1

2

Mapping of RNA modiﬁcations by direct Nanopore

sequencing and JACUSA2

3

Amina Lemsara1,2, Christoph Dieterich∗1,2,3, and Isabel Naarmann-de Vries1,2,3

1

Klaus Tschira Institute for Integrative Computational Cardiology, University

Heidelberg, 69120 Heidelberg, Germany

4

5

6

7

8

9

2

Department of Internal Medicine III (Cardiology, Angiology, and Pneumology),

University Hospital Heidelberg, 69120 Heidelberg, Germany

3

German Centre for Cardiovascular Research (DZHK)-Partner Site

Heidelberg/Mannheim, 69120 Heidelberg, Germany

1

0

Abstract

1

1

1

1

1

1

1

1

1

2

2

2

2

2

2

2

2

2

2

3

3

3

1

2

3

4

5

6

7

8

9

0

1

2

3

4

5

6

7

8

9

0

1

2

RNA modiﬁcations exist in all kingdom of life. Several diﬀerent

types of base or ribose modiﬁcations are now summarized under the

term the ”epitranscriptome”. With the advent of high-throughput se-

quencing technologies much progress has been made in understanding

RNA modiﬁcation biology and how these modiﬁcations can inﬂuence

many aspects of RNA life. The most widespread internal modiﬁcation

on mRNA is m6A, which has been implicated in physiological pro-

cesses as well as disease pathogenesis. Here, we provide a workﬂow for

the mapping of m6A sites using Nanopore direct RNA sequencing data.

Our strategy employs pairwise comparison of base calling error proﬁles

with JACUSA2. We outline a general strategy for RNA modiﬁcation

detection on mRNA and describe two speciﬁc use cases on m6A de-

tection in detail. Use case 1: a sample of interest with modiﬁcations

(e.g. ”wild type” sample) is compared to a sample lacking a speciﬁc

modiﬁcation type (e.g. ”knock out” sample, here METTL3-KO) or

Use case 2: a sample of interest with modiﬁcations is compared to

a sample lacking all modiﬁcations (e.g. in vitro transcribed cDNA).

We provide a detailed protocol on experimental and computational

aspects. Extensive online material provides a snakemake pipeline to

identify m6A positions in mRNA and to validate the results against

a miCLIP-derived m6A reference set. The general strategy is ﬂexible

and can be easily adapted by users in diﬀerent application scenarios.

∗Correspondence to: christoph.dieterich@uni-heidelberg.de

1



3

3

INTRODUCTION

3

3

3

3

3

3

4

4

4

4

4

4

4

4

4

4

5

5

5

5

5

5

5

5

5

5

6

6

6

6

6

6

6

6

6

6

7

7

7

7

7

7

4

5

6

7

8

9

0

1

2

3

4

5

6

7

8

9

0

1

2

3

4

5

6

7

8

9

0

1

2

3

4

5

6

7

8

9

0

1

2

3

4

5

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. Impres-

sive 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 diﬀer-

entiation, 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].](#br15) These modiﬁcations are catalyzed by a variety of

dedicated enzymes and can be divided into two classes: modiﬁcations of

cap-adjacent nucleotides and internal modiﬁcations.

In contrast to the m7G cap, the impact of internal modiﬁcations on gene

regulation has been less studied apart from RNA editing, which is mediated

by RNA deaminases (e.g. the ADAR family). The most widespread in-

ternal mRNA modiﬁcation is N6-methyladenosine (m6A). By modulating

the processing of mRNA, m6A can regulate a wide range of physiological

processes and its alteration has been linked to several diseases [Roignant](#br17)

[and Soller [2017],](#br17) [Zaccara et al. [2019],](#br18) [Shi et al. [2019].](#br17) The modiﬁcation is

catalyzed co-transcriptionally by a Mega-Dalton methyltransferase complex,

which includes the heterodimer METTL3-METTL14 and other associated

subunits [Garcias Morales and Reyes [2021].](#br16) This modiﬁcation is reversible

since two proteins of the AlkB-family of demethylases can remove m6A from

mRNA transcripts [[Jia et al., 2011,](#br16) [Zheng et al., 2013].](#br18) In mammals, m6A

preferentially localizes within long internal exons and at the beginning of

terminal exons at so-called DRACH motif (D = A/G/U, R = A/G, H =

A/C/U) sites [[Dominissini et al., 2012,](#br16) [Meyer et al., 2012,](#br17) [Ke et al., 2015].](#br16)

Once deposited, m6A is recognized by several reader proteins that can af-

fect the fate of mRNA transcripts in nearly every step of the mRNA life

cycle, including alternative splicing [[Adhikari et al., 2016,](#br15) [Roundtree et al.,](#br17)

[2017],](#br17) mRNA translation [[Wang et al., 2015]](#br18) and decay [[Wang et al., 2014,](#br18)

[Du et al., 2016,](#br16) [Roundtree et al., 2017].](#br17) The best-described readers are the

YTH domain family of proteins that decode the signal and mediate m6A

functions. By aﬀecting RNA structure, m6A can also indirectly inﬂuence

the association of additional RNA-binding proteins (RBPs) and the assem-

bly of larger messenger ribonucleoprotein particles (mRNPs) [[Patil et al.,](#br17)

[2018].](#br17)

2

7

7

7

7

8

8

8

8

8

8

8

8

6

7

8

9

0

1

2

3

4

5

6

7

Several approaches have been presented to map RNA modiﬁcations on

RNA. Herein, we focus on mRNA modiﬁcation site detection in general and

on m6A in particular where antibody-based protocols (miCLIP), methylation-

sensitive restriction enzyme assays (MazF) or transgenic approaches (TRIBE,

DART) have been presented to map m6A sites. All of the aforementioned

approaches rely on high-throughput short read sequencing on the Illumina

platform. This typically involves cDNA synthesis by reverse transcription

and PCR-based library ampliﬁcation. One recent addition to the toolbox of

RNA modiﬁcation mapping is direct RNA single molecule long read sequenc-

ing on the Oxford Nanopore Technologies platform (dRNA-seq). While our

software is able to deal with Illumina and Nanopore-based approaches, the

latter is the principal topic of this methods article.

8

8

8

9

MATERIALS

ONT direct RNA sequencing

9

9

0

1

This section summarizes all necessary consumables for direct RNA sequenc-

ing of poly-adenylated RNA (i.e. mRNA) on the MinION or similar device.

9

9

9

9

9

2

3

4

5

6

1. 500 ng polyA+ RNA isolated from total RNA e.g. with Oligotex

mRNA kit (#70022, Qiagen) or Dynabeads oligo dT25 beads (#61002,

Thermo Fisher Scientiﬁc) or in vitro transcriptome sample. Store RNA

at -80 ◦C and the mRNA puriﬁcation kit as recommended by the man-

ufacturer.

9

7

2. Nuclease-free water. Store at room temperature.

9

9

8

9

3. Direct RNA-sequencing kit (SQK-RNA002, Oxford Nanopore Tech-

nologies). Store at -20 ◦C.

1

1

00

01

4. NEBNext Quick Ligation Reaction Buﬀer (#B6058S, New England

Biolabs). Store at -20 ◦C.

1

02

5. T4 DNA Ligase (#M0202S, New England Biolabs). Store at -20 ◦C.

1

1

03

04

6. dNTP Mix (10 mM each, #R0191, Thermo Fisher Scientiﬁc). Store

at -20 ◦C.

1

1

05

06

7. SuperScript IV Reverse Transcriptase (#18090010, Thermo Fisher Sci-

entiﬁc). Store at -20 ◦C.

1

1

07

08

8. Agencourt RNAClean XP beads (#A63987, Beckman Coulter). Store

at 4 ◦C.

1

09

9. 70 % ethanol, freshly prepared.

3

1

1

10

11

10. Qubit dsDNA HS assay kit (#Q32854) and Qubit Fluorometer (Thermo

Fisher Scientiﬁc).

1

1

12

13

11. Flow cell priming kit (EXP-FLP002, Oxford Nanopore Technologies).

Store at -20 ◦C.

1

1

1

1

14

15

16

17

12. Thermocycler.

13. Gentle rotator mixer.

14. Magnetic stand for 1.5 ml tubes.

15. 1.5 ml DNA LoBind tubes (Eppendorf), 0.2 ml PCR tubes.

1

1

1

18

19

20

16. MinION or GridION sequencing device and MinION R9.4.1 Flow cells

(FLO-MIN106D, Oxford Nanopore Technologies). Store Flow cells at

4 ◦C.

1

21 Preparation of an in vitro transcriptome sample

1

1

1

1

22

23

24

25

1. 100 ng polyA+ RNA isolated from total RNA e.g. with Oligotex

mRNA kit (#70022, Qiagen) or Dynabeads oligo dT25 beads (#61002,

Thermo Fisher Scientiﬁc). Store RNA at -80 ◦C and the mRNA pu-

riﬁcation kit as recommended by the manufacturer

1

1

26

27

2. 10 µM oligo(dT)-VN RT primer.

TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTVN. Store at -20 ◦C.

1

1

28

29

3. 20 µM template switching oligo (TSO). ACTCTAATACGACTCAC-

TATAGGGAGAGGGCrGrG+G. Store at -20 ◦C.

1

1

30

31

4. 10 µM T7 extension primer. GCTCTAATACGACTCACTATAGG.

Store at -20 ◦C.

1

32

5. Nuclease-free water. Store at room temperature.

1

1

33

34

6. dNTP Mix (10 mM each, #R0191, Thermo Fisher Scientiﬁc). Store

at -20 ◦C.

1

1

35

36

7. Template Switching RT Enzyme Mix (#M0466S, New England Bio-

labs). Store at -20 ◦C.

1

1

37

38

8. Q5 Hot Start High-Fidelity 2X Master Mix (#M0494S, New England

Biolabs). Store at -20 ◦C.

1

1

39

40

9. RNase H (5,000 U/ml) (#M0297S, New England Biolabs). Store at

-20 ◦C.

4

1

1

1

41

42

43

10. NucleoSpin Gel and PCR Clean-up, Mini kit for gel extraction and

PCR clean up (#740609.50, Macherey-Nagel) or equivalent. Store at

room temperature.

1

1

44

45

11. MEGAscript T7 transcription kit (#AM1334, Thermo Fisher Scien-

tiﬁc). Store at -20 ◦C.

1

1

46

47

12. RNA Clean & Concentrator-25 kit (#R1017, Zymo Research). Store

at room temperature.

1

1

1

1

48

49

50

51

13. Thermocycler.

14. Table top centrifuge for 1.5 ml tubes.

15. Nanodrop spectrophotometer or equivalent.

16. 0.2 ml PCR tubes, 1.5 ml DNA LoBind tubes (Eppendorf).

1

52 Hardware requirements

1

1

1

1

1

53 All analyses have been performed/tested on two alternative hardware sys-

54 tems: a standard Linux desktop computer or an Apple iMac (Retina 5K,

55 ultimo 2014). The workﬂow requires a multi-core processor system with

56 minimal main memory of 16GB RAM and several GBs of free disk space

57 (depending on data set size).

1

58 Software dependencies and installation

1

1

1

1

1

1

1

1

1

59 Our analysis workﬂow has few requirements, which are detailed in Table 1.

60 Speciﬁcally, to execute our workﬂow, the following prerequisites are neces-

61 sary: a BASH shell, a J AVA runtime environment, a working PERL and

62 R installation. Additional i.e. non-standard software to process and map

63 Nanopore reads (bedtools, samtools and Minimap2) are obligatory. Ta-

64 ble 2 lists some additional R packages, which are required to run the R

65 code. Detailed installation instructions and corresponding workﬂow code

66 are deposited under [https://github.com/dieterich-lab/MiMB\_JACUSA2\_](https://github.com/dieterich-lab/MiMB_JACUSA2_chapter)

67 [chapter](https://github.com/dieterich-lab/MiMB_JACUSA2_chapter).

1

68 METHODS

1

1

1

1

1

1

69 Our workﬂow is based on the pairwise comparison of samples with diﬀer-

70 ent modiﬁcation status (Figure [1).](#br19) The sample of interest (yellow) may be

71 compared to diﬀerent samples lacking certain modiﬁcations. If available,

72 the wild type (WT) sample can be compared to a knock out (KO) sample

73 lacking speciﬁc enzymatic activities (green), as outlined in Use Case 1. Al-

74 ternatively, a sample lacking all modiﬁcations may be used for comparison

5

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

1

75 (blue). This may be either a simulated sample (i.e. with NanoSim) or an in

76 vitro transcribed sample derived from cDNA. Such an analysis is detailed in

77 Use Case 2. In any setting, JACUSA2 calculates scores for the Mismatch,

78 Insertion and Deletion rates of the pairwise comparisons as outlined above

79 (Figure [1,](#br19) right).

80

One feature of Nanopore sequencing is to read sequences as 5-mers, as

81 always ﬁve nucleotides are occupied by the pore protein (Figure [2).](#br20) Because

82 of this, a m6A modiﬁcation may aﬀect basecalling not only if the modiﬁed

83 nucleotide is in the central position, but also at neighboring positions (-2

84 to +2). To account for this, JACUSA2 scores for Deletion, Mismatch and

85 Insertion are calculated for the entire 5-mer context. Depending on the

86 modiﬁcation-speciﬁc signature, a Feature set can be selected to calculate

87 the ﬁnal JACUSA2 score (Figure [2).](#br20)

88

Our workﬂow can be divided into a wet-lab part (Figure [3A)](#br21) and a

89 computational part (Figure [3B).](#br21) Starting from total cellular RNA, polyA+

90 RNA is isolated and subjected to Nanopore direct RNA-sequencing. Guppy

91 basecalling can be done as well as live basecalling during sequencing on the

92 respective FAST5 ﬁles, which results in FASTQ output ﬁles (Figure [3A).](#br21)

93 FASTQ ﬁles are aligned to a reference sequence with Minimap2. SAMtools

94 is used to generate BAM ﬁles as input for JACUSA2 analysis, which yields

95 candidate m6A sites with the presented workﬂow in this chapter (Figure

96 [3B).](#br21) We will present all necessary experimental step for dRNA-seq in the

97 next section.

1

98 Nanopore direct RNA sequencing

1

2

2

2

2

2

99

00

01

02

03

04

1. Adjust 500 ng polyA+ RNA to a total volume of 9 µl with nuclease-

free water. Complete RT adapter ligation reaction (in 0.2 ml PCR

tube) with 3 µl NEBNext Quick Ligation Reaction Buﬀer, 0.5 µl

RNA CS (RCS, from SQK-RNA002), 1 µl RT-Adapter (RTA, from

SQK-RNA002) and 1.5 µl T4 DNA Ligase. Incubate 10 min at room

temperature.

2

2

2

05

06

07

2. Prepare reverse transcription master mix on ice during ligation: 9 µl

nuclease-free water, 2 µl 10 mM dNTPs, 8 µl 5x SuperScript IV ﬁrst

strand buﬀer, 4 µl 0.1 mM DTT.

2

2

2

2

08

09

10

11

3. Add the reverse transcription master mix to the ligation reaction and

mix by pipetting. Add 2 µl SuperScript IV reverse transcriptase and

mix by pipetting. Incubate in a thermocycler with the following pro-

◦

tocol: 50 min at 50 C, 10 min at 70 C, cool down to 4 C.

◦

◦

2

2

2

12

13

14

4. Let the Agencourt RNAClean XP beads come to room temperature

during reverse transcription. Carefully resuspend beads before use.

Transfer reaction to a 1.5 ml DNA LoBind tube and mix with 72 µl

6

2

2

15

16

Agencourt RNAClean XP beads. Incubate 5 min at room temperature

on a gentle rotator mixer.

2

2

2

2

17

18

19

20

5. Collect beads on a magnetic stand and remove supernatant. Wash

pelleted beads two times (30 sec) with 200 µl freshly prepared 70 %

ethanol. Remove supernatant. Spin sample down and place on magnet

again. Remove any residual ethanol.

2

2

2

2

21

22

23

24

6. Resuspend beads in 20 µl nuclease-free water by gentle ﬂicking and

incubate 5 min at room temperature on a gentle rotator mixer. Collect

beads on a magnetic stand and transfer 20 µl eluate in a fresh 1.5 ml

DNA LoBind tube.

2

2

2

2

25

26

27

28

7. For ligation of the RMX adapter, add the following to 20 µl eluate: 8

µl NEBNext Quick Ligation Reaction Buﬀer, 6 µl RMX (from SQK-

RNA002), 3 µl nuclease-free water, 3 µl T4 DNA Ligase. Mix by

pipetting and incubate 10 min at room temperature.

2

2

2

29

30

31

8. Add 40 µl carefully resuspended Agencourt RNAClean XP beads to

the reaction and mix by pipetting. Incubate 5 min at room tempera-

ture on a gentle rotator mixer.

2

2

2

2

32

33

34

35

9. Collect beads on a magnetic stand and remove supernatant. Wash

pelleted beads two times with 150 µl wash buﬀer (WSB, from SQK-

RNA002). Resuspend beads by ﬂicking, spin down and return to mag-

netic stand. Remove supernatant from pelleted beads.

2

2

2

2

36

37

38

39

10. Resuspend beads in 21 µl elution buﬀer (EB, from SQK-RNA002) by

gentle ﬂicking and incubate 5 min at room temperature on a gentle

rotator mixer. Pellet beads on a magnetic stand and transfer 21 µl

eluate in a fresh 1.5 ml DNA LoBind tube.

2

2

2

40

41

42

11. Quantify 1 µl of the library on a Qubit ﬂuorometer with the Qubit

dsDNA HS kit according to the manufacturerers protocol. Concentra-

tion should be usually in the range of 5 - 10 ng/µl.

2

2

2

2

43

44

45

46

12. Insert MinION R9.4.1 Flow cell in the MinION or GridION sequenc-

ing device and perform Flow cell check in the MinKNOW software.

For successful sequencing of mammalian polyA+ RNA at least 1,000

available pores are recommended.

2

2

2

2

2

2

2

47

48

49

50

51

52

53

13. Prepare Priming Mix by adding 30 µl ﬂush tether (FLT, from EXP-

FLP002) to a vial of ﬂush buﬀer (FB, from EXP-FLP002) and mix by

pipetting. Open priming port. Remove air bubble from priming port

by inserting the tip of a P1000 pipette into the priming port and slowly

dialing up, until a small volume of storage buﬀer enters the pipette

tip. Load 800 µl Priming Mix via the priming port and carefully avoid

introduction of air bubbles. Close the priming port and wait for 5 min.

7

2

2

2

2

2

2

54

55

56

57

58

59

14. Mix 20 µl library with 17.5 µl nuclease-free water and 37.5 µl RNA run-

ning buﬀer (RRB, from SQK-RNA002) and mix by pipetting. Open

the priming port and the sample port. Load 200 µl Priming Mix via

the priming port. Mix library by pipetting just before loading and

load dropwise via the sample port. Carefully avoid introduction of air

bubbles. Close the sample port and the priming port.

2

2

2

60

61

62

15. Start sequencing for 48 to 72 h in the MinKNOW software. Choose

direct RNA-sequencing kit and high-accuracy basecalling as parame-

ters.

2

63 Preparation of an in vitro transcriptome sample

2

2

2

2

64 The in vitro transcriptome sample is prepared based on a protocol published

65 by [Zhang et al. [2021]](#br18) with some modiﬁcations a detailed below. An in vitro

66 transcriptome lacks any RNA modiﬁcations and is a perfect reference sample

67 for RNA modiﬁcation mining.

2

2

2

2

68

69

70

71

1. Adjust 100 ng polyA+ RNA to a total volume of 6 µl with nuclease-

free water. Add 1 µl each of 10 µM oligo(dT)-VN RT primer and 10

mM dNTPs. Mix by pipetting and incubate in a thermocycler: 5 min

◦

at 75 C, 2 min at 42 C, cool to 4 C.

◦

◦

2

2

2

2

72

73

74

75

2. Assemble 2.5 µl 4x template switching RT buﬀer, 0.5 µl 20 µM TSO,

1 µl 10x template switching RT enzyme mix and mix by pipetting.

Combine with 6 µl RNA and incubate in a thermocycler: 90 min at

◦

42 C, 10 min at 68 C, cool to 4 C.

◦

◦

2

2

2

2

2

76

77

78

79

80

3. For Second strand synthesis add to First strand synthesis reaction: 50

µl Q5 Hot Start High-Fidelity 2X Master Mix, 5 µl RNase H, 2 µl 10

µM T7 extension primer, 33 µl nuclease-free water. Mix by pipetting

and incubate in a thermocycler: 15 min at 37 ◦C, 1 min at 95 ◦C, 10

◦

min at 65 C, cool to 4 C.

◦

2

2

2

2

81

82

83

84

4. Purify double stranded cDNA with NucleoSpin Gel and PCR Clean-up

kit according to the manufacturerers protocol and elute in 20 µl elution

buﬀer. Determine concentration on a Nanodrop spectrophotometer.

cDNA may be stored at -20 ◦C.

2

2

2

85

86

87

5. Combine 8 µl cDNA for in vitro transcription with 2 µl each of ATP,

GTP, CTP, UTP, 10x reaction buﬀer and enzyme mix from the MEGAscript

T7 transcription kit. Incubate 3 h at 37 ◦C.

2

2

88

89

6. Digest template DNA by addition of 1 µl Turbo DNase. Mix by pipet-

ting and incubate 15 min at 37 ◦C.

8

2

2

2

2

2

2

90

91

92

93

94

95

7. Adjust reaction volume to 100 µl with nuclease-free water and clean up

with RNA Clean & Concentrator-25 kit according to the manufactur-

ers protocol, using two volumes of adjusted RNA binding buﬀer (1:1

RNA binding buﬀer : ethanol). Elute RNA in 25 µl nuclease-free wa-

ter. Determine RNA concentration on a Nanodrop spectrophotometer.

Store at -80 ◦C.

2

96 Nanopore read processing

2

2

2

3

3

3

3

97

98

99

00

01

02

03

1. Base call the ionic current signal stored in FAST5 ﬁles using Guppy.

For the IVT sample, we applied real-time base calling with the MinKNOW-

embedded Guppy basecaller. Otherwise, Guppy basecaller software

can be used. In this case, the basecaller requires the path to FAST5

ﬁles, the output folder, and the conﬁg ﬁle or the ﬂowcell/kit combina-

tion. The output are FASTQ ﬁles that can be compressed using the

option ”–compress fastq”.

3

3

3

04

05

06

$ guppy\_basecaller --compress\_fastq -i path\_to\_fast5 -s path\_to\_output

-c config\_file.cfg --cpu\_threads\_per\_caller 14 --num\_callers

1

3

3

3

07

08

09

Set the number of threads ”cpu threads per caller” and the number

of parallel basecallers ”num caller” according to your resources. Ad-

ditional details can be found at <https://nanoporetech.com/>.

3

3

3

3

3

3

3

3

10

11

12

13

14

15

16

17

2. Align reads to the transcriptome using Minimap2 software. The out-

put is a SAM ﬁle that has to be converted to a compressed form as

BAM ﬁle using SAMtools command. The alignment requires a refer-

ence sequence. Here, we used GRCh38 Ensembl release 96 annotation

and FAS TA ﬁle. Pre-indexing of the human genome saves time dur-

ing read alignment. Please save the index with the option ”-d” before

read mapping and use the index instead of the reference ﬁle in the

minimap2 command line.

3

18

$ minimap2 -d reference.mmi reference.fa

3

3

3

3

3

3

3

3

3

3

19

20

21

22

23

24

25

26

27

28

For Direct RNA Sequencing, it is recommended to set a small k-mer

size ”-k [=14]” to enhance sensitivity. We recommend outputting only

primary alignments ”–secondary=no”. Use the parameter ’–MD’ to

add the reference sequence information to the alignment; this is neces-

sary for JACUSA2 downstream analysis. Adjust the number of threads

”-t” according to your resources. Check Minimap2 manual for more

details [[Min].](#br15) To enable spliced alignments, use the setting -ax splice

–junc-bed annotation.bed –junc-bonus where ”–junc-bonus” allows to

tune the bonus score and the BED ﬁle ”–junc-bed annotation.bed”

provides the splice junctions.

9



3

3

3

29

30

31

$ minimap2 -t 5 --MD -ax splice --junc-bonus 1 -k14 --secondary=no

--junc-bed final\_annotation\_96.bed -ub reference.mmi Reads.fastq.gz

|samtools view -bS > mapping.bam

3

3

32

33

The BED ﬁle can be generated from EnsEMBL GTF ﬁles using the

following command:

3

34

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

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56

57

58

59

60

3. Mapping RNA modiﬁcations using JACUSA2 pipeline: JACUSA2

[[Piechotta et al., 2021]](#br17) rapidly detects RNA modiﬁcations based on

a comparative strategy where read alignment features (mismatch, in-

sertion and deletion) of samples of interest are compared to a reference

sequence (call-1) or against reference samples without the correspond-

ing RNA modiﬁcation of interest (call-2). JACUSA2 processes repli-

cate experiments. The analysis of read alignment signatures is used

for RNA modiﬁcation detection. Particularly, we integrate JACUSA2

call-2 method with the downstream analysis in one workﬂow using

the Snakemake workﬂow manager [[K¨oster and Rahmann, 2012].](#br17) Our

Snakemake workﬂow encompasses several steps as shown in Figure

[4.](#br22) The workﬂow requires BAM ﬁles from 2 conditions as input. We

suggest to ﬁlter secondary and poor alignments beforehand. The out-

put of JACUSA2 call2 is preprocessed (get features) and subjected

to a machine learning step to extract and visualize modiﬁcation pat-

terns (resp. get pattern, visualize pattern) and make predictions (pre-

dict modiﬁcation). ”split trani test” rule allows splitting input data

into a training set and a test set. To use our snakemake-based JA-

CUSA2 pipeline a set of parameters should be deﬁned in the ”con-

ﬁg.yaml” ﬁle; mainly: the label of the analysis ’label’, the input bam

ﬁles under ’data’, the reference sequence ’reference’, a ﬁle containing

size of chromosomes ’chr size’, JACUSA2 jar ﬁle ’jar’, plus the path to

inputs and outputs under ’path inp’ and ’path out’ ﬁelds respectively.

We typically execute the workﬂow on a multi-core CPU system using

the following command by specifying the number of cores to be used

”–cores [=all]” and the rule name:

3

61

$ snakemake --cores all rule\_name

3

3

62

63

Please consult the Snakemake documentation for further details (see

<https://snakemake.readthedocs.io/en/stable/>).

3

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

3

3

65 The JACUSA2 workﬂow detects RNA modiﬁcations using direct RNA se-

66 quencing by comparing modiﬁed samples to unmodiﬁed control samples.

1

0



3

3

3

3

3

3

3

3

3

3

3

3

3

67 Here, we used a published dataset of HEK293 cell lines to map m6A mod-

68 iﬁcation [[Pratanwanich et al., 2021].](#br17) Our examples encompasses two con-

69 ditions: wild-type RNA (WT, modiﬁed RNAs) and RNA from METTL3

70 knockout cells (KO, m6A modiﬁcation is absentr). We use two replicates per

71 condition (see <https://doi.org/10.5281/zenodo.5913452>). The FASTQ

72 ﬁles are mapped using Minimap2 as described in the previous section. The

73 following analysis is validated against m6A sites consistently reported in

74 three miCLIP-based studies [Boulias et al. [2019],](#br15) [Koh et al. [2019],](#br16) [K¨ortel](#br17)

75 [et al. [2021]](#br17) (Figure [5).](#br23)

76

Starting with the preprocessed mapped reads as inputs (BAM ﬁles),

77 ’HEK293T-WT-rep2.bam’ and ’HEK293T-WT-rep3.bam’ represent the wild-

78 type replicates and ’HEK293T-KO-rep2.bam’ and ’HEK293T-KO-rep3.bam’

79 the control replicates,

3

3

80

81

1. Compute read error proﬁle with the jacusa2 call2 rule:

$ snakemake --cores all jacusa2\_call2

$

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

3

82

83

84

85

86

87

88

89

90

91

92

93

94

95

96

97

The method requires BAM ﬁles of the paired conditions and the cor-

responding library information ”-P1” and ”-P2”. In addition to the

mismatch score, add ”-D” and ”-I” to output the deletion and insertion

scores. JACUSA2 allows ﬁltering reads according to many parameters.

Here, we consider all sites with base calling quality ”-q [> 1]”, map-

ping quality ”-m [> 1] and read coverage ”-c [> 4]”. Here, we consider

ﬁltering sites within homopolymer regions ”-a [=Y]”. The output

(named here, ”Cond1vsCond2Call2.out”) consists of a read error pro-

ﬁle where the format is a combination of BED6 with JACUSA2 call-2

speciﬁc columns and common info columns: info, ﬁlter, and ref. Check

JACUSA2 manual for more details on JACUSA2 ﬁlter and output op-

tions [[JAC, 2021].](#br15) The number of threads can be customized via the

parameter ”-p”. All parameters related to the JACUSA2 method can

be added under the ﬁeld ”jacusa params” in the conﬁg ﬁle by setting

the name of the parameter followed by the corresponding value [key:

value]. Be aware to set all parameters before running the pipeline.

3

3

98

99

2. Process JACUSA2 output with the get features rule:

$ snakemake --cores all get\_features

4

4

4

4

4

4

4

00

01

02

03

04

05

06

we select all sites within 5-mer of a central nucleotide ’A’ ﬂanked by

2 random nucleotides (NNANN) and we ﬁlter out sites of the homo-

polymer regions. Then, we rebuild the tabular features such that the

observations are only sites with a reference base ’A’. Each site is char-

acterized by 15 features corresponding to the mismatch, insertion and

deletion scores for the observed site and its two ﬂanking positions

from both sides. The rule ”get features” performs the preprocessing

1

1



4

4

4

07

08

09

step. Use the parameter ’region’ with a ﬁle containing target 5-mers

to limit the analysis to speciﬁc sites. The output is an R object ”fea-

tures/features.rds”, representing the matrix of Sites×15 features.

4

4

10

11

3. Extract characteristic m6A modiﬁcation patterns with the get pattern

rule:

4

12

$ srun snakemake --cores all get\_pattern

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

4

13

14

15

16

17

18

19

20

21

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

We learn a model representing the m6A modiﬁcation patterns given

the matrix of Sites×Features. To this end, we employ non-negative

matrix factorization [(NMF)[Lee and Seung, 1999].](#br17) Brieﬂy, NMF fac-

torizes a non-negative data matrix X (here: n sites and m features)

into two non-negative matrices as X ≈ WH, such that W is an n × k

matrix containing basis vectors and H is an k × m matrix containing

coeﬃcient vectors. The coeﬃcient vectors and their combination can

be viewed as a pattern for m6A modiﬁcation. The rank of factorization

k is a critical parameter that aﬀects the performance substantially. We

suggest to select the rank k according to the method of [Frigyesi and](#br16)

[H¨oglund [2008]](#br16) by looking at silhouette [[Rousseeuw, 1987]](#br17) and cophe-

netic correlation [[Brunet et al., 2004]](#br16) indices. Accordingly, the perfor-

mance indices are computed for diﬀerent choices of rank (k < n, m)

and compared to the performance of a random permutation of the

original data. Subsequently, the chosen rank corresponds to the value

with the largest diﬀerence between slopes of the original and the ran-

domized data. Here, the unsupervised pattern training is based on the

consensus set of 1, 905 m6A sites reported in the three miCLIP-based

studies mentioned earlier. Based on the silhouette and cophenetic

correlation indices, we identiﬁed an optimal factorization rank of 6

(Figure [6A)](#br25). We then analyzed the identiﬁed patterns. According to

the membership indicator of each site in matrix W, more than 80% of

m6A modiﬁcation sites can be represented by ﬁve patterns (Patterns

1,2,3,4,6) (Figure [6B).](#br25) Interestingly, the linear combination of these

ﬁve patterns in Figure [6C](#br25) highlights the importance of position 3.

4

4

38

39

Multiple patterns and their combinations can be visualized using visualize pattern

rule. The corresponding outputs are under ”pattern/viz” folder.

4

4

40

41

$ srun snakemake --cores all visualize\_pattern

4. Predict m6A modiﬁcations with the predict modification rule:

4

42

$ srun snakemake --cores all predict\_modification

4

4

43

44

This rule uses patterns of 15 features to predict m6A modiﬁcation.

We examine the ability of prediction on a subset of data of more than

1

2



4

4

4

4

4

4

4

4

4

4

4

4

45

46

47

48

49

50

51

52

53

54

55

56

1.52 million sites including 17, 021 miCLIP m6A sites. We opt for

the linear combination of the ﬁve most relevant patterns described in

step 3. The empirical Cumulative Distribution Function (eCDF) of

the inferred scores shows a signiﬁcant diﬀerence between the diﬀer-

ent miCLIP m6A categories (miCLIP annotation) and the unmodiﬁed

sites (Figure [6D).](#br25) As the number of negative samples is much larger

than the number of positive samples, we consider the Positive Predic-

tive Value (PPV, TP/(TP+FP)) of our predictions. Here, Figure [6E](#br25)

shows that PPV increases with the score cut-oﬀ. The ﬁnal output is a

BED ﬁle containing the estimated scores as well as the corresponding

eCDF and PPV plots. The corresponding outputs are located under

a new folder called ”prediction”.

4

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

4

4

4

4

4

4

4

4

58 An alternative way to detect RNA modiﬁcations is to compare a modi-

59 ﬁed sample to an in-vitro transcribed (IVT) control sample. Therefore,

60 we benchmark JACUSA2 on a sample set of two replicates from wild-type

61 HEK293 cell lines (see Use Case 1m modiﬁed sample) [Pratanwanich et al.](#br17)

62 [[2021]](#br17) and a modiﬁcation-free IVT sample from HEK293 cDNA (control

63 sample) (see ”Preparation of an in vitro transcriptome sample”). All analy-

64 sis steps are identical to use case 1. We evaluate the analysis against miCLIP

65 m6A sites (Figure [5).](#br23)

4

4

4

4

4

4

66

67

68

69

70

71

1. Identify read error proﬁle: we use JACUSA2 call-2 with the same

parameters as the previously described case. The input BAM ﬁles

(HEK293T-WT-rep2.bam, HEK293T-WT-rep3.bam) and (HEK293T-

IVT-rep1.bam, HEK293T-IVT-rep2.bam) are associated to the wild-

type and IVT replicate samples respectively. All input ﬁles are avail-

able from <https://doi.org/10.5281/zenodo.5913452>

4

72

$ srun snakemake --cores all jacusa2\_call2

4

4

4

73

74

75

2. Process JACUSA2 output: we consider all sites corresponding to a

5-mer (NNANN) with a central A residue. We employ the Y ﬁlter to

excludes sites within homo-polymer regions.

4

76

$ srun snakemake --cores all get\_features

4

4

4

4

4

4

77

78

79

80

81

82

3. Extract m6A modiﬁcation pattern: Based on the silhouette and cophe-

netic correlation indices, we identiﬁed an optimal factorization rank of

6 (Figure [7A).](#br26) We determined the predominant factors from matrix

W. Accordingly, more than 80% of m6A modiﬁcation sites can be

represented by four patterns (Patterns: 1,2,3,6) (Figure [7B).](#br26) In agree-

ment with Use Case 1, the linear combination of the four patterns

1

3

4

4

83

84

conﬁrms the importance of position 3 and the implication of all scores

as shown in Figure [7C.](#br26)

4

85

$ srun snakemake --cores all get\_pattern

4

4

4

4

4

4

4

4

4

4

4

86

87

88

89

90

91

92

93

94

95

96

4. Predict m6A modiﬁcations: we evaluate the prediction ability of the

detected patterns on a test set of almost 1, 52 million sites where

17, 021 are miCLIP-m6A modiﬁed. We consider the linear combination

of the four most relevant patterns (1,2,3,6). Figure [7D](#br26) shows the eCDF

of the inferred scores. The diﬀerence between the cumulative distri-

bution of non miCLIP sites and miCLIP sites can be nicely observed,

while the PPV plot shows a lower performance as compared to Use

Case 1 (Figure [7E).](#br26) The decrease in performance is likely explained

by the absence of all modiﬁcations in the IVT samples. Additional

adenosine modiﬁcations such as RNA editing may be counted as false

positives.

4

97

$ srun snakemake --cores all predict\_modification

4

4

98 NOTES

99 Tips and Tricks

5

5

5

5

5

5

00

01

02

03

04

05

1. The reverse transcription step during library preparation is optional.

However, we recommend to include this step to ensure proper sequenc-

ing also of RNAs with secondary structures. Superscript IV reverse

transcriptase may be replaced by Superscript III reverse transcriptase,

which is used in the protocol provided by Oxford Nanopore Technolo-

gies.

5

5

5

5

5

06

07

08

09

10

2. The library preparation protocol contains two bead clean up steps. It

is important to remove ethanol and wash buﬀer completely. However,

beads should not be dried for several minutes. Directly add water

or elution buﬀer after washing to prevent sticking of the RNA to the

beads.

5

5

5

11

12

13

3. The default ﬁlter in current MinKNOW versions is a Q score of 9. For

direct RNA sequencing we recommend to adjust the output ﬁlter to a

minimum Q score of 7, as in previous MinKNOW versions.

5

5

5

14

15

16

4. During preparation of the in vitro transcriptome sample, in vitro tran-

scription and clean up kits may be replaced by equivalent products.

The protocol however has been tested only with the mentioned kits.

5

5

17

18

5. Conﬁguration of the pipeline should be handled via the conﬁg ﬁle:

config.yaml. All parameters should be set before executing rules.

1

4

5

5

5

19

20

21

6. Once the pipeline has run successfully you should expect the following

folders with the corresponding outputs in the output directory: bam,

jacusa, features, patterns, and prediction.

5

5

5

22

23

24

7. JACUSA2 call2 could be run separately using the command line as

described in JACUSA2 manual [[JAC, 2021],](#br15) then put the output under

a new folder with the name ’jacusa’ under the output directory.

5

5

5

5

5

25

26

27

28

29

8. In the snakemake pipeline, rules are linked so that the workﬂows

are determined from top (e.g. predict modiﬁcation) to bottom (e.g.

sort bam) and executed accordingly from bottom to top (Figure [4).](#br22)

Therefore, running for example ”predict modiﬁcation” rule leads to

executing all rules on its pipeline.

5

5

5

5

5

30

31

32

33

34

9. This workﬂow uses a subset of the input data that correspond to

known modiﬁed sites to derive characteristic read alignment proﬁles

by NMF (patterns). Alternatively, predeﬁned patterns could be used

via a preloaded NMF R object to compute scores for modiﬁcation site

prediction.

5

35 ACKNOWLEDGMENTS

5

5

5

36 The authors would like to thank Harald Wilhemi for testing the snakemake

37 pipeline. This work was supported by DFG SPP 1784 (DI1501/11-1) and

38 DFG TRR 319 - RMaP.

5

5

39 REFERENCES

40 Minimap2. <https://github.com/lh3/minimap2>. Accessed: 2022-01-19.

5

5

41 Jacusa2 manual. <https://github.com/dieterich-lab/JACUSA2>, 2021.

Accessed: 2022-01-15.

42

5

5

5

43 Samir Adhikari, Wen Xiao, Yong-Liang Zhao, and Yun-Gui Yang. m(6)a:

44

45

Signaling for mrna splicing. RNA biology, 13:756–759, September 2016.

ISSN 1555-8584. doi: 10.1080/15476286.2016.1201628.

5

5

5

5

46 Ina Anreiter, Quoseena Mir, Jared T. Simpson, Sarath C. Janga, and

47

48

49

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.

5

5

5

50 Konstantinos Boulias, Diana Toczydl owska-Socha, Ben R Hawley, Noa

51

52

Liberman, Ken Takashima, Sara Zaccara, Th´eo Guez, Jean-Jacques

Vasseur, Franc¸oise Debart, L Aravind, et al. Identiﬁcation of the m6am

1

5



5

5

53

54

methyltransferase pcif1 reveals the location and functions of m6am in the

transcriptome. Molecular cell, 75(3):631–643, 2019.

5

5

5

55 Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov.

56

57

Metagenes and molecular pattern discovery using matrix factorization.

Proceedings of the national academy of sciences, 101(12):4164–4169, 2004.

5

5

5

5

5

5

58 Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali

59

60

61

62

63

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.

5

5

5

5

64 Hao Du, Ya Zhao, Jinqiu He, Yao Zhang, Hairui Xi, Mofang Liu, Jinbiao

65

66

67

Ma, and Ligang Wu. Ythdf2 destabilizes m 6 a-containing rna through

direct recruitment of the ccr4–not deadenylase complex. Nature commu-

nications, 7(1):1–11, 2016.

5

5

5

68 Attila Frigyesi and Mattias H¨oglund. Non-negative matrix factorization for

69

70

the analysis of complex gene expression data: identiﬁcation of clinically

relevant tumor subtypes. Cancer informatics, 6:CIN–S606, 2008.

5

5

5

5

71 David Garcias Morales and Jos´e L. Reyes. A birds’-eye view of the activ-

72

73

74

ity 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.

5

5

5

5

5

75 Guifang Jia, Ye Fu, Xu Zhao, Qing Dai, Guanqun Zheng, Ying Yang,

76

77

78

79

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.

5

5

5

5

5

5

5

80 Shengdong Ke, Endalkachew A. Alemu, Claudia Mertens, Emily Conn Gant-

81

82

83

84

85

86

man, John J. Fak, Aldo Mele, Bhagwattie Haripal, Ilana Zucker-Scharﬀ,

Michael J. Moore, Christopher Y. Park, Cathrine Broberg V˚agbø, Anna

Kus´snierczyk, 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.

5

5

5

87 Casslynn WQ Koh, Yeek Teck Goh, and WS Sho Goh. Atlas of quantitative

88

89

single-base-resolution n 6-methyl-adenine methylomes. Nature communi-

cations, 10(1):1–15, 2019.

1

6

5

5

5

5

5

90 Nadine K¨ortel, Cornelia Ru¨ckl´e, You Zhou, Anke Busch, Peter Hoch-Kraft,

91

92

93

94

Reymond FX Sutandy, Jacob Haase, Mihika Pradhan, Michael Musheev,

Dirk Ostareck, et al. Deep and accurate detection of m6a rna modiﬁcations

using miclip2 and m6aboost machine learning. bioRxiv, pages 2020–12,

2021.

5

5

95 Johannes K¨oster and Sven Rahmann. Snakemake—a scalable bioinformatics

workﬂow engine. Bioinformatics, 28(19):2520–2522, 2012.

96

5

5

97 Daniel D Lee and H Sebastian Seung. Learning the parts of objects by

non-negative matrix factorization. Nature, 401(6755):788–791, 1999.

98

5

6

6

6

99 Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christo-

00

01

02

pher 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.

6

6

6

03 Deepak P Patil, Brian F Pickering, and Samie R Jaﬀrey. Reading m6a in

04

05

the transcriptome: m6a-binding proteins. Trends in cell biology, 28(2):

113–127, 2018.

6

6

06 Michael Piechotta, Qi Wang, Janine Altmu¨ller, and Christoph Dieterich.

Rna modiﬁcation mapping with jacusa2. bioRxiv, 2021.

07

6

6

6

6

6

08 Ploy N Pratanwanich, Fei Yao, Ying Chen, Casslynn WQ Koh, Yuk Kei

09

10

11

12

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.

6

6

6

6

13 Jean-Yves

Roignant

and

Matthias

Soller.

m,

14

15

16

javax.xml.bind.jaxbelement@8cec19d, a in mrna: An ancient mech-

anism 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.

6

6

6

17 Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic rna

18

19

modiﬁcations in gene expression regulation. Cell, 169:1187–1200, June

2017. ISSN 1097-4172. doi: 10.1016/j.cell.2017.05.045.

6

6

6

20 Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and

21

22

validation of cluster analysis. Journal of computational and applied math-

ematics, 20:53–65, 1987.

6

6

6

6

23 Hailing Shi, Jiangbo Wei, and Chuan He. Where, when, and how:

24

25

26

Context-dependent functions of rna methylation writers, readers, and

erasers. Molecular cell, 74:640–650, May 2019. ISSN 1097-4164. doi:

10.1016/j.molcel.2019.04.025.

1

7

6

6

6

6

27 Xiao Wang, Zhike Lu, Adrian Gomez, Gary C Hon, Yanan Yue, Dali Han,

28

29

30

Ye Fu, Marc Parisien, Qing Dai, Guifang Jia, et al. N 6-methyladenosine-

dependent regulation of messenger rna stability. Nature, 505(7481):117–

120, 2014.

6

6

6

6

31 Xiao Wang, Boxuan Simen Zhao, Ian A Roundtree, Zhike Lu, Dali Han,

32

33

34

Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He.

N6-methyladenosine modulates messenger rna translation eﬃciency. Cell,

161(6):1388–1399, 2015.

6

6

6

35 Sara Zaccara, Ryan J. Ries, and Samie R. Jaﬀrey. Reading, writing and

36

37

erasing mrna methylation. Nature reviews. Molecular cell biology, 20:608–

624, October 2019. ISSN 1471-0080. doi: 10.1038/s41580-019-0168-5.

6

6

6

6

38 Zhang Zhang, Tao Chen, Hong-Xuan Chen, Ying-Yuan Xie, Li-Qian Chen,

39

40

41

Yu-Li Zhao, Biao-Di Liu, Lingmei Jin, Wutong Zhang, Chang Liu,

et al. Systematic calibration of epitranscriptomic maps using a synthetic

modiﬁcation-free rna library. Nature Methods, 18(10):1213–1222, 2021.

6

6

6

6

6

6

6

6

6

42 Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min

43

44

45

46

47

48

49

50

Huang, Charles J. Li, Cathrine B. V˚agbø, Yue Shi, Wen-Ling Wang, Shu-

Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin

Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan,

Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne

Klungland, Yun-Gui Yang, and Chuan He. Alkbh5 is a mammalian rna

demethylase that impacts rna metabolism and mouse fertility. Molecular

cell, 49:18–29, January 2013. ISSN 1097-4164. doi: 10.1016/j.molcel.2012.

10.015.

1

8

6

51 FIGURES

Computational approach

Condition 1

Condition 2

Insertions

Deletions

Mismatches

Not modified

at all

Mismatches

Mapped

reads

KO

samples

G

GG

I

G

G

I

I

I

Reference

sequence

A

A

Missing specific

modification(s)

Mismatch

rate

**Samples**

**of**

**interest**

Insertion

rate

All

Deletion

rate

modifications

present

Simulated

samples

IVT

samples

Coverage

pro)le

Figure 1: General outline of RNA modiﬁcation detection by JA-

CUSA2. A key feature of our approach is that multiple replicates of

the same condition can be considered simulatenously. Samples of interests

where all modiﬁcations are present could be compared with either KO sam-

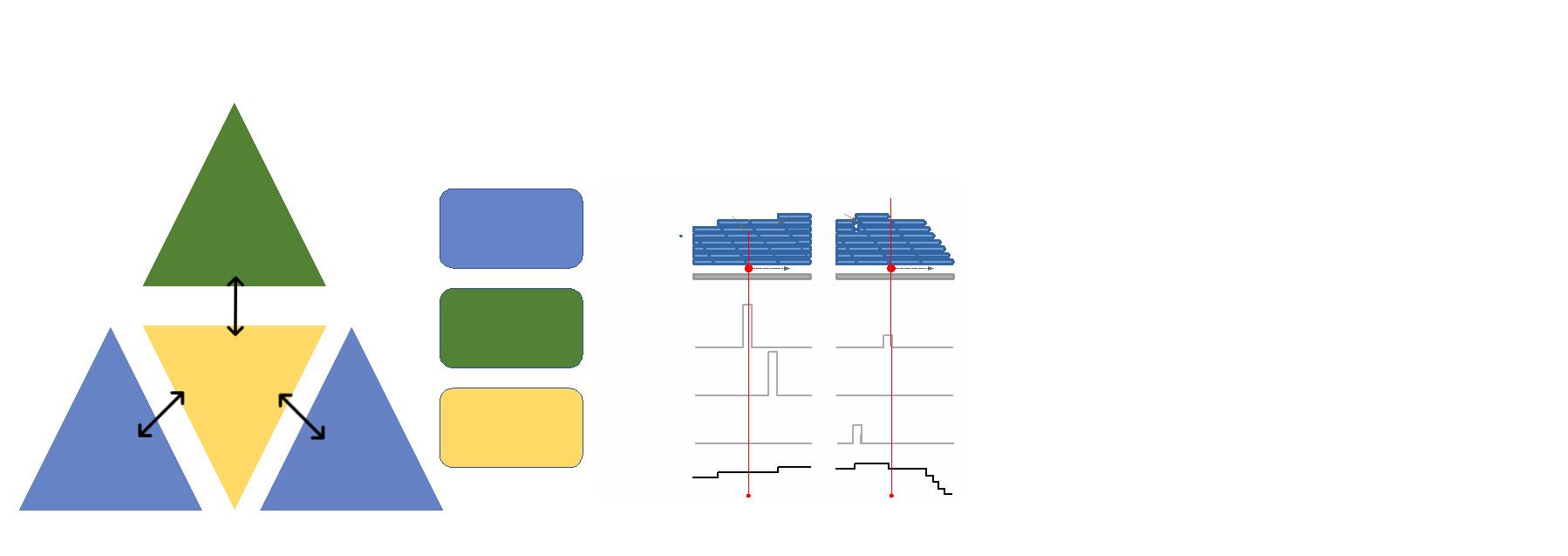
ples where the modiﬁcation of interest is missing or IVT/simulated samples

where all modiﬁcations are absent (left panel). Read stacks (in blue) are

compared head-to-head as shown on the right.

1

9



Deletion

Mismatch

Modification-specific pattern (Feature set)

JACUSA2 score

Insertion

Figure 2: Motivation of 5-mer context for RNA modiﬁcation map-

ping. The nanopore covers 5 consecutive RNA residues. That is why we

consider a 5-mer context and derive 3 principal features for every position

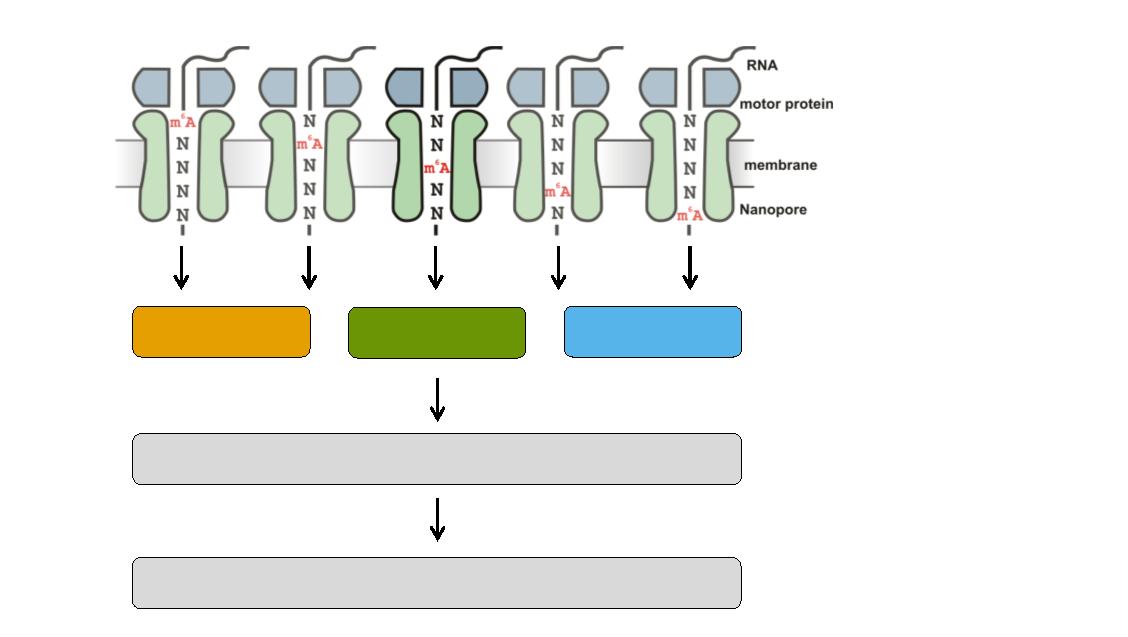
within a given 5-mer (15 features in total, with a central A residue in this

example). We evaluate each feature set by previously learned patterns and

compute a ﬁnal score for modiﬁcation site detection.

2

0



**A)**

**B)**

ONT direct

RNA-seq

Guppy

basecalling

polyA+

RNA

FASTQ

files

Total RNA

(SQK-RNA002)

(high accuracy)

FASTQ

files

minimap2

alignment

SAMtools

filtering

JACUSA2

analysis

Candidate

m6A sites

Figure 3: Experimental and computational workﬂow. A) Starting

from total cellular RNA, polyA+ RNA is isolated and subjected to Nanopore

direct RNA-sequencing. Guppy basecalling can be done as live basecalling

during sequencing or after the sequencing run from generated FAST5 ﬁles,

resulting in FASTQ output ﬁles. B) FASTQ ﬁles are aligned to a reference

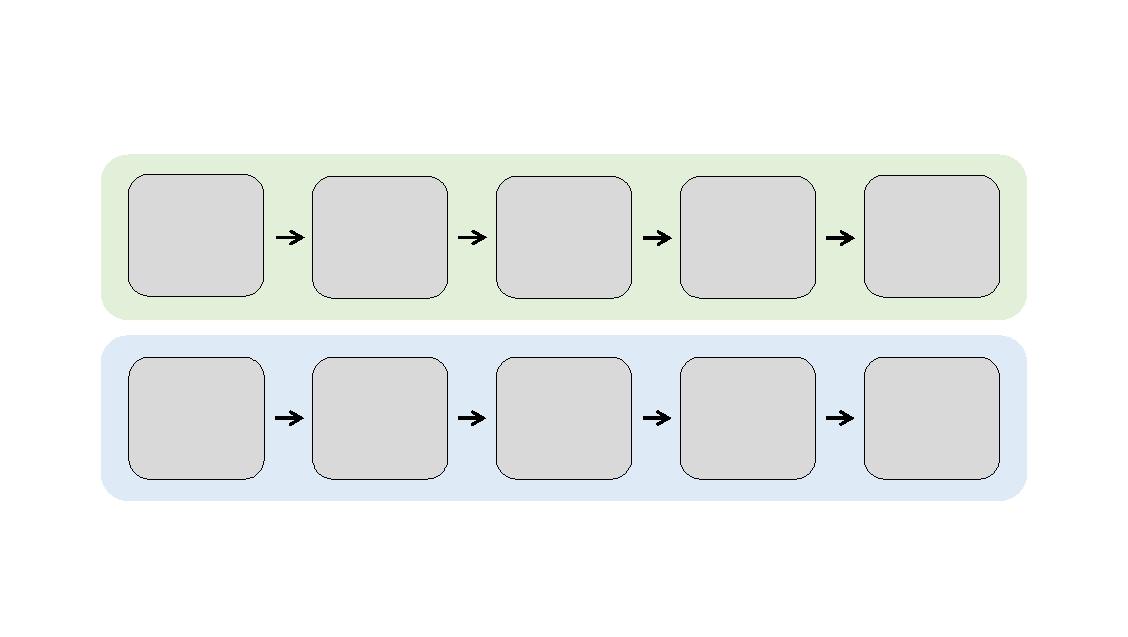
sequence with Minimap2. SAMtools is used to generate BAM ﬁles as input

for JACUSA2 analysis, which yields candidate m6A sites as described in

Figure 4.

2

1



sort\_bam

sort\_bam

sort\_bam

sort\_bam

Input: mapping\_cond1\_rep1.bam

Input: mapping\_cond1\_rep2.bam

Input: mapping\_cond2\_rep1.bam

Input: mapping\_cond2\_rep2.bam

ﬁlter\_bam

ﬁlter\_bam

ﬁlter\_bam

ﬁlter\_bam

jacusa2\_call2

get\_features

split\_train\_test

get\_pattern

predict\_modiﬁcation

visualize\_pattern

Figure 4: Computational workﬂow. Snakemake workﬂow for RNA mod-

iﬁcation detection based on JACUSA2 variant calling.

2

2

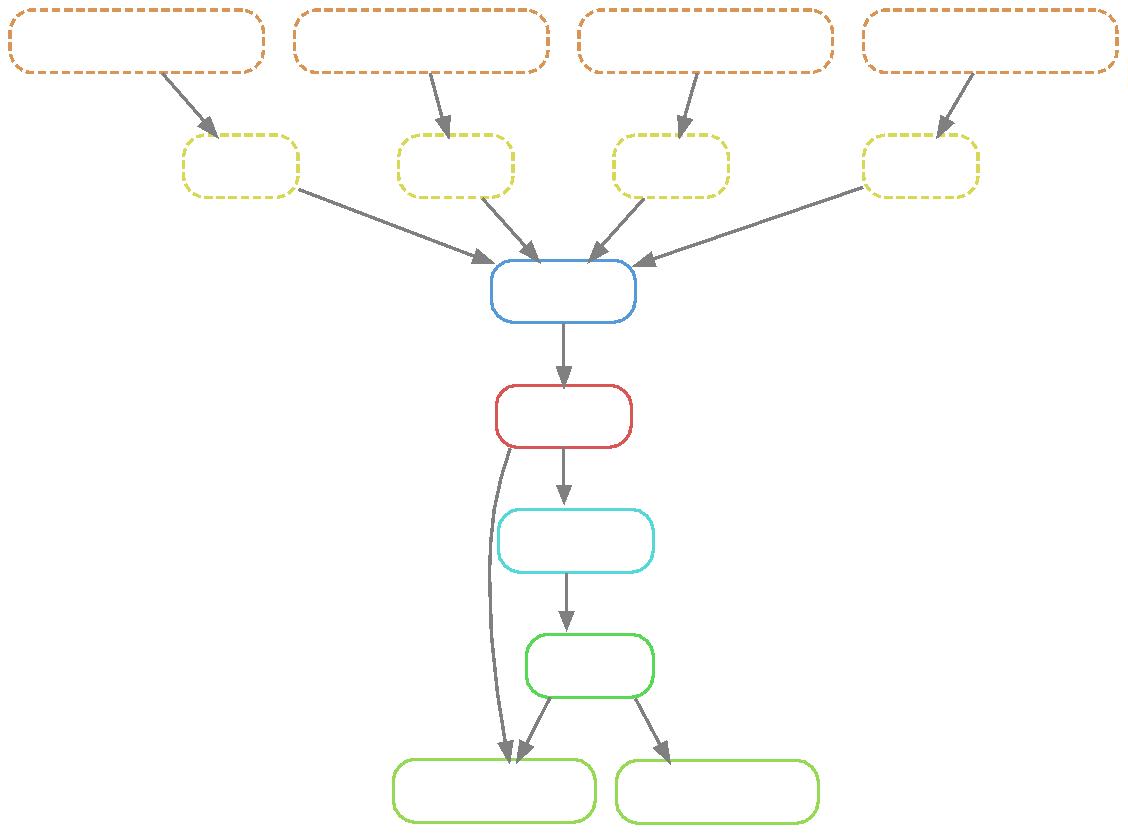
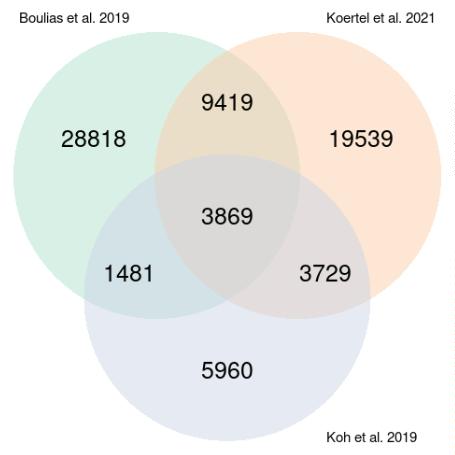


Figure 5: m6A sites reported in the three miCLIP-based studies

Boulias et al. [2019], Koh et al. [2019] and K¨ortel et al. [2021].

2

3



6

52 TABLES

Software Version

Description

Minimap2 <https://github.com/lh3/minimap2>

v2.22 or later

<https://lh3.github.io/minimap2/>

samtools [https://github.com/samtools/](https://github.com/samtools/samtools)

[samtools](https://github.com/samtools/samtools) v1.12 or later

<http://samtools.github.io/>

JAVA

<https://openjdk.java.net/>

021-07-20 - J AVA 11 or later

<https://www.r-project.org/>version The R Project for Statistical Comput-

.5.1 or later ing

<https://www.perl.org/>version 5.28.1 Perl is a highly capable, feature-rich

or later programming language

bedtools <https://github.com/arq5x/bedtools2>Perl is a highly capable, feature-rich

version 2.29.2 or later programming language

snakemake <https://snakemake.github.io/>version The Snakemake workﬂow management

11.0.12 OpenJDK Runtime Environment

2

R

3

PERL

6

.8.1 or later

system

Table 1: Software dependencies

R Pack- Version

ages

Description

ggplot2 [https://cran.r-project.org/web/](https://cran.r-project.org/web/packages/ggplot2/index.html)

-

ggplot2 is a system for declaratively

gg- creating graphics, based on The Gram-

mar of Graphics.

[packages/ggplot2/index.html](https://cran.r-project.org/web/packages/ggplot2/index.html)

plot2 3.3.0 or later

NMF

[https://cran.r-project.org/web/](https://cran.r-project.org/web/packages/NMF/index.html) Provides a framework to perform Non-

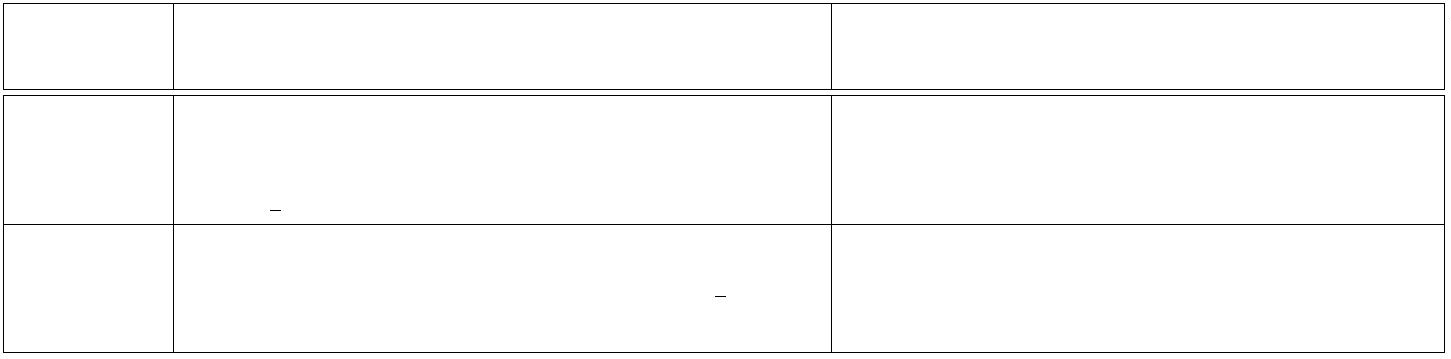
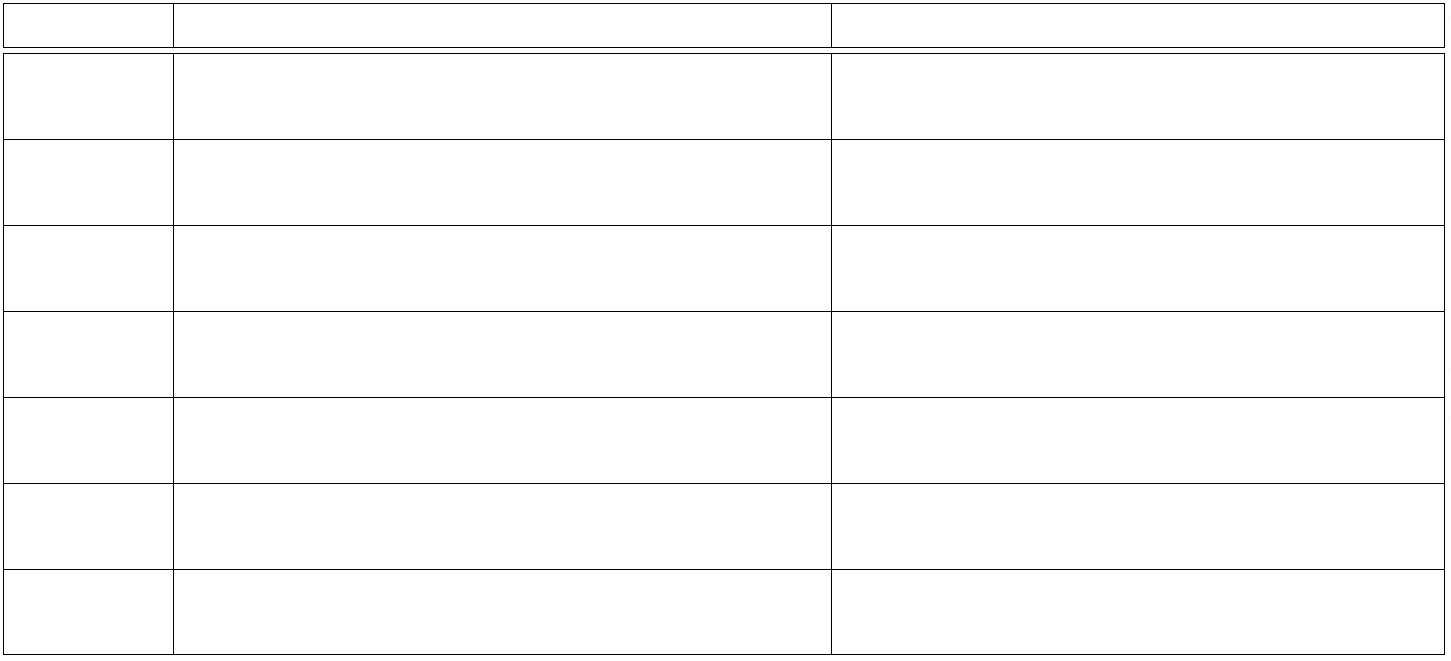
[packages/NMF/index.html](https://cran.r-project.org/web/packages/NMF/index.html) - NMF 0.22.0 negative Matrix Factorization (NMF).

or later

Table 2: R Package dependencies

2

4



A- NMF rank selection

B- NMF factorization

Basis components

Mixture coefficients

1

basis

Silhouette

Cophenetic

1

2

3

4

5

6

0

0

0

0

0

.8

.6

.4

.2

1

basis

1

2

3

4

5

6

1

2

3

4

5

6

0.8

0.6

0.4

0.2

0

consensus

2

3

4

5

6

7

8

9

10

1

2

3

4

5

6

Rank

C- Sum of top-ranked NMF patterns

Insertion

Deletion

Mismatch

Pos1 Pos2 Pos3 Pos4 Pos5

1

2

3

4

5

6

D- eCDF of sum of patterns scores

E- Overlap of m6A prediction and miCLIP sites

1

0

0

0

0

.00

.75

.50

.25

.00

2

1

0

5

100

75

50

25

0

CLIP

Boulias

Boulias,Koertel

Boulias,Koertel,Koh

Boulias,Koh

Koertel

10

5

Koertel,Koh

Koh

NoCLIP

0

0

10

20

30

0

10

20

30

score

Cutoff

Figure 6: Case 1. WT versus KO. A: NMF rank selection. Barplots rep-

resent diﬀerences in cophenetic correlation and silhouette scores for NMF

factorization of the original input and randomized data. Ranks with the

largest diﬀerences are determined for both measures, then the smallest rank

is selected for the NMF decomposition. B: NMF result represented by the

basis matrix W (”amplitude” matrix) and the coeﬃcient matrix H (”pat-

tern” matrix). The matrix H induces the RNA modiﬁcation pattern. C:

Barplot shows the sum of the top 5 patterns (y-axis) across the speciﬁc 5-

mer context (x-axis) and the score type: mismatch, deletion and insertion

(resp. black, orange and blue). The 5 patterns (coeﬃcient vectors: 1,2,3,4,6)

are selected according to the predominant columns in matrix W. D: Score

distribution inferred from the combined patterns, stratiﬁed by the diﬀerent

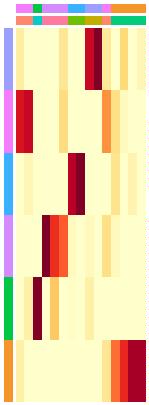
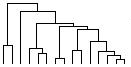
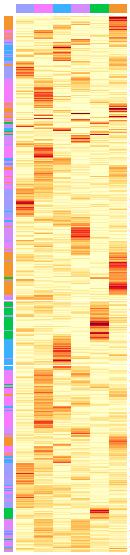
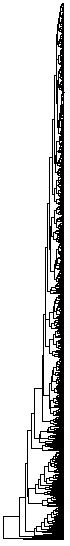
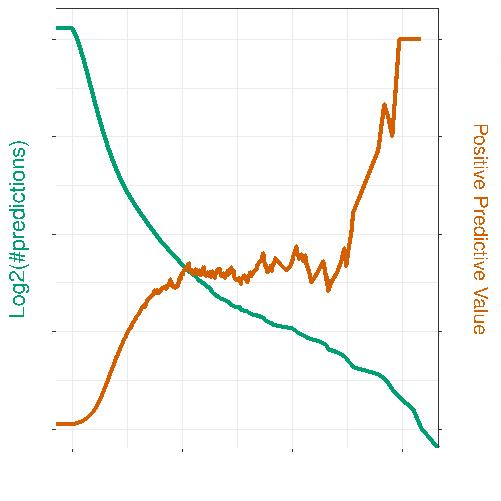
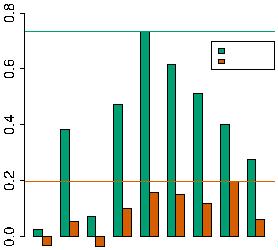
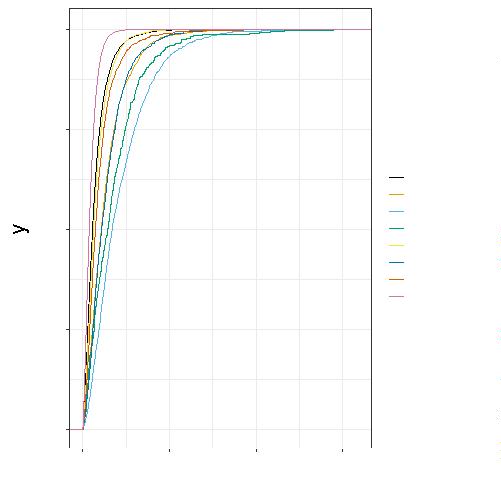
categories of miCLIP validated sites and non miCLIP sites. E: Number of

predicted m6A sites (green) and Positive Predictive Value (PPV) of pre-

dicted m6A sites that overlap with miCLIP sites (orange).

2

5



A- NMF rank selection

B- NMF factorization

Silhouette

Cophenetic

Basis components

Mixture coefficients

1

basis

1

2

3

4

5

6

0

0

0

0

0

.8

.6

.4

.2

1

basis

1

2

3

4

5

6

1

2

3

4

5

6

0.8

0.6

0.4

0.2

0

consensus

2

3

4

5

6

7

8

9

10

1

2

3

4

5

6

Rank

C- Sum of top-ranked NMF patterns

Insertion

Deletion

Mismatch

Pos1 Pos2 Pos3 Pos4 Pos5

1

2

3

4

5

6

D- eCDF of sum of patterns scores

E- Overlap of m6A prediction and miCLIP sites

1

0

0

0

0

.00

.75

.50

.25

.00

20

15

10

5

0

100

75

50

25

0

CLIP

Boulias

Boulias,Koertel

Boulias,Koertel,Koh

Boulias,Koh

Koertel

Koertel,Koh

Koh

NoCLIP

0

10

Score

20

0

10

20

Cutoff

Figure 7: Case 2. WT versus IVT. A: NMF rank selection. Barplots rep-

resent diﬀerences in cophenetic correlation and silhouette scores for NMF

factorization of the original input and randomized data. Ranks with the

largest value for both indices are determined, then the smallest rank is se-

lected for the NMF decomposition. B: NMF result represented by the basis

matrix W (”amplitude” matrix) and the coeﬃcient matrix H (”pattern”

matrix). The matrix H induces the RNA modiﬁcation pattern. C: Barplots

representing the linear combination of the top 4 patterns (y-axis) by 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 pat-

terns (coeﬃcient vectors: 1,2,3,6) are selected according to the predominant

columns in matrix W. D: Score distribution inferred from the combined

patterns, stratiﬁed by the diﬀerent categories of miCLIP validated sites and

non miCLIP sites. E: Number of predicted m6A sites (green) and Positive

Predictive Value (PPV) of predicted m6A sites that overlap with miCLIP

sites (orange).

2

6

