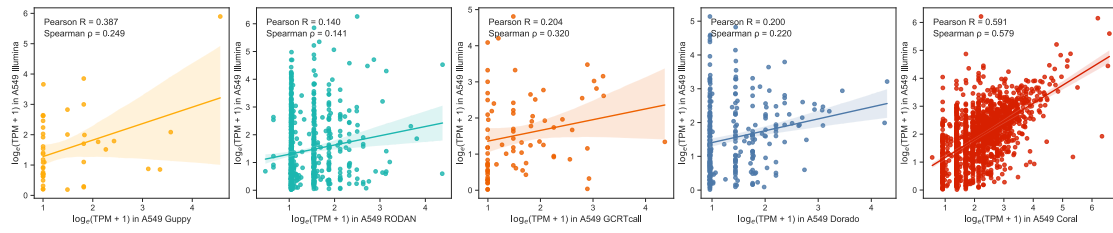


Fig. S13. Correlation between transcript expression estimates obtained from Coral, Guppy, RODAN, Dorado, and GCRTcall reads (using ESPRESSO) and those from Illumina RNA-seq reads (using Salmon), focusing on transcripts uniquely detected by each basecaller but also identified in the Illumina RNA-seq data. a, Results for the A549 dataset ($n = 1796, 434, 95, 54$ and 629 transcripts for Coral, Dorado, GCRTcall, Guppy, and RODAN respectively). b, Results for the HCT116 dataset ($n = 1506, 304, 78, 45$ and 838 transcripts for Coral, Dorado, GCRTcall, Guppy and RODAN respectively). The solid line represents the linear regression fit, and the shaded band indicates the 95% confidence interval. Pearson's correlation coefficient and Spearman's rank correlation coefficient are provided.

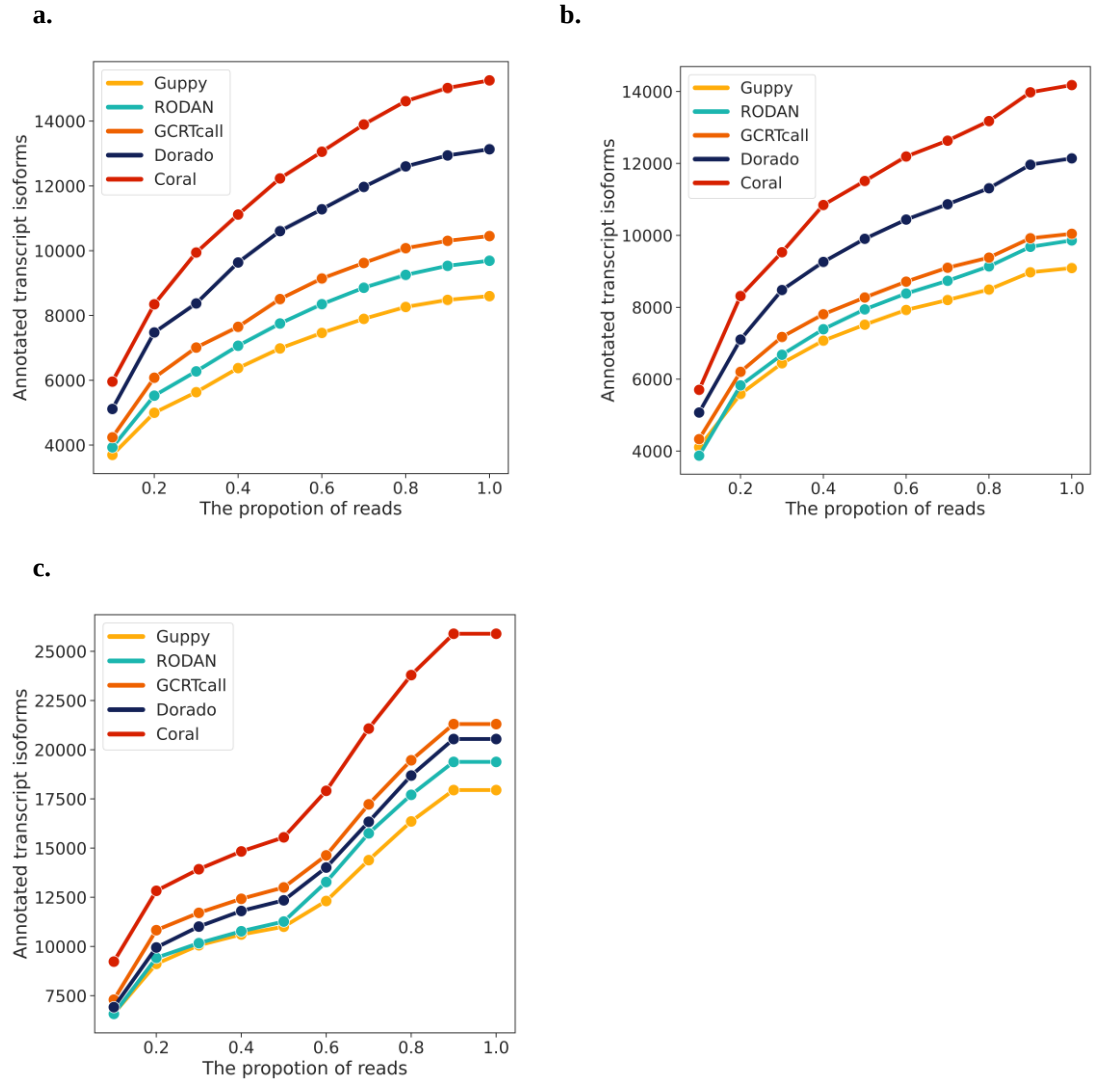


Fig. S14. Number of annotated human transcripts identified using randomly sampled reads at various read proportions. a, Results for the A549 dataset (100% corresponds to $n = 1254612$ reads). **b,** Results for the HCT116 dataset (100% corresponds to $n = 987488$ reads). **c,** Results for the NA12878 dataset (100% corresponds to $n = 8540683$ reads).

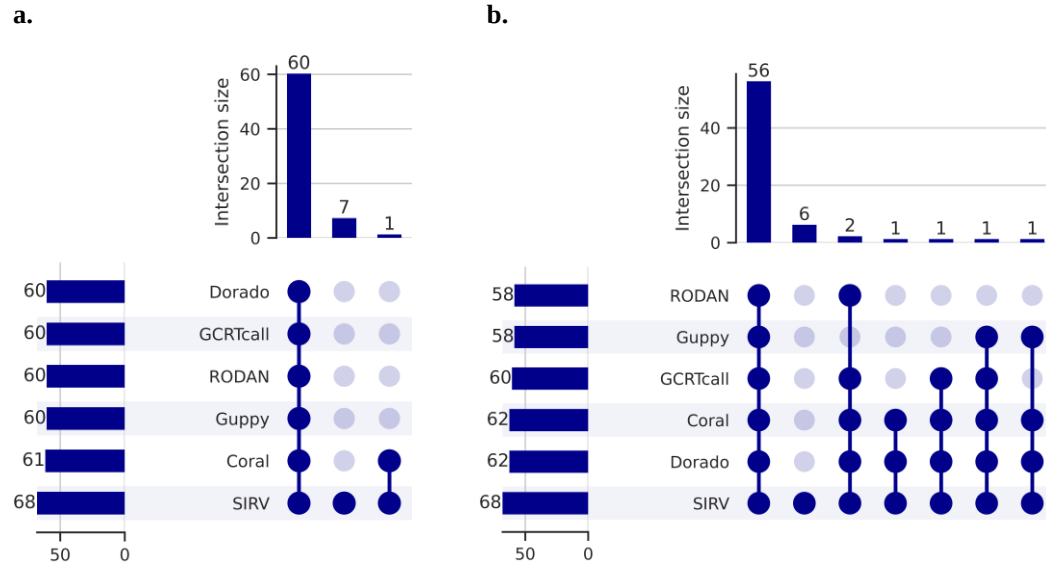


Fig. S15. The number of SIRV transcripts discovered by Coral, Guppy, RODAN, GCRTcall, and Dorado for another two samples HEK293T-2_Direct_RNA (a) and HEK293T-3_Direct_RNA (b). The analysis was based on $n = 68$ known SIRV transcripts. Upset plot displays the intersection size of SIRV transcripts discovered by the specific basecallers indicated by linked filled dots below each bar. The total number of SIRV transcripts identified by different basecallers is shown in the left-hand horizontal bars.

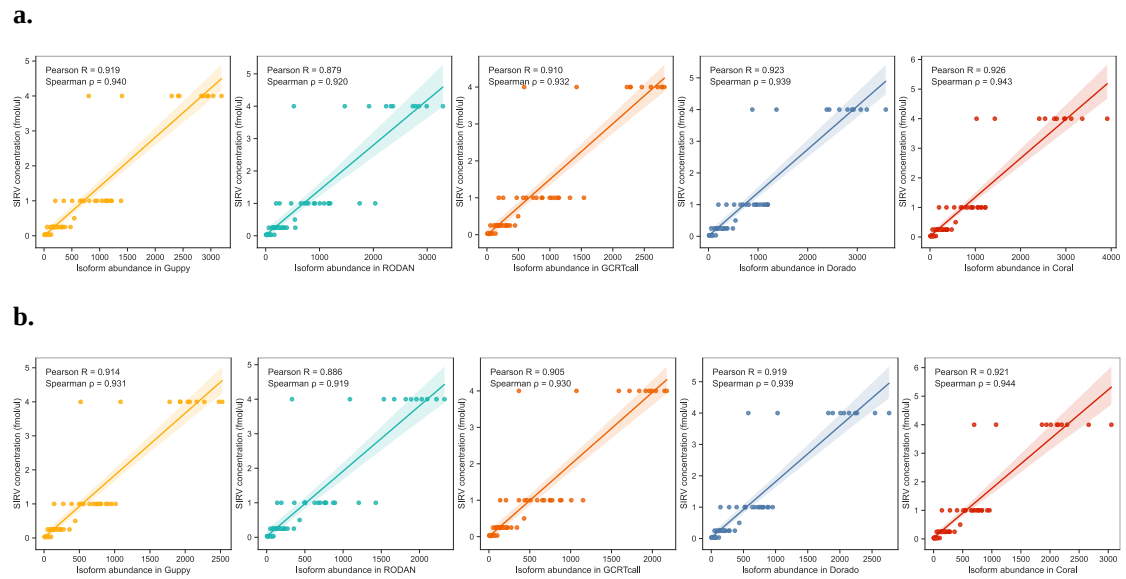
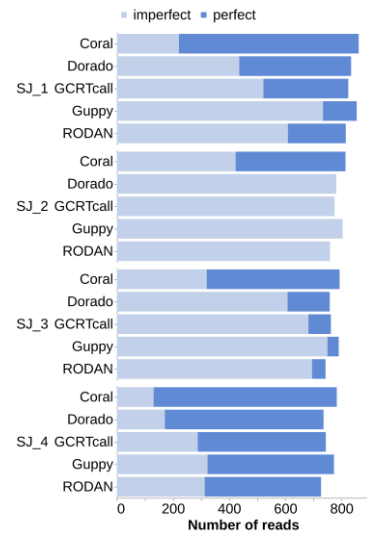
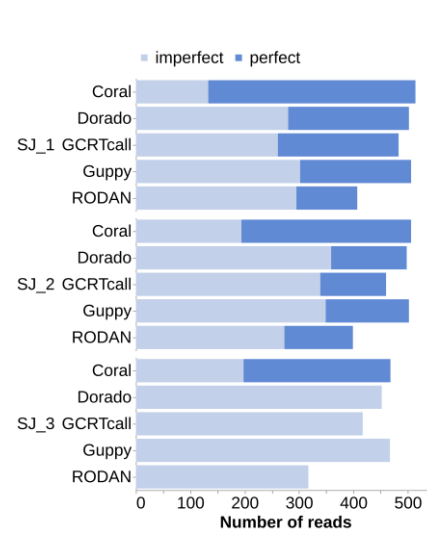


Fig. S16. Correlation between transcript expression estimates obtained from Coral, Guppy, RODAN, GCRTcall and Dorado reads (using ESPRESSO) and the SIRV truth concentration levels on another two HEK293T samples. a, Results for the sample HEK293T-2_Direct_RNA. **b,** Results for the sample HEK293T-3_Direct_RNA. The analysis was performed on $n = 68$ SIRV transcripts. The solid line represents the linear regression fit, and the shaded band indicates the 95% confidence interval. Pearson's correlation coefficient and Spearman's rank correlation coefficient are provided.

a. ENST00000305366.8 (A549)



b. ENST00000484558.3 (HCT116)



c. ENST00000376809.10 (NA12878)

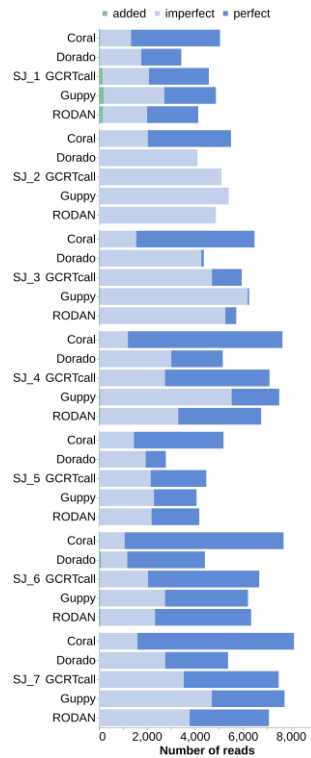


Fig. S17. Number of reads with perfect and imperfect alignments around splice junctions (SJs) for ENST00000305366.8 in the A549 dataset, ENST00000484558.3 in the HCT116 dataset, and ENST00000376809.10 in the NA12878 dataset. Perfect alignment refers to error-free (i.e., no mismatches, insertions, or deletions) within the 10-nt distance of the corresponding splice sites. The "added" refers to the splice junction that was missing in the original alignment but recovered after ESPRESSO correction.

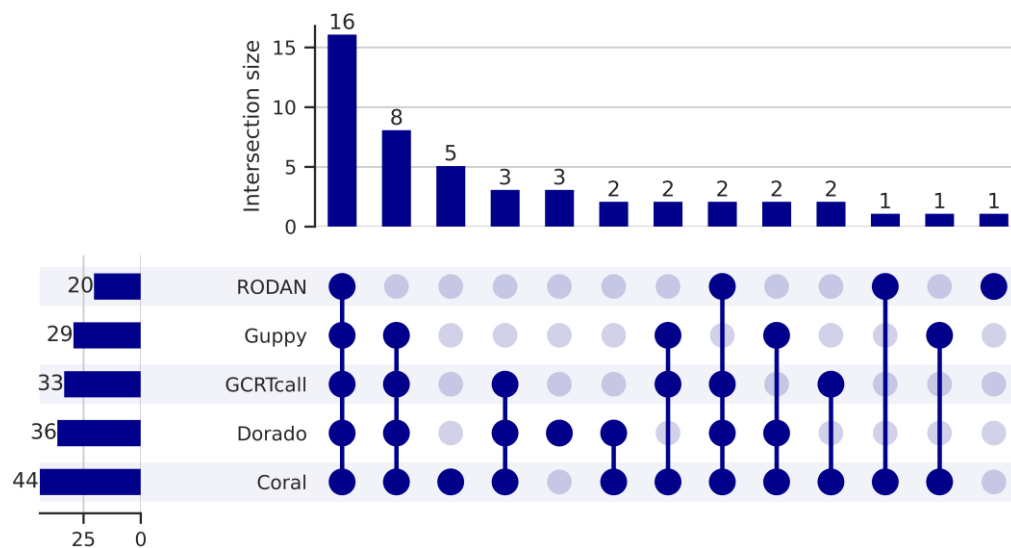


Fig. S18. Upset plot showing previously validated fusion genes rediscovered by Guppy, RODAN, GCRTcall, Dorado, and Coral using JAFFAL in the MCF-7 cancer line.

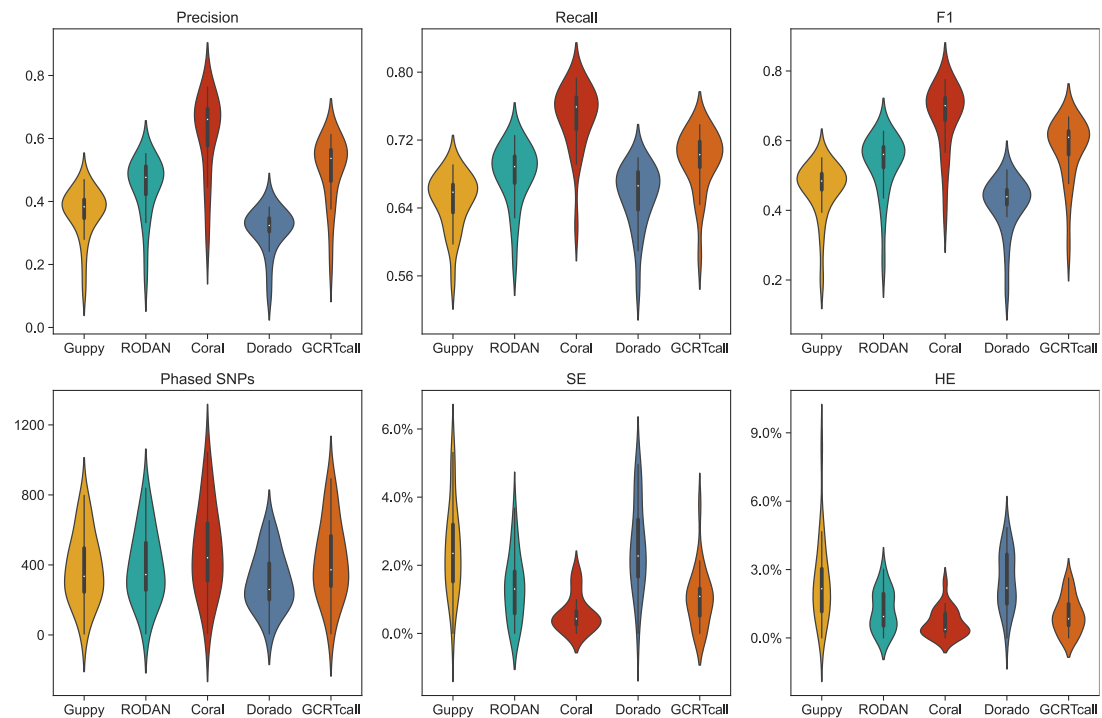


Fig. S19. Violin plots of SNP identification and haplotype performance compared to GIAB gold-standard across different chromosomes using ONT DRS data from GM12878. Precision, Recall, and F1 score were calculated by comparing the chromosome position, allele, and genotype between unphased SNPs against the GIAB gold standard. SE and HE, i.e., switch error rate and Hamming error rate, were calculated by comparing the haplotype differences between the phased blocks and the standard at the identical heterozygous variants. In the violin plots, the distribution of the data is shown by the kernel density estimate with an embedded boxplot: the box indicates the interquartile range (IQR), the white dot marks the median, and whiskers extend to $1.5 \times$ IQR. Statistics were derived from $n = 23$ chromosomes.

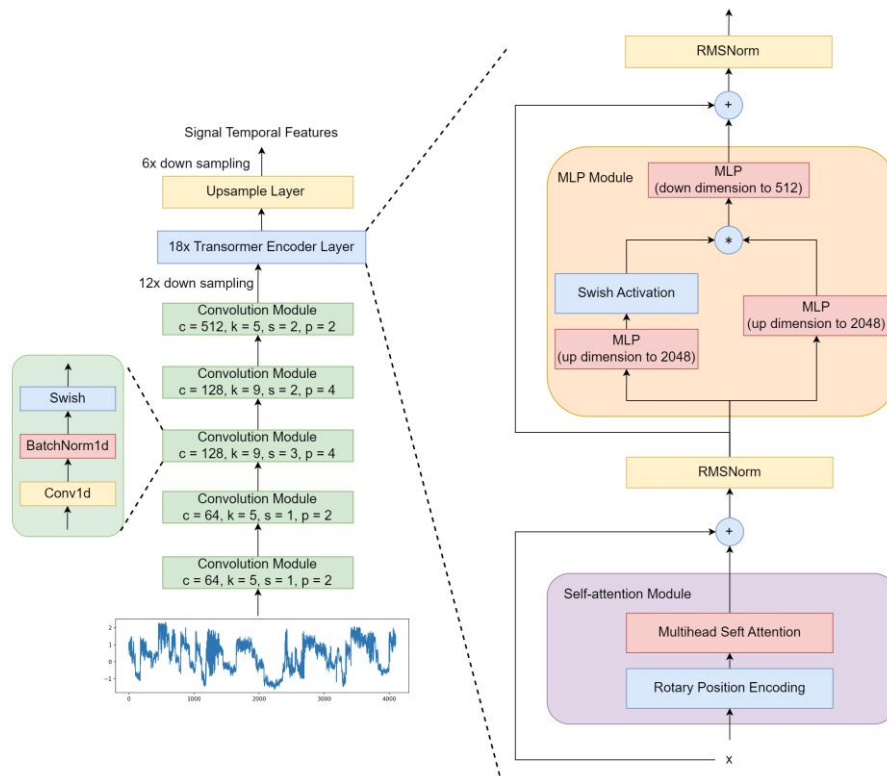


Fig. S20: Detailed architecture of CNN-Transformer Encoder in Coral model. The labels ('c', 'k', 's', and 'p') in the convolution module are represented as channel number, kernel size, stride size, and padding size respectively.

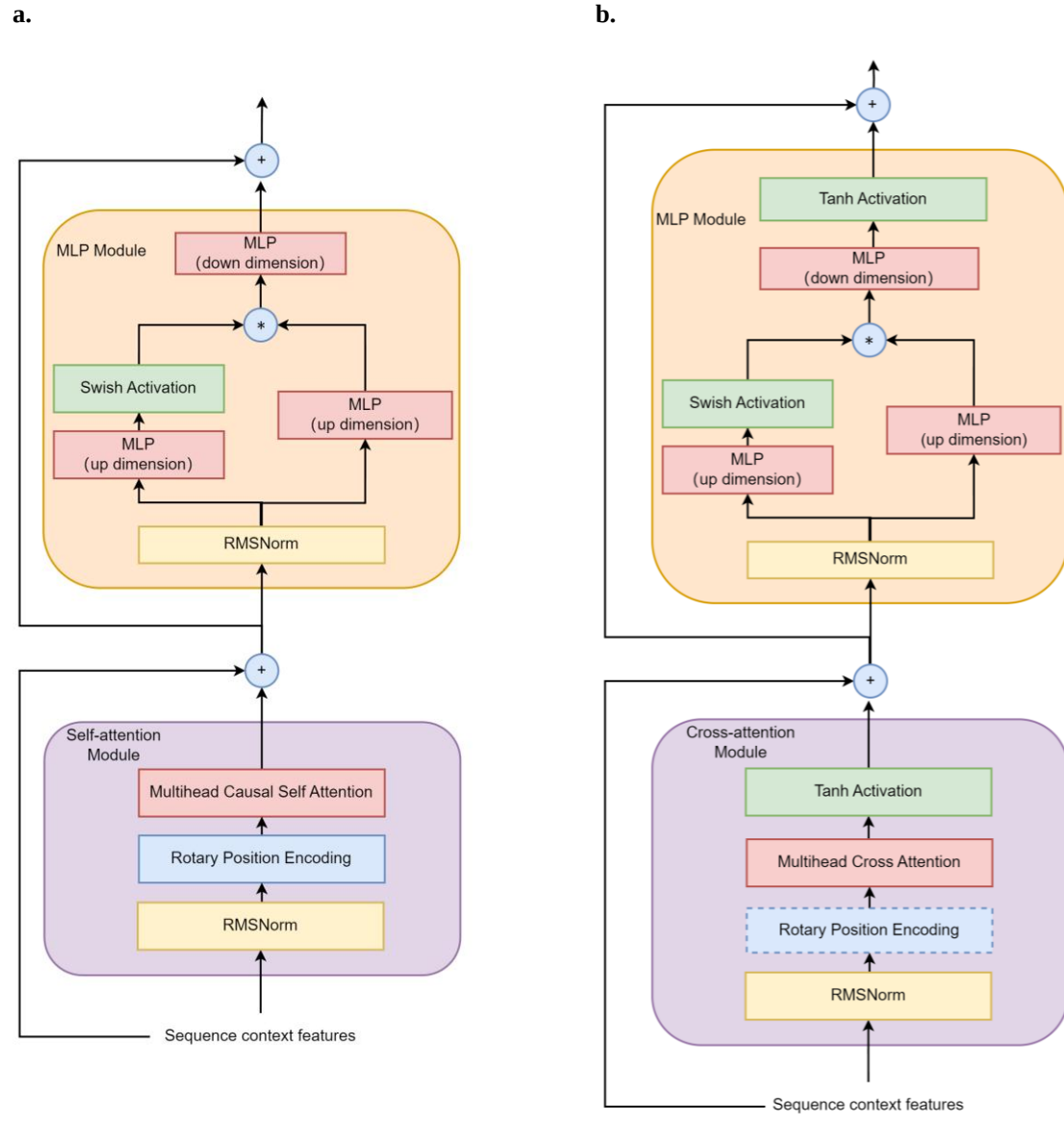


Fig. S21: Detailed architecture of Transformer decoder layer in Coral model. a, Architecture of transformer decoder layer with multi-head causal self attention. **b,** Architecture of transformer decoder layer with multi-head cross attention. Position encoding is optional in cross attention layer.

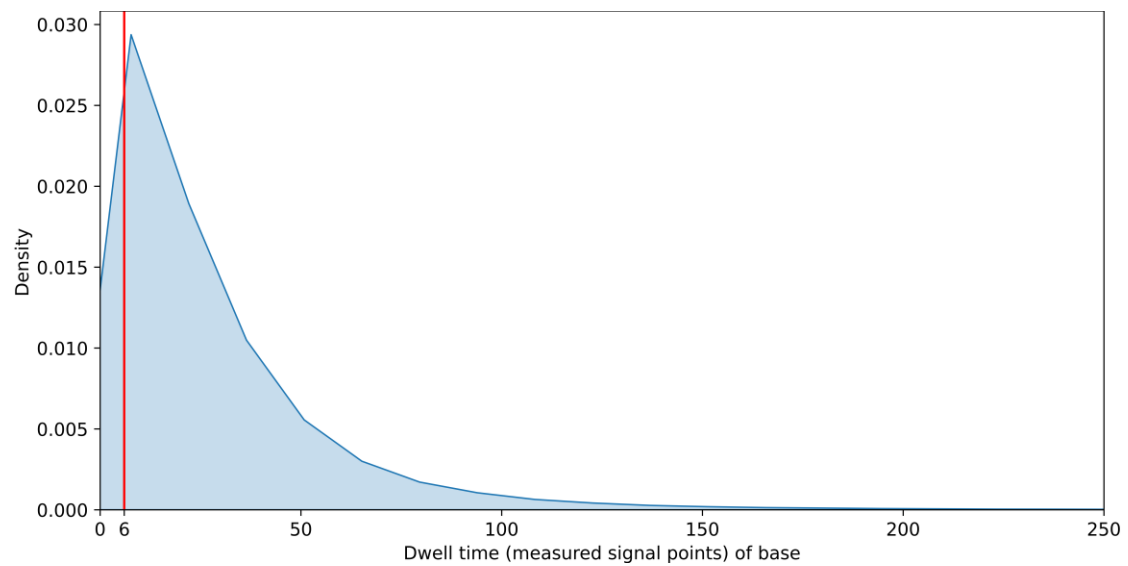


Fig. S22: Distribution of dwell times (measured signal points per base) calculated from one million bases in the HEK293T RNA004 dataset. Dwell times were obtained by aligning the ribonucleotide sequence to the raw electrical signal using the ont-remora tool.

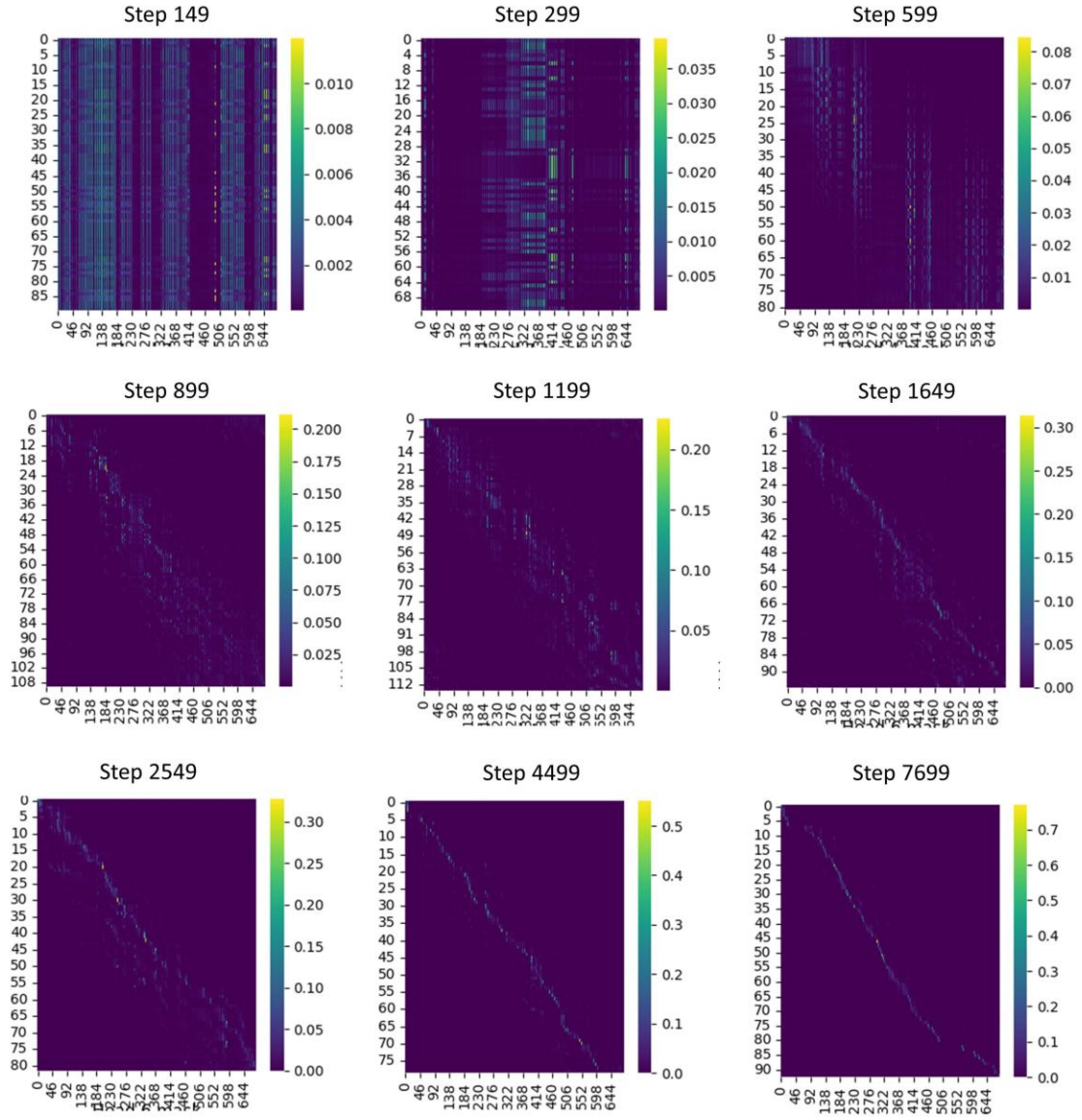


Fig. S23: Visualization of the cross-attention alignment matrix α_c of final cross-attention layer at different training steps. The matrix shows the alignment probability between the ribonucleotide sequence (y-axis) and signal representations (x-axis). As training progresses from Step 149 to Step 7699, the attention becomes increasingly focused and diagonally aligned.

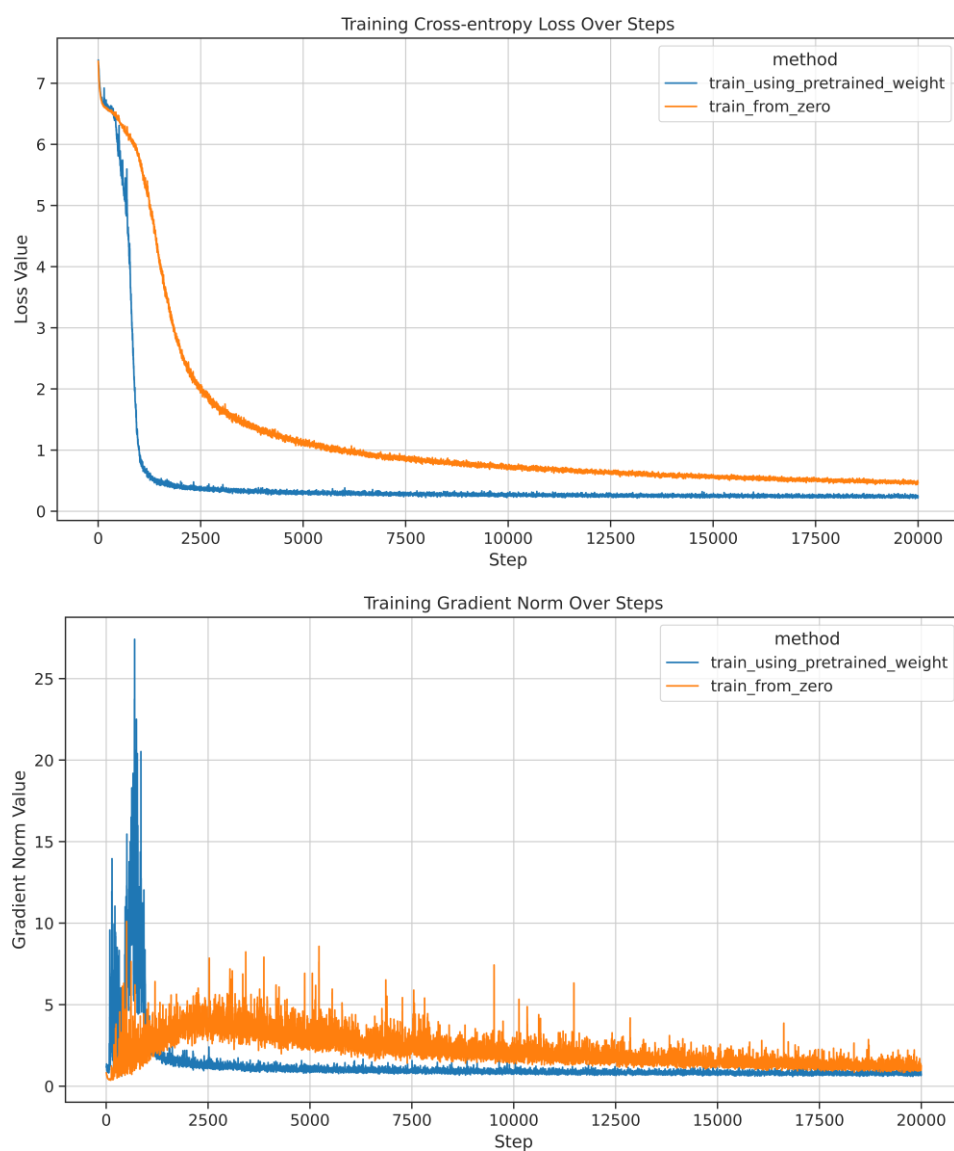


Fig. S24: Comparison of training cross-entropy loss (top) and gradient norm (below) for Coral-FAST with and without pre-trained weight initialization over training steps.