

# Mapping of RNA modifications by direct Nanopore sequencing and JACUSA2

Amina Lemsara<sup>1,2</sup>, Christoph Dieterich<sup>\*1,2,3</sup>, and Isabel Naarmann-de Vries<sup>1,2,3</sup>

<sup>1</sup>Klaus Tschira Institute for Integrative Computational Cardiology, University Heidelberg, 69120 Heidelberg, Germany

<sup>2</sup>Department of Internal Medicine III (Cardiology, Angiology, and Pneumology), University Hospital Heidelberg, 69120 Heidelberg, Germany

<sup>3</sup>German Centre for Cardiovascular Research (DZHK)-Partner Site Heidelberg/Mannheim, 69120 Heidelberg, Germany

## Abstract

RNA modifications exist in all kingdom of life. Several different types of base or ribose modifications are now summarized under the term the "epitranscriptome". With the advent of high-throughput sequencing technologies much progress has been made in understanding RNA modification biology and how these modifications can influence many aspects of RNA life. The most widespread internal modification on mRNA is m6A, which has been implicated in physiological processes as well as disease pathogenesis. Here, we provide a workflow for the mapping of m6A sites using Nanopore direct RNA sequencing data. Our strategy employs pairwise comparison of base calling error profiles with JACUSA2. We outline a general strategy for RNA modification detection on mRNA and describe two specific use cases on m6A detection in detail. **Use case 1:** a sample of interest with modifications (e.g. "wild type" sample) is compared to a sample lacking a specific modification type (e.g. "knock out" sample, here *METTL3*-KO) or **Use case 2:** a sample of interest with modifications is compared to a sample lacking all modifications (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 flexible and can be easily adapted by users in different application scenarios.

---

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

## 33 INTRODUCTION

34 Chemical modifications on DNA and histones, also known as epigenetics  
35 marks, strongly impact gene expression during cell differentiation and in  
36 several other biological programs. In the 1970s, it was recognized that RNA  
37 is also subjected to extensive covalent modification, and studies in the late  
38 1980s revealed the widespread deamination of bases (termed RNA editing),  
39 which can lead to recoding if it occurs within coding sequences. Impres-  
40 sive development in the RNA modification field occurred during the past  
41 eight years, with the discovery of an extensive layer of base modifications  
42 in mRNAs. These can influence gene expression and have been already  
43 shown to be involved in primary cellular programs such as stem cell differ-  
44 entiation, response to stress, and the circadian clock. The study of RNA  
45 modifications and their effects is now referred to as epitranscriptomics, and  
46 it reveals striking similarities to what is known for epigenomics. To date  
47 thirteen distinct modifications have been identified on mRNA transcripts  
48 [Anreiter et al., 2021]. These modifications are catalyzed by a variety of  
49 dedicated enzymes and can be divided into two classes: modifications of  
50 cap-adjacent nucleotides and internal modifications.

51 In contrast to the m7G cap, the impact of internal modifications on gene  
52 regulation has been less studied apart from RNA editing, which is mediated  
53 by RNA deaminases (e.g. the ADAR family). The most widespread in-  
54 ternal mRNA modification is N6-methyladenosine (m6A). By modulating  
55 the processing of mRNA, m6A can regulate a wide range of physiological  
56 processes and its alteration has been linked to several diseases Roignant  
57 and Soller [2017], Zaccara et al. [2019], Shi et al. [2019]. The modification is  
58 catalyzed co-transcriptionally by a Mega-Dalton methyltransferase complex,  
59 which includes the heterodimer METTL3-METTL14 and other associated  
60 subunits Garcias Morales and Reyes [2021]. This modification is reversible  
61 since two proteins of the AlkB-family of demethylases can remove m6A from  
62 mRNA transcripts [Jia et al., 2011, Zheng et al., 2013]. In mammals, m6A  
63 preferentially localizes within long internal exons and at the beginning of  
64 terminal exons at so-called DRACH motif (D = A/G/U, R = A/G, H =  
65 A/C/U) sites [Dominissini et al., 2012, Meyer et al., 2012, Ke et al., 2015].  
66 Once deposited, m6A is recognized by several reader proteins that can af-  
67 fect the fate of mRNA transcripts in nearly every step of the mRNA life  
68 cycle, including alternative splicing [Adhikari et al., 2016, Roundtree et al.,  
69 2017], mRNA translation [Wang et al., 2015] and decay [Wang et al., 2014,  
70 Du et al., 2016, Roundtree et al., 2017]. The best-described readers are the  
71 YTH domain family of proteins that decode the signal and mediate m6A  
72 functions. By affecting RNA structure, m6A can also indirectly influence  
73 the association of additional RNA-binding proteins (RBPs) and the assem-  
74 bly of larger messenger ribonucleoprotein particles (mRNPs) [Patil et al.,  
75 2018].

Several approaches have been presented to map RNA modifications on RNA. Herein, we focus on mRNA modification site detection in general and on m6A in particular where antibody-based protocols (miCLIP), methylation-sensitive restriction enzyme assays (MazF) or transgenic approaches (TRIBE, DART) have been presented to map m6A sites. All of the aforementioned approaches rely on high-throughput short read sequencing on the Illumina platform. This typically involves cDNA synthesis by reverse transcription and PCR-based library amplification. One recent addition to the toolbox of RNA modification mapping is direct RNA single molecule long read sequencing on the Oxford Nanopore Technologies platform (dRNA-seq). While our software is able to deal with Illumina and Nanopore-based approaches, the latter is the principal topic of this methods article.

## MATERIALS

### ONT direct RNA sequencing

This section summarizes all necessary consumables for direct RNA sequencing of poly-adenylated RNA (i.e. mRNA) on the MinION or similar device.

1. 500 ng polyA<sup>+</sup> RNA isolated from total RNA e.g. with Oligotex mRNA kit (Qiagen) or Dynabeads oligo dT<sub>25</sub> beads (Thermo Fisher Scientific) or *in vitro* transcriptome sample. Store RNA at -80 °C and the mRNA purification kit as recommended by the manufacturer.
2. Nuclease-free water. Store at room temperature.
3. Direct RNA-sequencing kit (SQK-RNA002, Oxford Nanopore Technologies). Store at -20 °C.
4. NEBNext Quick Ligation Reaction Buffer (New England Biolabs). Store at -20 °C.
5. T4 DNA Ligase (New England Biolabs). Store at -20 °C.
6. dNTP Mix (10 mM each). Store at -20 °C.
7. SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific). Store at -20 °C.
8. Agencourt RNAClean XP beads (Beckman Coulter). Store at 4 °C.
9. 70 % ethanol, freshly prepared.
10. Qubit dsDNA HS assay kit and Qubit Fluorometer (Thermo Fisher Scientific).

11. Flow cell priming kit (EXP-FLP002, Oxford Nanopore Technologies).  
Store at -20 °C.
12. Thermocycler.
13. Gentle rotator mixer.
14. Magnetic stand for 1.5 ml tubes.
15. 1.5 ml DNA LoBind tubes (Eppendorf), 0.2 ml PCR tubes.
16. MinION or GridION sequencing device and MinION R9.4.1 Flow cells  
(FLO-MIN106D, Oxford Nanopore Technologies). Store Flow cells at  
4 °C.

#### **Preparation of an *in vitro* transcriptome sample**

1. 100 ng polyA<sup>+</sup> RNA isolated from total RNA e.g. with Oligotex  
mRNA kit (Qiagen) or Dynabeads oligo dT<sub>25</sub> beads (Thermo Fisher  
Scientific). Store RNA at -80 °C and the mRNA purification kit as  
recommended by the manufacturer
2. 10 μM oligo(dT)-VN RT primer.  
TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN. Store at -20 °C.
3. 20 μM template switching oligo (TSO). ACTCTAATACGACTCAC-  
TATAGGGAGAGGGCrGrG+G. Store at -20 °C.
4. 10 μM T7 extension primer. GCTCTAATACGACTCACTATAGG.  
Store at -20 °C.
5. Nuclease-free water. Store at room temperature.
6. dNTP Mix (10 mM each). Store at -20 °C.
7. Template Switching RT Enzyme Mix (New England Biolabs). Store  
at -20 °C.
8. Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs).  
Store at -20 °C.
9. RNase H (5,000 U/ml) (New England Biolabs). Store at -20 °C.
10. NucleoSpin Gel and PCR Clean-up, Mini kit for gel extraction and  
PCR clean up (Macherey-Nagel) or equivalent. Store at room temper-  
ature.
11. MEGAscript T7 transcription kit (Thermo Fisher Scientific). Store at  
-20 °C.

- 141 12. RNA Clean & Concentrator-25 kit (Zymo Research). Store at room  
142 temperature.
- 143 13. Thermocycler.
- 144 14. Table top centrifuge for 1.5 ml tubes.
- 145 15. Nanodrop spectrophotometer or equivalent.
- 146 16. 0.2 ml PCR tubes, 1.5 ml DNA LoBind tubes (Eppendorf).

## 147 Hardware requirements

148 All analyses have been performed/tested on two alternative hardware sys-  
149 tems: a standard Linux desktop computer or an Apple iMac (Retina 5K,  
150 ultimo 2014). The workflow requires a multi-core processor system with  
151 minimal main memory of 16GB RAM and several GBs of free disk space  
152 (depending on data set size).

## 153 Software dependencies and installation

154 Our analysis workflow has few requirements, which are detailed in Table 2.  
155 Specifically, to execute our workflow, the following prerequisites are neces-  
156 sary: a BASH shell, a JAVA runtime environment, a working PERL and  
157 R installation. Additional i.e. non-standard software to process and map  
158 Nanopore reads (bedtools, samtools and Minimap2) are obligatory, while  
159 the installation of a Nanopore read simulator (NanoSim) is optional and de-  
160 pends on your use case. Table ?? lists some additional R packages, which are  
161 required to run the R code. Detailed instructions on how to setup are found  
162 under [https://github.com/dieterich-lab/MiMB\\_JACUSA2\\_chapter](https://github.com/dieterich-lab/MiMB_JACUSA2_chapter)

## 163 METHODS

164 Our workflow is based on the pairwise comparison of samples with differ-  
165 ent modification status (Figure 1). The sample of interest (yellow) may be  
166 compared to different samples lacking certain modifications. If available,  
167 the wild type (WT) sample can be compared to a knock out (KO) sample  
168 lacking specific enzymatic activities (green), as outlined in Use Case 1. Al-  
169 ternatively, a sample lacking all modifications may be used for comparison  
170 (blue). This may be either a simulated sample (i.e. with NanoSim) or an *in*  
171 *vitro* transcribed sample derived from cDNA. Such an analysis is detailed in  
172 Use Case 2. In any setting, JACUSA2 calculates scores for the Mismatch,  
173 Insertion and Deletion rates of the pairwise comparisons as outlined above  
174 (Figure 1, right).

labeling  
of Ta-  
bles in  
PDF  
doesn't  
seem  
to be  
correct

One feature of Nanopore sequencing is to read sequences as 5-mers, as always five nucleotides are occupied by the pore protein (Figure 2). Because of this, a m6A modification may affect basecalling not only if the modified nucleotide is in the central position, but also at neighboring positions (-2 to +2). To account for this, JACUSA2 scores for Deletion, Mismatch and Insertion are calculated for the 5-mer context. Depending on the modification-specific signature, a Feature set can be selected to calculate the final JACUSA2 score (Figure 2).

Our workflow can be divided into a wet-lab part (Figure 3A) and a computational part (Figure 3B). Starting from total cellular RNA, polyA<sup>+</sup> 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 files, resulting in FASTQ output files (Figure 3A). FASTQ files are aligned to a reference sequence with Minimap2. SAMtools is used to generate BAM files as input for JACUSA2 analysis, which yields candidate m4A sites (Figure 3B).

## Nanopore direct RNA sequencing

1. Adjust 500 ng polyA<sup>+</sup> RNA to a total volume of 9  $\mu$ l with nuclease-free water. Complete RT adapter ligation reaction (in 0.2 ml PCR tube) with 3  $\mu$ l NEBNext Quick Ligation Reaction Buffer, 0.5  $\mu$ l RNA CS (RCS, from SQK-RNA002), 1  $\mu$ l RT-Adapter (RTA, from SQK-RNA002) and 1.5  $\mu$ l T4 DNA Ligase. Incubate 10 min at room temperature.
2. Prepare reverse transcription master mix on ice during ligation: 9  $\mu$ l nuclease-free water, 2  $\mu$ l 10 mM dNTPs, 8  $\mu$ l 5x SuperScript IV first strand buffer, 4  $\mu$ l 0.1 mM DTT.
3. Add the reverse transcription master mix to the ligation reaction and mix by pipetting. Add 2  $\mu$ l SuperScript IV reverse transcriptase and mix by pipetting. Incubate in a thermocycler with the following protocol: 50 min at 50 °C, 10 min at 70 °C, cool down to 4 °C.
4. Let the Agencourt RNAClean XP beads come to room temperature during reverse transcription. Carefully resuspend beads before use. Transfer reaction to a 1.5 ml DNA LoBind tube and mix with 72  $\mu$ l Agencourt RNAClean XP beads. Incubate 5 min at room temperature on a gentle rotator mixer.
5. Collect beads on a magnetic stand and remove supernatant. Wash pelleted beads two times (30 sec) with 200  $\mu$ l freshly prepared 70 % ethanol. Remove supernatant. Spin sample down and place on magnet again. Remove any residual ethanol.

- 214 6. Resuspend beads in 20  $\mu$ l nuclease-free water by gentle flicking and  
215 incubate 5 min at room temperature on a gentle rotator mixer. Collect  
216 beads on a magnetic stand and transfer 20  $\mu$ l eluate in a fresh 1.5 ml  
217 DNA LoBind tube.
- 218 7. For ligation of the RMX adapter, add the following to 20  $\mu$ l eluate: 8  
219  $\mu$ l NEBNext Quick Ligation Reaction Buffer, 6  $\mu$ l RMX (from SQK-  
220 RNA002), 3  $\mu$ l nuclease-free water, 3  $\mu$ l T4 DNA Ligase. Mix by  
221 pipetting and incubate 10 min at room temperature.
- 222 8. Add 40  $\mu$ l carefully resuspended Agencourt RNAClean XP beads to  
223 the reaction and mix by pipetting. Incubate 5 min at room tempera-  
224 ture on a gentle rotator mixer.
- 225 9. Collect beads on a magnetic stand and remove supernatant. Wash  
226 pelleted beads two times with 150  $\mu$ l wash buffer (WSB, from SQK-  
227 RNA002). Resuspend beads by flicking, spin down and return to mag-  
228 netic stand. Remove supernatant from pelleted beads.
- 229 10. Resuspend beads in 21  $\mu$ l elution buffer (EB, from SQK-RNA002) by  
230 gentle flicking and incubate 5 min at room temperature on a gentle  
231 rotator mixer. Pellet beads on a magnetic stand and transfer 21  $\mu$ l  
232 eluate in a fresh 1.5 ml DNA LoBind tube.
- 233 11. Quantify 1  $\mu$ l of the library on a Qubit fluorometer with the Qubit  
234 dsDNA HS kit according to the manufacturerers protocol. Concentra-  
235 tion should be usually in the range of 5 - 10 ng/ $\mu$ l.
- 236 12. Insert MinION R9.4.1 Flow cell in the MinION or GridION sequenc-  
237 ing device and perform Flow cell check in the MinKNOW software.  
238 For successful sequencing of mammalian polyA<sup>+</sup> RNA at least 1,000  
239 available pores are recommended.
- 240 13. Prepare Priming Mix by adding 30  $\mu$ l flush tether (FLT, from EXP-  
241 FLP002) to a vial of flush buffer (FB, from EXP-FLP002) and mix by  
242 pipetting. Open priming port. Remove air bubble from priming port  
243 by inserting the tip of a P1000 pipette into the priming port and slowly  
244 dialing up, until a small volume of storage buffer enters the pipette  
245 tip. Load 800  $\mu$ l Priming Mix via the priming port and carefully avoid  
246 introduction of air bubbles. Close the priming port and wait for 5 min.
- 247 14. Mix 20  $\mu$ l library with 17.5  $\mu$ l nuclease-free water and 37.5  $\mu$ l RNA run-  
248 ning buffer (RRB, from SQK-RNA002) and mix by pipetting. Open  
249 the priming port and the sample port. Load 200  $\mu$ l Priming Mix via  
250 the priming port. Mix library by pipetting just before loading and  
251 load dropwise via the sample port. Carefully avoid introduction of air  
252 bubbles. Close the sample port and the priming port.

253 15. Start sequencing for 48 to 72 h in the MinKNOW software. Choose  
254 direct RNA-sequencing kit and high-accuracy basecalling as param-  
255 eters.

## 256 Preparation of an *in vitro* transcriptome sample

257 The *in vitro* transcriptome sample is prepared based on a protocol published  
258 by Zhang *et al.* Zhang et al. [2021] with some modifications.

- 259 1. Adjust 100 ng polyA<sup>+</sup> RNA to a total volume of 6  $\mu$ l with nuclease-  
260 free water. Add 1  $\mu$ l each of 10  $\mu$ M oligo(dT)-VN RT primer and 10  
261 mM dNTPs. Mix by pipetting and incubate in a thermocycler: 5 min  
262 at 75 °C, 2 min at 42 °C, cool to 4 °C.
- 263 2. Assemble 2.5  $\mu$ l 4x template switching RT buffer, 0.5  $\mu$ l 20  $\mu$ M TSO,  
264 1  $\mu$ l 10x template switching RT enzyme mix and mix by pipetting.  
265 Combine with 6  $\mu$ l RNA and incubate in a thermocycler: 90 min at  
266 42 °C, 10 min at 68 °C, cool to 4 °C.
- 267 3. For Second strand synthesis add to First strand synthesis reaction: 50  
268  $\mu$ l Q5 Hot Start High-Fidelity 2X Master Mix, 5  $\mu$ l RNase H, 2  $\mu$ l 10  
269  $\mu$ M T7 extension primer, 33  $\mu$ l nuclease-free water. Mix by pipetting  
270 and incubate in a thermocycler: 15 min at 37 °C, 1 min at 95 °C, 10  
271 min at 65 °C, cool to 4 °C.
- 272 4. Purify double stranded cDNA with NucleoSpin Gel and PCR Clean-up  
273 kit according to the manufacturerers protocol and elute in 20  $\mu$ l elution  
274 buffer. Determine concentration on a Nanodrop spectrophotometer.  
275 cDNA may be stored at -20 °C.
- 276 5. Combine 8  $\mu$ l cDNA for *in vitro* transcription with 2  $\mu$ l each of ATP,  
277 GTP, CTP, UTP, 10x reaction buffer and enzyme mix from the MEGAscript  
278 T7 transcription kit. Incubate 3 h at 37 °C.
- 279 6. Digest template DNA by addition of 1  $\mu$ l Turbo DNase. Mix by pipet-  
280 ting and incubate 15 min at 37 °C.
- 281 7. Adjust reaction volume to 100  $\mu$ l with nuclease-free water and clean up  
282 with RNA Clean & Concentrator-25 kit according to the manufactur-  
283 ers protocol, using two volumes of adjusted RNA binding buffer (1:1  
284 RNA binding buffer : ethanol). Elute RNA in 25  $\mu$ l nuclease-free wa-  
285 ter. Determine RNA concentration on a Nanodrop spectrophotometer.  
286 Store at -80 °C.



## 287 Nanopore read processing

- 288 1. Base call the ionic current signal stored in FAST5 files using Guppy.  
289 For the IVT sample, we applied real-time base calling with the MinKNOW-  
290 embedded Guppy basecaller. Otherwise, Guppy basecaller software  
291 can be used. In this case, the basecaller requires the path to FAST5  
292 files, the output folder, and the config file or the flowcell/kit combina-  
293 tion. The output are FASTQ files that can be compressed using the  
294 option “--compress\_fastq”.

```
295 $ guppy_basecaller --compress_fastq -i path_to_fast5 -s path_to_output  
296 -c config_file.cfg --cpu_threads_per_caller 14 --num_callers  
297 1
```

298 Set the number of threads “cpu\_threads\_per\_caller” and the number  
299 of parallel basecallers “num\_caller” according to your resources. Ad-  
300 ditional details can be found in Gup [2019].

- 301 2. Align reads to the transcriptome using Minimap2 software. The out-  
302 put is a SAM file that has to be converted to a compressed form as  
303 BAM file using SAMtools command. The alignment requires a ref-  
304 erence sequence. Here, we used GRCh38 Ensembl annotation and  
305 FASTA file release version 96. **To reduce the indexing time of the**  
306 **human genome, save the index with the option “-d” before the map-**  
307 **ping and use the index instead of the reference file in the minimap2**  
308 **command line.**

```
309 $ minimap2 -d reference.mmi reference.fa
```

310 **To enable spliced alignments, use the setting “-ax splice -junc-bed**  
311 **annotation.bed -junc-bonus” where “-junc-bonus” allows to tune the**  
312 **bonus score and the BED file “-junc-bed annotation.bed” provides the**  
313 **splice junctions. The BED file can be generated using the following**  
314 **command:**

```
315 $pafutils.js gff2bed annotation.gtf > annotation.bed
```

316 **Use “-ub” to allow alignment to both strands or ‘-uf’ to force the**  
317 **alignment to only forward strand. For Direct RNA Sequencing, it is**  
318 **recommended to set a small k-mer size “-k [=14]” to enhance sensitiv-**  
319 **ity. We recommend outputting primary alignments “-secondary=no”.**  
320 **Use the parameter ‘-MD’ to add the reference sequence information**  
321 **to the alignment; this is recommended for the downstream analysis.**  
322 **Customize the number of threads “-t” according to your resources.**  
323 **Check Minimap2 manual for more details [Min].**

```

324 $ minimap2 -t 5 --MD -ax splice --junc-bonus 1 -k14 --secondary=no
325 --junc-bed final_annotation_96.bed -ub reference.mmi Reads.fastq.gz
326 |samtools view -bS > mapping.bam

```

3. Map RNA modifications using JACUSA2 pipeline. JACUSA2 [Piechotta et al., 2021] rapidly detects RNA modifications based on a comparative strategy where the mapping features (mismatch, insertion and deletion) of a sample of interest are compared to a reference sequence (call-1) or against a sample without RNA modifications, e.g. a knock-out of an RNA modifying enzyme or an IVT (call-2). Moreover, it allows the integration of information from replicate experiments. **The output of JACUSA2 variant calling is a set of scores reflecting the read signatures involving mismatch, insertion and deletion. The analysis of read signature can be used for RNA modification detection. We integrate JACUSA2, in particular call-2 method, with the downstream analysis in one pipeline using the Python-based workflow management system Snakemake [Köster and Rahmann, 2012]. The Snakemake pipeline involves rules for the variant calling using JACUSA2 call-2, detection of RNA modification patterns, prediction of new modified sites and other intermediate rules as shown in Figure 4. The input of the pipeline are BAM files from paired conditions with different replicates. BAM files need to be sorted and may be subjected to many filters before being used by JACUSA2 call2 rule. Here, we suggest to filter out secondary and poor alignments. The output of JACUSA2 call2 is preprocessed (get\_features) and subjected to a learning process to extract and visualize modification patterns (resp. get\_pattern, visualize\_pattern) and make predictions (predict\_modification). "split\_train\_test" rule allows splitting input data into a training set and a test set. To use our snakemake-based JACUSA2 pipeline a set of parameters should be defined in the "config.yaml" file; mainly: the label of the analysis 'label', the input bam files under 'data', the reference sequence 'reference', a file containing size of chromosomes 'chr\_size', JACUSA2 jar file 'jar', plus the path to inputs and outputs under 'path\_inp' and 'path\_out' fields respectively. Further details on how to use JACUSA2 pipeline is presented within the use cases in the next section. The pipeline could be executed on a high-performance-computing cluster (HPC) using the following command by specifying the number of cores to be used "-cores [=all]" and the rule name:**

```

361 $ srun snakemake --cores all rule_name
362 Check Snakemake documentation for more details [sna].

```

## 363 Use Case 1: Comparison of wild-type and knock-out samples

364 The JACUSA2 workflow detects RNA modifications using direct RNA se-  
365 quencing by comparing a modified sample to an unmodified control sample.  
366 Here, we used a published dataset of HEK293 cell lines to map m6A modifi-  
367 cation [Pratanwanich et al., 2021]. The benchmark is composed of samples  
368 sets two conditions: wild-type cells (WT, modified RNAs) and Mettl3 knock-  
369 out cells (KO, unmodified RNAs) in two replicates (2 and 3). The FASTQ  
370 files are mapped using Minimap2 as described in the previous section. The  
371 following analysis is validated against m6A sites consistently reported in  
372 three miCLIP-based studies Boulias et al. [2019], Koh et al. [2019], Körtel  
373 et al. [2021] (Figure 5).

374 Starting with the preprocessed mapped reads as inputs (BAM files),  
375 'HEK293T-WT-rep2.bam' and 'HEK293T-WT-rep3.bam' represent the wild-  
376 type replicates and 'HEK293T-KO-rep2.bam' and 'HEK293T-KO-rep3.bam'  
377 the control replicates,

- 378 1. Identify read error profile: **use "jacusa2\_call2" rule to run JACUSA2**  
379 in pairwise condition mode (call-2). The method requires BAM files of  
380 the paired conditions and the corresponding library information "-P1"  
381 and "-P2". In addition to the mismatch score, add "-D" and "-I" to  
382 output the deletion and insertion scores. JACUSA2 allows filtering  
383 reads according to many parameters. Here, we consider all sites with  
384 base calling quality "-q [> 1]", mapping quality "-m [> 1]" and read  
385 coverage "-c [> 4]". Furthermore, it provides a filter feature to improve  
386 sensitivity. Here, **we consider filtering sites within homopolymer re-**  
387 **gions "-a [=Y]". The output (named here, "Cond1vsCond2Call2.out")**  
388 consists of a read error profile where the format is a combination  
389 of BED6 with JACUSA2 call-2 specific columns and common info  
390 columns: info, filter, and ref. Check JACUSA2 manual for more de-  
391 tails on JACUSA2 filter and output options [JAC, 2021]. The number  
392 of threads can be customized via the parameter "-p". **All parameters**  
393 **related to the JACUSA2 method can be added under the field "ja-**  
394 **cusa\_params" in the config file by setting the name of the parameter**  
395 **followed by the corresponding value [key: value]. Be aware to set all**  
396 **parameters before running the pipeline.**

```
397 $ srun snakemake --cores all jacusa2_call2 $
```

- 398 2. Preprocess JACUSA2 output: from JACUSA2 call-2 output, **we select**  
399 all sites within 5-mer of a central nucleotide 'A' flanked by 2 random  
400 nucleotides (NNANN) and **we filter out sites of the homo-polymer re-**  
401 **gions (JACUSA filter: Y). Then, we rebuild the tabular features such**  
402 **that the observations are only sites with a reference base 'A'. Each**  
403 **site is characterized by 15 features corresponding to the mismatch,**

404 insertion and deletion scores for the observed site and its two flank-  
 405 ing positions from both sides. The rule "get\_features" performs the  
 406 preprocessing step. Use the parameter 'region' with a file containing  
 407 target 5-mers to limit the analysis to specific sites. For comparison  
 408 reasons, we consider common sites between use cases 1 and 2 . The  
 409 output is an R object "features/features.rds", representing the matrix  
 410 of Sites $\times$ 15 features.

411 `$ srun snakemake --cores all get_features`

412 3. Extract m6A modification pattern: given the matrix of Sites $\times$ Features,  
 413 the next step is to learn a model representing the m6A modification  
 414 pattern. To this end, the conventional non-negative matrix factor-  
 415 ization (NMF) analysis is suggested [Lee and Seung, 1999]. Briefly,  
 416 NMF factorizes a non-negative data matrix  $X$  (here:  $n$  sites and  $m$   
 417 features) into two non-negative matrices as  $X \approx WH$ , such that  $W$   
 418 is an  $n \times k$  matrix containing basis vectors and  $H$  is an  $k \times m$  ma-  
 419 trix containing coefficient vectors. The coefficient vectors and their  
 420 combination can be viewed as a pattern for m6A modification. The  
 421 rank of factorization  $k$  is a critical parameter that affects the perfor-  
 422 mance substantially. We suggest to select the rank  $k$  according to  
 423 the method of Frigyesi and Höglund [2008] by looking at silhouette  
 424 [Rousseeuw, 1987] and cophenetic correlation [Brunet et al., 2004] in-  
 425 dices. Accordingly, the performance indices are computed for different  
 426 choices of rank ( $k < n, m$ ) and compared to the performance of a ran-  
 427 dom permutation of the original data. Subsequently, the chosen rank  
 428 corresponds to the value with the largest difference between slopes of  
 429 the original and the randomized data. Here, the unsupervised pattern  
 430 training is based on the consensus set of 1,905 m6A sites reported  
 431 in the three miCLIP-based studies mentioned earlier. Based on the  
 432 silhouette and cophenetic correlation indices, we identified an optimal  
 433 factorization rank of 6 (Figure 6A). We then analyzed the identified  
 434 patterns. According to the membership indicator of each site in ma-  
 435 trix  $W$ , more than 80% of m6A modification sites can be represented  
 436 by five patterns (Patterns 1,2,3,4,6) (Figure 6B). Interestingly, the  
 437 linear combination of these five patterns in Figure 6C highlights the  
 438 importance of position 3 and eventually the implication of all scores.

439 Using the JACUSA2 pipeline, run rule "get\_pattern" to generate pat-  
 440 terns and provide the set of modified sites as a ground truth under the  
 441 field "modified\_sites" in the config file. Here, the "miCLIP\_union.bed"  
 442 file contains the m6A sites from the three miCLIP-based studies. A  
 443 miCLIP annotation, reflecting the consensus sites, is added to each  
 444 site. A subset of modified sites can be used to generate patterns. Ac-  
 445 cordingly, the "internal\_pattern" field should refer to the annotation

Is this  
the  
reason  
why you  
chose  
to work  
on the  
three  
outputs  
together  
WT\_IV, WT\_KO,  
KO\_IVT

in Fig-  
ure 6C  
this is  
labeled  
sum

446 of selected sites from the "modified\_sites" file. Plus, multiple combi-  
447 nations of patterns can be defined and appended to the field "com-  
448 bined\_pattern" as new patterns. The corresponding outputs are under  
449 "patterns" folder.

```
450 $ srun snakemake --cores all get_pattern
```

451 The produced patterns and their combinations can be visualized using  
452 "visualize\_pattern" rule. The corresponding outputs are under "pat-  
453 tern/viz" folder.

```
454 $ srun snakemake --cores all visualize_pattern
```

455 4. Predict m6A modifications: the additive linear combination of the co-  
456 efficient vectors (patterns) with the 15 features can be used to predict  
457 m6A modification. We examine the ability of prediction on a subset of  
458 data of more than 1,52 million sites with 17,021 miCLIP m6A sites.  
459 We opt for the linear combination of the five most relevant patterns  
460 described in step 3. The empirical Cumulative Distribution Function  
461 (eCDF) of the inferred scores shows a significant difference between  
462 the different miCLIP m6A categories (miCLIP annotation) and the  
463 unmodified sites (Figure 6D). As the number of negative samples is  
464 much larger than the number of positive samples, we particularly rec-  
465 ommend investigating the Positive Predictive Value (PPV) of the pre-  
466 dictions. Here, Figure 6E shows a moderate PPV that increases with  
467 the cut-off.

468 To perform the prediction based on the selected patterns using the  
469 JACUSA2 pipeline, run rule "predict\_modification". The patterns  
470 can be generated from a subset of the input data according to the  
471 field "internal\_pattern" or predefined patterns indicated in the "exter-  
472 nal\_pattern" field. The output is a BED file containing the estimated  
473 scores as well as the corresponding eCDF and PPV plots. The corre-  
474 sponding outputs are located under a new folder called "prediction".  
475

```
476 $ srun snakemake --cores all predict_modification
```

## 477 Use Case 2: Comparison of wild-type and IVT samples

478 An alternative way to detect RNA modifications is to compare a modi-  
479 fied sample to an *in-vitro* transcribed (IVT) control sample. Therefore,  
480 we benchmark JACUSA2 on a sample set of two replicates (2 and 3) from  
481 wild-type HEK293 cell lines (modified sample) Pratanwanich et al. [2021]  
482 and a modification-free IVT sample from HEK293 cDNA (control sample)  
483 (see "Preparation of an *in vitro* transcriptome sample"). The analysis steps

484 are similar to case 1. We evaluate the analysis against miCLIP m6A sites  
485 (Figure 5).

486 1. Identify read error profile: we use JACUSA2 call-2 with the same  
487 parameters as the previously described case. The input BAM files  
488 (HEK293T-WT-rep2.bam, HEK293T-WT-rep3.bam) and (HEK293T-  
489 IVT-rep1.bam, HEK293T-IVT-rep2.bam) are associated to the wild-  
490 type and IVT replicate samples respectively.

491 `$ srun snakemake --cores all jacusa2_call2`

492 2. Preprocess JACUSA2 output: we select all sites within the specific 5-  
493 mer (NNANN) and we consider the Y filter that excludes sites within  
494 homo-polymer regions. Then, we extract 5-mer features such that the  
495 selected sites are represented by the Mismatch, Deletion and Insertion  
496 scores for the observed site and its two flanking positions from both  
497 sides.

498 `$ srun snakemake --cores all get_features`

499 3. Extract m6A modification pattern: using NMF factorization, we ex-  
500 tract patterns from the 1,905 sites reported as modified in the three  
501 miCLIP-based studies. Based on the silhouette and cophenetic corre-  
502 lation indices, we identified an optimal factorization rank of 6 (Figure  
503 7A). We determined the predominant factors from matrix  $W$ . Accord-  
504 ingly, more than 80% of m6A modification sites can be represented by  
505 four patterns (Patterns: 1,2,3,6) (Figure 7B). In agreement with Use  
506 Case 1, the linear combination of the four patterns confirms the im-  
507 portance of position 3 and the implication of all scores as shown in  
508 Figure 7C.

509 `$ srun snakemake --cores all get_pattern`

510 4. Predict m6A modifications: we evaluate the prediction ability of the  
511 detected patterns on a test set of almost 1,52 million sites where  
512 17,021 are miCLIP-m6A modified. We consider the linear combina-  
513 tion of the four most relevant patterns (1,2,3,6). Figure 7D shows the  
514 eCDF of the inferred scores. The difference between the cumulative  
515 distribution of non miCLIP sites and miCLIP sites can be nicely ob-  
516 served, while the PPV plot shows a lower performance as compared  
517 to Use Case 1 (Figure 7E). The decrease in performance is likely ex-  
518 plained by the absence of all modifications and not exclusively m6A in  
519 the control condition, which may induce noise to the score estimation  
520 by JACUSA2 call-2.

521 `$ srun snakemake --cores all predict_modification`

The first IVT run has rel. low coverage  $\rightarrow$  might this impact performance of UC2?

CD: to be confirmed

## NOTES

### Tips and Tricks

1. The reverse transcription step during library preparation is optional. However, we recommend to include this step to ensure proper sequencing also of RNAs with secondary structures. Superscript IV reverse transcriptase may be replaced by Superscript III reverse transcriptase, which is used in the protocol provided by Oxford Nanopore Technologies.
2. The library preparation protocol contains two bead clean up steps. It is important to remove ethanol and wash buffer completely. However, beads should not be dried for several minutes. Directly add water or elution buffer after washing to prevent sticking of the RNA to the beads.
3. The default filter in current MinKNOW versions is a Q score of 9. For direct RNA sequencing we recommend to adjust the output filter to a minimum Q score of 7, as in previous MinKNOW versions.
4. During preparation of the *in vitro* transcriptome sample, *in vitro* transcription and clean up kits may be replaced by equivalent products. The protocol however has been tested only with the mentioned kits.
5. Configuration of the pipeline should be handled via the config file. All parameters should be set before executing rules.
6. Once the pipeline has run successfully you should expect the following folders with the corresponding outputs in the output directory: bam, jacusa, features, patterns, and prediction.
7. JACUSA2 call2 could be run separately using the command line as described in JACUSA2 manual [JAC, 2021], then put the output under a new folder with the name 'jacusa' under the output directory.
8. In the snakemake pipeline, rules are linked so that the workflows are determined from top (e.g. predict\_modification) to bottom (e.g. sort\_bam) and executed accordingly from bottom to top (Figure 4). Therefore, running for example "predict\_modification" rule leads to executing all rules on its pipeline.
9. Patterns could be generated from a subset of the input data that correspond to known modified sites. Alternatively, predefined patterns as a NMF R object could be used as a prediction model.

## 557 ACKNOWLEDGMENTS

558 The authors would like to thank Harald Wilhemit for testing the snakemake  
559 pipeline. This work was supported by Informatics for Life funded by the  
560 Klaus Tschira Foundation.

CD:  
fund-  
ing?

## 561 REFERENCES

- 562 Minimap2. <https://github.com/lh3/minimap2>. Accessed: 2022-01-19.
- 563 Snakemake. <https://snakemake.readthedocs.io>. Accessed: 2022-01-26.
- 564 Basecalling with guppy. [https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/](https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/basecalling.rst)  
565 [basecalling.rst](https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/basecalling.rst), 2019. Accessed: 2022-01-19.
- 567 Jacusa2 manual. <https://github.com/dieterich-lab/JACUSA2>, 2021.  
568 Accessed: 2022-01-15.
- 569 Samir Adhikari, Wen Xiao, Yong-Liang Zhao, and Yun-Gui Yang. m(6)a:  
570 Signaling for mrna splicing. *RNA biology*, 13:756–759, September 2016.  
571 ISSN 1555-8584. doi: 10.1080/15476286.2016.1201628.
- 572 Ina Anreiter, Quoseena Mir, Jared T. Simpson, Sarath C. Janga, and  
573 Matthias Soller. New twists in detecting mrna modification dynamics.  
574 *Trends in biotechnology*, 39:72–89, January 2021. ISSN 1879-3096. doi:  
575 10.1016/j.tibtech.2020.06.002.
- 576 Konstantinos Boulas, Diana Toczydlowska-Socha, Ben R Hawley, Noa  
577 Liberman, Ken Takashima, Sara Zaccara, Théo Guez, Jean-Jacques  
578 Vasseur, Françoise Debart, L Aravind, et al. Identification of the m6am  
579 methyltransferase pcif1 reveals the location and functions of m6am in the  
580 transcriptome. *Molecular cell*, 75(3):631–643, 2019.
- 581 Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov.  
582 Metagenes and molecular pattern discovery using matrix factorization.  
583 *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.
- 584 Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali  
585 Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine  
586 Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and  
587 Gideon Rechavi. Topology of the human and mouse m6a rna methylomes  
588 revealed by m6a-seq. *Nature*, 485:201–206, April 2012. ISSN 1476-4687.  
589 doi: 10.1038/nature11112.
- 590 Hao Du, Ya Zhao, Jinqiu He, Yao Zhang, Hairui Xi, Mofang Liu, Jinbiao  
591 Ma, and Ligang Wu. Ythdf2 destabilizes m 6 a-containing rna through



592 direct recruitment of the ccr4–not deadenylase complex. *Nature commu-*  
593 *nications*, 7(1):1–11, 2016.

594 Attila Frigyesi and Mattias Höglund. Non-negative matrix factorization for  
595 the analysis of complex gene expression data: identification of clinically  
596 relevant tumor subtypes. *Cancer informatics*, 6:CIN–S606, 2008.

597 David Garcias Morales and José L. Reyes. A birds’-eye view of the activ-  
598 ity and specificity of the mrna m, javax.xml.bind.jaxbelement@6d66739e,  
599 a methyltransferase complex. *Wiley interdisciplinary reviews. RNA*, 12:  
600 e1618, January 2021. ISSN 1757-7012. doi: 10.1002/wrna.1618.

601 Guifang Jia, Ye Fu, Xu Zhao, Qing Dai, Guanqun Zheng, Ying Yang,  
602 Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, and Chuan He.  
603 N6-methyladenosine in nuclear rna is a major substrate of the obesity-  
604 associated fto. *Nature chemical biology*, 7:885–887, October 2011. ISSN  
605 1552-4469. doi: 10.1038/nchembio.687.

606 Shengdong Ke, Endalkachew A. Alemu, Claudia Mertens, Emily Conn Gant-  
607 man, John J. Fak, Aldo Mele, Bhagwattie Haripal, Ilana Zucker-Scharff,  
608 Michael J. Moore, Christopher Y. Park, Cathrine Broberg Vågbø, Anna  
609 Kussnierczyk, Arne Klungland, James E. Darnell, and Robert B. Darnell.  
610 A majority of m6a residues are in the last exons, allowing the potential  
611 for 3’ utr regulation. *Genes & development*, 29:2037–2053, October 2015.  
612 ISSN 1549-5477. doi: 10.1101/gad.269415.115.

613 Casslynn WQ Koh, Yeek Teck Goh, and WS Sho Goh. Atlas of quantitative  
614 single-base-resolution n 6-methyl-adenine methylomes. *Nature communi-*  
615 *cations*, 10(1):1–15, 2019.

616 Nadine Körtel, Cornelia Rücklé, You Zhou, Anke Busch, Peter Hoch-Kraft,  
617 Reymond FX Sutandy, Jacob Haase, Mihika Pradhan, Michael Musheev,  
618 Dirk Ostareck, et al. Deep and accurate detection of m6a rna modifications  
619 using miclip2 and m6aboost machine learning. *bioRxiv*, pages 2020–12,  
620 2021.

621 Johannes Köster and Sven Rahmann. Snakemake—a scalable bioinformatics  
622 workflow engine. *Bioinformatics*, 28(19):2520–2522, 2012.

623 Daniel D Lee and H Sebastian Seung. Learning the parts of objects by  
624 non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

625 Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christo-  
626 pher E. Mason, and Samie R. Jaffrey. Comprehensive analysis of mrna  
627 methylation reveals enrichment in 3’ utrs and near stop codons. *Cell*, 149:  
628 1635–1646, June 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.05.003.

629 Deepak P Patil, Brian F Pickering, and Samie R Jaffrey. Reading m6a in  
630 the transcriptome: m6a-binding proteins. *Trends in cell biology*, 28(2):  
631 113–127, 2018.

632 Michael Piechotta, Qi Wang, Janine Altmüller, and Christoph Dieterich.  
633 Rna modification mapping with jacusa2. *bioRxiv*, 2021.

634 Ploy N Pratanwanich, Fei Yao, Ying Chen, Casslynn WQ Koh, Yuk Kei  
635 Wan, Christopher Hendra, Polly Poon, Yeek Teck Goh, Phoebe ML Yap,  
636 Jing Yuan Chooi, et al. Identification of differential rna modifications  
637 from nanopore direct rna sequencing with xpore. *Nature Biotechnology*,  
638 39(11):1394–1402, 2021.

639 Jean-Yves Roignant and Matthias Soller. m,  
640 javax.xml.bind.jaxbelement@8cec19d, a in mrna: An ancient mech-  
641 anism for fine-tuning gene expression. *Trends in genetics : TIG*, 33:  
642 380–390, June 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2017.04.003.

643 Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic rna  
644 modifications in gene expression regulation. *Cell*, 169:1187–1200, June  
645 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.045.

646 Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and  
647 validation of cluster analysis. *Journal of computational and applied math-*  
648 *ematics*, 20:53–65, 1987.

649 Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how:  
650 Context-dependent functions of rna methylation writers, readers, and  
651 erasers. *Molecular cell*, 74:640–650, May 2019. ISSN 1097-4164. doi:  
652 10.1016/j.molcel.2019.04.025.

653 Xiao Wang, Zhike Lu, Adrian Gomez, Gary C Hon, Yanan Yue, Dali Han,  
654 Ye Fu, Marc Parisien, Qing Dai, Guifang Jia, et al. N 6-methyladenosine-  
655 dependent regulation of messenger rna stability. *Nature*, 505(7481):117–  
656 120, 2014.

657 Xiao Wang, Boxuan Simen Zhao, Ian A Roundtree, Zhike Lu, Dali Han,  
658 Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He.  
659 N6-methyladenosine modulates messenger rna translation efficiency. *Cell*,  
660 161(6):1388–1399, 2015.

661 Sara Zaccara, Ryan J. Ries, and Samie R. Jaffrey. Reading, writing and  
662 erasing mrna methylation. *Nature reviews. Molecular cell biology*, 20:608–  
663 624, October 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0168-5.

664 Zhang Zhang, Tao Chen, Hong-Xuan Chen, Ying-Yuan Xie, Li-Qian Chen,  
665 Yu-Li Zhao, Biao-Di Liu, Lingmei Jin, Wutong Zhang, Chang Liu,

666 et al. Systematic calibration of epitranscriptomic maps using a synthetic  
 667 modification-free rna library. *Nature Methods*, 18(10):1213–1222, 2021.

668 Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min  
 669 Huang, Charles J. Li, Cathrine B. Vågbø, Yue Shi, Wen-Ling Wang, Shu-  
 670 Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin  
 671 Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan,  
 672 Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne  
 673 Klungland, Yun-Gui Yang, and Chuan He. Alkbh5 is a mammalian rna  
 674 demethylase that impacts rna metabolism and mouse fertility. *Molecular*  
 675 *cell*, 49:18–29, January 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.  
 676 10.015.

## FIGURE CAPTIONS

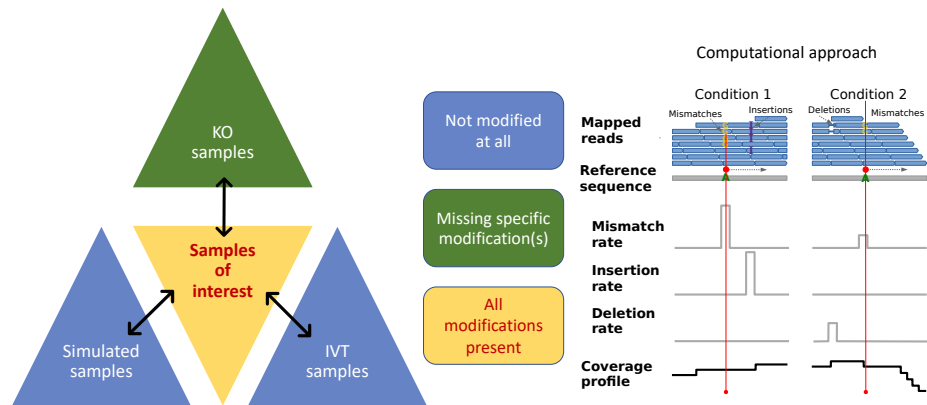
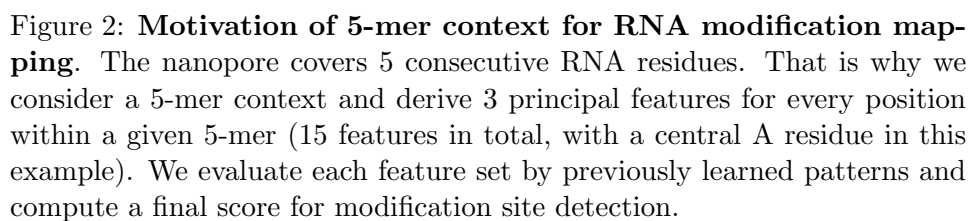


Figure 1: **General outline of RNA modification detection by JACUSA2.** A key feature of our approach is that multiple replicates can be compared as shown on the left. Samples of interests where all modifications are present could be compared with either KO samples where the modification of interest is missing or IVT/simulated samples where all modifications are absent. Read stacks (in blue) are compared head-to-head as shown on the right.



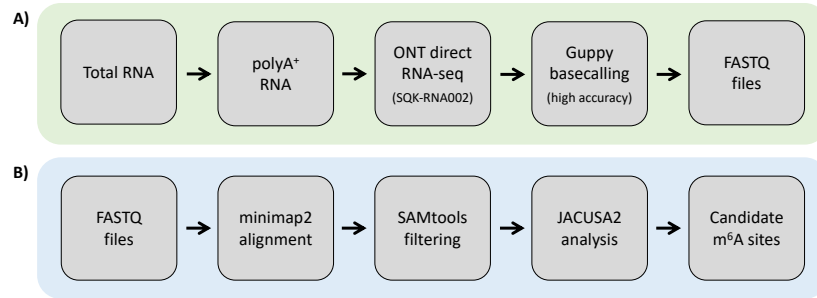


Figure 3: **Experimental and computational workflow.** A) Starting from total cellular RNA, polyA<sup>+</sup> RNA is isolated and subjected to Nanopore direct RNA-sequencing. Guppy basecalling can be done as live basecalling during sequencing or after the sequencing run from generated FAST5 files, resulting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

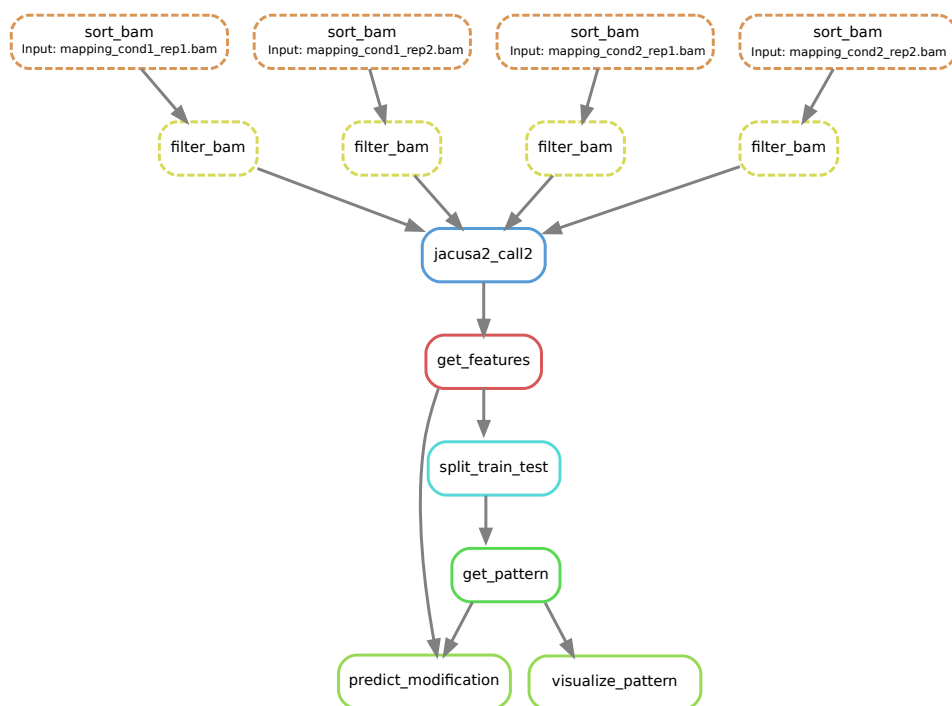


Figure 4: **Computational workflow.** Snakemake workflow for RNA modification detection based on JACUSA2 variant calling.

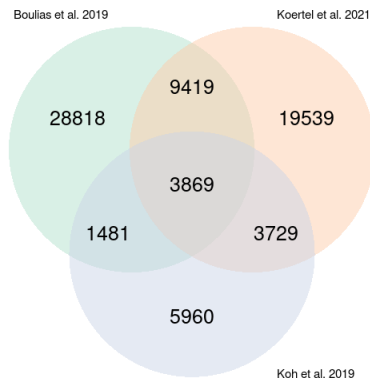


Figure 5: **m6A sites reported in the three miCLIP-based studies** Boulias et al. [2019], Koh et al. [2019] and Körtel et al. [2021].



Software	Version	Description
Minimap2	<a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a> v2.22 or later	<a href="https://lh3.github.io/minimap2/">https://lh3.github.io/minimap2/</a>
samtools	<a href="https://github.com/samtools/samtools">https://github.com/samtools/samtools</a> v1.12 or later	<a href="http://samtools.github.io/">http://samtools.github.io/</a>
JAVA	openjdk 11.0.12 2021-07-20 - JAVA 11 or later	OpenJDK Runtime Environment
R	<a href="https://www.r-project.org/">https://www.r-project.org/</a> version 3.5.1 or later	The R Project for Statistical Computing
PERL	<a href="https://www.perl.org/">https://www.perl.org/</a> version 5.28.1 or later	Perl is a highly capable, feature-rich programming language
BASH, sed, awk	should be part of your Linux distribution	Misc.
bedtools	<a href="https://github.com/arq5x/bedtools2">https://github.com/arq5x/bedtools2</a> version 2.29.2 or later	Perl is a highly capable, feature-rich programming language
NanoSim	<a href="https://github.com/bcgsc/NanoSim">https://github.com/bcgsc/NanoSim</a> version 3.0.2 or later (optional)	NanoSim is a fast and scalable read simulator that captures the technology-specific features of ONT data

Table 1: **Software dependencies** blubba

## 678 TABLE CAPTIONS

## 679 TABLES

R Pack- ages	Version	Description
ggplot2	<a href="https://cran.r-project.org/web/packages/ggplot2/index.html">https://cran.r-project.org/web/packages/ggplot2/index.html</a> - ggplot2_3.3.0 or later	ggplot2 is a system for declaratively creating graphics, based on The Grammar of Graphics.
NMF	<a href="https://cran.r-project.org/web/packages/NMF/index.html">https://cran.r-project.org/web/packages/NMF/index.html</a> - NMF_0.22.0 or later	Provides a framework to perform Non-negative Matrix Factorization (NMF).

Table 2: **R Package dependencies** blubba

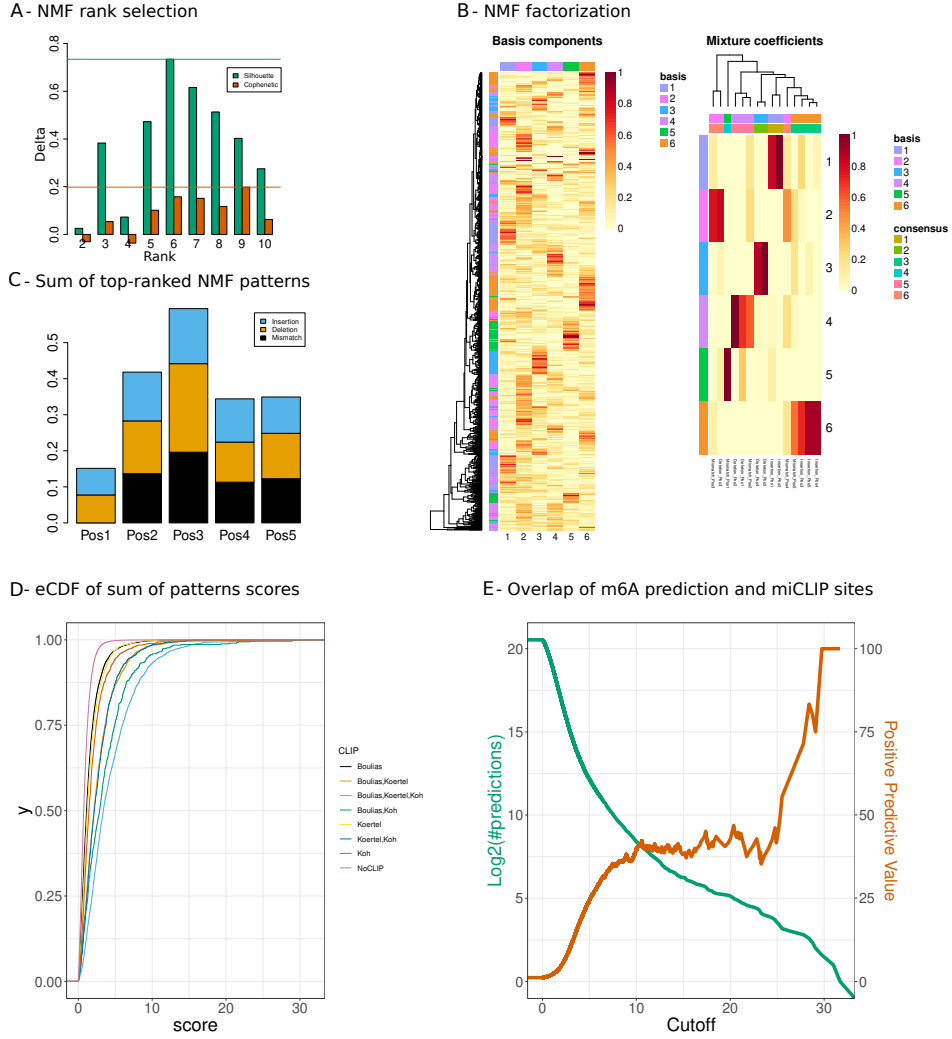


Figure 6: **Case 1. WT versus KO.** **A:** NMF rank selection. Barplots representing the difference of cophenetic correlation and silhouette indices between the NMF factorization of the original and the randomized data. Ranks with the largest value for both indices are determined, then the smallest rank is selected for the NMF decomposition. **B:** NMF result represented by the basis matrix  $W$  and the coefficient matrix  $H$ . The matrix  $H$  induces the RNA modification pattern. **C:** Barplots representing the linear combination of the top 5 patterns (y-axis) by the number of position in the specific 5-mer context (x-axis) and the score type: mismatch, deletion and insertion (resp. black, orange and blue). The 5 patterns (coefficient vectors: 1,2,3,4,6) are selected according to the predominant columns in matrix  $W$ . **D:** Score distribution inferred from the combined patterns, stratified by the different 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).

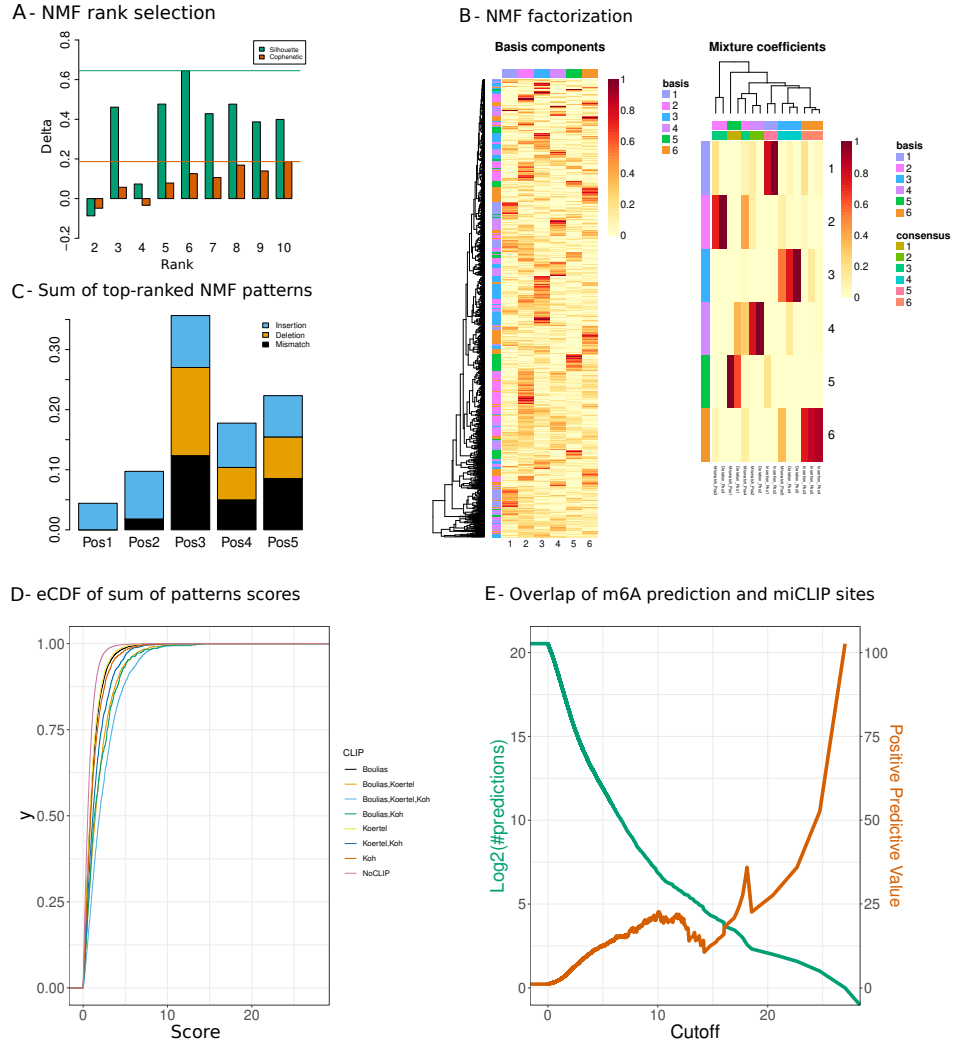


Figure 7: **Case 2. WT versus IVT.** **A:** NMF rank selection. Barplots representing the difference of cophenetic correlation and silhouette indices between the NMF factorization of the original and the randomized data. Ranks with the largest value for both indices are determined, then the smallest rank is selected for the NMF decomposition. **B:** NMF result represented by the basis matrix  $W$  and the coefficient matrix  $H$ . The matrix  $H$  induces the RNA modification pattern. **C:** Barplots representing the linear combination of the top 4 patterns (y-axis) by the number of position in the specific 5-mer context (x-axis) and the score type: mismatch, deletion and insertion (resp. black, orange and blue). The 4 patterns (coefficient vectors: 1,2,3,6) are selected according to the predominant columns in matrix  $W$ . **D:** Score distribution inferred from the combined patterns, stratified by the different 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).