1

2

Mapping of RNA modi cations by direct Nanopore

sequencing and JACUSA2

* Amina Lemsara1,2, Christoph Dieterich 1,2,3, and Isabel Naarmann-de Vries1,2,3
* 1Klaus Tschira Institute for Integrative Computational Cardiology, University

5 Heidelberg, 69120 Heidelberg, Germany

* 2Department of Internal Medicine III (Cardiology, Angiology, and Pneumology),
* University Hospital Heidelberg, 69120 Heidelberg, Germany
* 3German Centre for Cardiovascular Research (DZHK)-Partner Site

9 Heidelberg/Mannheim, 69120 Heidelberg, Germany

10

11

12

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

Abstract

RNA modi cations exist in all kingdom of life. Several di erent types of base or ribose modi cations are now summarized under the term the "epitranscriptome". With the advent of high-throughput se-quencing technologies much progress has been made in understanding RNA modi cation biology and how these modi cations can in uence many aspects of RNA life. The most widespread internal modi cation on mRNA is m6A, which has been implicated in physiological pro-cesses as well as disease pathogenesis. Here, we provide a work ow for the mapping of m6A sites using Nanopore direct RNA sequencing data. Our strategy employs pairwise comparison of base calling error pro les with JACUSA2. We outline a general strategy for RNA modi cation detection on mRNA and describe two speci c use cases on m6A de-tection in detail. Use case 1: a sample of interest with modi cations (e.g. "wild type" sample) is compared to a sample lacking a speci c modi cation type (e.g. "knock out" sample, here METTL3 -KO) or Use case 2: a sample of interest with modi cations is compared to a sample lacking all modi cations (e.g. in vitro transcribed cDNA). We provide a detailed protocol on experimental and computational aspects. Extensive online material provides a snakemake pipeline to identify m6A positions in mRNA and to validate the results against a miCLIP-derived m6A reference set. The general strategy is exible and can be easily adapted by users in di erent application scenarios.

Correspondence to: christoph.dieterich@uni-heidelberg.de

1

1. INTRODUCTION
2. Chemical modi cations on DNA and histones, also known as epigenetics
3. marks, strongly impact gene expression during cell di erentiation and in
4. several other biological programs. In the 1970s, it was recognized that RNA
5. is also subjected to extensive covalent modi cation, and studies in the late
6. 1980s revealed the widespread deamination of bases (termed RNA editing),
7. which can lead to recoding if it occurs within coding sequences. Impres-
8. sive development in the RNA modi cation eld occurred during the past
9. eight years, with the discovery of an extensive layer of base modi cations
10. in mRNAs. These can in uence gene expression and have been already
11. shown to be involved in primary cellular programs such as stem cell di er-
12. entiation, response to stress, and the circadian clock. The study of RNA
13. modi cations and their e ects is now referred to as epitranscriptomics, and
14. it reveals striking similarities to what is known for epigenomics. To date
15. thirteen distinct modi cations have been identi ed on mRNA transcripts
16. [Anreiter et al., 2021]. These modi cations are catalyzed by a variety of
17. dedicated enzymes and can be divided into two classes: modi cations of
18. cap-adjacent nucleotides and internal modi cations.
19. In contrast to the m7G cap, the impact of internal modi cations on gene
20. regulation has been less studied apart from RNA editing, which is mediated
21. by RNA deaminases (e.g. the ADAR family). The most widespread in-
22. ternal mRNA modi cation is N6-methyladenosine (m6A). By modulating
23. the processing of mRNA, m6A can regulate a wide range of physiological
24. processes and its alteration has been linked to several diseases Roignant
25. and Soller [2017], Zaccara et al. [2019], Shi et al. [2019]. The modi cation is
26. catalyzed co-transcriptionally by a Mega-Dalton methyltransferase complex,
27. which includes the heterodimer METTL3-METTL14 and other associated
28. subunits Garcias Morales and Reyes [2021]. This modi cation is reversible
29. since two proteins of the AlkB-family of demethylases can remove m6A from
30. mRNA transcripts [Jia et al., 2011, Zheng et al., 2013]. In mammals, m6A
31. preferentially localizes within long internal exons and at the beginning of
32. terminal exons at so-called DRACH motif (D = A/G/U, R = A/G, H =
33. A/C/U) sites [Dominissini et al., 2012, Meyer et al., 2012, Ke et al., 2015].
34. Once deposited, m6A is recognized by several reader proteins that can af-
35. fect the fate of mRNA transcripts in nearly every step of the mRNA life
36. cycle, including alternative splicing [Adhikari et al., 2016, Roundtree et al.,
37. 2017], mRNA translation [Wang et al., 2015] and decay [Wang et al., 2014,
38. Du et al., 2016, Roundtree et al., 2017]. The best-described readers are the
39. YTH domain family of proteins that decode the signal and mediate m6A
40. functions. By a ecting RNA structure, m6A can also indirectly in uence
41. the association of additional RNA-binding proteins (RBPs) and the assem-
42. bly of larger messenger ribonucleoprotein particles (mRNPs) [Patil et al.,
43. 2018].

2

1. Several approaches have been presented to map RNA modi cations on
2. RNA. Herein, we focus on mRNA modi cation site detection in general and
3. on m6A in particular where antibody-based protocols (miCLIP), methylation-
4. sensitive restriction enzyme assays (MazF) or transgenic approaches (TRIBE,
5. DART) have been presented to map m6A sites. All of the aforementioned
6. approaches rely on high-throughput short read sequencing on the Illumina
7. platform. This typically involves cDNA synthesis by reverse transcription
8. and PCR-based library ampli cation. One recent addition to the toolbox of
9. RNA modi cation mapping is direct RNA single molecule long read sequenc-
10. ing on the Oxford Nanopore Technologies platform (dRNA-seq). While our
11. software is able to deal with Illumina and Nanopore-based approaches, the
12. latter is the principal topic of this methods article.
13. MATERIALS
14. ONT direct RNA sequencing
15. This section summarizes all necessary consumables for direct RNA sequenc-
16. ing of poly-adenylated RNA (i.e. mRNA) on the MinION or similar device.
17. 1. 500 ng polyA+ RNA isolated from total RNA e.g. with Oligotex
18. mRNA kit (#70022, Qiagen) or Dynabeads oligo dT25 beads (#61002,
19. Thermo Fisher Scienti c) or in vitro transcriptome sample. Store RNA
20. at -80 C and the mRNA puri cation kit as recommended by the man-
21. ufacturer.
22. 2. Nuclease-free water. Store at room temperature.
23. 3. Direct RNA-sequencing kit (SQK-RNA002, Oxford Nanopore Tech-
24. nologies). Store at -20 C.
25. 4. NEBNext Quick Ligation Reaction Bu er (#B6058S, New England
26. Biolabs). Store at -20 C.
27. 5. T4 DNA Ligase (#M0202S, New England Biolabs). Store at -20 C.
28. 6. dNTP Mix (10 mM each, #R0191, Thermo Fisher Scienti c). Store
29. at -20 C.
30. 7. SuperScript IV Reverse Transcriptase (#18090010, Thermo Fisher Sci-
31. enti c). Store at -20 C.
32. 8. Agencourt RNAClean XP beads (#A63987, Beckman Coulter). Store
33. at 4 C.
34. 9. 70 % ethanol, freshly prepared.

3

1. 10. Qubit dsDNA HS assay kit (#Q32854) and Qubit Fluorometer (Thermo
2. Fisher Scienti c).
3. 11. Flow cell priming kit (EXP-FLP002, Oxford Nanopore Technologies).
4. Store at -20 C.
5. 12. Thermocycler.
6. 13. Gentle rotator mixer.
7. 14. Magnetic stand for 1.5 ml tubes.
8. 15. 1.5 ml DNA LoBind tubes (Eppendorf), 0.2 ml PCR tubes.
9. 16. MinION or GridION sequencing device and MinION R9.4.1 Flow cells
10. (FLO-MIN106D, Oxford Nanopore Technologies). Store Flow cells at
11. 4 C.
12. Preparation of an in vitro transcriptome sample
13. 1. 100 ng polyA+ RNA isolated from total RNA e.g. with Oligotex
14. mRNA kit (#70022, Qiagen) or Dynabeads oligo dT25 beads (#61002,
15. Thermo Fisher Scienti c). Store RNA at -80 C and the mRNA pu-
16. ri cation kit as recommended by the manufacturer
17. 2. 10 M oligo(dT)-VN RT primer.
18. TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN. Store at -20 C.
19. 3. 20 M template switching oligo (TSO). ACTCTAATACGACTCAC-
20. TATAGGGAGAGGGCrGrG+G. Store at -20 C.
21. 4. 10 M T7 extension primer. GCTCTAATACGACTCACTATAGG.
22. Store at -20 C.
23. 5. Nuclease-free water. Store at room temperature.
24. 6. dNTP Mix (10 mM each, #R0191, Thermo Fisher Scienti c). Store
25. at -20 C.
26. 7. Template Switching RT Enzyme Mix (#M0466S, New England Bio-
27. labs). Store at -20 C.
28. 8. Q5 Hot Start High-Fidelity 2X Master Mix (#M0494S, New England
29. Biolabs). Store at -20 C.
30. 9. RNase H (5,000 U/ml) (#M0297S, New England Biolabs). Store at
31. -20 C.

4

1. 10. NucleoSpin Gel and PCR Clean-up, Mini kit for gel extraction and
2. PCR clean up (#740609.50, Macherey-Nagel) or equivalent. Store at
3. room temperature.
4. 11. MEGAscript T7 transcription kit (#AM1334, Thermo Fisher Scien-
5. ti c). Store at -20 C.
6. 12. RNA Clean & Concentrator-25 kit (#R1017, Zymo Research). Store
7. at room temperature.
8. 13. Thermocycler.
9. 14. Table top centrifuge for 1.5 ml tubes.
10. 15. Nanodrop spectrophotometer or equivalent.
11. 16. 0.2 ml PCR tubes, 1.5 ml DNA LoBind tubes (Eppendorf).
12. Hardware requirements
13. All analyses have been performed/tested on two alternative hardware sys-
14. tems: a standard Linux desktop computer or an Apple iMac (Retina 5K,
15. ultimo 2014). The work ow requires a multi-core processor system with
16. minimal main memory of 16GB RAM and several GBs of free disk space
17. (depending on data set size).
18. Software dependencies and installation
19. Our analysis work ow has few requirements, which are detailed in Table 1.
20. Speci cally, to execute our work ow, the following prerequisites are neces-
21. sary: a BASH shell, a JAVA runtime environment, a working PERL and
22. R installation. Additional i.e. non-standard software to process and map
23. Nanopore reads (bedtools, samtools and Minimap2) are obligatory. Ta-
24. ble 2 lists some additional R packages, which are required to run the R
25. code. Detailed installation instructions and corresponding work ow code
26. are deposited under [https://github.com/dieterich-lab/MiMB\_JACUSA2\_](https://github.com/dieterich-lab/MiMB_JACUSA2_chapter)
27. [chapter](https://github.com/dieterich-lab/MiMB_JACUSA2_chapter).
28. METHODS
29. Our work ow is based on the pairwise comparison of samples with di er-
30. ent modi cation status (Figure 1). The sample of interest (yellow) may be
31. compared to di erent samples lacking certain modi cations. If available,
32. the wild type (WT) sample can be compared to a knock out (KO) sample
33. lacking speci c enzymatic activities (green), as outlined in Use Case 1. Al-
34. ternatively, a sample lacking all modi cations may be used for comparison

5

1. (blue). This may be either a simulated sample (i.e. with NanoSim) or an in
2. vitro transcribed sample derived from cDNA. Such an analysis is detailed in
3. Use Case 2. In any setting, JACUSA2 calculates scores for the Mismatch,
4. Insertion and Deletion rates of the pairwise comparisons as outlined above
5. (Figure 1, right).
6. One feature of Nanopore sequencing is to read sequences as 5-mers, as
7. always ve nucleotides are occupied by the pore protein (Figure 2). Because
8. of this, a m6A modi cation may a ect basecalling not only if the modi ed
9. nucleotide is in the central position, but also at neighboring positions (-2
10. to +2). To account for this, JACUSA2 scores for Deletion, Mismatch and
11. Insertion are calculated for the entire 5-mer context. Depending on the
12. modi cation-speci c signature, a Feature set can be selected to calculate
13. the nal JACUSA2 score (Figure 2).
14. Our work ow can be divided into a wet-lab part (Figure 3A) and a
15. computational part (Figure 3B). Starting from total cellular RNA, polyA+
16. RNA is isolated and subjected to Nanopore direct RNA-sequencing. Guppy
17. basecalling can be done as well as live basecalling during sequencing on the
18. respective FAST5 les, which results in FASTQ output les (Figure 3A).
19. FASTQ les are aligned to a reference sequence with Minimap2. SAMtools
20. is used to generate BAM les as input for JACUSA2 analysis, which yields
21. candidate m6A sites with the presented work ow in this chapter (Figure
22. 3B). We will present all necessary experimental step for dRNA-seq in the
23. next section.
24. Nanopore direct RNA sequencing
25. 1. Adjust 500 ng polyA+ RNA to a total volume of 9 l with nuclease-
26. free water. Complete RT adapter ligation reaction (in 0.2 ml PCR
27. tube) with 3 l NEBNext Quick Ligation Reaction Bu er, 0.5 l
28. RNA CS (RCS, from SQK-RNA002), 1 l RT-Adapter (RTA, from
29. SQK-RNA002) and 1.5 l T4 DNA Ligase. Incubate 10 min at room
30. temperature.
31. 2. Prepare reverse transcription master mix on ice during ligation: 9 l
32. nuclease-free water, 2 l 10 mM dNTPs, 8 l 5x SuperScript IV rst
33. strand bu er, 4 l 0.1 mM DTT.
34. 3. Add the reverse transcription master mix to the ligation reaction and
35. mix by pipetting. Add 2 l SuperScript IV reverse transcriptase and
36. mix by pipetting. Incubate in a thermocycler with the following pro-
37. tocol: 50 min at 50 C, 10 min at 70 C, cool down to 4 C.
38. 4. Let the Agencourt RNAClean XP beads come to room temperature
39. during reverse transcription. Carefully resuspend beads before use.
40. Transfer reaction to a 1.5 ml DNA LoBind tube and mix with 72 l

6

215

216

Agencourt RNAClean XP beads. Incubate 5 min at room temperature on a gentle rotator mixer.

1. 5. Collect beads on a magnetic stand and remove supernatant. Wash
2. pelleted beads two times (30 sec) with 200 l freshly prepared 70 %
3. ethanol. Remove supernatant. Spin sample down and place on magnet
4. again. Remove any residual ethanol.
5. 6. Resuspend beads in 20 l nuclease-free water by gentle icking and
6. incubate 5 min at room temperature on a gentle rotator mixer. Collect
7. beads on a magnetic stand and transfer 20 l eluate in a fresh 1.5 ml
8. DNA LoBind tube.
9. 7. For ligation of the RMX adapter, add the following to 20 l eluate: 8
10. l NEBNext Quick Ligation Reaction Bu er, 6 l RMX (from SQK-
11. RNA002), 3 l nuclease-free water, 3 l T4 DNA Ligase. Mix by
12. pipetting and incubate 10 min at room temperature.
13. 8. Add 40 l carefully resuspended Agencourt RNAClean XP beads to
14. the reaction and mix by pipetting. Incubate 5 min at room tempera-
15. ture on a gentle rotator mixer.
16. 9. Collect beads on a magnetic stand and remove supernatant. Wash
17. pelleted beads two times with 150 l wash bu er (WSB, from SQK-
18. RNA002). Resuspend beads by icking, spin down and return to mag-
19. netic stand. Remove supernatant from pelleted beads.
20. 10. Resuspend beads in 21 l elution bu er (EB, from SQK-RNA002) by
21. gentle icking and incubate 5 min at room temperature on a gentle
22. rotator mixer. Pellet beads on a magnetic stand and transfer 21 l
23. eluate in a fresh 1.5 ml DNA LoBind tube.
24. 11. Quantify 1 l of the library on a Qubit uorometer with the Qubit
25. dsDNA HS kit according to the manufacturerers protocol. Concentra-
26. tion should be usually in the range of 5 - 10 ng/ l.
27. 12. Insert MinION R9.4.1 Flow cell in the MinION or GridION sequenc-
28. ing device and perform Flow cell check in the MinKNOW software.
29. For successful sequencing of mammalian polyA+ RNA at least 1,000
30. available pores are recommended.
31. 13. Prepare Priming Mix by adding 30 l ush tether (FLT, from EXP-
32. FLP002) to a vial of ush bu er (FB, from EXP-FLP002) and mix by
33. pipetting. Open priming port. Remove air bubble from priming port
34. by inserting the tip of a P1000 pipette into the priming port and slowly
35. dialing up, until a small volume of storage bu er enters the pipette
36. tip. Load 800 l Priming Mix via the priming port and carefully avoid
37. introduction of air bubbles. Close the priming port and wait for 5 min.

7

1. 14. Mix 20 l library with 17.5 l nuclease-free water and 37.5 l RNA run-
2. ning bu er (RRB, from SQK-RNA002) and mix by pipetting. Open
3. the priming port and the sample port. Load 200 l Priming Mix via
4. the priming port. Mix library by pipetting just before loading and
5. load dropwise via the sample port. Carefully avoid introduction of air
6. bubbles. Close the sample port and the priming port.
7. 15. Start sequencing for 48 to 72 h in the MinKNOW software. Choose
8. direct RNA-sequencing kit and high-accuracy basecalling as parame-
9. ters.
10. Preparation of an in vitro transcriptome sample
11. The in vitro transcriptome sample is prepared based on a protocol published
12. by Zhang et al. [2021] with some modi cations a detailed below. An in vitro
13. transcriptome lacks any RNA modi cations and is a perfect reference sample
14. for RNA modi cation mining.
15. 1. Adjust 100 ng polyA+ RNA to a total volume of 6 l with nuclease-
16. free water. Add 1 l each of 10 M oligo(dT)-VN RT primer and 10
17. mM dNTPs. Mix by pipetting and incubate in a thermocycler: 5 min
18. at 75 C, 2 min at 42 C, cool to 4 C.
19. 2. Assemble 2.5 l 4x template switching RT bu er, 0.5 l 20 M TSO,
20. 1 l 10x template switching RT enzyme mix and mix by pipetting.
21. Combine with 6 l RNA and incubate in a thermocycler: 90 min at
22. 42 C, 10 min at 68 C, cool to 4 C.
23. 3. For Second strand synthesis add to First strand synthesis reaction: 50
24. l Q5 Hot Start High-Fidelity 2X Master Mix, 5 l RNase H, 2 l 10
25. M T7 extension primer, 33 l nuclease-free water. Mix by pipetting
26. and incubate in a thermocycler: 15 min at 37 C, 1 min at 95 C, 10
27. min at 65 C, cool to 4 C.
28. 4. Purify double stranded cDNA with NucleoSpin Gel and PCR Clean-up
29. kit according to the manufacturerers protocol and elute in 20 l elution
30. bu er. Determine concentration on a Nanodrop spectrophotometer.
31. cDNA may be stored at -20 C.
32. 5. Combine 8 l cDNA for in vitro transcription with 2 l each of ATP,
33. GTP, CTP, UTP, 10x reaction bu er and enzyme mix from the MEGAscript
34. T7 transcription kit. Incubate 3 h at 37 C.
35. 6. Digest template DNA by addition of 1 l Turbo DNase. Mix by pipet-
36. ting and incubate 15 min at 37 C.

8

1. 7. Adjust reaction volume to 100 l with nuclease-free water and clean up
2. with RNA Clean & Concentrator-25 kit according to the manufactur-
3. ers protocol, using two volumes of adjusted RNA binding bu er (1:1
4. RNA binding bu er : ethanol). Elute RNA in 25 l nuclease-free wa-
5. ter. Determine RNA concentration on a Nanodrop spectrophotometer.
6. Store at -80 C.
7. Nanopore read processing
8. 1. Base call the ionic current signal stored in FAST5 les using Guppy.
9. For the IVT sample, we applied real-time base calling with the MinKNOW-
10. embedded Guppy basecaller. Otherwise, Guppy basecaller software
11. can be used. In this case, the basecaller requires the path to FAST5
12. les, the output folder, and the con g le or the owcell/kit combina-
13. tion. The output are FASTQ les that can be compressed using the
14. option "{compress fastq".

304

305

306

307

308

309

* guppy\_basecaller --compress\_fastq -i path\_to\_fast5 -s path\_to\_output -c config\_file.cfg --cpu\_threads\_per\_caller 14 --num\_callers

1

Set the number of threads "cpu threads per caller" and the number of parallel basecallers "num caller" according to your resources. Ad-ditional details can be found at <https://nanoporetech.com/>.

1. 2. Align reads to the transcriptome using Minimap2 software. The out-
2. put is a SAM le that has to be converted to a compressed form as
3. BAM le using SAMtools command. The alignment requires a refer-
4. ence sequence. Here, we used GRCh38 Ensembl release 96 annotation
5. and FASTA le. Pre-indexing of the human genome saves time dur-
6. ing read alignment. Please save the index with the option "-d" before
7. read mapping and use the index instead of the reference le in the
8. minimap2 command line.

318

319

320

321

322

323

324

325

326

327

328

$ minimap2 -d reference.mmi reference.fa

For Direct RNA Sequencing, it is recommended to set a small k-mer size "-k [=14]" to enhance sensitivity. We recommend outputting only primary alignments "{secondary=no". Use the parameter ’{MD’ to add the reference sequence information to the alignment; this is neces-sary for JACUSA2 downstream analysis. Adjust the number of threads "-t" according to your resources. Check Minimap2 manual for more details [Min]. To enable spliced alignments, use the setting -ax splice {junc-bed annotation.bed {junc-bonus where "{junc-bonus" allows to tune the bonus score and the BED le "{junc-bed annotation.bed" provides the splice junctions.

9

329

330

331

332

333

334

* minimap2 -t 5 --MD -ax splice --junc-bonus 1 -k14 --secondary=no --junc-bed final\_annotation\_96.bed -ub reference.mmi Reads.fastq.gz |samtools view -bS > mapping.bam

The BED le can be generated from EnsEMBL GTF les using the following command:

$paftools.js gff2bed annotation.gtf > annotation.bed

1. 3. Mapping RNA modi cations using JACUSA2 pipeline: JACUSA2
2. [Piechotta et al., 2021] rapidly detects RNA modi cations based on
3. a comparative strategy where read alignment features (mismatch, in-
4. sertion and deletion) of samples of interest are compared to a reference
5. sequence (call-1) or against reference samples without the correspond-
6. ing RNA modi cation of interest (call-2). JACUSA2 processes repli-
7. cate experiments. The analysis of read alignment signatures is used
8. for RNA modi cation detection. Particularly, we integrate JACUSA2
9. call-2 method with the downstream analysis in one work ow using
10. the Snakemake work ow manager [Koster and Rahmann, 2012]. Our
11. Snakemake work ow encompasses several steps as shown in Figure
12. 4. The work ow requires BAM les from 2 conditions as input. We
13. suggest to lter secondary and poor alignments beforehand. The out-
14. put of JACUSA2 call2 is preprocessed (get features) and subjected
15. to a machine learning step to extract and visualize modi cation pat-
16. terns (resp. get pattern, visualize pattern) and make predictions (pre-
17. dict modi cation). "split trani test" rule allows splitting input data
18. into a training set and a test set. To use our snakemake-based JA-
19. CUSA2 pipeline a set of parameters should be de ned in the "con-
20. g.yaml" le; mainly: the label of the analysis ’label’, the input bam
21. les under ’data’, the reference sequence ’reference’, a le containing
22. size of chromosomes ’chr size’, JACUSA2 jar le ’jar’, plus the path to
23. inputs and outputs under ’path inp’ and ’path out’ elds respectively.
24. We typically execute the work ow on a multi-core CPU system using
25. the following command by specifying the number of cores to be used
26. "{cores [=all]" and the rule name:

361

362

363

$ snakemake --cores all rule\_name

Please consult the Snakemake documentation for further details (see <https://snakemake.readthedocs.io/en/stable/>).

1. Use Case 1: Comparison of wild-type and knock-out samples
2. The JACUSA2 work ow detects RNA modi cations using direct RNA se-
3. quencing by comparing modi ed samples to unmodi ed control samples.

10

1. Here, we used a published dataset of HEK293 cell lines to map m6A mod-
2. i cation [Pratanwanich et al., 2021]. Our examples encompasses two con-
3. ditions: wild-type RNA (WT, modi ed RNAs) and RNA from METTL3
4. knockout cells (KO, m6A modi cation is absentr). We use two replicates per
5. condition (see <https://doi.org/10.5281/zenodo.5913452>). The FASTQ
6. les are mapped using Minimap2 as described in the previous section. The
7. following analysis is validated against m6A sites consistently reported in
8. three miCLIP-based studies Boulias et al. [2019], Koh et al. [2019], Kortel
9. et al. [2021] (Figure 5).
10. Starting with the preprocessed mapped reads as inputs (BAM les),
11. ’HEK293T-WT-rep2.bam’ and ’HEK293T-WT-rep3.bam’ represent the wild-
12. type replicates and ’HEK293T-KO-rep2.bam’ and ’HEK293T-KO-rep3.bam’
13. the control replicates,
14. 1. Compute read error pro le with the jacusa2 call2 rule:

381

382

383

384

385

386

387

388

389

390

391

392

393

394

395

396

397

$ snakemake --cores all jacusa2\_call2 $

The method requires BAM les of the paired conditions and the cor-responding library information "-P1" and "-P2". In addition to the mismatch score, add "-D" and "-I" to output the deletion and insertion scores. JACUSA2 allows ltering reads according to many parameters. Here, we consider all sites with base calling quality "-q [> 1]", map-ping quality "-m [> 1] and read coverage "-c [> 4]". Here, we consider ltering sites within homopolymer regions "-a [=Y]". The output (named here, "Cond1vsCond2Call2.out") consists of a read error pro-le where the format is a combination of BED6 with JACUSA2 call-2 speci c columns and common info columns: info, lter, and ref. Check JACUSA2 manual for more details on JACUSA2 lter and output op-tions [JAC, 2021]. The number of threads can be customized via the parameter "-p". All parameters related to the JACUSA2 method can be added under the eld "jacusa params" in the con g le by setting the name of the parameter followed by the corresponding value [key: value]. Be aware to set all parameters before running the pipeline.

398 2. Process JACUSA2 output with the get features rule:

399

400

401

402

403

404

405

406

$ snakemake --cores all get\_features

we select all sites within 5-mer of a central nucleotide ’A’ anked by 2 random nucleotides (NNANN) and we lter out sites of the homo-polymer regions. Then, we rebuild the tabular features such that the observations are only sites with a reference base ’A’. Each site is char-acterized by 15 features corresponding to the mismatch, insertion and deletion scores for the observed site and its two anking positions from both sides. The rule "get features" performs the preprocessing

11

407

408

409

step. Use the parameter ’region’ with a le containing target 5-mers to limit the analysis to speci c sites. The output is an R object "fea-tures/features.rds", representing the matrix of Sites 15 features.

1. 3. Extract characteristic m6A modi cation patterns with the get pattern
2. rule:

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

$ srun snakemake --cores all get\_pattern

We learn a model representing the m6A modi cation patterns given the matrix of Sites Features. To this end, we employ non-negative matrix factorization (NMF)[Lee and Seung, 1999]. Brie y, NMF fac-torizes a non-negative data matrix X (here: n sites and m features) into two non-negative matrices as X W H, such that W is an n k matrix containing basis vectors and H is an k m matrix containing coe cient vectors. The coe cient vectors and their combination can be viewed as a pattern for m6A modi cation. The rank of factorization k is a critical parameter that a ects the performance substantially. We suggest to select the rank k according to the method of Frigyesi and

Hoglund [2008] by looking at silhouette [Rousseeuw, 1987] and cophe-netic correlation [Brunet et al., 2004] indices. Accordingly, the perfor-mance indices are computed for di erent choices of rank (k < n; m) and compared to the performance of a random permutation of the original data. Subsequently, the chosen rank corresponds to the value with the largest di erence between slopes of the original and the ran-domized data. Here, the unsupervised pattern training is based on the consensus set of 1; 905 m6A sites reported in the three miCLIP-based studies mentioned earlier. Based on the silhouette and cophenetic correlation indices, we identi ed an optimal factorization rank of 6 (Figure 6A). We then analyzed the identi ed patterns. According to the membership indicator of each site in matrix W , more than 80% of m6A modi cation sites can be represented by ve patterns (Patterns 1,2,3,4,6) (Figure 6B). Interestingly, the linear combination of these ve patterns in Figure 6C highlights the importance of position 3.

Multiple patterns and their combinations can be visualized using visualize pattern rule. The corresponding outputs are under "pattern/viz" folder.

$ srun snakemake --cores all visualize\_pattern

441 4. Predict m6A modi cations with the predict modification rule:

442

443

444

$ srun snakemake --cores all predict\_modification

This rule uses patterns of 15 features to predict m6A modi cation. We examine the ability of prediction on a subset of data of more than

12

445

446

447

448

449

450

451

452

453

454

455

456

1:52 million sites including 17; 021 miCLIP m6A sites. We opt for the linear combination of the ve most relevant patterns described in step 3. The empirical Cumulative Distribution Function (eCDF) of the inferred scores shows a signi cant di erence between the di er-ent miCLIP m6A categories (miCLIP annotation) and the unmodi ed sites (Figure 6D). As the number of negative samples is much larger than the number of positive samples, we consider the Positive Predic-tive Value (PPV, TP/(TP+FP)) of our predictions. Here, Figure 6E shows that PPV increases with the score cut-o . The nal output is a BED le containing the estimated scores as well as the corresponding eCDF and PPV plots. The corresponding outputs are located under a new folder called "prediction".

1. Use Case 2: Comparison of wild-type and IVT samples
2. An alternative way to detect RNA modi cations is to compare a modi-
3. ed sample to an in-vitro transcribed (IVT) control sample. Therefore,
4. we benchmark JACUSA2 on a sample set of two replicates from wild-type
5. HEK293 cell lines (see Use Case 1m modi ed sample) Pratanwanich et al.
6. [2021] and a modi cation-free IVT sample from HEK293 cDNA (control
7. sample) (see "Preparation of an in vitro transcriptome sample"). All analy-
8. sis steps are identical to use case 1. We evaluate the analysis against miCLIP
9. m6A sites (Figure 5).
10. 1. Identify read error pro le: we use JACUSA2 call-2 with the same
11. parameters as the previously described case. The input BAM les
12. (HEK293T-WT-rep2.bam, HEK293T-WT-rep3.bam) and (HEK293T-
13. IVT-rep1.bam, HEK293T-IVT-rep2.bam) are associated to the wild-
14. type and IVT replicate samples respectively. All input les are avail-
15. able from <https://doi.org/10.5281/zenodo.5913452>
16. $ srun snakemake --cores all jacusa2\_call2
17. 2. Process JACUSA2 output: we consider all sites corresponding to a
18. 5-mer (NNANN) with a central A residue. We employ the Y lter to
19. excludes sites within homo-polymer regions.
20. $ srun snakemake --cores all get\_features
21. 3. Extract m6A modi cation pattern: Based on the silhouette and cophe-
22. netic correlation indices, we identi ed an optimal factorization rank of
23. 6 (Figure 7A). We determined the predominant factors from matrix
24. W . Accordingly, more than 80% of m6A modi cation sites can be
25. represented by four patterns (Patterns: 1,2,3,6) (Figure 7B). In agree-
26. ment with Use Case 1, the linear combination of the four patterns

13

483

484

485

con rms the importance of position 3 and the implication of all scores as shown in Figure 7C.

$ srun snakemake --cores all get\_pattern

1. 4. Predict m6A modi cations: we evaluate the prediction ability of the
2. detected patterns on a test set of almost 1; 52 million sites where
3. 17; 021 are miCLIP-m6A modi ed. We consider the linear combination
4. of the four most relevant patterns (1,2,3,6). Figure 7D shows the eCDF
5. of the inferred scores. The di erence between the cumulative distri-
6. bution of non miCLIP sites and miCLIP sites can be nicely observed,
7. while the PPV plot shows a lower performance as compared to Use
8. Case 1 (Figure 7E). The decrease in performance is likely explained
9. by the absence of all modi cations in the IVT samples. Additional
10. adenosine modi cations such as RNA editing may be counted as false
11. positives.
12. $ srun snakemake --cores all predict\_modification
13. NOTES
14. Tips and Tricks
15. 1. The reverse transcription step during library preparation is optional.
16. However, we recommend to include this step to ensure proper sequenc-
17. ing also of RNAs with secondary structures. Superscript IV reverse
18. transcriptase may be replaced by Superscript III reverse transcriptase,
19. which is used in the protocol provided by Oxford Nanopore Technolo-
20. gies.
21. 2. The library preparation protocol contains two bead clean up steps. It
22. is important to remove ethanol and wash bu er completely. However,
23. beads should not be dried for several minutes. Directly add water
24. or elution bu er after washing to prevent sticking of the RNA to the
25. beads.
26. 3. The default lter in current MinKNOW versions is a Q score of 9. For
27. direct RNA sequencing we recommend to adjust the output lter to a
28. minimum Q score of 7, as in previous MinKNOW versions.
29. 4. During preparation of the in vitro transcriptome sample, in vitro tran-
30. scription and clean up kits may be replaced by equivalent products.
31. The protocol however has been tested only with the mentioned kits.
32. 5. Con guration of the pipeline should be handled via the con g le:
33. config.yaml. All parameters should be set before executing rules.

14

1. 6. Once the pipeline has run successfully you should expect the following
2. folders with the corresponding outputs in the output directory: bam,
3. jacusa, features, patterns, and prediction.
4. 7. JACUSA2 call2 could be run separately using the command line as
5. described in JACUSA2 manual [JAC, 2021], then put the output under
6. a new folder with the name ’jacusa’ under the output directory.
7. 8. In the snakemake pipeline, rules are linked so that the work ows
8. are determined from top (e.g. predict modi cation) to bottom (e.g.
9. sort bam) and executed accordingly from bottom to top (Figure 4).
10. Therefore, running for example "predict modi cation" rule leads to
11. executing all rules on its pipeline.
12. 9. This work ow uses a subset of the input data that correspond to
13. known modi ed sites to derive characteristic read alignment pro les
14. by NMF (patterns). Alternatively, prede ned patterns could be used
15. via a preloaded NMF R object to compute scores for modi cation site
16. prediction.
17. ACKNOWLEDGMENTS
18. The authors would like to thank Harald Wilhemi for testing the snakemake
19. pipeline. This work was supported by DFG SPP 1784 (DI1501/11-1) and
20. DFG TRR 319 - RMaP.
21. REFERENCES
22. Minimap2. <https://github.com/lh3/minimap2>. Accessed: 2022-01-19.
23. Jacusa2 manual. <https://github.com/dieterich-lab/JACUSA2>, 2021.
24. Accessed: 2022-01-15.
25. Samir Adhikari, Wen Xiao, Yong-Liang Zhao, and Yun-Gui Yang. m(6)a:
26. Signaling for mrna splicing. RNA biology, 13:756{759, September 2016.
27. ISSN 1555-8584. doi: 10.1080/15476286.2016.1201628.
28. Ina Anreiter, Quoseena Mir, Jared T. Simpson, Sarath C. Janga, and
29. Matthias Soller. New twists in detecting mrna modi cation dynamics.
30. Trends in biotechnology, 39:72{89, January 2021. ISSN 1879-3096. doi:
31. 10.1016/j.tibtech.2020.06.002.
32. Konstantinos Boulias, Diana Toczydlowska-Socha, Ben R Hawley, Noa
33. Liberman, Ken Takashima, Sara Zaccara, Theo Guez, Jean-Jacques
34. Vasseur, Francoise Debart, L Aravind, et al. Identi cation of the m6am

15

1. methyltransferase pcif1 reveals the location and functions of m6am in the
2. transcriptome. Molecular cell, 75(3):631{643, 2019.
3. Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov.
4. Metagenes and molecular pattern discovery using matrix factorization.
5. Proceedings of the national academy of sciences, 101(12):4164{4169, 2004.
6. Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali
7. Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine
8. Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and
9. Gideon Rechavi. Topology of the human and mouse m6a rna methylomes
10. revealed by m6a-seq. Nature, 485:201{206, April 2012. ISSN 1476-4687.
11. doi: 10.1038/nature11112.
12. Hao Du, Ya Zhao, Jinqiu He, Yao Zhang, Hairui Xi, Mofang Liu, Jinbiao
13. Ma, and Ligang Wu. Ythdf2 destabilizes m 6 a-containing rna through
14. direct recruitment of the ccr4{not deadenylase complex. Nature commu-
15. nications, 7(1):1{11, 2016.
16. Attila Frigyesi and Mattias Hoglund. Non-negative matrix factorization for
17. the analysis of complex gene expression data: identi cation of clinically
18. relevant tumor subtypes. Cancer informatics, 6:CIN{S606, 2008.
19. David Garcias Morales and Jose L. Reyes. A birds’-eye view of the activ-
20. ity and speci city of the mrna m, javax.xml.bind.jaxbelement@6d66739e,
21. a methyltransferase complex. Wiley interdisciplinary reviews. RNA, 12:
22. e1618, January 2021. ISSN 1757-7012. doi: 10.1002/wrna.1618.
23. Guifang Jia, Ye Fu, Xu Zhao, Qing Dai, Guanqun Zheng, Ying Yang,
24. Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, and Chuan He.
25. N6-methyladenosine in nuclear rna is a major substrate of the obesity-
26. associated fto. Nature chemical biology, 7:885{887, October 2011. ISSN
27. 1552-4469. doi: 10.1038/nchembio.687.
28. Shengdong Ke, Endalkachew A. Alemu, Claudia Mertens, Emily Conn Gant-
29. man, John J. Fak, Aldo Mele, Bhagwattie Haripal, Ilana Zucker-Schar ,
30. Michael J. Moore, Christopher Y. Park, Cathrine Broberg Vagb , Anna
31. Kussnierczyk, Arne Klungland, James E. Darnell, and Robert B. Darnell.
32. A majority of m6a residues are in the last exons, allowing the potential
33. for 3’ utr regulation. Genes & development, 29:2037{2053, October 2015.
34. ISSN 1549-5477. doi: 10.1101/gad.269415.115.
35. Casslynn WQ Koh, Yeek Teck Goh, and WS Sho Goh. Atlas of quantitative
36. single-base-resolution n 6-methyl-adenine methylomes. Nature communi-
37. cations, 10(1):1{15, 2019.

16

1. Nadine Kortel, Cornelia Ruckle, You Zhou, Anke Busch, Peter Hoch-Kraft,
2. Reymond FX Sutandy, Jacob Haase, Mihika Pradhan, Michael Musheev,
3. Dirk Ostareck, et al. Deep and accurate detection of m6a rna modi cations
4. using miclip2 and m6aboost machine learning. bioRxiv, pages 2020{12,
5. 2021.
6. Johannes Koster and Sven Rahmann. Snakemake|a scalable bioinformatics
7. work ow engine. Bioinformatics, 28(19):2520{2522, 2012.
8. Daniel D Lee and H Sebastian Seung. Learning the parts of objects by
9. non-negative matrix factorization. Nature, 401(6755):788{791, 1999.
10. Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christo-
11. pher E. Mason, and Samie R. Ja rey. Comprehensive analysis of mrna
12. methylation reveals enrichment in 3’ utrs and near stop codons. Cell, 149:
13. 1635{1646, June 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.05.003.
14. Deepak P Patil, Brian F Pickering, and Samie R Ja rey. Reading m6a in
15. the transcriptome: m6a-binding proteins. Trends in cell biology, 28(2):
16. 113{127, 2018.
17. Michael Piechotta, Qi Wang, Janine Altmuller, and Christoph Dieterich.
18. Rna modi cation mapping with jacusa2. bioRxiv, 2021.
19. Ploy N Pratanwanich, Fei Yao, Ying Chen, Casslynn WQ Koh, Yuk Kei
20. Wan, Christopher Hendra, Polly Poon, Yeek Teck Goh, Phoebe ML Yap,
21. Jing Yuan Chooi, et al. Identi cation of di erential rna modi cations
22. from nanopore direct rna sequencing with xpore. Nature Biotechnology,
23. 39(11):1394{1402, 2021.

613 Jean-Yves Roignant and Matthias Soller. m,

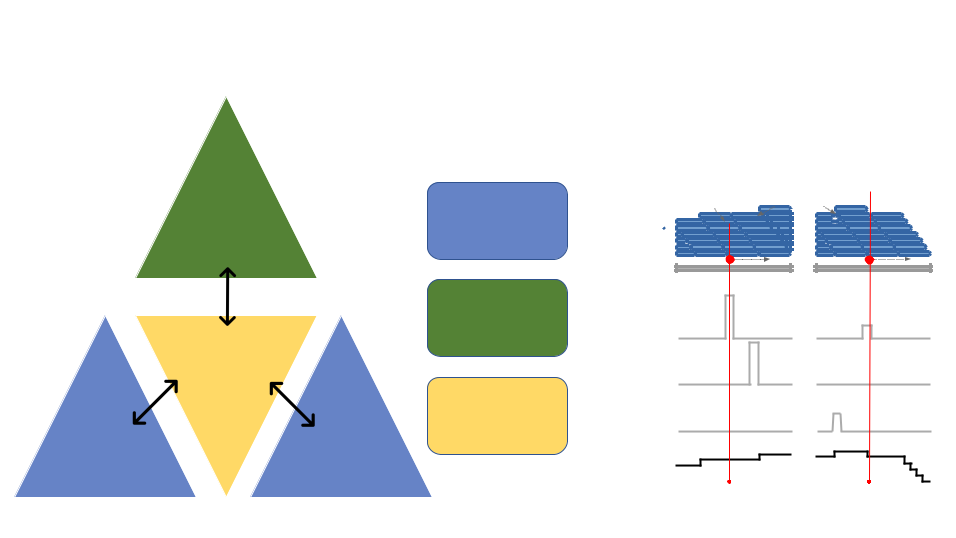
1. javax.xml.bind.jaxbelement@8cec19d, a in mrna: An ancient mech-
2. anism for ne-tuning gene expression. Trends in genetics : TIG, 33:
3. 380{390, June 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2017.04.003.
4. Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic rna
5. modi cations in gene expression regulation. Cell, 169:1187{1200, June
6. 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.045.
7. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and
8. validation of cluster analysis. Journal of computational and applied math-
9. ematics, 20:53{65, 1987.
10. Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how:
11. Context-dependent functions of rna methylation writers, readers, and
12. erasers. Molecular cell, 74:640{650, May 2019. ISSN 1097-4164. doi:
13. 10.1016/j.molcel.2019.04.025.

17

1. Xiao Wang, Zhike Lu, Adrian Gomez, Gary C Hon, Yanan Yue, Dali Han,
2. Ye Fu, Marc Parisien, Qing Dai, Guifang Jia, et al. N 6-methyladenosine-
3. dependent regulation of messenger rna stability. Nature, 505(7481):117{
4. 120, 2014.
5. Xiao Wang, Boxuan Simen Zhao, Ian A Roundtree, Zhike Lu, Dali Han,
6. Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He.
7. N6-methyladenosine modulates messenger rna translation e ciency. Cell,
8. 161(6):1388{1399, 2015.
9. Sara Zaccara, Ryan J. Ries, and Samie R. Ja rey. Reading, writing and
10. erasing mrna methylation. Nature reviews. Molecular cell biology, 20:608{
11. 624, October 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0168-5.
12. Zhang Zhang, Tao Chen, Hong-Xuan Chen, Ying-Yuan Xie, Li-Qian Chen,
13. Yu-Li Zhao, Biao-Di Liu, Lingmei Jin, Wutong Zhang, Chang Liu,
14. et al. Systematic calibration of epitranscriptomic maps using a synthetic
15. modi cation-free rna library. Nature Methods, 18(10):1213{1222, 2021.
16. Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min
17. Huang, Charles J. Li, Cathrine B. Vagb , Yue Shi, Wen-Ling Wang, Shu-
18. Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin
19. Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan,
20. Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne
21. Klungland, Yun-Gui Yang, and Chuan He. Alkbh5 is a mammalian rna
22. demethylase that impacts rna metabolism and mouse fertility. Molecular
23. cell, 49:18{29, January 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.
24. 10.015.

18

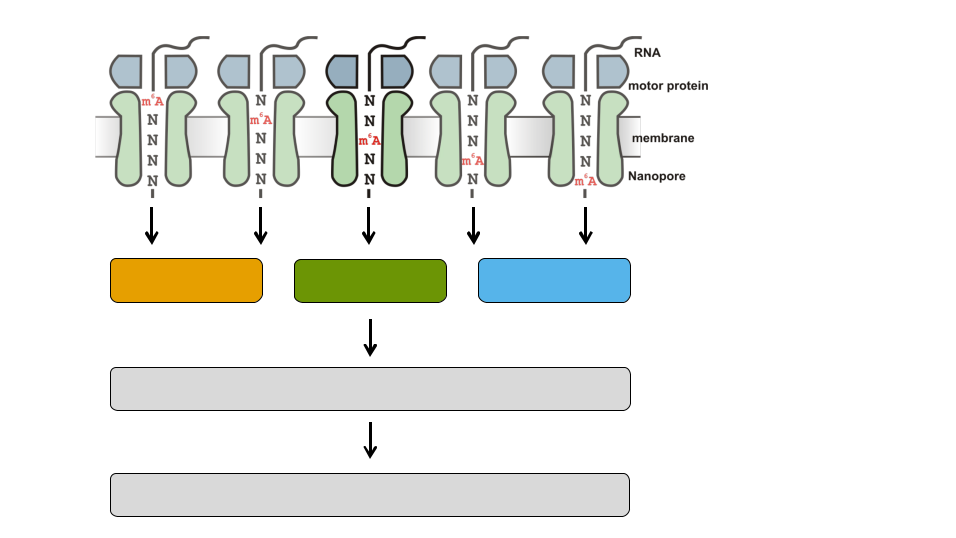
1. FIGURES



|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  | Computational approach | | | |
|  |  |  |  | Condition 1 | | Condition 2 | |
|  |  | Not modified | **Mapped** | Mismatches | Insertions | Deletions | Mismatches |
|  | KO |  |  |
|  | **G** | **I** |  | **G** |
|  |  | **reads** | **G** | **I** |  | **G** |
|  |  | at all |  | **G** | **I** |  |
|  | samples |  | **G** | **I** |  |  |
|  |  |  |  |  |  |
|  |  |  | **Reference** | **A** |  |  | **A** |
|  |  |  | **sequence** |  |  |
|  |  |  |  |  |  |  |
|  |  | Missing specific | **Mismatch** |  |  |  |  |
|  |  | modification(s) |  |  |  |  |
|  | **Samples** | **rate** |  |  |  |  |
|  |  |  |  |  |  |  |
|  | **of** |  | **Insertion** |  |  |  |  |
|  | **interest** | All | **rate** |  |  |  |  |
|  |  |  |  |  |  |
|  |  | **Deletion** |  |  |  |  |
|  |  | modifications |  |  |  |  |
|  |  | **rate** |  |  |  |  |
| Simulated | IVT | present |  |  |  |  |  |
| samples | samples |  | **Coverage** |  |  |  |  |
|  | **pro)le** |  |  |  |  |
|  |  |  |  |  |  |  |

Figure 1: General outline of RNA modi cation detection by JA-CUSA2. A key feature of our approach is that multiple replicates of the same condition can be considered simulatenously. Samples of interests where all modi cations are present could be compared with either KO sam-ples where the modi cation of interest is missing or IVT/simulated samples where all modi cations are absent (left panel). Read stacks (in blue) are compared head-to-head as shown on the right.

19



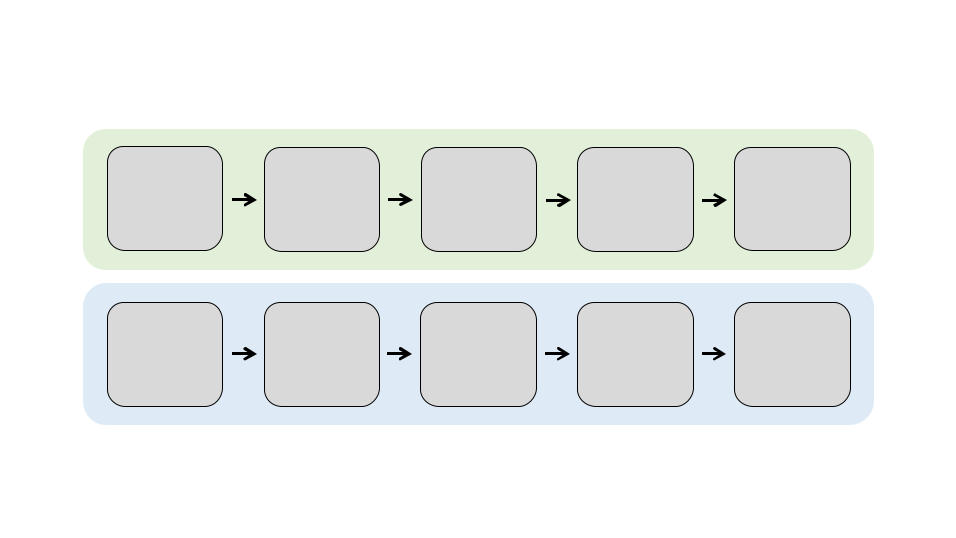
Deletion Mismatch Insertion

Modification-specific pattern (Feature set)

JACUSA2 score

Figure 2: Motivation of 5-mer context for RNA modi cation map-ping. The nanopore covers 5 consecutive RNA residues. That is why we consider a 5-mer context and derive 3 principal features for every position within a given 5-mer (15 features in total, with a central A residue in this example). We evaluate each feature set by previously learned patterns and compute a nal score for modi cation site detection.

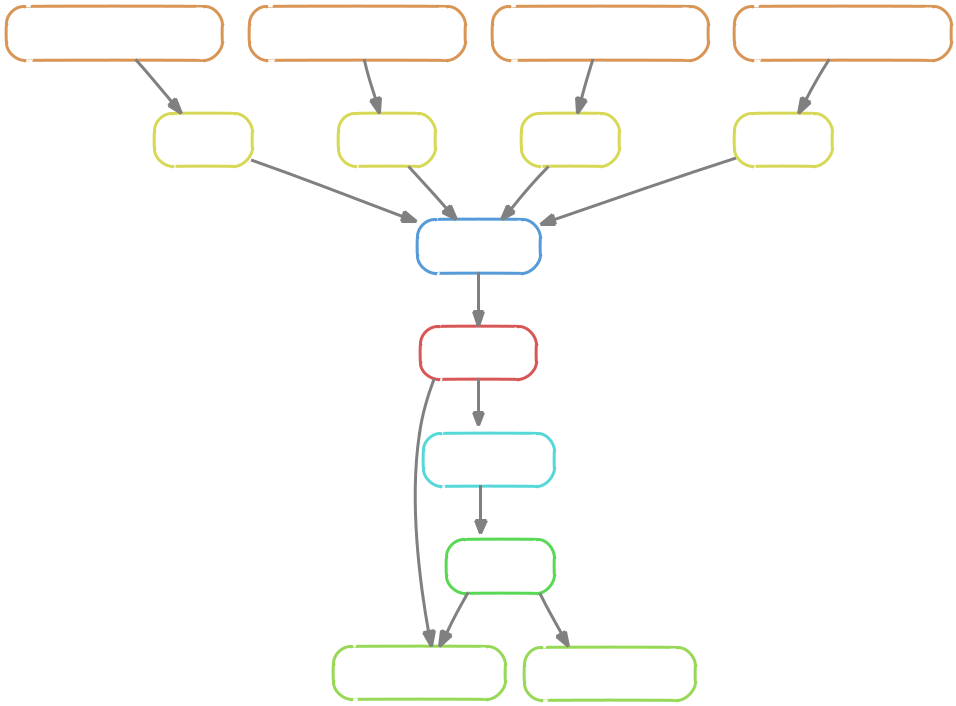
20



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **A)** |  |  |  |  |
| Total RNA | polyA+ | ONT direct | Guppy | FASTQ |
| RNA-seq | basecalling |
| RNA | files |
|  | (SQK-RNA002) | (high accuracy) |
|  |  |  |
| **B)** |  |  |  |  |
| FASTQ | minimap2 | SAMtools | JACUSA2 | Candidate |
| files | alignment | filtering | analysis | m6A sites |

Figure 3: Experimental and computational work ow. A) Starting from total cellular RNA, polyA+ RNA is isolated and subjected to Nanopore direct RNA-sequencing. Guppy basecalling can be done as live basecalling during sequencing or after the sequencing run from generated FAST5 les, resulting in FASTQ output les. B) FASTQ les are aligned to a reference sequence with Minimap2. SAMtools is used to generate BAM les as input for JACUSA2 analysis, which yields candidate m6A sites as described in Figure 4.

21



|  |  |  |  |
| --- | --- | --- | --- |
| sort\_bam | sort\_bam | sort\_bam | sort\_bam |
| Input: mapping\_cond1\_rep1.bam | Input: mapping\_cond1\_rep2.bam | Input: mapping\_cond2\_rep1.bam | Input: mapping\_cond2\_rep2.bam |
| filter\_bam | filter\_bam | filter\_bam | filter\_bam |

jacusa2\_call2

get\_features

split\_train\_test

get\_pattern

predict\_modification visualize\_pattern

Figure 4: Computational work ow. Snakemake work ow for RNA mod-i cation detection based on JACUSA2 variant calling.

22

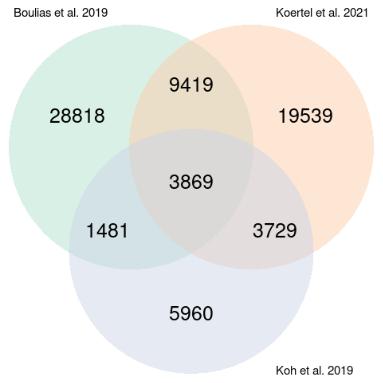


Figure 5: m6A sites reported in the three miCLIP-based studies Boulias et al. [2019], Koh et al. [2019] and Kortel et al. [2021].

23

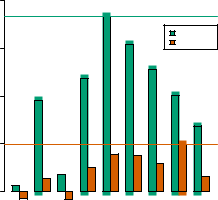
1. TABLES

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Software | Version | | |  |  |  |  | Description |
|  |  |  |  | |  |  |  |  |
|  |  |  |  | |  |  | |  |
| Minimap2 | <https://github.com/lh3/minimap2> | | | | | |  | <https://lh3.github.io/minimap2/> |
|  | v2.22 or later | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| samtools | [https://github.com/samtools/](https://github.com/samtools/samtools) | | |  |  |  |  | <http://samtools.github.io/> |
|  | [samtools](https://github.com/samtools/samtools) v1.12 or later | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| JAVA | <https://openjdk.java.net/> | | | 11.0.12 | | |  | OpenJDK Runtime Environment |
|  | 2021-07-20 - JAVA 11 or later | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| R | <https://www.r-project.org/> | | | version | | |  | The R Project for Statistical Comput- |
|  | 3.5.1 or later | | |  |  |  |  | ing |
|  |  |  |  |  |  |  |  |  |
| PERL | <https://www.perl.org/>version | | | 5.28.1 | | |  | Perl is a highly capable, feature-rich |
|  | or later | | |  |  |  |  | programming language |
|  |  |  |  | |  |  |  |  |
| bedtools | <https://github.com/arq5x/bedtools2> | | | | | |  | Perl is a highly capable, feature-rich |
|  | version 2.29.2 or later | | |  |  |  |  | programming language |
|  |  |  |  | |  |  |  |  |
| snakemake | <https://snakemake.github.io/>version | | | | | |  | The Snakemake work ow management |
|  | 6.8.1 or later | | |  |  |  |  | system |
|  |  |  |  | |  |  |  |  |
|  |  |  | Table 1: Software dependencies | | | | |  |
|  |  |  |  |  |  |  |  |  |
| R Pack- | Version | | |  |  |  |  | Description |
| ages |  |  |  |  |  |  |  |  |
|  |  |  |  | |  |  |  |  |
|  |  |  |  | |  |  | |  |
| ggplot2 | [https://cran.r-project.org/web/](https://cran.r-project.org/web/packages/ggplot2/index.html) | | | | | |  | ggplot2 is a system for declaratively |
|  | [packages/ggplot2/index.html](https://cran.r-project.org/web/packages/ggplot2/index.html) | | |  | - gg- | |  | creating graphics, based on The Gram- |
|  | plot2 |  | 3.3.0 or later |  |  |  |  | mar of Graphics. |
|  |  |  |  |  |  |
|  |  | | | |  |  |  |  |
| NMF | [https://cran.r-project.org/web/](https://cran.r-project.org/web/packages/NMF/index.html) | | | | | |  | Provides a framework to perform Non- |
|  | [packages/NMF/index.html](https://cran.r-project.org/web/packages/NMF/index.html) - NMF | | | |  | 0.22.0 |  | negative Matrix Factorization (NMF). |
|  | or later | | |  |  |  |  |  |
|  |  |  |  | |  |  |  |  |
|  |  |  | Table 2: R Package dependencies | | | | |  |

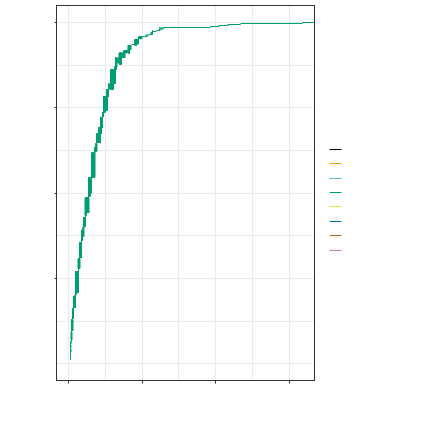
24

A - NMF rank selection B - NMF factorization

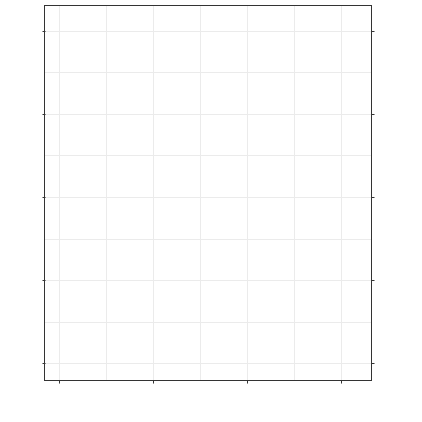
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0.8 |  |  |  |  |  |  |  |  |  | **Basis components** | | | | **Mixture coefficients** | | | |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.6 |  |  |  |  |  |  |  | Silhouette | |  |  |  | 1 | **basis** |  |  |  |  |  |
|  |  |  |  |  |  |  | Cophenetic | |  |  |  |  | 1 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | 0.8 | 2 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  | 3 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Delta0.4 |  |  |  |  |  |  |  |  |  |  |  |  |  | 4 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  | 0.6 | 5 |  |  |  | 1 | **basis** |
|  |  |  |  |  |  |  |  |  |  |  |  |  | 6 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 1 |
|  |  |  |  |  |  |  |  |  |  |  |  | 0.4 |  |  |  | 1 | 0.8 | 2 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 0.2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 3 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 4 |
|  |  |  |  |  |  |  |  |  |  |  |  | 0.2 |  |  |  |  | 0.6 | 5 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 2 |  | 6 |
| 0.0 |  |  |  |  |  |  |  |  |  |  |  |  | 0 |  |  |  | 0.4 | **consensus** |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |  |  |  |  |  |  |  |  |  | 1 |
|  |  |  |  |  |  |  |  |  |  | 2 |
|  |  |  |  | Rank | |  |  |  |  |  |  |  |  |  |  |  |  | 0.2 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 3 | 3 |
| C- Sum of top-ranked NMF patterns | | | | | | | | | |  |  |  |  |  |  |  | 0 | 4 |
|  |  |  |  |  |  |  |  | 5 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 6 |
|  |  |  |  |  |  |  |  |  | Insertion |  |  |  |  |  |  |  | 4 |  |  |
| 50. |  |  |  |  |  |  |  |  | Deletion |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  | Mismatch |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 40. |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 5 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 30. |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | 6 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 1 20.0. |  |  |  |  |  |  |  |  |  |  |  |  |  | Pos3Mismatch | Pos3DeletionPos1MismatchPos2DeletionPos1Deletion | Pos2MismatchPos4DeletionPos5DeletionPos1InsertionPos2InsertionPos4MismatchPos5MismatchPos3InsertionPos5Insertion | Pos4Insertion |  |  |
| 00. | Pos1 | | Pos2 | | Pos3 | | Pos4 | | Pos5 | 1 2 | 3 | 4 5 | 6 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |



D- eCDF of sum of patterns scores



E- Overlap of m6A prediction and miCLIP sites



|  |
| --- |
| y |

1.00

0.75

0.50

0.25

0.00 

0

1. score20

30

CLIP

Boulias

Boulias,Koertel

Boulias,Koertel,Koh

Boulias,Koh

Koertel

Koertel,Koh

Koh

NoCLIP

|  |
| --- |
| Log2(#predictions) |

20

15

10

5

0

0

1. Cutoff 20

30

100

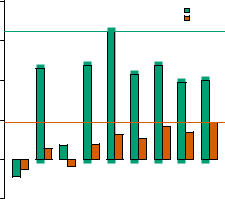
|  |  |
| --- | --- |
| 75 | Positive |
| ValuePredictive |
| 50 |
|  |
| 25 |  |
| 0 |  |

Figure 6: Case 1. WT versus KO. A: NMF rank selection. Barplots rep-resent di erences in cophenetic correlation and silhouette scores for NMF factorization of the original input and randomized data. Ranks with the largest di erences are determined for both measures, then the smallest rank is selected for the NMF decomposition. B: NMF result represented by the basis matrix W ("amplitude" matrix) and the coe cient matrix H ("pat-tern" matrix). The matrix H induces the RNA modi cation pattern. C: Barplot shows the sum of the top 5 patterns (y-axis) across the speci c 5-mer context (x-axis) and the score type: mismatch, deletion and insertion (resp. black, orange and blue). The 5 patterns (coe cient vectors: 1,2,3,4,6) are selected according to the predominant columns in matrix W. D: Score distribution inferred from the combined patterns, strati ed by the di erent categories of miCLIP validated sites and non miCLIP sites. E: Number of predicted m6A sites (green) and Positive Predictive Value (PPV) of pre-dicted m6A sites that overlap with miCLIP sites (orange).

25

A - NMF rank selection B - NMF factorization

|  |  |
| --- | --- |
| 0.8 | Silhouette |
| Cophenetic |



|  |
| --- |
| 0.6 |

|  |
| --- |
| 0.0 0.2 0.4 Delta |

|  |  |
| --- | --- |
| 0.−2 | 2345678910 |
|  | Rank |

C - Sum of top-ranked NMF patterns

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  | Insertion | |
| 300. |  |  |  |  | Deletion | |
|  |  |  |  | Mismatch | |
| 200. |  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 100. |  |  |  |  |  |
| 000. | Pos1 | Pos2 | Pos3 | Pos4 | Pos5 |
|  |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Basis components** | | | | | | **Mixture coefficients** | | | |  |  |
|  |  |  |  |  | 1 | **basis** |  |  |  |  |  |
|  |  |  |  |  |  | 1 |  |  |  |  |  |
|  |  |  |  |  | 0.8 | 2 |  |  |  |  |  |
|  |  |  |  |  | 3 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  | 4 |  |  |  |  |  |
|  |  |  |  |  | 0.6 | 5 |  |  |  | 1 | **basis** |
|  |  |  |  |  |  | 6 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | 1 |
|  |  |  |  |  | 0.4 |  |  |  | 1 | 0.8 | 2 |
|  |  |  |  |  |  |  |  |  |  |  | 3 |
|  |  |  |  |  |  |  |  |  |  |  | 4 |
|  |  |  |  |  | 0.2 |  |  |  |  | 0.6 | 5 |
|  |  |  |  |  |  |  |  |  | 2 |  | 6 |
|  |  |  |  |  | 0 |  |  |  | 0.4 | **consensus** |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  | 1 |
|  |  |  |  |  |  |  |  |  |  | 0.2 | 2 |
|  |  |  |  |  |  |  |  |  | 3 | 3 |
|  |  |  |  |  |  |  |  |  |  | 4 |
|  |  |  |  |  |  |  |  |  |  | 0 | 5 |
|  |  |  |  |  |  |  |  |  |  | 6 |
|  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  | 4 |  |  |
|  |  |  |  |  |  |  |  |  | 5 |  |  |
|  |  |  |  |  |  |  |  |  | 6 |  |  |
|  |  |  |  |  |  | Pos3Mismatch | Pos3DeletionPos1MismatchPos1DeletionPos4MismatchPos2MismatchPos2DeletionPos1InsertionPos2InsertionPos5MismatchPos4DeletionPos5DeletionPos3Insertion | Pos5Insertion | Pos4Insertion |  |  |
| 1 | 2 | 3 | 4 | 5 | 6 |  |  |  |  |  |  |

|  |
| --- |
| y |

1.00

0.75

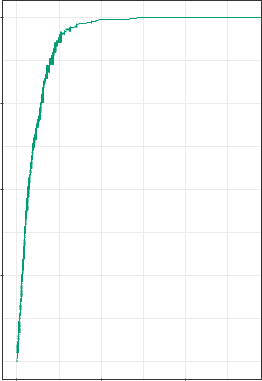
0.50

0.25

0.00 

0

10Score 20



CLIP

Boulias

Boulias,Koertel

Boulias,Koertel,Koh

Boulias,Koh

Koertel

Koertel,Koh

Koh

NoCLIP

E- Overlap of m6A prediction and miCLIP sites

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 20 |  |  |  |  |  |  |  | 100 | ValuePredictivePositive |
|  |  |  |  |  |  |  |  |
| Log2(#predictions) | 15 |  |  |  |  |  |  |  | 75 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
| 10 |  |  |  |  |  |  |  | 50 |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | 5 |  |  |  |  |  |  |  | 25 |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  | 0 |  |  |  |  |  |  |  | 0 |  |
|  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
|  |  | 0 | | 10 Cutoff | | 20 | |  |  |  |

Figure 7: Case 2. WT versus IVT. A: NMF rank selection. Barplots rep-resent di erences in cophenetic correlation and silhouette scores for NMF factorization of the original input and randomized data. Ranks with the largest value for both indices are determined, then the smallest rank is se-lected for the NMF decomposition. B: NMF result represented by the basis matrix W ("amplitude" matrix) and the coe cient matrix H ("pattern" matrix). The matrix H induces the RNA modi cation pattern. C: Barplots representing the linear combination of the top 4 patterns (y-axis) by by the number of position in the speci c 5-mer context (x-axis) and the score type: mismatch, deletion and insertion (resp. black, orange and blue). The 4 pat-terns (coe cient vectors: 1,2,3,6) are selected according to the predominant columns in matrix W. D: Score distribution inferred from the combined patterns, strati ed by the di erent categories of miCLIP validated sites and non miCLIP sites. E: Number of predicted m6A sites (green) and Positive Predictive Value (PPV) of predicted m6A sites that overlap with miCLIP

|  |  |
| --- | --- |
| sites (orange). | 26 |