

Mapping of RNA modifications by direct Nanopore sequencing and JACUSA2

Christoph Dieterich^{*1,2,3}, Amina Lemsara^{1,2}, and Isabel Naarmann-de Vries^{1,2,3}

¹Klaus Tschira Institute for Integrative Computational Cardiology, University Heidelberg, 69120 Heidelberg, Germany

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

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

Abstract

to be written

Keywords: Bayesian, 10X Genomics, Cell barcode assignment, Nonsense-mediated mRNA decay (NMD)

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), 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.

*christoph.dieterich@uni-heidelberg.de

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

55 Several approaches have been presented to map RNA modifications on
 56 RNA. Herein, we focus on mRNA modification site detection in general and
 57 on m6A in particular where antibody-based protocols (miCLIP), methylation-
 58 sensitive restriction enzyme assays (MazF) or transgenic approaches (TRIBE,
 59 DART) have been presented. All of the aforementioned approaches rely on
 60 high-throughput sequencing on the Illumina platform. This typically in-
 61 volves cDNA synthesis by reverse transcription and PCR-based library am-
 62 plification. One recent addition to the tool is direct RNA single molecule
 63 sequencing on the Oxford Nanopore Technology platform. While our software
 64 workflow is able to deal with Illumina and Nanopore-based approaches, the
 65 latter is the principal topic of our methods article.

66 MATERIALS

67 ONT direct RNA sequencing

- 68 1. 500 ng polyA⁺ RNA isolated from total RNA e.g. with Oligotex
 69 mRNA kit (Qiagen) or Dynabeads oligo dT₂₅ beads (Thermo Fisher
 70 Scientific) or *in vitro* transcriptome sample. Store RNA at -80 °C and
 71 the mRNA purification kit as recommended by the manufacturer.

- 72 2. Nuclease-free water. Store at room temperature.
- 73 3. Direct RNA-sequencing kit (SQK-RNA002, Oxford Nanopore Tech-
74 nologies). Store at -20 °C.
- 75 4. NEBNext Quick Ligation Reaction Buffer (New England Biolabs).
76 Store at -20 °C.
- 77 5. T4 DNA Ligase (New England Biolabs). Store at -20 °C.
- 78 6. dNTP Mix (10 mM each). Store at -20 °C.
- 79 7. SuperScript IV Reverse Transcriptase (Thermo Fisher Scientific). Store
80 at -20 °C.
- 81 8. Agencourt RNAClean XP beads (Beckman Coulter). Store at 4 °C.
- 82 9. 70 % ethanol, freshly prepared.
- 83 10. Qubit dsDNA HS assay kit and Qubit Fluorometer (Thermo Fisher
84 Scientific).
- 85 11. Flow cell priming kit (EXP-FLP002, Oxford Nanopore Technologies).
86 Store at -20 °C.
- 87 12. Thermocycler.
- 88 13. Gentle rotator mixer.
- 89 14. Magnetic stand for 1.5 ml tubes.
- 90 15. 1.5 ml DNA LoBind tubes (Eppendorf), 0.2 ml PCR tubes.
- 91 16. MinION or GridION sequencing device and MinION R9.4.1 Flow cells
92 (FLO-MIN106D, Oxford Nanopore Technologies). Store Flow cells at
93 4 °C.

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

- 95 1. 100 ng polyA⁺ RNA isolated from total RNA e.g. with Oligotex
96 mRNA kit (Qiagen) or Dynabeads oligo dT₂₅ beads (Thermo Fisher
97 Scientific). Store RNA at -80 °C and the mRNA purification kit as
98 recommended by the manufacturer
- 99 2. 10 μM oligo(dT)-VN RT primer. TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN.
100 Store at -20 °C.
- 101 3. 20 μM template switching oligo (TSO). ACTCTAATACGACTCAC-
102 TATAGGGAGAGGGCrGrG+G. Store at -20 °C.

- 103 4. 10 μ M T7 extension primer. GCTCTAATACGACTCACTATAGG.
104 Store at -20 °C.
- 105 5. Nuclease-free water. Store at room temperature.
- 106 6. dNTP Mix (10 mM each). Store at -20 °C.
- 107 7. Template Switching RT Enzyme Mix (New England Biolabs). Store
108 at -20 °C.
- 109 8. Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs).
110 Store at -20 °C.
- 111 9. RNase H (5,000 U/ml) (New England Biolabs). Store at -20 °C.
- 112 10. NucleoSpin Gel and PCR Clean-up, Mini kit for gel extraction and
113 PCR clean up (Macherey-Nagel) or equivalent. Store at room temper-
114 ature.
- 115 11. MEGAscript T7 transcription kit (Thermo Fisher Scientific). Store at
116 -20 °C.
- 117 12. RNA Clean & Concentrator-25 kit (Zymo Research). Store at room
118 temperature.
- 119 13. Thermocycler.
- 120 14. Table top centrifuge for 1.5 ml tubes.
- 121 15. Nanodrop spectrophotometer or equivalent.
- 122 16. 0.2 ml PCR tubes, 1.5 ml DNA LoBind tubes (Eppendorf).

123 **Hardware requirements**

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

129 **Software dependencies and installation**

130 Our analysis workflow has few requirements, which are detailed in Table 2.
131 Specifically, to execute our workflow, the following prerequisites are neces-
132 sary: a BASH shell, a JAVA runtime environment, a working PERL and
133 R installation. Additional i.e. non-standard software to process and map
134 Nanopore reads (bedtools, samtools and Minimap2) are obligatory, while

135 the installation of a Nanopore read simulator (NanoSim) is optional and de-
136 pends on your use case. Table ?? lists some additional R packages, which are
137 required to run the R code. Detailed instructions on how to setup are found
138 under https://github.com/dieterich-lab/MiMB_JACUSA2_chapter

139 METHODS

140 Overview Figure 1

141 Nanopore direct RNA sequencing

- 142 1. Adjust 500 ng polyA⁺ RNA to a total volume of 9 μ l with nuclease-
143 free water. Complete RT adapter ligation reaction (in 0.2 ml PCR
144 tube) with 3 μ l NEBNext Quick Ligation Reaction Buffer, 0.5 μ l
145 RNA CS (RCS, from SQK-RNA002), 1 μ l RT-Adapter (RTA, from
146 SQK-RNA002) and 1.5 μ l T4 DNA Ligase. Incubate 10 min at room
147 temperature.
- 148 2. Prepare reverse transcription master mix on ice during ligation: 9 μ l
149 nuclease-free water, 2 μ l 10 mM dNTPs, 8 μ l 5x SuperScript IV first
150 strand buffer, 4 μ l 0.1 mM DTT.
- 151 3. Add the reverse transcription master mix to the ligation reaction and
152 mix by pipetting. Add 2 μ l SuperScript IV reverse transcriptase and
153 mix by pipetting. Incubate in a thermocycler with the following pro-
154 tocol: 50 min at 50 °C, 10 min at 70 °C, cool down to 4 °C.
- 155 4. Let the Agencourt RNAClean XP beads come to room temperature
156 during reverse transcription. Carefully resuspend beads before use.
157 Transfer reaction to a 1.5 ml DNA LoBind tube and mix with 72 μ l
158 Agencourt RNAClean XP beads. Incubate 5 min at room temperature
159 on a gentle rotator mixer.
- 160 5. Collect beads on a magnetic stand and remove supernatant. Wash
161 pelleted beads two times (30 sec) with 200 μ l freshly prepared 70 %
162 ethanol. Remove supernatant. Spin sample down and place on magnet
163 again. Remove any residual ethanol.
- 164 6. Resuspend beads in 20 μ l nuclease-free water by gentle flicking and
165 incubate 5 min at room temperature on a gentle rotator mixer. Collect
166 beads on a magnetic stand and transfer 20 μ l eluate in a fresh 1.5 ml
167 DNA LoBind tube.
- 168 7. For ligation of the RMX adapter, add the following to 20 μ l eluate: 8
169 μ l NEBNext Quick Ligation Reaction Buffer, 6 μ l RMX (from SQK-
170 RNA002), 3 μ l nuclease-free water, 3 μ l T4 DNA Ligase. Mix by
171 pipetting and incubate 10 min at room temperature.

- 172 8. Add 40 μ l carefully resuspended Agencourt RNAClean XP beads to
173 the reaction and mix by pipetting. Incubate 5 min at room tempera-
174 ture on a gentle rotator mixer.
- 175 9. Collect beads on a magnetic stand and remove supernatant. Wash
176 pelleted beads two times with 150 μ l wash buffer (WSB, from SQK-
177 RNA002). Resuspend beads by flicking, spin down and return to mag-
178 netic stand. Remove supernatant from pelleted beads.
- 179 10. Resuspend beads in 21 μ l elution buffer (EB, from SQK-RNA002) by
180 gentle flicking and incubate 5 min at room temperature on a gentle
181 rotator mixer. Pellet beads on a magnetic stand and transfer 21 μ l
182 eluate in a fresh 1.5 ml DNA LoBind tube.
- 183 11. Quantify 1 μ l of the library on a Qubit fluorometer with the Qubit
184 dsDNA HS kit according to the manufacturerers protocol. Concentra-
185 tion should be usually in the range of 5 - 10 ng/ μ l.
- 186 12. Insert MinION R9.4.1 Flow cell in the MinION or GridION sequenc-
187 ing device and perform Flow cell check in the MinKNOW software.
188 For successful sequencing of mammalian polyA⁺ RNA at least 1,000
189 available pores are recommended.
- 190 13. Prepare Priming Mix by adding 30 μ l flush tether (FLT, from EXP-
191 FLP002) to a vial of flush buffer (FB, from EXP-FLP002) and mix by
192 pipetting. Open priming port. Remove air bubble from priming port
193 by inserting the tip of a P1000 pipette into the priming port and slowly
194 dialing up, until a small volume of storage buffer enters the pipette
195 tip. Load 800 μ l Priming Mix via the priming port and carefully avoid
196 introduction of air bubbles. Close the priming port and wait for 5 min.
- 197 14. Mix 20 μ l library with 17.5 μ l nuclease-free water and 37.5 μ l RNA run-
198 ning buffer (RRB, from SQK-RNA002) and mix by pipetting. Open
199 the priming port and the sample port. Load 200 μ l Priming Mix via
200 the priming port. Mix library by pipetting just before loading and
201 load dropwise via the sample port. Carefully avoid introduction of air
202 bubbles. Close the sample port and the priming port.
- 203 15. Start sequencing for 48 to 72 h in the MinKNOW software. Choose
204 direct RNA-sequencing kit and high-accuracy basecalling as paramet-
205 ers. We recommend to adjust the output filter to a minimum Q score
206 of 7 (instead of 9).

207 Preparation of an *in vitro* transcriptome sample

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

- 210 1. Adjust 100 ng polyA⁺ RNA to a total volume of 6 μ l with nuclease-
211 free water. Add 1 μ l each of 10 μ M oligo(dT)-VN RT primer and 10
212 mM dNTPs. Mix by pipetting and incubate in a thermocycler: 5 min
213 at 75 °C, 2 min at 42 °C, cool to 4 °C.
- 214 2. Assemble 2.5 μ l 4x template switching RT buffer, 0.5 μ l 20 μ M TSO,
215 1 μ l 10x template switching RT enzyme mix and mix by pipetting.
216 Combine with 6 μ l RNA and incubate in a thermocycler: 90 min at
217 42 °C, 10 min at 68 °C, cool to 4 °C.
- 218 3. For Second strand synthesis add to First strand synthesis reaction: 50
219 μ l Q5 Hot Start High-Fidelity 2X Master Mix, 5 μ l RNase H, 2 μ l 10
220 μ M T7 extension primer, 33 μ l nuclease-free water. Mix by pipetting
221 and incubate in a thermocycler: 15 min at 37 °C, 1 min at 95 °C, 10
222 min at 65 °C, cool to 4 °C.
- 223 4. Purify double stranded cDNA with NucleoSpin Gel and PCR Clean-up
224 kit according to the manufacturerers protocol and elute in 20 μ l elution
225 buffer. Determine concentration on a Nanodrop spectrophotometer.
226 cDNA may be stored at -20 °C.
- 227 5. Combine 8 μ l cDNA for *in vitro* transcription with 2 μ l each of ATP,
228 GTP, CTP, UTP, 10x reaction buffer and enzyme mix from the MEGAscript
229 T7 transcription kit. Incubate 3 h at 37 °C.
- 230 6. Digest template DNA by addition of 1 μ l Turbo DNase. Mix by pipet-
231 ting and incubate 15 min at 37 °C.
- 232 7. Adjust reaction volume to 100 μ l with nuclease-free water and clean up
233 with RNA Clean & Concentrator-25 kit according to the manufactur-
234 ers protocol, using two volumes of adjusted RNA binding buffer (1:1
235 RNA binding buffer : ethanol). Elute RNA in 25 μ l nuclease-free wa-
236 ter. Determine RNA concentration on a Nanodrop spectrophotometer.
237 Store at -80 °C.

238 Nanopore read processing

- 239 1. Following standard steps, base call the ionic current signal stored in
240 FAST5 file using Guppy. The output is FASTQ files. For the IVT read-
241 out, we adopted real-time base calling with the MinKNOW-embedded
242 Guppy basecaller. Otherwise, Guppy basecaller software could be
243 used. Guppy basecaller requires the path to FAST5 files, the output
244 path, and the config file or the flowcell/kit combination and outputs
245 a compressed fastq files using the option `-compress_fastq`. Additional
246 details can be found in Gup [2019].

```

247 $ guppy_basecaller --compress_fastq -i path_to_fast5 -s path_to_output
248 -c config_file.cfg

```

- 249 2. Align reads to the transcriptome using Minimap2 software with the
250 recommended setting for DirectRNA Sequencing (-ax map-ont). The
251 output is a SAM file that have to be converted to a compressed form
252 as a BAM file using SAMtools command. The alignment requires
253 the reference transcriptome/ genome. We used GRCh38 Ensembl an-
254 notations and FASTA file release version 96(). Check the Minimap2
255 manual for more details Min.

```

256 $ minimap2 --secondary=no -ax map-ont -uf -k14 reference.fasta
257 Reads.fastq |samtools view -bS > mapping.bam

```

- 258 3. JACUSA2 requires sorted and indexed BAM files. To sort and create
259 a BAM file index use the following SAMtools commands.

```

260 $ samtools sort mapping.bam mapping.sorted.bam
261 $ samtools index mapping.sorted.bam

```

AL: Ex-
plain
the rest
of the
parameters

262 Use Case 1: Comparison of wildtype and knock-out samples

263 The conventional way to detect RNA modification using direct RNA se-
264 quencing is to compare a modified sample to unmodified control sample. To
265 assess the ability of JACUSA2 RNA modification detection in this case, we
266 used a published dataset of HEK293 cell line Pratanwanich et al. [2021].
267 The benchmark is composed of two samples from two conditions: wild type
268 cells (modified RNAs) and Mettl3 knockout cells (unmodified RNAs) with
269 two replicates (2 and 3). The FASTQ files are preprocessed and mapped
270 according to the steps described in the previous section.

271 Given the preprocessed mapped reads as input (BAM files) 'HEK293T-
272 WT-rep2.bam, HEK293T-WT-rep3.bam representing the wildtype replicates
273 and HEK293T-KO-rep2.bam and HEK293T-KO-rep3.bam as the control
274 replicates,

- 275 1. Identify read error profile: run JACUSA2 in pairwise conditions mode
276 (call-2). The method requires the BAM files of the paired conditions,
277 the corresponding library information, and the output file name. Plus,
278 it allows filtering reads according to many parameters. Here, we con-
279 sider all sites with read coverage > 4. The output consists of a read
280 error profile where the format is a combination of BED6 with JA-
281 CUSA2 call-2 specific columns and common info columns: info, filter,
282 and ref. Check JACUSA2 manual for more details on JACUSA2 filter
283 and output options JAC [2021].


```

284 $ JACUSA2 2.0.0-RC22 call-2 -m 1 -q 1 -c 4 -p 10 -D -I -a D,Y
285 -P1 FR-SECONDSTRAND -P2 FR-SECONDSTRAND -r WT_vs_KO_call2_result.out
286 HEK293T-WT-rep2.bam, HEK293T-WT-rep3.bam HEK293T-KO-rep2.bam,
287 HEK293T-KO-rep3.bam

288 2. Preprocess JACUSA2 output: from the JACUSA2 call-2 output, select
289 all sites within 5mer of a central nucleotide 'A' flanked by 2 random nu-
290 cleotides (NNANN) and filter out sites of the homo-polymer regions
291 (JACUSA filter: Y). The 'README_processing.sh' bash script per-
292 forms the preprocessing step by providing the JAUSA2 call-2 output,
293 the reference genome and the output folder. The output of the prepro-
294 cessing is a text file 'call2_SitesExt2_indel.slim2.txt' containing tabular
295 features of the selected sites. Eventually, sites are characterized by the
296 main scores: mismatch, deletion, and insertion and additional infor-
297 mation: reference base, strand and position number within the specific
298 5mer context.

299 $ bash README_processing.sh WT_vs_KO_call2_result.out hg38.genome
300 GRCh38_96.fa path_to_output.

301 Then, using the R script 'HEK293_data_prep.R', rebuild the tabular
302 features such that the observations are only sites with a reference base
303 A. Each site is characterized by 15 features corresponding to the mis-
304 match, insertion and deletion scores for the observed site and its two
305 flanking positions. The output is an R object 'BigTable.rds', repre-
306 senting the matrix of Sites $\times$ 15 features. Be aware to precise the path
307 to outputs that contains already the preprocessed data and provide
308 the sample's name as a label of the analysis.

309 $ Rscript HEK293_data_prep.R path_to_output WT_vs_KO_call2_result.out

310 3. Extract m6A modification pattern: given the matrix of Sites $\times$ Features,
311 the next step is to learn a model representing the m6A modification
312 pattern. To this end, the conventional non-negative matrix factor-
313 ization (NMF) analysis is suggested Lee and Seung [1999]. Briefly,
314 NMF factorizes a non-negative data matrix  $X$  (here:  $n$  sites and  $m$ 
315 features) into two non-negative matrices as  $X \approx WH$ , such that  $W$ 
316 is an  $n \times k$  matrix containing basis vectors and  $H$  is an  $k \times m$  ma-
317 trix containing coefficient vectors. The coefficient vectors and their
318 combination can be viewed as a pattern for m6A modification. The
319 R script 'HEK293_data_prep_step2.R' allows generating patterns from
320 a subset of the data related to previously reported m6A sites. Here,
321 the unsupervised pattern training is based on 2401 common reported
322 m6A sites in Koh et al. [2019] . Based on the Silhouette and Coph-
323 netic Correlation indices January 20, 2022AL: explain more, we could

```

AL: ref-
erences

identify an optimal factorization rank of 7 (figure 4A). 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 2,3,4,6,7) (figure 4B). Interestingly, the linear combination of these five patterns in figure (4C) highlights the importance of position 3 and eventually the implication of all scores.

```
$ Rscript HEK293_data_prep_step2.R path_to_output miCLIP_union.bed
```

The 'miCLIP_union.bed' file contains all m6A sites reported in Koh et al. [2019] ().

4. Predict m6A modification: the additive linear combination of the coefficient vectors (patterns) with features can be used to predict m6A modification. We examine the ability of prediction on a subset of data of more than 1,98 million sites with 22248 miCLIP m6A validated sites . We opt for the linear combination of the five important patterns described in the previous section. The empirical Cumulative Distribution Function (eCDF) of the inferred scores shows clearly a significant difference between the different miCLIP m6A categories and the unmodified sites (figure 4D). As the number of negative samples is much larger than the number of positive samples, we particularly recommend investigating the Positive Predictive Value (PPV) of the predictions. Here, (figure 4E) shows a moderate PPV that increases with the cut-off. The R script 'HEK293_data_prep_step3.R' allows generating the scores, eCDF probabilities of modification, and the corresponding eCDF and PPV plots.

AL: reference + description

AL: reference + description of categories

```
$ Rscript HEK293_data_prep_step3.R path_to_output miCLIP_union.bed
```

Use Case 2: Comparison of wildtype and IVT samples

An alternative way to detect RNA modification is to compare modified sample to an *in-vitro* (IVT) synthesized control sample. Therefore, we benchmark JACUSA2 on a sample set of wildtype HEK293 cell line (modified sample) with two replicates (2 and 3) from Pratanwanich et al. [2021] and a modification-free RNA synthesized sample (control sample). The analysis steps are similar to case 1.

1. Identify read error profile: we use JACUSA2 call-2 with the same parameters as the previously described case. The input BAM files are associated to the replicates of the wildtype and IVT replicate samples.

```
$ JACUSA2 2.0.0-RC22 call-2 -m 1 -q 1 -c 4 -p 10 -D -I -a D,Y
-P1 FR-SECONDSTRAND
```

```

362 -P2 FR-SECONDSTRAND -r WT_vs_IVT_call2_result.out HEK293T-WT-rep2.bam,
363 HEK293T-WT-rep3.bam HEK293T-IVT-rep1.bam, HEK293T-IVT-rep2.bam

364 2. Preprocess JACUSA2 output: we select all sites within the specific
365 5mer (NNANN) and we consider the Y filter that excludes sites within
366 the homo-polymer regions.

367 $ bash README_processing.sh WT_vs_IVT_call2_result.out hg38.genome
368 GRCh38_96.fa path_to_output.

369 Then, we extract 5mer features such that the selected sites are repre-
370 sented by the three scores: Mismatch, Deletion and Insertion for the
371 observed site and its two flanking positions.

372 $ Rscript HEK293_data_prep.R path_to_output WT_vs_IVT_call2_result.out

373 3. Extract m6A modification pattern: using NMF factorization, we ex-
374 tract patterns from 2401 m6A sites (). Based on the silhouette and
375 cophenetic Correlation indices, we could identify an optimal factor-
376 ization rank of 6 (figure 5A). We determined the predominant factors
377 from matrix  $W$ . Herein, more than 80% of m6A modification sites
378 can be represented by four patterns (Patterns: 1,2,3,6) (figure 5E).
379 In agreement with case 1, the linear combination of the four patterns
380 confirms the importance of position 3 and the implication of all scores
381 as shown in figure (5C).

382 $ Rscript HEK293_data_prep_step2.R path_to_output miCLIP_union.bed

383 4. Predict m6A modifications: we evaluate the prediction ability of the
384 detected patterns on a test set of 22248 m6A sites. We consider the
385 linear combination of the four patterns (1,2,3,6). Figure 5D shows the
386 eCDF of the inferred scores. The difference between the cumulative
387 distribution of non miCLIP sites and miCLIP sites can be nicely ob-
388 served, while, the PPV plot shows a lower performance as compared
389 to case 1 (figure 5E). The decrease in performance is likely explained
390 by the absence of all modifications and not exclusively m6A in the
391 control condition which induces noise to the score estimation by JA-
392 CUSA2 call2 method .

393 $ Rscript HEK293_data_prep_step3.R path_to_output miCLIP_union.bed

```

AL: ref-
erence

CD:to
be con-
firmed

NOTES

Tips and Tricks

ACKNOWLEDGMENTS

The authors would like to thank Etienne Boileau, Thiago Britto Borges, Tobias Jakobi for proof-reading and comments. The authors are grateful to Marek Franitza for running the experiments on the 10x platform and to Christian Becker for running ONT sequencing. This work was supported by Informatics for Life funded by the Klaus Tschira Foundation.

REFERENCES

- Minimap2. <https://github.com/lh3/minimap2>. Accessed: 2022-01-19.
- Basecalling with guppy. <https://github.com/metagenomics/denbi-nanopore-training/blob/master/docs/basecalling/basecalling.rst>, 2019. Accessed: 2022-01-19.
- Jacusa2 manual. <https://github.com/dieterich-lab/JACUSA2>, 2021. Accessed: 2022-01-15.
- Samir Adhikari, Wen Xiao, Yong-Liang Zhao, and Yun-Gui Yang. m(6)a: Signaling for mrna splicing. *RNA biology*, 13:756–759, September 2016. ISSN 1555-8584. doi: 10.1080/15476286.2016.1201628.
- Ina Anreiter, Quoseena Mir, Jared T. Simpson, Sarath C. Janga, and Matthias Soller. New twists in detecting mrna modification dynamics. *Trends in biotechnology*, 39:72–89, January 2021. ISSN 1879-3096. doi: 10.1016/j.tibtech.2020.06.002.
- Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and Gideon Rechavi. Topology of the human and mouse m6a rna methylomes revealed by m6a-seq. *Nature*, 485:201–206, April 2012. ISSN 1476-4687. doi: 10.1038/nature11112.
- David Garcias Morales and José L. Reyes. A birds’-eye view of the activity and specificity of the mrna m, javax.xml.bind.jaxbelement@6d66739e, a methyltransferase complex. *Wiley interdisciplinary reviews. RNA*, 12: e1618, January 2021. ISSN 1757-7012. doi: 10.1002/wrna.1618.
- Guifang Jia, Ye Fu, Xu Zhao, Qing Dai, Guanqun Zheng, Ying Yang, Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, and Chuan He.

428 N6-methyladenosine in nuclear rna is a major substrate of the obesity-
429 associated fto. *Nature chemical biology*, 7:885–887, October 2011. ISSN
430 1552-4469. doi: 10.1038/nchembio.687.

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

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

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

443 Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christo-
444 pher E. Mason, and Samie R. Jaffrey. Comprehensive analysis of mrna
445 methylation reveals enrichment in 3’ utrs and near stop codons. *Cell*, 149:
446 1635–1646, June 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.05.003.

447 Ploy N Pratanwanich, Fei Yao, Ying Chen, Casslynn WQ Koh, Yuk Kei
448 Wan, Christopher Hendra, Polly Poon, Yeek Teck Goh, Phoebe ML Yap,
449 Jing Yuan Chooi, et al. Identification of differential rna modifications
450 from nanopore direct rna sequencing with xpore. *Nature Biotechnology*,
451 39(11):1394–1402, 2021.

452 Jean-Yves Roignant and Matthias Soller. m,
453 javax.xml.bind.jaxbelement@8cec19d, a in mrna: An ancient mech-
454 anism for fine-tuning gene expression. *Trends in genetics : TIG*, 33:
455 380–390, June 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2017.04.003.

456 Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic rna
457 modifications in gene expression regulation. *Cell*, 169:1187–1200, June
458 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.045.

459 Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how:
460 Context-dependent functions of rna methylation writers, readers, and
461 erasers. *Molecular cell*, 74:640–650, May 2019. ISSN 1097-4164. doi:
462 10.1016/j.molcel.2019.04.025.

463 Sara Zaccara, Ryan J. Ries, and Samie R. Jaffrey. Reading, writing and
464 erasing mrna methylation. *Nature reviews. Molecular cell biology*, 20:608–
465 624, October 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0168-5.

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

470 Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min
471 Huang, Charles J. Li, Cathrine B. Vågbø, Yue Shi, Wen-Ling Wang, Shu-
472 Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin
473 Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan,
474 Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne
475 Klungland, Yun-Gui Yang, and Chuan He. Alkbh5 is a mammalian rna
476 demethylase that impacts rna metabolism and mouse fertility. *Molecular*
477 *cell*, 49:18–29, January 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.
478 10.015.

FIGURE CAPTIONS

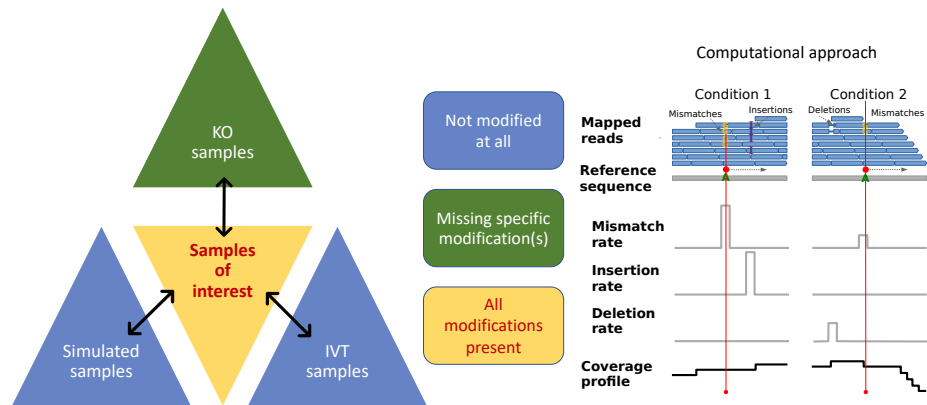


Figure 1: **General outline of RNA modification detection by JACUSA2.** A key feature of our approach is that multiple replicates can be compared as shown on the left. Samples of interests where all modifications are present could be compared with either KO samples where the modification of interest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

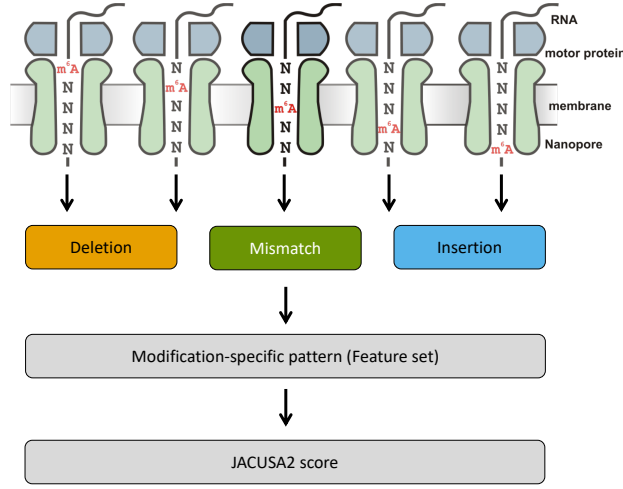


Figure 2: **Motivation of 5mer context for RNA modification mapping.** The nanopore covers 5 consecutive RNA residues. That is why we consider a 5mer context and derive 3 principal features for every position within a given 5 mer (15 features in total, with a central A residue in this example). We evaluate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

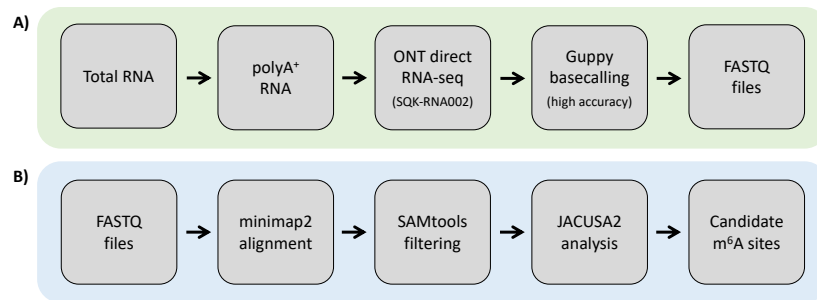


Figure 3: **Experimental and computational workflow.** tbd

Software	Version	Description
Minimap2	https://github.com/lh3/minimap2 v2.22 or later	https://lh3.github.io/minimap2/
samtools	https://github.com/samtools/samtools v1.12 or later	http://samtools.github.io/
JAVA	openjdk 11.0.12 2021-07-20 - JAVA 11 or later	OpenJDK Runtime Environment
R	https://www.r-project.org/ version 3.5.1 or later	The R Project for Statistical Computing
PERL	https://www.perl.org/ 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	https://github.com/arq5x/bedtools2 version 2.29.2 or later	Perl is a highly capable, feature-rich programming language
NanoSim	https://github.com/bcgsc/NanoSim 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

480 TABLE CAPTIONS

481 TABLES

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

Table 2: **R Package dependencies** blubba

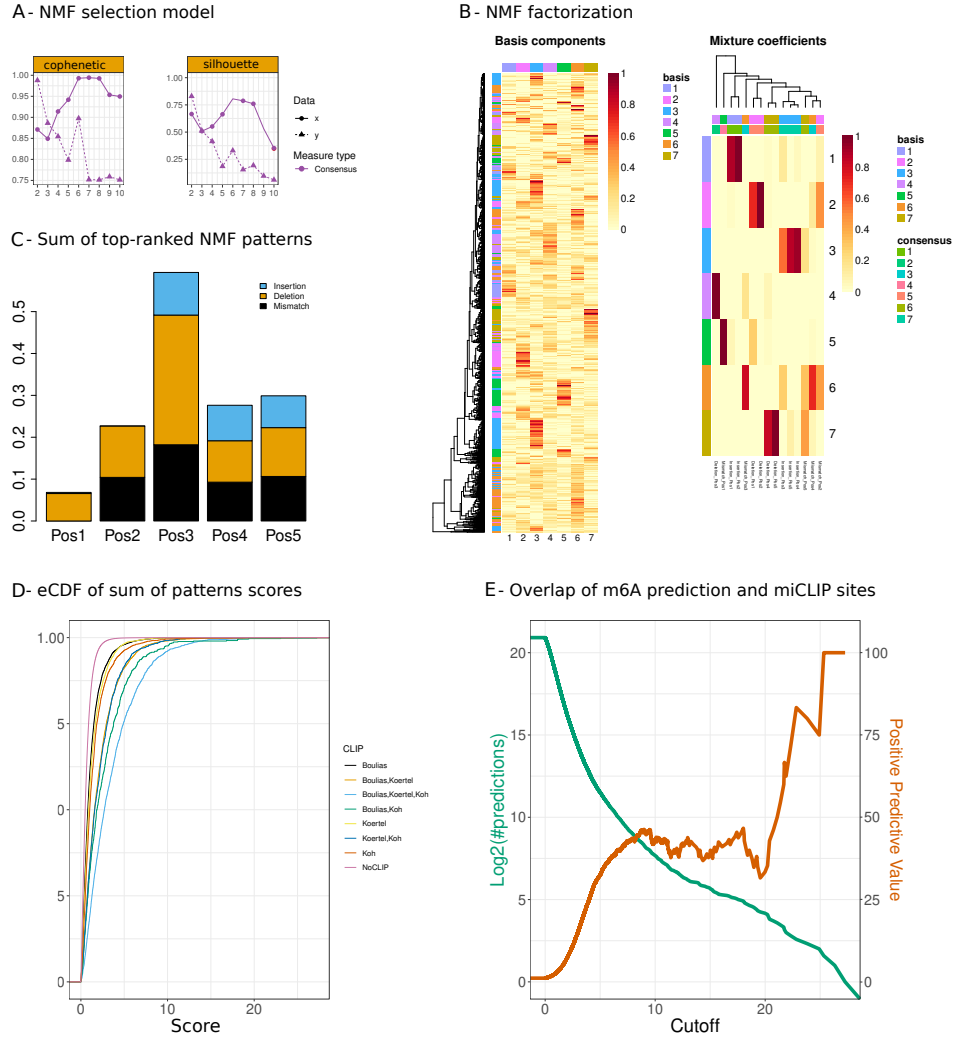
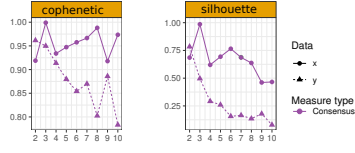
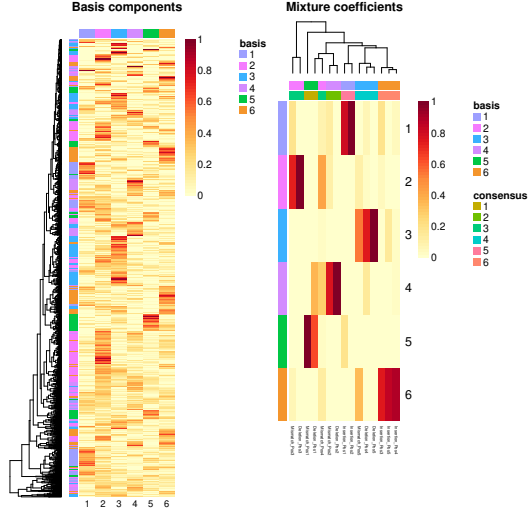


Figure 4: Case 01. WT versus KO. A: NMF rank selection. Comparison of cophenetic correlation and silhouette resulting from the NMF factorization of the original and the randomized data **B:** NMF result represented by the basis matrix W and the coefficient matrix H . The matrix H induces the RNA modification pattern. **B:** Barplots representing the linear combination of the top 5 patterns (y-axis) by position in the specific 5mer context (x-axis) and the score type: mismatch, deletion and insertion (resp. black, orange and blue). The 5 patterns (coefficient vectors: 2,3,4,6,7) are selected according to the predominant columns in matrix W . **C:** Score distribution inferred from the combined patterns, stratified by the different categories of miCLIP validated sites and non miCLIP sites. **D:** Number of predicted m6A sites (green) and Positive Predictive Value (PPV) of predicted m6A sites that overlap with miCLIP sites (orange).

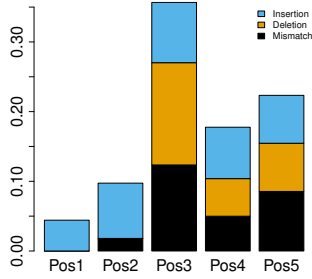
A - NMF selection model



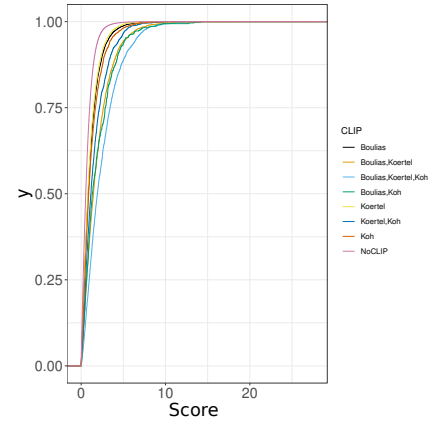
B - NMF factorization



C - Sum of top-ranked NMF patterns



D - eCDF of sum of patterns scores



E - Overlap of m6A prediction and miCLIP sites

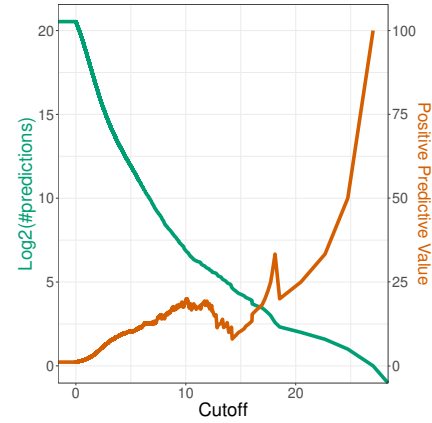


Figure 5: **Case 02. WT versus IVT.** **A:** NMF rank selection. Comparison of cophenetic correlation and silhouette resulting from the NMF factorization of the original and the randomized data **B:** NMF result represented by the basis matrix W and the coefficient matrix H . The matrix H induces the RNA modification pattern. **B:** Barplots representing the linear combination of the top 4 patterns (y-axis) by position in the specific 5mer 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 . **C:** Score distribution inferred from the combined patterns, stratified by the different categories of miCLIP validated sites and non miCLIP sites. **D:** Number of predicted m6A sites (green) and Positive Predictive Value (PPV) of predicted m6A sites that overlap with miCLIP sites (orange).