NEURAL INFORMATION PROCESSING SYSTEMS

Aalto-yliopisto

NanoBaseLib

A Multi-task Benchmark Dataset for Nanopore Sequencing

Guangzhao Cheng[1], Chengbo Fu[1], Lu Cheng[1,2]

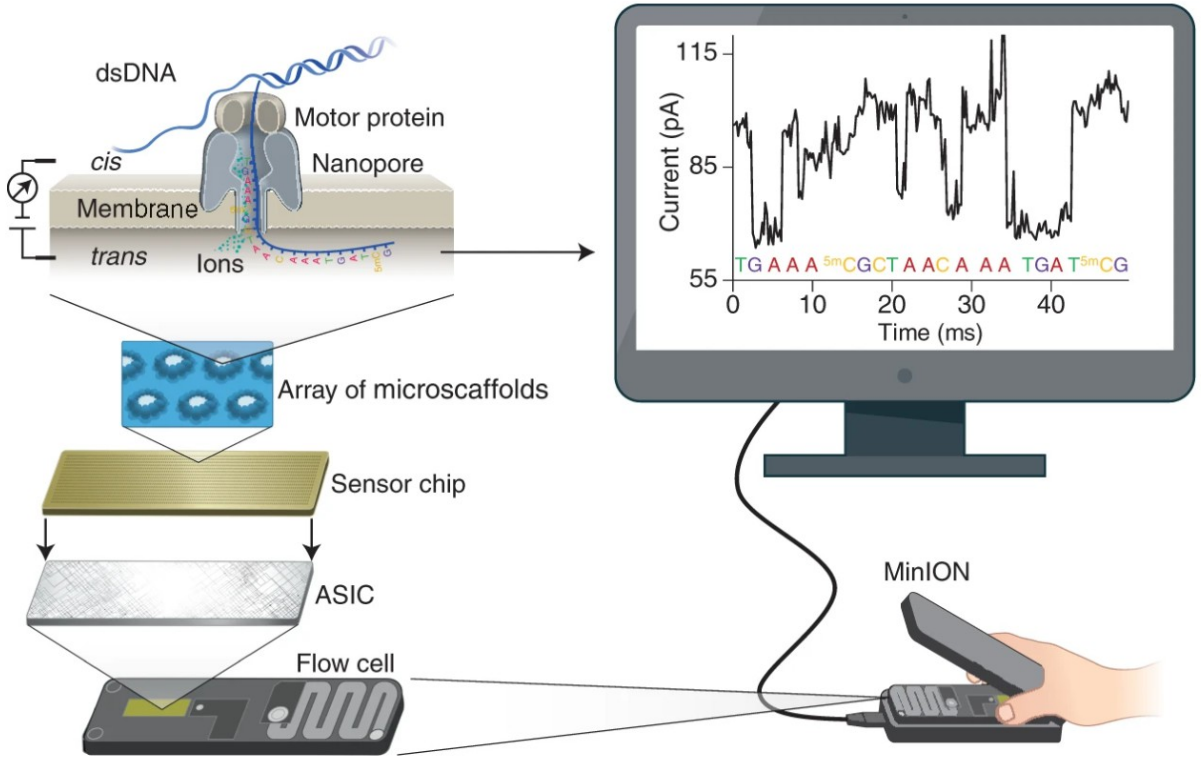
[1] Department of Computer Science, Aalto University, Finland
[2] Institute of Biomedicine, University of Eastern Finland, Finland

Contents

- Background
- Motivation
- NanoBaseLib: Dataset
- NanoBaseLib: Benchmark
  - Base Calling (BC)
  - PolyA Detection (PD)
  - Segmentation and event Alignment (SA)
  - Modification Detection (MD)
- Summary

Background: Nanopore Sequencing

- Nanopore sequencing is the third-generation sequencing technology.

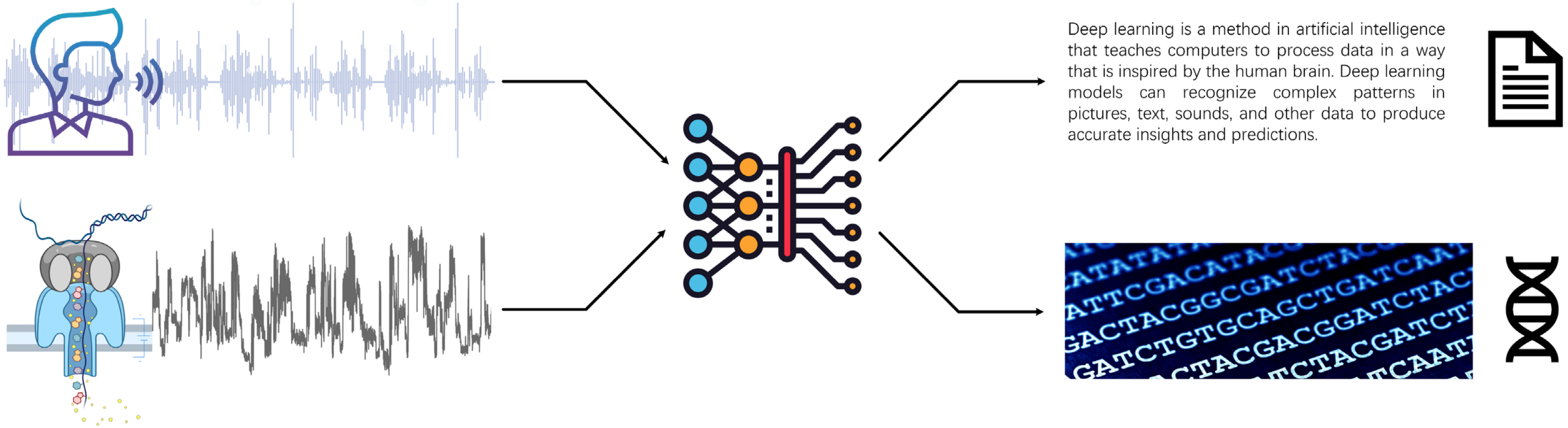


(Source o the figure: Wang, Yunhao, et al. 2021)

Aalto-yliopisto

Background: Nanopore Sequencing

- Base Calling (BC) is the **seq2seq** model similar with Automatic Speech Recognition (ASR).



Deep learning is a method in artificial intelligence that teaches computers to process data in a way that is inspired by the human brain. Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions.

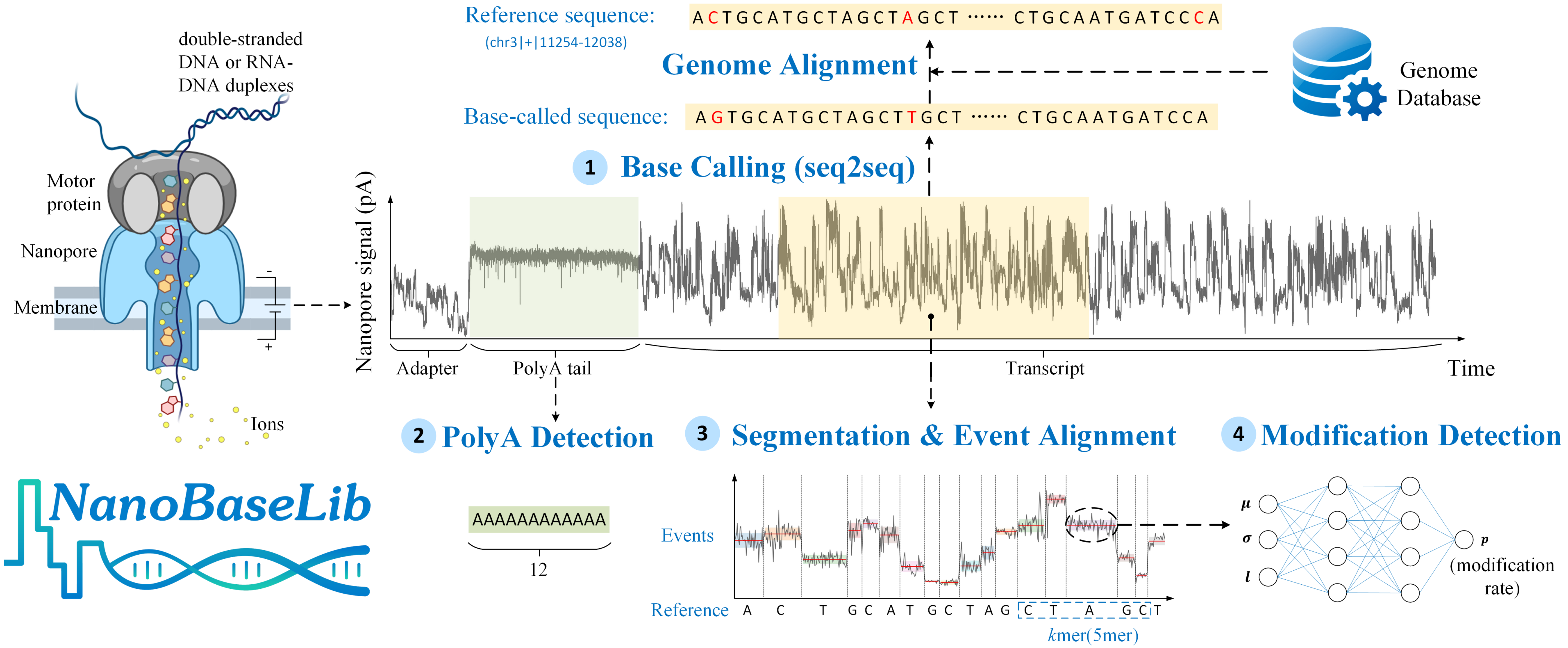
Automatic Speech Recognition (ASR)

Base Calling (BC)

Motivation

- Dataset
  - Performance improvement needs larger dataset
  - Dataset are scattered in different repositories
  - ONT has released various versions of the pores (R6 ~ R10.4)
- Analysis
  - Different bioinformatic preprocessing workflows
  - Undermines the benchmarking fairness
- Ground truth
  - Various tasks requires lots of domain knowledge

NanoBaseLib: One Dataset, Multiple Tasks

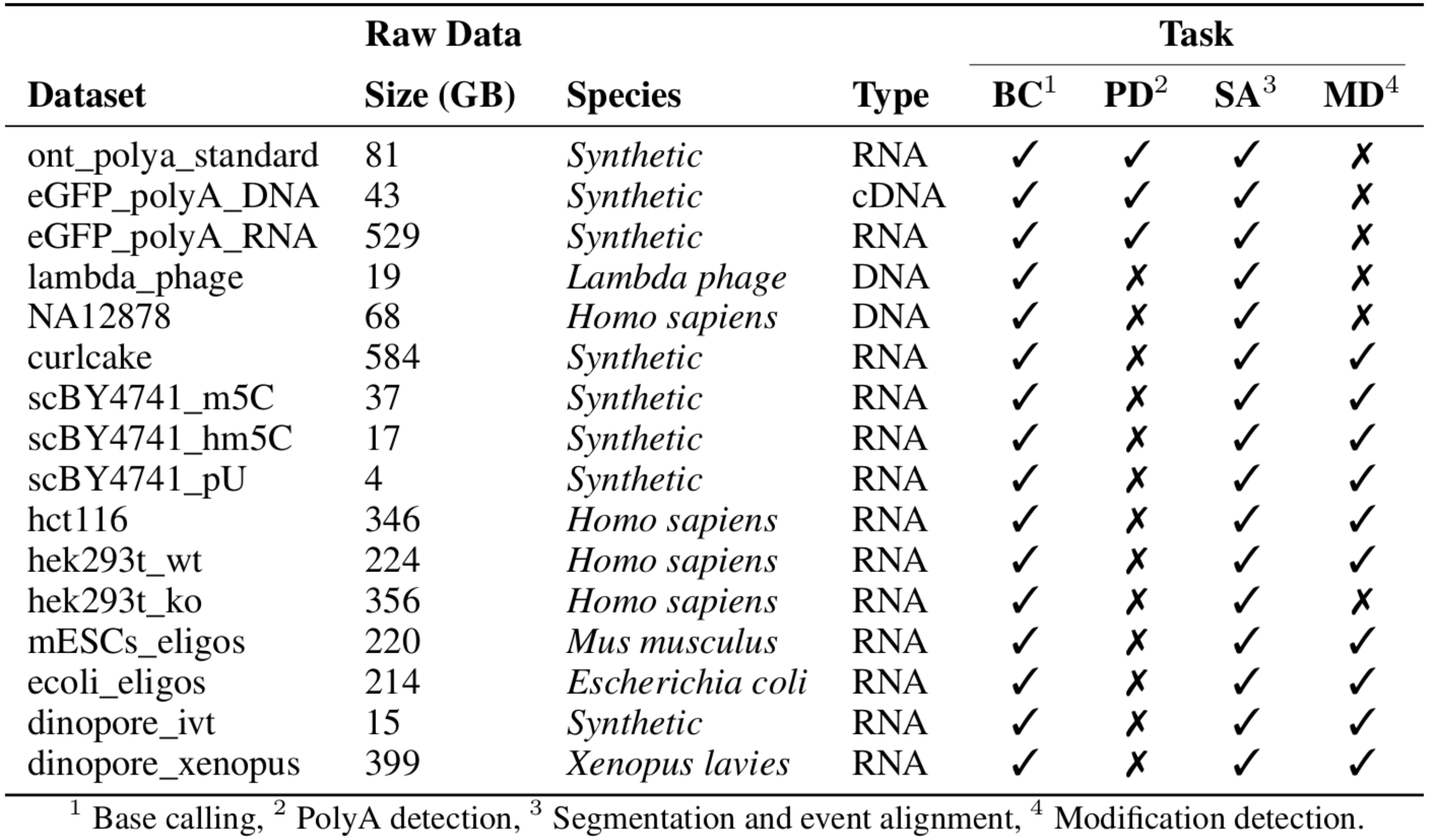


NanoBaseLib

Aalto-yliopisto

NanoBaseLib: Dataset

- https://nanobaselib.github.io/dataset.html



	Raw Data			Task			
Dataset	**Size (GB)**	**Species**	**Type**	**BC**[1]	**PD**[2]	**SA**[3]	**MD**[4]
ont_polya_standard	81	*Synthetic*	RNA	✓	✓	✓	✗
eGFP_polyA_DNA	43	*Synthetic*	cDNA	✓	✓	✓	✗
eGFP_polyA_RNA	529	*Synthetic*	RNA	✓	✓	✓	✗
lambda_phage	19	*Lambda phage*	DNA	✓	✗	✓	✗
NA12878	68	*Homo sapiens*	DNA	✓	✗	✓	✗
curlcake	584	*Synthetic*	RNA	✓	✗	✓	✓
scBY4741_m5C	37	*Synthetic*	RNA	✓	✗	✓	✓
scBY4741_hm5C	17	*Synthetic*	RNA	✓	✗	✓	✓
scBY4741_pU	4	*Synthetic*	RNA	✓	✗	✓	✓
hct116	346	*Homo sapiens*	RNA	✓	✗	✓	✓
hek293t_wt	224	*Homo sapiens*	RNA	✓	✗	✓	✓
hek293t_ko	356	*Homo sapiens*	RNA	✓	✗	✓	✗
mESCs_eligos	220	*Mus musculus*	RNA	✓	✗	✓	✓
ecoli_eligos	214	*Escherichia coli*	RNA	✓	✗	✓	✓
dinopore_ivt	15	*Synthetic*	RNA	✓	✗	✓	✓
dinopore_xenopus	399	*Xenopus lavies*	RNA	✓	✗	✓	✓

[1] Base calling, [2] PolyA detection, [3] Segmentation and event alignment, [4] Modification detection.

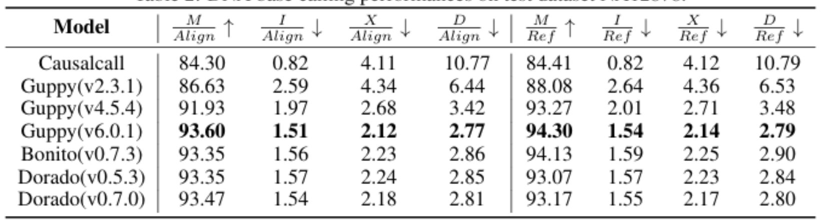
(Modifications: m6A, m5C, hm5C, inosine, pseudouridine)

Benchmark: Base Calling (BC)

- Input: raw current signal sequence
- Output: DNA/RNA base sequence

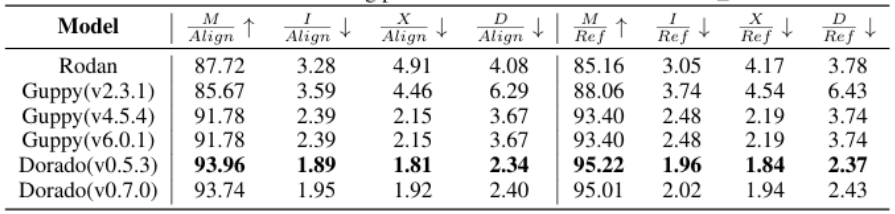
Software	Developer		Type		Architecture		
	ONT	Third-party	DNA	RNA	Convolution	Encoder	Decoder
Causalcall	✗	✓	✓	✗	Causal Dilated CNN →	→	CTC
Rodan	✗	✓	✗	✓	CNN →	→	CTC
Bonito	✓	✗	✓	✓	CNN →	LSTM →	CTC - CRF
Dorado	✓	✗	✓	✓	CNN →	Bi-LSTM →	CTC - CRF

Table 2: DNA base calling performances on test dataset NA12878.



Model	$\frac{M}{Align}\uparrow$	$\frac{I}{Align}\downarrow$	$\frac{X}{Align}\downarrow$	$\frac{D}{Align}\downarrow$	$\frac{M}{Ref}\uparrow$	$\frac{I}{Ref}\downarrow$	$\frac{X}{Ref}\downarrow$	$\frac{D}{Ref}\downarrow$
Causalcall	84.30	0.82	4.11	10.77	84.41	0.82	4.12	10.79
Guppy(v2.3.1)	86.63	2.59	4.34	6.44	88.08	2.64	4.36	6.53
Guppy(v4.5.4)	91.93	1.97	2.68	3.42	93.27	2.01	2.71	3.48
Guppy(v6.0.1)	**93.60**	**1.51**	**2.12**	**2.77**	**94.30**	**1.54**	**2.14**	**2.79**
Bonito(v0.7.3)	93.35	1.56	2.23	2.86	94.13	1.59	2.25	2.90
Dorado(v0.5.3)	93.35	1.57	2.24	2.85	93.07	1.57	2.23	2.84
Dorado(v0.7.0)	93.47	1.54	2.18	2.81	93.17	1.55	2.17	2.80

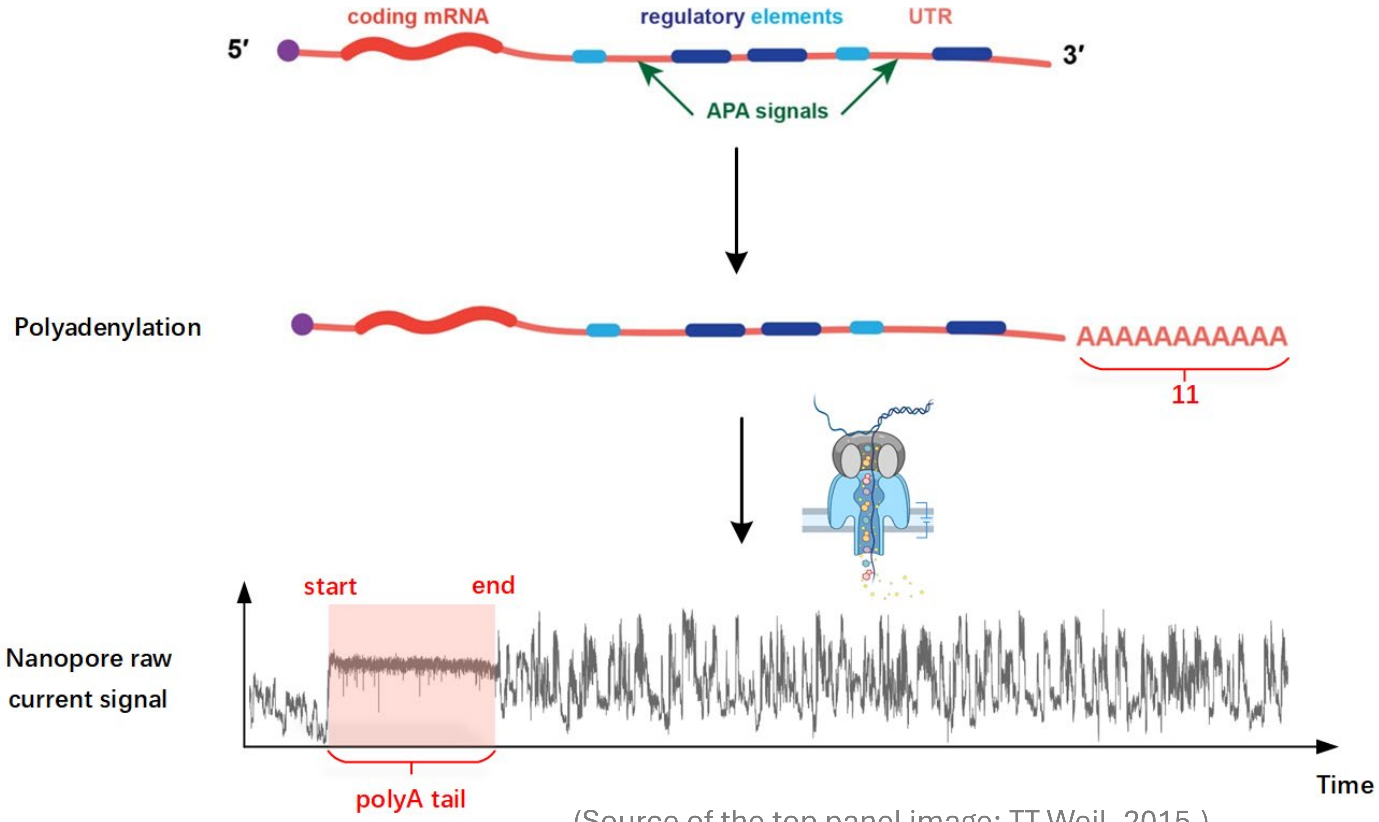
Table 3: RNA base calling performances on test dataset hek293t_wt.



Model	$\frac{M}{Align}\uparrow$	$\frac{I}{Align}\downarrow$	$\frac{X}{Align}\downarrow$	$\frac{D}{Align}\downarrow$	$\frac{M}{Ref}\uparrow$	$\frac{I}{Ref}\downarrow$	$\frac{X}{Ref}\downarrow$	$\frac{D}{Ref}\downarrow$
Rodan	87.72	3.28	4.91	4.08	85.16	3.05	4.17	3.78
Guppy(v2.3.1)	85.67	3.59	4.46	6.29	88.06	3.74	4.54	6.43
Guppy(v4.5.4)	91.78	2.39	2.15	3.67	93.40	2.48	2.19	3.74
Guppy(v6.0.1)	91.78	2.39	2.15	3.67	93.40	2.48	2.19	3.74
Dorado(v0.5.3)	**93.96**	**1.89**	**1.81**	**2.34**	**95.22**	**1.96**	**1.84**	**2.37**
Dorado(v0.7.0)	93.74	1.95	1.92	2.40	95.01	2.02	1.94	2.43

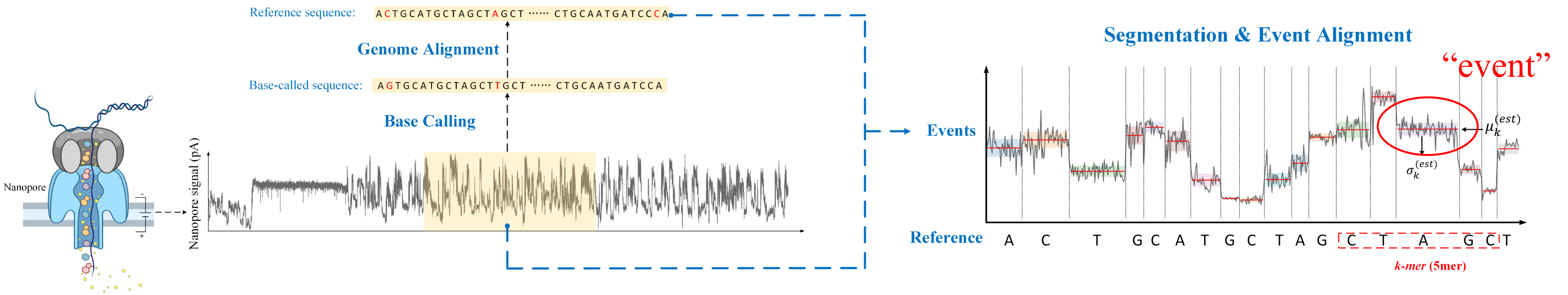
Benchmark: PolyA Detection (PD)

- Polyadenylation is the addition of a poly(A) tail to an RNA transcript, typically a messenger RNA (mRNA).



(Source of the top panel image: TT Weil, 2015.)

Benchmark: Segmentation and event Alignment (SA)

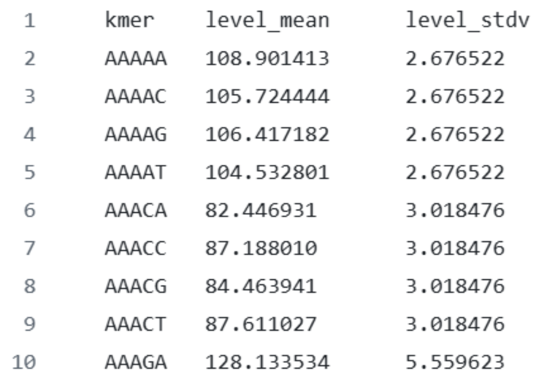


- Evaluation metrics:

$$\hat{\sigma} = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{1}{K_n}\sum_{k=1}^{K_n}\sigma_{s_{n,k}}^{(est)}\right\}$$

$$\hat{L} = \frac{1}{N}\sum_{n=1}^{N}\left\{\frac{1}{K_n}\sum_{k=1}^{K_n}\log\mathcal{N}\left(\mu_{s_{n,k}}^{(est)}|\mu_{s_{n,k}}^{(ref)},\sigma_{s_{n,k}}^{(ref)}\right)\right\}$$

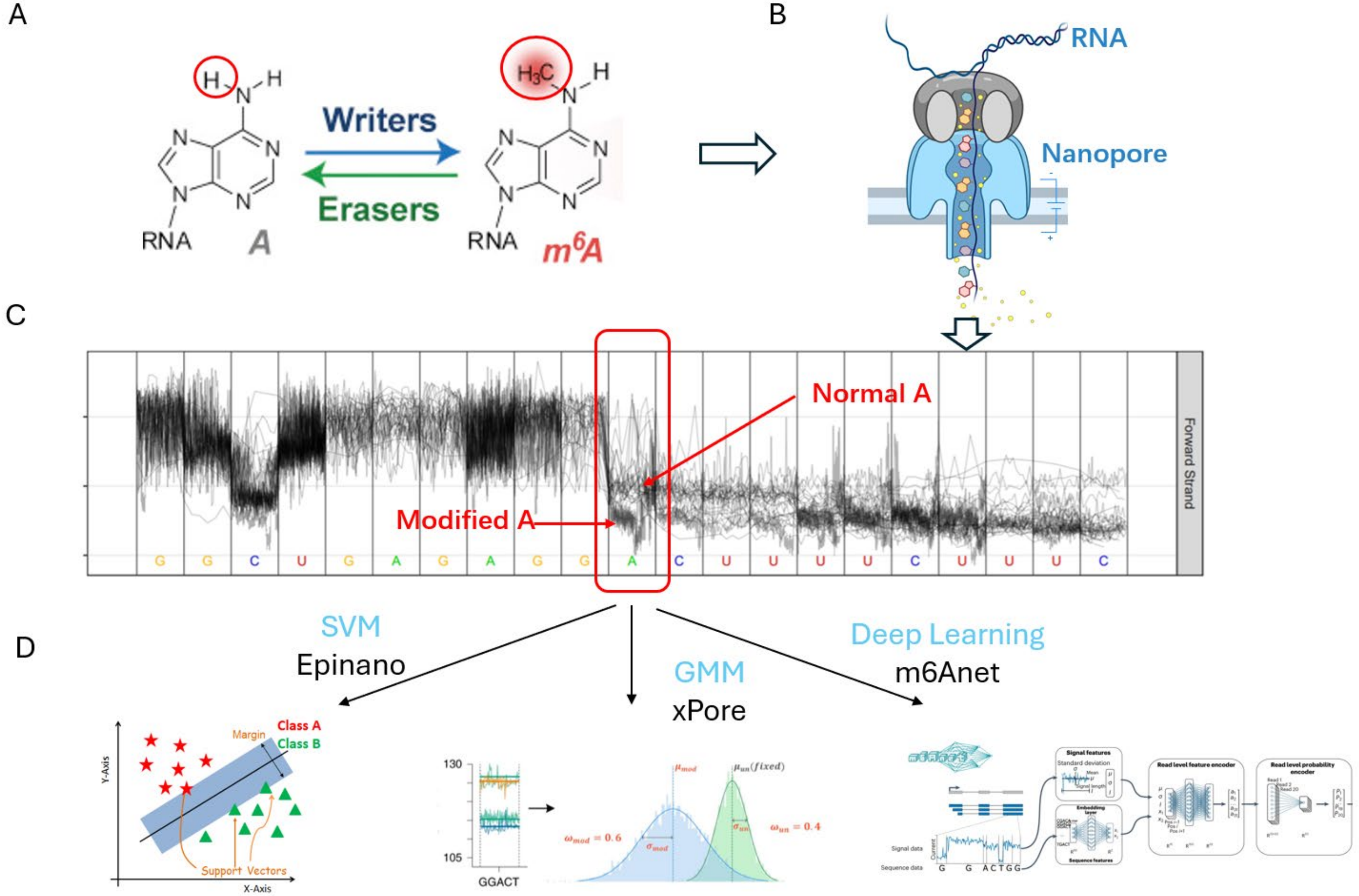
"standard" kmer parameter table from ONT[1]



kmer	level_mean	level_stdv
AAAAA	108.901413	2.676522
AAAAC	105.724444	2.676522
AAAAG	106.417182	2.676522
AAAAT	104.532801	2.676522
AAACA	82.446931	3.018476
AAACC	87.188010	3.018476
AAACG	84.463941	3.018476
AAACT	87.611027	3.018476
AAAGA	128.133534	5.559623

[1] https://github.com/nanoporetech/kmer_models

Benchmark: Modification Detection (MD)



(Source of the Figure A: Hailing Shi, et al. 2019. Source of the Figure D: Pratanwanich, Ploy N., et al. 2021, Hendra, Christopher, et al. 2022)

Summary

- One Dataset ➔ Multiple Tasks

NanoBaseLib Website !

✓ NanoBaseLib is a comprehensive dataset integrating 16 public datasets with over 30 million reads.

✓ NanoBaseLib is a benchmark platform covering 4 critical tasks.

✓ NanoBaseLib is a software package designed to incorporate new datasets efficiently.

Aalto-yliopisto