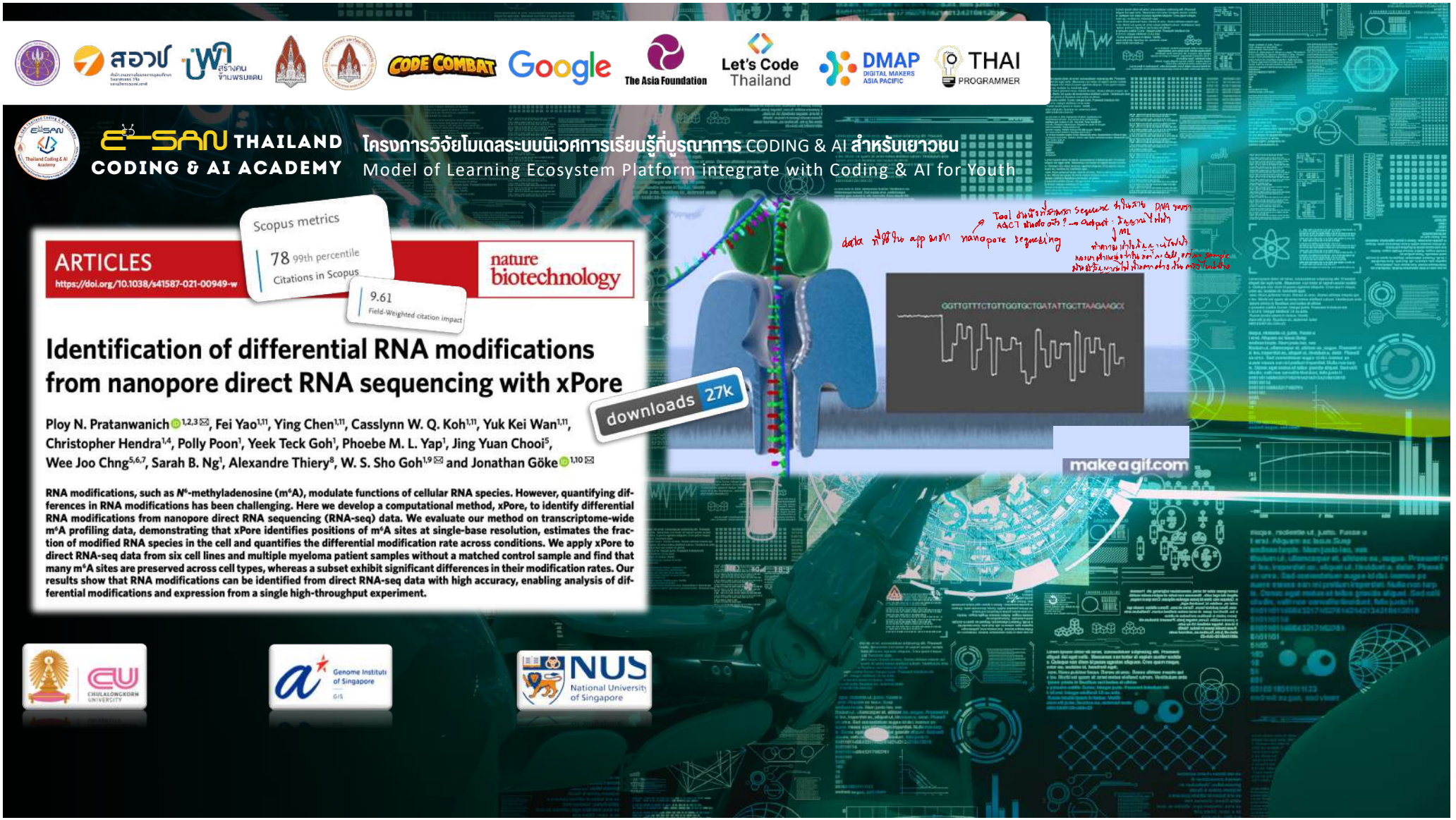
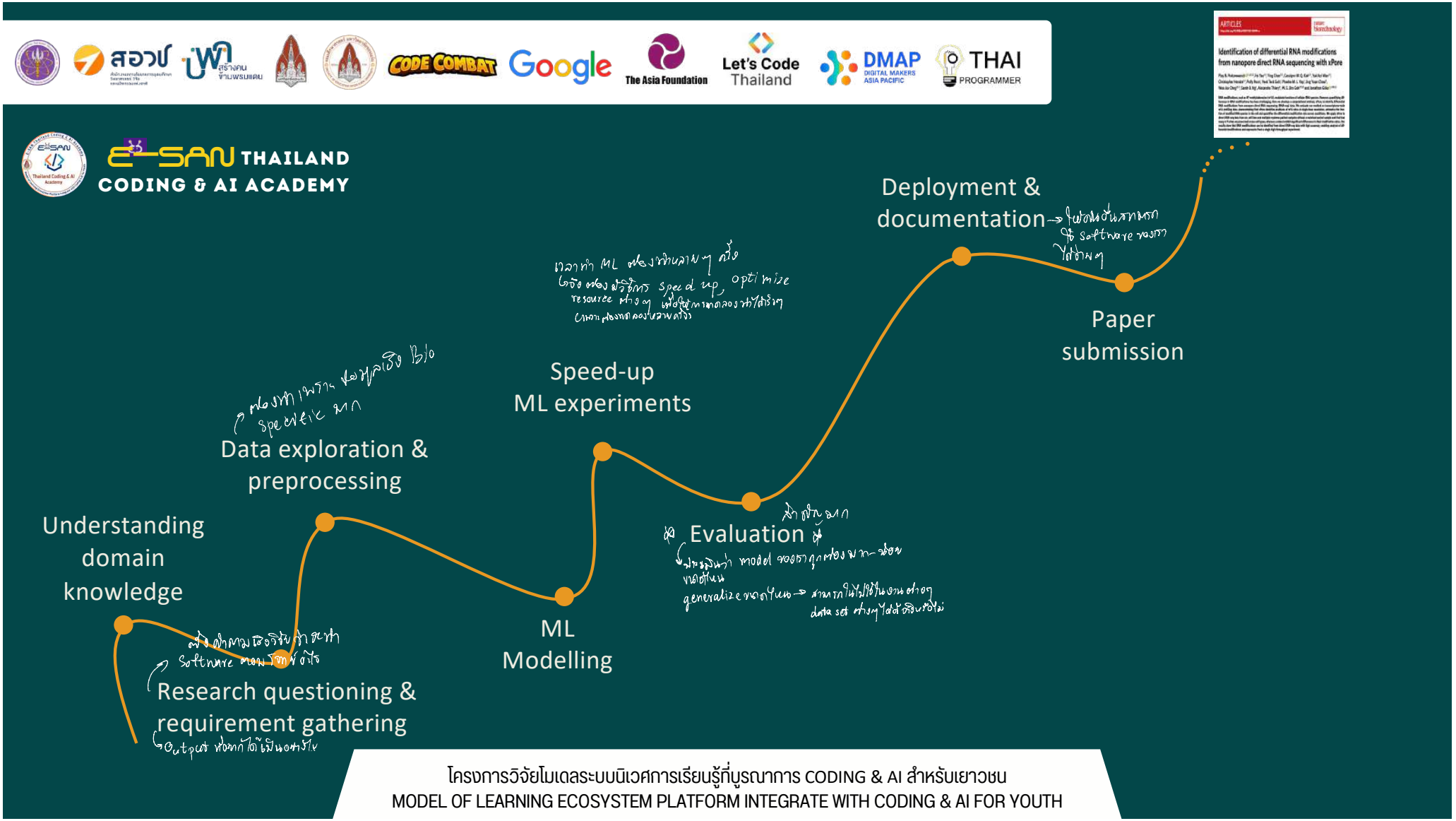

E-SAN THAILAND
CODING & AI ACADEMY
โครงการวิจัยโมเดลระบบนิเวศการเรียนรู้ที่บูรณาการ CODING & AI สำหรับเยาวชน
Model of Learning Ecosystem Platform integrate with Coding & AI for Youth
โครงการย่อยที่ 6
การพัฒนาเยาวชนเพื่อเข้าสู่วิชาชีพขั้นสูงด้าน Coding & AI
ร่วมกับ Coding Entrepreneur & Partnership: Personal AI
developed since 2019-2021
xPore
Application of AI for Bioinformatic
AI-Powered App
for Bioinformaticians
ผศ. ดร.นฤมล ประทานวณิช
โครงการย่อยที่ 6

งานวิจัย RNA modification

Published : Nature biotechnology

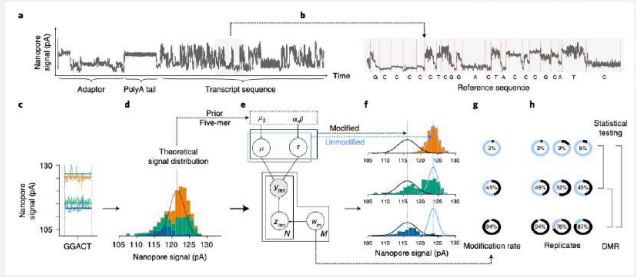
cited: 73 times

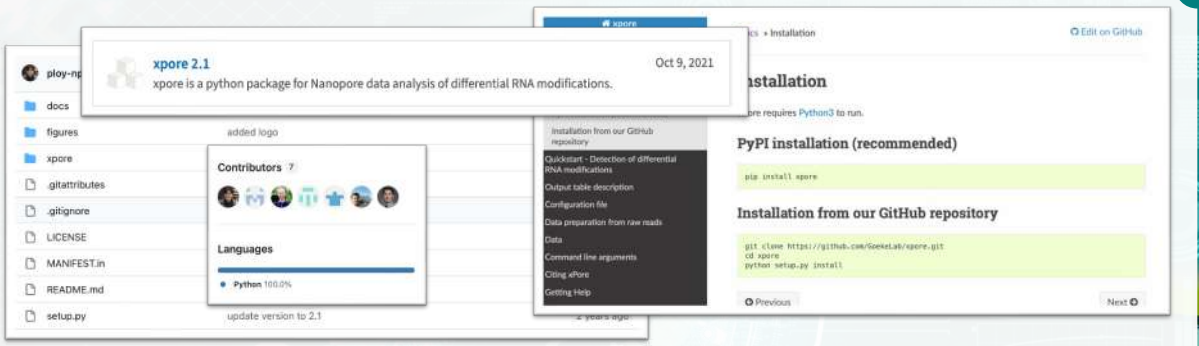


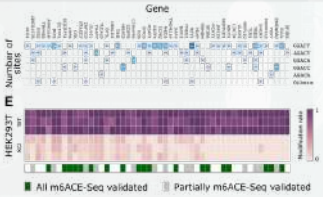


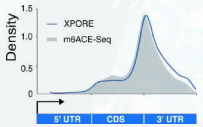
โครงการวิจัยโมเดลระบบนิเวศการเรียนรู้ที่บูรณาการ CODING & AI สำหรับเยาวชน
MODEL OF LEARNING ECOSYSTEM PLATFORM INTEGRATE WITH CODING & AI FOR YOUTH

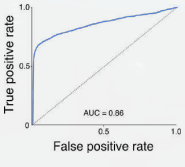
Course Overview

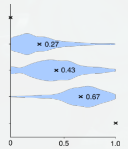


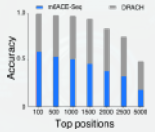


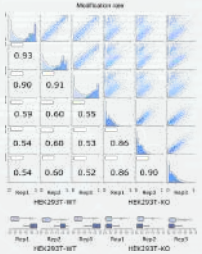


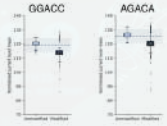












XPORE

E-SAN THAILAND CODING & AI ACADEMY

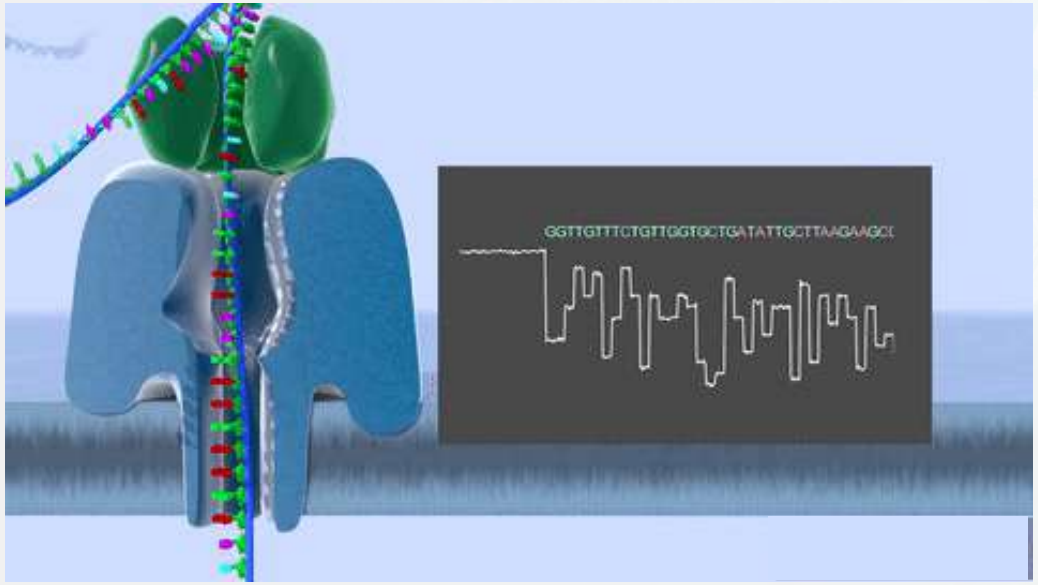
โครงการวิจัยโมเดลระบบนิเวศการเรียนรู้ที่บูรณาการ CODING & AI สำหรับเยาวชน
Model of Learning Ecosystem Platform integrate with Coding & AI for Youth

Outline

การพัฒนาเยาวชนเพื่อเข้าสู่วิชาชีพขั้นสูงด้าน Coding & AI **ร่วมกับ** Coding **Entrepreneur & Partnership: Personal AI**

1. Problem Statement ระบุปัญหา
2. Data Collection and Preparation จัดเก็บ data → หาจากไหน, เตรียมข้อมูล อย่างไร
3. Bayesian [Multi-Sample] Gaussian Mixture Modelling
   Try Coding
   Core ML technique in X-pere
4. Evaluation → Model ประสิทธิภาพ อย่างไร
5. Visualization and Presentation
6. Future Work

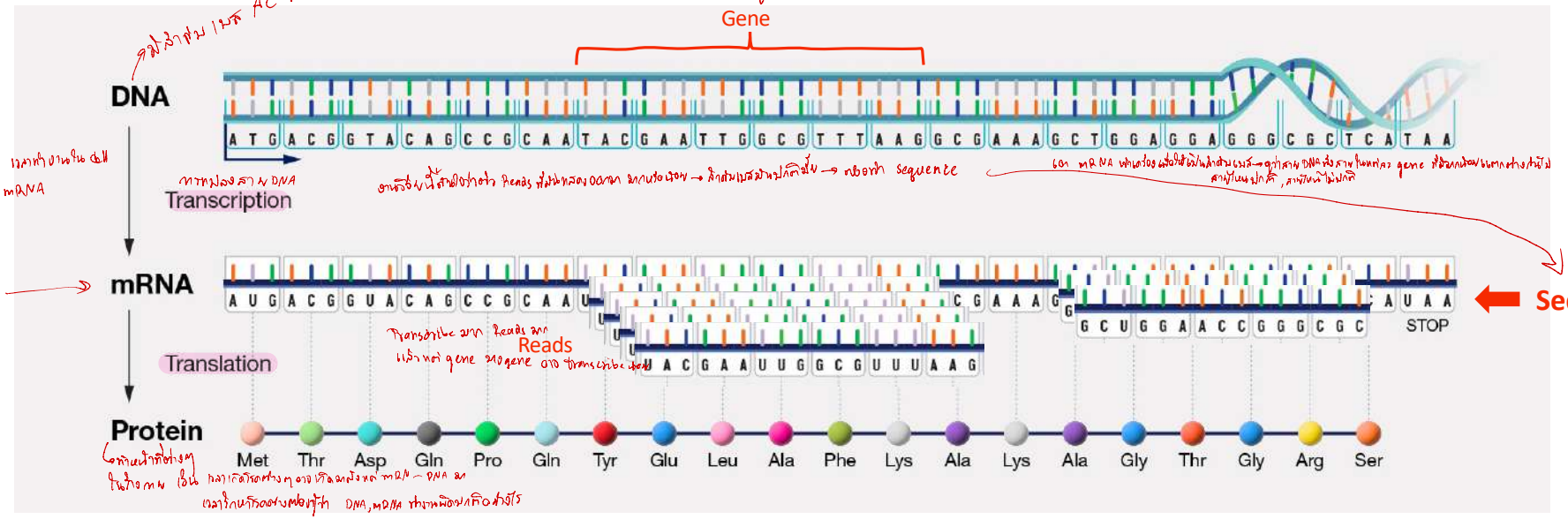
1. Problem Statement



- **Nanopore Sequencing**
- **RNA Modification**
- **Inputs & Outputs**
- **Research Objectives**

Central Dogma

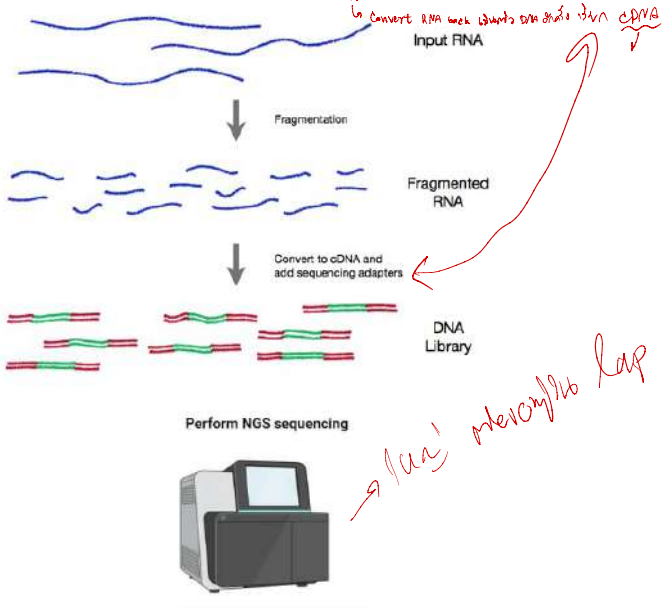
Domain knowledge



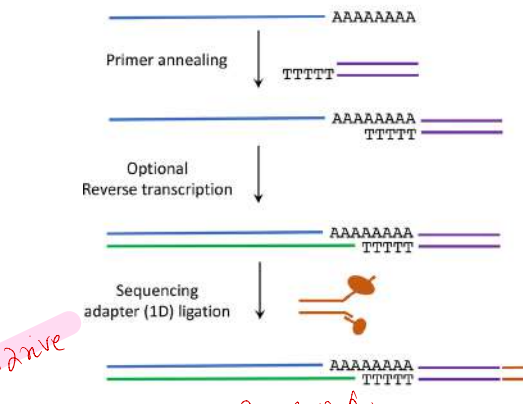
Source: https://www.genome.gov/genetics-glossary/Central-Dogma

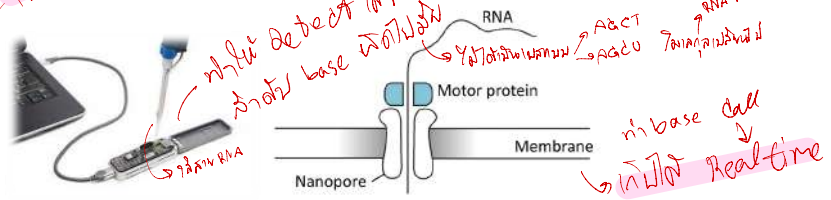
RNA Sequencing

Then

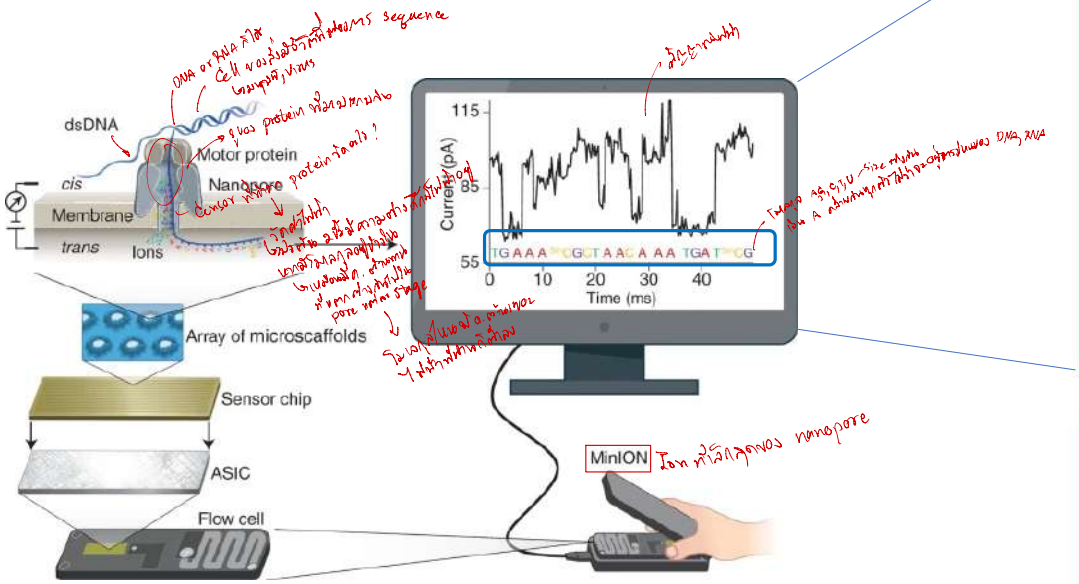


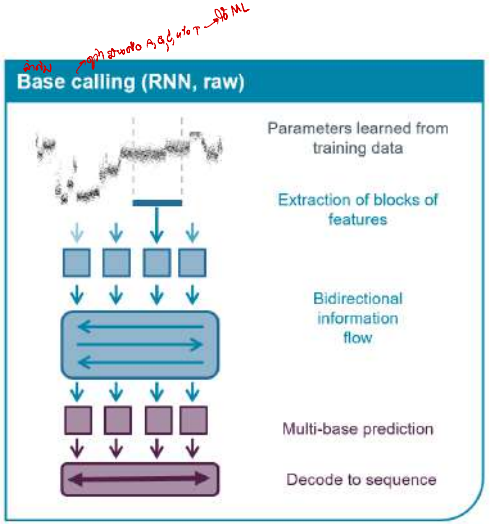
Now





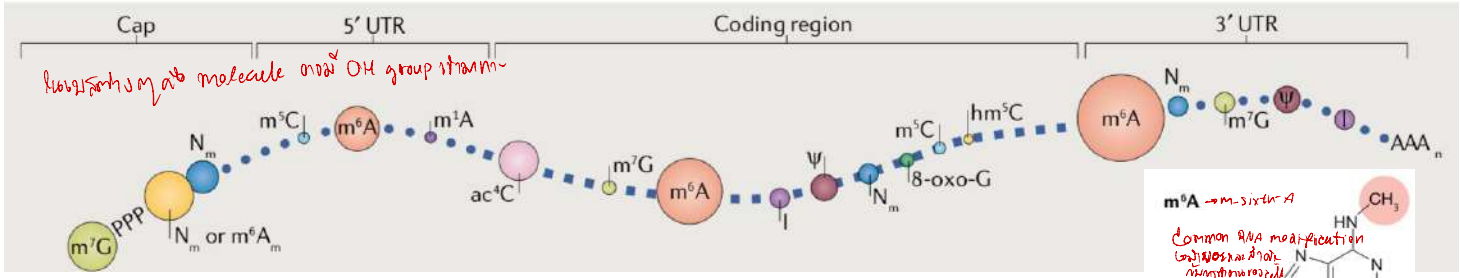
RNA Sequencing



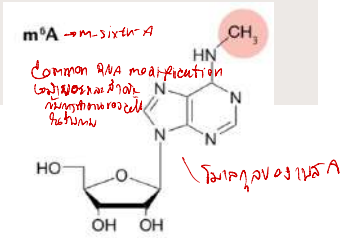


Ref: Yunhao Wang, *et al.*, "Nanopore sequencing technology, bioinformatics and applications", *Nature Biotechnology* (2021)

RNA modifications



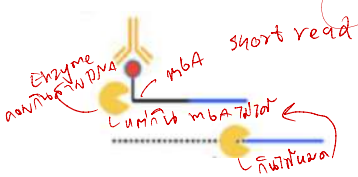
Ref: Zaccara, Sara, Ryan J. Ries, and Samie R. Jaffrey. *Nature Reviews Molecular Cell Biology* (2019)



Splicing | RNA Instability | Translation | Disease-related

Single-base-resolution CLIP-based detection methods

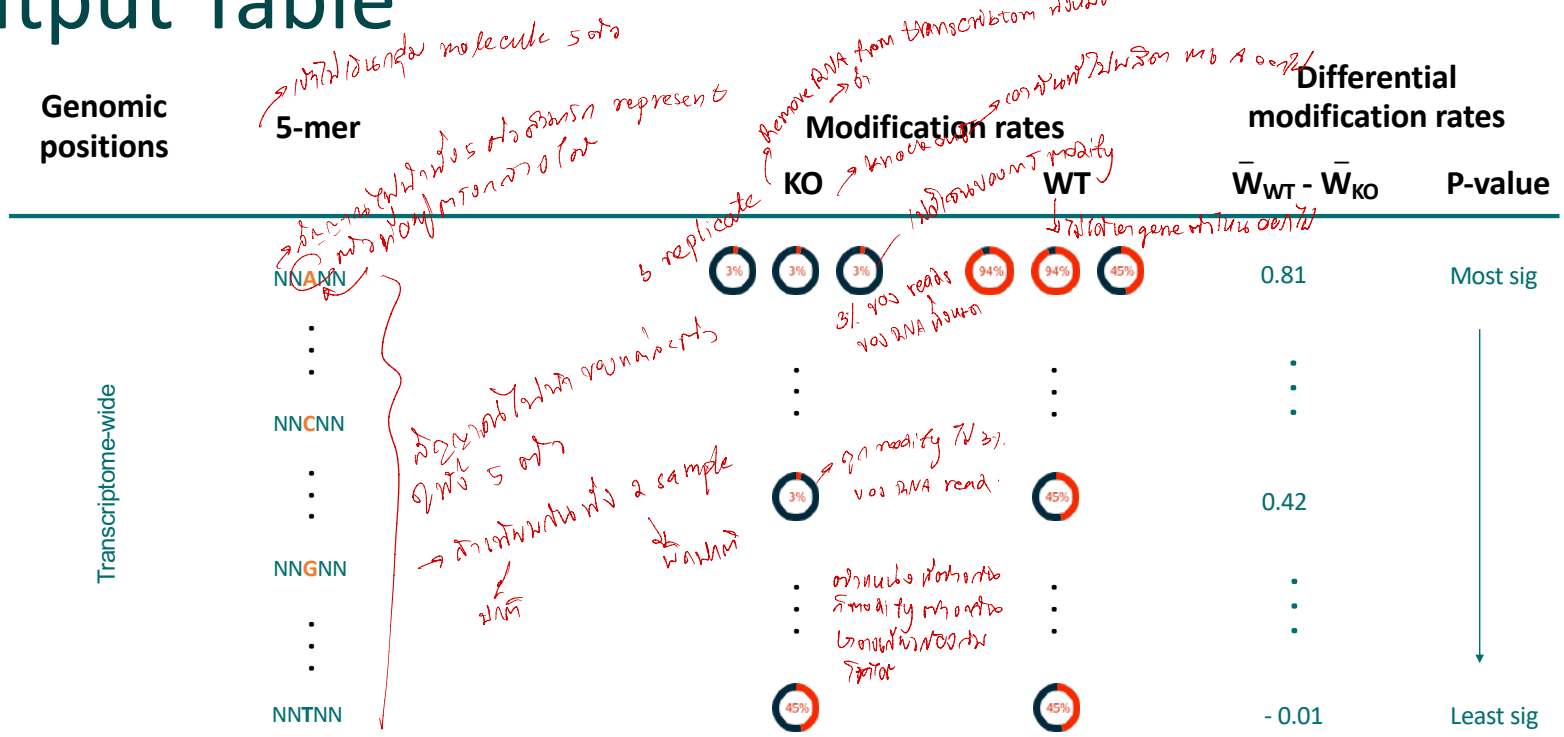
*Use **antibodies** to induce **truncations** or mutations at m6A sites during **reverse transcription**.*



m6ACE-Seq

Ref: Koh, Casslynn WQ, Yeek Teck Goh, and WS Sho Goh. *Nature Communications* 10.1 (2019)

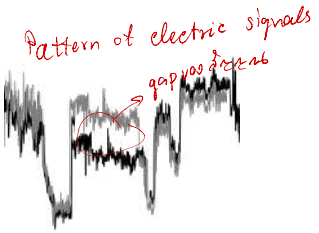
Output Table



Research Objectives

GGACT
GGm6ACT

Nanopore Sequencing



XPORE

- Locate modified positions
- *Quantify fraction of modified reads -- modification rate*

Signal-level modification detection methods

EpiNano

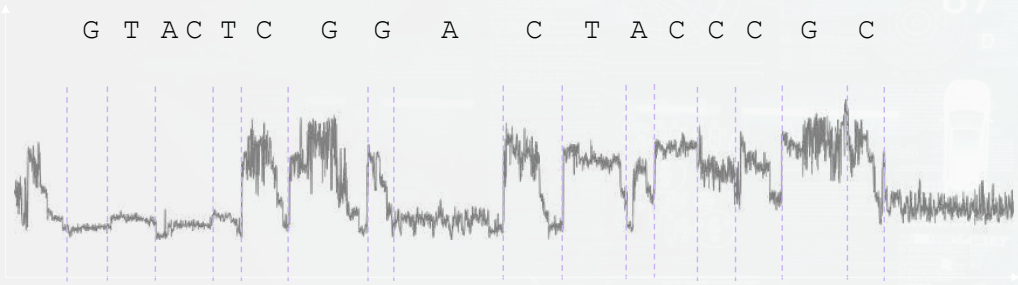
MINES

- m6A
- Training data required.

Tombo

- All modification types.
- No training data required.

2. Data Collection and Preparation



- **Nanopore Raw Signal Data**
- **Sequencing Data**
- **Genome Browser**
- **Nanopore Data Pipeline**

file from Nanopore Sequencer.

FAST5

- Raw signal - Sequencing output
- Intensity level (pA) → Current, Pico Ampere
- HDF5 format (binary), storing large and complex data

1 Fast5 file = 1 RNA read (1 การอ่านครั้งนึง)

store in dictionary file ({})

```
HDF5 "GISPC936_20181120_FAK27249_MN18749_sequencing_run_SHO_20112018_Empty
GROUP "/" {
   ATTRIBUTE "file_version" {
      DATATYPE  H5T_IEEE_F64LE
      DATASPACE  SCALAR
      DATA {
      (0): 0.6
      }
   }
   GROUP "PreviousReadInfo" {
      ATTRIBUTE "previous_read_id" {
         DATATYPE  H5T_STRING {
            STRSIZE 38;
            STRPAD H5T_STR_NULLTERM;
            CSET H5T_CSET_ASCII;
            CTYPE H5T_C_S1;
         }
         DATASPACE  SCALAR
         DATA {
         (0): "ac7312ce-d058-4382-a6c6-8471302869b9"
         }
      }
      ATTRIBUTE "previous_read_number" {
         DATATYPE  H5T_STD_U32LE
         DATASPACE  SCALAR
         DATA {
         (0): 976
         }
      }
   }
   GROUP "Raw" {
      GROUP "Reads" {
         GROUP "Read_984" {
            ATTRIBUTE "duration" {
               DATATYPE  H5T_STD_U32LE
               DATA {
               (0): 12639754
               }
            }
            DATASET "Signal" {
               DATATYPE  H5T_STD_I16LE
               DATASPACE  SIMPLE { ( 76256 ) / ( H5S_UNLIMITED ) }
               DATA {
               (0): 595, 492, 497, 502, 500, 499, 514, 495, 515, 512, 531,
               (11): 529, 515, 483, 497, 529, 510, 521, 524, 525, 523, 514,
               (22): 519, 517, 512, 520, 522, 519, 521, 517, 535, 514, 505,
               (33): 537, 527, 512, 521, 528, 523, 530, 530, 529, 529, 521,
               (44): 527, 515, 537, 522, 512, 485, 480, 481, 478, 465, 467,
               (55): 472, 476, 463, 469, 476, 454, 450, 446, 468, 471, 470,
               (66): 466, 468, 467, 466, 460, 458, 466, 467, 464, 465, 467,
               (77): 465, 459, 476, 470, 477, 460, 486, 470, 485, 486, 468,
               (88): 475, 470, 472, 472, 468, 456, 457, 452, 448, 440, 440,
               (99): 473, 470, 454, 442, 448, 449, 455, 461, 443, 455, 448,
               (110): 449, 444, 462, 456, 461, 459, 467, 459, 461, 458, 472,
               (121): 461, 463, 467, 456, 471, 468, 471, 475, 467, 466, 471,
               (132): 477, 459, 473, 482, 466, 477, 470, 461, 464, 452, 454,
               (143): 457, 468, 457, 466, 472, 474, 441, 456, 470, 467, 444,
               (154): 442, 455, 451, 456, 470, 469, 473, 479, 478, 468, 472,
               (165): 462, 466, 458, 435, 436, 464, 467, 455, 462, 463, 471,
               (176): 455, 459, 446, 460, 442, 453, 465, 465, 480, 465, 478,
               (187): 467, 475, 483, 512, 502, 539, 521, 506, 521, 523, 516,
               (198): 518, 511, 514, 518, 530, 516, 528, 503, 503, 510, 524,
               (209): 529, 526, 513, 504, 469, 476, 472, 470, 468, 476, 476,
               (220): 476, 471, 459, 457, 432, 443, 472, 466, 477, 467, 471,
               (231): 470, 474, 449, 468, 456, 457, 460, 459, 459, 456, 469,
               (242): 457, 469, 475, 468, 465, 465, 463, 446, 455, 458, 461,
               (253): 456, 448, 446, 462, 444, 464, 462, 469, 479, 471, 502,
```

FASTQ

- Basecalled sequence
- Text format:
  - Name/ID, starting with "@"
  - Sequence
  - Optional info, starting with "+"
  - Quality of the sequence, encoding the probability error

```
@2fdda8fc-621a-415e-baaf-1b088a91302d runid=385c19b6dda90bee8f8b01b21643004d7cb1d235 sampleid=SHO_20112018_EmptyE2-9 read=65062 ch=111 start_time=2018-11-20T22:13:23Z
CAACAUCGCGGCCGUGGCGGGAUGCUGUGGGAGGGGUCAUCCUGCCCAGGCCGGGCCGGCUGCCUACCCUUCCAGACUGUCCGCAUCGGCUCCAAGGCUGCUACGCCACCGUCGUGUGAUGCCAUCAGCGGCAGAAGCUGAUGGCUGUGACUGAAUAUCCCCCGACAGCCAUCCU
UCCCAUCAUGCCUGCCACUUCCAGCCCCCACAGGAAGAUAGGCCUCAUCAGGCUUUUCCGCCAGAUAGCAGCAGUUUUCCAGGACAACCGAAUGAUAGCGUUGCCAGAAUGUGGCUUGAGUGCAGAGGGACAAGCUUUUUCGUGCGACACCAGCUUGCGGAAACACAAGAUCCUGA
UGAAGGUUUCCCCAACCAGGUCCUGAAGCCCUUCCUGAGGAUUCCAAGUACCAAAAUCUGCUGCCCCUUUUGUGGGGCACUAUCGCUGCUGGUCAAAUGAAGAGCCCGGGUCAAAGAUGGUCACGGAUCUUUAAGGACUGUGCCAUCACCCUGCCGCUGCAUGGGUGGGCGGCAUU
GAUGACACCAUCCAGCAGGCGGCUUUAUCAACUACUCCUUGACCCAGCCUGCCCUGGUGCAGGAGCUUGUAGGAGGCCUCACCUGCCACAGCCCAGACCCACUCCCUGCUCCAGCACCAGCCCCACCGUUGACCACCCCUGUUGGACCAGUACAUCAGAGAGCAACGCGCGAGAAG
GAUUUGUCAUGCGGCCAAUGGGUCAGAUCCUGACACUGGGUUCCGGACUUGCAGCAGCCUGUUCGGCCAGCCCCUGCGCAUAAAUACAUUUGUGUUAUUGGCUGUGCUCUCCUUGCCGAGCUCGUGGAAGAACAUUGUCGAGGAGUGUUUGUCACUUGGUUCACUAGAGGUAAUGA
UAUUGUCGUGUAGGUAAUGGGCACUUGGAGAUGCAGAGGAUUCCAUUUCAGAUGUCAGUCACCGAAUCUCGUCCUUAGUUUCCCUAUUGGGACGUGAUAUGGGAGCAAAGACUUCCAUUACCCAGGUCCAAGGCAGAGAGAGAGAUCUGAAAAGAUAGGGCUAUUGUCCCCUGGCC
CUUGGUCACUGCCUUUUUGCUGGCACGACCUGAGCCUCCCUCCUUGGGGCACAACCUAUCCCAUUGCCACAGUAGCUCAUCAAGCAGUUGAGUGCUGAGAAUGGCACACCUGGUGAGAGCCUGUUGUGUGCCAGGCUUUGUGUGUGAGUGCUGUCAUGUAUUAGUUCCUUACUGCU
GACCACUGUACCCAUAUCACAGAGAAGGAGCAGAGAAAUUAAGUGGGCUUGCUCCACAAGGUCAUGCAGUUAGUAAGUGGCAGACUACGGACUUGAACCAAGCCUCUGCUUUGAGUCGCGUCCUGAAUUUUUCACUAGAGCUCACCACAUUUCAGGUUACCCAGAAGUGGGUCCA
UCCAUCCAGGUGUGCUUGGAUGUUAGUUCCACCCUCGAGGUGUCUGUGAAAAGUUGGGAGCAUUGUUUUUAUAAAAAUGAAUAUUUAAAAAACC
+
%$$$$$%%&(,)++),-(-7.+/0)))&..,%*,0/&$$&),*(/6.)')%''5--.+,)&((-(*)2++)-.+-.354688.+'')*1)(*0-,0.,'($)',+*-*/3--)))+(+-.23(&+&&'*0'%%'(/'((+-,-2++,./,++,,2231.:42($&''(+/4.+-+'
)'-)**),,.07110-'%&()+..4783-++&(*&*(+*+./0.1*,,,)(.''/.**1(+-(((&*+)*,++-)*.***').&&),+)*2-.01.,))'&())7,,-.,+)+1)&&%%&%)++*)*0*)+.*''),*)''''))**,-.2,+*,)&&&,(3/-,,-/-.-/1)+,
+,.++-).(197/.+++)3)*-+*-0)*+2-'',(,+)*0)+,+/-4,))*3)42,+'.++-0-22'+,)-,02172()*&(&%&&&))+,2***&$&+.,,,+/0)&+/*+,5.**+1.*'(*,1,.%+)+&(-0,-..-/51.&$$&-.&*+,.1/.-()*/.&'+)),.'')-
((%')(,.0(())&**)1.*)+(*,),-0/-(**)'(*$&'(,8-(*9-909://420/,).*)*.1+*&$')%*+..))(-/002/))(&(/*//-.4,)*'/51-.,.611/0+/..,922(#''%$)''0,*/61*,,4,7-,1..2/...*-,*)***-84-,*)++,+1)5
++23,+)))((%(*1*1,+,,+&)+/333,+++,))(()-'**3+6(&'(((*+,03:5-,)-$(,.706-792,-,*))(25/)**+,+((&).--2.240/132)('&,+-0.--..**)(&'(.+.-2)%$()0)+**+,%'%&'*0(+*)*0-0+*-'((1+)')++52/0*
)',,0.+'&(%$&)&**),,,,((204.,+,,+*(&&,(*7261//3*,./**++,25-+*,,))-&%+,,22/13-0)3)(2.#((()--'(),-,,(')++')'/..(%%$%'+'()%&12()0.,*-,'/&%''''&&(()-.))*341,)('(01)*+-0.*,66.'%().
*0-.+''&%$()''('('))(&/,()('&)(()'),&',+$'+(+*0.,,*--&*+,''0-&&(,./4/+**+-1-.*,%,+-,,/0+'%&(()'('%,.//-2-+*(//.9431/.++'-%()*(*()**/++-'*/.)-,)()&%($(*,'$%%(*+*+*&/2,),),'()))*
*-0)'&)''%'/.(''&&')*&'&*+/-,)(''&'12,/.3+)((**-/.,()'*$'+,*.*+575645810.,-))'*'&(*)'&&%)4,+.5./0-..#%)+++'+.(((*'-,()'&)*0.+-10)/+)'%$''(*++*)&$(($%&'*(%'))))%(0-.01,+()+)'*/(
%&)%&-.),,)**)'',//0-.4-'+),')$(+)()((.+)+'&)())*66.,/--1+*+*)+*('*('&&%,+5.*',/,((-+()2/3+)+$
```

1 RNA read

FASTA

- Reference sequence
- Text format:
  - Sequence ID, starting with ">", optionally followed by other attributes
  - Sequence

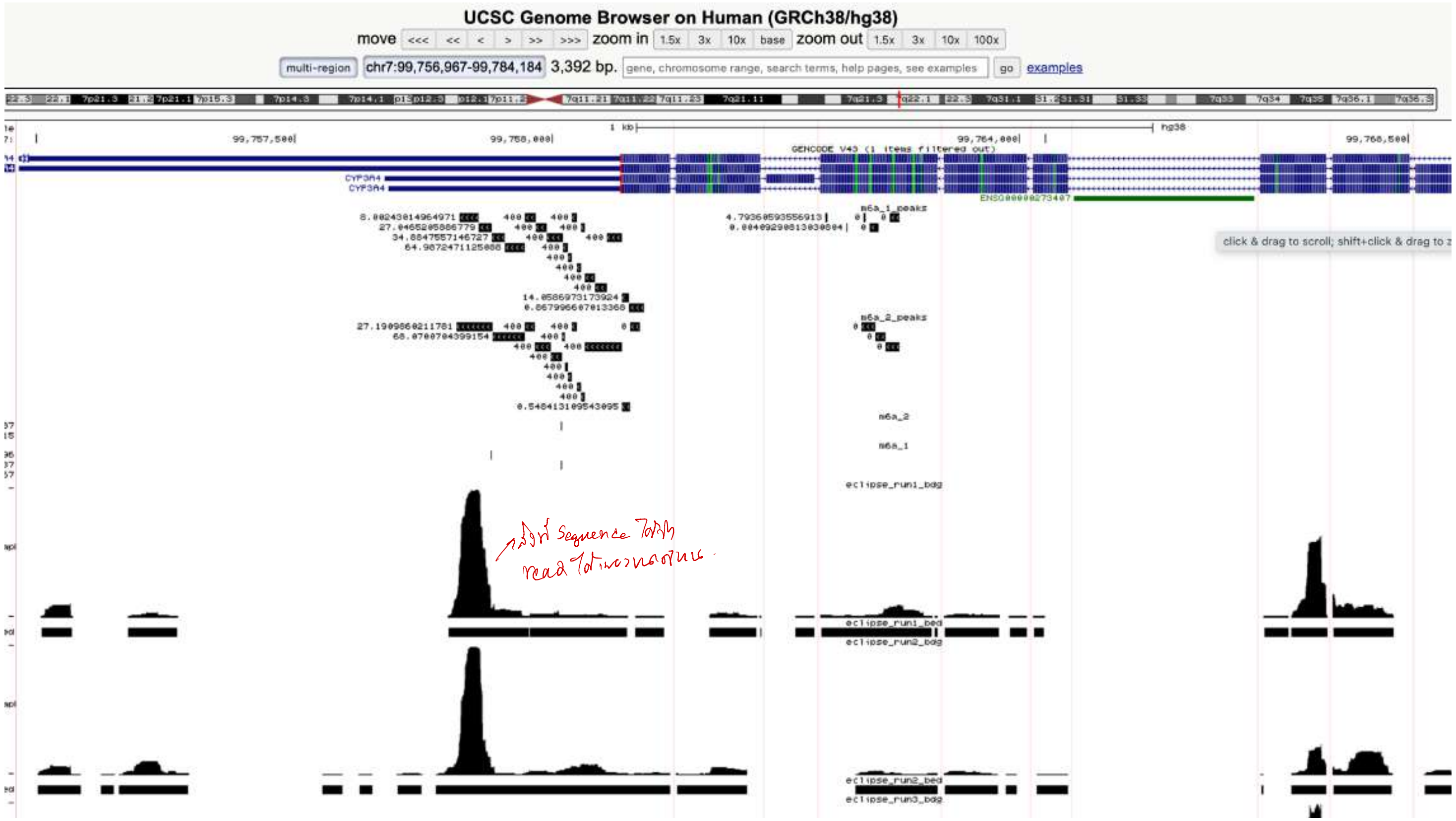
```
>ENST00000480901.1 cdna chromosome:GRCh38:17:47828308:47831525:-1 gene:ENSG00000159111.12 gene_biotype:protein_coding transcript_biotype:retained_intron gene_symbol:MRPL10 desc
ription:mitochondrial ribosomal protein L10 [Source:HGNC Symbol;Acc:HGNC:14055]
TTCTTCCGGTGGAGATGGCTGCGGCCGTGGCGGGGATGCTGCGAGGGGGTCTCCTGCCCC
AGGCGGGTAAGGAGTGGCCCAGGTCCTCACGGCGTGTCTTGCGGCCGCTCTCTAGTCCTC
ATCTGCCCTCCTCTACTACTGATTCTTCCCATAATCTCTGACCCCAGCTAGATCGCTGGC
CTCCTTACCCCGTCCAGTTCCTTGTGACTCGACTGGCCGGCTGCCTACCCTCCAGACTGT
CCGCTATGGCTCCAAGGCTGTTACCCGCCACCGTCGTGTGATGCACTTTCAGCGGCAGAA
GCTGATGGCTGTGACTGAATATATCCCCCCGAAACCAGCCATCCACCCATCATGCCTGCC
ATCTCCTCCCAGCCCCCCACAGGAGGTAAGGAGGAATTTGGGTACATGTCACTTGGTGGT
GGGATGGTGGATTAAAGTAATCTTGTCTCTGGCCATAGTGAAGTAGGACACTCAGCCATT
GTCATGCACGTCATTATTTCAGTTTGACTGCCTGATCCAGATATTTTAAGATGAAATCCG
CACTTGATTCTGTATTGGCTTTTGGGCTCTGGATTGGG
```

THAILAND CODING & AI ACADEMY

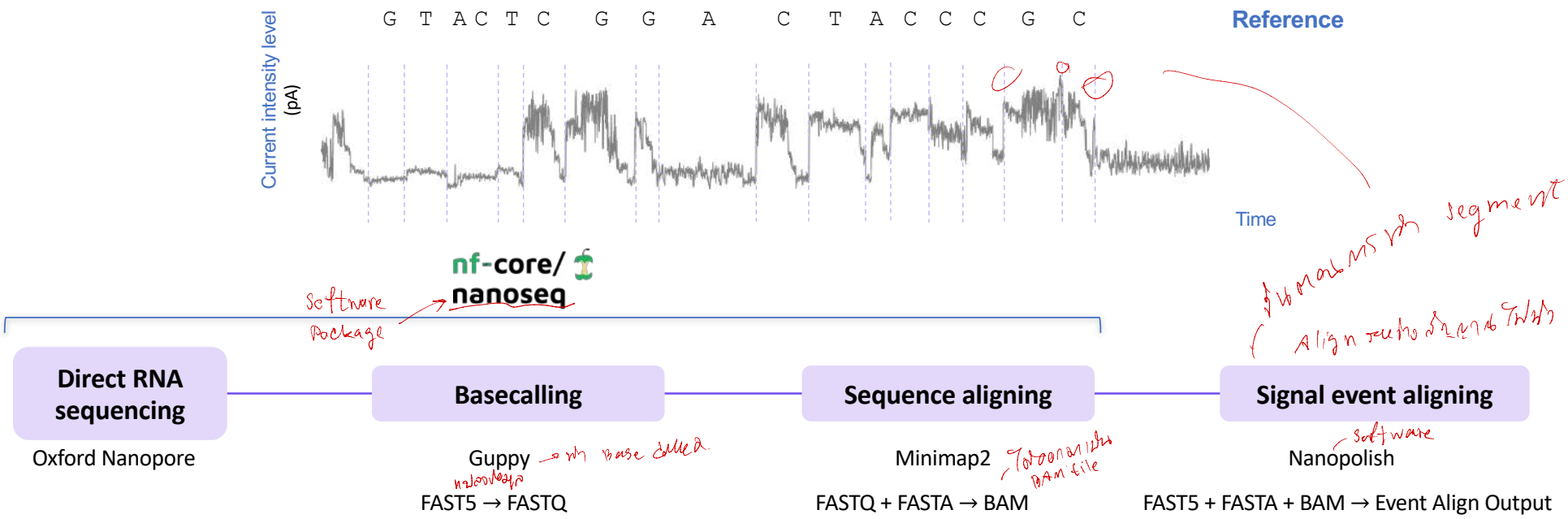
BAM / SAM

- Alignment results (FASTQ aligned with FASTA)
- BAM – Binary / SAM – Text

```
2fdda8fc-621a-415e-baaf-1b088a91302d    0       ENST00000351111.6       13      13      7S11M1D13M1I6M1I5M1D6M1I18M1I14M1D2M1I16M2D19M1D1M1D31M2D12M3D9M1I17M4D7M1D11M2D27M3D36M
1D3M1D15M1D10M1I13M1I13M1I28M1D29M1D31M1D13M1I3M2D10M1I12M1D7M2D7M1I7M1I16M2I12M1I5M1I21M2D8M2D21M2D10M1D10M3D25M2D41M1D8M1I29M2I13M1D7M1D10M3D16M2I15M1D16M1I17M1D25M3D2M1I3M1I
13M1I2M3D8M2D15M2D7M3I14M4I5M1D51M1I10M1D8M1D13M2I12M1D21M6I6M1D29M1D14M1I7M1I5M3D29M2I20M1D11M2I13M2I36M1I11M1D14M1D14M2D38M1I8M3I28M2I2M1D15M1D11M1D2M1D15M1D12M1I6M2I21M1D4M3
D26M2D16M3D10M1D12M1I7M2D8M13S  *       0       0       CAACATCGCGGCCGTGGCGGATGCTGTGGGAGGGGTCATCCTGCCCAGGCCGGGCCGGCTGCCTACCCTTCCAGACTGTCCGCATCGGCTCCAAGGCTGCTACGCCACCGTCGTGTGAT
GCCATCAGCGGCAGAAGCTGATGGCTGTGACTGAATATCCCCCGACAGCCATCCTTCCCATCATGCCTGCCACTTCCAGCCCCACAGGAAGATAGGCCTCATCAGGCTTTTCCGCCAGATAGCAGCAGTTTTCCAGGACAACCGAATGATAGCGTTGCCAGAATGTGGCTTGAG
TGCAGAGGGACAAGCTTTTTCGTGCGACACCAGCTTGCGGAAACACAAGATCCTGATGAAGGTTTCCCCAACCAGGTCCTGAAGCCCTTCCTGAGGATTCCAAGTACCAAAATCTGCTGCCCCTTTTGTGGGGCACTATCGCTGCTGGTCAAATGAAGAGCCCGGGTCAAAGATGG
TCACGGATCTTTAAGGACTGTGCCATCACCCTGCCGCTGCATGGGTGGGCGGCATTGATGACACCATCCAGCAGGCGGCTTTATCAACTACTCCTTGACCCAGCCTGCCCTGGTGCAGGAGCTTGTAGGAGGCCTCACCTGCCACAGCCCAGACCCACTCCCTGCTCCAGCACCAG
CCCCACCGTTGACCACCCCTGTTGGACCAGTACATCAGAGAGCAACGCGCGAGAAGGATTTGTCATGCGGCCAATGGGTCAGATCCTGACACTGGGTTCCGGACTTGCAGCAGCCTGTTCGGCCAGCCCCTGCGCATAAATACATTTGTGTTATTGGCTGTGCTCTCCTTGCCGAG
CTCGTGGAAGAACATTGTCGAGGAGTGTTTGTCACTTGGTTCACTAGAGGTAATGATATTGTCGTGTAGGTAATGGGCACTTGGAGATGCAGAGGATTCCATTTCAGATGTCAGTCACCGAATCTCGTCCTTAGTTTCCCTATTGGGACGTGATATGGGAGCAAAGACTTCCATTA
CCCAGGTCCAAGGCAGAGAGAGAGATCTGAAAAGATAGGGCTATTGTCCCCTGGCCCTTGGTCACTGCCTTTTTGCTGGCACGACCTGAGCCTCCCTCCTTGGGGCACAACCTATCCCATTGCCACAGTAGCTCATCAAGCAGTTGAGTGCTGAGAATGGCACACCTGGTGAGAGC
CTGTTGTGTGCCAGGCTTTGTGTGTGAGTGCTGTCATGTATTAGTTCCTTACTGCTGACCACTGTACCCATATCACAGAGAAGGAGCAGAGAAATTAAGTGGGCTTGCTCCACAAGGTCATGCAGTTAGTAAGTGGCAGACTACGGGACTTGAACCAAGCCTCTGCTTTGAGTCGC
GTCCTGAATTTTTCACTAGAGCTCACCACATTTCAGGTTACCCAGAAGTGGGTCCATCCATCCAGGTGTGCTTGGATGTTAGTTCCACCCTCGAGGTGTCTGTGAAAAGTTGGGAGCATTGTTTTTATAAAAATGAATATTTAAAAAACC  %$$$$$%%&(,)++),-(-7.+/0
)))&..,%*,0/&$$&),*(/6.)')%''5--.+,)&((-(*)2++)-.+-.354688.+'')*1)(*0-,0.,'($)',+*-*/3--)))+(+-.23(&+&&'*0'%%'(/'((+-,-2++,./,++,,2231.:42($&''(+/4.+-+')'-)**),,.07110-'%&()+..
4783-++&(*&*(+*+./0.1*,,,)(.''/.**1(+-(((&*+)*,++-)*.***').&&),+)*2-.01.,))'&()7,,-.,+)+1)&&%%&%)++*)*0*)+.*''),*)''''))**,-.2,+*,)&&&,(3/-,,-/-.-/1)+,+,.++-).(197/.+++)3)*-+*
-0)*+2-'',(,+)*0)+,+/-4,))*3)42,+'.++-0-22'+,)-,02172()*&(&%&&&))+,2***&$&+.,,,+/0)&+/*+,5.**+1.*'(*,1,.%+)+&(-0,-..-/51.&$$&-.&*+,.1/.-()*/.&'+)),.'')-((%')(,.0(())&**)1.*)+(*
,),-0/-(**)'(*$&'(,8-(*9-909://420/,).*)*.1+*&$')%*+..))(-/002/))(&(/*//-.4,)*'/51-.,.611/0+/..,922(#''%$)''0,*/61*,,4,7-,1..2/...*-,*)***-84-,*)++,+1)5++23,+)))((%(*1*1,+,,+&)
+/333,+++,))(()-'**3+6(&'(((*+,03:5-,)-$(,.706-792,-,*))(25/)**+,+((&).--2.240/132)('&,+-0.--..**)(&'(.+.-2)%$()0)+**+,%'%&'*0(+*)*0-0+*-'((1+)')++52/0*)',,0.+'&(%$&)&**),,,,((
204.,+,,+*(&&,(*7261//3*,./**++,25-+*,,))-&%+,,22/13-0)3)(2.#((()--'(),-,,(')++')'/..(%%$%'+'()%&12()0.,*-,'/&%'''''&&(()-.))*341,)('(01)*+-0.*,66.'%().*0-.+''&%$()''('('))(&/,
()('&)(()'),&',+$'+(+*0.,,*--&*+,''0-&&(,./4/+**+-1-.*,%,+-,,/0+'%&(()'('%,.//-2-+*(//.9431/.++'-%()*(*()**/++-'*/.)-,)()&%($(*,'$%%(*+**&/2,),),'()))**-0)'&)''%'/.(''&&')*&'&
*+/-,)(''&'12,/.3+)((**-/.,()'*$'+,*.*+575645810.,-))'*'&(*)'&&%)4,+.5./0-..#%)+++'+.(((*'-,()'&)*0.+-10)/+)'%$''(*++*)&$(($%&'*(%'))))%(0-.01,+()+)'*/(%&)%&-.),,)**)'',//0-.4-
'+),')$(+)()((.+)+'&)())*66.,/--1+*+*)+*('*('&&%,+5.*',/,((-+()2/3+)+$  NM:i:212        ms:i:1794       AS:i:1794       nn:i:0  ts:A:+  tp:A:P  cm:i:50 s1:i:490        s2:i:47d
e:f:0.1026
```


UCSC Genome Browser on Human (GRCh38/hg38)
move
zoom in
zoom out
multi-region
chr7:99,756,967-99,784,184
3,392 bp.
go
examples
click & drag to scroll; shift+click & drag to z

Nanopore pre-processing pipeline for signal-level data analysis



https://xpore.readthedocs.io/en/latest/

Data preparation from raw reads

1. After obtaining fast5 files, the first step is to basecall them. Below is an example script to run Guppy basecaller. You can find more detail about basecalling at Oxford nanopore Technologies:

```
guppy_basecaller -i </PATH/TO/FAST5> -s </PATH/TO/FASTQ> --flowcell <FLOWCELL_ID> --kit <KI
```

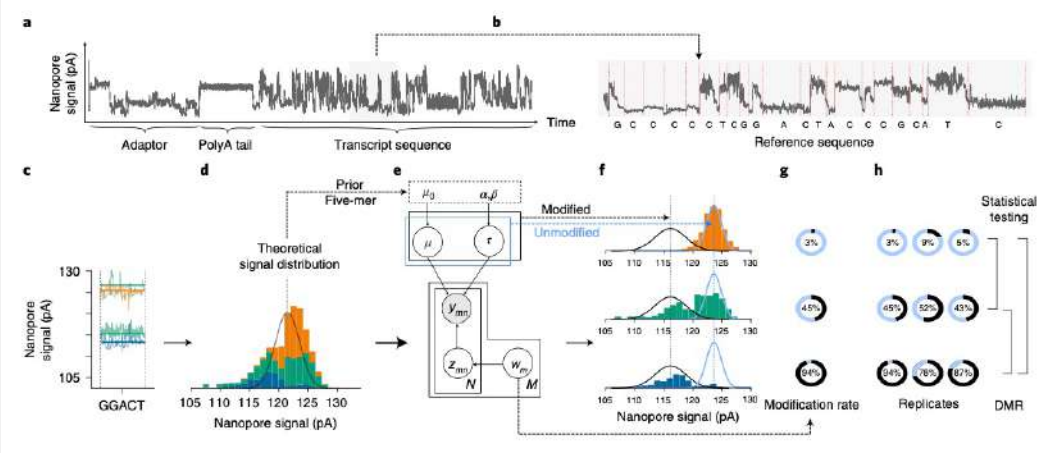
2. Align to transcriptome:

```
minimap2 -ax map-ont -uf -t 3 --secondary=no <MMI> <PATH/TO/FASTQ.GZ> > <PATH/TO/SAM> 2>> <
samtools view -Sb <PATH/TO/SAM> | samtools sort -o <PATH/TO/BAM> - &>> <PATH/TO/BAM_LOG>
samtools index <PATH/TO/BAM> &>> <PATH/TO/BAM_INDEX_LOG>
```

3. Resquiggle using nanopolish eventalign:

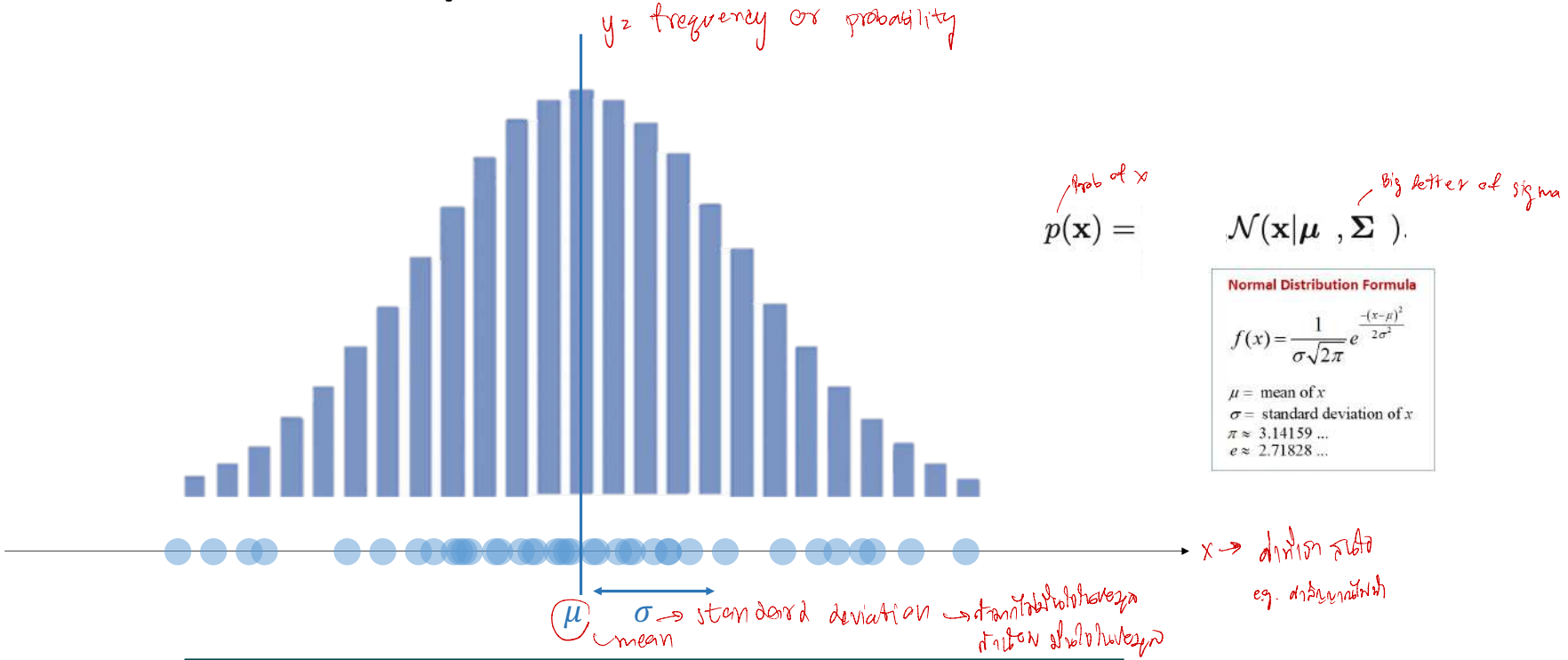
```
nanopolish index -d <PATH/TO/FAST5_DIR> <PATH/TO/FASTQ_FILE>
nanopolish eventalign --reads <PATH/TO/FASTQ_FILE> \
--bam <PATH/TO/BAM_FILE> \
--genome <PATH/TO/FASTA_FILE \
--signal-index \
--scale-events \
--summary <PATH/TO/summary.txt> \
--threads 32 > <PATH/TO/eventalign.txt>
```

3. Bayesian [Multi-Sample] Gaussian Mixture Modelling

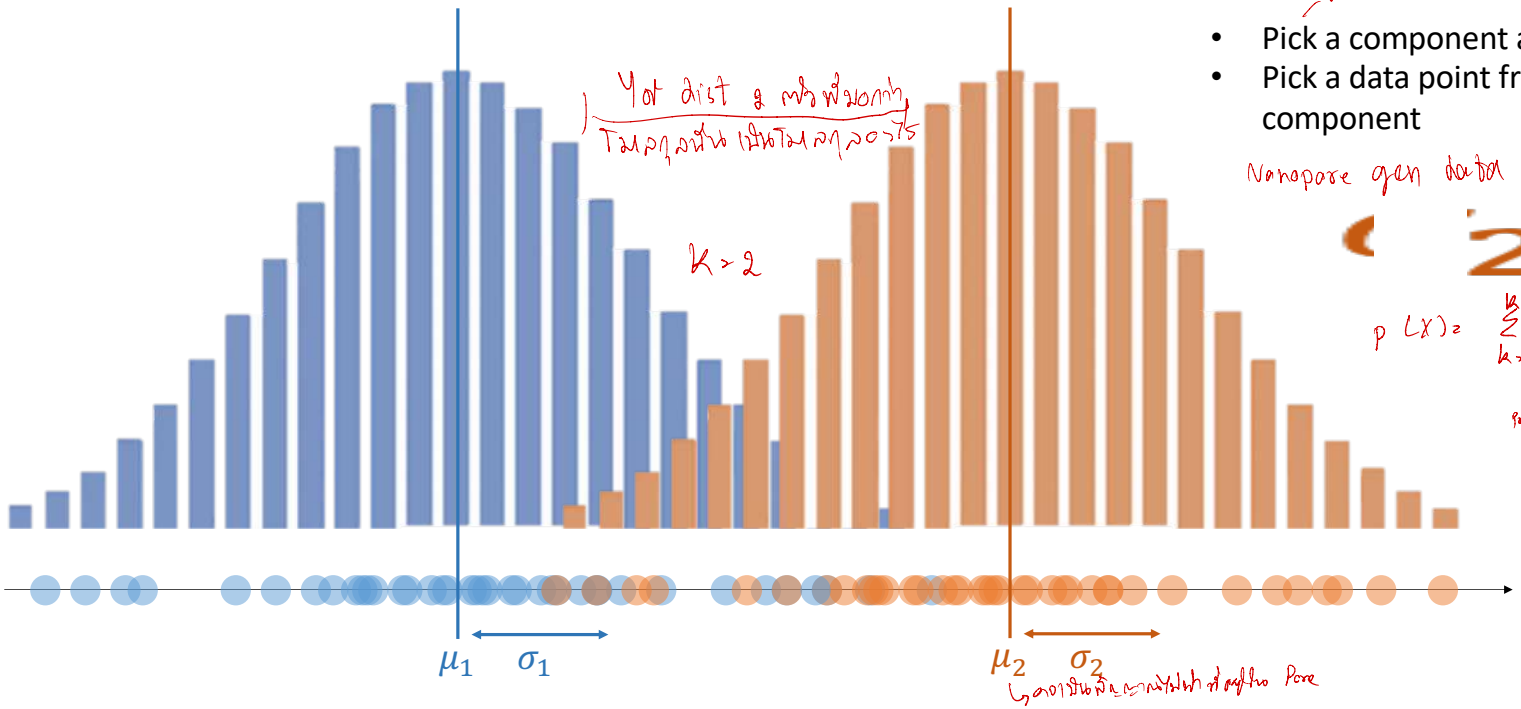


- [Bayesian] GMM
- Where did the idea come from?
- How Multi-Sample?
- Why Bayesian?
- Speed-Up ML Experiments

Bayesian Multi-Sample Gaussian Mixture Model



Bayesian Multi-Sample Gaussian Mixture Model



- Pick a component at random
- Pick a data point from the chosen component

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

What is GMM?

Assumption how data are **generated** as follows

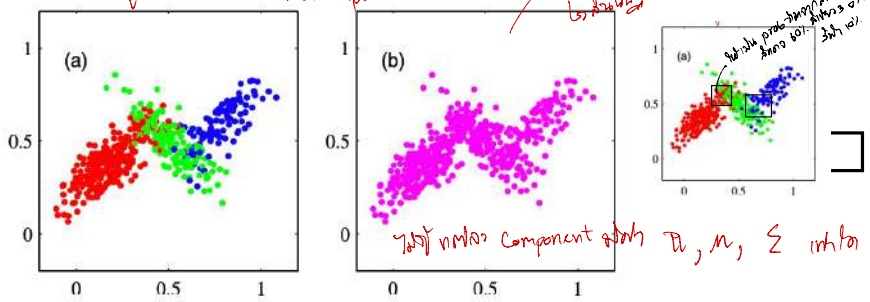
- There are K components
- Each component is defined as a Gaussian distribution
- Pick a component at random
- Pick a data point from the chosen component

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

θ

Data

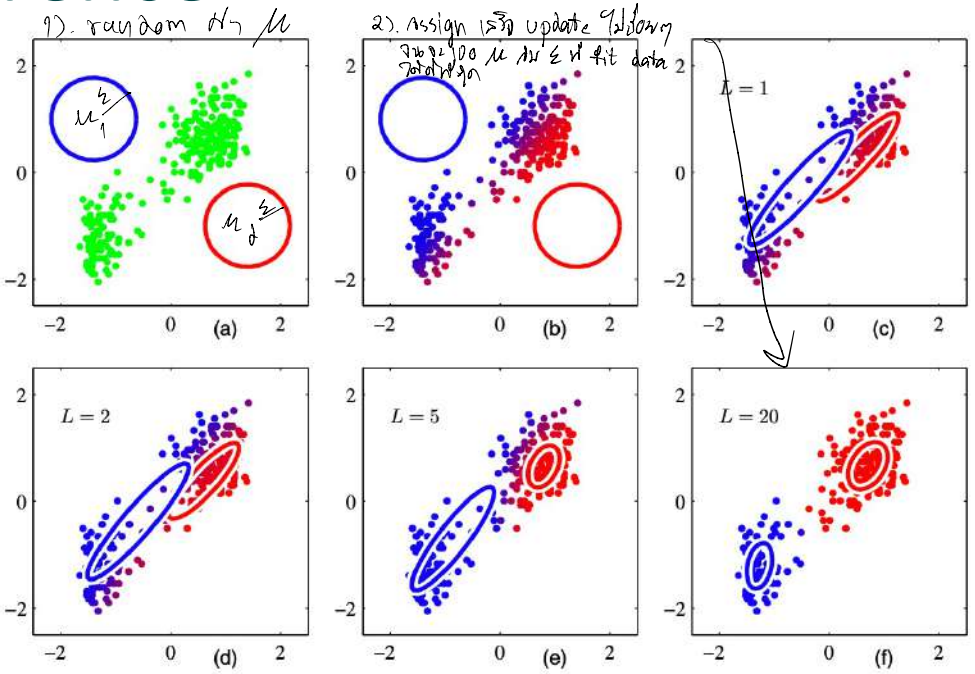
Parameter = π, μ, Σ



Source: Christopher M. Bishop, "Pattern Recognition and Machine Learning", 2006

GMM Inference

Iterative algorithm



Source: Christopher M. Bishop, “Pattern Recognition and Machine Learning”, 2006

model แรก ๆ ของ

Generative AI

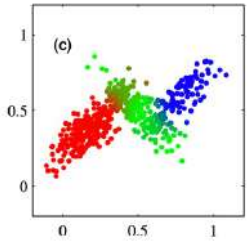
39067

ML ที่ใช้ Density

GMM
as a **Density Estimator**

Model
how data are generated

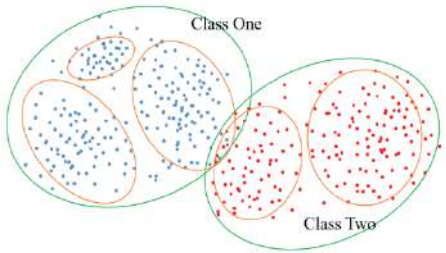
Assumption



Clustering

ลด data ก่อน → ทำ Clustering ก็ดี

(One-Calss) Classification

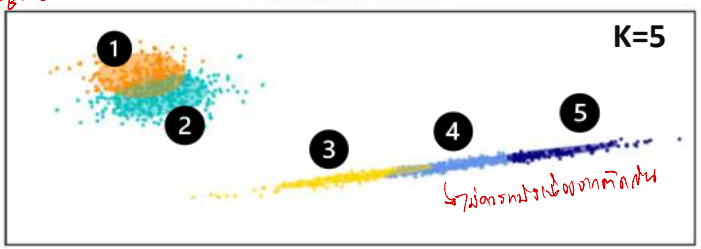


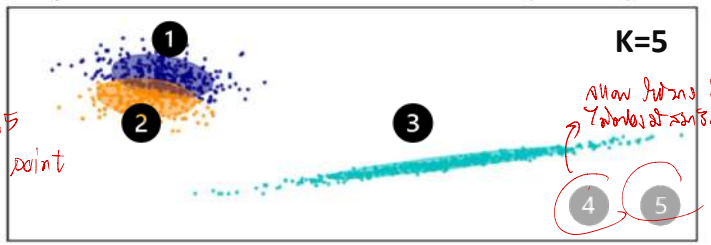
Bayesian Multi-Sample Gaussian Mixture Model

Learning algorithm for making inference on the **latent** variables

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

π_k : mixing parameter, Θ : $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k$, Data





Point estimate = Maximum Likelihood

Θ = argmax P(Data | Θ)
Θ

Posterior = Likelihood x Prior

P(Θ | Data) = P(Data | Θ) x P(Θ)

Frequentist vs Bayesian

$P(Data \mid \Theta)$

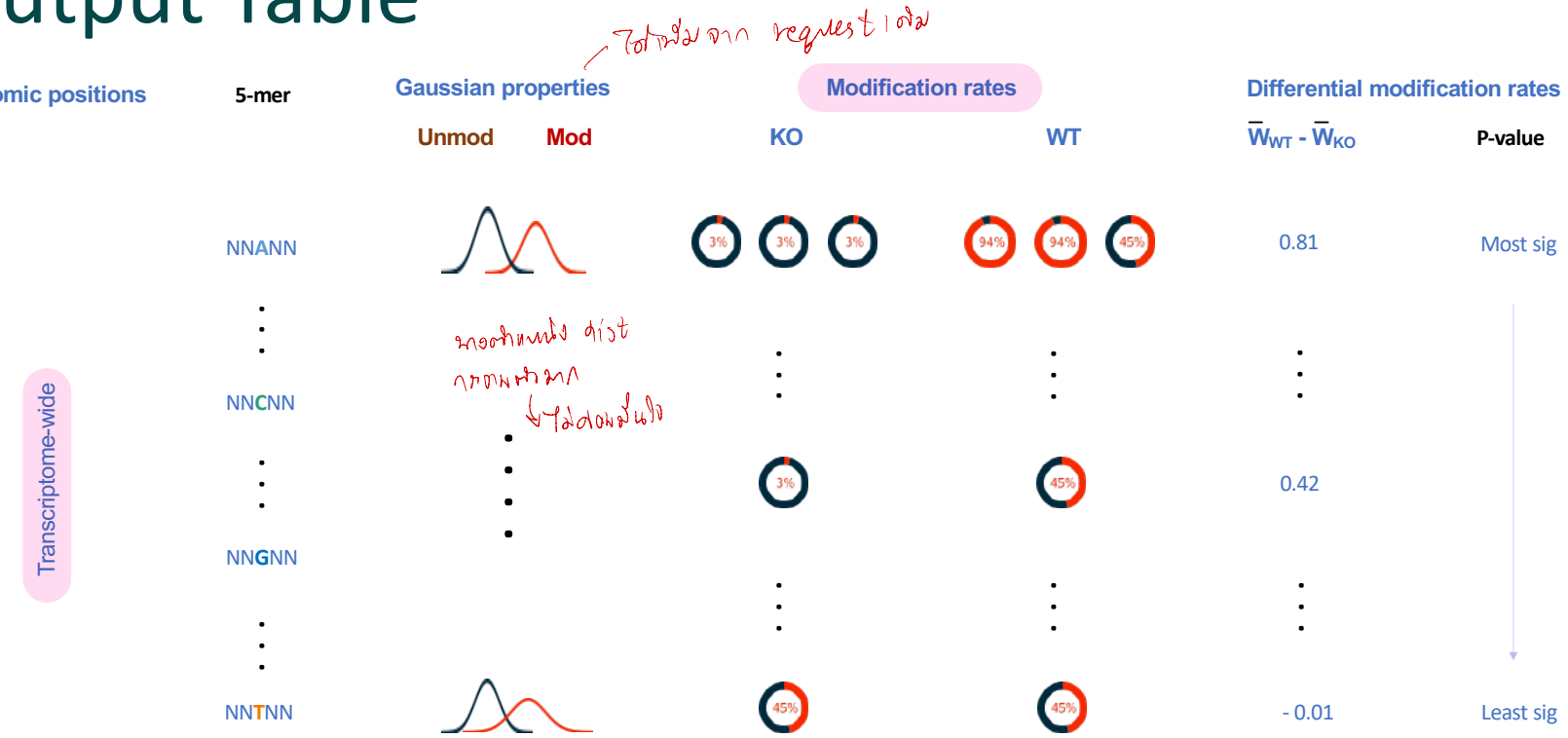
$P(\Theta \mid Data) = P(Data \mid \Theta) \times P(\Theta)$

Aspect	Frequentist	Bayesian
Probability interpation	Long term frequency	Posterior
Treatment of parameters	Fixed / Point estimates	Random / Probability distributions
Prior information	No	Yes
Sample size requirement	Larger	Smaller
Interpretation of results	Focused on the observed data	In the context of prior beliefs and their updates based on the observed data
Computational complexity	Simpler	More complex

Bayesian Multi-Sample Gaussian Mixture Model

- Each site has 2 distributions at maximum
  - Unmodified
  - Modified by only one modification type
- To accommodate comparison **across many samples**
- Nanopolish Eventalign assume a Gaussian distribution for each signal event (kmer)
- Mean and variance of unmodified k-mer is estimated by the eventalign algorithm *in priori*
- Fast → **Parallelisation**

Output Table



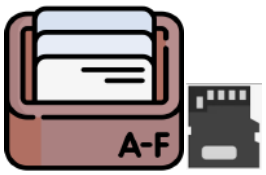
Speed-Up ML Experiments

Automated ML models

- Hyper-parameter settings
- Multiple datasets
- Different models / methods

- Config file
- Python packaging

- Parallelization



- File indexing

Why config files?

- Automating tasks
- Centralised configuration
- Documentation
- Portability

YAML, JSON, TOML, and INI are the popular and standardised formats of configuration files

```
xpore diffmod --config Hek293T_config.yml
```

Configuration file

xpore / xpore / diffmod / configurator.py

Code | Blame | 78 lines (63 loc) · 2.68 KB | Raw

```python
import yaml
import os
from collections import defaultdict

from ..utils import misc

def get_condition_run_name(condition_name,run_name):
    return '-'.join([condition_name,run_name])

class Configurator(object):
    def __init__(self, config_filepath):
        self.filepath = os.path.abspath(config_filepath)
        self.filename = self.filepath.split('/')[-1]
        self.yaml = yaml.safe_load(open(self.filepath, 'r'))

    def get_paths(self):
        paths = {}

        if 'prior' in self.yaml:
            paths['model_kmer'] = os.path.abspath(self.yaml['prior'])
        else:
            paths['model_kmer'] = os.path.join(os.path.dirname(__file__),'model_kmer.csv')

        paths['out_dir'] = os.path.join(os.path.abspath(self.yaml['out']))
        paths.update(misc.makedirs(paths['out_dir'],sub_dirs=['models']))
        paths['model_filepath'] = os.path.join(paths['out_dir'], 'models', '%s.model')
        return paths
```

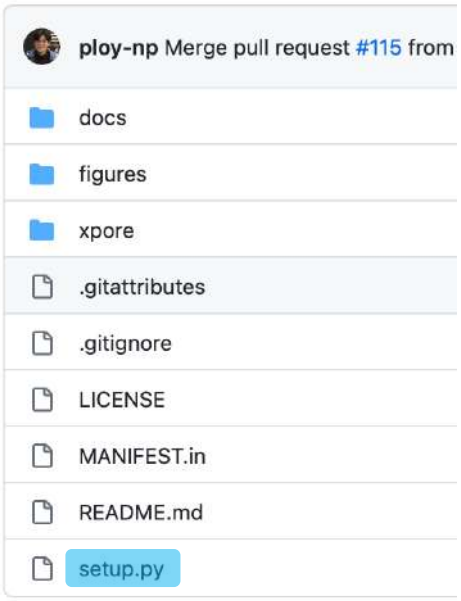
```python
config = Configurator(config_filepath)
paths = config.get_paths()
data_info = config.get_data_info()
method = config.get_method()
criteria = config.get_criteria()
prior_params = config.get_priors()
```

output, input ตัวอย่างไหน

ทดลอง 10 หมวด ใน 10 files

```
data:
    <CONDITION_NAME_1>:
        <REP1>: <DIR_PATH_TO_DATA_JSON>
        ...

    <CONDITION_NAME_2>:
        <REP1>: <DIR_PATH_TO_DATA_JSON>
        ...

    ...

out: <DIR_PATH_FOR_OUTPUTS>

criteria:
    readcount_min: <15>
    readcount_max: <1000>

method:
    # To speed up xpore-diffmod, you can use a statistical test (currently only t-test is imple
    # to remove positions that are unlikely to be differentially modified. So, xpore-diffmod w
    # those significant positions by the statistical test -- usually the P_VALUE_THRESHOLD ver
    # If you want xPore to test every genomic/transcriptomic position, please remove this pref
    prefiltering:
        method: t-test
        threshold: <P_VALUE_THRESHOLD>

    # Here are the parameters for Bayesian inference. The default values shown in <> are used,
    max_iters: <500>
    stopping_criteria: <0.00001>
```

Python Packaging




```python
"""Setup for the xpore package."""

from setuptools import setup,find_packages

__pkg_name__ = 'xpore'


with open('README.md') as f:
    README = f.read()

setup(
    author="Ploy N. Pratanwanich",
    maintainer_email="naruemon.p@chula.ac.th",
    name=__pkg_name__,
    license="MIT",
    description='xpore is a python package for Nanopore data analysis of differential RNA modifications.',
    version='v2.1',
    long_description=README,
    long_description_content_type='text/markdown',
    url='https://github.com/GoekeLab/xpore',
    packages=find_packages(),
    include_package_data=True,
    install_requires=[
            'numpy>=1.18.0',
            'pandas>=0.25.3',
            'scipy>=1.4.1',
            'PyYAML',
            'h5py>=2.10.0',
            'pyensembl>=1.8.5',
            'ujson>=4.0.1'
            ],
    python_requires=">=3.8",
    entry_points={'console_scripts': ["xpore={}.scripts.xpore:main".format(__pkg_name__)]},
    classifiers=[
        # Trove classifiers
        # (https://pypi.python.org/pypi?%3Aaction=list_classifiers)
        'Development Status :: 1 - Planning',
        'License :: OSI Approved :: MIT License',
        'Programming Language :: Python',
        'Programming Language :: Python :: 3.8',
        'Topic :: Software Development :: Libraries',
        'Topic :: Scientific/Engineering :: Bio-Informatics',
        'Intended Audience :: Science/Research',
    ],
)
```

Parallelization / Multiprocessing

```
import multiprocessing
```

When Data are Too Big to Fit in the Memory

ถ้าเราสามารถอ่านของ file → access file ตรงๆได้

data.index

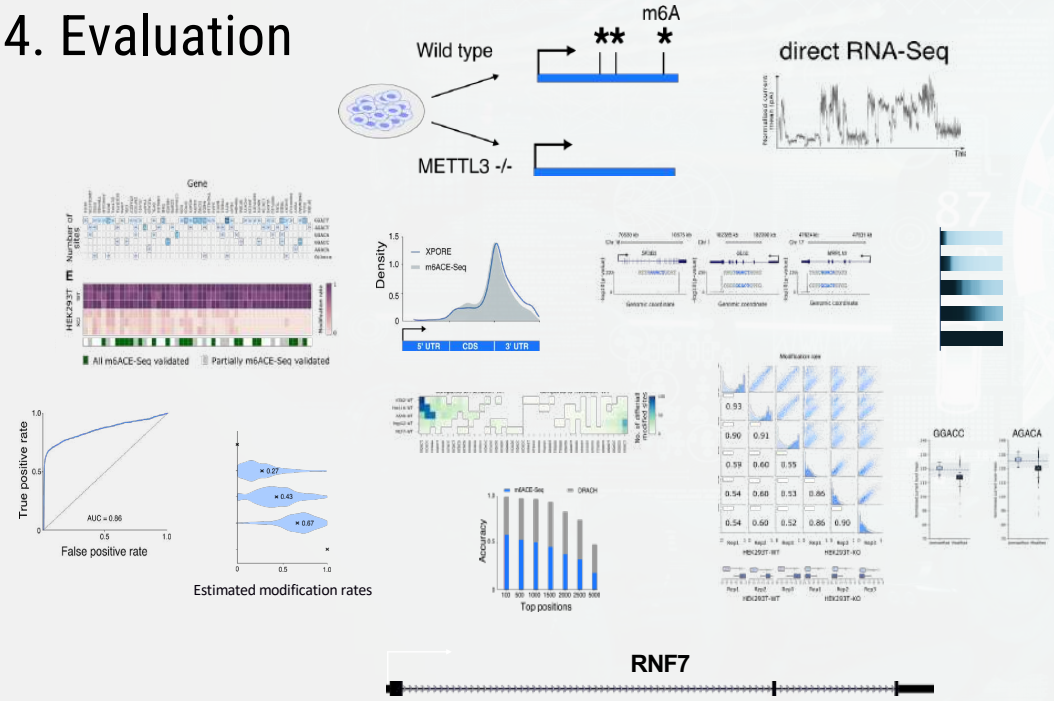
gene_id	start_idx	stop_idx
ENGxx1	0	16856
ENGxx2	16857	29435
…	…	…

data.json

{'ENGxx1': [123,110,…]}, {'ENGxx1': [123,110,…]}, {…}

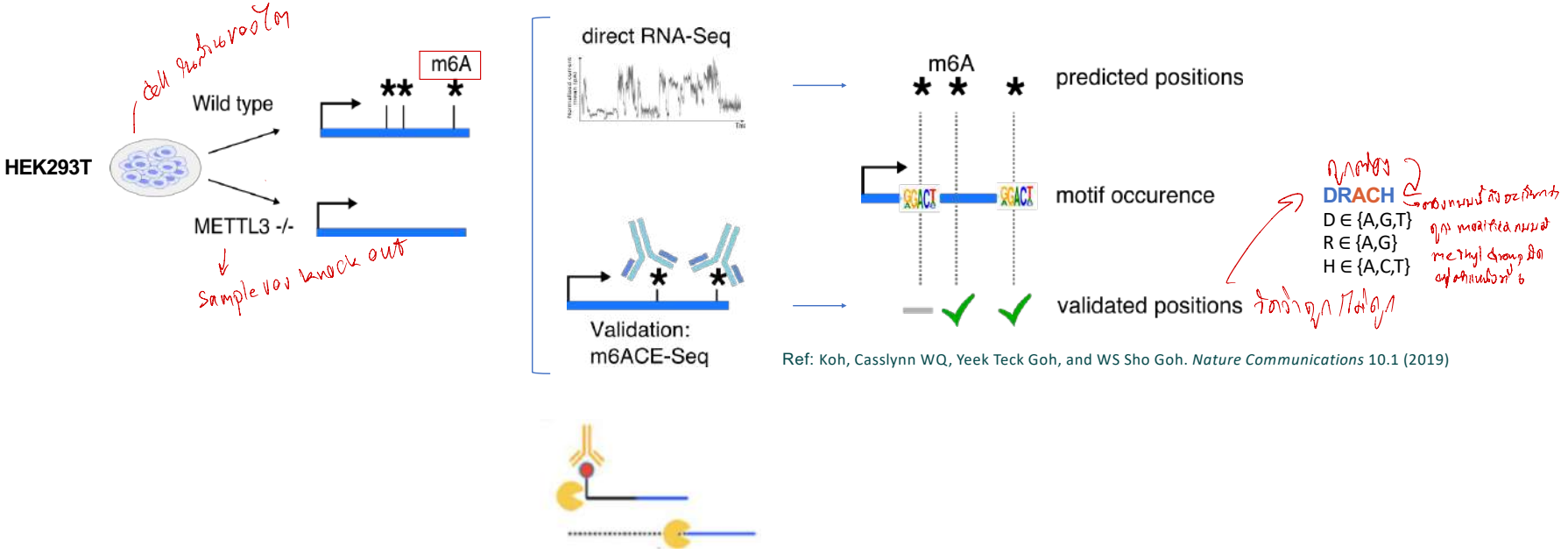
ใช้ผลลัพธ์จาก model มาวัดว่าดี Evaluate อย่างไร

4. Evaluation



- **Experiment setup**
- **Validation**
- **Applicability**
- **Discovery**

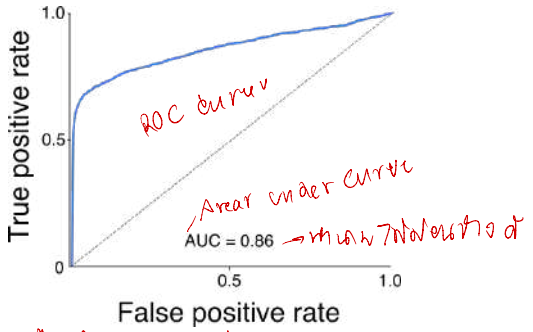
Experiment Setup



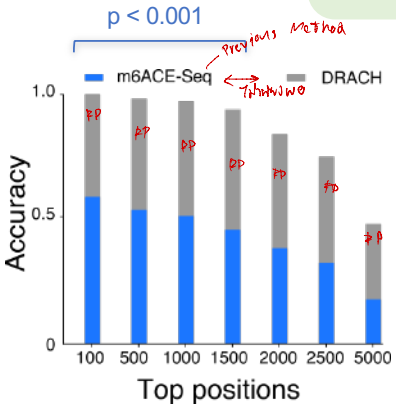
Ref: Koh, Casslynn WQ, Yeek Teck Goh, and WS Sho Goh. *Nature Communications* 10.1 (2019)

Validation: m6A calling

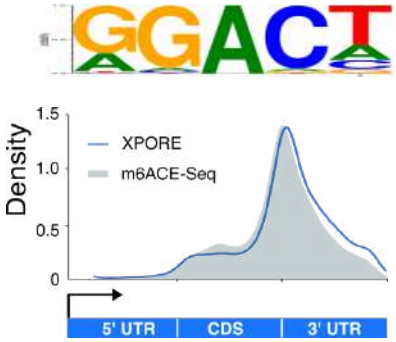
In bioinformatics,
- as the **labels are incomplete**,
- the **false positives may not be wrong**.
- **Analysis on the predictions** is usually required to **get more insights**.



- ~1 million sites were tested.
- xPore achieves AUCROC of 86% to call differentially m6A sites.



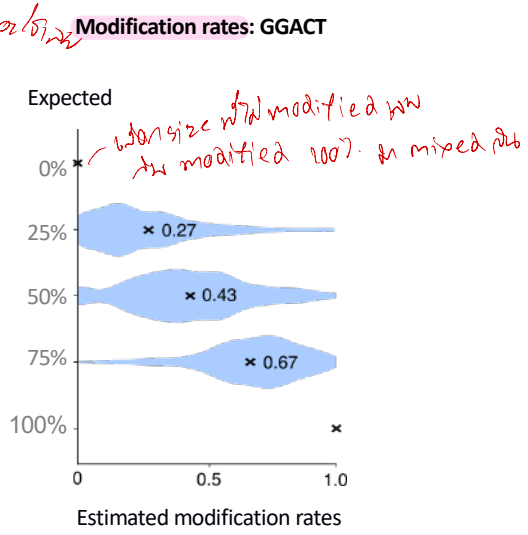
- Around half were identified by m6ACE-Seq.
- With m6ACE-Seq + DRACH, the accuracy is up to >95%.
- dRNA-Seq helps identify a different set of modified sites that had been otherwise missed.



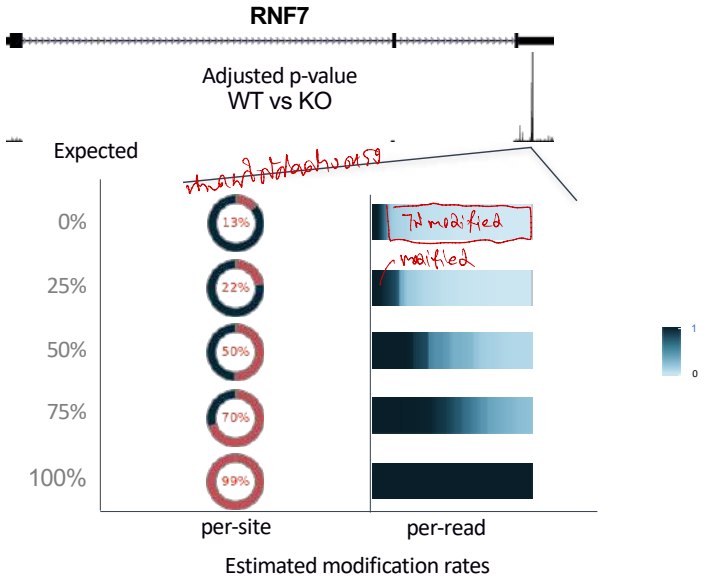
- m6A motifs e.g. GGACT, AGACT are confirmed.
- xPore can facilitate motif discovery in any other pairwise comparison.
- The differentially modified sites are also enriched at stop codons.

Validation: m6A stoichiometry quantification

- xPore models all RNA mixture samples at once.



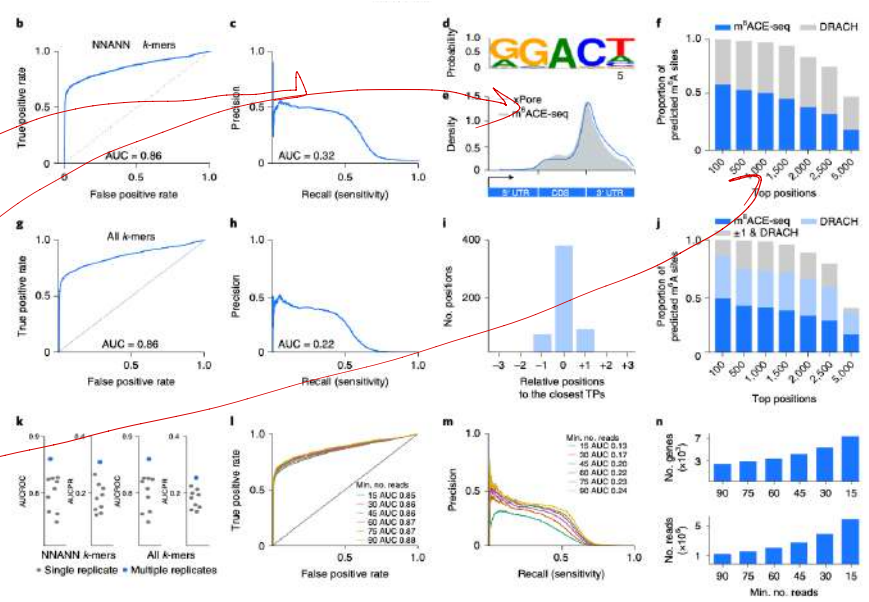
- Estimated modification rates closely match to the expected.



- Modification rates estimated by xPore can be interpretable as fractions of modified reads in a cell.
- This allows the analysis of differential modifications.

Validation: ML Metrics & Result Analysis

Fig. 2 | Detection of m6A sites in the human transcriptome.



ML Metrics

- ROC Curve
- Precision-Recall Curve
- Accuracy

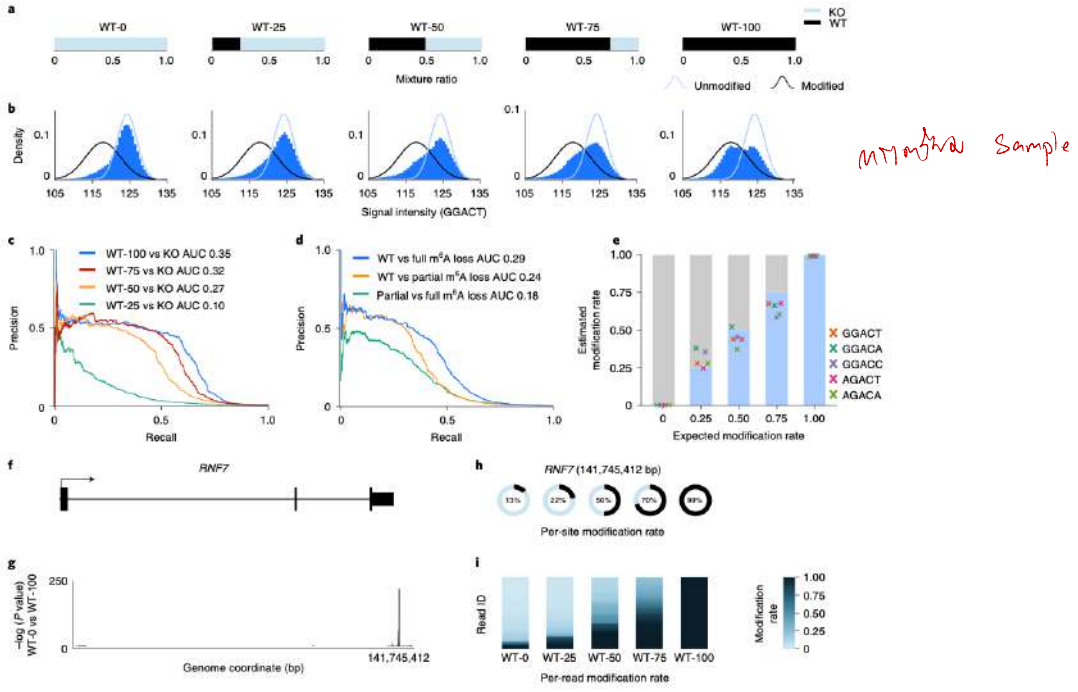
Analysis

- Domain-specific evaluation
- Effects of the data size

Validation: ML Metrics & Result Analysis

Aims 2nd research Objective

Fig. 3 | xPore modification-rate estimates correspond to the fraction of modified RNA species in the cell



Mutation Sample

ML Metrics

- ROC Curve
- Precision-Recall Curve
- Accuracy

Analysis

- Domain-specific evaluation
- Effects of the data size

Applicability: Full Dataset

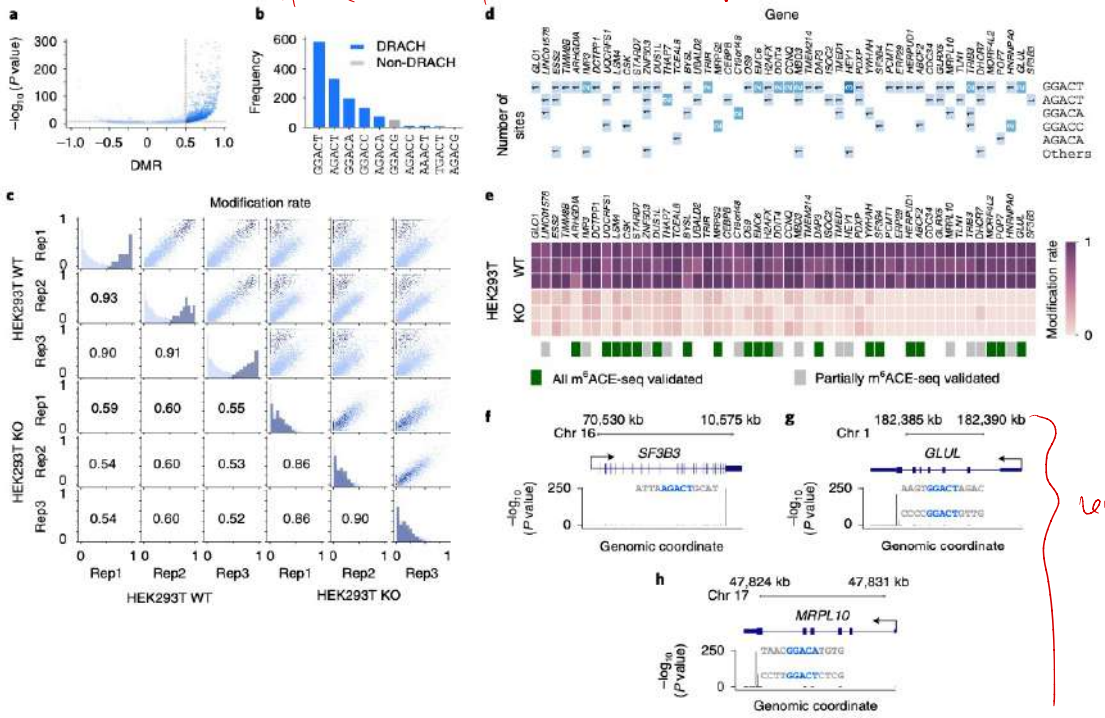
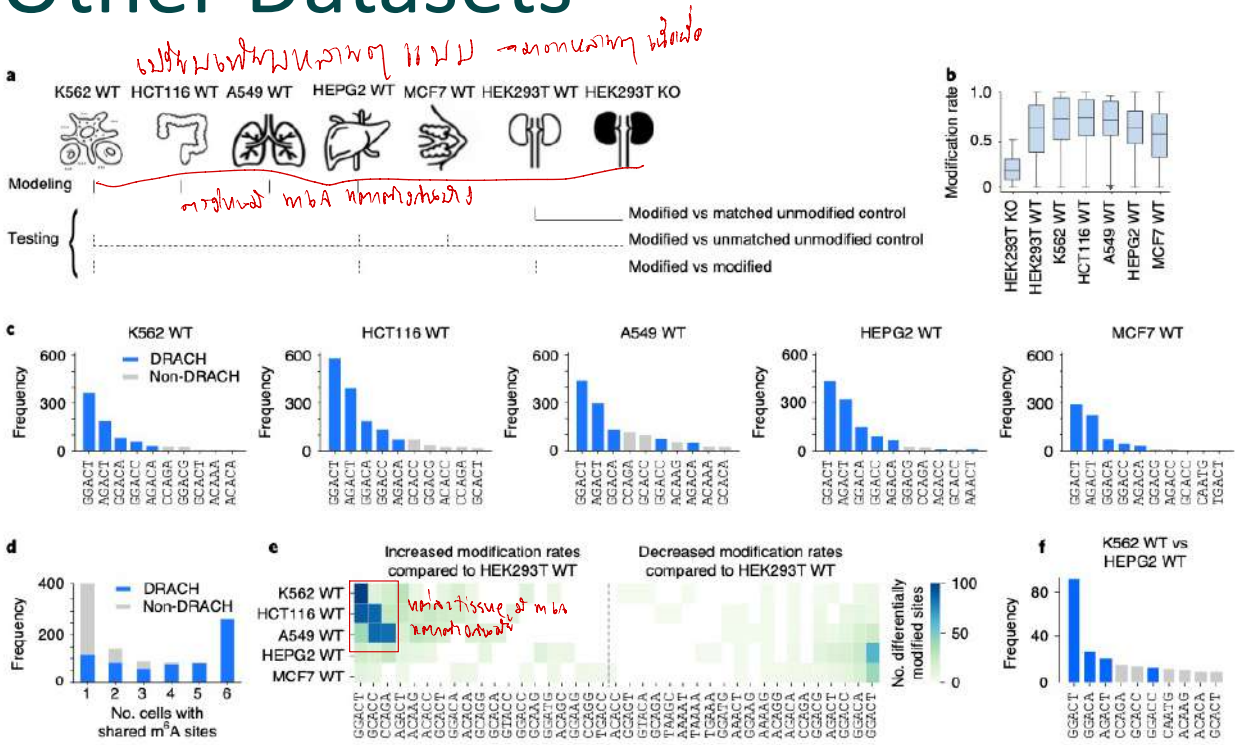


Fig. 4 | Transcriptome-wide identification of differentially modified positions.

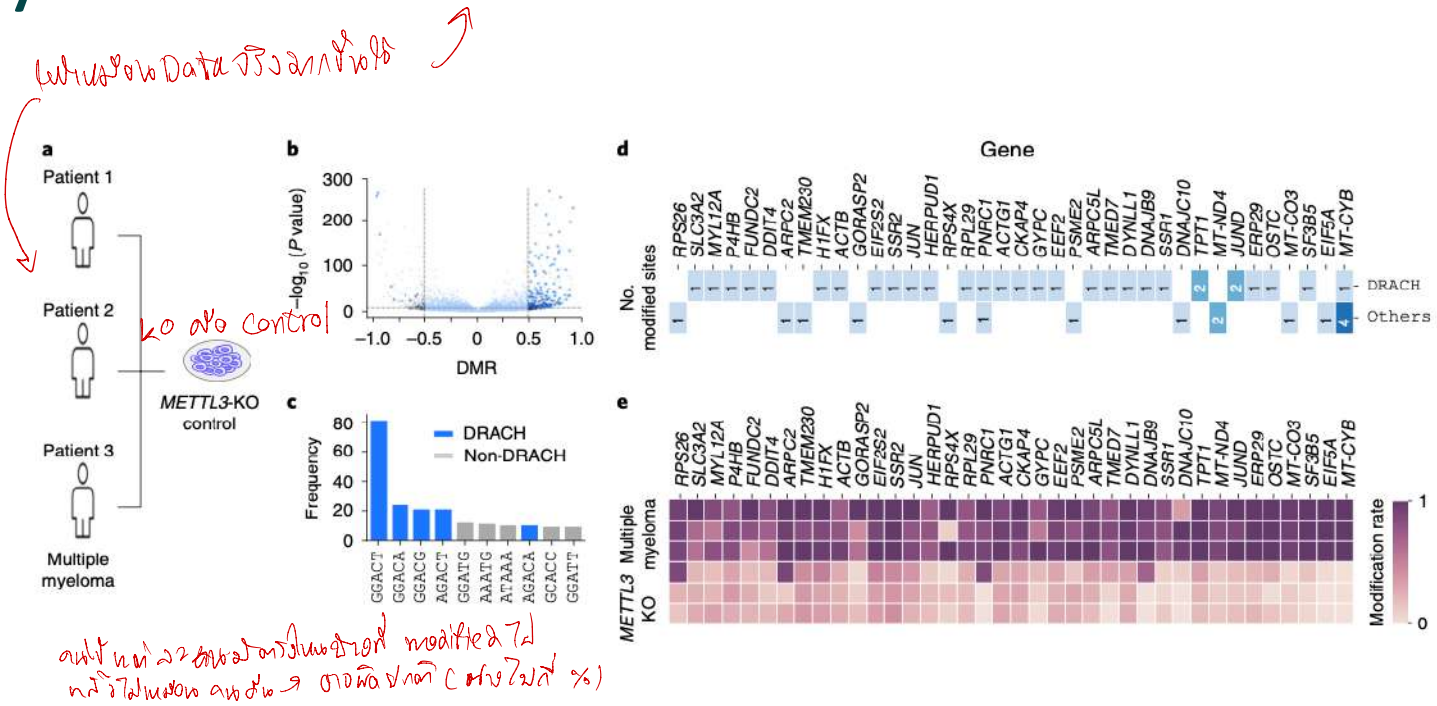
Applicability: Other Datasets

Fig. 5 | Identification of m6A sites across different tissues and cell lines.



Applicability: Clinical Data

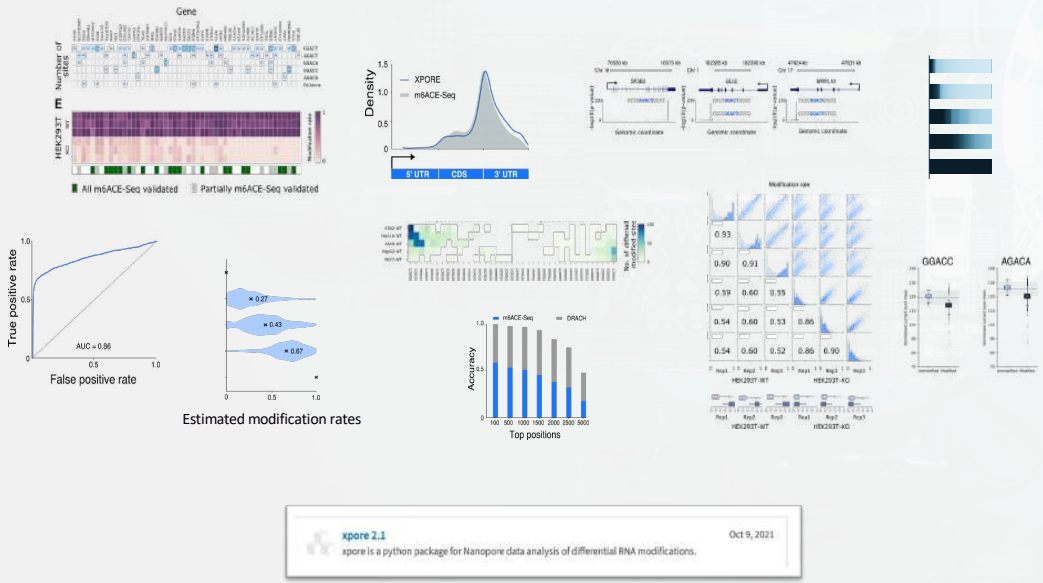
Fig. 6 | Identification of m6A in clinical samples using direct RNA-seq.



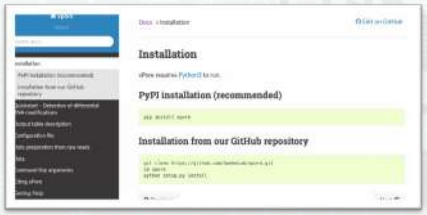
Evaluation: Keys Takeaway

- Validation
  - Using appropriate ML metrics
  - Analyzing the results to get more insights
- Comparison with other state-of-the-art methods
- Applicability
  - External / Other data
  - Human evaluation
  - Discovery

5. Visualization and Presentation



- Storylining
- Choosing the Right Plots
- Source Code
- Online Documentation



Storylining

Method overview

- **Fig. 1 | Schematic workflow: quantification of RNA modifications from direct RNA-seq data using xPore**

Validation

- **Fig. 2 | Detection of m6A sites in the human transcriptome.**
  - xPore: identification of differential RNA modifications.
  - xPore identifies m6A sites at single-base resolution.
  - Replicates increase precision.
  - Pooling data increases sensitivity.
- **Fig. 3 | xPore modification-rate estimates correspond to the fraction of modified RNA species in the cell**
  - xPore identifies modified positions with low stoichiometry.
  - Quantitative estimation of RNA-modification rates.

Applicability & Discovery

- **Fig. 4 | Transcriptome-wide identification of differentially modified positions.**
  - DMRs as estimates of effect size.
- **Fig. 5 | Identification of m6A sites across different tissues and cell lines.**
  - Identification of m6A across genetically diverse cell lines.
  - Variation of m6A across different cell lines.
- **Fig. 6 | Identification of m6A in clinical samples using direct RNA-seq.**
  - Identification of m6A in clinical cancer samples.

Choosing the Right Plots

Fig. 1 | Schematic workflow: quantification of RNA modifications from direct RNA-seq data using xPore

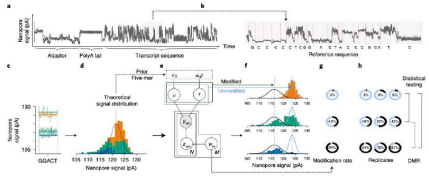


Fig. 2 | Detection of m6A sites in the human transcriptome.

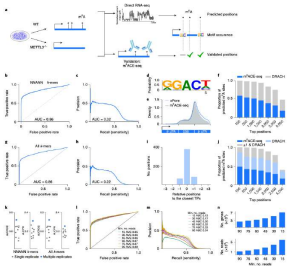


Fig. 3 | xPore modification-rate estimates correspond to the fraction of modified RNA species in the cell

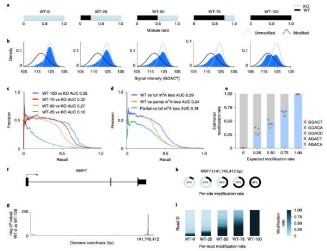


Fig. 4 | Transcriptome-wide identification of differentially modified positions.

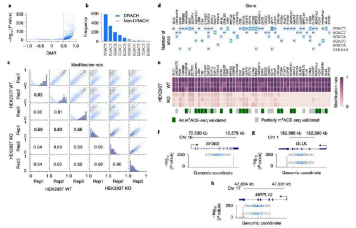


Fig. 5 | Identification of m6A sites across different tissues and cell lines.

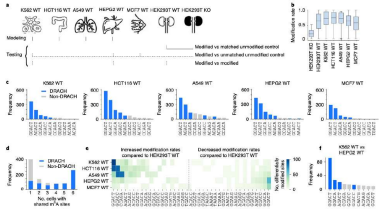
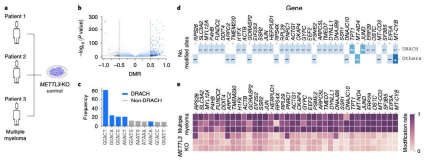


Fig. 6 | Identification of m6A in clinical samples using direct RNA-seq.



Choosing the Right Plots

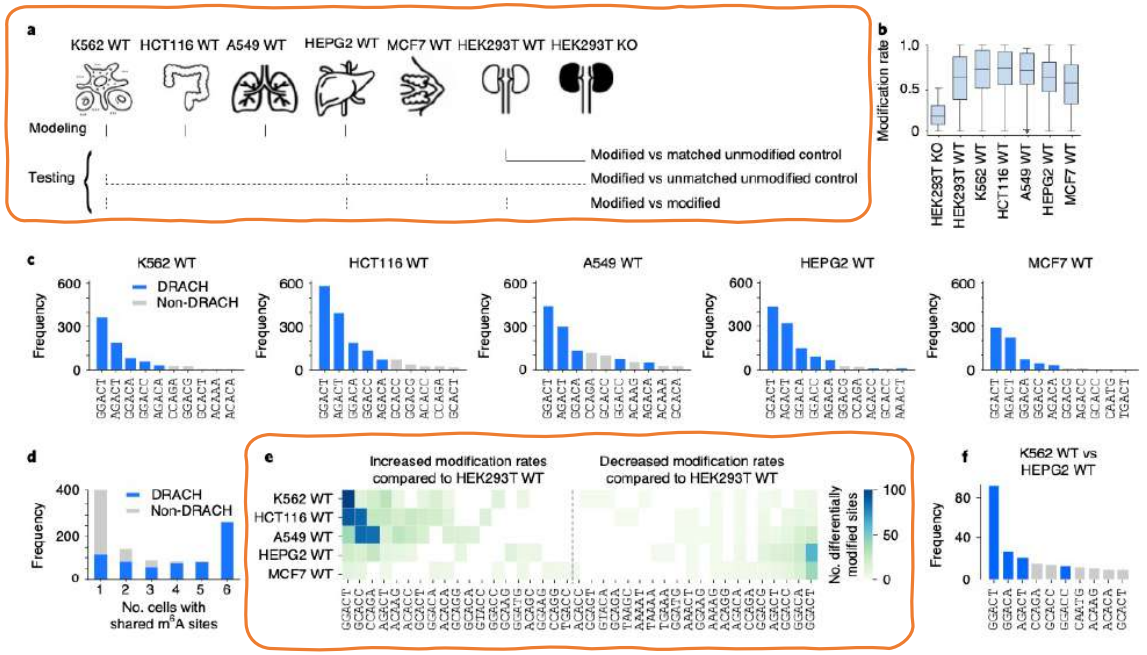


Fig. 5 | Identification of m6A sites across different tissues and cell lines.

3 Key Success to Develop AI-Powered Apps

1. Alignment with the actual needs
2. Sufficient generalization and evaluation
3. Simple deployment and serving
   - Online documentation
   - Easy installation
   - Source code
   - Data availability
   - Lightweight
   - Fast

https://github.com/GoekeLab/xpore

ARTICLES nature biotechnology

Identification of differential RNA modifications from nanopore direct RNA sequencing with xPore

Ploy N. Pratanwanich, Fei Yao, Ying Chen, Casslynn W. Q. Koh, Yuk Kei Wan, Christopher Hendra, Polly Poon, Yeek Teck Goh, Phoebe M. L. Yap, Jing Yuan Chooi, Wee Joo Chng, Sarah B. Ng, Alexandre Thiery, W. S. Sho Goh and Jonathan Göke

Scopus metrics

78 99th percentile
Citations in Scopus

9.61
Field-Weighted citation impact

downloads 27k

xpore.readthedocs.io/

python machine-learning rna-seq nanopore genomics rna transcriptomics modification nanopore-sequencing rna-modifications

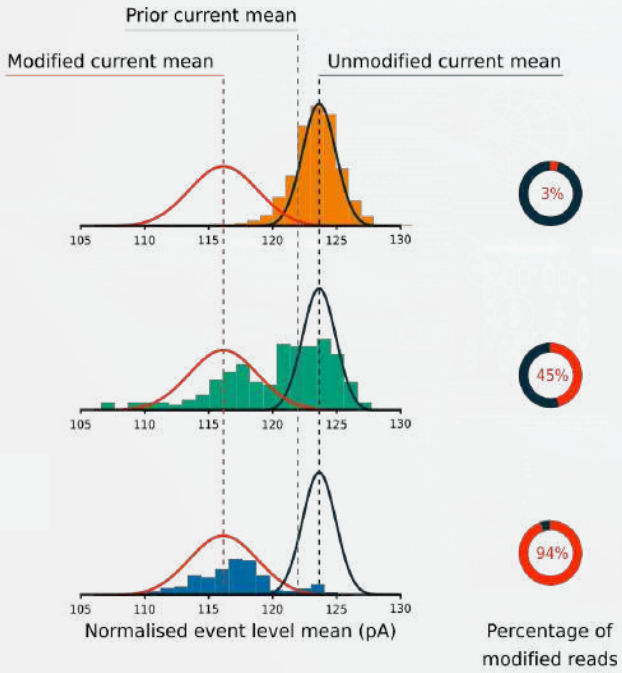
Readme
MIT license
Activity
121 stars
9 watching
22 forks
Report repository

Releases 9

xPore v2.1 Latest
on Oct 9, 2021

+ 8 releases

6. Future Work



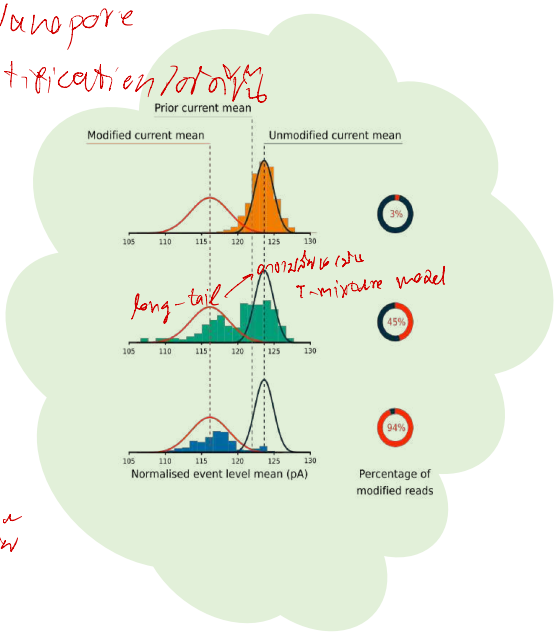
- **Identifying the Limitations**
- **Considering Changes in the Future**

Future Work

Domain-Oriented

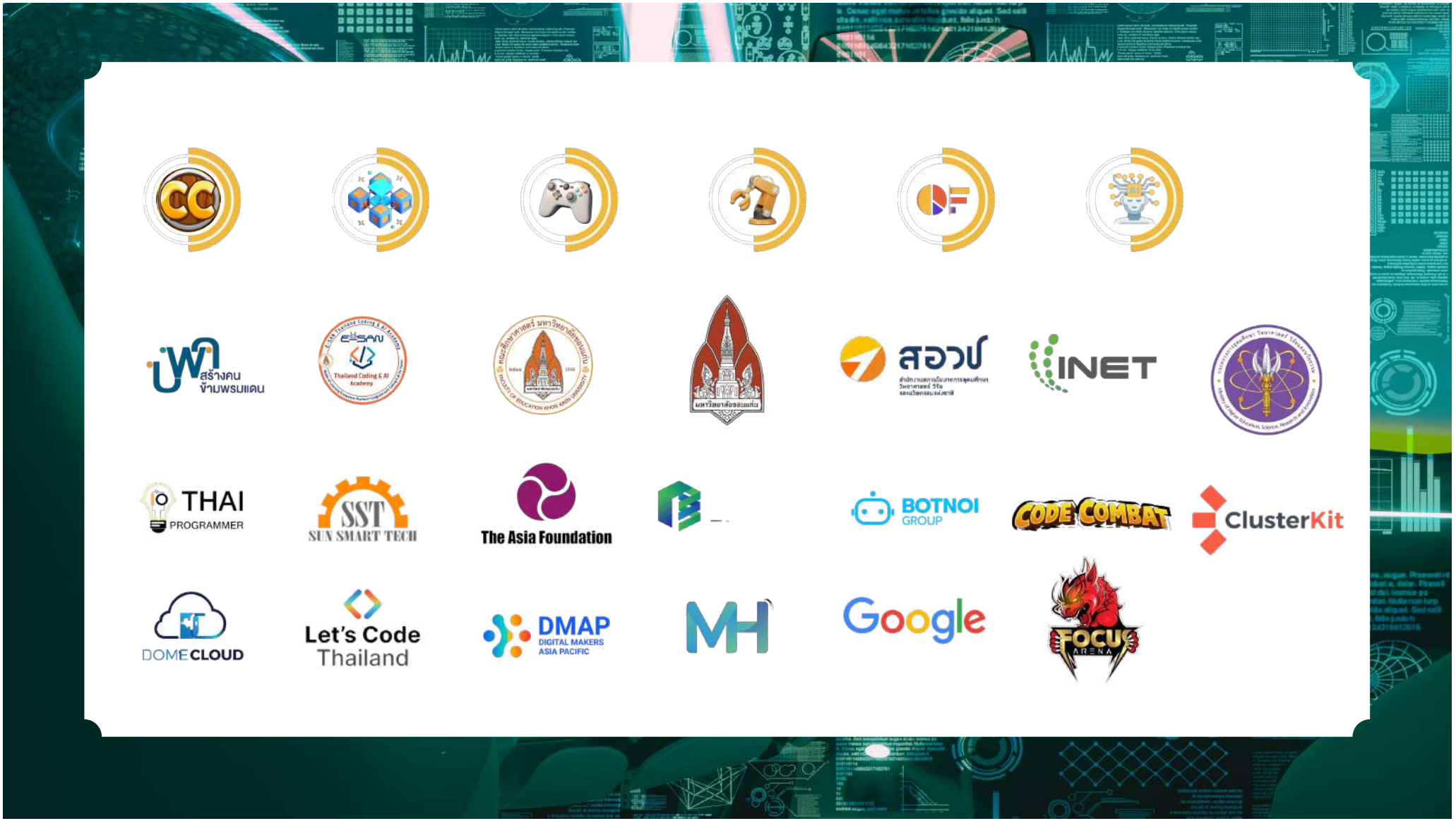
- m6anet
- <Gaussian> mixture model
- Interpretability
  - Modification or basecalled errors
- End-to-end
  - Why?

  → Nanopolish eventalign / Guppy basecaller are subjet to change



Method-Oriented

- Deep autoencoder + GMM
- CNN + GMM
- **Other models + GMM**


สร้างคน
ข้ามพรมแดน
Thailand Coding & AI Academy
สอวช
INET
THAI
PROGRAMMER
SST
SUN SMART TECH
The Asia Foundation
BOTNOI
GROUP
CODE COMBAT
ClusterKit
DOMECLOUD
Let's Code
Thailand
DMAP
DIGITAL MAKERS
ASIA PACIFIC
Google
FOCUS
ARENA