Direct RNA sequencing with modifications

- Jannes Spangenberg

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Jannes Spangenberg
jannes.spangenberg@uni-jena.de

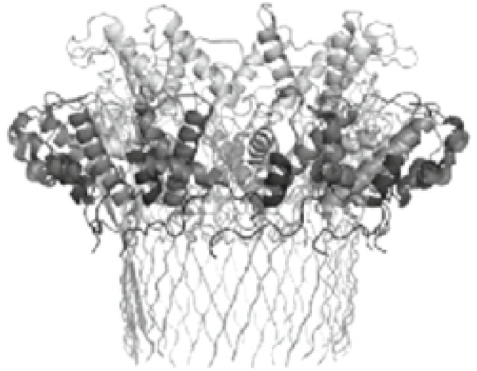
Presentation outline

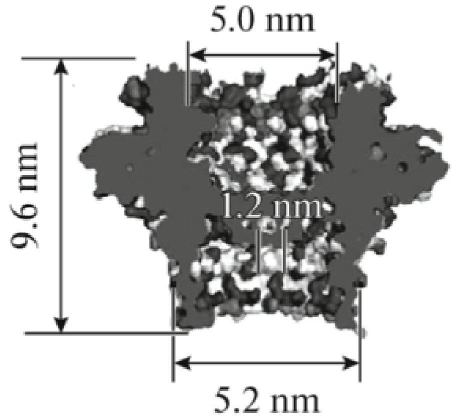
- Oxford Nanopore Technologies direct RNA sequencing
  - Challenges and problems
- RNA modification prediction using neural networks
- Current projects and ideas

What is a nanopore?

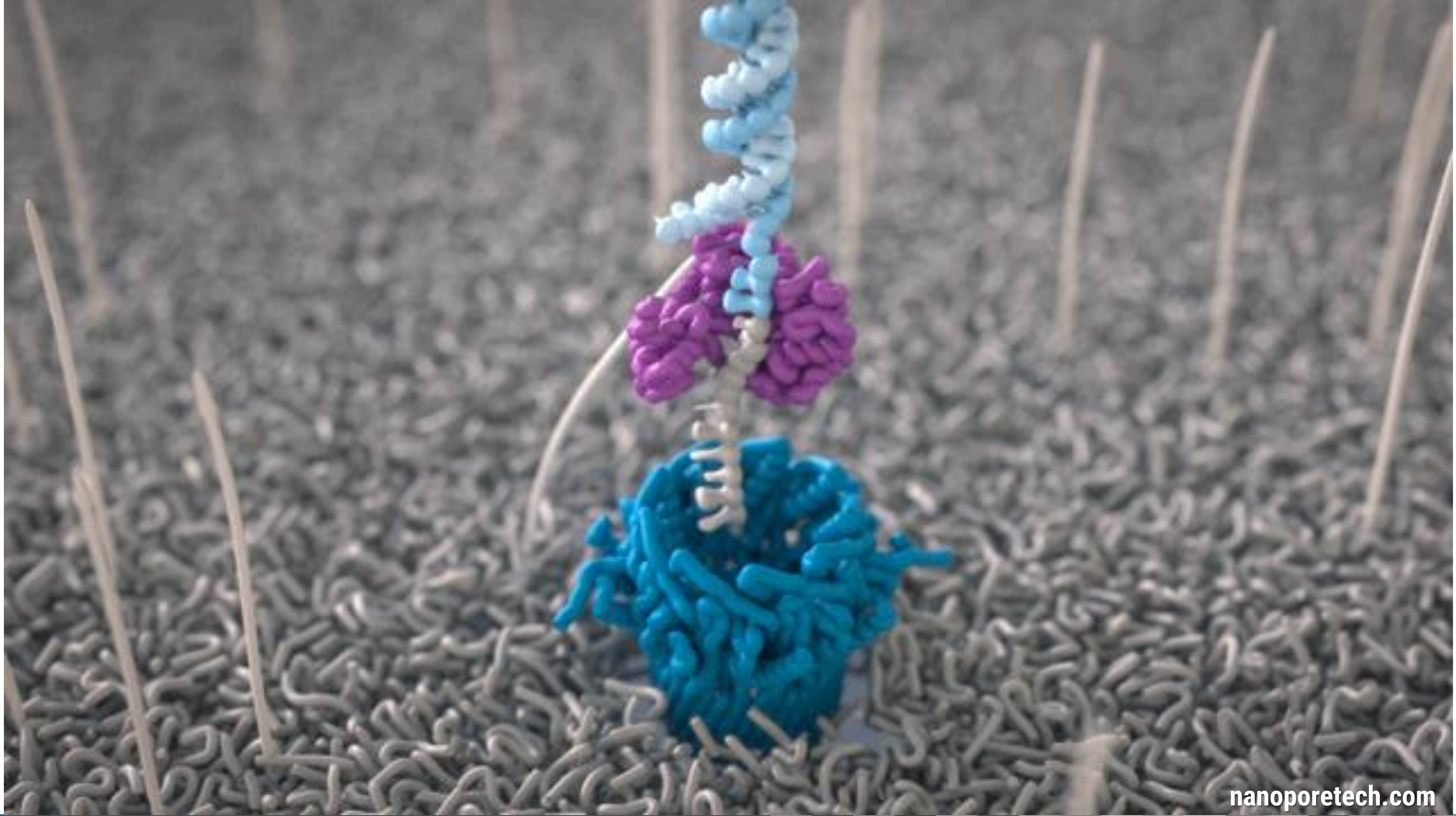
- Nanometer-sized pore made up of proteins
- Based on bacterial membrane pore complexes
- Improved by deliberate mutation to measure nucleotides
- Pore resides in a membrane

R9 pore

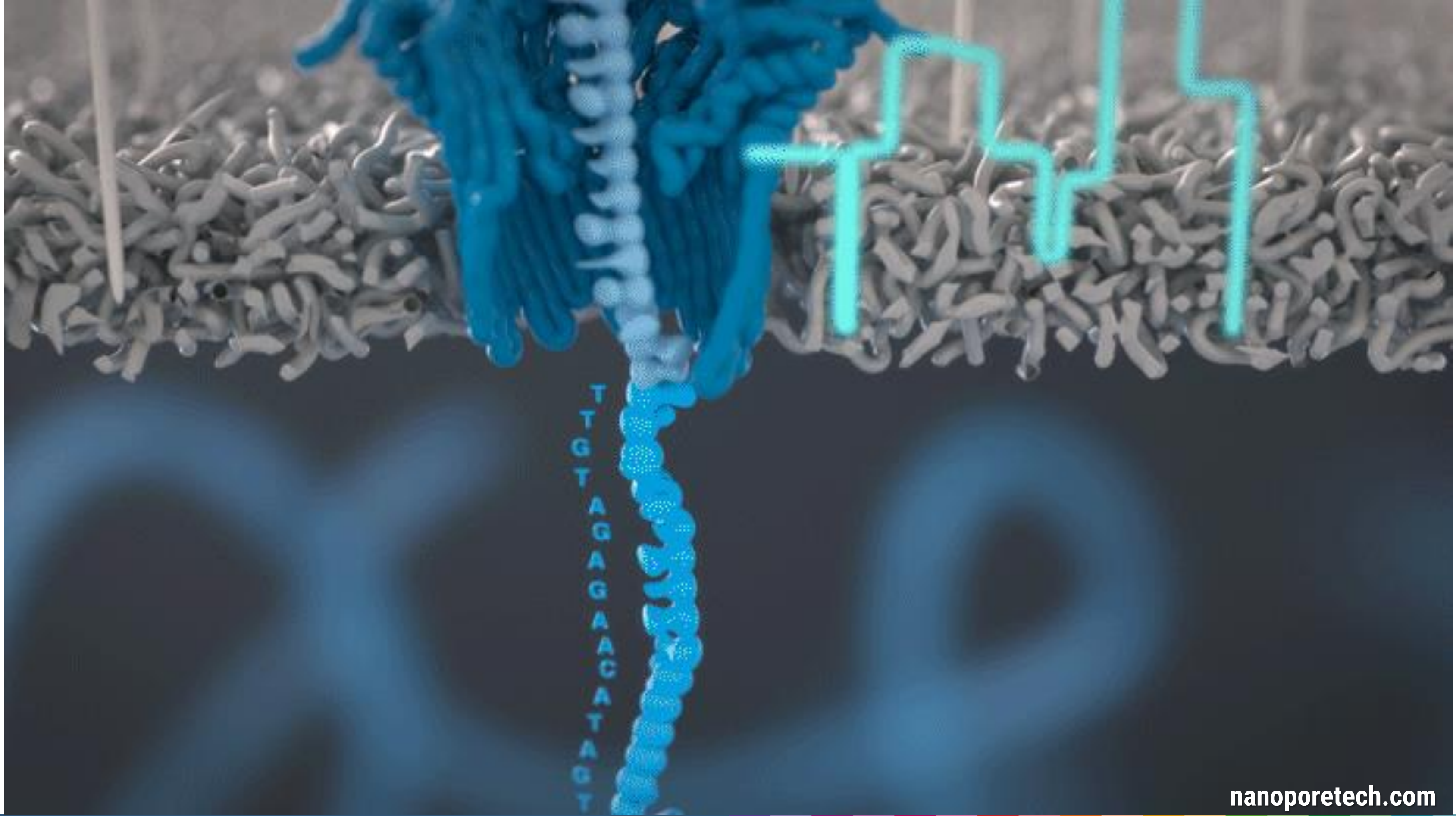




Barkova, D.V., Andrianova, M.S., Komarova, N.V. *et al.* Channel and Motor Proteins for Translocation of Nucleic Acids in Nanopore Sequencing. Moscow Univ. Chem. Bull. 75, 149–161 (2020).

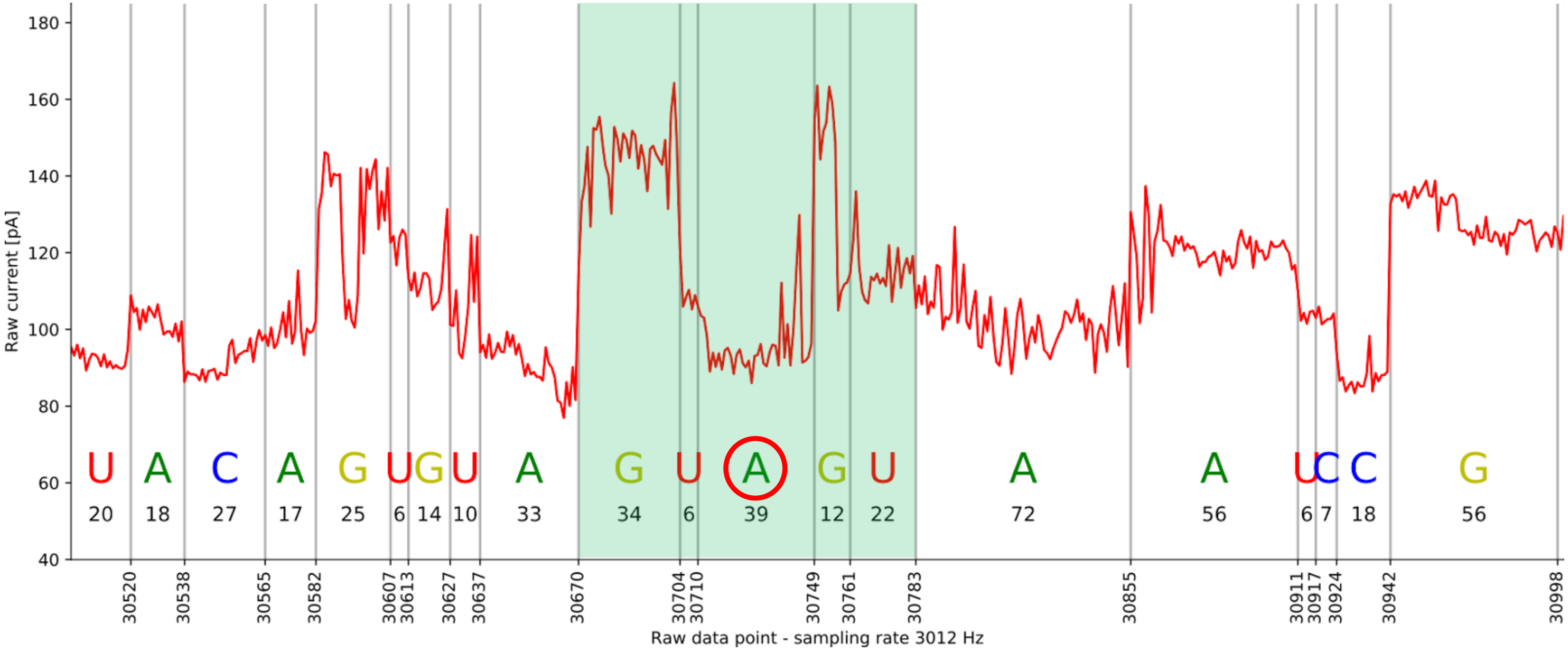


nanoporetech.com

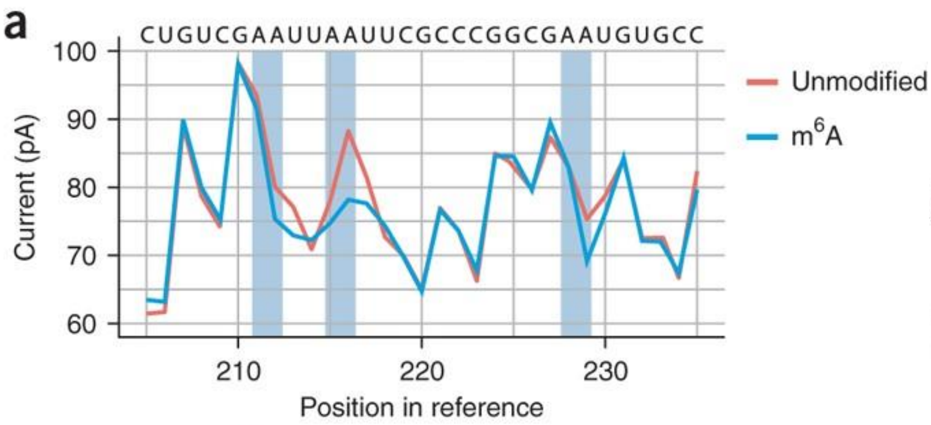


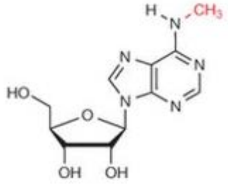
nanoporetech.com

Raw signal of a RNA

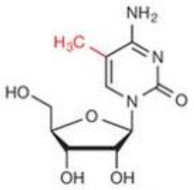


RNA modifications

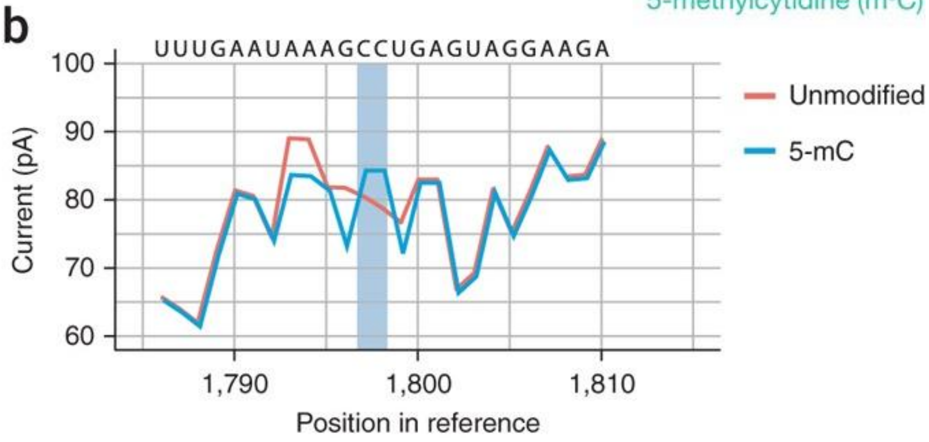




*N*6-methyladenosine (m^6A)

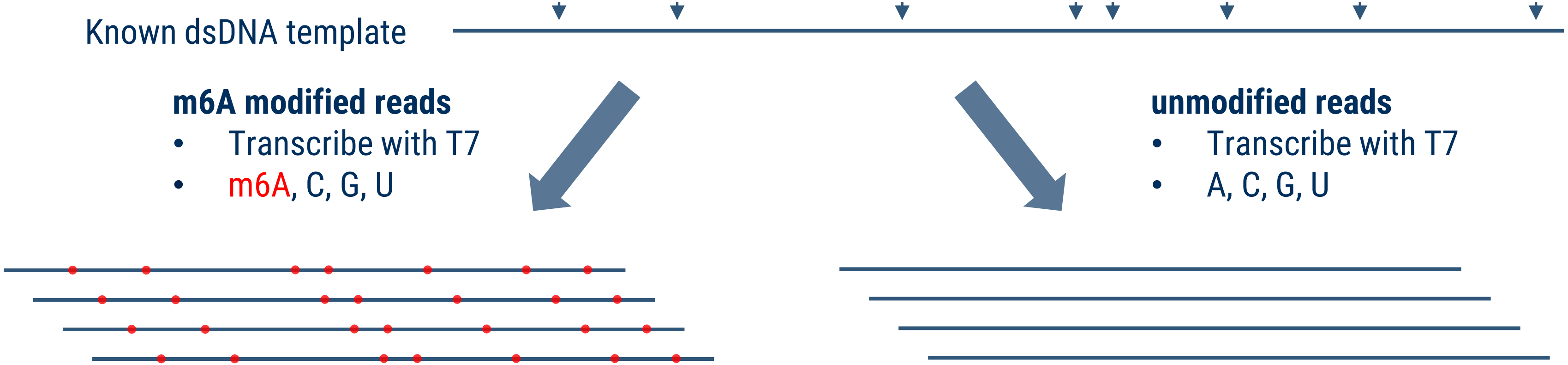


5-methylcytidine (m^5C)



Garalde, D., Snell, E., Jachimowicz, D. *et al.* Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods 15, 201–206 (2018)

Creating training data for m6A detection via *in vitro* transcription (IVT)



Known dsDNA template

m6A modified reads

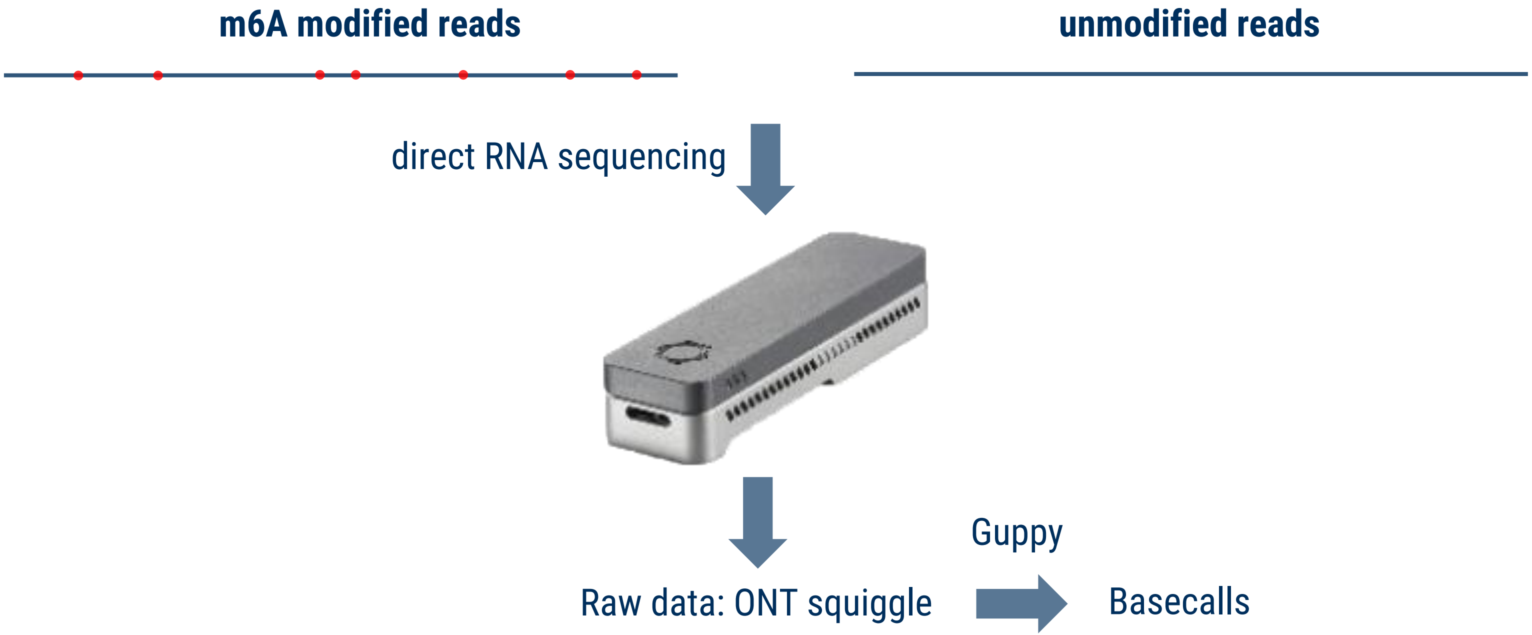
- Transcribe with T7
- m6A, C, G, U

unmodified reads

- Transcribe with T7
- A, C, G, U

Data design from: Liu, H., Begik, O., Lucas, M.C. *et al.* Accurate detection of m6A RNA modifications in native RNA sequences. Nat Commun 10, 4079 (2019)

Creating training data for m6A detection via *in vitro* transcription (IVT)



m6A modified reads

unmodified reads

direct RNA sequencing

Raw data: ONT squiggle

Guppy

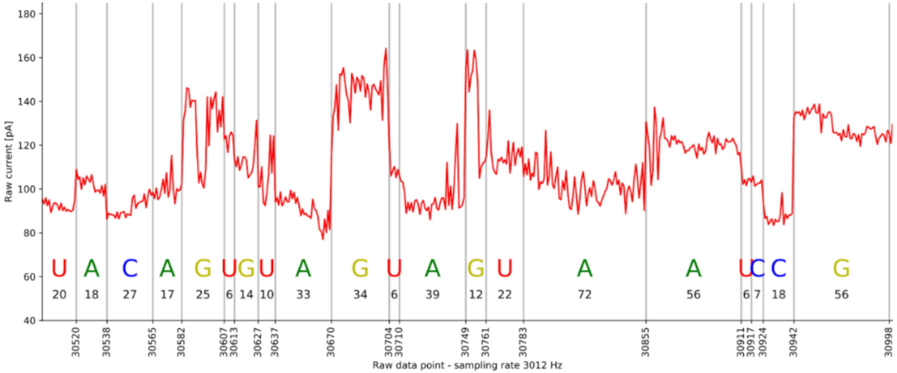
Basecalls

Creating training data for m6A detection via *in vitro* transcription (IVT)

- Raw data: ONT squiggle
- Basecalled reads
- Reference sequence
- Mapping

→

Resquiggling with **nanopolish eventalign**
(basecalling error correction and signal segmentation)

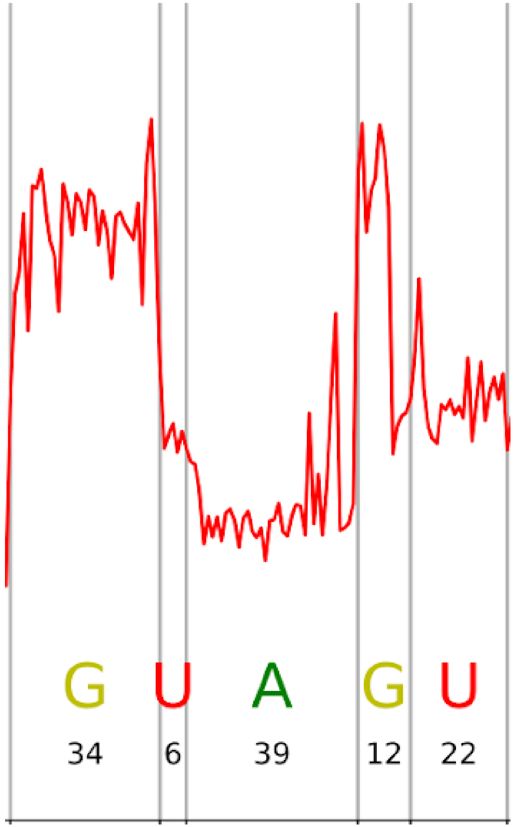


Loman, N., Quick, J. & Simpson, J. A complete bacterial genome assembled de novo using only nanopore sequencing data. Nat Methods 12, 733–735 (2015).

What is the input?

Sample: 5mer containing a m6A or canonical A in the middle

- Features:
  - Corrected basecalls
    - Use an embedding layer to encode the bases (A, C, G, U)
    - m6A and A will have the same encoding
  - Segmented signal
    - Extract the segments and interpolate them to a given size
  - Segment size
  - Trace from Guppy (base transition probabilities, bad resolution, 1 trace every 10th measured datapoint)



Results on datasets

- Train on IVT dataset from Liu *et al.*
- Test on our IVT dataset (different dataset, similar design)
- Test on *in vivo* dataset from Göke *et al.*
  - very bad m6A detection
  - Not transferable to *in vivo* data…?

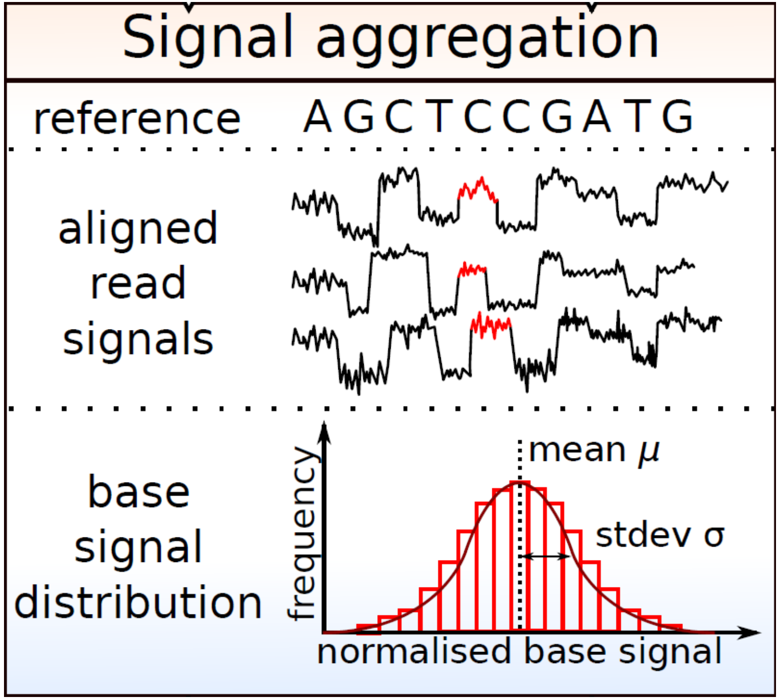
Dataset	Samples #	Acc.
Training on IVT of Liu *et al.* (80:20 split)	4‘888‘798 • Mod: 2‘444‘399 • Can: 2‘444‘399	0.95 on 20% split
Testing on IVT of Manja *et al.*	1’394’076 • Mod: 697’038 • Can: 697’038	0.73
Testing for TP on *in vivo* data from Göke *et al.*	Mod: 1‘252‘679	0.25

Challenges

- How can we design the *in vitro* transcription experiments as natural as possible and still know which positions are modified?
- Where/How can we get ground truth for *in vivo* data for modifications?
- Do you know or have ONT data with modifications and have a ground truth that we could use?
- Which input features should be used and how should they be provided/feeded to the model?

Magnipore (not published yet)

- Compares two samples sequenced with ONT
- Collect signals per reference positions from reads to calculate signal distributions
- Compare these signal distributions between the samples per position
- Look for significant signal differences
- Differences can originate from mutations or modifications



Isotopic labeling with D2O

- Detect deuterium labeled nucleotide sequences with ONT
- Isotopes are much smaller modifications
- Currently we see minor changes between H2O and D2O
- The signal-to-noise ratio is currently too small for accurate detection

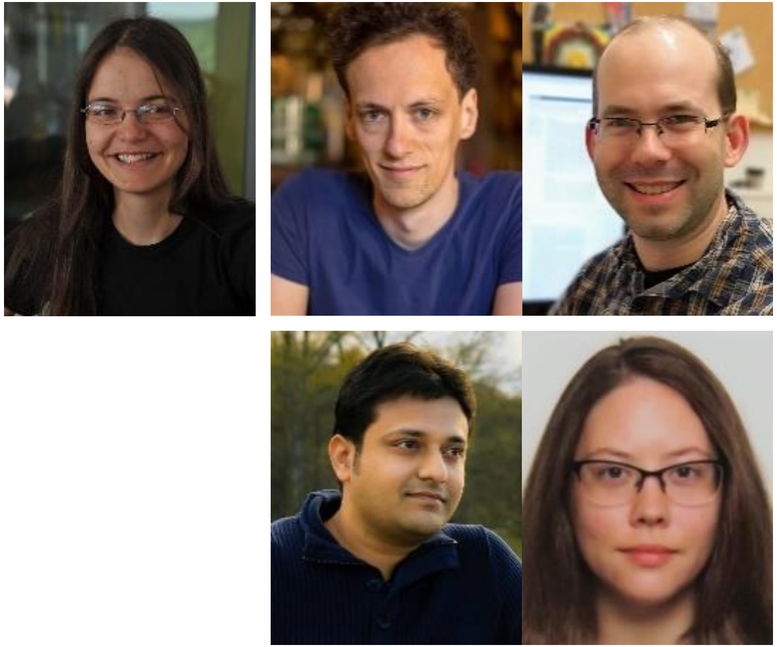
Thanks to:

- Manja Marz
- Christian Höner zu Siederdissen
- Sebastian Krautwurst

- Wetlab:
  - Akash Srivastava
  - Milena Žarković

and you!

Thanks for your attention!



Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

AquaDiva

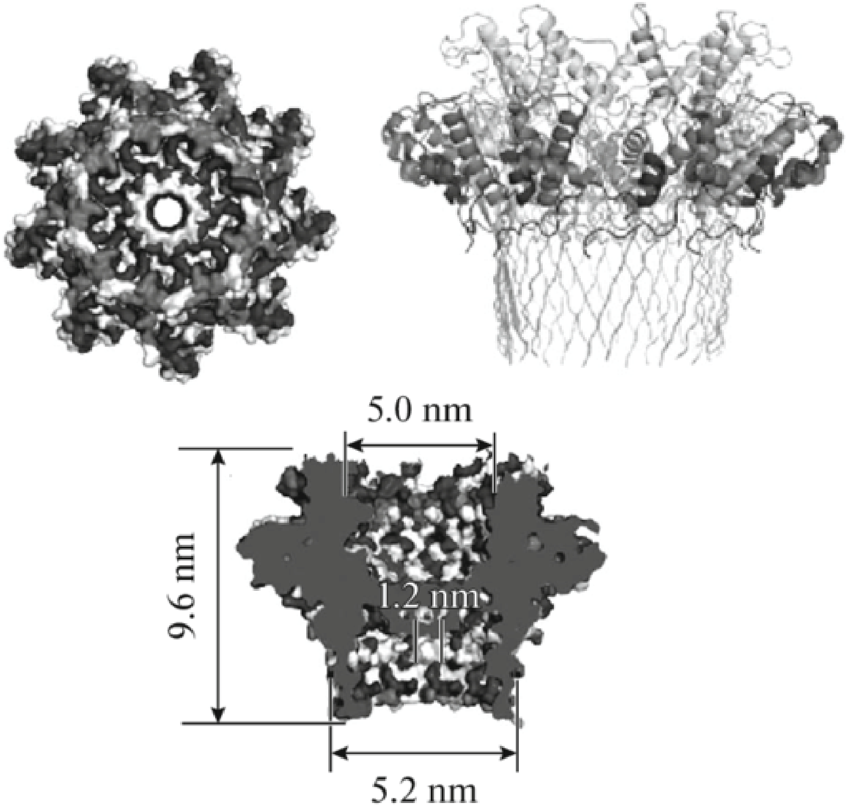
Thüringer Zentrum für Lernende Systeme und Robotik

Funding (in part):

- German DFG Collaborative Research Centre AquaDiva (CRC 1076 AquaDiva)
- The German state of Thuringia via the Thüringer Aufbaubank (2021 FGI 0009)

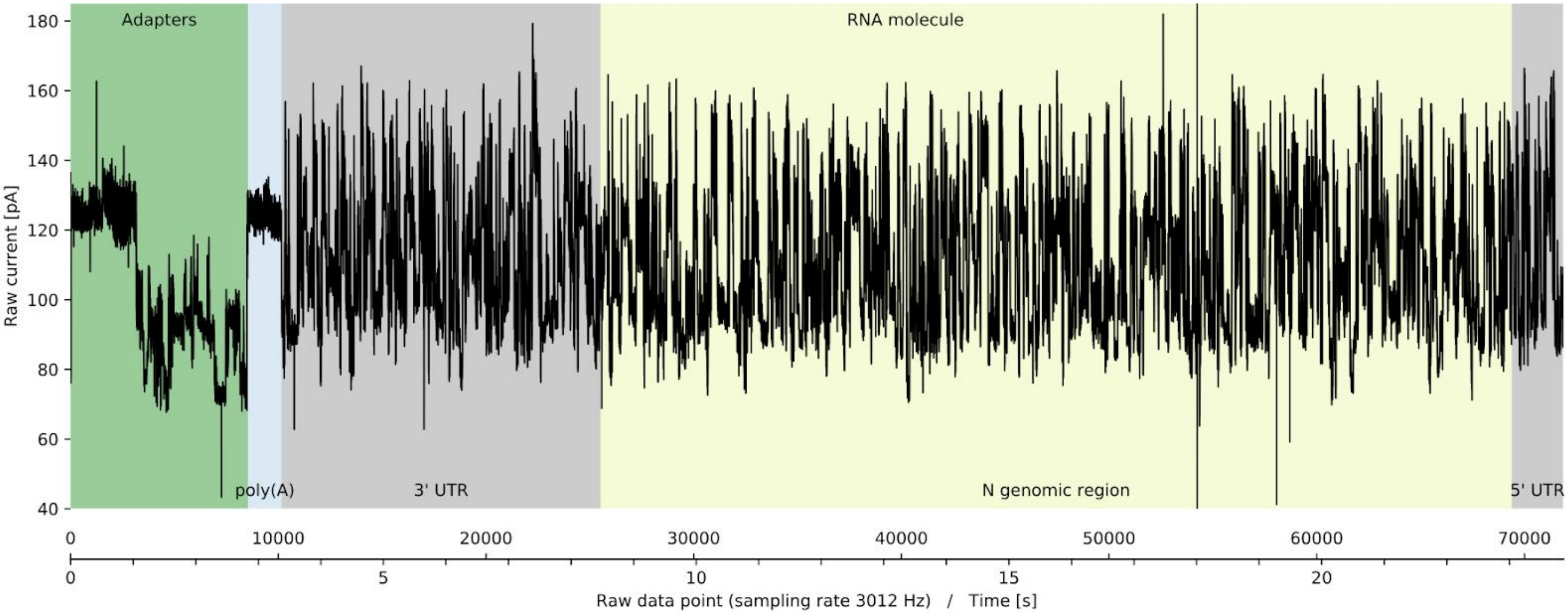
What is a nanopore?

- Voltage is applied to the membrane
- Measured current is characteristic to the molecules at the most narrow part of the pore
- 5 nucleotides are measured at once
- Measurements are influenced by the molecules in the pore, the pore, the sensor, the flowcell and the sequencing kit/protocol



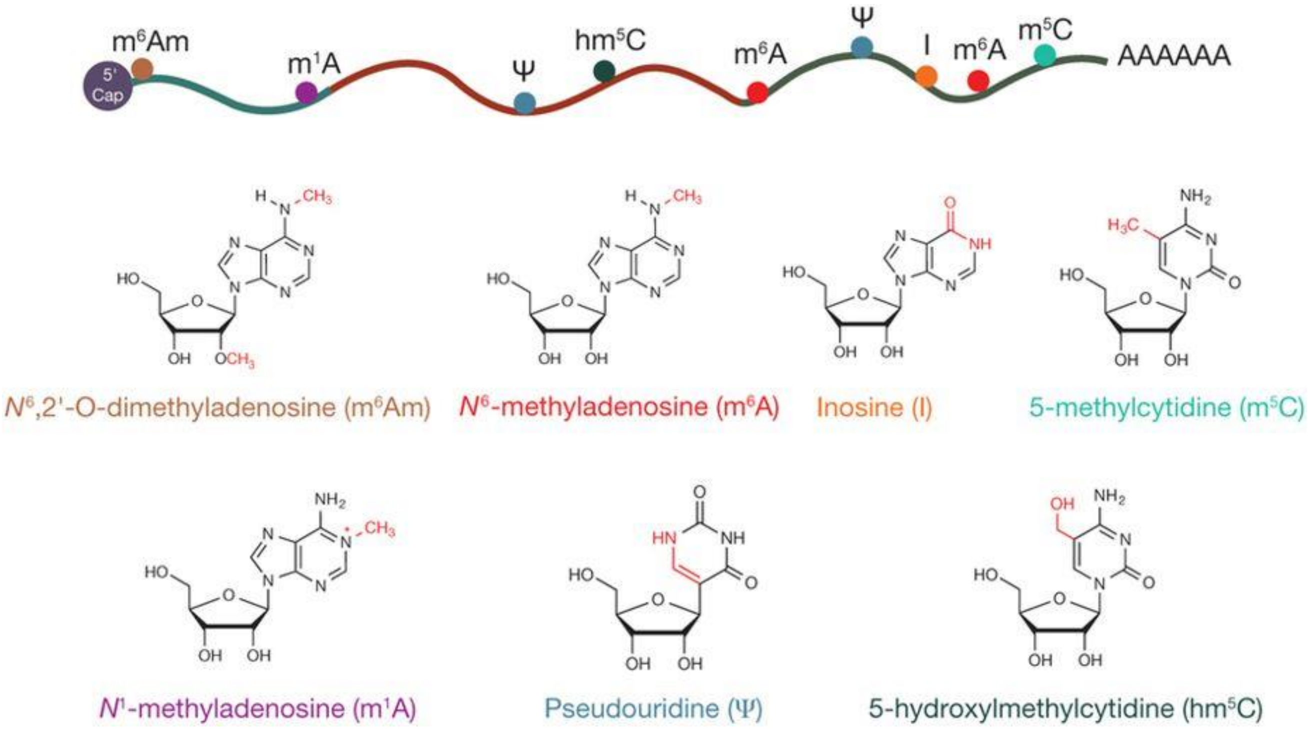
Barkova, D.V., Andrianova, M.S., Komarova, N.V. *et al.* Channel and Motor Proteins for Translocation of Nucleic Acids in Nanopore Sequencing. Moscow Univ. Chem. Bull. 75, 149–161 (2020).

Raw signal of a RNA



Viehweger *et al*., Direct RNA nanopore sequencing of full-length coronavirus genomes provides novel insights into structural variants and enables modification analysis, Genome Research 2019

RNA modifications

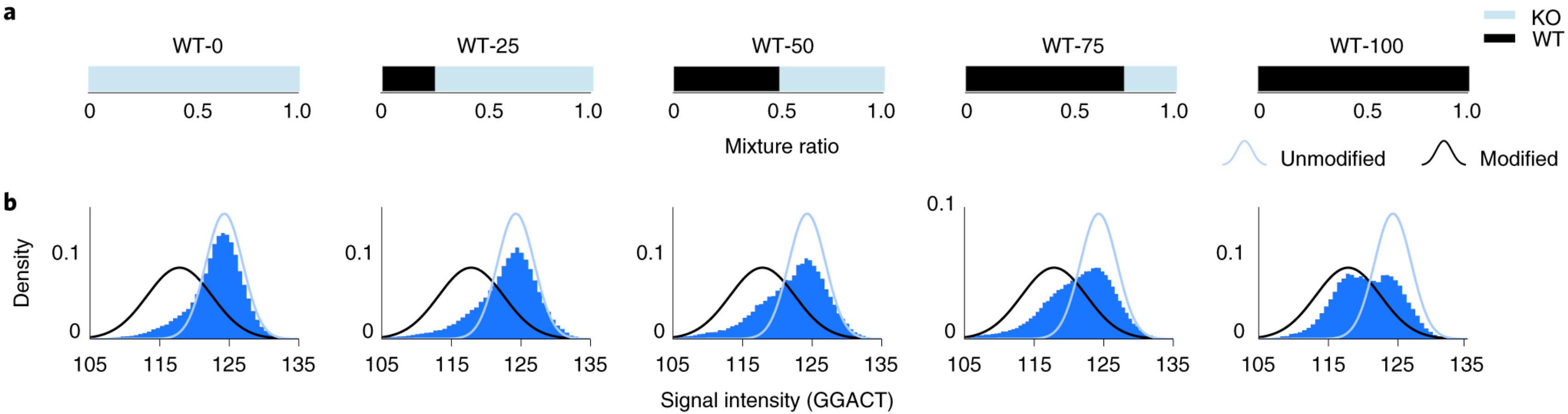

5' Cap
m^{6}Am
m^{1}A
Ψ
hm^{5}C
m^{6}A
Ψ
I
m^{6}A
m^{5}C
AAAAAA
N^6,2'-O-dimethyladenosine (m^{6}Am)
N^6-methyladenosine (m^{6}A)
Inosine (I)
5-methylcytidine (m^{5}C)
N^1-methyladenosine (m^{1}A)
Pseudouridine (Ψ)
5-hydroxylmethylcytidine (hm^{5}C)

Signal normalisation

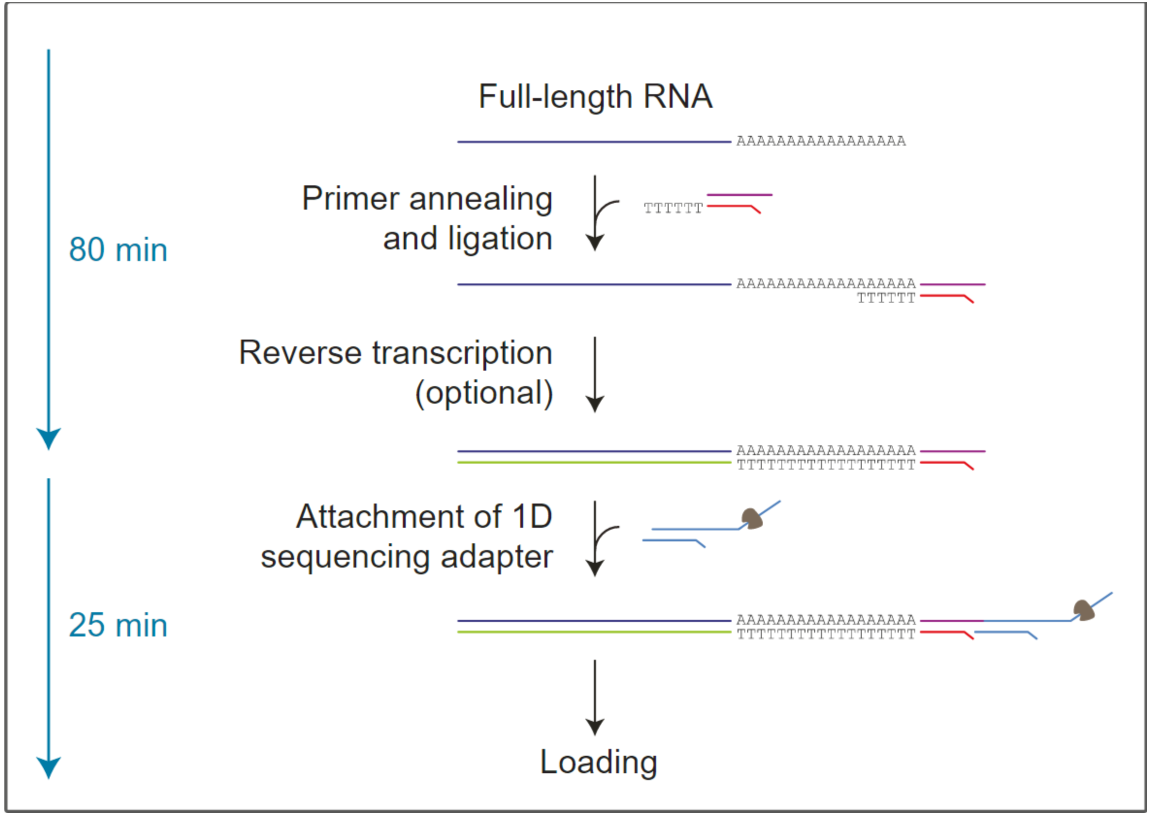
- Make the sequencing comparable across different
  - Pores/sensors
  - Flowcells
  - Sequencing protocols/experiments

$$normalised\ signal = \frac{signal\ - median(signal)}{median\_absolute\_deviation(signal)}$$

RNA modifications



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



Full-length RNA

Primer annealing and ligation

Reverse transcription (optional)

Attachment of 1D sequencing adapter

Loading

80 min

25 min

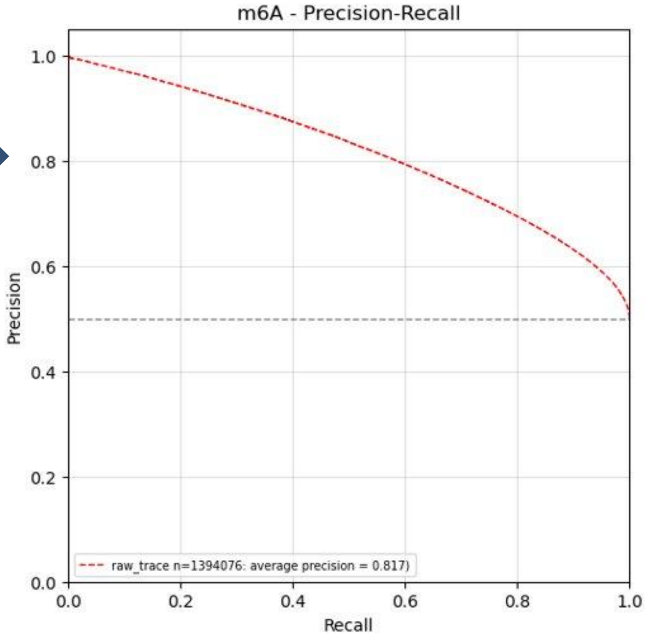
nanoporetech.com

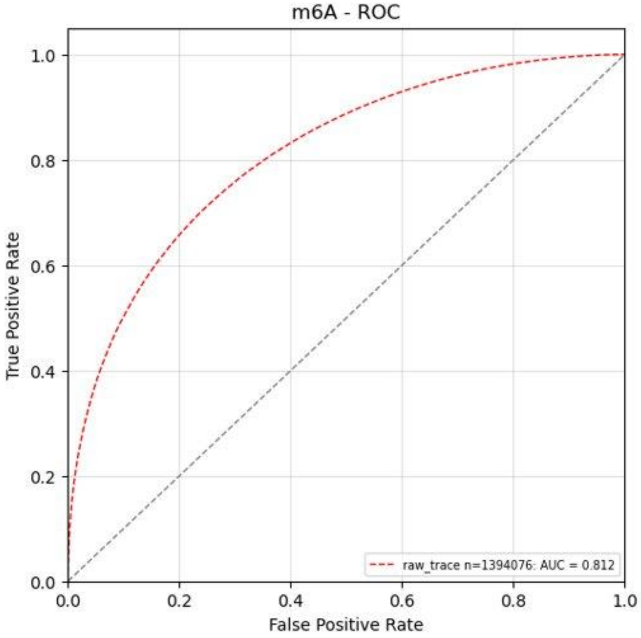
Results

- $Precision = \frac{TP}{TP+FP}$
- $Recall = \frac{TP}{TP+FN}$

- $True\ positive\ rate = recall$
- $False\ positive\ rate = \frac{FP}{FP+TN}$

- Train on IVT dataset from Liu et. al.
- Test on our IVT dataset (different dataset, similar design) →
- Test on in vivo dataset from Goeke et. al.
  - very bad m6A detection
  - Not transferable to in vivo data…?





Datasets

IVT datasets	Liu *et al.*	Manja *et al.*
# Reads	88,819	39,219
# Canonical	47,325	9,542
# Modified	11,921	26,337

In vivo datasets by Göke *et al.*	Wild Type			Knockout		
	1	2	3	1	2	3
# Reads	2,389,434	3,302,095	1,124,426	3,476,668	4,265,961	3,993,818

Neural Network

- Embedding layer for the base encoding
- Transformer layer for the signal processing
- Linear fully connected layers
- One output value from a sigmoid function predicting the modification status for the A in the input sample
  - Output $< 0.5 =$ unmodified
  - Output $\geq 0.5 =$ modified