

# 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

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.

---

<sup>\*</sup>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<sup>+</sup> RNA isolated from total RNA e.g. with Oligotex  
 69 mRNA kit (Qiagen) or Dynabeads oligo dT<sub>25</sub> 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<sup>+</sup> RNA isolated from total RNA e.g. with Oligotex  
96 mRNA kit (Qiagen) or Dynabeads oligo dT<sub>25</sub> beads (Thermo Fisher  
97 Scientific). Store RNA at -80 °C and the mRNA purification kit as  
98 recommended by the manufacturer
- 99 2. 10  $\mu$ M oligo(dT)-VN RT primer. TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN.  
100 Store at -20 °C.
- 101 3. 20  $\mu$ M template switching oligo (TSO). ACTCTAATACGACTCAC-  
102 TATAGGGAGAGGGCrGrG+G. Store at -20 °C.

- 103 4. 10  $\mu$ 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](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<sup>+</sup> RNA to a total volume of 9  $\mu$ l with nuclease-  
143 free water. Complete RT adapter ligation reaction (in 0.2 ml PCR  
144 tube) with 3  $\mu$ l NEBNext Quick Ligation Reaction Buffer, 0.5  $\mu$ l  
145 RNA CS (RCS, from SQK-RNA002), 1  $\mu$ l RT-Adapter (RTA, from  
146 SQK-RNA002) and 1.5  $\mu$ l T4 DNA Ligase. Incubate 10 min at room  
147 temperature.
- 148 2. Prepare reverse transcription master mix on ice during ligation: 9  $\mu$ l  
149 nuclease-free water, 2  $\mu$ l 10 mM dNTPs, 8  $\mu$ l 5x SuperScript IV first  
150 strand buffer, 4  $\mu$ l 0.1 mM DTT.
- 151 3. Add the reverse transcription master mix to the ligation reaction and  
152 mix by pipetting. Add 2  $\mu$ 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  $\mu$ 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  $\mu$ 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  $\mu$ 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  $\mu$ 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  $\mu$ l eluate: 8  
169  $\mu$ l NEBNext Quick Ligation Reaction Buffer, 6  $\mu$ l RMX (from SQK-  
170 RNA002), 3  $\mu$ l nuclease-free water, 3  $\mu$ l T4 DNA Ligase. Mix by  
171 pipetting and incubate 10 min at room temperature.

- 172 8. Add 40  $\mu\text{l}$  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  $\mu\text{l}$  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  $\mu\text{l}$  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  $\mu\text{l}$   
182 eluate in a fresh 1.5 ml DNA LoBind tube.
- 183 11. Quantify 1  $\mu\text{l}$  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/ $\mu\text{l}$ .
- 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<sup>+</sup> RNA at least 1,000  
189 available pores are recommended.
- 190 13. Prepare Priming Mix by adding 30  $\mu\text{l}$  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  $\mu\text{l}$  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  $\mu\text{l}$  library with 17.5  $\mu\text{l}$  nuclease-free water and 37.5  $\mu\text{l}$  RNA run-  
198 ning buffer (RRB, from SQK-RNA002) and mix by pipetting. Open  
199 the priming port and the sample port. Load 200  $\mu\text{l}$  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<sup>+</sup> RNA to a total volume of 6  $\mu$ l with nuclease-  
211 free water. Add 1  $\mu$ l each of 10  $\mu$ 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  $\mu$ l 4x template switching RT buffer, 0.5  $\mu$ l 20  $\mu$ M TSO,  
215 1  $\mu$ l 10x template switching RT enzyme mix and mix by pipetting.  
216 Combine with 6  $\mu$ 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  $\mu$ l Q5 Hot Start High-Fidelity 2X Master Mix, 5  $\mu$ l RNase H, 2  $\mu$ l 10  
220  $\mu$ M T7 extension primer, 33  $\mu$ 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  $\mu$ l elution  
225 buffer. Determine concentration on a Nanodrop spectrophotometer.  
226 cDNA may be stored at -20 °C.
- 227 5. Combine 8  $\mu$ l cDNA for *in vitro* transcription with 2  $\mu$ 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  $\mu$ l Turbo DNase. Mix by pipet-  
231 ting and incubate 15 min at 37 °C.
- 232 7. Adjust reaction volume to 100  $\mu$ 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  $\mu$ 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 stored in FAST5  
240 file using Guppy. The output is FASTQ files. For the IVT readout, we  
241 adopted real-time base calling with the integrated MinKNOW Guppy.  
242 Otherwise, Guppy base caller software could be used. The base caller  
243 requires the path to FAST5 files, the output path, and the config file or  
244 the flowcell/kit combination. More details can be found in Piechotta  
245 [a].
- ```

246 $ guppy_basecaller --compress_fastq -i path_to_fast5 -s path_to_output
247 --cpu_threads_per_caller 14 --num_callers 1 -c config_file.cfg

```

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

```
255 $ minimap2 --secondary=no -ax map-ont -uf -k14 reference.fasta  

  256 Reads.fastq |samtools view -bS > mapping.bam
```

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

```
259 $ samtools sort mapping.bam mapping.sorted.bam  

  260 $ samtools index mapping.sorted.bam
```

Explain  
the rest  
of con-  
sidered  
parame-  
ters

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

262 We used a published dataset of Hek293 cell line Pratanwanich et al. [2021].  
 263 The benchmark is composed of two samples from two conditions: wild type  
 264 cells (modified RNAs) and Mettl3 knockout cells (unmodified RNAs) with  
 265 two replicates (2 and 3). The FASTQ files are preprocessed and mapped  
 266 according to the steps described above.

267 Given the preprocessed mapped reads as input (BAM files) 'HEK293T-  
 268 WT-rep2.bam, HEK293T-WT-rep3.bam representing the wild type repli-  
 269 cates and HEK293T-KO-rep2.bam and HEK293T-KO-rep3.bam as the con-  
 270 trol replicates,

271 1. Identify read error profile: run JACUSA2 in paired samples mode (call-  
 272 2). The method requires setting the BAM files of the paired conditions,  
 273 the corresponding library information, and the output file. Plus, it  
 274 allows filtering reads according to many parameters. Here, we consider  
 275 all sites with read coverage > 4. The output consists of a read error  
 276 profile where the format is a combination of BED6 with JACUSA2  
 277 call-2 specific columns and common info columns: "info", "filter", and  
 278 "ref". Check JACUSA2 manual for more details on JACUSA2 filter  
 279 and output options Piechotta [b].

```
280 $ JACUSA2 2.0.0-RC22 call-2 -m 1 -q 1 -c 4 -p 10 -D -I -a D,Y  

  281 -P1 FR-SECONDSTRAND -P2 FR-SECONDSTRAND -r WT_vs_KO_call2_result.out  

  282 HEK293T-WT-rep2.bam, HEK293T-WT-rep3.bam HEK293T-KO-rep2.bam,  

  283 HEK293T-KO-rep3.bam
```

284 2. Preprocess JACUSA2 output: given the JACUSA2 output, select non-  
 285 overlapping sites of homo-polymer regions (JACUSA filter: Y) and



286 within a 5mer of a central nucleotide A flanked by 2 adjacent random  
287 nucleotides (NNANN). Selected sites are characterized by the inser-  
288 tion, deletion and mismatch scores, and the position number within the  
289 specific 5mer context. The 'README\_processing.sh' bash script per-  
290 forms the preprocessing step and produces a text file 'call2\_SitesExt2\_indel\_slim2.txt'  
291 containing tabular features of the selected sites and a separate file  
292 representing the 5mer bases 'checkMotif\_reformat.txt'. The path to  
293 outputs could be specified within the command.

```
294 $ bash README_processing.sh WT_vs_KO_RC22_call2_result.out
295 hg38.genome GRCh38_96.fa path_to_output.
```

296 3. Extract 5mer features: rebuild the tabular features such that the  
297 scores: mismatch, insertion and deletion are represented for each po-  
298 sition of the specific 5mer (NNANN). To do so, run the R script  
299 'HEK293\_data\_prep.R'. This produces an R object named 'BigTable.rds',  
300 representing the matrix of Sites $\times$ 15 features corresponding to the mis-  
301 match, insertion and deletion scores for the observed site and its two  
302 flanking positions. Be aware to precise the path to outputs that con-  
303 tains already the preprocessed data and provide the sample's name as  
304 a label of the analysis.

```
305 $ Rscript HEK293_data_prep.R path_to_output WT_vs_KO_RC22_call2_result.out
```

306 4. Extract m6A modification pattern: given the matrix of Sites $\times$ Features,  
307 the next step is to learn a model representing the m6A modification  
308 pattern. To this end, the conventional non-negative matrix factor-  
309 ization (NMF) analysis is suggested Lee and Seung [1999]. Briefly,  
310 NMF factorizes a non-negative data matrix  $X$  (here:  $n$  sites and  $m$   
311 features) into two non-negative matrices as  $X \approx WH$ , such that  $W$   
312 is an  $n \times k$  matrix containing basis vectors and  $H$  is an  $k \times m$  ma-  
313 trix containing coefficient vectors. The coefficient vectors and their  
314 combination can be viewed as a pattern for m6A modification. The  
315 R script 'HEK293\_data\_prep\_step2.R' allows generating patterns from  
316 a subset of the data related to previously reported m6A sites. Here,  
317 the unsupervised pattern training is based on 2401 common reported  
318 m6A sites in Koh et al. [2019] . Based on the Silhouette and Cophe-  
319 netic Correlation indices, we could identify an optimal factorization  
320 rank of 7. We then analyzed the identified patterns. According to  
321 the membership indicator of each site in matrix  $W$ , more than 80% of  
322 m6A modification sites can be represented by five patterns (Patterns  
323 2,3,4,6,7). The linear combination of these five patterns in fig (4A)  
324 highlights the importance of position 3 and eventually the implication  
325 of all scores.

references

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

327 The 'miCLIP\_union.bed' file contains all m6A sites reported in Koh  
328 et al. [2019] ().

reference

329 5. Predict m6A modification: The additive linear combination of the co-  
330 efficient vectors (patterns) with features can be used to predict m6A  
331 modification. We examine the ability of prediction on a subset of data  
332 of more than 1,98 million sites with 22248 miCLIP m6A validated sites  
333 . We opt for the linear combination of the five important patterns de-  
334 scribed in the previous section. The empirical Cumulative Distribution  
335 Function (eCDF) of the inferred scores shows a significant difference  
336 between the different miCLIP m6A categories and the unmodified sites  
337 (fig. 4B). As the number of negative samples is much larger than the  
338 number of positive samples, we particularly recommend investigating  
339 the Positive Predictive Value (PPV) of the predictions. Here, (fig. 4C)  
340 shows a moderate PPV that increases with the cut-off. The R script  
341 'HEK293\_data\_prep\_step3.R' allows generating the scores, eCDF prob-  
342 abilities of modification, and the corresponding eCDF and PPV plots.

reference  
+  
descrip-  
tion of  
cate-  
gories

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

## 344 Use Case 2: Comparison of wildtype and IVT samples

345 The second benchmarck is composed of wildtype HEK293 cell line from  
346 Pratanwanich et al. [2021] and the synthetised IVT samples described in  
347 section 2. The analysis steps are similar to the first case.

348 1. Identify read error profile: given the synthetised sequence, run JA-  
349 CUSA2 on paired conditions mode with the same parameters as the  
350 previously described case.

351 `$ JACUSA2 2.0.0-RC22 call-2 -m 1 -q 1 -c 4 -p 10 -D -I -a D,Y -P1 FR-SECONDSTRAND`  
352 `-P2 FR-SECONDSTRAND -r WT_vs_realIVT_v202_call12_result.out HEK293T-WT-rep2.bam,`  
353 `HEK293T-IVT-rep1.bam,HEK293T-IVT-rep2.bam`

354 2. Preprocess JACUSA2 output: select the 5mer specific sites (NNANN)  
355 considering the Y filter as follows.

356 `$ bash README_processing.sh WT_vs_IVT_RC22_call12_result.out hg38.genome GRCh38.p12`

357 3. Extract 5mer features using the following command:

358 `$ Rscript Code/HEK293_data_prep.R path_to_output WT_vs_IVT_RC22_call12_result.out`

359 4. Extract m6A modification patterns based on 2401 m6A sites ([reference](#)). From  
 360 the Silhouette and Cophenetic Correlation indices, we could identify  
 361 an optimal factorization rank of 6 (fig. 5).

```
362 $ Rscript HEK293_data_prep_step2.R path_to_output miCLIP_union_flat_exclude_Y_c
```

363 5. Predict m6A modifications: we examine the prediction of modification  
 364 using the detected patterns on the test set of 22248 m6A sites. So here,  
 365 we plot the eCDF of pattern 6 scores by category (fig. 5)

```
366 $ Rscript HEK293_data_prep_step3.R path_to_output miCLIP_union_flat_exclude_Y_c
```

## 367 NOTES

### 368 Tips and Tricks

## 369 ACKNOWLEDGMENTS

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

## 375 REFERENCES

- 376 Minimap2. <https://github.com/lh3/minimap2>. Accessed: 2022-01-19.
- 377 Samir Adhikari, Wen Xiao, Yong-Liang Zhao, and Yun-Gui Yang. m(6)a:  
 378 Signaling for mrna splicing. *RNA biology*, 13:756–759, September 2016.  
 379 ISSN 1555-8584. doi: 10.1080/15476286.2016.1201628.
- 380 Ina Anreiter, Quoseena Mir, Jared T. Simpson, Sarath C. Janga, and  
 381 Matthias Soller. New twists in detecting mrna modification dynamics.  
 382 *Trends in biotechnology*, 39:72–89, January 2021. ISSN 1879-3096. doi:  
 383 10.1016/j.tibtech.2020.06.002.
- 384 Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali  
 385 Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine  
 386 Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and  
 387 Gideon Rechavi. Topology of the human and mouse m6a rna methylomes  
 388 revealed by m6a-seq. *Nature*, 485:201–206, April 2012. ISSN 1476-4687.  
 389 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. N6-methyladenosine in nuclear rna is a major substrate of the obesity-associated fto. *Nature chemical biology*, 7:885–887, October 2011. ISSN 1552-4469. doi: 10.1038/nchembio.687.

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

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

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

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

Michael Piechotta, a.

Michael Piechotta, b.

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

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

426 Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic rna  
427 modifications in gene expression regulation. *Cell*, 169:1187–1200, June  
428 2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.045.

429 Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how:  
430 Context-dependent functions of rna methylation writers, readers, and  
431 erasers. *Molecular cell*, 74:640–650, May 2019. ISSN 1097-4164. doi:  
432 10.1016/j.molcel.2019.04.025.

433 Sara Zaccara, Ryan J. Ries, and Samie R. Jaffrey. Reading, writing and  
434 erasing mrna methylation. *Nature reviews. Molecular cell biology*, 20:608–  
435 624, October 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0168-5.

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

440 Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min  
441 Huang, Charles J. Li, Cathrine B. Vågbø, Yue Shi, Wen-Ling Wang, Shu-  
442 Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin  
443 Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan,  
444 Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne  
445 Klungland, Yun-Gui Yang, and Chuan He. Alkbh5 is a mammalian rna  
446 demethylase that impacts rna metabolism and mouse fertility. *Molecular*  
447 *cell*, 49:18–29, January 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.  
448 10.015.

## FIGURE CAPTIONS

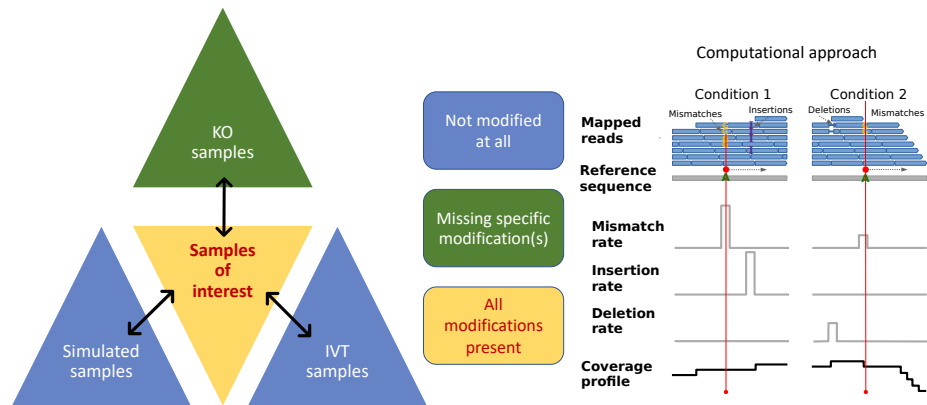


Figure 1: **General outline of RNA modification detection by JACUSA2.** A key feature of our approach is that multiple replicates can be compared as shown on the left. Samples of interests where all modifications are present could be compared with either KO samples where the modification of interest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

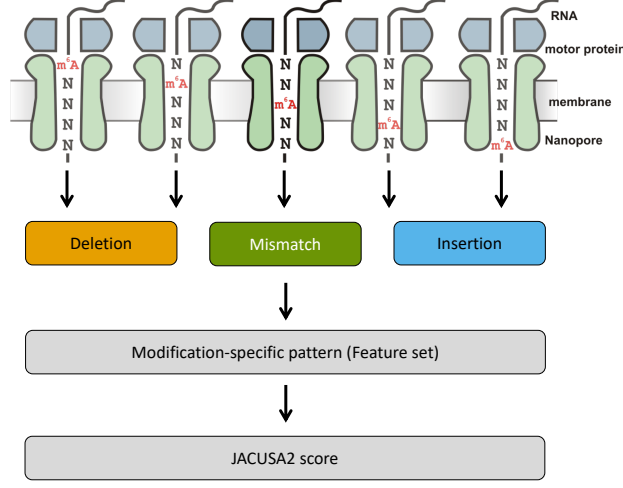


Figure 2: **Motivation of 5mer context for RNA modification mapping.** The nanopore covers 5 consecutive RNA residues. That is why we consider a 5mer context and derive 3 principal features for every position within a given 5 mer (15 features in total, with a central A residue in this example). We evaluate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

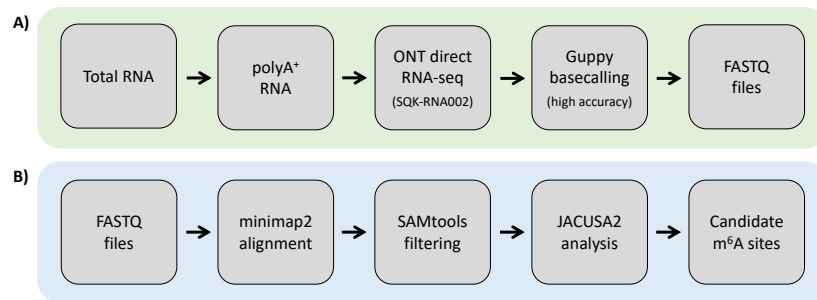


Figure 3: **Experimental and computational workflow.** tbd



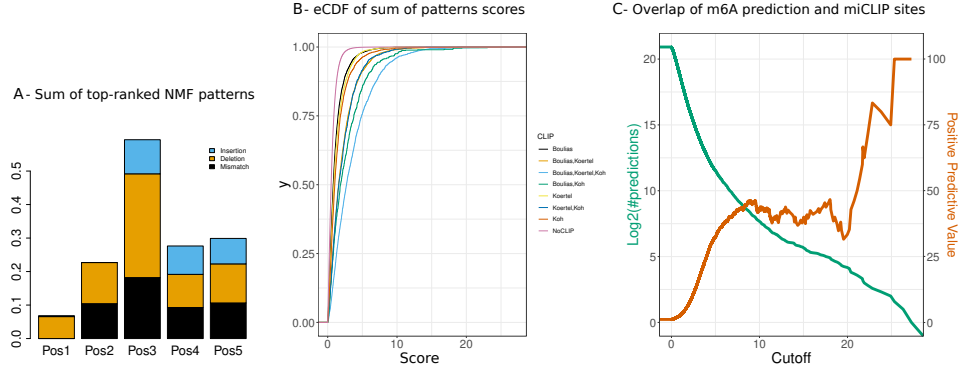


Figure 4: **Case 01. WT versus KO.** **A:** Barplots representing the linear combination of the top 5 patterns (y-axis) by position in the specific 5mer context (x-axis) and the score type: mismatch, deletion and insertion (resp. black, orange and blue). The 5 patterns (coefficient vectors: 2,3,4,6,7) are selected according to the predominant columns in matrix W. **B:** Score distribution inferred from the combined patterns, stratified by the different categories of miCLIP validated sites and non miCLIP sites. **C:** Number of predicted m6A sites (green) and Positive Predictive Value (PPV) of predicted m6A sites that overlap with miCLIP sites (orange).

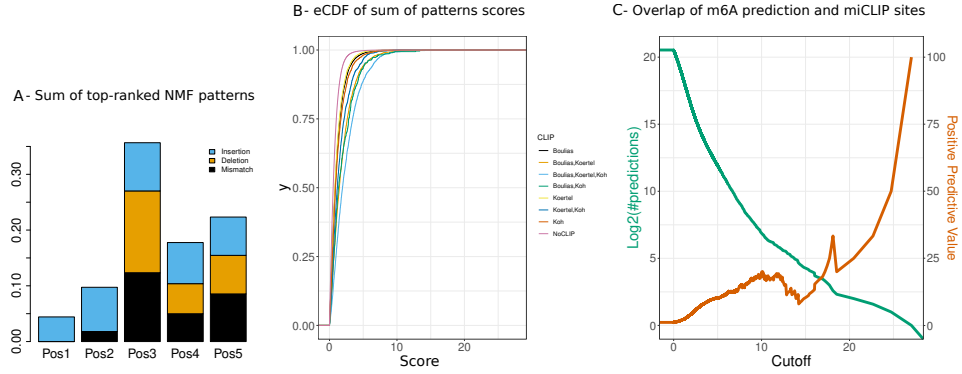


Figure 5: **Case 02. WT versus IVT.** **A:** Barplots representing the linear combination of the top 4 patterns (y-axis) by position in the specific 5mer context (x-axis) and the score type: mismatch, deletion and insertion (resp. black, orange and blue). The 4 patterns (coefficient vectors: 1,2,3,6) are selected according to the predominant columns in matrix  $W$ . **B:** Score distribution inferred from the combined patterns, stratified by the different categories of miCLIP validated sites and non miCLIP sites. **C:** Number of predicted m6A sites (green) and Positive Predictive Value (PPV) of predicted m6A sites that overlap with miCLIP sites (orange).

| Software       | Version                                                                                                              | Description                                                                                              |
|----------------|----------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------|
| Minimap2       | <a href="https://github.com/lh3/minimap2">https://github.com/lh3/minimap2</a><br>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><br>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><br>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><br>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

## 450 TABLE CAPTIONS

## 451 TABLES

| R Pack-<br>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