

# Mapping of RNA modifications by direct Nanopore sequencing and JACUSA2

Christoph Dieterich<sup>\*1,2,3</sup>, Amina Lemsara<sup>1,2</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

Cellular RNA is modified by different types of chemical modifications, which are now summarized as the "epitranscriptome". With the advent of high-throughput sequencing technologies much progress has been made in understanding the mechanisms of RNA modification biology and how these modifications can influence gene expression. 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 in Nanopore direct RNA sequencing data, which employs pairwise comparison by JACUSA2. We describe two exemplary Use Cases in detail: a sample of interest ("wild type") may be either compared to a sample lacking a specific modification type ("knock out", Use Case 1) or to a sample lacking all modifications ("IVT", Use Case 2). We provide a detailed guidance for preprocessing of Nanopore reads and provide a snakemake pipeline to map m6A and validate the results against a consensus miCLIP-derived m6A site set.

**Keywords:** RNA modification, Nanopore sequencing, m6A

## INTRODUCTION

Chemical modifications on DNA and histones, also known as epigenetics marks, strongly impact gene expression during cell differentiation and in several other biological programs. In the 1970s, it was recognized that RNA is also subjected to extensive covalent modification, and studies in the late 1980s revealed the widespread deamination of bases (termed RNA editing),

---

\*christoph.dieterich@uni-heidelberg.de

which can lead to recoding if it occurs within coding sequences. Impressive development in the RNA modification field occurred during the past eight years, with the discovery of an extensive layer of base modifications in mRNAs. These can influence gene expression and have been already shown to be involved in primary cellular programs such as stem cell differentiation, response to stress, and the circadian clock. The study of RNA modifications and their effects is now referred to as epitranscriptomics, and it reveals striking similarities to what is known for epigenomics. To date thirteen distinct modifications have been identified on mRNA transcripts [Anreiter et al., 2021]. These modifications are catalyzed by a variety of dedicated enzymes and can be divided into two classes: modifications of cap-adjacent nucleotides and internal modifications.

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

78 and PCR-based library amplification. One recent addition to the toolbox  
79 of RNA modification mapping is direct RNA single molecule long read se-  
80 quencing on the Oxford Nanopore Technologies platform. While our soft-  
81 ware workflow is able to deal with Illumina and Nanopore-based approaches,  
82 the latter is the principal topic of this methods article.

## 83 MATERIALS

### 84 ONT direct RNA sequencing

- 85 1. 500 ng polyA<sup>+</sup> RNA isolated from total RNA e.g. with Oligotex  
86 mRNA kit (Qiagen) or Dynabeads oligo dT<sub>25</sub> beads (Thermo Fisher  
87 Scientific) or *in vitro* transcriptome sample. Store RNA at -80 °C and  
88 the mRNA purification kit as recommended by the manufacturer.
- 89 2. Nuclease-free water. Store at room temperature.
- 90 3. Direct RNA-sequencing kit (SQK-RNA002, Oxford Nanopore Tech-  
91 nologies). Store at -20 °C.
- 92 4. NEBNext Quick Ligation Reaction Buffer (New England Biolabs).  
93 Store at -20 °C.
- 94 5. T4 DNA Ligase (New England Biolabs). Store at -20 °C.
- 95 6. dNTP Mix (10 mM each). Store at -20 °C.
- 96 7. SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific). Store  
97 at -20 °C.
- 98 8. Agencourt RNAClean XP beads (Beckman Coulter). Store at 4 °C.
- 99 9. 70 % ethanol, freshly prepared.
- 100 10. Qubit dsDNA HS assay kit and Qubit Fluorometer (Thermo Fisher  
101 Scientific).
- 102 11. Flow cell priming kit (EXP-FLP002, Oxford Nanopore Technologies).  
103 Store at -20 °C.
- 104 12. Thermocycler.
- 105 13. Gentle rotator mixer.
- 106 14. Magnetic stand for 1.5 ml tubes.
- 107 15. 1.5 ml DNA LoBind tubes (Eppendorf), 0.2 ml PCR tubes.
- 108 16. MinION or GridION sequencing device and MinION R9.4.1 Flow cells  
109 (FLO-MIN106D, Oxford Nanopore Technologies). Store Flow cells at  
110 4 °C.

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

- 112 1. 100 ng polyA<sup>+</sup> RNA isolated from total RNA e.g. with Oligotex  
113 mRNA kit (Qiagen) or Dynabeads oligo dT<sub>25</sub> beads (Thermo Fisher  
114 Scientific). Store RNA at -80 °C and the mRNA purification kit as  
115 recommended by the manufacturer
- 116 2. 10 μM oligo(dT)-VN RT primer.  
117 TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN. Store at -20 °C.
- 118 3. 20 μM template switching oligo (TSO). ACTCTAATACGACTCAC-  
119 TATAGGGAGAGGGCrGrG+G. Store at -20 °C.
- 120 4. 10 μM T7 extension primer. GCTCTAATACGACTCACTATAGG.  
121 Store at -20 °C.
- 122 5. Nuclease-free water. Store at room temperature.
- 123 6. dNTP Mix (10 mM each). Store at -20 °C.
- 124 7. Template Switching RT Enzyme Mix (New England Biolabs). Store  
125 at -20 °C.
- 126 8. Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs).  
127 Store at -20 °C.
- 128 9. RNase H (5,000 U/ml) (New England Biolabs). Store at -20 °C.
- 129 10. NucleoSpin Gel and PCR Clean-up, Mini kit for gel extraction and  
130 PCR clean up (Macherey-Nagel) or equivalent. Store at room temper-  
131 ature.
- 132 11. MEGAscript T7 transcription kit (Thermo Fisher Scientific). Store at  
133 -20 °C.
- 134 12. RNA Clean & Concentrator-25 kit (Zymo Research). Store at room  
135 temperature.
- 136 13. Thermocycler.
- 137 14. Table top centrifuge for 1.5 ml tubes.
- 138 15. Nanodrop spectrophotometer or equivalent.
- 139 16. 0.2 ml PCR tubes, 1.5 ml DNA LoBind tubes (Eppendorf).

## 140 Hardware requirements

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

## 146 Software dependencies and installation

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

## 156 METHODS

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

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

176 Our workflow can be divided into a wet-lab part (Figure 3A) and a  
177 computational part (Figure 3B). Starting from total cellular RNA, polyA<sup>+</sup>  
178 RNA is isolated and subjected to Nanopore direct RNA-sequencing. Guppy

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

179 basecalling can be done as live basecalling during sequencing or after the  
180 sequencing run from generated FAST5 files, resulting in FASTQ output files  
181 (Figure 3A). FASTQ files are aligned to a reference sequence with Minimap2.  
182 SAMtools is used to generate BAM files as input for JACUSA2 analysis,  
183 which yields candidate m4A sites (Figure 3B).

## 184 Nanopore direct RNA sequencing

- 185 1. Adjust 500 ng polyA<sup>+</sup> RNA to a total volume of 9  $\mu$ l with nuclease-  
186 free water. Complete RT adapter ligation reaction (in 0.2 ml PCR  
187 tube) with 3  $\mu$ l NEBNext Quick Ligation Reaction Buffer, 0.5  $\mu$ l  
188 RNA CS (RCS, from SQK-RNA002), 1  $\mu$ l RT-Adapter (RTA, from  
189 SQK-RNA002) and 1.5  $\mu$ l T4 DNA Ligase. Incubate 10 min at room  
190 temperature.
- 191 2. Prepare reverse transcription master mix on ice during ligation: 9  $\mu$ l  
192 nuclease-free water, 2  $\mu$ l 10 mM dNTPs, 8  $\mu$ l 5x SuperScript IV first  
193 strand buffer, 4  $\mu$ l 0.1 mM DTT.
- 194 3. Add the reverse transcription master mix to the ligation reaction and  
195 mix by pipetting. Add 2  $\mu$ l SuperScript IV reverse transcriptase and  
196 mix by pipetting. Incubate in a thermocycler with the following pro-  
197 tocol: 50 min at 50 °C, 10 min at 70 °C, cool down to 4 °C.
- 198 4. Let the Agencourt RNAClean XP beads come to room temperature  
199 during reverse transcription. Carefully resuspend beads before use.  
200 Transfer reaction to a 1.5 ml DNA LoBind tube and mix with 72  $\mu$ l  
201 Agencourt RNAClean XP beads. Incubate 5 min at room temperature  
202 on a gentle rotator mixer.
- 203 5. Collect beads on a magnetic stand and remove supernatant. Wash  
204 pelleted beads two times (30 sec) with 200  $\mu$ l freshly prepared 70 %  
205 ethanol. Remove supernatant. Spin sample down and place on magnet  
206 again. Remove any residual ethanol.
- 207 6. Resuspend beads in 20  $\mu$ l nuclease-free water by gentle flicking and  
208 incubate 5 min at room temperature on a gentle rotator mixer. Collect  
209 beads on a magnetic stand and transfer 20  $\mu$ l eluate in a fresh 1.5 ml  
210 DNA LoBind tube.
- 211 7. For ligation of the RMX adapter, add the following to 20  $\mu$ l eluate: 8  
212  $\mu$ l NEBNext Quick Ligation Reaction Buffer, 6  $\mu$ l RMX (from SQK-  
213 RNA002), 3  $\mu$ l nuclease-free water, 3  $\mu$ l T4 DNA Ligase. Mix by  
214 pipetting and incubate 10 min at room temperature.

- 215 8. Add 40  $\mu\text{l}$  carefully resuspended Agencourt RNAClean XP beads to  
216 the reaction and mix by pipetting. Incubate 5 min at room tempera-  
217 ture on a gentle rotator mixer.
- 218 9. Collect beads on a magnetic stand and remove supernatant. Wash  
219 pelleted beads two times with 150  $\mu\text{l}$  wash buffer (WSB, from SQK-  
220 RNA002). Resuspend beads by flicking, spin down and return to mag-  
221 netic stand. Remove supernatant from pelleted beads.
- 222 10. Resuspend beads in 21  $\mu\text{l}$  elution buffer (EB, from SQK-RNA002) by  
223 gentle flicking and incubate 5 min at room temperature on a gentle  
224 rotator mixer. Pellet beads on a magnetic stand and transfer 21  $\mu\text{l}$   
225 eluate in a fresh 1.5 ml DNA LoBind tube.
- 226 11. Quantify 1  $\mu\text{l}$  of the library on a Qubit fluorometer with the Qubit  
227 dsDNA HS kit according to the manufacturerers protocol. Concentra-  
228 tion should be usually in the range of 5 - 10 ng/ $\mu\text{l}$ .
- 229 12. Insert MinION R9.4.1 Flow cell in the MinION or GridION sequenc-  
230 ing device and perform Flow cell check in the MinKNOW software.  
231 For successful sequencing of mammalian polyA<sup>+</sup> RNA at least 1,000  
232 available pores are recommended.
- 233 13. Prepare Priming Mix by adding 30  $\mu\text{l}$  flush tether (FLT, from EXP-  
234 FLP002) to a vial of flush buffer (FB, from EXP-FLP002) and mix by  
235 pipetting. Open priming port. Remove air bubble from priming port  
236 by inserting the tip of a P1000 pipette into the priming port and slowly  
237 dialing up, until a small volume of storage buffer enters the pipette  
238 tip. Load 800  $\mu\text{l}$  Priming Mix via the priming port and carefully avoid  
239 introduction of air bubbles. Close the priming port and wait for 5 min.
- 240 14. Mix 20  $\mu\text{l}$  library with 17.5  $\mu\text{l}$  nuclease-free water and 37.5  $\mu\text{l}$  RNA run-  
241 ning buffer (RRB, from SQK-RNA002) and mix by pipetting. Open  
242 the priming port and the sample port. Load 200  $\mu\text{l}$  Priming Mix via  
243 the priming port. Mix library by pipetting just before loading and  
244 load dropwise via the sample port. Carefully avoid introduction of air  
245 bubbles. Close the sample port and the priming port.
- 246 15. Start sequencing for 48 to 72 h in the MinKNOW software. Choose  
247 direct RNA-sequencing kit and high-accuracy basecalling as paramet-  
248 ers.

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

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

- 252 1. Adjust 100 ng polyA<sup>+</sup> RNA to a total volume of 6  $\mu$ l with nuclease-  
253 free water. Add 1  $\mu$ l each of 10  $\mu$ M oligo(dT)-VN RT primer and 10  
254 mM dNTPs. Mix by pipetting and incubate in a thermocycler: 5 min  
255 at 75 °C, 2 min at 42 °C, cool to 4 °C.
- 256 2. Assemble 2.5  $\mu$ l 4x template switching RT buffer, 0.5  $\mu$ l 20  $\mu$ M TSO,  
257 1  $\mu$ l 10x template switching RT enzyme mix and mix by pipetting.  
258 Combine with 6  $\mu$ l RNA and incubate in a thermocycler: 90 min at  
259 42 °C, 10 min at 68 °C, cool to 4 °C.
- 260 3. For Second strand synthesis add to First strand synthesis reaction: 50  
261  $\mu$ l Q5 Hot Start High-Fidelity 2X Master Mix, 5  $\mu$ l RNase H, 2  $\mu$ l 10  
262  $\mu$ M T7 extension primer, 33  $\mu$ l nuclease-free water. Mix by pipetting  
263 and incubate in a thermocycler: 15 min at 37 °C, 1 min at 95 °C, 10  
264 min at 65 °C, cool to 4 °C.
- 265 4. Purify double stranded cDNA with NucleoSpin Gel and PCR Clean-up  
266 kit according to the manufacturerers protocol and elute in 20  $\mu$ l elution  
267 buffer. Determine concentration on a Nanodrop spectrophotometer.  
268 cDNA may be stored at -20 °C.
- 269 5. Combine 8  $\mu$ l cDNA for *in vitro* transcription with 2  $\mu$ l each of ATP,  
270 GTP, CTP, UTP, 10x reaction buffer and enzyme mix from the MEGAscript  
271 T7 transcription kit. Incubate 3 h at 37 °C.
- 272 6. Digest template DNA by addition of 1  $\mu$ l Turbo DNase. Mix by pipet-  
273 ting and incubate 15 min at 37 °C.
- 274 7. Adjust reaction volume to 100  $\mu$ l with nuclease-free water and clean up  
275 with RNA Clean & Concentrator-25 kit according to the manufactur-  
276 ers protocol, using two volumes of adjusted RNA binding buffer (1:1  
277 RNA binding buffer : ethanol). Elute RNA in 25  $\mu$ l nuclease-free wa-  
278 ter. Determine RNA concentration on a Nanodrop spectrophotometer.  
279 Store at -80 °C.

## 280 Nanopore read processing

- 281 1. Base call the ionic current signal stored in FAST5 files using Guppy.  
282 For the IVT sample, we applied real-time base calling with the MinKNOW-  
283 embedded Guppy basecaller. Otherwise, Guppy basecaller software  
284 can be used. In this case, the basecaller requires the path to FAST5  
285 files, the output folder, and the config file or the flowcell/kit combina-  
286 tion. The output are FASTQ files that can be compressed using the  
287 option "--compress\_fastq".



```

288 $ guppy_basecaller --compress_fastq -i path_to_fast5 -s path_to_output
289 -c config_file.cfg --cpu_threads_per_caller 14 --num_callers
290 1
291 Set the number of threads "cpu_threads_per_caller" and the number
292 of parallel basecallers "num_caller" according to your resources. Ad-
293 ditional details can be found in Gup [2019].

```

2. Align reads to the transcriptome using Minimap2 software. The output is a SAM file that has to be converted to a compressed form as BAM file using SAMtools command. The alignment requires a reference sequence. Here, we used GRCh38 Ensembl annotation and FASTA file release version 96. **To reduce the indexing time of the human genome, save the index with the option "-d" before the mapping and use the index instead of the reference file in the minimap2 command line.**

```

302 $ minimap2 -d reference.mmi reference.fa

```

**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 file "-junc-bed annotation.bed" provides the splice junctions. The BED file can be generated using the following command:**

```

308 $paftools.js gff2bed annotation.gtf > annotation.bed

```

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

```

317 $ minimap2 -t 5 --MD -ax splice --junc-bonus 1 -k14 --secondary=no
318 --junc-bed final_annotation_96.bed -ub reference.mmi Reads.fastq.gz
319 |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:

```
$ srun snakemake --cores all rule_name
```

Check Snakemake documentation for more details [sna].

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

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

367 Starting with the preprocessed mapped reads as inputs (BAM files),  
 368 'HEK293T-WT-rep2.bam' and 'HEK293T-WT-rep3.bam' represent the wild-  
 369 type replicates and 'HEK293T-KO-rep2.bam' and 'HEK293T-KO-rep3.bam'  
 370 the control replicates,

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

390 \$ srun snakemake --cores all jacusa2\_call2 \$

391 2. Preprocess JACUSA2 output: from JACUSA2 call-2 output, we select  
 392 all sites within 5-mer of a central nucleotide 'A' flanked by 2 random  
 393 nucleotides (NNANN) and we filter out sites of the homo-polymer re-  
 394 gions (JACUSA filter: Y). Then, we rebuild the tabular features such  
 395 that the observations are only sites with a reference base 'A'. Each  
 396 site is characterized by 15 features corresponding to the mismatch,  
 397 insertion and deletion scores for the observed site and its two flank-  
 398 ing positions from both sides. The rule "get\_features" performs the  
 399 preprocessing step. Use the parameter 'region' with a file containing  
 400 target 5-mers to limit the analysis to specific sites. For comparison  
 401 reasons, we consider common sites between use cases 1 and 2. The  
 402 output is an R object "features/features.rds", representing the matrix  
 403 of Sites×15 features.

404 \$ srun snakemake --cores all get\_features

405 3. Extract m6A modification pattern: given the matrix of Sites×Features,  
 406 the next step is to learn a model representing the m6A modification

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

pattern. To this end, the conventional non-negative matrix factorization (NMF) analysis is suggested [Lee and Seung, 1999]. Briefly, NMF factorizes a non-negative data matrix  $X$  (here:  $n$  sites and  $m$  features) into two non-negative matrices as  $X \approx WH$ , such that  $W$  is an  $n \times k$  matrix containing basis vectors and  $H$  is an  $k \times m$  matrix containing coefficient vectors. The coefficient vectors and their combination can be viewed as a pattern for m6A modification. The rank of factorization  $k$  is a critical parameter that affects the performance substantially. We suggest to select the rank  $k$  according to the method of Frigyesi and Höglund [2008] by looking at silhouette [Rousseeuw, 1987] and cophenetic correlation [Brunet et al., 2004] indices. Accordingly, the performance indices are computed for different 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 difference between slopes of the original and the randomized 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 identified an optimal factorization rank of 6 (Figure 6A). We then analyzed the identified patterns. According to the membership indicator of each site in matrix  $W$ , more than 80% of m6A modification sites can be represented by five patterns (Patterns 1,2,3,4,6) (Figure 6B). Interestingly, the linear combination of these five patterns in Figure 6C highlights the importance of position 3 and eventually the implication of all scores.

Using the JACUSA2 pipeline, run rule "get\_pattern" to generate patterns and provide the set of modified sites as a ground truth under the field "modified\_sites" in the config file. Here, the "miCLIP\_union.bed" file contains the m6A sites from the three miCLIP-based studies. A miCLIP annotation, reflecting the consensus sites, is added to each site. A subset of modified sites can be used to generate patterns. Accordingly, the "internal\_pattern" field should refer to the annotation of selected sites from the "modified\_sites" file. Plus, multiple combinations of patterns can be defined and appended to the field "combined\_pattern" as new patterns. The corresponding outputs are under "patterns" folder.

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

The produced 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
```

in Figure 6C this is labeled sum

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

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

469 \$ srun snakemake --cores all predict\_modification

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

471 An alternative way to detect RNA modifications is to compare a modi-  
 472 fied sample to an *in-vitro* transcribed (IVT) control sample. Therefore,  
 473 we benchmark JACUSA2 on a sample set of two replicates (2 and 3) from  
 474 wild-type HEK293 cell lines (modified sample) Pratanwanich et al. [2021]  
 475 and a modification-free IVT sample from HEK293 cDNA (control sample)  
 476 (see "Preparation of an *in vitro* transcriptome sample"). The analysis steps  
 477 are similar to case 1. We evaluate the analysis against miCLIP m6A sites  
 478 (Figure 5).

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

484 \$ srun snakemake --cores all jacusa2\_call2

485 2. Preprocess JACUSA2 output: we select all sites within the specific 5-  
 486 mer (NNANN) and we consider the Y filter that excludes sites within

The first IVT run has a relatively low coverage -> might this impact performance of UC2?

487 homo-polymer regions. Then, we extract 5-mer features such that the  
488 selected sites are represented by the Mismatch, Deletion and Insertion  
489 scores for the observed site and its two flanking positions from both  
490 sides.

491 `$ srun snakemake --cores all get_features`

492 3. Extract m6A modification pattern: using NMF factorization, we ex-  
493 tract patterns from the 1,905 sites reported as modified in the three  
494 miCLIP-based studies. Based on the silhouette and cophenetic corre-  
495 lation indices, we identified an optimal factorization rank of 6 (Figure  
496 7A). We determined the predominant factors from matrix  $W$ . Accord-  
497 ingly, more than 80% of m6A modification sites can be represented by  
498 four patterns (Patterns: 1,2,3,6) (Figure 7B). In agreement with Use  
499 Case 1, the linear combination of the four patterns confirms the im-  
500 portance of position 3 and the implication of all scores as shown in  
501 Figure 7C.

502 `$ srun snakemake --cores all get_pattern`

503 4. Predict m6A modifications: we evaluate the prediction ability of the  
504 detected patterns on a test set of almost 1,52 million sites where  
505 17,021 are miCLIP-m6A modified. We consider the linear combina-  
506 tion of the four most relevant patterns (1,2,3,6). Figure 7D shows the  
507 eCDF of the inferred scores. The difference between the cumulative  
508 distribution of non miCLIP sites and miCLIP sites can be nicely ob-  
509 served, while the PPV plot shows a lower performance as compared  
510 to Use Case 1 (Figure 7E). The decrease in performance is likely ex-  
511 plained by the absence of all modifications and not exclusively m6A in  
512 the control condition, which may induce noise to the score estimation  
513 by JACUSA2 call-2 .

514 `$ srun snakemake --cores all predict_modification`

CD:to  
be con-  
firmed

## 515 NOTES

### 516 Tips and Tricks

517 1. The reverse transcription step during library preparation is optional.  
518 However, we recommend to include this step to ensure proper sequenc-  
519 ing also of RNAs with secondary structures. Superscript IV reverse  
520 transcriptase may be replaced by Superscript III reverse transcriptase,  
521 which is used in the protocol provided by Oxford Nanopore Technolo-  
522 gies.

- 523 2. The library preparation protocol contains two bead clean up steps. It  
524 is important to remove ethanol and wash buffer completely. However,  
525 beads should not be dried for several minutes. Directly add water  
526 or elution buffer after washing to prevent sticking of the RNA to the  
527 beads.
- 528 3. The default filter in current MinKNOW versions is a Q score of 9. For  
529 direct RNA sequencing we recommend to adjust the output filter to a  
530 minimum Q score of 7, as in previous MinKNOW versions.
- 531 4. During preparation of the *in vitro* transcriptome sample, *in vitro* tran-  
532 scription and clean up kits may be replaced by equivalent products.  
533 The protocol however has been tested only with the mentioned kits.
- 534 5. Configuration of the pipeline should be handled via the config file. All  
535 parameters should be set before executing rules.
- 536 6. Once the pipeline has run successfully you should expect the following  
537 folders with the corresponding outputs in the output directory: bam,  
538 jacusa, features, patterns, and prediction.
- 539 7. JACUSA2 call2 could be run separately using the command line as  
540 described in JACUSA2 manual [JAC, 2021], then put the output under  
541 a new folder with the name 'jacusa' under the output directory.
- 542 8. In the snakemake pipeline, rules are linked so that the workflows  
543 are determined from top (e.g. predict\_modification) to bottom (e.g.  
544 sort\_bam) and executed accordingly from bottom to top (Figure 4).  
545 Therefore, running for example "predict\_modification" rule leads to  
546 executing all rules on its pipeline.
- 547 9. Patterns could be generated from a subset of the input data that  
548 correspond to known modified sites. Alternatively, predefined patterns  
549 as a NMF R object could be used as a prediction model.

## 550 ACKNOWLEDGMENTS

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

## 554 REFERENCES

- 555 Minimap2. <https://github.com/lh3/minimap2>. Accessed: 2022-01-19.
- 556 Snakemake. <https://snakemake.readthedocs.io>. Accessed: 2022-01-26.

CD:  
fund-  
ing?

557 Basecalling with guppy. [https://github.com/metagenomics/](https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/basecalling.rst)  
558 [denbi-nanopore-training/blob/master/docs/basecalling/](https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/basecalling.rst)  
559 [basecalling.rst](https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/basecalling.rst), 2019. Accessed: 2022-01-19.

560 Jacusa2 manual. <https://github.com/dieterich-lab/JACUSA2>, 2021.  
561 Accessed: 2022-01-15.

562 Samir Adhikari, Wen Xiao, Yong-Liang Zhao, and Yun-Gui Yang. m(6)a:  
563 Signaling for mrna splicing. *RNA biology*, 13:756–759, September 2016.  
564 ISSN 1555-8584. doi: 10.1080/15476286.2016.1201628.

565 Ina Anreiter, Quoseena Mir, Jared T. Simpson, Sarath C. Janga, and  
566 Matthias Soller. New twists in detecting mrna modification dynamics.  
567 *Trends in biotechnology*, 39:72–89, January 2021. ISSN 1879-3096. doi:  
568 10.1016/j.tibtech.2020.06.002.

569 Konstantinos Boulas, Diana Toczyłowska-Socha, Ben R Hawley, Noa  
570 Liberman, Ken Takashima, Sara Zaccara, Théo Guez, Jean-Jacques  
571 Vasseur, Françoise Debart, L Aravind, et al. Identification of the m6am  
572 methyltransferase pcif1 reveals the location and functions of m6am in the  
573 transcriptome. *Molecular cell*, 75(3):631–643, 2019.

574 Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov.  
575 Metagenes and molecular pattern discovery using matrix factorization.  
576 *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.

577 Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali  
578 Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine  
579 Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and  
580 Gideon Rechavi. Topology of the human and mouse m6a rna methylomes  
581 revealed by m6a-seq. *Nature*, 485:201–206, April 2012. ISSN 1476-4687.  
582 doi: 10.1038/nature11112.

583 Hao Du, Ya Zhao, Jinqiu He, Yao Zhang, Hairui Xi, Mofang Liu, Jinbiao  
584 Ma, and Ligang Wu. Ythdf2 destabilizes m 6 a-containing rna through  
585 direct recruitment of the ccr4-not deadenylase complex. *Nature commu-*  
586 *nications*, 7(1):1–11, 2016.

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

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



594 Guifang Jia, Ye Fu, Xu Zhao, Qing Dai, Guanqun Zheng, Ying Yang,  
595 Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, and Chuan He.  
596 N6-methyladenosine in nuclear rna is a major substrate of the obesity-  
597 associated fto. *Nature chemical biology*, 7:885–887, October 2011. ISSN  
598 1552-4469. doi: 10.1038/nchembio.687.

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

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

609 Nadine Körtel, Cornelia Rücklé, You Zhou, Anke Busch, Peter Hoch-Kraft,  
610 Reymond FX Sutandy, Jacob Haase, Mihika Pradhan, Michael Musheev,  
611 Dirk Ostareck, et al. Deep and accurate detection of m6a rna modifications  
612 using miclip2 and m6aboost machine learning. *bioRxiv*, pages 2020–12,  
613 2021.

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

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

618 Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christo-  
619 pher E. Mason, and Samie R. Jaffrey. Comprehensive analysis of mrna  
620 methylation reveals enrichment in 3’ utrs and near stop codons. *Cell*, 149:  
621 1635–1646, June 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.05.003.

622 Deepak P Patil, Brian F Pickering, and Samie R Jaffrey. Reading m6a in  
623 the transcriptome: m6a-binding proteins. *Trends in cell biology*, 28(2):  
624 113–127, 2018.

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

627 Ploy N Pratanwanich, Fei Yao, Ying Chen, Casslynn WQ Koh, Yuk Kei  
628 Wan, Christopher Hendra, Polly Poon, Yeek Teck Goh, Phoebe ML Yap,  
629 Jing Yuan Chooi, et al. Identification of differential rna modifications  
630 from nanopore direct rna sequencing with xpore. *Nature Biotechnology*,  
631 39(11):1394–1402, 2021.

632 Jean-Yves Roignant and Matthias Soller. m,  
 633 javax.xml.bind.jaxbelement@8cec19d, a in mrna: An ancient mech-  
 634 anism for fine-tuning gene expression. *Trends in genetics : TIG*, 33:  
 635 380–390, June 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2017.04.003.

636 Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic rna  
 637 modifications in gene expression regulation. *Cell*, 169:1187–1200, June  
 638 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.045.

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

642 Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how:  
 643 Context-dependent functions of rna methylation writers, readers, and  
 644 erasers. *Molecular cell*, 74:640–650, May 2019. ISSN 1097-4164. doi:  
 645 10.1016/j.molcel.2019.04.025.

646 Xiao Wang, Zhike Lu, Adrian Gomez, Gary C Hon, Yanan Yue, Dali Han,  
 647 Ye Fu, Marc Parisien, Qing Dai, Guifang Jia, et al. N 6-methyladenosine-  
 648 dependent regulation of messenger rna stability. *Nature*, 505(7481):117–  
 649 120, 2014.

650 Xiao Wang, Boxuan Simen Zhao, Ian A Roundtree, Zhike Lu, Dali Han,  
 651 Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He.  
 652 N6-methyladenosine modulates messenger rna translation efficiency. *Cell*,  
 653 161(6):1388–1399, 2015.

654 Sara Zaccara, Ryan J. Ries, and Samie R. Jaffrey. Reading, writing and  
 655 erasing mrna methylation. *Nature reviews. Molecular cell biology*, 20:608–  
 656 624, October 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0168-5.

657 Zhang Zhang, Tao Chen, Hong-Xuan Chen, Ying-Yuan Xie, Li-Qian Chen,  
 658 Yu-Li Zhao, Biao-Di Liu, Lingmei Jin, Wutong Zhang, Chang Liu,  
 659 et al. Systematic calibration of epitranscriptomic maps using a synthetic  
 660 modification-free rna library. *Nature Methods*, 18(10):1213–1222, 2021.

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

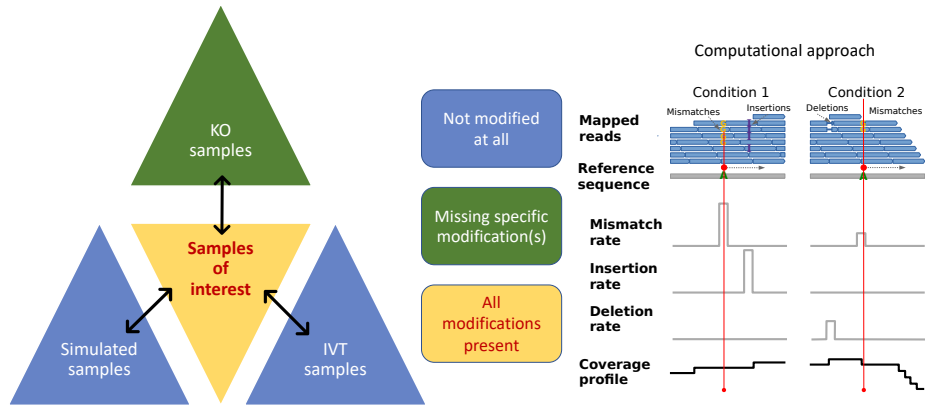


Figure 1: **General outline of RNA modification detection by JACUSA2.** A key feature of our approach is that multiple replicates can be compared as shown on the left. Samples of interests where all modifications are present could be compared with either KO samples where the modification of interest 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.

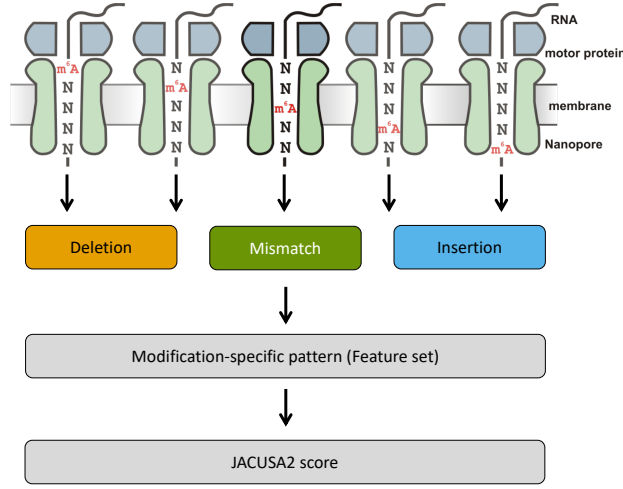


Figure 2: **Motivation of 5-mer context for RNA modification mapping.** The nanopore covers 5 consecutive RNA residues. That is why we consider a 5-mer context and derive 3 principal features for every position within a given 5-mer (15 features in total, with a central A residue in this example). We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

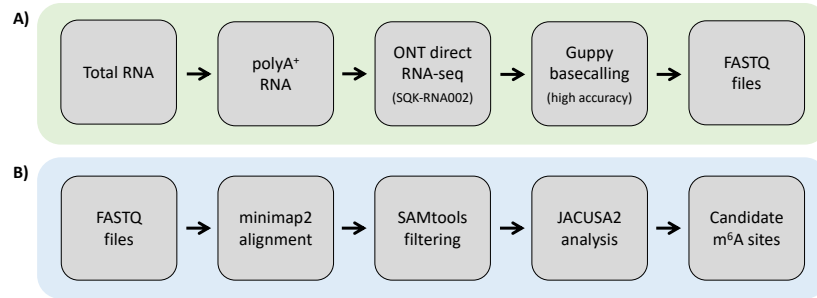


Figure 3: **Experimental and computational workflow.** A) Starting from total cellular RNA, polyA<sup>+</sup> RNA is isolated and subjected to Nanopore direct RNA-sequencing. Guppy basecalling can be done as live basecalling during sequencing or after the sequencing run from generated FAST5 files, resulting 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.

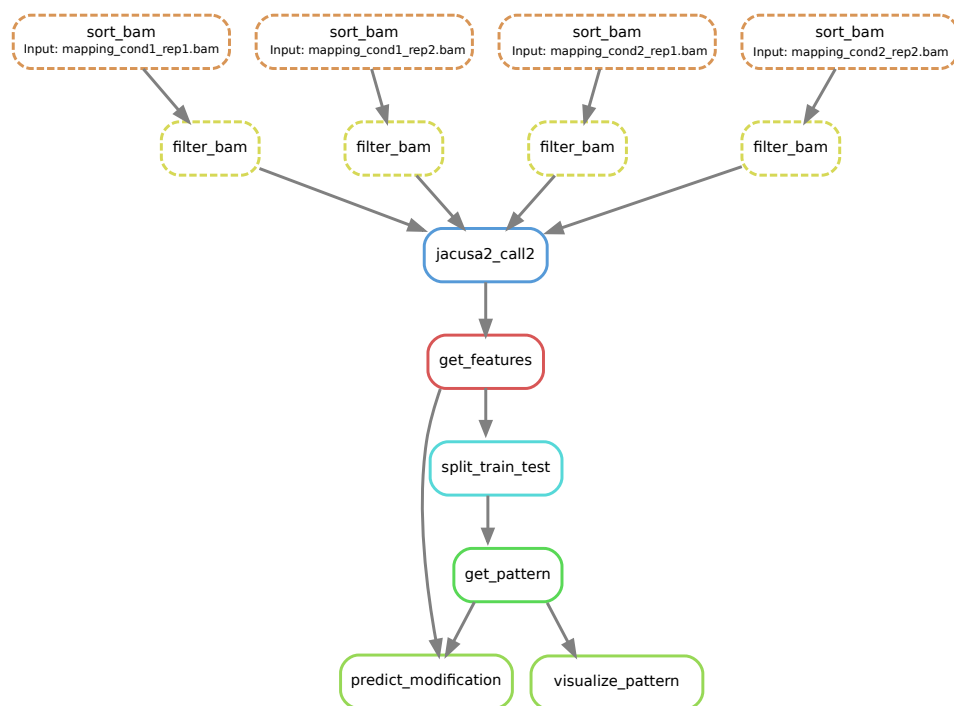


Figure 4: **Computational workflow.** Snakemake workflow for RNA modification detection based on JACUSA2 variant calling.

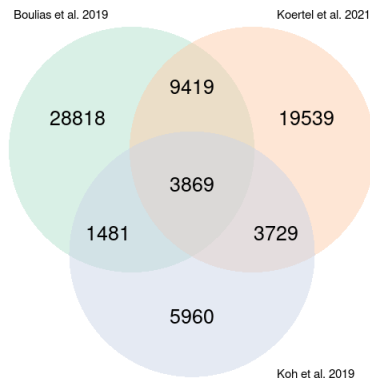


Figure 5: **m6A sites reported in the three miCLIP-based studies** Boulias et al. [2019], Koh et al. [2019] and Körtel et al. [2021].

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

## 671 TABLE CAPTIONS

## 672 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



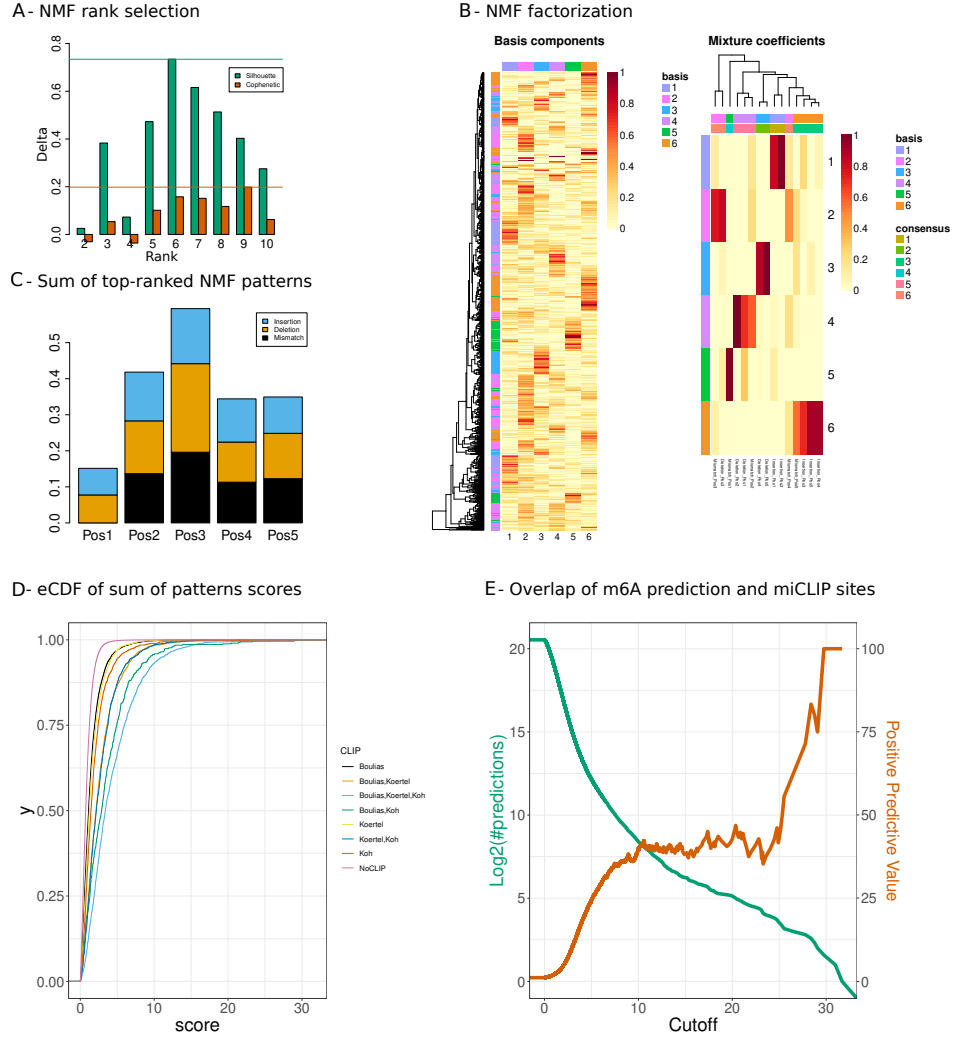


Figure 6: **Case 1. WT versus KO.** **A:** NMF rank selection. Barplots representing the difference of cophenetic correlation and silhouette indices between the NMF factorization of the original and the randomized data. Ranks with the largest value for both indices are determined, then the smallest rank 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).

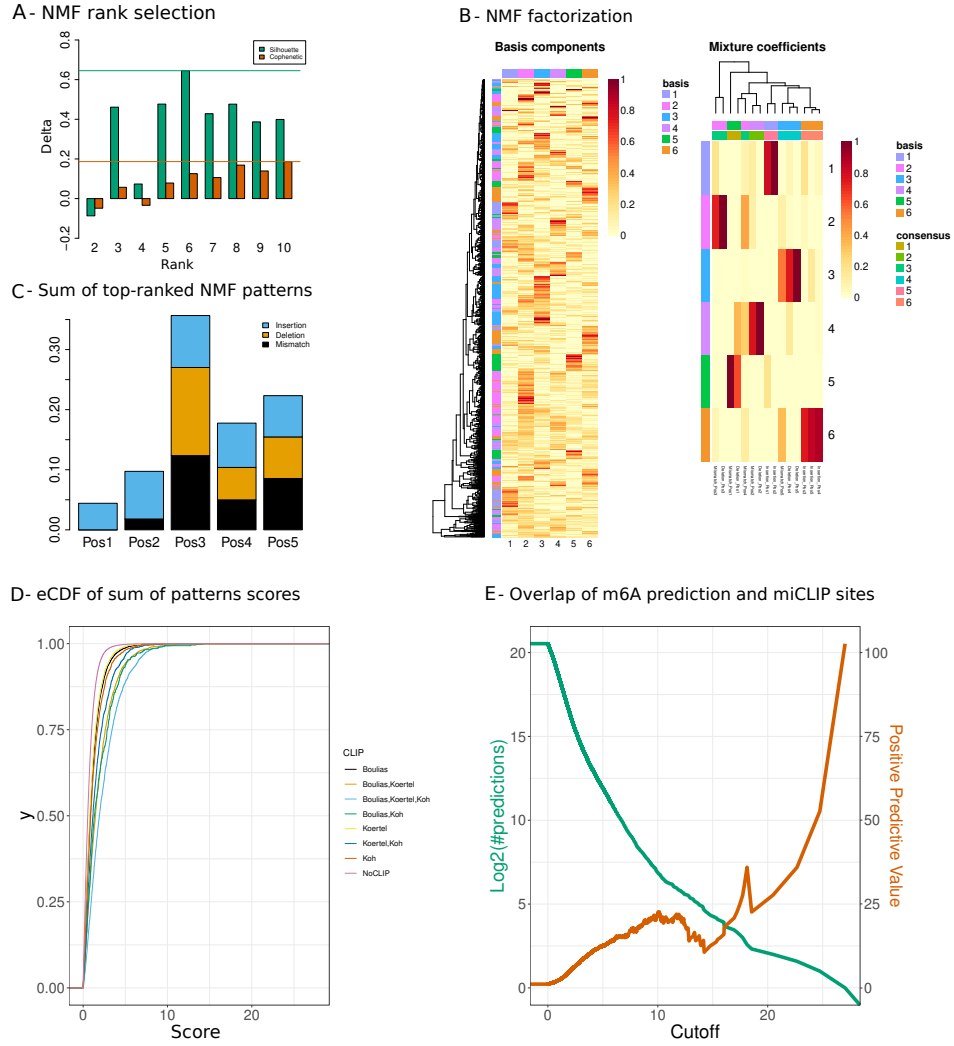


Figure 7: **Case 2. WT versus IVT.** **A:** NMF rank selection. Barplots representing the difference of cophenetic correlation and silhouette indices between the NMF factorization of the original and the randomized data. Ranks with the largest value for both indices are determined, then the smallest 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).