

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 or 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. For the IVT readout, we adopted real-time
241 base calling with the MinKNOW-embedded Guppy basecaller. Other-
242 wise, Guppy basecaller software can be used; in this case, the basecaller
243 requires the path to FAST5 files, the output folder, and the config file
244 or the flowcell/kit combination. The output is FASTQ files that can
245 be compressed using the option "--compress-fastq".

```

246 $ guppy_basecaller --compress_fastq -i path_to_fast5 -s path_to_output
247 -c config_file.cfg --cpu_threads_per_caller 14 --num_callers
248 1
249 Set the number of threads "cpu_threads_per_caller" and the number
250 of parallel basecallers "num_caller" according to your resources. Ad-
251 ditional details can be found in Gup [2019].

252 2. Align reads to the transcriptome using Minimap2 software. The out-
253 put is a SAM file that has to be converted to a compressed form as
254 BAM file using SAMtools command. The alignment requires the ref-
255 erence sequence. Here, we used GRCh38 Ensembl annotation and
256 FASTA file release version 96. For Direct RNA Sequencing, it is rec-
257 ommended to use the default setting "-ax map-ont", "-uf" to force the
258 alignment to the forward strand of the reference, and a small k-mer
259 size "-k [=14]" to enhance sensitivity. We recommend to output pri-
260 mary alignments and up to "-N [=100]" top secondary alignments if
261 the ratio of their chaining scores compared to the corresponding pri-
262 mary alignments is higher than "-p [=1]" . Customize the number of
263 threads "-t" according to your resources. Check Minimap2 manual for
264 more details [Min].

265 $ minimap2 -t 5 -ax map-ont -uf -k14 -p 1.0 -N 100 reference.fasta
266 Reads.fastq |samtools view -bS > mapping.bam

267 3. Map RNA modifications using JACUSA2 software [Piechotta et al.,
268 2021]. JACUSA2 rapidly detects RNA modifications based on a com-
269 parative strategy where the mapping features (mismatch, insertion and
270 deletion) of a sample of interest is compared to a reference sequence
271 (call-1) or against a sample without RNA modifications, e.g. a knock-
272 out of an RNA modifying enzyme or an IVT (call-2). Moreover, it
273 allows the integration of information from replicate experiments. Fur-
274 ther details on how to use JACUSA2 is presented in the next section.
275 Beforehand, JACUSA2 requires sorted and indexed BAM files. To sort
276 a BAM file and create a BAM file index use the following SAMtools
277 commands.

278 $ samtools sort mapping.bam mapping.sorted.bam
279 $ samtools index mapping.sorted.bam

```

CD: not sure about min-
imap2 step specifically these 2 parameters -N and -p. need to be validated

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

281 The conventional way to detect RNA modifications using direct RNA se-
 282 quencing is to compare a modified sample to an unmodified control sample.
 283 To assess the ability of JACUSA2 in this case, we used a published dataset

284 of HEK293 cell lines to detect m6A modification [Pratanwanich et al., 2021].
 285 The benchmark is composed of two samples from two conditions: wild-type
 286 cells (modified RNAs) and Mettl3 knockout cells (unmodified RNAs) with
 287 two replicates (2 and 3). The FASTQ files are preprocessed and mapped
 288 according to the steps described in the previous section. The analysis is vali-
 289 dated against reported m6A sites in the three miCLIP-based studies Boulias
 290 et al. [2019], Koh et al. [2019], Körtel et al. [2021] (figure 4).

291 Given the preprocessed mapped reads as inputs (BAM files) 'HEK293T-
 292 WT-rep2.bam, HEK293T-WT-rep3.bam representing the wildtype replicates
 293 and HEK293T-KO-rep2.bam and HEK293T-KO-rep3.bam as the control
 294 replicates,

- 295 1. Identify read error profile: run JACUSA2 in pairwise conditions mode
 296 (call-2). The method requires BAM files of the paired conditions, the
 297 corresponding library information "-P1" and "-P2", and the output
 298 file name "-r". In addition to the mismatch score, add "-D" and
 299 "-I" to output the deletion and insertion scores. JACUSA2 allows
 300 filtering reads according to many parameters. Here, we consider all
 301 sites with base calling quality "-q [> 1]", mapping quality "-m [>
 302 1]" and read coverage "-c [> 4]". Plus, it provides a filter feature
 303 to improve sensitivity. Here, we consider the filter of sites within
 304 homopolymer regions "-a [=Y]". The output consists of a read error
 305 profile where the format is a combination of BED6 with JACUSA2
 306 call-2 specific columns and common info columns: info, filter, and
 307 ref. Check JACUSA2 manual for more details on JACUSA2 filter and
 308 output options [JAC, 2021]. The number of threads can be customized
 309 via the parameter "-p".

```
310 $ JACUSA2 2.0.0-RC22 call-2 -q 1 -m 1 -c 4 -p 10 -D -I -a Y
311 -P1 FR-SECONDSTRAND -P2 FR-SECONDSTRAND -r WT_vs_KO_call2_result.out
312 HEK293T-WT-rep2.bam, HEK293T-WT-rep3.bam HEK293T-KO-rep2.bam,
313 HEK293T-KO-rep3.bam
```

- 314 2. Preprocess JACUSA2 output: from JACUSA2 call-2 output, select all
 315 sites within 5-mer of a central nucleotide 'A' flanked by 2 random nu-
 316 cleotides (NNANN) and filter out sites of the homo-polymer regions
 317 (JACUSA filter: Y). 'README_processing.sh' bash script performs
 318 the preprocessing step by providing the output of JAUSA2 call-2, the
 319 reference genome, and the output folder. The output of the prepro-
 320 cessing is a text file 'call2_SitesExt2_indel_slim2.txt' containing tabular
 321 features of the selected sites. Eventually, sites are characterized by the
 322 main scores: mismatch, deletion, and insertion and additional infor-
 323 mation: reference base, strand and position number within the specific
 324 5-mer context.

```

325 $ bash README_processing.sh WT_vs_KO_call2_result.out hg38.genome
326 GRCh38_96.fa path_to_output.

```

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

```

335 $ Rscript HEK293_data_prep.R path_to_output WT_vs_KO_call2_result.out

```

336 3. Extract m6A modification pattern: given the matrix of Sites \times Features,
337 the next step is to learn a model representing the m6A modification
338 pattern. To this end, the conventional non-negative matrix factoriza-
339 tion (NMF) analysis is suggested [Lee and Seung, 1999]. Briefly, NMF
340 factorizes a non-negative data matrix X (here: n sites and m features)
341 into two non-negative matrices as $X \approx WH$, such that W is an $n \times k$
342 matrix containing basis vectors and H is an $k \times m$ matrix containing
343 coefficient vectors. The coefficient vectors and their combination can
344 be viewed as a pattern for m6A modification. The rank of factorization
345 k is a critical parameter that affects the performance substantially. We
346 suggest to select the rank k according to the method of Frigyesi and
347 Höglund [2008] by looking at silhouette [Rousseeuw, 1987] and cophe-
348 netic correlation [Brunet et al., 2004] indices. Accordingly, the perfor-
349 mance indices are computed for different choices of rank ($k < n, m$)
350 and compared to the performance of a random permutation of the
351 original data. Subsequently, the chosen rank corresponds to the value
352 with the largest difference between slopes of the original and the ran-
353 domized data. Here, the unsupervised pattern training is based on the
354 consensus set of 2,401 m6A sites reported in the three miCLIP-based
355 studies mentioned earlier. Based on the silhouette and cophenetic cor-
356 relation indices, we could identify an optimal factorization rank of 7
357 (figure 5A). We then analyzed the identified patterns. According to
358 the membership indicator of each site in matrix W , more than 80% of
359 m6A modification sites can be represented by five patterns (Patterns
360 2,3,4,6,7) (figure 5B). Interestingly, the linear combination of these
361 five patterns in figure (5C) highlights the importance of position 3
362 and eventually the implication of all scores.

363 Use the R script 'HEK293_data_prep_step2.R' to generate patterns
364 and provide the set of modified sites as a ground truth. Here, the
365 'miCLIP_union.bed' file contains the m6A sites from three miCLIP-

366 based studies. A miCLIP annotation, reflecting studies (hence, the
367 consensus) wherein the modification is reported, is added to each site.

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

369 4. Predict m6A modification: the additive linear combination of the co-
370 efficient vectors (patterns) with the 15 features can be used to predict
371 m6A modification. We examine the ability of prediction on a sub-
372 set of data of more than 1,98 million sites with 22,248 miCLIP m6A
373 sites. We opt for the linear combination of the five important patterns
374 described in step 3. The empirical Cumulative Distribution Function
375 (eCDF) of the inferred scores shows clearly a significant difference be-
376 tween the different miCLIP m6A categories (miCLIP annotation) and
377 the unmodified sites (figure 5D). As the number of negative samples
378 is much larger than the number of positive samples, we particularly
379 recommend investigating the Positive Predictive Value (PPV) of the
380 predictions. Here, figure 5E shows a moderate PPV that increases
381 with the cut-off. Use the R script 'HEK293_data_prep_step3.R' to es-
382 timate scores and eCDF probability of modification from the selected
383 patterns and to produce the corresponding PPV plot.

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

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

386 An alternative way to detect RNA modification is to compare a modi-
387 fied sample to an *in-vitro* (IVT) synthesized control sample. Therefore,
388 we benchmark JACUSA2 on a sample set of wild-type HEK293 cell lines
389 (modified sample) with two replicates (2 and 3) from Pratanwanich et al.
390 [2021] and a modification-free RNA synthesized sample (control sample).
391 The analysis steps are similar to case 1. We evaluate the analysis against
392 miCLIP m6A sites (figure 4).

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

396 `$ JACUSA2 2.0.0-RC22 call-2 -m 1 -q 1 -c 4 -p 10 -D -I -a D,Y`
397 `-P1 FR-SECONDSTRAND`
398 `-P2 FR-SECONDSTRAND -r WT_vs_IVT_call2_result.out HEK293T-WT-rep2.bam,`
399 `HEK293T-WT-rep3.bam HEK293T-IVT-rep1.bam, HEK293T-IVT-rep2.bam`

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

```

403 $ bash README_processing.sh WT_vs_IVT_call2_result.out hg38.genome
404 GRCh38_96.fa path_to_output.
405
406 Then, we extract 5-mer features such that the selected sites are rep-
407 resented by the three scores: mismatch, deletion and insertion for the
408 observed site and its two flanking positions.
409
410 $ Rscript HEK293_data_prep.R path_to_output WT_vs_IVT_call2_result.out
411
412 3. Extract m6A modification pattern: using NMF factorization, we ex-
413 tract patterns from 1,905 sites reported as modified in the three miCLIP-
414 based studies. Based on the silhouette and cophenetic correlation in-
415 dices, we could identify an optimal factorization rank of 6 (figure 6A).
416 We determined the predominant factors from matrix  $W$ ; accordingly,
417 more than 80% of m6A modification sites can be represented by four
418 patterns (Patterns: 1,2,3,6) (figure 6B). In agreement with case 1, the
419 linear combination of the four patterns confirms the importance of
420 position 3 and the implication of all scores as shown in figure (6C).
421
422 $ Rscript HEK293_data_prep_step2.R path_to_output miCLIP_union.bed
423
424 4. Predict m6A modifications: we evaluate the prediction ability of the
425 detected patterns on a test set of almost 1,52 million sites where
426 17,021 are miCLIP-m6A modified. We consider the linear combina-
427 tion of the four patterns (1,2,3,6). Figure 6D shows the eCDF of the
428 inferred scores. The difference between the cumulative distribution of
429 non miCLIP sites and miCLIP sites can be nicely observed, while, the
430 PPV plot shows a lower performance as compared to case 1 (figure
431 6E). The decrease in performance is likely explained by the absence
432 of all modifications and not exclusively m6A in the control condition,
433 which may induce noise to the score estimation by JACUSA2 call-2 .
434
435 $ Rscript HEK293_data_prep_step3.R path_to_output miCLIP_union.bed

```

CD:to
be con-
firmed

430 NOTES

431 Tips and Tricks

432 ACKNOWLEDGMENTS

433 The authors would like to thank Etienne Boileau, Thiago Britto Borges,
434 Tobias Jakobi for proof-reading and comments. The authors are grateful
435 to Marek Franitza for running the experiments on the 10x platform and to
436 Christian Becker for running ONT sequencing. This work was supported by
437 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.
- Konstantinos Boulas, Diana Toczyłowska-Socha, Ben R Hawley, Noa Liberman, Ken Takashima, Sara Zaccara, Théo Guez, Jean-Jacques Vasseur, Françoise Debart, L Aravind, et al. Identification of the m6am methyltransferase pcif1 reveals the location and functions of m6am in the transcriptome. *Molecular cell*, 75(3):631–643, 2019.
- Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the national academy of sciences*, 101(12):4164–4169, 2004.
- 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.
- Attila Frigyesi and Mattias Höglund. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. *Cancer informatics*, 6:CIN-S606, 2008.
- 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.

475 N6-methyladenosine in nuclear rna is a major substrate of the obesity-
476 associated fto. *Nature chemical biology*, 7:885–887, October 2011. ISSN
477 1552-4469. doi: 10.1038/nchembio.687.

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

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

488 Nadine Körtel, Cornelia Rücklé, You Zhou, Anke Busch, Peter Hoch-Kraft,
489 Reymond FX Sutandy, Jacob Haase, Mihika Pradhan, Michael Musheev,
490 Dirk Ostareck, et al. Deep and accurate detection of m6a rna modifications
491 using miclip2 and m6aboost machine learning. *bioRxiv*, pages 2020–12,
492 2021.

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

495 Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christo-
496 pher E. Mason, and Samie R. Jaffrey. Comprehensive analysis of mrna
497 methylation reveals enrichment in 3’ utrs and near stop codons. *Cell*, 149:
498 1635–1646, June 2012. ISSN 1097-4172. doi: 10.1016/j.cell.2012.05.003.

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

501 Ploy N Pratanwanich, Fei Yao, Ying Chen, Casslynn WQ Koh, Yuk Kei
502 Wan, Christopher Hendra, Polly Poon, Yeek Teck Goh, Phoebe ML Yap,
503 Jing Yuan Chooi, et al. Identification of differential rna modifications
504 from nanopore direct rna sequencing with xpore. *Nature Biotechnology*,
505 39(11):1394–1402, 2021.

506 Jean-Yves Roignant and Matthias Soller. m,
507 javax.xml.bind.jaxbelement@8cec19d, a in mrna: An ancient mech-
508 anism for fine-tuning gene expression. *Trends in genetics : TIG*, 33:
509 380–390, June 2017. ISSN 0168-9525. doi: 10.1016/j.tig.2017.04.003.

510 Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic rna
511 modifications in gene expression regulation. *Cell*, 169:1187–1200, June
512 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.045.

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

516 Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how:
517 Context-dependent functions of rna methylation writers, readers, and
518 erasers. *Molecular cell*, 74:640–650, May 2019. ISSN 1097-4164. doi:
519 10.1016/j.molcel.2019.04.025.

520 Sara Zaccara, Ryan J. Ries, and Samie R. Jaffrey. Reading, writing and
521 erasing mrna methylation. *Nature reviews. Molecular cell biology*, 20:608–
522 624, October 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0168-5.

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

527 Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min
528 Huang, Charles J. Li, Cathrine B. Vågbø, Yue Shi, Wen-Ling Wang, Shu-
529 Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin
530 Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan,
531 Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne
532 Klungland, Yun-Gui Yang, and Chuan He. Alkbh5 is a mammalian rna
533 demethylase that impacts rna metabolism and mouse fertility. *Molecular*
534 *cell*, 49:18–29, January 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.
535 10.015.

FIGURE CAPTIONS

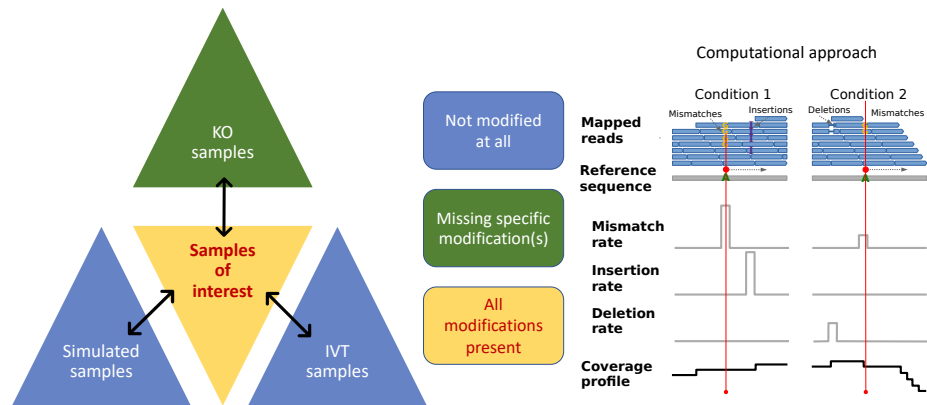


Figure 1: **General outline of RNA modification detection by JACUSA2.** A key feature of our approach is that multiple replicates can be compared as shown on the left. Samples of interests where all modifications are present could be compared with either KO samples where the modification of interest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

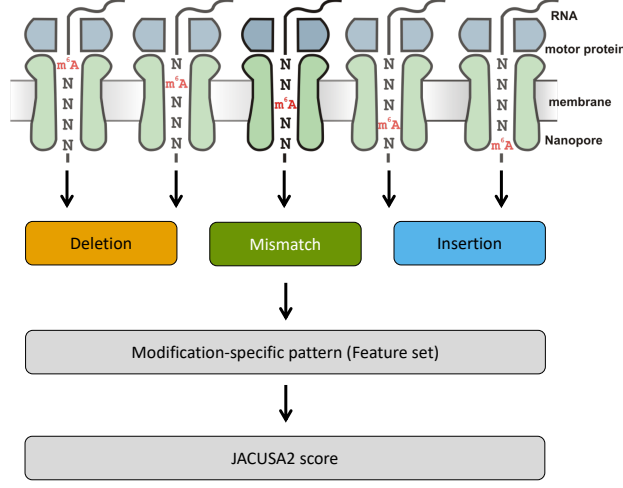


Figure 2: **Motivation of 5-mer context for RNA modification mapping.** The nanopore covers 5 consecutive RNA residues. That is why we consider a 5-mer context and derive 3 principal features for every position within a given 5-mer (15 features in total, with a central A residue in this example). We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

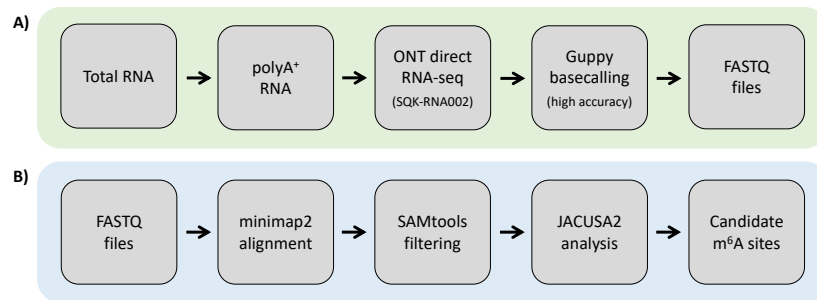


Figure 3: **Experimental and computational workflow.** tbd

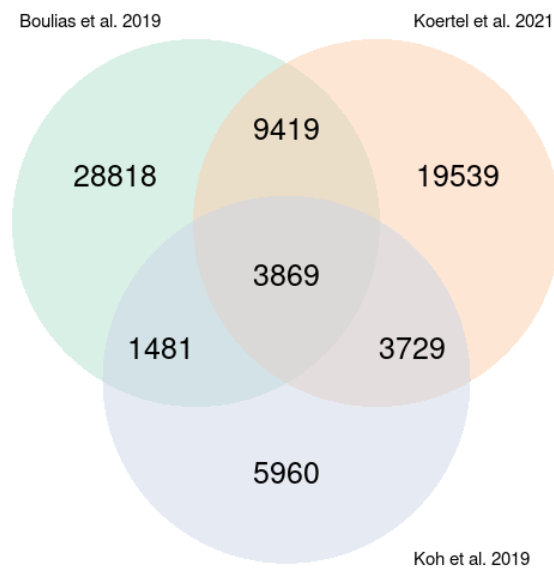


Figure 4: m6A sites reported in the three miCLIP-based studies: Boulias et al. [2019], Koh et al. [2019] and Körtel et al. [2021].

Software	Version	Description
Minimap2	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

537 TABLE CAPTIONS

538 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

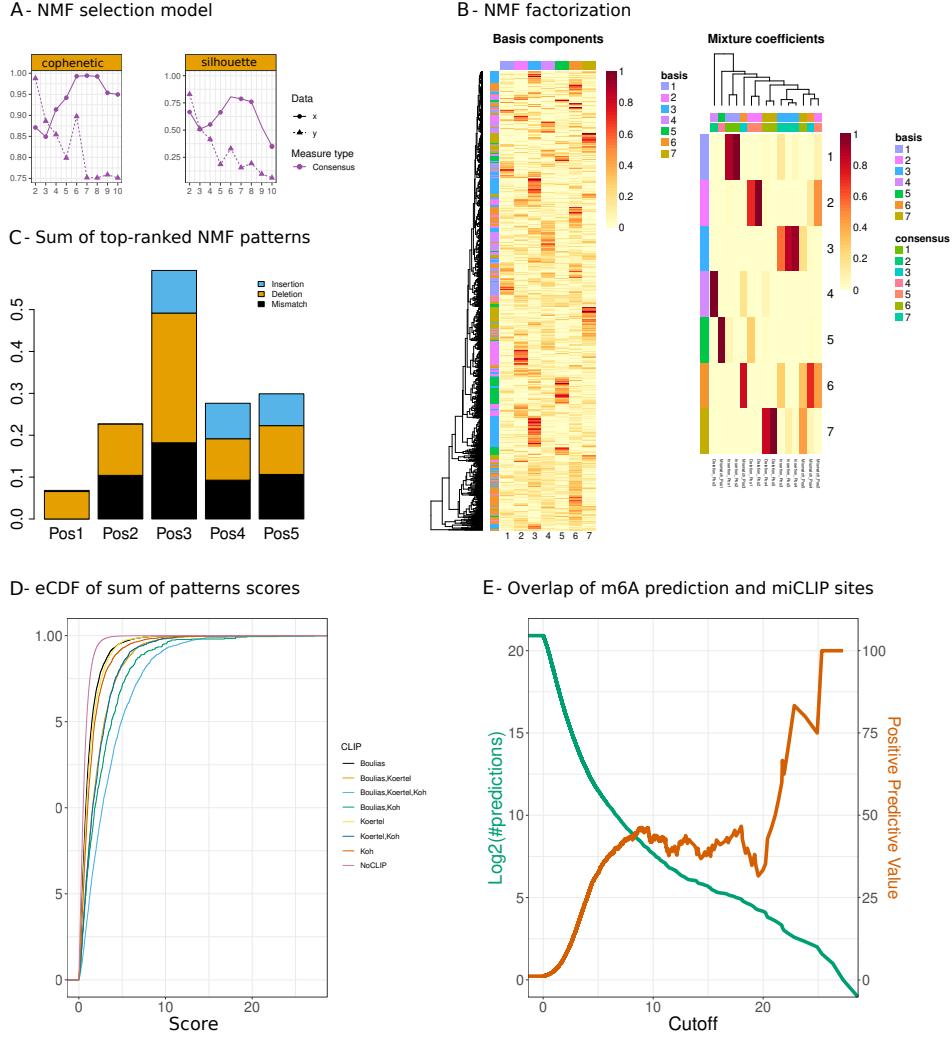


Figure 5: **Case 1. WT versus KO.** **A:** NMF rank selection. Comparison of cophenetic correlation and silhouette indices obtained from the NMF factorization of the original and the randomized data **B:** NMF result represented by the basis matrix W and the coefficient matrix H . The matrix H induces the RNA modification pattern. **B:** Barplots representing the linear combination of the top 5 patterns (y-axis) by 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: 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).

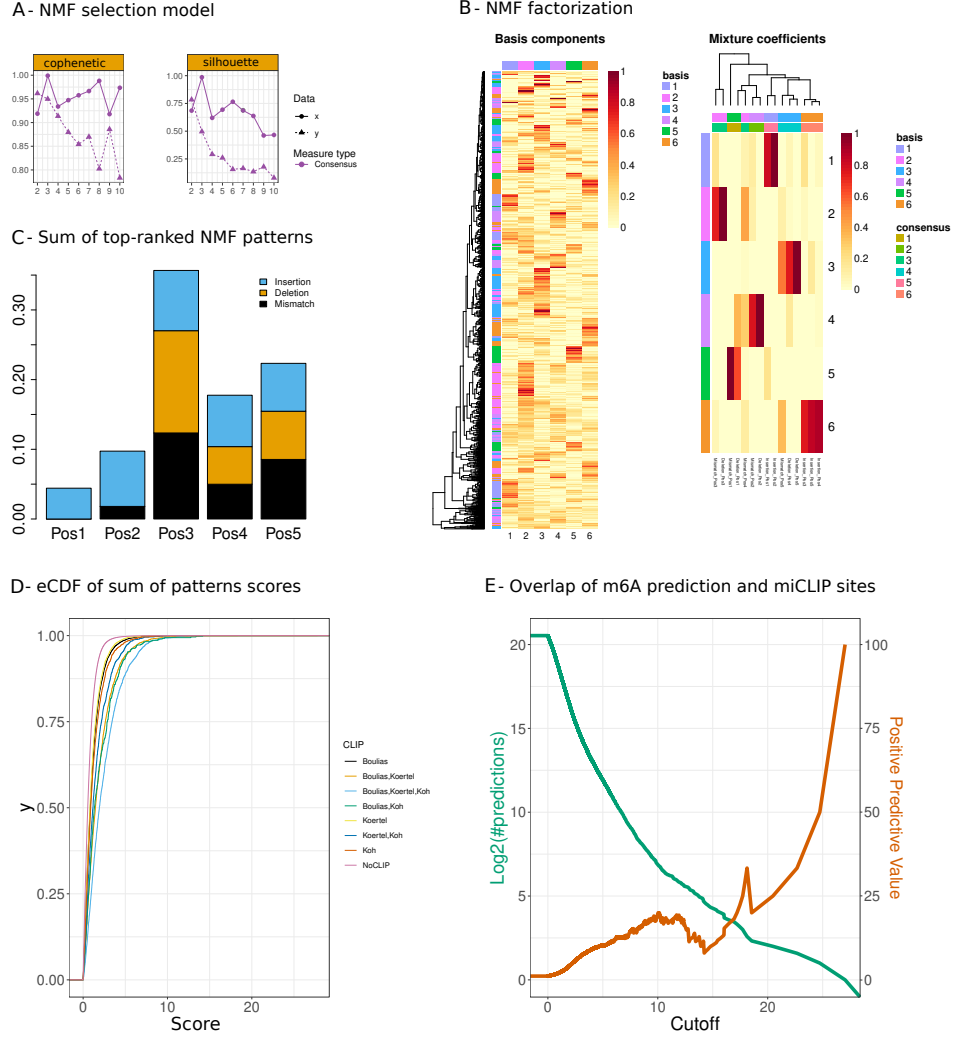


Figure 6: **Case 2. WT versus IVT.** **A:** NMF rank selection. Comparison of cophenetic correlation and silhouette indices resulting from the NMF factorization of the original and the randomized data **B:** NMF result represented by the basis matrix W and the coefficient matrix H . The matrix H induces the RNA modification pattern. **B:** Barplots representing the linear combination of the top 4 patterns (y-axis) by 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 . **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).