Optimisation of Next Generation sequencing methodologies for RNA modifications detection

Dissertation zur Erlangung des Grades
« Doktor der Naturwissenschaften »
im Promotionsfach Pharmazie

am Fachbereich Chemie, Pharmazie, Geographie und Geowissenschaften
der Johannes Gutenberg-University Mainz

Florian Pichot
Geboren 01.09.1995 in Poitiers (Frankreich)

Mainz, Juli 2023



Dekan: Prof. Dr. Eva Rentschler

1. Berichterstatter:

2. Berichterstatter:

Tag der mündlichen Prüfung: 8. Dezember 2023

D77 (Dissertation Mainz)

FB09 Fachbereich
Chemie, Pharmazie, Geographie und
Geowissenschaften

Versicherung

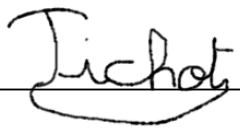
für das Gesuch um Zulassung zur Promotion im Fachbereich 09 an der Johannes Gutenberg-Universität Mainz

Hiermit versichere ich Pichot Florian
Name, Vorname

gemäß § 13 Abs. (3), 3-7 der Promotionsordnung vom 18.10.2021 i.d.F. vom 17.08.2022, dass:

- die eingereichte wissenschaftliche Arbeit noch an keiner anderen deutschen oder ausländischen Hochschule oder vergleichbaren Einrichtung zur Erlangung eines akademischen Grades eingereicht wurde
- ich noch kein Promotions- oder ein vergleichbares Graduierungsverfahren im Promotionsfach erfolglos beendet habe
- ich noch kein Promotions- oder ein vergleichbares Graduierungsverfahren im Fach der Promotion erfolgreich abgeschlossen habe
- die wissenschaftliche Arbeit selbstständig verfasst und ausschließlich die angegebenen Quellen und Hilfsmittel verwendet wurden
- keine entgeltliche Hilfe Dritter, insbesondere Promotionsberatung oder -vermittlung in Anspruch genommen wurde.

25.08.2023
Datum

Unterschrift 

« La joie est dans le risque de faire du neuf. »

Denis Diderot

Joy is within the risk of making something new.

Contents

Contents v
Abstract vii
Zusammenfassung ix
Abbreviations table xi
I. Introduction 1
1.1 The different RNA species and their functions 1
1.1.1 Ribosomal RNA (rRNA) 1
1.1.2 Transfer RNA (tRNA) 2
1.1.3 Messenger RNA (mRNA) 2
1.1.4 RNA regulators (siRNA, miRNA, tRFs) 2
1.1.5 small nucleolar RNA (snoRNA) 3
1.2 Post-transcriptional modifications 3
1.2.1 2'-O-methylation 5
RNA ribose methylation (2'-O-methylation): occurrence, biosynthesis and biological functions. [DOI: 10.1016/j.bbagrm.2018.11.009] 6
1.2.2 Pseudouridine (Psi) 30
1.2.3 N6-methlyadenosine (m^6A) 30
1.2.4 Impact on diseases 30
1.3 Detection of RNA modifications 31
1.3.1 Physico-chemical detection methods 31
1.3.2 Sequencing 32
RT-signatures 33
Chemical adducts for RT stops induction 34
Antibody-based methods 34
Chemical treatment for modifications' signals extraction 34
Nanopore 35
II. Motivation and objectives 37
III. Contribution to papers 39
IV. Results and Discussion 42
Chapter 1: Illumina-adapted RiboMethSeq method 42
4.1.1 Next-Generation Sequencing-Based RiboMethSeq Protocol for Analysis of tRNA 2'-O-Methylation. [DOI: 10.3390/biom7010013] 42

4.1.2 Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2'-O-Methylations in RNA by RiboMethSeq. [DOI: 10.3389/fgene.2020.00038] 71

4.1.3 Non-Redundant tRNA Reference Sequences for Deep Sequencing Analysis of tRNA Abundance and Epitranscriptomic RNA Modifications. [DOI: 10.3390/genes12010081] 96

4.1.4 2'-O-methylation within prokaryotic and eukaryotic tRNA inhibits innate immune activation by endosomal Toll-like receptors but does not affect recognition of whole organisms. [DOI: 10.1261/rna.070243.118] 127

4.1.5 Machine learning algorithm for precise prediction of 2'-O-methylation (Nm) sites from experimental RiboMethSeq datasets. [DOI: 10.1016/j.ymeth.2022.03.007] .. 152

Chapter 2: Creation of alternative detection methods 188

4.2.1 HydraPsiSeq: a method for systematic and quantitative mapping of pseudouridines in RNA. [DOI: 10.1093/nar/gkaa769] 188

4.2.2 Data Analysis Pipeline for Detection and Quantification of Pseudouridine (ψ) in RNA by HydraPsiSeq. [DOI: 10.1007/978-1-0716-2962-8_14] 229

4.2.3 NOseq: amplicon sequencing evaluation method for RNA m^6A sites after chemical deamination. [DOI: 10.1093/nar/gkaa1173] 245

Final chapter: Compiled application of the developed methods 281

4.3 Quantification of substoichiometric modification reveals global tsRNA hypomodification, preferences for angiogenin-mediated tRNA cleavage, and idiosyncratic epitranscriptomes of human neuronal cell-lines. [DOI: 10.1016/j.csbj.2022.12.020] 281

V. Conclusion and outlook 345

Optimisation of RiboMethSeq in adequation with biological studies 345

Importance of statistical analyses applications for error minimization 345

Creation of modifications detection techniques for new epitranscriptomics discoveries 347

Tool sharpening leading to cutting-edge epitranscriptomics discoveries 348

Bibliography 349

Curriculum Vitae 397

Abstract

Ribonucleic acid (RNA) is an essential biomolecule in the domain of life. As a mediator between the genetic information contained in deoxyribonucleic acid (DNA) and the cell, its regulation has a major impact on the homeostasis of living organisms. The discovery and identification of chemical modifications at nucleotide level in RNA has made it possible to study a new layer of regulation of these molecules. In view of the importance of these chemical alterations within the biology domain, a specific term for the study of these modifications has been coined: epitranscriptomic. More than 170 modifications have now been listed, and their analysis has led to the discovery of numerous links between these alterations and various diseases such as viral infections, cancers and neurodegenerative diseases. Epitranscriptomic therefore holds out the hope of developing innovative treatments for yet incurable diseases.

However, of the large number of modifications identified to date, few methods exist that can locate them precisely in RNA sequences. In order to overcome these limitations, the work presented here concerns the development and optimisation of methods for detecting modifications through high-throughput sequencing. More specifically, the focus of this work is on the computational processing of sequencing data in order to perfect the detection of these modified nucleotides.

As a first step, an application of a method for detecting 2'-O-methylations via high-throughput sequencing entitled RiboMethSeq was carried out on tRNA samples from *Escherichia coli* and *Saccharomyces cerevisiae* with the aim of studying the role of these modifications in innate immunity. This study uncovered interesting details about the regulation of immunity, but also revealed detection limitations on the part of RiboMethSeq. These limitations are mainly due to the non-optimal parameterisation of the various data processing stages, but also to the lack of transfer RNA (tRNA) sequence references suitable for epitranscriptomic studies. These limitations have been specifically addressed and a thorough optimisation of these two concepts has been implemented, enabling a more in-depth analysis of the links between 2'-O-methylations and innate immunity in a second study. Finally, with the aim of pushing back the limits of 2'-O-methylation detection, the potential of deep learning algorithms for detecting modifications is explored using RiboMethSeq ribosomal RNA (rRNA) data as a training set for Random Forest algorithm.

Secondly, development of alternative methods for the detection of two modifications, N6-methlyadenosine (m^6A) and pseudouridine (Psi) is carried out. Existing detection methods for these modifications have limitations due to the nature of the approach chosen. In order to fill these gaps, two alternatives processes based on the induction of chemical signatures emitted by these modifications were proposed and then applied to biological samples, proving their robustness of detection as well as their quality of quantification of these biomolecules of interest.
Finally, this work concludes with the simultaneous use of three detection methods - RiboMethSeq, AlkAnilineSeq and BisulfiteSeq - on the same set of brain cell samples in the

context of studies on neurodegenerative diseases. The combination of these methods enabled five human tRNA modifications to be mapped, and allowing their respective quantification according to the nature of the cell tissue and stress conditions. This work has enabled us to confirm in greater detail observations already seen in the literature, but also to highlight a still little-studied modification, Dihydrouridine, as a potential determining factor in tRNA fragmentation.

Zusammenfassung

Ribonukleinsäure (RNA) ist ein wesentliches Biomolekül im Bereich des Lebens. Als Vermittler zwischen der in der Desoxyribonukleinsäure (DNA) enthaltenen genetischen Information und der Zelle hat ihre Regulierung einen großen Einfluss auf die Homöostase lebender Organismen. Die Entdeckung und Identifizierung chemischer Veränderungen auf Nukleotidebene in der RNA hat es ermöglicht, eine neue Ebene der Regulierung dieser Moleküle zu untersuchen. Angesichts der Bedeutung dieser chemischen Veränderungen im Bereich der Biologie wurde ein spezieller Begriff für die Untersuchung dieser Modifikationen geprägt: Epitranskriptomik. Mehr als 170 Modifikationen wurden inzwischen aufgelistet, und ihre Analyse hat zur Entdeckung zahlreicher Verbindungen zwischen diesen Veränderungen und verschiedenen Krankheiten wie Virusinfektionen, Krebs und neurodegenerativen Erkrankungen geführt. Die Epitranskriptomik birgt daher die Hoffnung auf die Entwicklung innovativer Behandlungen für bisher unheilbare Krankheiten.

Von der großen Zahl der bisher identifizierten Modifikationen gibt es jedoch nur wenige Methoden, mit denen sie in RNA-Sequenzen genau lokalisiert werden können. Um diese Einschränkungen zu überwinden, befassen sich die hier vorgestellten Arbeiten mit der Entwicklung und Optimierung von Methoden zum Nachweis von Modifikationen durch Hochdurchsatz-Sequenzierung. Genauer gesagt liegt der Schwerpunkt dieser Arbeit auf der computergestützten Verarbeitung von Sequenzierungsdaten, um den Nachweis dieser modifizierten Nukleotide zu perfektionieren.

In einem ersten Schritt wurde eine Methode zum Nachweis von 2'-O-Methylierungen mittels Hochdurchsatzsequenzierung (RiboMethSeq) auf transfer-RNA (tRNA) Proben aus *Escherichia coli* und *Saccharomyces cerevisiae* angewandt, um die Rolle dieser Modifikationen in der angeborenen Immunität zu untersuchen. Diese Studie enthüllte interessante Details über die Regulierung der Immunität, zeigte aber auch die Nachweisgrenzen von RiboMethSeq auf. Diese Einschränkungen sind hauptsächlich auf die nicht optimale Parametrisierung der verschiedenen Datenverarbeitungsstufen zurückzuführen, aber auch auf den Mangel an tRNA-Sequenzreferenzen, die für epitranskriptomische Studien geeignet sind. Diese Einschränkungen wurden gezielt angegangen, und es wurde eine gründliche Optimierung dieser beiden Konzepte vorgenommen, so dass in einer zweiten Studie die Zusammenhänge zwischen 2'-O-Methylierungen und angeborener Immunität eingehender analysiert werden können. Mit dem Ziel, die Grenzen der Erkennung von 2'-O-Methylierungen zu verschieben, wird schließlich das Potenzial von Deep-Learning-Algorithmen für die Erkennung von Modifikationen unter Verwendung von RiboMethSeq ribosomale RNA-Daten als Trainingsset für den Random Forest-Algorithmus erforscht.

Zweitens werden alternative Methoden für den Nachweis von zwei Modifikationen, N6-Methlyadenosin (m^6A) und Pseudouridin (Psi), entwickelt. Bestehende Nachweismethoden für diese Modifikationen weisen aufgrund der Art des gewählten Ansatzes Beschränkungen auf. Um diese Lücken zu schließen, wurden zwei alternative Verfahren vorgeschlagen, die

auf der Induktion chemischer Signaturen beruhen, die von diesen Modifikationen ausgehen, und dann auf biologische Proben angewandt, wobei ihre Robustheit bei der Erkennung sowie ihre Qualität bei der Quantifizierung dieser Biomoleküle von Interesse bewiesen wurden.
Die Arbeit schließt mit der gleichzeitigen Anwendung von drei Nachweismethoden - RiboMethSeq, AlkAnilineSeq und BisulfiteSeq - auf denselben Satz von Gehirnzellproben im Rahmen von Studien über neurodegenerative Erkrankungen. Die Kombination dieser Methoden ermöglichte es, fünf menschliche tRNA-Modifikationen zu kartieren und ihre jeweilige Quantifizierung je nach Art des Zellgewebes und der Stressbedingungen zu bestimmen. Diese Arbeit hat es uns ermöglicht, bereits in der Literatur gemachte Beobachtungen zu bestätigen, aber auch eine noch wenig untersuchte Modifikation, Dihydrouridin, als potenziell entscheidenden Faktor bei der tRNA-Fragmentierung herauszustellen.

Abbreviations table

DNA	Deoxyribonucleic Acid
2D	Two dimensional
3D	Three-dimensional
Ab	Antibody
AIDS	Acquired Immune Deficiency Syndrome
ALS	Amyotrophic lateral sclerosis
Am	2’-O-methyladenosine
ANG	Angiogenin
AUC	Area Under Curve
BMSC	Bone Marrow Stem Cells
BS	Bisulfite-Sequencing
Cm	2’-O-methylcytidine
CMC/CMCT	1-cyclohexyl-(2-morpholinoethyl)carbodiimide metho-p-toluene sulfonate
D	Dihydrouridine
DMF	Dimethylformamide
DMSO	Dimethyl sulfoxide
E. coli	*Escherichia coli*
FDR	False Discovery Rate
FTD	Frontotemporal Dementia
Gm	2’-O-methylguanosine
HIV	Human Immunodeficiency Virus
IGV	Integrative Genomics Viewer
KO	Knock-Out
LC/HPLC	Liquid Chromatography/High-Performance Liquid Chromatography
LC-MS/MS	Liquid Chromatography-tandem Mass Spectrometry
LSU	Large Subunit
m^1A	N1-methyladenosine
m^1G	N1-methylguanosine
m^3C	N3-methylcytidine
m^5C	5-Methylcytidine
m^6A	N6-methyladenosine
m^7G	N7-methylguanosine
MCC	Matthews correlation coefficient
MeRIP	Methylated (m6A and others) RNA immunoprecipitation
miRNA	Micro-Ribonucleic Acid
mln	millions
mRNA	Messenger Ribonucleic Acid
MS	Mass Spectrometry
MZ	MZ-294 human glioblastoma cell line
NGS	Next-Generation Sequencing

PCR	Polymerase Chain reaction
PD	Parkinson’s disease
RNA	Ribonucleic Acid
RNAi	Ribonucleic Acid Interference
RNA-seq	Ribonucleic Acid Sequencing
RNP	Ribonucleoprotein
ROC	Receiver Operating Characteristic
rRNA	Ribosomal Ribonucleic Acid
RT	Reverse Transcriptase
S. cerevisiae	*Saccharomyces cerevisiae*
SCARLET	site-specific cleavage and radioactive-labelling followed by ligation-assisted extraction and thin-layer chromatography
SH	SH-SY5Y human neuroblastoma cell line
siRNA	Small Interfering Ribonucleic Acid
snoRNA	Small nucleolar Ribonucleic Acid
snoRNP	Small nucleolar Ribonucleoprotein
SNP	Single-Nucleotide Polymorphism
SSU	Small Subunit
TGF-b	transforming growth factor b
tiRNA	tRNA halves
TLC	Thin Layer Chromatography
TNR	specificity, selectivity or true negative rate
TPR	sensitivity or true positive rate
tRFs	tRNA fragments
tRNA	Transfer Ribonucleic Acid
tsRNA	tRNA-derived small Ribonucleic Acid
Um	2’-O-methyluridine
ψ/Psi	Pseudouridine

I. Introduction

In 1958, Crick explained his central dogma[1] which would become the pinnacle of molecular biology. Deoxyribonucleic acid (DNA), support of inherited genetic information is transcribed to ribonucleic acid (RNA), which can be translated to proteins, vital regulators of organisms' homeostasis, in a one-way manner. From this dogma and its controversies have given rise to numerous studies about the cross-talks between these three major biological molecules and their contribution to life. Nowadays, progress in molecular biology are proving that RNA and proteins can regulate genetic expression[2, 3], proteins can regulate RNA through degradation[4] and RNA and proteins can form complexes known as ribonucleoprotein particles (RNPs) essential for translation and post-transcriptional control (splicing[5] and polyadenylation[6] for example).

These recent studies shed light on RNA as an intermediary between genetic expression and organism survival. Some viruses are not using DNA as genetic support but RNA[7, 8], thus highlighting RNA as a central molecular piece across species and evolution. The various roles and functions of RNA have been thoroughly studied for the past decades, yielding new insights into diseases machinery and creation of RNA-based drugs such as mRNA vaccines[9] beneficial to healthcare worldwide. Yet, the whole contribution of RNA to life is still not fully understood, and its remaining mysteries carry on a bright potential to increase life expectancy.

1.1 The different RNA species and their functions

Through its diversified functions of genetic information holder, DNA/RNA regulators and chemical reaction catalysts, RNA is divided into multiple subtypes according to their main role. A non-exhaustive list on essential RNA types is given in the following section.

1.1.1 Ribosomal RNA (rRNA)

Ribosomal RNA is the most abundant RNA species in cells[10]. rRNA forms complexes with ribosomal proteins to form the ribosome which carries out cells' translation, the protein synthesis. It is composed of two units known as Small subunit (SSU) and Large subunit (LSU). SSU contains one rRNA molecule whereas LSU contains two rRNA molecules for prokaryotes and three for eukaryotes. rRNAs denominations depend on their size in Svedberg units[11] and thus, are different across species.

Being the catalytic core of the translation, stability of the ribosome is critical for correct protein assembly[12] and so cell survival. Different regulation mechanisms such as epitranscriptomic modifications ensure rRNA correct folding to fulfil its duties[13].

1.1.2 Transfer RNA (tRNA)

The second most abundant RNA species[10], tRNA is also involved in protein synthesis. tRNAs carry the different amino acids which are the molecular units of proteins to the ribosomal machinery. tRNAs molecules are chemically stable through their well-characterized structure in a cloverleaf shape. tRNA structure can be divided into five distinct parts: the acceptor stem, the D-loop, the anticodon loop, the variable loop, and the TψC-loop.

Each of these structures has an important structural role for the chemical stabilization of the molecule and its recognition by tRNA-binding proteins and the ribosome. Of note, tRNAs possess a high number of modified nucleosides to ensure their correct folding to enable them to function[14, 15].

Inside the ribosome, the tRNA anticodon loop binds to messenger RNAs (mRNAs) by a 3-nucleotide hybridisation interaction known as “codon/anticodon interaction”. This leads to the delivery of specific amino acids from the tRNA to synthesise, amino acid by amino acid, a functional protein.

At the same time, to guarantee that each tRNA molecule carries the correct amino acid among the twenty available, different cellular regulatory mechanisms carry out numerous aminoacylation quality controls[16, 17], mainly through the correct identification of tRNA molecule thanks to the anticodon loop and the acceptor stem. These quality controls allow the trustworthy decoding of the information contained inside mRNA to properly vehicle genomic instruction.

1.1.3 Messenger RNA (mRNA)

Third most abundant RNA molecule[10], mRNAs carry genetic information thanks to the transcription of DNA genes by RNA polymerases.
These molecules undergo different maturation steps from their pre-mRNA form to their mature form, including splicing, 5’cap addition and polyadenylation, which enables modulation of gene expression [18–20].
To ensure genetic expression quality, mature mRNA is still heavily regulated, through degradation enzymes known as nucleases[21], sensors such as Toll-like receptors[22] and RNA regulators like small interfering RNA (siRNA) and microRNA (miRNA).

1.1.4 RNA regulators (siRNA, miRNA, tRFs)

RNA regulators are short, single-stranded RNA molecules that can specifically bind to mRNA to prevent the latter from being translated. This gene silencing is called RNA interference (RNAi) and is mainly centred around two molecules: siRNA and miRNA. Those two short RNA molecules (around 20 nucleotides long) can silence either one or multiple mRNAs, respectively for siRNA and miRNA[23].

Along the same lines, tRNA-derived small RNAs (tsRNA) are products of tRNA fragmentation performed by RNA nucleases. These fragments can effect gene silencing by binding to mRNA like siRNA and miRNA[24]. Depending on their site of origin, different types of tsRNAs have been named such as tRNA halves (tiRNAs) produced by the nuclease angiogenin[25]. These tsRNAs play a role in stress response by regulating apoptosis and cell proliferation during oxidative stress[26]. They are also implicated in various diseases such as viral infections, cancers and neurodegenerative diseases[27], making them interesting molecules for physiological and pathological studies.

1.1.5 small nucleolar RNA (snoRNA)

Finally, snoRNAs are eukaryotic RNA molecules in which a portion of their sequence carries evolutionarily conserved motifs that are complementary to specific RNA sequences (mainly rRNA) forming a double-stranded RNA molecule[28] SnoRNAs bind to proteins in order to form RNA-protein complexes also known as small nucleolar ribonucleoprotein (snoRNP). Two classes of snoRNAs have been identified: C/D box snoRNA and H/ACA box snoRNA and respectively, forms C/D box snoRNP and H/ACA box snoRNP when they are associated with their specific proteins[29]. Through the binding of the complementary sequence of the snoRNAs acting as a guide, C/D box snoRNP carries on the 2'-O-methylation of RNA molecules while H/ACA box snoRNP carries on the pseudouridylation, two post-transcriptional modifications.

Those modifications are important RNA regulators, from their structure maintenance thanks to RNA nucleotide chemistry alterations to chemical signalisation for RNA-proteins interactions affecting RNA and/or protein fate. Due to their crucial role for RNA functions, detailed insights about them allow better understanding in cells' functionality and will be further discussed in this manuscript.

1.2 Post-transcriptional modifications

The chemical structure of RNA classifies it as a nucleic acid. A nucleic acid is a polymer of multiple nucleotides, each of them composed of three specific molecular components: a nucleobase, a sugar backbone known as ribose and one phosphate group. A nucleotide without a phosphate group is called a nucleoside.

The four RNA canonical nucleobases - Adenine (A), Cytosine (C), Guanine (G) and Uracil (U) (see Figure 1) - are the unitary carriers of genetic information. The succession of these nucleobases generates coded genetic instructions for the cells.

Due to their information mediator function, most RNAs are single-stranded molecules to simplify their interactions with other molecules such as DNA and proteins.

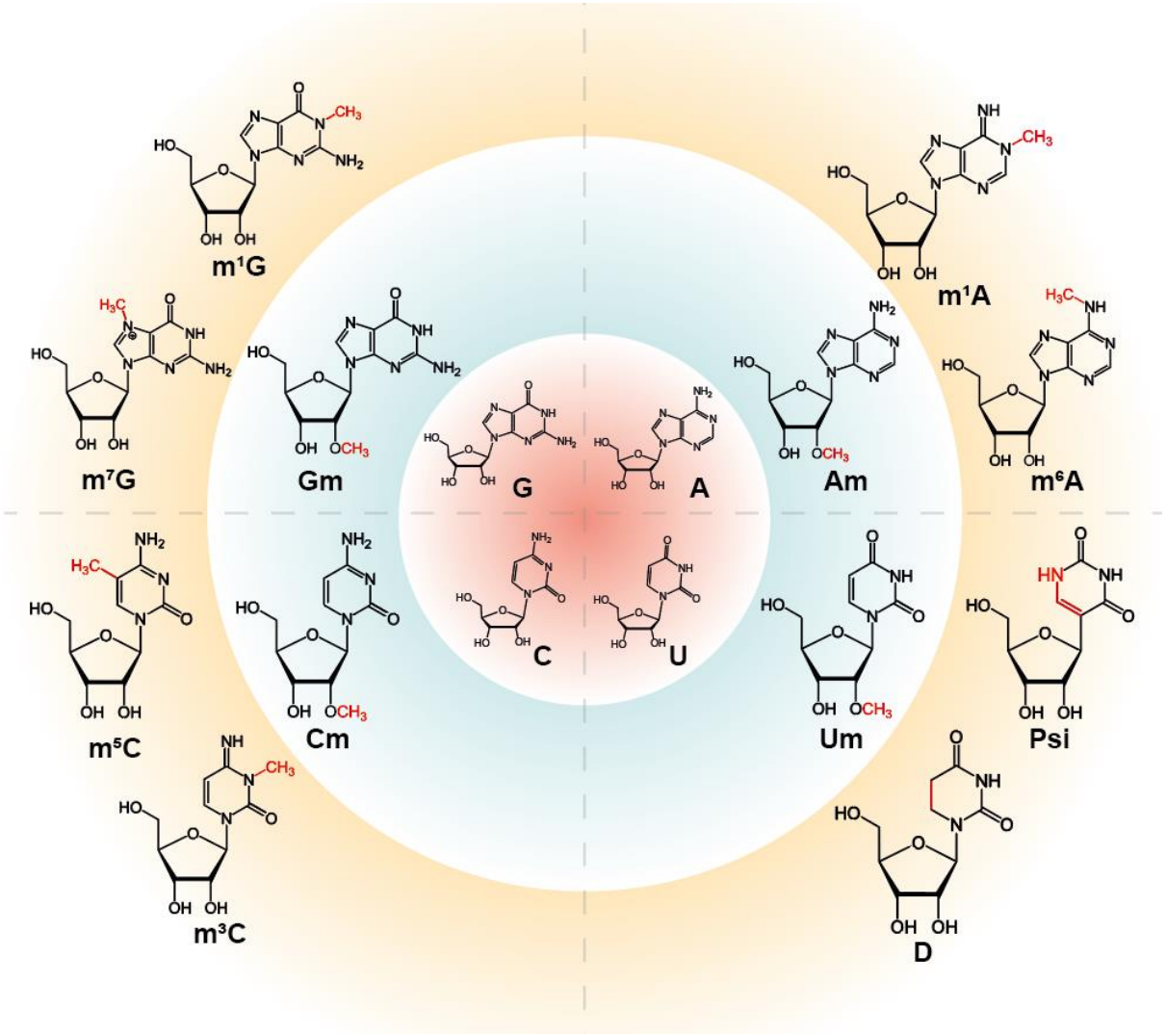


Figure 1: Chemical structure of abundant posttranscriptional modifications

Inner circle in red contains the chemical structure of the four canonical nucleotide Guanosine (G), Adenosine (A), Uridine (U), and Cytidine (C). Median circle in blue contains the four main 2'-O-methylation modifications chemical structure, namely 2'-O-methylguanosine (Gm), 2'-O-methyladenosine (Am), 2'-O-methyluridine (Um). Outer circle in orange contains chemical structure of recurrently seen RNA modifications which are N7-methylguanosine (m^7G), N1-methylguanosine (m^1G), N1-methyladenosine (m^1A), N6-methyladenosine (m^6A), pseudouridine (Psi), Dihydrouridine (D), N3-methylcytidine (m^3C) and 5-methylcytidine (m^5C). Chemical structures representations are open content available on the Internet (CC BY-SA 3.0 licence) which has been adapted to highlight the different modification's locations in red.

However, these different molecular components can be further altered by cellular enzymes in order to change the chemical properties of RNA. These changes are called post-transcriptional modifications and more than 170 modifications have been identified so far[30]. These modifications are essential for cells' homeostasis and their absence may be lethal for them depending on their functions. A few modifications are further discussed below as examples.

1.2.1 2'-O-methylation

2'-O-methylation consist of methyl groups added to the oxygen in the 2' position of the nucleotide's ribose (see Figure 1).
This modification helps the chemical stabilization of RNA and plays an important role in innate immunity, since 2'-O-methylated RNAs escape the immune response, protecting them from degradation. This biological function is of interest for viral infections and auto-immune diseases.
Further details will be given through the review below.

RNA ribose methylation (2'-O-methylation): occurrence, biosynthesis and biological functions. [DOI: 10.1016/j.bbagrm.2018.11.009]

Lilia AYADI[1], Adeline GALVANIN[1], Florian PICHOT[2], Virginie MARCHAND[2], Yuri MOTORIN[1*]

[1] UMR7365 IMoPA CNRS-Lorraine University, Biopôle, 9 avenue de la forêt de haye, 54505 Vandoeuvre-les-Nancy, France

[2] UMS2008 IBSLor CNRS-INSERM-Lorraine University, Biopôle, 9 avenue de la forêt de haye, 54505 Vandoeuvre-les-Nancy, France

Abstract

Methylation of riboses at 2'-OH group is one of the most common RNA modifications found in number of cellular RNAs from almost any species which belong to all three life domains. This modification was extensively studied for decades in rRNAs and tRNAs, but recent data revealed the presence of 2'-O-methyl groups also in low abundant RNAs, like mRNAs.

Ribose methylation is formed in RNA by two alternative enzymatic mechanisms: either by stand-alone protein enzymes or by complex assembly of proteins associated with snoRNA guides (sno(s)RNPs). In that case one catalytic subunit acts at various RNA sites, the specificity is provided by base pairing of the sno(s)RNA guide with the target RNA. In this review we compile available information on 2'-OH ribose methylation in different RNAs, enzymatic machineries involved in their biosynthesis and dynamics, as well as on the physiological functions of these modified residues.

Introduction and historical overview of early studies on RNA 2'-O-methylation

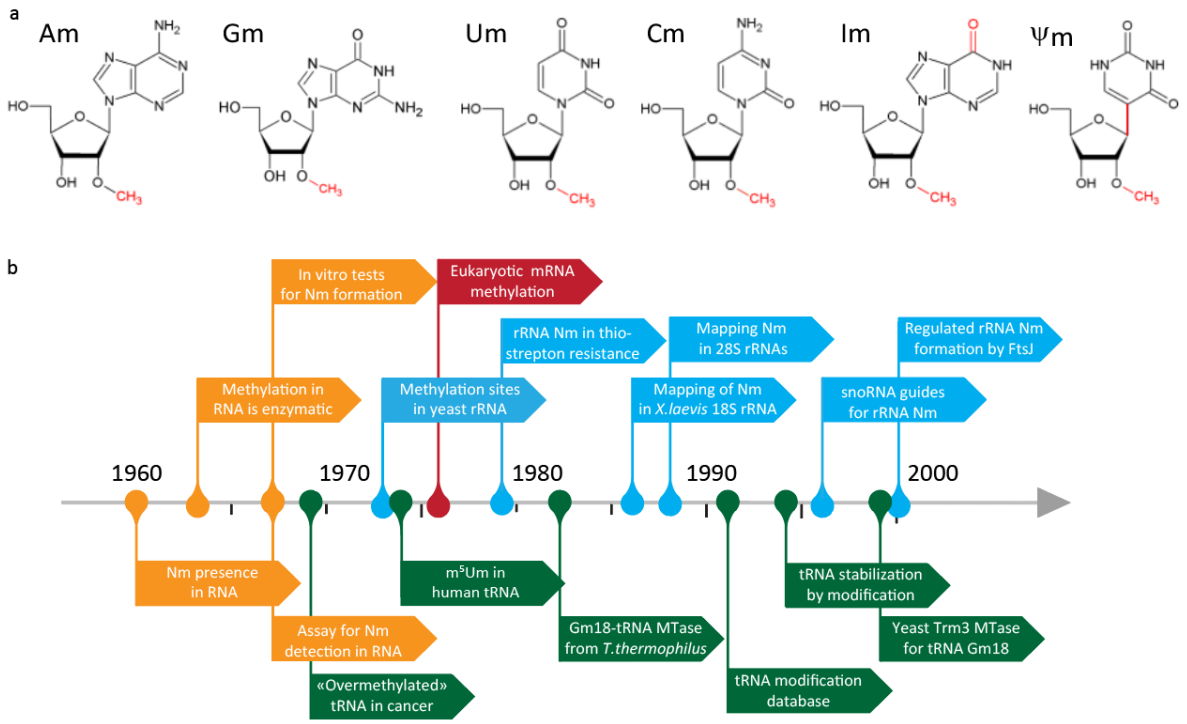


Figure 1 a) Chemical structures of common RNA 2'-O-methylated nucleotides (Am, Gm, Um, Cm, Im and ψm), b) hallmarks of studies on RNA 2'-O-methylation from 1960 to 2000.

Ribose 2'-O-methylation is one of the most common and widespread type of RNA modification commonly found in in ribosomal RNA (rRNA), transfer RNA (tRNA), small nucleolar RNA (snRNA) and, as reported recently, in messenger RNA (mRNA). This modification consists of the methylation of the ribose 2'-OH moiety (enzymatic transfer

of methyl CH_3-group from methyl donor to RNA substrate) and can therefore occur in all 4 nucleotides and other non-canonical nucleotides (Figure 1a). Despite their discovery already in the 60's, RNA 2'-O-methylations remained for a long time enigmatic, without any attributed cellular function. Only recently, complete distribution profiles of these residues have been reported, especially for low abundant RNAs and some physiological functions become clear. In this review, we provide a summary of the current knowledge on the presence and precise locations of 2'-O-methylations in different bacterial, archaeal and eukaryal RNAs. Enzymatic mechanisms involved in their biosynthesis and biological functions of 2'-O-methylation, as improvement of RNA stability and fine modulation of its conformation, as well as its importance in ribosomal translation and innate immunity are discussed. Methods for detection of those modified nucleotides are briefly discussed here since covered by other recent reviews[31–34].
Biological RNA methylation was extensively studied already in 60-70's (Figure 1b) and established the presence of methyl groups in tRNAs and rRNAs[35], but not in bacterial mRNAs[36]. Major efforts were concentrated on mapping of those modified residues in highly abundant RNAs, like rRNAs[37, 38] from different model organisms, as well as some isolated tRNAs[39]. This provided the first ideas on the distribution and the exact locations of 2'-O-methylations[40, 41], even if only a subset of sites has been precisely mapped at that time, due to technological limitations.

In parallel, biochemical approaches to study RNA modification were established [38, 42, 43] using undermethylated RNA from methionine-starved cells for *in vitro* tests for RNA modification[44]. The major impediment for detailed biochemical characterization was the lack of appropriate unmodified RNA substrates. Biochemical characterization with "undermodified" or inappropriately modified heterologous RNAs as substrates was thus extensively used[45, 46].

Mid and late 70's brought even more advances in the field, by measuring of exact stoichiometry of modified nucleotides (including 2'-O-methylations) in eukaryotic rRNAs[47, 48]. 2'-O-methylation profile of tRNAs from both selected bacterial species and eukaryotes was more precisely established[49–51]. These studies also confirmed earlier observations[52] on the existence of hypo(or hyper) methylated isoforms of tRNAs[53]. In addition to stable and abundant RNAs, 2'-O-methylations were found in U2 snRNA[54] and in various preparations of eukaryotic mRNAs[55–58].

Functions of these modified residues in various RNAs still remained largely unknown, but some influence on RNA conformation was already noticed[59, 60], and pioneering studies on antibiotic resistance demonstrated importance of ribose methylation for thiostrepton resistance in *Streptomyces azureus*[61, 62].

Mapping of 2'-O-methylated residues by next-generation sequencing

RiboMethSeq, 2'OMe-seq, Nm-Seq

In 2015, the group of H. Nielsen published a high-throughput approach for mapping of ribose methylation on an Ion Torrent sequencing platform[63]. One year later, we

adapted the protocol to an Illumina sequencing platform and reduced significantly the required RNA amounts (1,000x less), thereby allowing first analyses of clinical samples[64–66]. Both protocols are based on a reduced cleavage of the RNA phosphodiester bond that is 3' adjacent to a 2'-O-methylation in RNA. A fragmentation of RNA in alkaline conditions produces a collection of RNA fragments starting or ending at all RNA positions except at position +1 (5'-end) or at the 2'-O-methylated residue (3'-end). Therefore, after fragments end-repair, library preparation and sequencing, 5'- and 3'- end coverage is calculated and transformed into scores defining the level of methylation at a single nucleotide resolution. Another method called 2'OMe-seq based on RT-arrest at 2'-O-methylated residues at low dNTP concentration was developed allowing mapping of previously unknown positions in mouse rRNA[67]. Finally, 2 other groups independently developed a new strategy to map Nm in low abundant RNAs. These methods (RibOxi-Seq and Nm-Seq) rely on the differential reactivity of 2'-O-methylated and 2'-hydroxylated nucleosides to periodate oxidative cleavage, and on a selective enrichment of 2'-O-methylated sites[68, 69].

Presence of 2'-O-methylated residues in cellular RNAs

tRNA 2'-O-methylation

Transfer RNA (tRNA) molecules contain multiple 2'-O-methylated residues[30], but the number of 2'-O-methylated sites has evolved from a simple to a more complex pattern from bacteria to human.

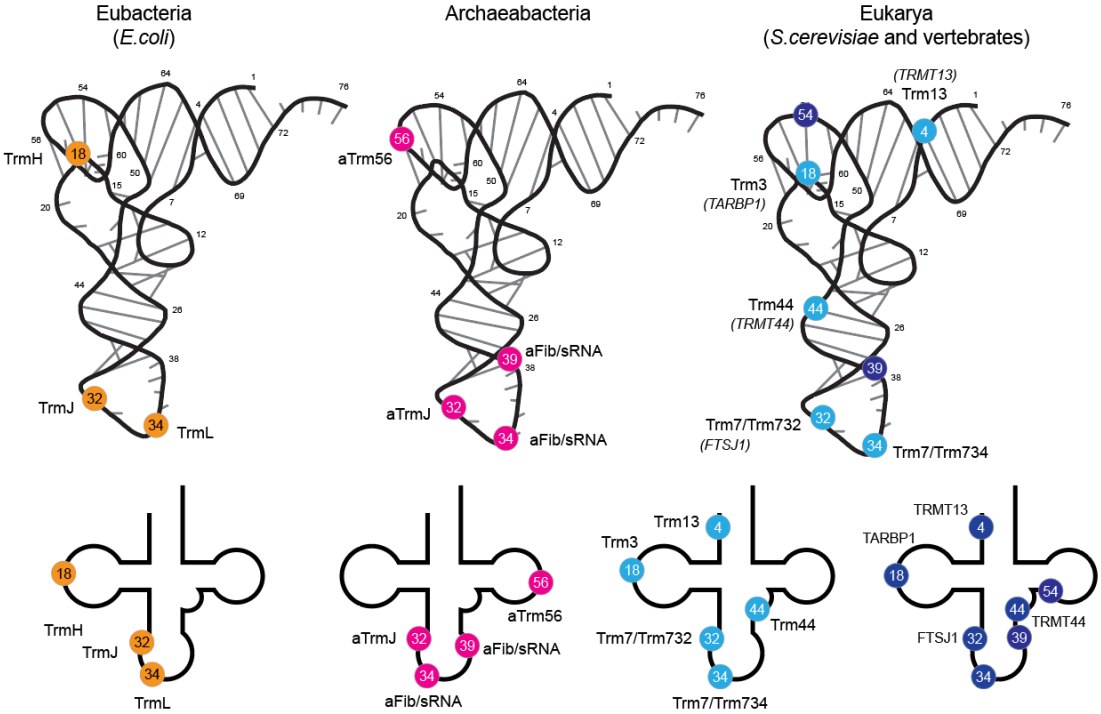


Figure 2 Top. 2'-O-methylated positions in tRNAs from *Eubacteria* (*E. coli*), *Archaea* (*H. volcanii*) and *Eukarya* (*S. cerevisiae* and vertebrates). *S. cerevisiae* sites are shown in light blue, additional human sites in dark blue. Bottom part represents tRNA cloverleaf structure with the indicated modified positions.

Bacterial tRNA 2'-O-methylation

Three different 2'-O-methylated positions have been discovered in tRNA from the model organism *E. coli* (Figure 2, Supplementary Table 1)[30]. The major site is located in the D-loop at position 18, where the highly conserved G residue is converted to a 2'-O-methylated guanine (Gm) in many tRNA species. The two other 2'-O-methylated residues located in the anticodon loop at positions 32 and 34 (tRNA wobble position) correspond to Cm and Um/cmnm^{5}Um residues and are present in a subset of tRNAs.

All studied bacterial species seem to contain a 2'-O-methylation at Cm34 or Um34, whereas Gm18 is present only in Gram-negative bacteria, with the exception of the Gram-positive *Geobacillus sp*.[70]. Modified Cm32 is present in all Gram-negative species, and is also probably conserved in Gram-positive bacteria, but data are sparse [50]. Except for *E. coli* and more generally Gram-negative species, bacterial 2'-O-methylation in tRNA is rather poorly characterized and data for the majority of pathogenic bacteria are still lacking.

Archaeal tRNA 2'-O-methylation

In archaea, although tRNA 2'-O-methylation is poorly studied, four positions were found to be modified: the universally conserved positions 32 and 34, as well as two additional positions 39 and 56. The latter is unique to archaea[71]. Compared to eubacteria, archaea lack Gm18[72, 73].

Eukaryotic tRNA 2'-O-methylation

In yeast *S. cerevisiae*, in addition to the conserved Gm18, Cm32 and Nm34 residues[74], 2 other positions are 2'-O-methylated: Nm4 and Um44. Um44 is frequent in cytoplasmic tRNA containing a long variable loop such as tRNA^{Ser} [75]. Concerning Nm4, this modification is the only one known to be in the middle of a duplex region in tRNA[76]. Methylations Cm32 and Nm34 in the anticodon loop are prerequisite for m^{1}G synthesis at position 37, which is an intermediate step in the biosynthesis of yW37 (wybutosine) [77, 78].

In vertebrates, the 5 positions described above for other species are similar with additional complexity: Cm/Um4; Gm18; Cm/Um/Ψm32; Cm/Gm34 and Um44. Moreover, two other sites are characterized for vertebrate tRNA 2'-O-methylation. The first one is Tm(m^{5}Um)54 in the TΨ-loop that sometimes replaces the universally conserved T54 modified residue[79]. The second one is Ψm39[30, 80]. In higher eukaryotes, 2'-O-methylations are often part of a more complex modified residue, where base modification is coupled with ribose methylation (like f^{5}Cm[81], ncm^{5}Um[82] and m^{1}Am[83]).

rRNA 2'-O-methylation

2'-O-ribose methylation is one of the major modifications (with pseudouridine) found in rRNA[84, 85] and this through all three life domains. The number of 2'-O-methylations in rRNA increases in the most complex organisms being up to ten-fold higher in human

in comparison to bacteria, however, despite the differences in the machinery involved in their biosynthesis, many positions are evolutionarily conserved between bacteria and archaea/eukaryotes. Most of the time 2'-O-methylations occur in functionally important domains of the two ribosomal subunits, essentially in proximity of the Peptidyl Transferase Center (PTC) and of the Decoding Center (DC)[85, 86].

Bacterial rRNA 2'-O-methylation

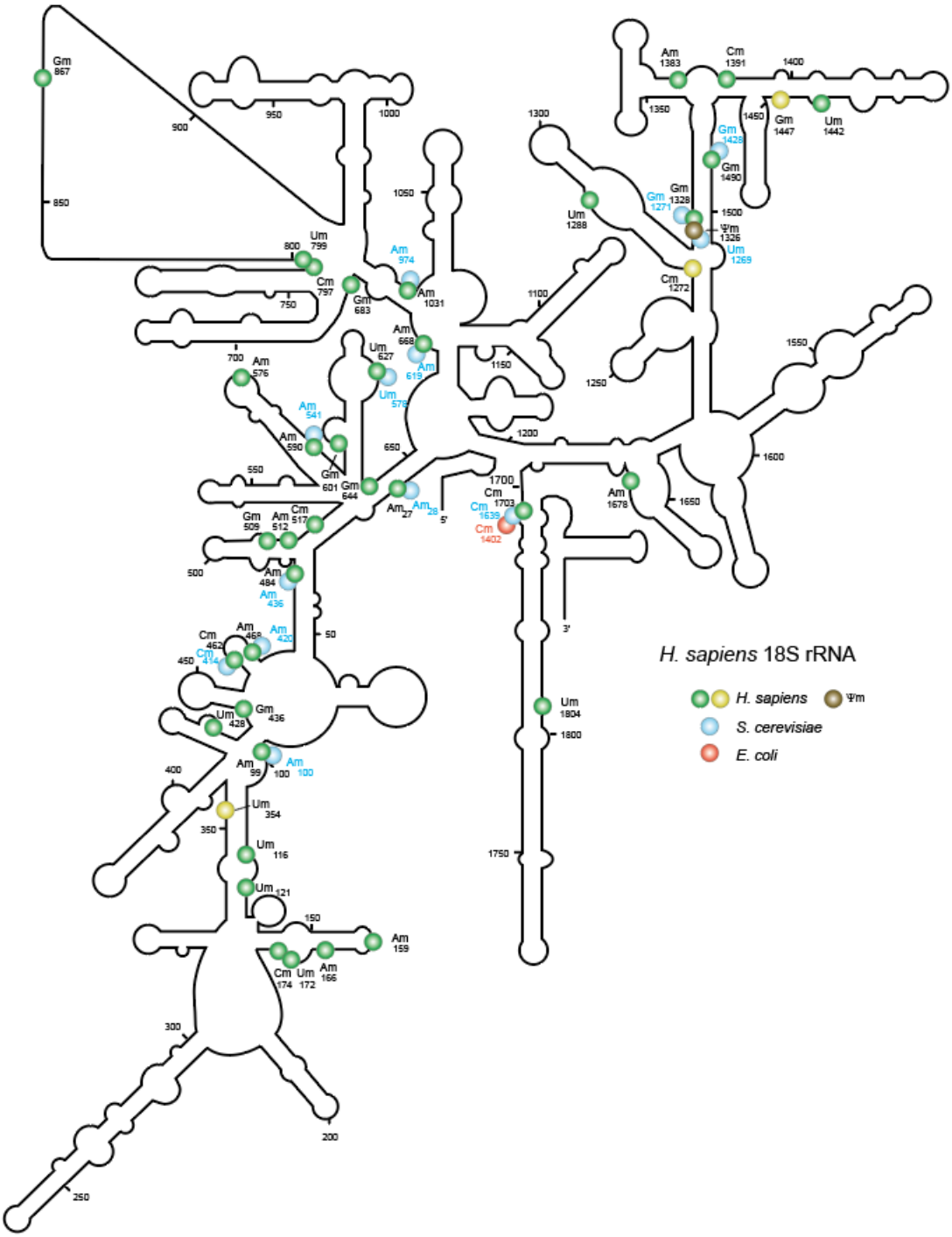


Figure 3a) Human 18S rRNA 2’-O-methylation. The 2D structure of human 18S rRNA is represented by wire. 2’-O-methylated residues and their numbering are shown. Previously annotated human 2’-O-methylated sites (https://people.biochem.umass.edu/fournierlab/3dmodmap/main.php) are in green, newly identified positions are in yellow. 2’-O-methylated pseudouridine at position 1326 is in brown color. Conservation with *S.cerevisiae* (in blue) and *E.coli* (in red) 2’-O-methylation is shown.

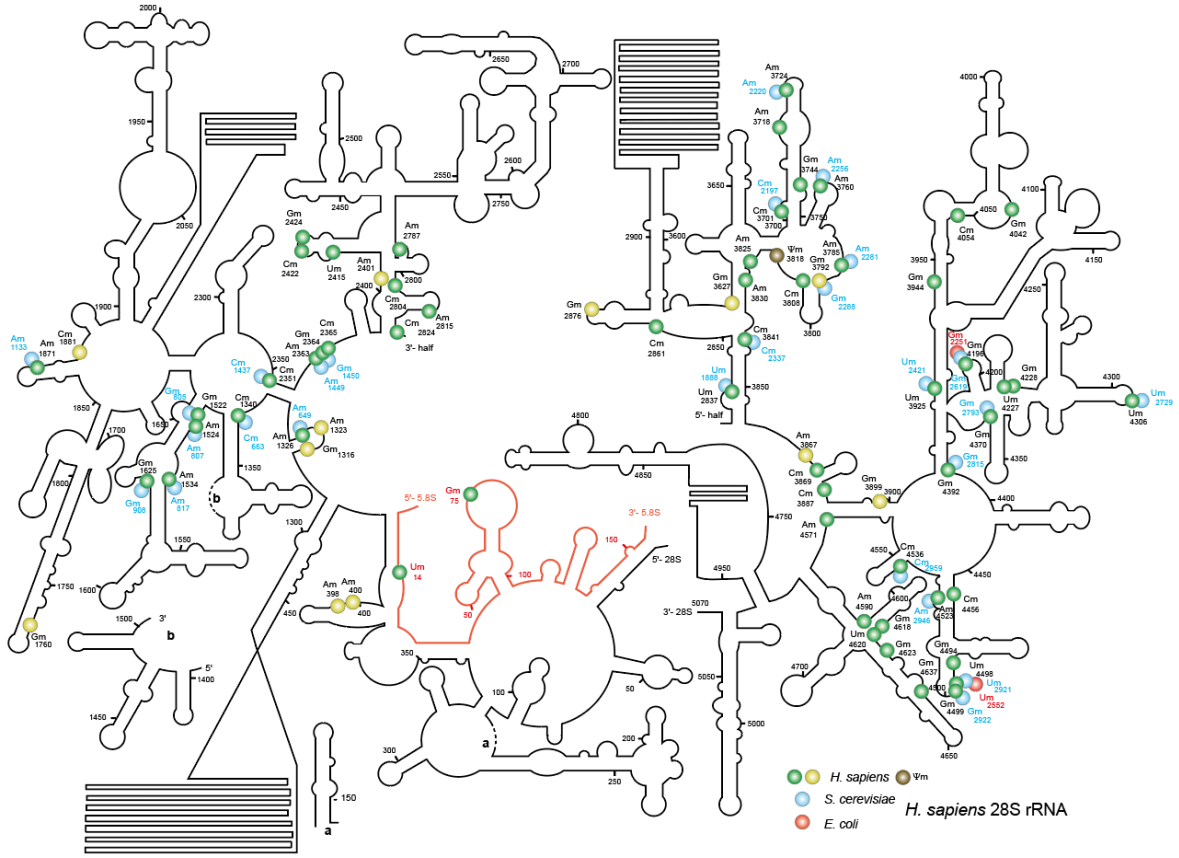


Figure 3b) Human 28S and 5.8S rRNA 2’-O-methylations. The 2D structure of human 28S rRNA is represented by wire, 5.8S rRNA is in red. Previously annotated human 2’-O-methylated sites (https://people.biochem.umass.edu/fournierlab/3dmodmap/main.php) are in green, newly identified positions are in yellow. Numbering corresponds to “new” reference NR_046235. 2’-O-methylated pseudouridine at position 3818 is in brown color. Conservation with *S.cerevisiae* (in blue) and *E.coli* (in red) 2’-O-methylation sites is shown.

In contrast to eukaryotes, where rRNA 2'-O-methylations are quite numerous, such residues are relatively rare in bacterial rRNAs.

Ribose methylation was essentially found at four highly conserved nucleotides: C1402 of the SSU-16S rRNA and G2251, C2498 and U2552 of the LSU-23S rRNA (for review[87]). Almost all bacteria have a modified 16S-Cm1402, which participates in P-site formation and seems to improve the precision of initiation codon selection[88]. In addition, crystal structure of the *E. coli* 30S subunit complex revealed that Cm1402 is stabilized in a C3'-endo conformation to promote its interaction with the mRNA.

All three 23S rRNA 2'-O-methylated residues Gm2251, Cm2498, Um2552 are located in the PTC. While Gm2251 and Cm2498 appear to be dispensable for cell viability[89], Um2552 is essential and is 2'-O-methylated in the majority of living organisms from bacteria to human, suggesting its important function.

- Gm2251 stays in close contact with the CCA-end of the P-site bound tRNA[90]. Bacteria lacking this modification do not show any phenotype under normal growth conditions[91]. However in yeast, the same modification present in mitochondrial rRNA (Supplementary Table 2) is required for cell viability[92].

- Um2552 is one of the conserved nucleotides in the ribosomal A site and interacts with aminoacyl-tRNA. Binding of the CCA-end of tRNA to the A-site of the ribosome involves a Watson-Crick base pairing between G2553 of the 23S rRNA and C75 of tRNA[93] and methylation of the neighboring U2552 may influence the binding of aminoacyl tRNA to the A-site[94]. Since A- and P-loops of the 23S rRNA contain Gm2251 and Um2552, respectively, one can propose that they may perform similar functions regarding the binding of tRNA to A- and P-sites[87]. Although the 2'-O-methylation of U2552 is not necessary for the function of *in vitro* reconstituted ribosomes, the substitution of uridine nucleotide or deletion of the corresponding methyltransferase affect growth and ribosome assembly[95–97].

In general, 2'-O-methylated residues found in *E. coli* rRNA are conserved, since in *Thermus thermophilus*, a relatively distant bacterium, a single ribose-methylated residue, m^4Cm1402 was observed in 16S rRNA[98], while *Mycobacteria smegmatis* SN2 rRNAs seem to contain one additional site of 2'-O-methylation[99].

Archaeal rRNA 2'-O-methylations

A map of predicted archaeal rRNA 2'-O-methylations has been established on the basis of high-throughput analysis of highly expressed sRNAs from 7 different archaeal species[100]. It shows that 2'-O-methylations are clearly clustered in functional areas of rRNA, suggesting their contribution to the folding, structural stabilization, assembly and function of the ribosome. However, most of predicted 2'-O-methylation positions were not validated experimentally. Interestingly, hyperthermophilic archaea *N. equitans* and *I. hospitalis* contain additional 2'-O-methylation residues within 150 3'-end nucleotides of 16S rRNA[101].

Yeast *S. cerevisiae* rRNA 2'-O-methylations

To date a total of 55 2'-O-methylated sites have been found in yeast rRNA[63, 64, 102, 103]: 18 in the SSU-18S rRNA (8 Am, 3 Cm, 6 Gm and 1 Um), and 37 in the LSU-25S (12 Am, 7 Cm, 10 Gm and 8 Um). No 2'-O-methylation sites have been described for 5.8S and 5S rRNAs. Orthogonal methods have been used to validate and/or quantify some of the modified sites by deletion of genes encoding the corresponding C/D-box snoRNA guides[63, 64] or by nucleoside analysis[104]. A clear mapping of 2'-O-methylated riboses on both 2D and 3D structures of both yeast 18S and 25S rRNA is now available and showed that the majority of modified residues are located close to the functional area of the ribosome: the DC (equivalent of A-site of 40S) and the PTC (A and P site of 60S), as well as in the inter-subunit region of the ribosome[13]. Position 2347 of 25S rRNA is 2'-O-methylated either on a U or a partially pseudouridylated residue[104, 105].

Concerning yeast mitochondrial rRNA 2'-O-ribose methylation, two positions Gm2270 and Um2791, which are conserved in other species have been detected in the LSU-21S rRNA by primer extension experiments[106].

Higher eukaryotic rRNA 2'-O-methylation in human and mouse models

Recently, high-throughput and MS approaches have been applied to different cell lines to map and quantify 2'-O-methylations in human rRNAs[67, 68, 107–110]. In this review, we have compiled all these data to give an updated view of 2'-O-methylations sites in human rRNAs (Supplementary Table 2)

(i) As many as 46 potentially 2'-O-methylated positions were reported in human SSU-18S rRNA by different studies. Majority of them (41 sites) were recently confirmed by different high-throughput sequencing methods, by MS and/or by fibrillarin knockdown (Figure 3a and Supplementary Table 2)[68, 107–109, 111, 112]. Methylation was not confirmed for nucleotides G1536, U1602 and U1668[68, 107] initially reported by Maden[112]. U1668 was probably incorrectly mapped and the corresponding GCC-U/Ψm-GC sequence should be assigned to position 18S-1232[107], but validation is missing for this position. MS-mapping also reported Cm621[110], but this position was newer found in other studies.

(ii) Precise comparison of human LSU-28S rRNA modifications is complicated due to various reference sequences used in independent studies. Early studies used 28S rRNA entry GeneBank U13369 as a reference sequence, but recent high-throughput approaches were referenced to NR_046235, which has multiple insertions of nucleotides. Thus, both “new” and “old” numbering schemes are used in parallel in different studies.

Mapping of 2'-O-methylations in human 28S rRNA by Maden in 80's reported over 50 modification sites, however, the exact location of several nucleotides remained unclear[113]. Later, additional positions have been mapped by RT primer extension at low dNTP concentration[114]. Application of recent high-throughput mapping

technologies mostly confirmed previous results and allowed to extend this list to 67 modification sites[108, 110, 111, 113, 114] (Figure 3b, Supplementary Table 2). The signal for Am1323 was detected in 2'OMet-Seq and RibOxi-Seq experiments[68, 111] and its partial methylation was confirmed by MS analysis[110], however no RiboMeth-Seq signal was found at this position[107, 108].

Three previously reported sites were not retained in the final list:

– Um1760/1773 (old/new numbering), found to be modified at 70% by MS analysis[115], but never seen by other approaches.

– Cm2279/2292 (old/new numbering), reported by 2'OMet-Seq[111] and predicted to be methylated by SNORD48[116]. This position was not confirmed by other approaches, RibOxi-Seq, RiboMeth-Seq and MS[68, 107, 108, 115].

– Gm2435/2448 (old/new numbering), captured in interaction with SNORD2 in 4 different CLIP experiments in HEK293 cells[117], but having only low RiboMeth-Seq scores, indicating partial methylation, if any. Not confirmed by global MS analysis[115].

(iii) Two 2'-O-methylated sites (Um14 and Gm75) are present in human 5.8S rRNA. The first evidence for their presence was given already in 1975[118]. In the nucleolus of HeLa cells Um75 is fully methylated, whereas the partially methylated Um14 is further modified in the cytoplasm after incorporation of the 5.8S rRNA into mature ribosomes[119].

In conclusion, human rRNAs totalize altogether 110 2'-O-methylated sites (41 sites in SSU-18S rRNA, 67 sites in LSU-28S and 2 in 5.8S rRNA) (Supplementary Table 2).

Until now, very little information was available on mouse rRNA 2'-O-methylation. This was investigated only recently[111] using rRNAs from E14 mouse embryonic stem cells (ESCs) and globally both 2'-O-methylated positions and their methylation status are quite well conserved between human and mouse. This also correlates with a strong conservation of snoRNAs between the two species. The only exception found yet is Gm4166/4196 which was not found in mouse and SNORD31 which is required as guide for this methylation in human cells has no equivalent in mouse.

rRNA modification in vertebrate mitochondria

To date, only three positions were found to be 2'-O-methylated in hamster[120] and human[121, 122] mitochondrial rRNA. These positions (Gm1145, Um1369 and Gm1370) are localized in the peptidyl transferase center of LSU-16S mitochondrial rRNA. Whereas the two first are conserved in other species, the third one was only found in vertebrate LSU-16S mitochondrial rRNA (Supplementary Table 2).

rRNA 2'-O-methylation in other eukaryotic organisms

In addition to vertebrates, several studies performed analysis of 2'-O-methylation sites in other lower and higher eukaryotes, the most studied are frog *Xenopus laevis,* plant *Arabidopsis thaliana,* protist *Euglena gracilis* and worm *Caenorhabditis elegans*. Only

partial rRNA modification maps have been established for those species, but a general conservation of 2'-O-methylation sites with human and yeast was noticed[112, 113, 123–125].

snRNA/snoRNA 2'-O-methylation

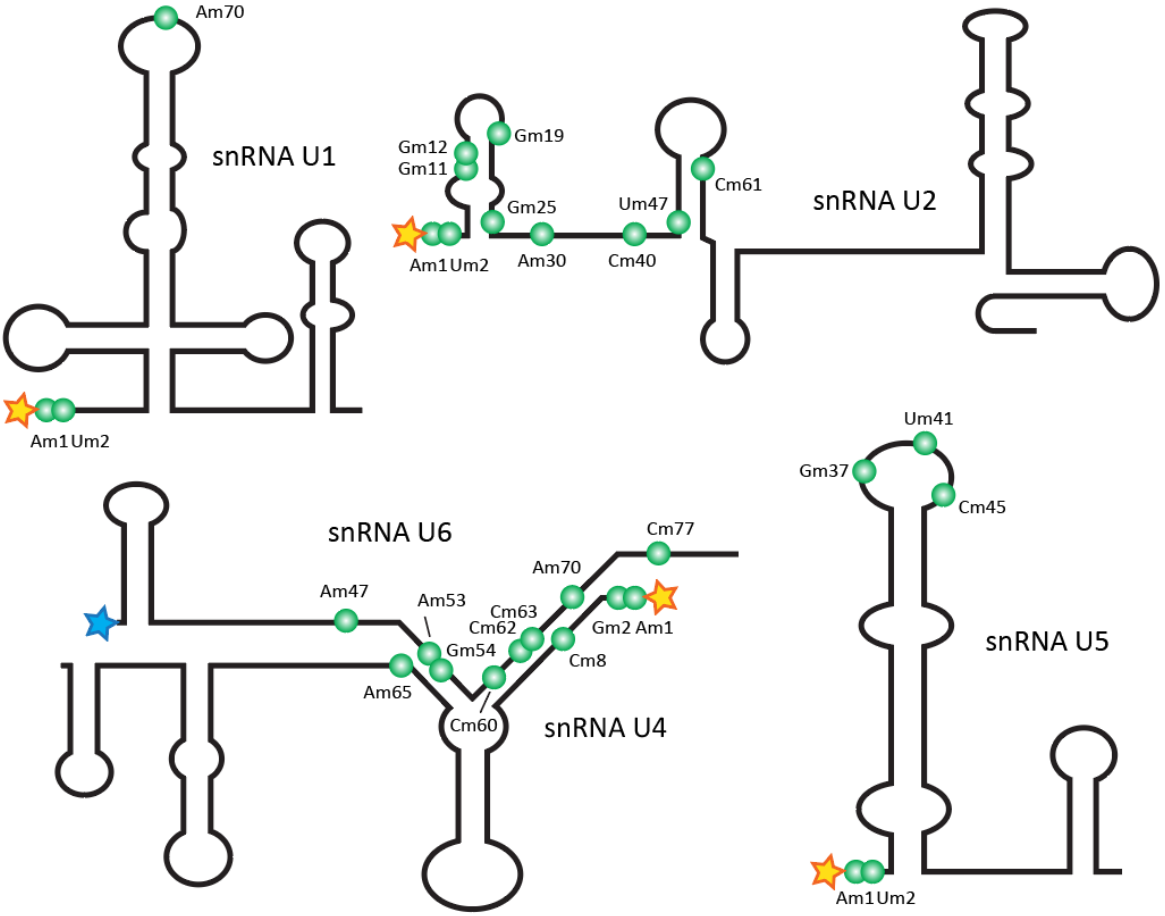


Figure 4 Reported human snRNA 2'-O-methylations. $m^{2,2,7}$G-Cap is shown as a star. γ-methyl cap of U6 snRNA is represented by a blue star. 2'-O-methylation sites are indicated by green circles.

Spliceosomal snRNAs harbor numerous 2'-O-methylated residues (Figure 4). All snRNAs (except U6 and U6atac) possess a trimethylguanosine ($m^{2,2,7}$G) cap structure followed by two 2'-O-methylated residues that are introduced by the enzymes CMTr1 and CMTr2 during cap formation[126, 127]. In addition, snRNAs contain internal 2'-O-methylated residues clustered in functionally important regions that participate in base-pairing with pre-mRNAs or with other snRNAs[54, 128]. In mammalian cells, internal snRNA modifications appear to be catalyzed by small Cajal body specific RNAs (scaRNAs) (for review[129]). Recently, using the RiboMethSeq approach, it was proposed that most of ribose methylations of snRNAs are almost fully modified and required for splicing activity[130]. However, the methylation pattern may vary depending on the conditions used and may contribute to splicing modulation[131].

So far, 2'-O-methylated residues have not been described as part of minor snRNAs,

scaRNAs or snoRNAs, but it will of interest to investigate these RNA species by high-throughput technologies.

2'-O-methylations in siRNA/piRNA/miRNA

In eukaryotes, small non-coding RNAs are central components of the RNA silencing pathway and play important roles in many biological processes. Their levels are tightly controlled *in vivo* through the balance of their biogenesis and turnover. Indeed, perturbations in their abundance are associated with many developmental and physiological defects. One of the major mechanisms controlling this balance is the 2'-O-methylation on the 2'-OH of the 3'-terminal ribose in small RNAs (for review[132, 133]).

2'-O-methylations in mRNAs

Ribose 2'-O-methylation in eukaryotic mRNAs was known from the 70's[55–57], precisely on the first and sometimes second cap-proximal nucleotides (N1 and N2, where N corresponds to any nucleotide), resulting in Cap1- (m7GpppNmN) or Cap2- (m7GpppNmNm) mRNA. N1 (and N2) methylation was shown to serve as a molecular signature of "self" to escape detection by the immune innate system (see below).

More recently, with the development of high-throughput Nm-seq technology (based on periodate oxidative cleavage), it was proposed that internal ribose 2'-O-methylation is also present in mRNAs[69]. Thousands of Nm sites were detected with slight enrichment in coding regions (CDS), and with Um as a dominant Nm modification (64% and 78% of all Nm sites in HeLa and HEK mRNA, respectively). Another manuscript deposited in BioRxiv also goes in the same direction and identifies thousands of Nm sites in yeast mRNA, suggesting that Nm sites may be prevalent in mRNAs. The presence of these numerous sites needs to be confirmed by orthogonal methods. This may be achieved also by the identification of the enzymes/snoRNAs responsible for these modifications. Some human mRNAs were captured in chimeras with snoRNAs[134], but do not show evidences of 2'-O-methylation, probably reflecting other roles of snoRNAs in mRNA function[135].

Enzymatic mechanisms of 2'-O-methylation

Protein stand-alone enzymes

Methyltransferases (MTases) specific to RNA belong to a vast group of MTases that catalyze the transfer of a CH3- group (Me-group) from a methyl donor to a biomolecule[136–138]. The almost universal methyl donor in all three kingdoms of life is S-adenosyl-L-methionine (SAM or AdoMet). There are 6 MTase types according to their targets: histones, protein N-terminal aminoacid, DNA, RNA, natural product and non SAM-dependent methyltransferases. This last group uses N^5,N^{10}-methylenetetrahydrofolate (CH_2H_4- folate) as a methyl donor. All MTases share the same core structure of a mixed seven stranded β-sheets. Five SAM-binding classes of MTases (I–V) differs by the presence of characteristic motifs and the global organization of the protein fold[138]. Although a lot of known MTases, such as DNA-MTases, are in the Class I, most 2'-O-MTases belong to the superfamily of SPOUT

MTases (class IV). In the past RNA 2'-O-MTases were separated into two groups: Rossman-like fold MTases (RMT) and SPOUT, but recently it has been demonstrated that they share a common evolutionary origin and they all form a single superfamily[136, 139–141].

Bacterial tRNA 2'-O-methyltransferases

Enzymes involved in tRNA 2'-O-methylation are rather well studied in *E. coli*. Each known position is formed by a specific tRNA MTase. TrmH (or SpoU) is responsible for Gm18 and was found in the *gmk-rpoZ-spoT-spoU-recG* operon[142, 143]. This enzyme is conserved in Gram negative species and in species living in harsh conditions such as *T. thermophilus*[144]. However, no TrmH homologs were reported for Gram positive species such as *Staphylococcus aureus*, *Lactococcus lactis* or *Bacillus subtilis*. Crystal structures of TrmH bound or not to SAM have been established and by extensive mutagenesis studies, catalytic site and enzyme reaction mechanism have been characterized. Four amino acids of the catalytic site (Asn35, Arg41, Glu124 and Asn152) are involved in tRNA binding. In addition, Asn35 is also responsible for SAM release[140, 142, 145].

TrmJ (or YfhQ or TrMet5Xm32) is involved in the formation of Cm32 in $tRNA^{Ser}_1$ and $tRNA^{Gln}_2$ [146]. This enzyme also belongs to the SPOUT class of MTases[147]. The 3D structure and catalytic site of *E. coli* TrmJ have been characterized. tRNA recognition by TrmJ involves not only SPOUT domain, but also additional parts of the protein[148]. Characterization of TrmJ from *Pseudomonas aeruginosa* revealed that, in addition to Cm32 and Um32, this MTase catalyzes also Am32 formation in $tRNA^{Pro}$(GGG)[149].

TrmL (yibK) is responsible for 2'-O-methylation at position 34 in two *E. coli* $tRNA^{Leu}$ (anticodons CAA and UAA). Cm34 is formed in $tRNA^{Leu}$(CAA) and Um34 is introduced in an hypermodified nucleotide in another isoacceptor $tRNA^{Leu}$($cmnm^5UmAA$)[150]. TrmL methylates only pyrimidine residues at position 34 and requires 2-methylthio-N^6-isopentenyadenosine (ms^2I^6A) at position 37 for activity[151]. Sequence homology search using TrmL as a query, revealed homologous proteins in *L. lactis* and *B. subtilis*, but not in *S. aureus*[152].

Archaeal tRNA 2'-O-methyltransferases

In contrast to bacteria, archaeal species use both stand-alone enzymes and C/D-box small RNA (sRNA) guided machinery for tRNA 2'-O-methylation, sometimes with overlapping substrate and site specificity[153]. The universally conserved Cm32 and the archaea-specific position Cm56 are catalyzed by stand-alone archaeal MTases belonging to the SPOUT family.

In *Sulfolobus acidocaldarius*, 2'-O-methylation at position 32 is catalyzed by a homolog of the bacterial TrmJ. However, while in *E. coli* TrmJ modifies tRNAs without any specificity for a given nucleotide, its archaeal homolog can act only at the ribose of a cytidine[147]. Enzyme aTrm56 methylates the ribose of a cytidine residue and is found in almost all sequenced archaeal genomes, except for the crenarchaeon *Pyrobaculum*

aerophilum in which Cm56-forming activity is provided by a C/D-box sRNP[154]. Similarly, Cm34 formation in *Haloferax volcanii* elongator $tRNA^{Met}$ is guided by a C/D-box sRNA targeting the tRNA intron-containing precursor[155] and Cm34 and Um39 in pre-$tRNA^{Trp}$ are formed by another C/D- box sRNP machinery[156].

Eukaryotic tRNA 2'-O-methyltransferases

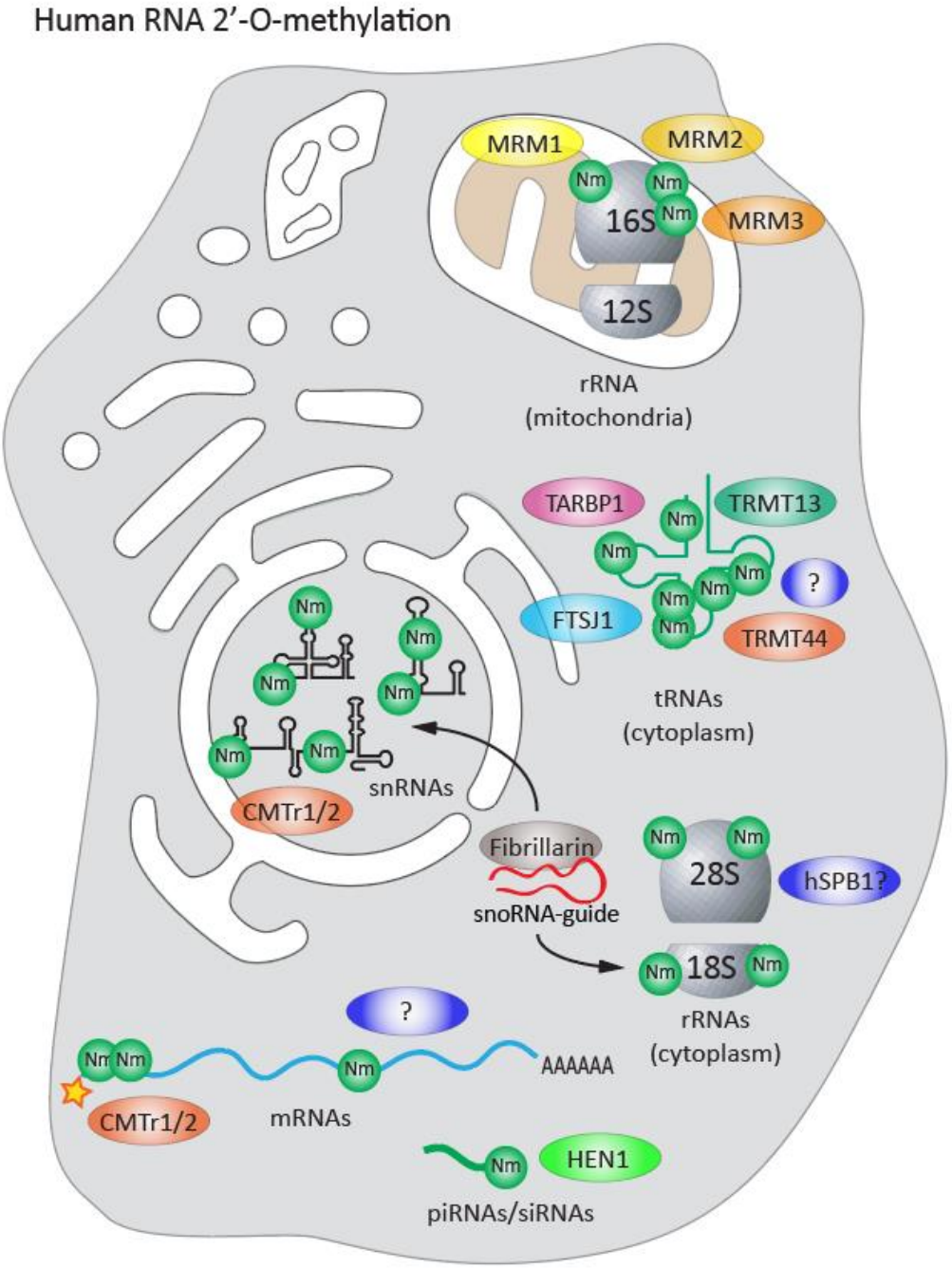


Figure 5 General view of 2'-O-methylation machinery in human cells. 2'-O-methylation sites are in green circles. Enzymes (if known) are shown.

The formation of the highly conserved Gm18 in *S. cerevisiae* is catalyzed by Trm3 (encoded by ORF YDL112w). Genetic depletion of Trm3 led to a complete absence of Gm18 in tRNAs, while other 2'-O-methylations in tRNA or rRNA were not affected[157]. Two *S. cerevisiae* 2'-O-methylations Cm32 and Nm34 present in tRNA anticodon loop share a common MTase Trm7[74, 158] (encoded by YBR061C), that requires a protein partner to specify which position is methylated[159]. When Trm7 is interacting with Trm732 (encoded by YMR259c) Cm32 is formed, while when Trm7 is associated to Trm734 (or Rtt10), nucleotide at position 34 is methylated[78]. However, Cm32 and Nm34 seem to be formed by different catalytic mechanisms since a specific mutation in Trm7 abolishes formation of only Gm34 in $tRNA^{Phe}$ [160]. Formation of Um44 is catalyzed by Trm44 (ORF YPL030w) which is found in metazoans and fungi, but not in plants[75]. Finally, position 4 in *S. cerevisiae* tRNAs is methylated by Trm13 (encoded by YOL125w)[76] which belongs to Rossmann-fold family of MTases[161] (Figure 5).

In Drosophila, tRNA 2'-O-methylation profile has not been deeply characterized. Only a putative tRNA:(Cm32/Gm34)-2'-O-methyltransferase has been identified by sequence homology, but no experimental validation was performed. As observed in human tRNAs, 2'-O-methyl-ribothymidine (Tm, m^5Um) was found to replace ribothymidine in drosophila $tRNA^{Lys}_2$ and RNA^{Lys}_5 species[51].

In human, positions of tRNA 2'-O-methylation are mostly known, but the corresponding enzymes were not systematically characterized (Figure 5). Sequence homology search using yeast MTases revealed homologous enzymes for all of them (Supplementary Table 1), but no activity test was performed for these proteins[162]. Only human protein homologous to yeast Trm7 2'-O-MTase was characterized and called FTSJ1. Indeed, human cells with mutations in FTJS1 nearly completely lack both Cm32 and Gm34[160].

Stand-alone rRNA 2'-O-methyltransferases

Ribose methylations in *E. coli* rRNAs (C1402 of the SSU-16S rRNA and G2251, C2498 and U2552 of the LSU-23S rRNA)[87] are formed by stand-alone MTases RsmI (YraL), RlmB (YfjH), RlmM (YgdE), RlmE (RrmJ or Ftsj), respectively (Supplementary Table 1)[89, 91]. While RlmB and RlmE belong to the SPOUT family, RlmM and RsmI have a Rossman-like fold structure[163]. Crystal structures of RlmM-SAM[164], RsmI-SAM-rRNA[165], RlmB[166] and RlmE[167] were determined and allow a better understanding recognition of their substrates and organization of catalytic domains.

In archaea and eukarya, the implication of stand-alone MTases in rRNA ribose methylation remains an exception. Indeed, this function is mainly insured by C/D-box s(sno)RNPs. Interestingly, a recent report described a newly discovered stand-alone activity of archaeal Fib–Nop5 heterodimer that can perform 2'-O-methylation of SSU-16S and LSU-23S rRNAs in *Pyrococcus abyssi* [168].

In yeast *S. cerevisiae*, Spb1 protein was shown to methylate both G2922 and U2921 residues in late 27S pre-rRNA precursor, the latter is also modified by a guide snR52-

snoRNP[169, 170]. Spb1 corresponds to the bacterial RrmJ (specific to U2552 in LSU-23S rRNA). The human homolog of yeast Spb1 is FTSJ3. It was demonstrated that FTSJ3 is involved in pre-rRNA processing *via* its interaction with NIP7: a trans-acting factor required for eukaryotic ribosome biogenesis[171]. A recent study deposited in BioRxiv claimed that Sbp1 has 2'-O-MTase activity and is responsible for methylation of thousands sites in yeast mRNA, suggesting that human FTSJ3 may have the same function.

Interestingly, the yeast enzyme Trm7 which is normally required for 2'-O-methylation at positions 32 and 34 in several tRNAs, has been recently reported to be directly or indirectly involved in methylation of five positions in *S. cerevisiae* 25S rRNAs: Am649, Cm663, Gm805, Am807 and Um2421[172]. These positions were found to be significantly hypo-methylated in a ΔTrm7 mutant. Furthermore, the most affected position, Cm663, was hypo-methylated in mutants lacking Rtt10, one of the two heterodimerization partners for Trm7. The closest homolog of eukaryotic Trm7 is a rRNA:MTase in bacteria[173].

In human mitochondria, three closely related 2′-O-ribose MTases (MRM1, MRM2/FTSJ2, and MRM3/RNMTL1) modify three positions of the peptidyl transferase center of LSU 16S mt-rRNA, Gm1145, Um1369 and Gm1370, respectively (Supplementary Table 2)[121, 122, 174]. MRM1 and MRM2 have bacterial and yeast homologs, whereas MRM3 appears to have evolved later in higher eukaryotes[122, 173]. Defective mitochondrial rRNA methyltransferase MRM2 causes MELAS-like clinical syndrome[175].

In yeast, PET56 is a ribose MTase that modifies the conserved nucleotide G2270 in the peptidyl transferase center of the mitochondrial LSU-21S rRNA. Loss of this modification leads to a defect in assembly of mitochondrial large ribosomal subunits[106]. Mrm2p, which is the ortholog of FtsJ/RrmJ in *E. coli,* catalyzes the formation of Um2791 in LSU-21S rRNA[173] (Supplementary Table 2).

siRNA/piRNA/miRNA 2'-O-methylation enzymes

In Arabidopsis, 2'-O-methylation at the 3'-end of RNA is performed by the RNA methyltransferase Hua Enhancer 1 (HEN1), which methylates both miRNA/miRNA and siRNA/siRNA duplexes with a preference for 21-24 nt RNA duplexes with a 2 nt overhangs[176, 177]. In contrast to plants, HEN1 homologs in animals act on single-stranded RNAs such as germline piRNAs and Ago2-associated siRNAs (but not Ago1-associated miRNA) as well as tRNA-derived sncRNAs[178] to deposit a 2'-O-methyl group on their 3' termini[178, 179]. Interestingly, the substrates of HEN1 are quite different between plants and other species and this was recently proposed to be due to a conserved FxPP motif that is essential for substrate recognition[178].

In different organisms, HEN1 mutant phenotypes showed reduction of small RNA and piRNA abundance and pleiotropic developmental defects (for more details, see review of [132]). 3'-end methylation has two major roles: 1) to protect small RNAs from 3'-

terminal uridylation and 2) to inhibit 3'->5' exonucleolytic degradation[180], thereby maintaining an equilibrium in the levels of small non-coding RNAs. Interestingly, HEN1 homolog is also present in bacteria[181]. Pnkp/Hen1 system adds a methyl group to the 2'-OH group on the original RNA cleavage site and thus prevents it from further cleavage (for review[182]).

snoRNA (sRNA) guided machinery for ribose methylation

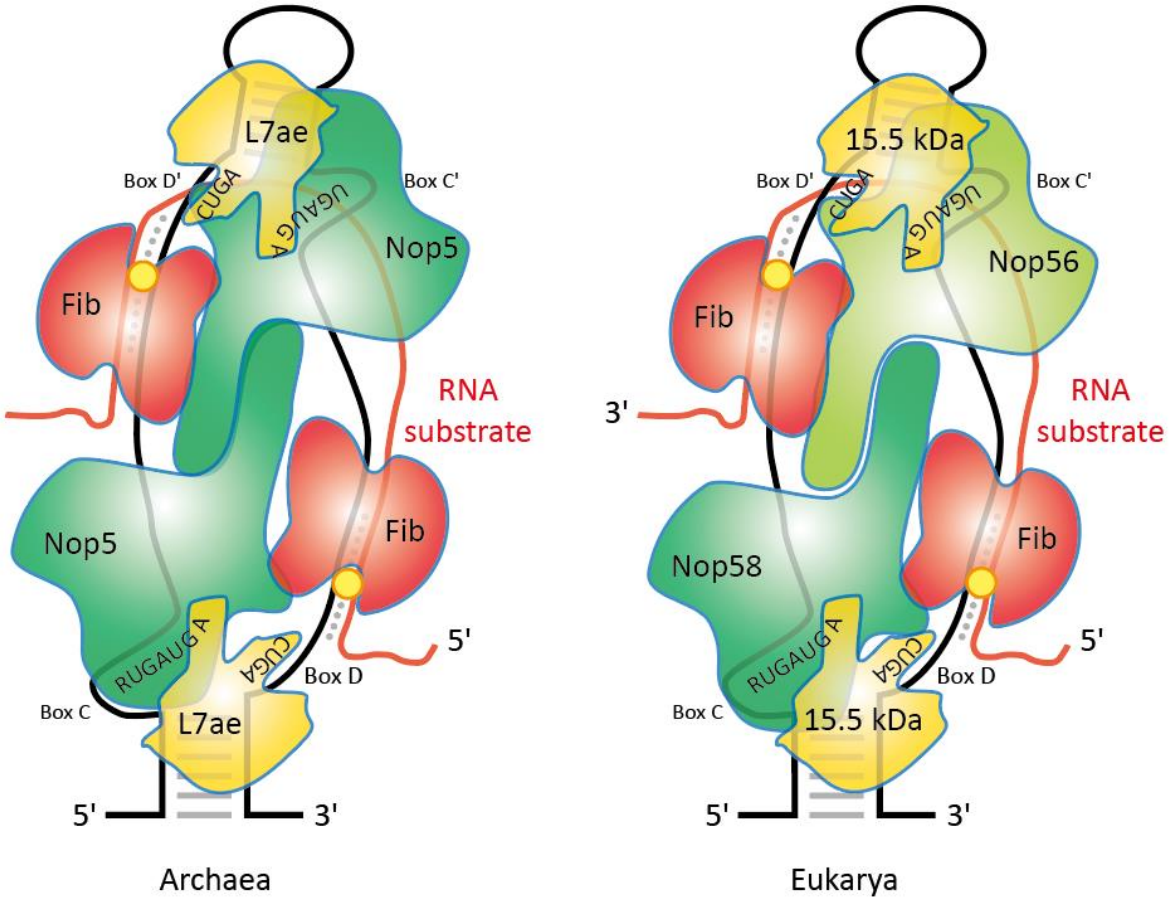


Figure 6 Structure of C/D-box sno(s)RNPs in different species.

In archaea and eukaryotes, 2'-O-methylations in rRNA are introduced by one class of structurally and functionally conserved small RNA-protein complexes (named C/D-box sRNPs in archaea) or small nucleolar RNA-protein complexes (named C/D-box snoRNP in eukaryotes)[183–186]. Each C/D-box s(no)RNP is formed by one specific small RNA (sRNA or snoRNA) which targets a specific nucleotide that will be modified and a set of three (in archaea) or four (in eukaryotes) common proteins including one SAM-dependent methyltransferase called fibrillarin (Nop1 in yeast). The other protein components of the s(no)RNP are Nop56, Nop58 (Nop5 in archaea is related to the two proteins) and 15.5K (Snu13p in yeast and L7Ae in Archaea) Figure 6. Many mechanistic studies were carried out in Archaea in order to understand the catalytic mechanism of sRNP-catalyzed ribose methylation. This was made possible because, in contrast to their eukaryotic homologs, the methylation C/D-box sRNP complex has

a capacity to assemble *in vitro*[187]. L7Ae binds first to the helix-internal loop–helix structure K-turn of the RNA guide[188]. Then the C-terminal domain of Nop5 binds to the L7Ae–RNA complex, while the N-terminal domain of Nop5 interacts with fibrillarin[189].

The length of C/D-box s(no)RNAs varies from approximately 50 to 300 nts, being shorter in Archaea (50-70 nts). The s(no)RNAs are characterized by two well-conserved sequences called C box (RUGAGA) and D box (CUGA) elements located respectively at the 5' and 3' extremities of the RNA and less conserved internal ones called C' and D'. These motifs fold into a kink-turn structure (C/D-box) and a kink-loop structure (C'/D'-box) allowing binding of the core proteins and the positioning of fibrillarin for catalysis[190]. By sequence complementarity between the guide sequence and the RNA target, snoRNA guide places the catalytic subunit for methylation. The guide sequence is longer in eukaryotes than in archaea, 10-21 nts and 8-12 nts in length, respectively. It is possible that base pairing rules for C/D-box sno(s)RNAs may be modulated under stress conditions, allowing modification at additional sites in RNA.

Unlike eukaryotes, archaeal C/D-box sRNP particles combine two methylation events at the same time on two distinct substrate RNA positions[191]. This particularity has no known explanation yet, but this arrangement is probably allowing coordination of methylation states of the two sites[192].

Archaea C/D-box sRNAs are highly abundant and ensure folding and stabilization of RNA molecules through modification of numerous nucleotides. In order to identify 2'-O-methylation targets of C/D-box sRNAs in archaea, 18 different C/D box sRNAs were first captured by co-immunoprecipitation with fibrillarin and Nop5 from *Sulfolobus acidocaldarius*[187]. Thereafter a global profiling of small RNAs from archaeal model organisms has been obtained using global RNA-Seq technology and numerous C/D box sRNA sequences were experimentally verified[193, 194].

In yeast, except for the residue 25S-Gm2922 which is methylated both by the stand-alone methyltransferase Sbp1 and a snoRNP, all other 2'-O-methylations are essentially carried out by 47 C/D-box snoRNPs each directing the methylation of one or two positions[185]. Two snoRNAs, U24 and snR40, have been found to target 3 positions each[103]. Conversely, one rRNA position can be methylated by more than one snoRNP.

To generate an updated catalog of human C/D-box snoRNAs that direct 2′-O-ribose methylation in human rRNA, a high-throughput variant of the reverse-transcriptase-based method was developed (termed RimSeq)[195].

In conclusion, 2'-O-methylation in tRNA is mostly done by protein stand-alone enzymes, except some specific archaeal species which are modified using C/D-box sRNA guides. In contrast, rRNA 2'-O-methylation is insured by stand-alone enzymes only in bacteria and eukaryotic organelles, while the same type of modification is formed in archaeal and eukaryotic rRNA by snoRNA-guided machinery.

Dynamics of RNA 2'-O-methylation

Dynamics of RNA methylation as well as incomplete methylation levels in different RNA species, depending on cell origin or tissue, were long standing observations already from early studies in 60's to 80's[52, 53, 196]. Various differentially modified tRNA species were isolated from different sources, but their relevance to any biological function was not established. Only very recently this idea came back to limelight and several studies clearly established that modification profile of tRNAs[197–199] and also rRNAs[109, 200] can be modulated during stress, or depending on the physiological status of the cell.

These observations experimentally demonstrate fine tuning of the translational machinery in response to stress and even precise modulation of specific mRNA translation in function of the RNA modification profile. 2'-O-Methylations, together with other modified nucleotides, present in tRNAs and rRNAs play an important role in such adaptation mechanisms.

Novel RNP complexes termed regulatory RNPs (regRNPs) have been described to provide a mechanism by which ribosome heterogeneity is accomplished[201]. Indeed, three C/D-box scaRNA (2, 9 and 17) which naturally guide the 2'-O-methylations of specific sites on U2, U4 and U12 snRNAs, respectively, give rise to stable, nucleolus enriched RNA fragments[202] that can impact the activity of snoRNPs by base pairing with several different snoRNAs. At least three 2'-O-methylation sites in human rRNA (28S-Gm3923, 18S-Um1804 and 18S-Am484) were shown to be differentially modified in endogenous ribosomes with a variability >25%[107, 111, 203] and are predicted to be regulated *via* interaction of regRNP 2, 9 and 17 with HBII82B, snoRD20 and snoRD16, respectively. It is the case of regRNP17 that impacts the level of 18S-Am484 in rRNA *via* interaction with snoRD16, by disrupting snoRD16/18S rRNA interaction[201].

Biological functions of 2'-O-methylated residues

The exact biological roles and functions of ribose methylation in RNA are still to a large extent unknown. It was demonstrated that Nm nucleotides are involved in RNA stability, and also play a role in more subtle functions such as translational modulation and are implicated in immune evasion mechanisms.

Physical properties of 2'-O-methylation

The presence of methyl group at the ribose 2'-OH considerably affects the physico-chemical properties of the modified nucleoside(tide). Formally, base-pairing of 2'-O-Me nucleotides is not directly affected, even if stabilization of the typical A-type RNA helix is observed[60, 204]. In contrast, the hydrogen bonding, typical for unmodified 2'-OH of the ribose, is abolished by methylation, as well as the nucleophilic character of the O2'. This leads to highly increased stability of 2'-O-methylated RNA against alkaline (and even enzymatic) hydrolysis[40, 205]. In addition, the absence of the proton at 2'-OH negatively affects protein binding, as was demonstrated for components of

mRNA polyadenylation machinery[206] and may also be the case for exon-junction complex deposition[207, 208]. Ribose methylation also affects the RNA-DNA duplex formation, this property is used to limit off-target effects of siRNAs (see for example[209]).

RNA stabilization

Post-transcriptional modifications of RNA have been shown to exert noticeable effects on stability of hyperthermophilic archaeal tRNAs[210]. The association of both base modification and 2'-O-methylation at the same nucleotide, which is the case of nucleosides ac^4Cm and $m^2{}_2Gm$, unique to the archaeal hyperthermophiles, seems to have additive effects to confer conformational stability essential for tRNA function[204]. This stabilization is particularly important for hyperthermophilic archaea to counterbalance RNA unfolding at elevated growth temperatures. Interestingly, tRNAs from mesophilic archaea, which grow at moderate temperatures, generally have only one of these ribose and base modifications. A relationship between the *B. stearothermophilus* growth temperature and *in vivo* tRNA methylation profile was established. It was shown that the nature of the modified nucleotides in tRNA remains similar between 50°C and 70°C, but the rate of global methylation increases 1.4 times essentially due to 2'-O-methylation[211]. As at 70°C the global tRNA structure is in an open form, RNA might be more accessible to respective MTases, that would serve as a kind of protection to prevent tRNA degradation by nucleases.

Ribosomal RNA from bacterial and archaeal hyperthermophiles shows the same tendency of an increased level of 2'-O-ribose methylation proportional to the growth temperature. This suggests a potential role of these rRNA modifications in RNA stabilization under extreme conditions. The first post-transcriptional modification map for the 16S rRNA including identity and sequence location of modified residues was described for the bacterial thermophile, *T. thermophilus* grown at 75°C[98]. Similarly, archaea that are growing at higher temperatures contain more guide C/D-box sRNAs and, in consequence, more 2'-O-methylations[100, 212, 213].
Interestingly, a protective effect of 2'-O-methylation against age-related neurodegenerative events has been observed in Drosophila. With age, miRNAs containing 2'-O-methylation at their 3'-end are preferentially loaded by Ago2 and thus are protected from degradation[214].

Resistance determinant for antibiotics

Antibiotics targeting the ribosome principally interact with the central loop of the 23S rRNA domain V in the 50S subunit, this interaction has been shown to directly inhibit peptidyl transferase activity and, in consequence, to reduce protein synthesis[215]. To overcome this effect, bacteria have developed various strategies counteracting the effects of antibiotics. One of the strategies used by bacteria is to mutate their rRNA, or, as an alternative, to introduce a methyl group directly at the nucleotides implicated in interaction with the drug/antibiotic[216]. Indeed, it is expected that chemical changes of the RNA constituting the drug binding site can directly influence the strength of

binding to the ribosomes[217, 218]. The first evidence of the involvement of rRNA ribose methylation in antibiotic resistance was brought already in 1979 by the study of thiostrepton producing soil bacteria *S. azureus*[62]. Thiostrepton is an antibiotic that inhibits bacterial protein synthesis by binding to the 50S subunit. To prevent its death *S. azureus* expresses a specific rRNA:2'-O-MTase able to methylate adenosine A1067 (*E. coli* ribosome numbering) in the 23S rRNA, presumably during 50S subunit assembly, and thus prevents thiostrepton binding[62, 219].

More recently, by biochemical *in vitro* experiments and mass spectrometry, it was demonstrated that the rRNA:MTase AviRb which provides avilomycin resistance to the producer strain *Streptomyces viridochromogenes* Tü57, 2'-O-methylates position U2479 in the 23S rRNA[220].

Another example is given by Cfr (chloramphenicol-florfenicol resistance) protein which is a rRNA:MTase homologous to a protein superfamily called Radical SAM[221] that targets nucleotide LSU-23S A2503 likely on the ribose and inhibits ribose methylation at nucleotide C2498, thereby causing resistance to chloramphenicol, florfenicol and clindamycin[222]. Nucleotide A2503 is already known to be methylated to m^2A, but surprisingly, primer extension experiments showed that the stop at m^2A2503 is apparently also dNTP concentration-dependent, indicating 2'-O-ribose methylation.

Global screening of rRNA 2'-O-methylation functions in *S. cerevisiae* with deletion of 20 individual C/D-box snoRNA genes revealed that the loss of individual methylated sites affects adaptation to stress conditions, as well as to intraribosomal antibiotics binding[223]. For the moment, underlying molecular mechanisms affecting stress adaptation remain unclear, but one can imagine enhanced selective translation of stress-related mRNAs.

RNA repair to counteract ribotoxins

HEN1 is adding a 2'-O-methyl group at the 3'-end of small RNA in plants and animals. Interestingly, HEN1 homolog was also found in bacteria where it associates to a protein phosphatase called Pnkp. Bacterial complex Pnkp/Hen1 is able to repair tRNAs cleaved by ribotoxins[181]. The bacterial Pnkp/Hen1 system in addition to other RNA repair system is able to add an extra methyl group to the 2'-OH group on the original RNA cleavage site, therefore preventing the repaired RNA from new cleavage at the site of repair (for review[182]).

Functions of 2'-O-methylation in mRNA splicing and translation

Eukaryotic splicing machinery is sensitive to the methylation status of both snRNAs and mRNAs. Injection of unmodified U2 snRNA into U2-depleted Xenopus oocytes does not rescue splicing, indicating that U2 snRNA modifications, including 10 2'-O-methylations are essential[130]. Later, the region required for U2 snRNP assembly was narrowed down to 20 first 5'-terminal nucleotides containing three pseudouridines and 5 2'-O-methylations[224]. However, the exact nature of essential residue(s) remains unclear. Artificial insertion of 2'-O-me at the mRNA branch point in Xenopus oocytes

activates cryptic branch points in mRNA, thus changing splicing pattern. If no cryptic branch points are present in vicinity, splicing is prevented, even if mRNA is still assembled into pre-catalytic spliceosome[225].

2'-O-methylated residues in mRNAs play different functions depending on their proximity or not to the cap. For instance, it was shown that nucleotide N1 methylation serves as a molecular signature of "self" and thus many viruses produce Cap1-mRNA through the action of host or virally encoded MTases[226, 227], thereby escaping from detection by the immune system[228]. In contrast, mRNAs that do not contain any 2'-O-methylation close to the cap are marked as "non-self" and trigger type I IFN response. This mechanism is driven in part by the IFIT proteins (IFN-induced protein with tetratricopeptide repeats-1) which were shown to be effectors of the host immune innate response by selectively inhibiting translation of viral mRNAs[229]. Indeed, X-ray structural analysis recently revealed that these proteins specifically recognize foreign capped mRNAs that do not contain 2'-O-methylated residues and bind to the cap and four additional nucleotides thereby competing with the translation initiation factor eIF4F[229]. The RNA sensor MDA5 is also probably involved in the recognition of viral RNAs that do not contain 2'-O-methylated residues[230].

Recently, numerous internal 2′-O-methylation sites were discovered within the human mRNA coding regions and may affect translation either by changing tRNA–mRNA or mRNA–ribosome interactions, resulting in altered dynamics of translation elongation. The first piece of evidence came from an elegant *in vitro* study where the authors showed that the presence of a 2′-O-methylated mRNA codon decreased translation efficiency, especially when this modification is present at the second position of a codon[231]. This was further confirmed by both single-molecule and bulk kinetic methods, which showed that decrease in translation efficiency is not due to disruption of codon–anticodon RNA-duplex formation, but because it impairs efficient decoding of the modified codon and results in excessive rejection of cognate tRNA during proofreading. Therefore, the presence of 2'-O-methylations in some mRNAs may selectively decrease translation elongation[232].

Importance of 2'-O-methylations for immunostimulation/suppression

In the field of innate immunity, Toll-like receptors (TLRs) play a crucial role. They recognize microbial associated molecular patterns (MAMPs) and thus activate immune cells. TLR3, 7, 8, 9 and 13 (mouse specific) are known to be activated by nucleic acids. Indeed, TLR3 is specific for double stranded RNA, mouse TLR13 for bacterial 23S rRNA and TLR 9 for unmethylated CpG-DNA. Moreover, TLR7 and TLR8 are activated by single stranded RNA of bacterial origin. In humans, TLR7 is expressed only in plasmacytoid dendritic and B cells and induces the production of type I interferon such as IFN-α, whereas TLR8 is found in monocytes, myeloid dendritic cells and regulatory T cells and its activation leads the production of cytokines[233]. The innate immune system is able to make a difference between "self" and "non-self" RNA molecules due to specific RNA modifications, spatial restrictions and sequence

positions in nucleic acids. It is now well established that 2'-O-methylations act as immunosuppressors. *In vitro* studies showed that 2'-O-methylations incorporated in *in vitro* RNA transcripts, siRNA or 18S rRNA prevent TLR activation[233, 234]. In fact, in native RNA, it was demonstrated that Gm18 in bacterial *E. coli* tRNA is responsible for escape from the immune response and that bacterial $tRNA^{Tyr}$ plays a major role[235, 236]. The molecular mechanism of immunosupression is still poorly understood. 2'-O-methylated tRNA is a TLR7 antagonist[237] with a recognition pattern of 2 nucleotides including Gm18 and its downstream neighboring nucleotide[238]. In addition, it was shown that 2'-O-methylation acts as an inhibitor for TLR7, but not for TLR8[233]. However, other bacterial species that lack Gm18, such as Gram positive strains (e.g. *S. aureus*) also inhibit TLR7 response. This shows that 2'-O-methylations at other positions in tRNA may also have immunosuppression properties[236].

2'-O-methylations in cancer and development

Importance of rRNA 2'-O-methylations was established by suppression of snoRNA expression by morpholino antisense oligonucleotides in zebrafish model. Suppression of U26, U44 and U78 snoRNAs (corresponding to 28S-Am398, 18S-Am163 and 28S-Gm3745) reduces rRNA modification at the corresponding sites and led to severe morphological defects and embryonic lethality, suggesting the essential role of those rRNA 2'-O-methylations in vertebrate development[238]. It was recently shown that increased levels of C/D box snoRNA/snoRNP correlated with increased 2'-O-methylation in rRNA contribute to leukaemogenesis[239], probably through increased protein synthesis and impairment of translation fidelity. Therefore, these variations of 2'-O-methylation levels of rRNA contribute to ribosome heterogeneity. In parallel, impairment of Fibrillarin in human cells shows that it correlates with a global 2'-O-methylation rRNA decrease perturbing translation fidelity[108, 109]. These accumulating evidences show that controlling 2'-O-methylation levels of rRNA probably may play a key role in ribosome heterogeneity and cancer progression and outcome. Further studies will be necessary to decipher how 2'-O-methylation levels of rRNA tightly regulate cell's fate.

Conclusion

RNA modifications play an essential role in gene expression regulation, by altering and fine tuning of mRNA, rRNA, tRNA and ncRNA properties. Despite the decades of considerable efforts, our understanding of RNA modification functions remains incomplete. Recent studies aiming to precisely locate RNA modifications and specifically RNA 2'-O-methylations now bring new insights in the field, providing a solid basis for future development. The ultimate step is the precise mapping of such residues in representative cellular RNAs and further understanding of physiological properties of post-transcriptional RNA alterations.

Acknowledgments

Y.M was supported by ANR-DFG grant HTRNA-Mod ANR-13-ISV8-0001 and Research funds from Lorraine Région (France). This work was performed in the COST EPITRAN (CA16120) framework.

-- End of article --

1.2.2 Pseudouridine (Psi)

This peculiar modification is formed by isomerisation of uridine, forming a C-C bond instead of C-N bond between the ribose and the nucleobase (see Figure 1). Thanks to this unique chemical bonding, this modification helps to stabilize overall RNA structure, especially in tRNAs[240].

This was the 1st modified nucleotide discovered[241] and is the 2nd most abundant modifications in rRNAs after 2'-O-methylation.

This modification is either catalyzed by H/ACA box snoRNP for rRNAs but also by stand-alone proteins such as PUS family enzymes[242].

Pseudouridine dysregulation is associated to certain types of cancer and deterioration of neurodevelopment[243]. This isomerised nucleobase is also known for being an immune response suppressor, and is therefore used in mRNA vaccines design[244] to protect them from the immune system. However, the biological functions of pseudouridine are not yet fully understood but show promises for therapeutic innovations[243].

1.2.3 N6-methlyadenosine (m^6A)

Nucleobase methylation is a common RNA modification such as 5-methylcytidine (m^5C), N7-methylguanosine (m^7G) and N6-methyladenosine (m^6A) (see Figure 1). The latter occurs at the position 6 of the nucleobase adenine. A methylation at this precise location changes the primary amine group of the adenine into a secondary amine which can prevent its deamination under certain conditions. A deamination of adenine deamination would result in formation of an inosine[245], changing the base chemical properties and hydrogen bonding, thus making this methylation a nucleobase chemical stabilizer.

The regulation of m^6A is thoughtfully explained in the literature[246, 247]. First, m^6A is catalyzed by "writers" which are methyltransferases with the most prominent ones being the enzymes from the METTL family. From then, the m^6A interacts with "reader" proteins such as the ones from the YTH family which promote or decay m^6A-containing RNA metabolism. In the meantime, this methylation can be reversed due to "eraser" demethylases such as FTO, enabling dynamic regulation of this modification.

m^6A is predominant in mRNA and as such, an important regulator of gene expression[248]. Apart from this role, numerous studies display m^6A interaction with cancer[249–251] and RNA self-recognition by the immune system[252, 253] making it an interesting therapeutic target.

1.2.4 Impact on diseases

As seen through these three examples, RNA modifications are important modulators of the function of RNAs bearing them. These chemical elements impact the RNA transcription, translation, splicing and stability. It has been proven that modifications such as 2'-O-methylation play an important role in viral infections and innate immune responses[254]. A recent study on Human Immunodeficiency Virus (HIV) RNA proved

its ability to recruit FTSJ3, a 2'-O-methyltransferase, to escape the innate immune system allowing it to remain in infected hosts[255]. Disruption of HIV-1 genome 2'-O-methylation allows decay of this viral RNA[256] preventing the infection. FTSJ3 inhibitors development would then be a promising way to prevent or even cure Acquired Immune Deficiency Syndrome (AIDS).
In the meantime, there have been numerous reports of modification malfunctions due to variations in the content of modifications and differential proteins expression of modification writers, readers and erasers, which are leading to the exarcerbation of diseases[257, 258]. Studies about modification writers' impacts in cancer have opposite outcomes since they act either as a promoter of oncogenic targets or as a tumor suppressor, leading to confusion about their role in the field[258].
This is why epitranscriptomics studies about modifications regulation and dysfunction occurring in diseases would truly help the medical and pharmaceutical field to find new cures.

1.3 Detection of RNA modifications

The investigation of the links between RNA modifications and varied diseases is mainly attributable to elaboration of modifications detection techniques. By detecting their presence and/or absence, information about molecular cellular mechanisms can be elucidated. Creation of efficient RNA modification detection methods is crucial for investigation of disease pathways and innovative drug design.

1.3.1 Physico-chemical detection methods

The first detection method used was Thin Layer Chromatography (TLC) to identify and characterize canonical nucleobases and their derivatives[259]. This method is based on elution of the different nucleotides on a plate from digested RNA. Depending on the affinity of each nucleobase for the solvent, they migrate at different speeds, allowing their separation and identification. Since they are colourless but absorb UV-light at 260 nm, the detection and relative quantification of the nucleotides are done by UV-spectrometry. From then, this method has been improved to increase its sensitivity and application, by performing a two-dimensional TLC (2D-TLC)[260] or modified to give rise to capillary electrophoresis[261] in order to reduce the amount of input material. Recently, a TLC-based technique termed "site-specific cleavage and radioactive-labelling followed by ligation-assisted extraction and thin-layer chromatography" (SCARLET) allows precise identification and quantification of a nucleotide of interest[262].
In parallel, liquid chromatography-tandem mass spectrometry (LC-MS/MS) is a powerful tool for nucleosides characterisation thanks to its high sensitivity and accuracy for mass calculations[263, 264]. RNA is digested into a mixture of nucleosides (such as for TLC techniques), separated by a liquid chromatography column, and deduction of their mass by the mass spectrometer. Moreover, the nucleoside can be

further fragmented, usually resulting in the separation between the ribose and the nucleobase, allowing a distinct mass calculation for both. From their mass, the chemical formula of a nucleoside can be deduced which allows the reconstruction of its molecular structure for its identification. Depending on the desired outcome, LC-MS/MS can be used to detect nucleosides even at femtomolar amounts[264] or for identification of yet unknown modifications[265].

Both TLC-related and LC-MS/MS techniques are powerful tools for RNA modification studies. However, due to their need for nucleoside digestion for characterisation, these techniques cannot properly identify the location of the modifications in the transcriptome. Even though SCARLET can specifically select a single nucleotide of interest, the application of this method to a whole sequence would be laborious and time-consuming[266]. For better insights on RNA modifications, mapping tools need to be conceived to localise the presence or absence of epitranscriptomic markers at nucleotide resolution.

1.3.2 Sequencing

RNA sequencing allows decoding of all nucleobases of RNA sequences. Enhancement to this technology has led to the development of Next-Generation Sequencing (NGS) machines enabling the deciphering of millions of sequences in a single run.
For instance, Illumina's sequencing technology is based on Sequencing by Synthesis (SBS) for high-throughput nucleotide detection. SBS is based on Polymerase Chain reaction (PCR) amplification of DNA strand clusters and the detection of fluorescent emitters attached to the nucleotides. DNA strands from one or more samples are arranged on a counting cell. Identical DNA strands form clusters and are replicated several times by bridge PCR. During this *de novo* synthesis, each of the four nucleotides - A, C, G, T - has a fluorescent marker that enables it to be identified. When one of the four nucleotide binds to its complementary strand during PCR amplification, its attributed fluorophore is detected to identify it. This process is repeated simultaneously for all the clusters in the counting cell until each strand of fixed DNA has been fully sequenced. This simultaneous detection of several thousand strands of DNA means that around ten human genomes can be sequenced in the space of a few days.
Even though the sequencing machines are built to detect the canonical nucleobases A/C/G/T in DNA, fine adjustment of the method allows detection of modified nucleosides also in RNA. These developed sequencing detection methods follow roughly the same data processing (see Figure 2).

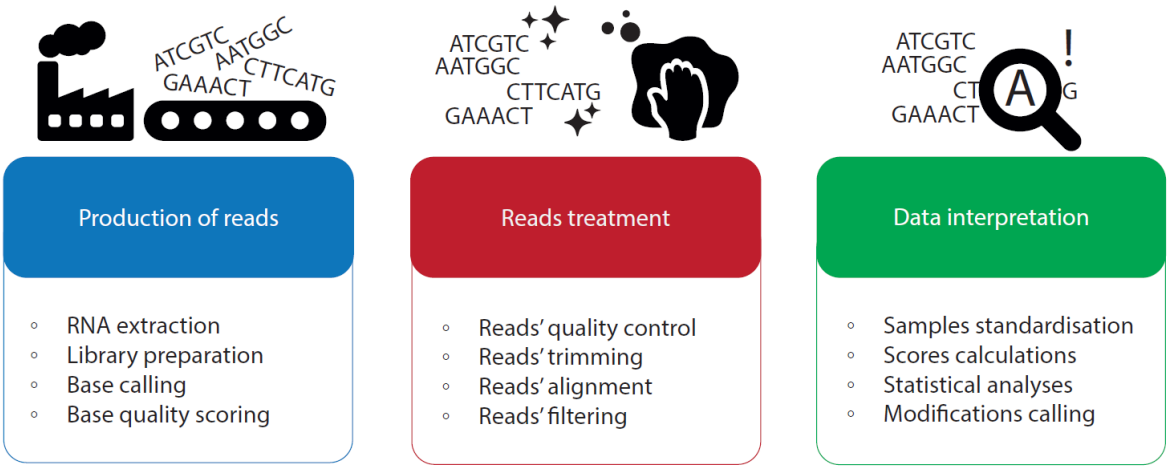


Figure 2: General procedure of Next-Generation Sequencing RNA modification detection methods

The three common stages of sequencing detection methods procedure: Production of reads, reads treatment and data interpretation are detailed with their respective steps.

In short, RNA sequences of interest are extracted and reverse transcribed to a cDNA library which is inserted to a sequencing device. The latter assigns the nucleobases of the sequences of interest based on the signals it receives, creating "reads", and assigns to each identified nucleobase a quality score. This quality score (also known as Phred scores)[267] enables the quality control of the reads before their trimming (removal of the adapter sequences used for the library preparation) and alignment (assembly of the different reads to a sequence of reference). From these steps, optional filtering reads steps can occur before standardization of the data for their interpretation. Then, each sequencing detection method calculates their related metrics called "scores" for the detection of modifications. Finally, the different scores are processed through statistical analyses to differentiate significant signals compared to the noise in order to pinpoint the location of the modifications in a sequence of interest.

The main difference between the varied sequencing methods is the detection approach. This can be mainly translated to personalisation of the library preparation, the scores calculations and the statistical analyses. To get a better idea about these variences, multiples sequencing detection methodologies will be described below.

RT-signatures

Some modifications such as N1-methyladenosine (m^1A) and 1-methylguanosine (m^1G) can naturally blocks the reverse transcriptase (RT) during library preparation due to the presence of modifications on the Watson-Crick's edge (see Figure 1). This induces reads to stop specifically at the modified sites, thereby allowing their detection. This approach has the potential to detect multiple modified sites at once. However, the exact nature of the modification is difficult to clearly identify (differentiate m^1G from another G nucleobase methylation for example). An exception to this statement can

be made by using Bayesian and/or deep-learning approach which can find subtle signal differences from similar modifications[268].

Although this approach can only be used for naturally occurring modifications blocking the RT, it is possible to induce it to non RT-signatures sensitives modifications by using chemical adducts.

Chemical adducts for RT stops induction

One example of such a methodology is the entitled Pseudo-Seq using carbodiimide N-cyclohexyl-N′-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate (CMC) adducts[269]. This adduct specifically binds to pseudouridines residues through click chemistry, which subsequently results in RT stops during library preparation. This allows detection of pseudouridine though RT-signatures approach whereas natural pseudouridine does not hinders RT. Such approach allows detection of modifications usually unseen by standard RT-signatures methods. However, since RT-signatures can detect multiple modifications at the same times, negative controls (samples without CMC addicts) are needed to ensure correct detection of the modification of interest. Also, if the click chemistry reaction is partial, i.e., some pseudouridines residues are not linked to CMC addicts, the modification quantification could be difficult. Quantification of the modification remains an important aspect of detection techniques to correctly assess the differences between total absence and decrease of modification level for physiological studies.

Antibody-based methods

An example of sensitive detection and quantification sequencing method is m^{6}A-seq[270] and its derivatives MeRIP-Seq and miCLIP-m^{6}A. They have based their detection of m^{6}A through use of antibodies anti-m^{6}A. By immunoprecipitation of m^{6}A-bearing RNA molecules and comparing it to a negative control, these methods can pinpoint locations of RNA regions enriched with m^{6}A and even quantify then with use of spike-in RNA which serves as internal standard. However, the precise location of m^{6}A sites at nucleotide resolutions is difficult and can be subject of controversies depending on the statistical analysis parameters used for m^{6}A-peak attribution. Also, controversies lie in the antibodies used for these methods since their specificity is up to debate[271, 272]. Creation of an antibody-independent m^{6}A sequencing detection method would then be a suitable alternative approach for studies about this modification.

Chemical treatment for modifications' signals extraction

Another kind of sensitive and quantitative sequencing methods are the ones based on using modifications unique chemistry as a signature to identify them through chemical treatment. This detection approach is based either on chemical protection induced by

the modification, or on chemical conversion specifically targeting the modification of interest.

To illustrate a method based on chemical protection, RiboMethSeq method[63] use the 2'-O-methylation chemical protection property against alkaline cleavage in order to detect them. Subjecting RNA sequences to alkaline conditions results in random fragmentation of the molecule except at the vicinity of the 2'-O-methylation which hinders it. Assessing the nucleotide protection against alkaline cleavage allows then the detection and even the quantification of the 2'-O-methylation through RiboMethSeq scores calculations. Indeed, amplitude of this protection is proportional to the 2'-O-methylation stoichiometry providing its quantification[64]. Similar feats can be achieved by the m^5C detection method BisulfiteSeq[273] which uses the deamination protection provided by this modification to detect it.

Finally, to illustrate a method based on modifications chemical targeting, AlkAnilineSeq methods based its detection on intrinsic fragility of a subset of modifications such as m^7G towards alkaline conditions[274]. This induces formation of abasic sites which are later specifically cleaved by aniline allowing modification detection.

Aniline-induced cleavage, such as most of the techniques based on chemical treatment approach, allows key relative quantification of modification[275]. However, alteration of RNA chemistry can induce false-positives hits appearance in sequencing data and unspecific reaction with other modifications than the ones of interest[276]. Even though these issues can be anticipated and minimized, they must be dealt with care to perform right data interpretation and projects' application.

Nanopore

Recently, an innovative sequencing technology entitled Nanopore sequencing[277] emerged with potential to detect and characterized most if not all modification site in one run[277–279]. This technology consists of passing RNA sequences through a nanopore, thus generating flow current disruptions. By analysing the disruptions signals strengths, they can be link to the (modified) nucleotide which produced them. This involves base calling is also simultaneously modifications calling with this approach, but this needs strong Bayesian and/or deep-learning tool to achieve this[278].

However, this technology is still fresh, and as it is currently it carries on high potential but not all the experience and application to exploit it to its full yet. Nevertheless, validation of modifications calls from nanopore sequencing can also be done by other approaches cited earlier. Moreover, methods and statistical tools developed for other approaches could be use in the future for this dream technology such as the strategies of modifications calling from RT-signature methods or enhancing modifications signals through RNA chemical treatment.

Thanks to diversified approaches and optimized detection of modifications by RNA sequencing, the next step in epitransciptomics could be the collaboration and cross-talks between different methods to explore unprecedented sights in biological and pharmaceutical studies.

II. Motivation and objectives

Since the discoveries of new nucleobases due to enzymatic chemical modifications, an increasing number of techniques for the characterization and detection of those elements has been established. Deregulation of these modifications are linked to increased sensitivity to diseases, from viral infections to tumor development. Developed methods for epitranscriptomic studies should be then thoroughly optimised to assure clear and efficient data gathering on post-transcriptional markers biogenesis, regulations and malfunctions. This latter would help the pharmaceutical field to find new insights and cures for specific diseases.

In order to better understand the interactions between RNA modifications and diseases, studying RNA 2'-O-methylation would be beneficial to the field. A high-throughput sequencing detection method for 2'-O-methylation, such as an Illumina-adapted RiboMethSeq protocol, would be suitable for this task This would further expand the reach of discovery related to 2'-O-methlyation biological functions.
In the meantime, however, the epitranscriptomic field also needs optimised and efficient detection methods to avoid false-positives hits which can hinder correct interpretation of the different biological and case studies. While using the same approach, lab groups often get strongly divergent results due to the different parameters and threshold used[272]. Every sequencing detection method follows roughly the same procedure (Figure 2). But each step has distinctive parameters and approaches, which increases the dispersion of results between research groups, to the detriment of potential new discoveries. As an example, not all research groups are using the same sequence references for rRNAs or tRNAs[280, 281], creating confusion to modified sites attribution and comparisons.
To address this, an in-depth evaluation of the various analysis parameters, from trimming to statistical methods application, ought to be carried out for as many sequencing detection methods as possible. With this intention in mind, a holistic study and optimisation of RiboMethSeq parameters would minimize the potential errors which could occur with this method. Also, a semi-automated method for generating tRNA references adapted for epitranscriptomic studies would maximize the discoveries potential on tRNAs and enable researchers to base the allocation of modified sites in tRNAs on a common reference. In order to ensure transparency and reproducibility on the optimisation, online publications of the various scripts would be an advantage. Finally, application of machine learning algorithms, known for their strength to detect patterns in data, for automated interpretation of epitranscriptomicx data would considerably improve the capacity to efficiently detect modifications if set-up correctly.

In parallel, providing innovative approach or alternatives for modifications detection would be advantageous for expanding the arsenal of research's tools to decipher RNA (multiple) functions.
For exemple, sequencing detection methods for pseudouridine are based on chemical adducts on pseudouridylated residues occurring in RT stops[269]. The latter allows good detection of these modifications but these methods are only semi-quantitative which reduces the potential for data interpretation. That is why a method allowing detection and quantification of pseudouridine would be beneficial to the field. With this idea in mind, RNA hydrazine treatment shows potential in that regard since it generates ring-opening of the pyrimidine base of uridines but not on pseudouridines, distinguishing these two residues[282]. An essay regarding this detection methodology based on hydrazine with its detailed bioinformatic pipeline workflow should be undertaken.
Concerning m^6A, the main methods for its detection rely mainly on commercial antibodies[270], which do not allow reliable quantification at nucleotide level. Moreover, unwanted antibody cross-reactivity with another modification can occur as it has been the case for similar methods for m^1A detection[283]. An antibody-independent method for m^6A detection would overcome the limitations of current techniques and could be used to enhance the mapping of new m^6A sites in the epitranscriptome. That is why a method based on chemical distinction between unmodified adenosine and m^6A, such as bisulfite sequencing, would be optimal to complement current techniques.

Finally, simultaneous application of distinctive optimised pipelines on research projects would unravel the mechanisms by which modifications are regulated, thanks to an analysis of a detailed mapping of the epitrancriptome. Results compilation from various detection methods on the same material ought to be tested in order to find new cures to pathologies.

III. Contribution to papers

Pichot F, Marchand V, Helm M, Motorin Y. Machine learning algorithm for precise prediction of 2'-O-methylation (Nm) sites from experimental RiboMethSeq datasets. Methods. 2022 Jul;203:311-321. doi: 10.1016/j.ymeth.2022.03.007. Epub 2022 Mar 18. PMID: 35314341.

F. Pichot designed and performed the machine learning algorithms implementation from their adaptation to RiboMethSeq data to models' performance assessment with Y. Motorin and shared contribution with V. Marchand, M. Helm and Y. Motorin for manuscript writing.

Contribution as first author: 75%

Pichot F, Hogg MC, Marchand V, Bourguignon V, Jirström E, Farrell C, Gibriel HA, Prehn JHM, Motorin Y, Helm M. Quantification of substoichiometric modification reveals global tsRNA hypomodification, preferences for angiogenin-mediated tRNA cleavage, and idiosyncratic epitranscriptomes of human neuronal cell-lines. Comput Struct Biotechnol J. 2022 Dec 19;21:401-417. doi: 10.1016/j.csbj.2022.12.020. PMID: 36618980; PMCID: PMC9798144.

F. Pichot designed and perfomed most of the sequencing analyses with M. Helm, Y. Motorin, M.C. Hogg, and H.A. Gibriel. M.C. Hogg provided the samples and V. Marchand and V. Bourgignon the sequencing data. E. Jirström and C. Farrell performed the immunocytochemistry experiment. The manuscript was written by F. Pichot, M. Helm, Y. Motorin, V. Marchand and J.H.M. Prehn.

Contribution as first author: 70%

Pichot F, Marchand V, Helm M, Motorin Y. Non-Redundant tRNA Reference Sequences for Deep Sequencing Analysis of tRNA Abundance and Epitranscriptomic RNA Modifications. Genes (Basel). 2021 Jan 10;12(1):81. doi: 10.3390/genes12010081. PMID: 33435213; PMCID: PMC7827920.

F. Pichot and Y. Motorin designed and performed the creation of the R pipeline for generation of collapsed tRNA sequences references including their validation from experimental data provided by V. Marchand. F. Pichot published the different scripts online. The manuscript was written by all the authors.

Contribution as first author: 60%

Pichot F, Marchand V, Ayadi L, Bourguignon-Igel V, Helm M, Motorin Y. Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2'-O-Methylations in RNA by RiboMethSeq. Front Genet. 2020 Feb 13;11:38. doi: 10.3389/fgene.2020.00038. PMID: 32117451; PMCID: PMC7031861.

F. Pichot designed and performed the optimisation of RiboMethSeq pipeline including their validation from experimental data provided by V. Marchand, L. Ayadi and V. Bourguignon-Igel. F. Pichot published the different scripts online. The manuscript was written by F. Pichot, Y. Motorin, V. Marchand and M. Helm.

Contribution as first author: 60%

Pichot F, Marchand V, Helm M, Motorin Y. Data Analysis Pipeline for Detection and Quantification of Pseudouridine (ψ) in RNA by HydraPsiSeq. Methods Mol Biol. 2023;2624:207-223. doi: 10.1007/978-1-0716-2962-8_14. PMID: 36723818.

F. Pichot designed and performed the HydraPsiSeq pipeline creation and data analysis. F. Pichot published the different HydraPsiSeq scripts online. The manuscript was written by F. Pichot, Y. Motorin, V. Marchand and M. Helm.

Contribution as first author: 50%

Marchand V, Pichot F, Thüring K, Ayadi L, Freund I, Dalpke A, Helm M, Motorin Y. Next-Generation Sequencing-Based RiboMethSeq Protocol for Analysis of tRNA 2'-O-Methylation. Biomolecules. 2017 Feb 9;7(1):13. doi: 10.3390/biom7010013. PMID: 28208788; PMCID: PMC5372725.

F. Pichot designed and performed the RiboMethSeq pipeline creation and sequencing data analysis.

Contribution as co-author: 30%

Marchand V, Pichot F, Neybecker P, Ayadi L, Bourguignon-Igel V, Wacheul L, Lafontaine DLJ, Pinzano A, Helm M, Motorin Y. HydraPsiSeq: a method for systematic and quantitative mapping of pseudouridines in RNA. Nucleic Acids Res. 2020 Nov 4;48(19):e110. doi: 10.1093/nar/gkaa769. PMID: 32976574; PMCID: PMC7641733.

F. Pichot designed and performed the HydraPsiSeq pipeline creation and data analysis.

Contribution as co-author: 25%

Werner S, Galliot A, Pichot F, Kemmer T, Marchand V, Sednev MV, Lence T, Roignant JY, König J, Höbartner C, Motorin Y, Hildebrandt A, Helm M. NOseq: amplicon sequencing evaluation method for RNA m^6A sites after chemical deamination. Nucleic Acids Res. 2021 Feb 26;49(4):e23. doi: 10.1093/nar/gkaa1173. PMID: 33313868; PMCID: PMC7913672.

F. Pichot designed the R pipeline creation for m^6A detection which has been later adapted in Python by A. Galliot.

Contribution as co-author: 15%

Freund I, Buhl DK, Boutin S, Kotter A, Pichot F, Marchand V, Vierbuchen T, Heine H, Motorin Y, Helm M, Dalpke AH, Eigenbrod T. 2'-*O*-methylation within prokaryotic and eukaryotic tRNA inhibits innate immune activation by endosomal Toll-like receptors but does not affect recognition of whole organisms. RNA. 2019 Jul;25(7):869-880. doi: 10.1261/rna.070243.118. Epub 2019 Apr 24. PMID: 31019095; PMCID: PMC6573781.

F. Pichot designed and performed the RiboMethSeq pipeline creation and sequencing data analysis with Y. Motorin.

Contribution as co-author: 15%

Ayadi L, Galvanin A, Pichot F, Marchand V, Motorin Y. RNA ribose methylation (2'-O-methylation): Occurrence, biosynthesis and biological functions. Biochim Biophys Acta Gene Regul Mech. 2019 Mar;1862(3):253-269. doi: 10.1016/j.bbagrm.2018.11.009. Epub 2018 Dec 17. PMID: 30572123.

F. Pichot modeslty contributed to the section about "Higher eukaryotic rRNA 2'-O-methylation in human and mouse models" by pinpointing the conserved 2'-O-methylated sites between mice and humans.

Contribution as co-author: 10%

IV. Results and Discussion

Chapter 1: Illumina-adapted RiboMethSeq method

4.1.1 Next-Generation Sequencing-Based RiboMethSeq Protocol for Analysis of tRNA 2'-O-Methylation. [DOI: 10.3390/biom7010013]

Virginie MARCHAND[1,2], Florian PICHOT[1], Kathrin THÜRING[3], Lilia AYADI[1], Isabel FREUND[4], Alexander DALPKE[4], Mark HELM[3], Yuri MOTORIN[1,2]

[1] IMoPA UMR7365 CNRS-UL, Biopole Lorraine University, 54505 Vandoeuvre-les-Nancy, France

[2] Next-Generation Sequencing core facility, FR3209 BMCT CNRS-UL, Biopole Lorraine University, 54505 Vandoeuvre-les-Nancy, France

[3] Institute of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, 55128 Mainz, Germany

[4] Dept. of Infectious Diseases, Medical Microbiology and Hygiene, Ruprecht-Karls University Heidelberg, 69120 Heidelberg, Germany

Overview

The biological role of tRNA 2'-O-methylation as a signature for cellular immune system evasion is in dire need of in-depth characterisation in order to better understand the distinction between self and non-self recognition, intrinsically linked to viral infections and autoimmune diseases. Therefore, the ability to specifically detect tRNA 2'-O-methylation at nucleotide-resolution is required, though such available detection methods are few. Next-Generation Sequencing technology (such as Illumina sequencing technology) is the prime candidate to reach that goal thanks to its capability to gain nucleotide information on billions of sequences in a single run. From this perspective, we present the first application of an Illumina-based protocol for 2'-O-methylation detection on tRNAs.

This work has been based on the first application of an Illumina-based RiboMethSeq protocol on rRNA[64]. At this time, the data analysis was done on Excel which was sufficient for simple rRNA analysis, but lacked computational power for broader projects such as tRNA analyses. This is why a new pipeline needed to be designed based on R software for its statistical applications, graphics design and relative programming language simplicity. To ensure its application to concrete biological studies, this pipeline has been then tested on *Escherichia coli* and *Saccharomyces cerevisiae* data from Pr. Dalpke's group, investigating the immunostimulation regulation done by 2'-O-methylated sites in tRNAs.

We have shown the potential and limitations of Illumina-based RiboMethSeq tRNA analyses through the investigation of known 2'-O-sites in *E. coli* and *S. cerevisiae*, with a validation through specific removal of 2'-O-methylated guanosine on position 18 (Gm18) of a tRNA subset. In fact, mutant strains for TrmH in *E. coli* and Trm3 in *S. cerevisiae* which are enzymes both responsible for the 2'-O-methylation catalysis of Gm18 have been realised and subjected to sequencing. Then, the data analysis showed presence of Gm18 in wild-type strains, but not on mutant strains, while other 2'-O-methylated sites displayed no changes, highlighting the robustness of the method.

However, this application has also pictured two main issues for RiboMethSeq tRNA analyses which are the lack of adapted tRNA references available and multiple false-positives hits due to the irregularity of the sequencing profile of tRNAs. These concerns would need to be closely inspected and corrected for further RiboMethSeq applications.

For this work, I have overseen the creation of the RiboMethSeq pipeline in R programming language for sequencing data import and filtering, RiboMethSeq scores calculations used for 2'-O-methylation detection and graphs production. I have also analysed the tRNA samples displayed in this paper and interpreted them.

Abstract

Analysis of RNA modifications by traditional physico-chemical approaches is labor-intensive, requires substantial amounts of input material and only allows site-by-site measurements. Recent development of qualitative and quantitative approaches based on next-generation sequencing (NGS) opens new perspectives for analysis of various cellular RNA species. Illumina-sequencing based RiboMethSeq protocol was initially developed and successfully applied for mapping of rRNA 2'-O-methylations. This method also gives excellent results in quantitative analysis of rRNA modifications in different species and under varying growth conditions. However, until now RiboMethSeq was only employed for rRNA and the whole sequencing and analysis pipeline was only adapted to this long and rather conserved RNA species. A deep understanding of RNA modification functions requires large and global analysis datasets for other important RNA species, namely for tRNAs, which are well known to contain a great variety of functionally important modified residues. Here we evaluated the application of RiboMethSeq protocol for analysis of tRNA 2'-O-methylation in E. coli and in S. cerevisiae. After a careful optimization of the bioinformatic pipeline, RiboMethSeq proved to be suitable for relative quantification of methylation rates for known modified positions in different tRNA species.

1. Introduction

RNA modification is a complex step in post-transcriptional maturation of cellular RNAs and includes numerous chemical alterations of initially incorporated parental nucleotides. Among these various chemical reactions, the most frequent is the transfer of a methyl(CH_3)-group from a methylation donor (generally S-adenosyl-L-methionine, abbreviated SAM or AdoMet) to different positions on nucleobases or to the 2'-OH of the ribose[136, 162, 284–286]. Methylated nucleotides are found in every studied RNA species and their presence allows fine modulation (tuning) of RNA properties in different cellular processes[287–290]. One of the best known RNA species are tRNAs whose properties and functions have now been studied for over 50 years. These relatively small RNA molecules have been extensively analyzed for their RNA modification profile and a wealth of information is now available for tRNA from bacteria, lower eukaryotes (mostly yeasts and fungi) and mammals (human, rat, rabbit, beef)[291]. This already served as a basis for extensive search of the corresponding enzymes and enzymatic machineries involved in modification of tRNA species. Thanks to these efforts, most of the enzymes for tRNA modification in model organisms were identified and their specificity characterized[292–294]. It is noteworthy that archaeal species somehow escaped this thorough analysis and only a few tRNA species are included in available databases. Archaeal enzymes are mostly characterized by sequence homology followed by functional tests to confirm the initial hypothesis on their predicted activity (aTrm56)[295, 296].

One very common methylation in many RNA species is 2'-O-ribose methylation (2'-O-Me). This type of methylations touches all 4 canonical nucleotides in RNA and even many base-modified residues, the most frequent probably being 2'-O-methylated pseudouridine (Ψm).

Bacterial tRNA 2'-O-methylation

Early on, analysis of tRNA species from the model bacterial organism *Escherichia coli* demonstrated the presence of three possible positions for 2'-O-methylation (Table 1). A major site is located in D-loop at position 18, where the highly conserved G residue is frequently converted to Gm, while two others are located in the anticodon loop at positions 32 and 34 (wobble) where Cm and Um/cmnm^{5}Um residues are frequently present (Figure 1). *E. coli* tRNA has no other reported sites, while other bacterial species like *B. subtilis* and *M. capricolum* seem to lack Gm18, but may contain Gm34 (*B. subtilis*) or cmnm^{5}Um34 (*M. capricolum*) at the wobble position of the anticodon.

Formation of Gm18 in bacterial tRNA is catalyzed by a specific tRNA:Gm18-methyltransferases (MTases) from the SpoU (now renamed to TrmH) family, which had already been predicted to be an RNA:MTase[297–299], prior to experimental validation. This enzyme was found in the *gmk-rpoZ-spoT-spoU-recG* operon in *E. coli*[142] and was also conserved in related bacterial species and in extreme thermophile *T. thermophilus*[144].

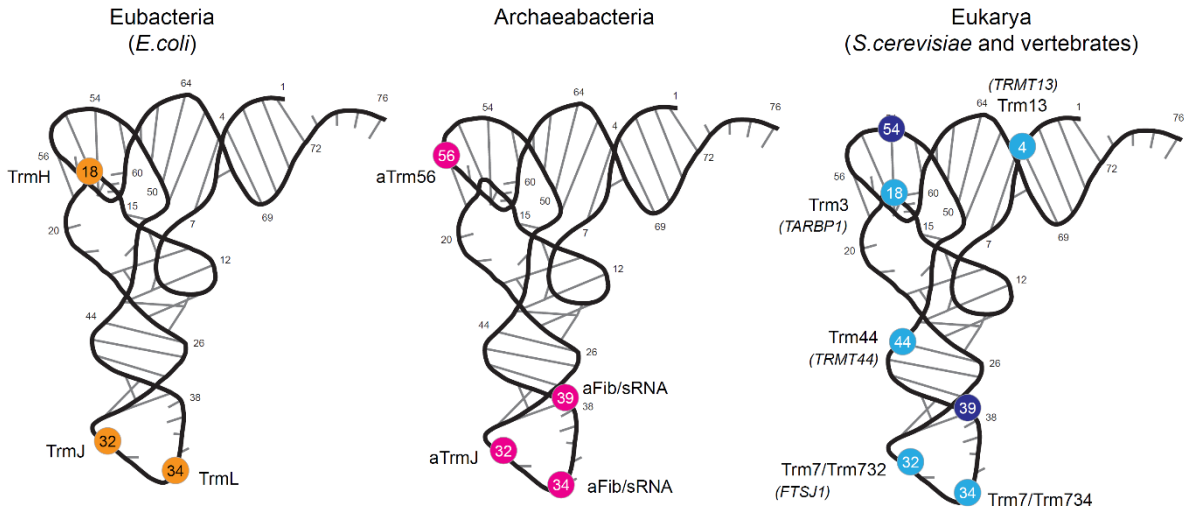


Figure 1. Known positions of tRNA 2'-O-methylations and corresponding enzymes in Eubacteria (*E. coli/B. subtilis*), Archaea (data almost exclusively from *H. volcanii*) and Eukarya (*S. cerevisiae* and various vertebrates). tRNAs are represented in folded 3D structures. 2'-O-methylated nucleotides are shown as spheres. Light blue color corresponds to yeast modifications, additional vertebrate positions are shown in dark blue. Names of homologous human proteins are in italics.

Table 1. Known tRNA:2'-O-methyltransferases in model organisms (*E. coli*, *H. volcanii*, S. cerevisiae, *H. sapiens*) **and their tRNA substrates.**

Organism	Enzyme	Enzyme alt name	Position	Modification	tRNA	Anticodon	tRNAdb ref
E. coli	TrmH	spoU	18	Gm	Ile	k^2CAU	tdbR00000160
					Leu	cmnm^{5}UmAA	tdbR00000220
					Leu	CAG	tdbR00000221
					Leu	GAG	tdbR00000222
					Met	ac^{4}CAU	tdbR00000276
					Gln	CUG	tdbR00000332
					Gln	U*UG	tdbR00000333
					Ser	CGA	tdbR00000390
					Ser	GGA	tdbR00000392-3
					Ser	cmo^{5}UGA	tdbR00000394
					Tyr	QUA	tdbR00000545-46
					Leu	CmAA	tdbR00000595
	TrmJ	YfhQ, TrMet(Xm32)	32	Cm	Ser	cmo^{5}UGA	tdbR00000394
					Trp	CCA	tdbR00000484
					Ini	CAU	tdbR00000510-11
				Um	Gln	CUG	tdbR00000332
					Gln	U*UG	tdbR00000333
					Pro	CGG	tdbR00000623
	TrmL	yibK	34	Cm	Leu	CmAA	tdbR00000595
				cmnm5Um	Leu	cmnm^{5}UmAA	tdbR00000220
H. volcanii	aTrmJ	HVO_2906	32	Cm	Lys	U*UU	tdbR00000175
					Lys	ac^{4}CUU	tdbR00000176
					Trp	CmCA	tdbR00000478
					Tyr	GUA	tdbR00000541
	sRNA/aFib		34	Cm	Met	CmAU	tdbR00000271
			34	Cm	Trp	CmCA	tdbR00000478
			39	Um	Trp	CmCA	tdbR00000478
	aTrm56	HVO_1173	56	Cm	All known tRNAs		
S. cerevisiae	Trm13	YOL125w	4	Cm	Gly	GCC	tdbR00000129
				Cm	Pro	U*GG	tdbR00000323-4
				Am	His	GUG	tdbR00000145
	Trm3	YDL112w	18	Gm	Leu	CAA	tdbR00000249
					Leu	UAG	tdbR00000250
					Leu	UAA	tdbR00000251
					Trp	CmCA	tdbR00000494
					Ser	IGA	tdbR00000407-8
					Ser	U*GA	tdbR00000406
					His	GUG	tdbR00000145
					Tyr	GΨA	tdbR00000555
	Trm7/Trm732	YBR061C/YMR259c	32	Cm	Trp	CmCA	tdbR00000494
					Phe	GmAA	tdbR00000083-4

					Leu	UAA	tdbR00000251
	Trm7/Trm734	YBR061C/Rtt10	34	Cm	Trp	CmCA	tdbR00000494
				Gm	Phe	GmAA	tdbR00000083-4
				cmnm^{5}Um	Lys	cmnm^{5}UmUU	tdbR00000193
	Trm44	YPL030w	44	Um	Ser	IGA	tdbR00000407-8
					Ser	U*GA	tdbR00000406
H. sapiens	hTrm13	CCDC76	4	Cm	Gly	CCC	tdbR00000135
				Um	Gly	GCC	tdbR00000136
	hTrm3	TARBP1	18	Gm	Asn	NUG	tdbR00000306
					Gln	U*UG	tdbR00000345
					Gln	U*UG	tdbR00000346
					Ser	UGA	tdbR00000428
	hTrm7	FTSJ1	32	Cm	Phe	GmAA	tdbR00000103
					Asn	NUG	tdbR00000306
					Gln	U*UG	tdbR00000345
					Gln	U*UG	tdbR00000346
				Um	Ala	IGC	tdbR00000016
					Ala	IGC	tdbR00000017
			34	Cm	Met	CmAU	tdbR00000289
				Gm	Phe	GmAA	tdbR00000103
	unknown		39	ψm	Gly	CCC	tdbR00000135
	hTrm44	METTL19	44	Um	Leu	NAA	tdbR00000269
					Ser	UGA	tdbR00000428
	unknown		54	m^{5}Um	Glu	CUC	tdbR00000062

U* - unknown modified uridine, Q – queuosine, I – inosine, Ψ – pseudouridine, N – unknown residue, other modified nucleotides abbreviated according to their conventional names, tRNA reference numbers are from tRNAdb.

Recognition of tRNA substrates by TrmH was extensively investigated and the enzymatic mechanism elucidated by a structure-based site-directed mutagenesis study[140]. The TrmH protein from the thermophilic bacterium *Aquifex aeolicus* has been crystallized and its structure studied at high resolution[300].

Two other important sites of 2'-O-methylation in *E. coli* tRNA are located in the anticodon loop, at positions 32 and 34. Formation of Cm32 and Um32 tRNA$^{Ser}_1$ and tRNA$^{Gln}_2$ is catalyzed by YfhQ (renamed now to TrmJ or TrMet(Xm32))[146] which belongs to the SPOUT family, in contrast to eukaryotic TRM7 from the RFM group (see below)[147].

Finally, the enzyme encoded by yibK, named TrmL, is a part of a more complex pathway[151] designed for biosynthesis of a complex, multistep modification of the nucleotide at wobble position 34 in two leucyl isoacceptor tRNAs. Depending on tRNA species, TrmL catalyzes formation of Cm34 or Um34 in tRNA^{Leu}(CmAA) and tRNA^{Leu}(cmnm^{5}UmAA)[150]. TrmL methylates pyrimidines, but not purine residues at the wobble position, and the 2'-O-methylation depends on prior N^6-isopentenyladenosine modification at position 37.

In other bacteria, enzymatic tRNA modification systems targeting the ribose-2'OH are only partially characterized. Recently, a tRNA MTase TrmJ in *Pseudomonas*

aeruginosa PA14 was found to be involved in resistance to oxidative stress. TrmJ catalyzes formation of Cm34, Um34, and, in contrast to *E. coli*, an additional Am34 in $tRNA^{Pro}$(GGG)[149].

Archaeal tRNA 2’-O-methylation

Archaeal tRNA 2’-O-methylation is only partially characterized up to now. In contrast to Bacteria, where only proteinaceous enzymes modify tRNAs, archaeal organisms use both stand-alone protein enzymes and C/D snoRNA guided machinery (Figure 1 and Table 1), sometimes with overlapping substrate and site specificity[153].

Methylation of the universally conserved methylated position at position 32 in the tRNA anticodon in Archaea *S. acidocaldarius* is catalyzed by a homolog of the bacterial TrmJ. However, while all four canonical nucleosides are substrates of the *E. coli* enzyme, the archaeal TrmJ can only methylate the ribose of a cytidine[147].

In contrast to Eubacteria, methylation of the wobble residue C34 of *H. volcanii* elongator $tRNA^{Met}$ is guided by a box C/D RNA targeting the tRNA intron-containing precursor[155]. 2'-O-methylation of both nucleotides C34 and U39 in *H. volcanii* pre-$tRNA^{Trp}$ also requires the box C/D ribonucleoprotein (RNP) machinery. Methylation occurs sequentially with the C/D-box sRNA guide in *trans* which would also eliminate the obligatory refolding of the pre-tRNA before splicing[156]. Most euryarchaeal $tRNA^{Trp}$ genes require a box C/D guide RNA within their intron for specifying methylation at these two sites[154].

The archaeal-specific modification Cm56 is formed by the SPOUT S-adenosylmethionine (AdoMet)-dependent MTase aTrm56, which is found only in archaea[72, 73]. aTrm56 is present in almost all the archaeal genomes sequenced up to now, except in the crenarchaeon *P. aerophilum* in which the tRNA Cm56 2'-O-methylation is provided by a C/D sRNP[154].

Eukaryotic tRNA 2’-O-methylation

In yeast and human tRNAs 2’-O-methylations occurs at five rather conserved positions and two minor sites, the best conserved locations being Nm4, Gm18, Cm32/Nm34 and Um44 (Table 1).

Formation of the highly conserved Gm18 in *Saccharomyces cerevisiae* tRNAs is catalyzed by the enzyme encoded by ORF YDL112w (now renamed “Trm3”). Genetic depletion of YDL112w resulted in the complete absence of Gm18 in all tRNAs that naturally contain this modification. Other tRNA ribose methylations and the complex pattern of rRNA ribose methylations were not affected[157].

Anticodon 2’-O-modifications both at positions 32 (Cm32) and 34 (Nm34) in yeast tRNAs require a unique MTase Trm7 (encoded by YBR061C)[159], which also tightly interacts with ORF YMR259c (now named Trm732) for 2'-O-methylation of C32, and with Rtt10 (named Trm734) for 2'-O-methylation of N34 in substrate tRNAs[78]. Moreover, the presence of Trm7-catalyzed modifications in tRNA anticodon loop is required for efficient modification of m^1G at position 37, in the biosynthesis pathway leading to yW (wyebutosine)[77, 78]. Activities towards Cm32 and Nm34 seem to be

independent, since a specific mutation in Trm7 selectively abolishes formation of Gm34 only[160].

2’-O-methylated Um44 is frequent among eukaryotic cytoplasmic tRNAs bearing a long variable loop. Yeast ORF YPL030w (now renamed to Trm44) was identified by screening for Um44:MTase activity of yeast genomic library of affinity purified proteins. Consistent with the conservation of Um44 in eukaryotic tRNAs, Trm44 is found in metazoans and fungi, but surprisingly, absent in plants[75].

The last 2'-O-methylation site in yeast tRNAs was found at position 4 and it is the only 2'-O-methylation that occurs in the middle of a duplex region in tRNA. It was shown that *S. cerevisiae* ORF YOL125w (renamed to Trm13) is required and sufficient for this modification[76].

The tRNA 2’-O-methylation profile in vertebrates is somehow similar to the one observed for lower eukaryotes like yeasts, with conserved sites at position Cm/Um4, Gm18, Cm/Um/Ψm32, Cm/Gm34 and Um44. One notable exception is the presence of Tm(m^5Um)54 in the TΨ-loop, this nucleotide occasionally replaces the universally conserved T54[79]. Another minor site is Ψm39 (Figure 1 and Table 1).

In addition to common simple 2’-O-methylated residues, tRNAs from higher eukaryotes contain other more complex modified residues, where base modification is coupled with ribose methylation. For example, 2'-O-methyl-5-formyl modification of cytidine (f^5Cm) was found at position 34 in bovine $tRNA^{Leu}$ and confirmed for rabbit and lamb. Thus f^5Cm could be a general feature of cytoplasmic $tRNAs^{Leu}$ in mammals[81].

Studies of higher eukaryotic RNA 2’-O-methylation enzymes have a long-standing history[301]. Already in 80s, the enzymatic activity acting at Gm34 in $tRNA^{Phe}$ was detected using micro-injection into *Xenopus laevis* oocytes[158]. However, systematic identification of the corresponding enzymes has started only recently.

The only higher eukaryotic tRNA-specific 2’-O-MTase characterized up to now is a human homolog of yeast Trm7. Mutations in human FTSJ1, the Trm7 homolog, are present in two genetically independent cell lines of patients and cause nonsyndromic X-linked intellectual disability (NSXLID). Cells with loss-of-function FTSJ1 mutations nearly completely lack Cm32 as well as Gm34, moreover, an NSXLID patient with a novel FTSJ1-p.A26P missense allele specifically lacked Gm34, while levels of Cm32 and o^2yW37 remained normal[160].

A homology search using yeast tRNA:MTases as probes revealed the presence of predicted homologs for all four known yeast tRNA:2’-O-MTases[162](see Table 1), however, activity of most of these proteins was not experimentally confirmed up to now.

Functions of tRNA 2’-O-methylation in tRNA stability and immunostimulation

Ribose 2’-O-methylation has been known for a long time to affect stability and physico-chemical properties of RNA chain. 2'-O-methylation of pyrimidines stabilizes the C3'-*endo* sugar conformation (and thus A-type RNA helix) primarily due to the

steric repulsion among the 2-carbonyl group, the 2'-O-methyl group and the 3'-phosphate group in the C2'-*endo* form[302].

Archaea-specific hypermodified residues N4-acetyl-2'-O-methylcytidine (ac^4Cm), 5-methyl-2'-O-methylcytidine (m^5Cm) and N2-dimethyl-2'-O-methylguanosine ($m^2{}_2Gm$) were found to stabilize the C3'-*endo* form and therefore cause "conformational rigidity" as attested by proton NMR spectroscopy. In particular, the ac^4Cm was found to be extremely rigid due to additive effects of the N^4-acetylation and 2'-O-methylation[204].

In addition to purely structural roles and general protection of RNA molecules from alkaline hydrolysis, tRNA 2’-O-methylation has an important functional role in fine-tuning of bacterial co-habitations with human (and other mammalian) organisms. It was noticed a long time ago that preparations of *S. typhimurium* tRNA and rRNA provide RNA-dependent immunogenic activity[303]. More recently, it was clearly demonstrated that immunostimulation by tRNA is inhibited by tRNA:Gm18-MTase activity, and even by the single methyl group on the 2'-oxygen of Gm18 as a natural modification in native tRNA[235, 304]. This might explain differential immunostimulatory effects of different bacterial strains and it was discussed that tRNA Gm18 modification might be used as immune evasion mechanism. Of note, Gm18-modified tRNA acted as TLR7 antagonist and blocked activation by RNA from plasmacytoid dendritic cells[235, 303].

In addition, Gm18 also potently inhibited TLR7-independent human monocyte activation by RNA derived from a variety of bacterial strains. Antagonizing effects of Gm18-modified RNA are due to competition with stimulatory RNA for receptor binding. The effect of immunostimulation/suppression mediated by Gm18 in *E. coli* tRNA was observed for human and murine innate immune receptors TLR7/8 whereas the mouse specific receptor TLR13 was insensitive to tRNA Gm18[236], but inhibited by N^6-methylation of rRNA (A2085 in *S. aureus*/A2058 *E. coli*)[305].

Analytical methods for detection of 2’-O-methylation in RNA

Despite the growing interest in detection and quantification of 2’-O-methylated residues, no rapid and high-throughput analytical technologies have been applied to tRNAs, mostly due to their small size, very stable secondary and tertiary structure, and an important proportion of other modified RNA residues. Reverse-transcription (RT) at low dNTP concentrations is not well suitable for tRNAs, since natural RT-pauses are frequent and RT-arresting modified residues prevent generation of long primer extension products[306, 307]. High-Pressure Liquid Chromatography (HPLC) and HPLC coupled with mass-spectrometry (MS) related techniques require purification of tRNA species to near homogeneity, thus limiting their application to rather global analysis of RNA modification profile[308–310]. Other approaches were also proposed, but did not become popular in the field[311].

Recently, high-throughput techniques have been developed for analysis of a small subset of RNA modifications, namely m^6A, m^1A, pseudouridine and m^5C[312–316]. These techniques allow transcriptome-wide screening, but quantification of modification is

uncertain and, when possible, requires complex calibration procedures. Only one type of modifications in RNA can be detected and precisely quantified by NGS, due to induced resistance of phosphodiester bond to cleavage in the presence of neighboring 2'-O-me group[63, 64, 107]. Thus, measurement of protection against nucleolytic cleavage provides an excellent and quantitative way to measure 2'-O-modification.

The aim of the present study was to adapt the previously published RiboMethSeq protocol[64] to 2'-O-methylation analysis of tRNAs. The main difficulties are related to the short length of most tRNA species, their stable 2D and 3D RNA structures, the existence of very similar closely related tRNA isoacceptors, and the presence of numerous other modified nucleotides affecting both efficiency of RNA cleavage as well as primer extension by RT.

2. Results and Discussion

We first evaluated the results from application of the standard RiboMethSeq protocol (previously adapted to rRNA) for comparatively short tRNA sequences from *E. coli* and *S. cerevisiae*. The RiboMethSeq protocol[64] is based on high-throughput analysis of an alkaline RNA digest and is able to precisely map resistant phosphodiester bonds, with such sites mostly corresponding to natural 2'-O-methylated 5'-adjacent riboses. However, spontaneous structure- or sequence-dependent resistance to alkaline hydrolysis may generate false-positive hits in the analysis.

One essential step of the RiboMethSeq protocol consists in the generation of a suitable library of RNA fragments after alkaline treatment. Short and structured RNA species (like tRNAs) are comparatively resistant to alkaline hydrolysis and their fragmentation conditions have to be optimized to get sufficient amounts of short 20-40 nt fragments. Preliminary tests indicated that hydrolysis times of 6-10 min were appropriate for *E. coli* tRNAs, while a somewhat longer treatment of 8-12 min was required for *S. cerevisiae* tRNA fraction (Figure 2). These tRNA fragments were converted to libraries and subjected to single-read Illumina sequencing for 50 nt.

With standard treatment parameters, the RiboMethSeq procedure uses reads of 17 nt minimum in length, which are aligned to the reference sequence with End-to-End option. Only 5'-end extremities are used for analysis with calculation interval of 6 neighboring nucleotides. Application of these parameters for tRNA 2'-methylation analysis demonstrated, in principle, detection of modified nucleotides, but also showed the necessity for a careful optimization of all treatment steps of the analysis pipeline.

The necessity for optimization is related to short and quite similar tDNA sequences used for alignment, to very irregular alkaline cleavage profile of these small and strongly structured RNA molecules, and certainly to the presence of numerous other modified nucleotides with RT-arresting properties, decreasing even further the number of available sequencing reads at certain tDNA regions.

Taking into account all these difficulties, we proceeded for optimization of all treatment steps (Figure 3), the most important concerning the used read length, selection of only uniquely aligned reads, simultaneous usage of 5'- and 3'-end

information, as well as calculation of scores for shorter neighboring regions, and finally a careful optimization of the reference tDNA dataset. All these optimization steps are discussed below.

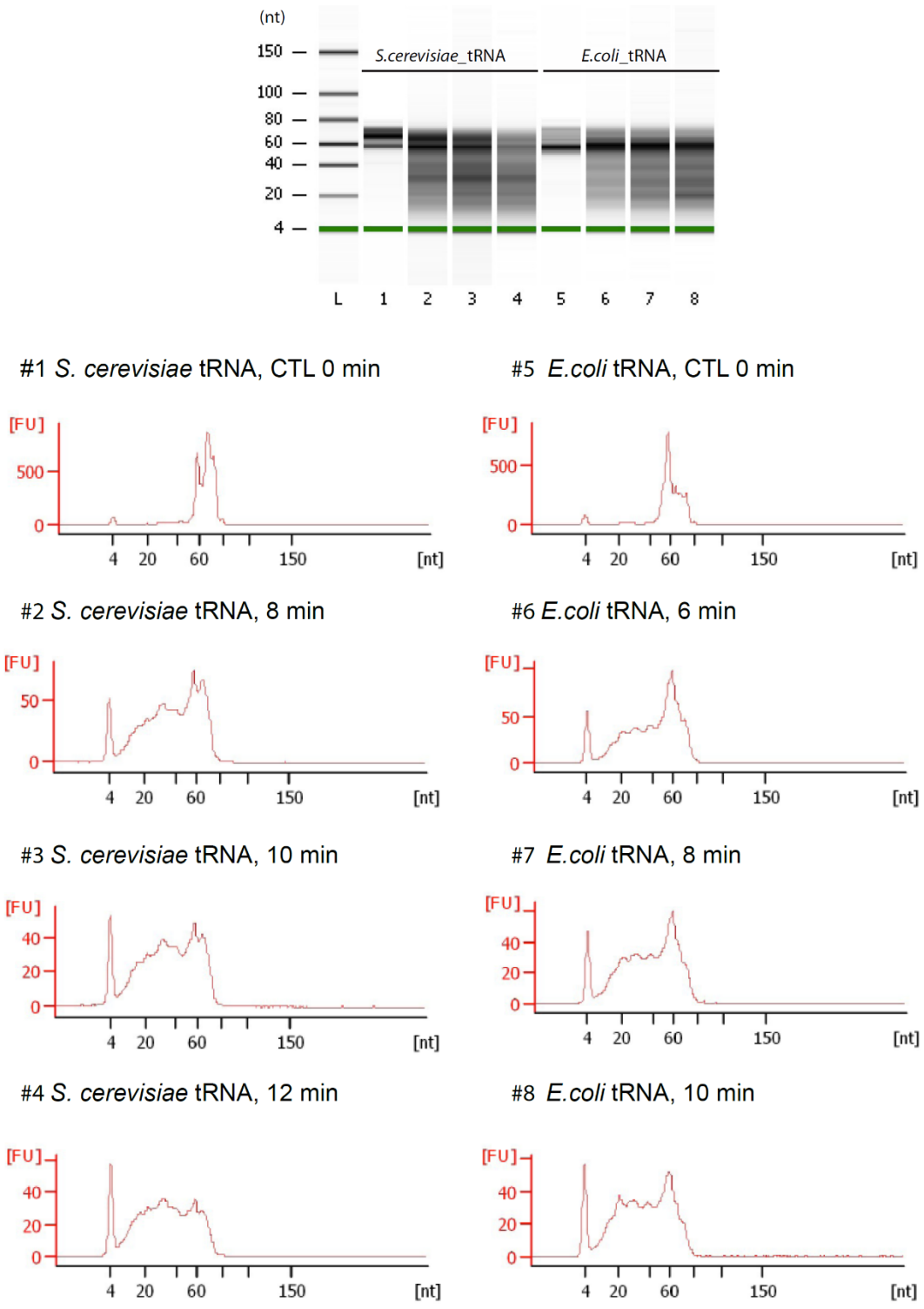


Figure 2 Fragmentation profiles for *E. coli* and *S. cerevisiae* total tRNA fractions. Alkaline hydrolysis was performed for 6 to 12 min (as indicated on the traces). Fragments size was analyzed by capillary electrophoresis on Bioanalyzer 2100 (Agilent) Small RNA Chip.

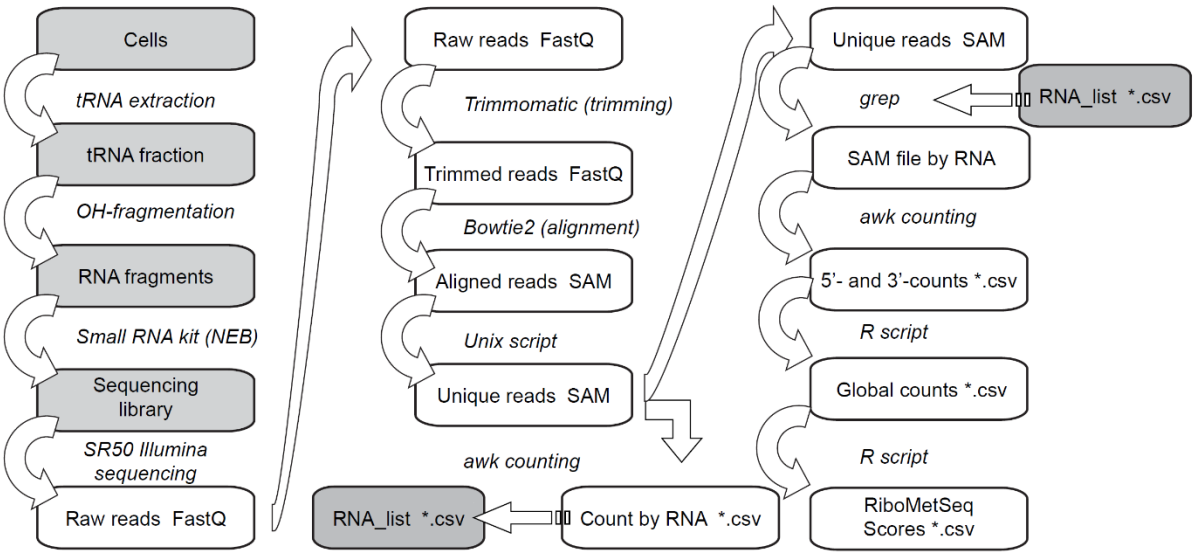


Figure 3. Optimized treatment pipeline for tRNA 2'-O-methylation analysis by high-throughput sequencing.

Construction of the reference tDNA dataset

Initial tDNA datasets for alignment were constructed from sequences of modified tRNA species deposited in tRNAdb (http://trna.bioinf.uni-leipzig.de), converted to DNA sequences in Fasta format. Inspection of the sequencing reads alignment to this tDNA reference revealed that, even at a minimum length of 17 nt, a significant number of reads mapped to multiple locations in tRNAs (Figure 4). This was mostly due to almost identical tDNA species present in dataset and also to the presence of highly repetitive regions. In order to obtain better certitude in read mapping, we first decided to reduce and simplify the tDNA reference datasets for analysis.

In order to identify closely related tDNA sequences, the Levenshtein distance was calculated for complete tDNA dataset for known modified tRNA species, obtained from tRNAdb.

Distances between tDNA sequences are given in Figure S1, calculation was performed for the dataset before and after collapsing of closely related sequences. tDNA sequences with Levenshtein distance <4 (corresponds to maxi 3 nucleotide replacements) were collapsed into one unique sequence and an undefined nucleotide "N" was placed at the divergent positions. Other closely related pairs were retained in the dataset. At the same time, observed sequencing non-conformities and errors were corrected and instances of the A-derived, but C-basepairing modified nucleotide inosine (I) present in some tRNA species, were replaced by a G in the reference tDNA dataset. The same strategy was applied to *E. coli* and *S. cerevisiae* tDNA datasets. Details of the adaptation of both tRNA datasets are given in Table S1 and S2.

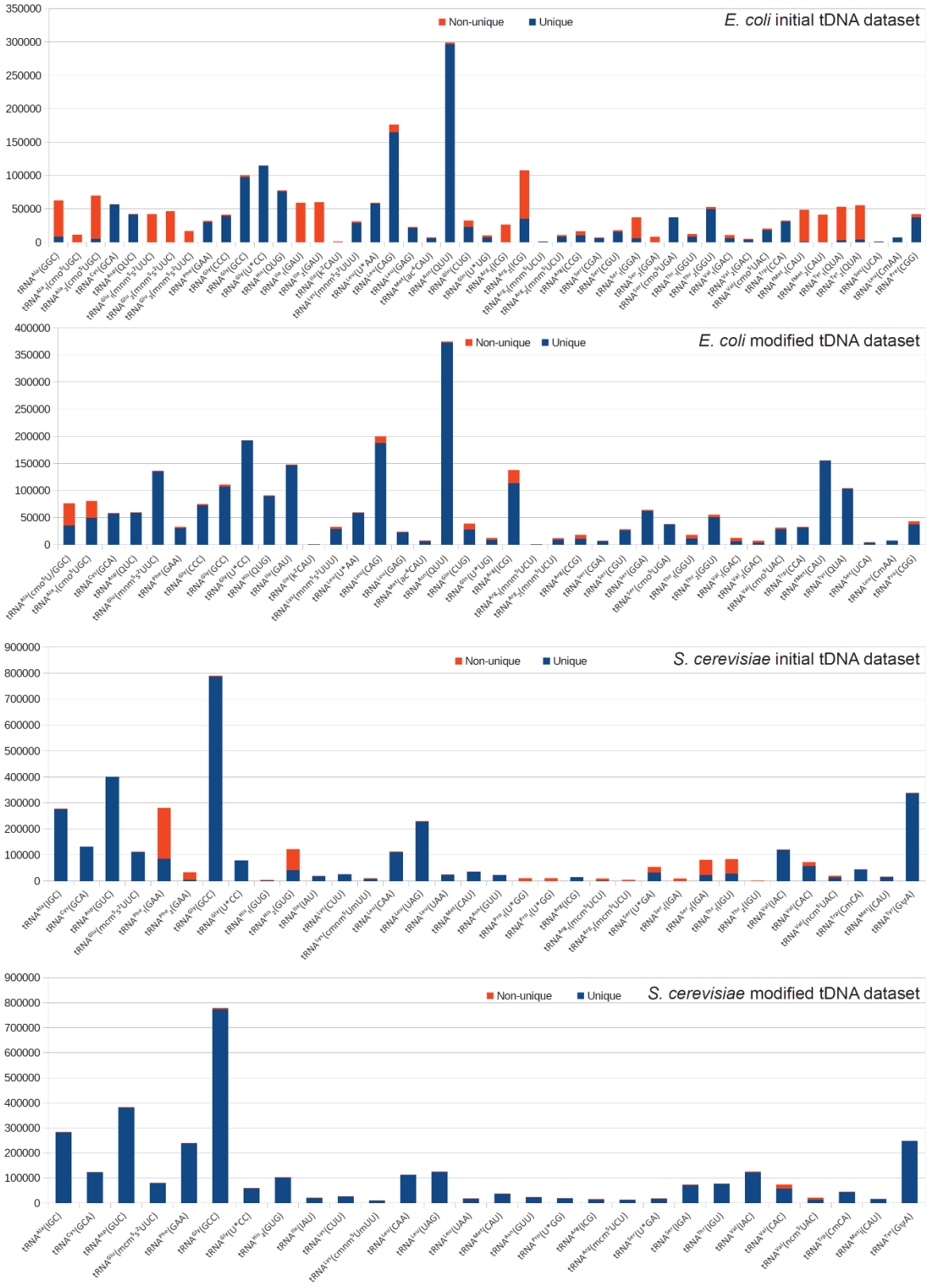


Figure 4. Proportion of non-uniquely mapped reads with initial and collapsed (optimized) tDNA datasets for each tRNA species. For both datasets, number of uniquely and non-uniquely mapped reads was calculated for each tRNA species and plotted as stacked bars.

Alignment and analysis

For alignment to optimized the tDNA dataset, raw read trimming was performed to keep all sequences >8nt. Of note, this length, while unsuitable for large genomic alignments, is still appropriate to a limited size tDNA “genome”. In order to gain additional information on both, the 5’- and the 3’-end positions for every aligned read, only trimmed (with adapter removed) sequences were retained for subsequent alignment. This reduced the number of usable reads, but still provided some gain in information, since both 5’- and 3’-ends locations were used.

Bowtie2 alignment parameters were adjusted for short input sequences (L=10) in End-to-End mode. In order to avoid contamination by ambiguously (multiply) mapped sequencing reads, only uniquely aligned sequences were taken for further analysis. Details for number of retained reads at every step are given in Table S3. Positions of 5’- and 3’-ends of every read were extracted after conversion of aligned *.sam file to bed format with bedtools.

Positions of 5’- and 3’-ends were cumulated together to obtain global end-coverage profile and this data was used to calculate RiboMethSeq scores (ScoreMEAN (derived from ScoreMAX), and scoresA, B and MethScore(C)) as described previously[63, 64]. Since the length of tRNA sequences is rather limited and the cleavage profile of tRNAs is highly irregular, only 2 neighboring nucleotides on each side were taken into account (instead of 6 in the standard RiboMethSeq approach).

As a result of this optimization, over 95% of trimmed reads were unambiguously mapped to tDNA reference sequences and a better homogeneity of the coverage was obtained. All tRNAs (except one species in *E. coli* and one in *S. cerevisiae*) received sufficient coverage and further analysis of 2’-O-methylation could be performed.

Profiles for tRNAs from WT and ΔTrmH *E. coli* mutant strains

Analysis of *E. coli* tRNA 2’-O-methylation was performed in parallel for wild-type (WT) strain and for a strain with deletion of the nonessential *trmH* gene responsible for Gm18 formation in bacterial tRNAs. *E. coli* tRNAs were reported to contain 2’-O-methylated residues at three different locations (Gm18, Cm32 and Cm/Um/cmnm^{5}Um34, see Introduction). As expected, RiboMethSeq efficiently detected alkali-resistant positions, which mostly correspond to 2’-O-methylations, to other modified RNA residues affecting cleavage or phosphorylation/ligation efficiency for adapter, or to naturally insensitive non-modified residues. The present goal was to provide a set of potentially modified positions to be verified by independent approaches, and also to gain relative quantification of the methylation level for known 2’-O-methylated nucleotides.

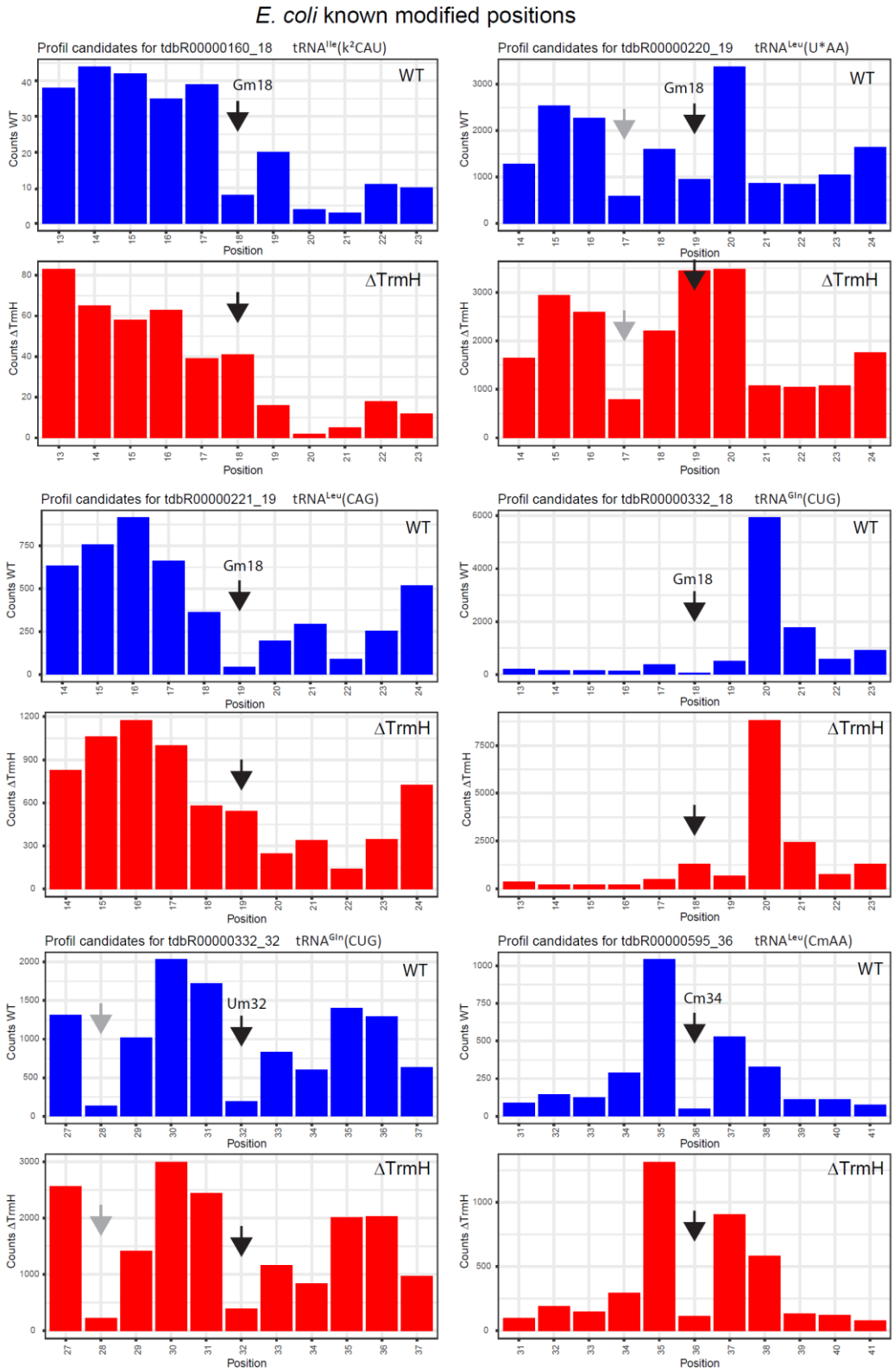


Figure 5. Representative cleavage profiles for 6 *E. coli* tRNA species for WT and ΔTrmH strains. Cumulated number of 5'- and 3'-end extremities at every position is traced as vertical bars. All profiles are centered for known 2'-O-methylated position shown by black arrow, false-positive alkali-resistant positions are indicated by gray arrows.

Representative cleavage profiles for selection of well-detectable 2'-O-methylated sites are shown in Figure 5, and the corresponding scores are listed in Table 2. All previously reported 2'-O-methylated sites gave rather high values of ScoreMEAN, while ScoreA[63] varied considerably. Comparison of both scores obtained for WT and ΔTrmH strains demonstrated that the value of the score considerably decreased for all TrmH-dependent 2'-O-methylations (Gm18), but remained very stable for all 2'-O-methylated residues at positions 32 and 34, catalyzed by TrmJ and TrmL (Figure 1 and Table 1). Inspection of the corresponding cleavage profiles confirmed that cleavage efficiency at all TrmH-catalyzed positions drastically increased in the ΔTrmH strain. This clearly demonstrated that tRNA 2'-O-methylation can be detected and even precisely quantified using RiboMethSeq, even when the RNA cleavage profile is highly irregular. However, multiple false-positive hits were also detected in various tRNAs (Figure S2), some of them in proximity to highly accessible tRNA nucleotides or to some of the numerous modified residues in tRNA, namely pseudouridine or m^7G46.

Table 2 Calculated Scores mean and MethScore for known 2'-O-methylated positions in E. coli and S. cerevisiae tRNAs extracted from WT and the strain lacking tRNA:Gm18-MTase (TrmH or TRM3)

Escherichia coli

				WT *E.coli*		ΔTrmH *E.coli*		Diffe-rence	
tdbR number	tRNA	Modifi-cation	pos*	Mean	Meth-Score	Mean	Meth-Score	Mean	Meth-Score
tdbR00000160	$tRNA^{Ile}_{(k^2CAU)}$	Gm18	18	0,92	0,68	0,29	-0,37	**0,63**	**1,05**
tdbR00000332	$tRNA^{Gln}_{(CUG)}$	Gm18	18	0,97	0,85	0,94	0,79	0,03	0,06
tdbR00000333	$tRNA^{Gln}_{(U*UG)}$	Gm18	18	0,81	0,7	0,76	0,67	0,05	0,04
tdbR00000390	$tRNA^{Ser}_{(CGA)}$	Gm18	18	0,6	0,25	0,24	-0,3	**0,35**	**0,55**
tdbR00000392-393	$tRNA^{Ser}_{(GGA)}$	Gm18	18	0,86	0,76	0,26	0,32	**0,6**	**0,44**
tdbR00000394	$tRNA^{Ser}_{(cmo^5UGA)}$	Gm18	18	0,98	0,88	0,94	0,8	0,04	0,08
tdbR00000545-546	$tRNA^{Tyr}_{(QUA)}$	Gm18	18	0,93	0,84	0,35	0,17	**0,58**	**0,67**
tdbR00000220	$tRNA^{Leu}_{(cmnm^5UmAA)}$	Gm18	19	0,87	0,42	0,46	-0,78	**0,42**	**1,21**
tdbR00000221	$tRNA^{Leu}_{(CAG)}$	Gm18	19	0,97	0,89	0,35	-0,02	**0,62**	**0,9**
tdbR00000222	$tRNA^{Leu}_{(GAG)}$	Gm18	19	0,99	0,95	0,25	-0,48	**0,74**	**1,44**
tdbR00000276	$tRNA^{Met}_{(ac^4CAU)}$	Gm18	19	0,58	0,42	0,75	0,44	-0,17	-0,02
tdbR00000595	$tRNA^{Leu}_{(CmAA)}$	Gm18	19	0,93	0,9	0,73	0,67	0,19	0,22
tdbR00000332	$tRNA^{Gln}_{(CUG)}$	Um32	32	0,97	0,85	0,94	0,79	0,03	0,06
tdbR00000333	$tRNA^{Gln}_{(U*UG)}$	Um32	32	0,81	0,7	0,76	0,67	0,05	0,04
tdbR00000484	$tRNA^{Trp}_{(CCA)}$	Cm32	33	0,99	0,92	0,99	0,91	0	0,01
tdbR00000394	$tRNA^{Ser}_{(cmo^5UGA)}$	Cm32	34	0,98	0,88	0,94	0,8	0,04	0,08
tdbR00000510-511	$tRNA^{Met}_{i(CAU)}$	Cm32	34	0,97	0,84	0,97	0,85	0	-0,02
tdbR00000623	$tRNA^{Pro}_{(CGG)}$	Um32	34	0,98	0,93	0,97	0,92	0	0,01
tdbR00000595	$tRNA^{Leu}_{(CmAA)}$	Cm34	36	1	0,94	0,99	0,88	0,01	0,05
tdbR00000220	$tRNA^{Leu}_{(cmnm^5UmAA)}$	$cmnm^5Um34$	36	1	0,94	0,98	0,81	0,02	0,13

Saccharomyces cerevisiae

				WT *S.cerevisiae*		ΔTRM3 *S.cerevisiae*		Difference	
tdbR number	tRNA	Modification	pos*	Mean	Meth-Score	Mean	Meth-Score	Mean	Meth-Score
tdbR00000129	$tRNA^{Gly}$(GCC)	Cm4	5	0,89	0,63	0,93	0,76	-0,04	-0,13
tdbR00000323-4	$tRNA^{Pro}$(U*GG)	Cm4	5	0,95	0,83	0,95	0,81	0,01	0,02
tdbR00000145	$tRNA^{His}$(GUG)	Am5	6	0,98	0,89	0,96	0,85	0,01	0,03
tdbR00000249	$tRNA^{Leu}$(CAA)	Gm18	18	0,99	0,89	0,84	0,03	0,15	**0,87**
tdbR00000250	$tRNA^{Leu}$(UAG)	Gm18	18	0,93	0,84	-0,35	-0,46	**1,28**	**1,29**
tdbR00000251	$tRNA^{Leu}$(UAA)	Gm18	18	0,93	0,86	-0,16	-0,14	**1,10**	**1,01**
tdbR00000406	$tRNA^{Ser}$(U*GA)	Gm18	18	0,91	0,85	0,75	0,57	**0,15**	**0,29**
tdbR00000407-8	$tRNA^{Ser}$(IGA)	Gm18	18	0,95	0,81	-0,35	-0,80	**1,30**	**1,61**
tdbR00000494	$tRNA^{Trp}$(CmCA)	Gm18	18	0,93	0,81	0,26	-0,05	**0,67**	**0,87**
tdbR00000145	$tRNA^{His}$(GUG)	Gm18	19	0,91	0,80	-0,98	-0,84	**1,89**	**1,64**
tdbR00000555	$tRNA^{Tyr}$(GψA)	Gm18	19	0,96	0,92	0,09	0,13	**0,87**	**0,79**
tdbR00000494	$tRNA^{Trp}$(CmCA)	Cm32	32	0,99	0,86	0,99	0,88	0,00	-0,02
tdbR00000083-4	$tRNA^{Phe}$(GmAA)	Cm32	33	0,98	0,79	0,98	0,81	0,00	-0,03
tdbR00000251	$tRNA^{Leu}$(UAA)	Cm32	34	0,92	0,55	0,95	0,69	-0,03	-0,14
tdbR00000494	$tRNA^{Trp}$(CmCA)	Cm34	34	0,99	0,87	0,99	0,89	0,00	-0,03
tdbR00000083-4	$tRNA^{Phe}$(GmAA)	Gm34	35	0,96	0,74	0,93	0,63	0,03	0,10
tdbR00000193	$tRNA^{Lys}$($cmnm^5UmUU$)	$cmnm^5Um34$	35	0,64	0,30	0,58	0,26	0,06	0,04
tdbR00000406	$tRNA^{Ser}$(U*GA)	Um44	45	0,93	0,86	0,93	0,83	-0,01	0,03
tdbR00000407-8	$tRNA^{Ser}$(IGA)	Um44	45	0,90	0,78	0,86	0,71	0,04	0,07

tdbR identifiers are taken from tRNAdb, when collapsed sequence is used, both numbers are indicated
* position corresponds to the real position of the signal in the sequence of tRNA, thus it appears at pos +1 to 2'-O-methylated residue. U* – unknown modified uridine, Q – queuosine, I – inosine, ψ – pseudouridine

E. coli $tRNA^{Met}$(ac^4CAU), tdbR00000276, (highlighted in grey in Table 2) was previously reported to contain at least partial Gm18 methylation. Analysis of the RiboMethSeq results for this tRNA demonstrated that rather low scores were obtained for potential Gm18 and, in addition, no difference was detected between WT and ΔTrmH strains. In order to check the real existence of Gm modification in this tRNA, it was purified to near homogeneity by hybridization with biotinylated DNA oligonucleotide and subjected to nucleotide analysis by HPLC-MS/MS (Figure S4a,b). As clearly seen in the figure, only trace amounts of Gm were detected, confirming the absence of Gm18 in *E. coli* $tRNA^{Met}$(ac^4CAU). Following up on our previous studies on the immunostimulatory effect of a Gm18 modification in *E. coli* tRNA[235] we tested the native tRNA for its effect on TLR7 mediated stimulation of interferon emission from peripheral blood mononuclear cells (PBMCs). Indeed, the native tRNA induced substantial release of alpha-interferon, which would be in contrast to a possible presence of Gm18, but is in excellent agreement with its absence (Figure S4c). These findings illustrate, that the mapping of ribose methylated residues in tRNA is accurate to a point that reports from the literature can be efficiently corrected.

Analysis of 2'-O-methylation in WT and ΔTRM3 yeast mutant strains

The optimized treatment pipeline described above was also applied for analysis of alkaline hydrolysis-resistant positions in yeast total tRNAs extracted from WT and ΔTrm3-strains. All 19 known positions of 2'-O-Me in yeast gave strong and consistent value of ScoreMEAN (median value 0.93, s.d. 0.077), as well as MethScore (median 0.83, s.d. 0.15). All scores for Gm18 demonstrated considerable decrease (average ScoreMEAN -0.037 and MethScore -0.1) in ΔTrm3 yeast strain compared to the WT cells. In contrast, calculated values of scores for other known 2'-O-methylated positions (Nm5-6, Nm32-34, Nm45) were not at all affected in ΔTrm3 strain (See Table 2). Representative cleavage profiles are given in Figure 6 and false-positive hits in Figure S3.

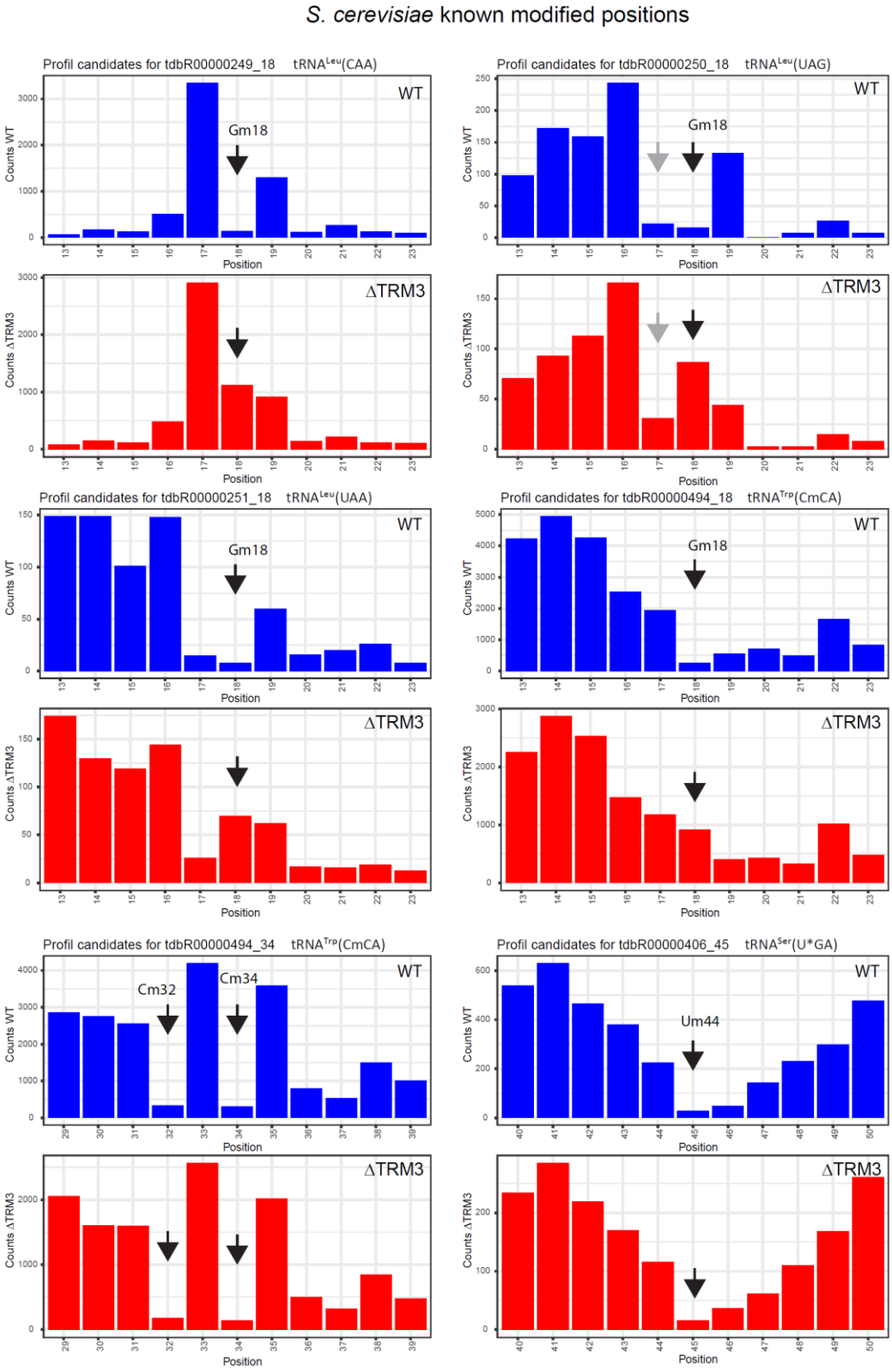


Figure 6. Representative cleavage profiles for 6 *S. cerevisiae* tRNA species for WT and ΔTRM3 strains. Cumulated number of 5'- and 3'-end extremities at every position is traced as vertical bars. All profiles are centered for known 2'-O-methylated position shown by black arrow, false-positive alkali-resistant positions are indicated by gray arrows.

Limitations

The results obtained in this study demonstrate that RiboMethSeq can be successfully applied not only to rRNA sequences in different species, but also to tRNAs, including also relative quantification of 2'-O-methylation levels in this type of RNA. However, rather important sequencing coverage is required for reliable analysis and the cleavage profiles for tRNAs remain extremely irregular and thus generate an important number of false-positive hits. Some of them correspond to other modified tRNA nucleotides, and independent approaches are required to distinguish real RNA modification and phosphodiester bonds that are alkaline resistant for other reasons, such as three-dimensional structure. The ScoreMEAN parameter provides rather good sensitivity (all known 2'-O-methylated positions have high ScoreMEAN value), but poor selectivity, while the opposite is true for ScoreA. Other proposed scoring schemes are even less discriminative and were not used here.

The use of cell lines or strains devoid of selected RNA-modification enzymes provided considerable support in the distinction between truly ribose-methylated positions from false-positive hits, since only signals corresponding to real modified nucleotides are affected, while signals for false-positives and other unaffected modified nucleotides remain generally stable. For a more precise quantification of RNA modification level, the calibration of RiboMethSeq signals with unmodified RNA species extracted from KO strain is strongly recommended.

Conclusion

The alkaline-hydrolysis based RiboMethSeq method (Figure 3) can be successfully applied to the analysis of known tRNA 2'-O-methylations and their dynamics or variations under different conditions. Quantification of the modification level is possible with careful calibration, obtained either with unmodified tRNAs extracted from a relevant KO strain or, if not available, with synthetic RNA transcript with identical sequence. The candidates for novel modified positions revealed by this high-throughput technique should be further validated by orthogonal approaches, either by physico-chemical methods like HPLC-MS(MS) analysis or by primer extension at low dNTP concentrations, approach specific for RNA 2'-O-methylations. In the case of Gm18 in *E. coli* $tRNA^{Met}(ac^4CAU)$, where we have exercised such a validation, the RiboMethSeq method proved accurate and led to correction of data from the literature. As this example has demonstrated, RiboMethSeq may be used for the correct prediction of immunosilencing properties of RNA molecules based on the functional effect of ribose methylation on the recognition by components of the innate immune system.

3. Materials and Methods

tRNA sources

S. cerevisiae parental strain (Y00000, strain BY4741; S288C isogenic yeast strain: MATa; his3Δ1; leu2Δ0; met15Δ0; ura3Δ0) and MTase mutant (Y03809, MATa; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; YDL112w::kanMX4) deficient for Trm3 (Gm18-2'-O-MTase) were obtained from Euroscarf, Johann Wolfgang Goethe-University Frankfurt and grown at 30°C in YPD medium (Carl Roth, Karlsruhe, Germany) to mid-log phase.

E. coli parental strain (BW25113; CGSC#7636; F-, Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ, rph-1, Δ(rhaD-rhaB)568, hsdR514) and trmH (Gm18-2'-O-MTase) deficient mutant (JW3626-1; CGSC#11693; F-, Δ(araD-araB)567, ΔlacZ4787(::rrnB-3), λ-, rph-1, ΔtrmH755::kan, Δ(rhaD-rhaB)568, hsdR514) were purchased from *E. coli* Genetic Stock Center CGSC, Yale University and grown to mid-log phase at 37°C in LB medium (AppliChem, Darmstadt, Germany). Knockout strains were grown with 200 µg/ml geneticin (G418) (Carl Roth, Karlsruhe, Germany) and 50 µg/ml kanamycin (AppliChem, Darmstadt, Germany), respectively.

Isolation of tRNA

For optimal amounts of RNA, bacteria and yeast were pelleted and frozen at -80°C overnight and thereafter lysed in 40 mg/ml lysozyme (Serva, Heidelberg, Germany) or 50 U/ml lyticase (Sigma Aldrich, St. Louis USA), respectively. Total yeast and bacterial RNA were isolated using TRIzol reagent (Thermo Fisher Scientific, Waltham, USA) according to the manufacturer's instructions. Total RNA preparations were separated by denaturating PAGE (10% acrylamide, 8 M urea); tRNA containing bands were cut out and shaken in 0.5 M sodium acetate overnight. tRNA was precipitated with ice cold ethanol and resuspended in nuclease-free water. The concentration of RNA was controlled via UV absorbance.

tRNA fragmentation and analysis

tRNA (about 100 ng) hydrolysis was performed in 50 mM bicarbonate buffer pH 9.2 for different times (6-10 min for *E. coli*) or (8-12 min for *S. cerevisiae*) at 95°C. The reaction was stopped by ethanol precipitation using 3 M Na-OAc, pH 5.2 and glycoblue as a carrier in liquid nitrogen. After centrifugation, the pellet was washed with 80% ethanol and resuspended in nuclease-free water. The sizes of generated RNA fragments were assessed by capillary electrophoresis using a Small RNA chip on Bioanalyzer 2100 (Agilent, USA).

Library preparation

RNA fragments without any gel-purification step were directly 3'-end dephosphorylated using 5 U of Antarctic Phosphatase (New England Biolabs, UK) for 30 min at 37°C. After inactivation of the phosphatase for 5 min at 70°C, RNA fragments were phosphorylated at the 5'-end using T4 PNK and 1 mM ATP for 1h at 37°C. End-

repaired RNA fragments were then purified using RNeasy MinElute Cleanup kit (QIAGEN, Germany) according to the manufacturer's recommendations except that the volume of 96% ethanol was adjusted for RNA binding. Elution was performed in 10 µl of nuclease-free water. RNA fragments were converted to library using NEBNext® Small RNA Library kit (ref#E7330S, New England Biolabs, UK or equivalent from Illumina, USA) following the manufacturer's instructions. DNA library quality was assessed using a High Sensitivity DNA chip on a Bioanalyzer 2100. Library quantification was done using a fluorometer (Qubit 2.0 fluorometer, Invitrogen, USA).

Library mix for sequencing was adjusted to obtain env 10-15 mln of raw reads for each library, this number was found to be sufficient for 5000-8000 nt length reference sequence (end-coverage ~1000x).

Sequencing

Library sequencing was performed using Illumina HiSeq 1000 sequencer in a single read mode for 50 nt with SBS kit v3. Primary analysis of sequencing quality was done with RTA 2.12 software, to insure >Q30 quality score for >95% of obtained sequences. Following SR50 sequencing run, demultiplexing was done with BclToFastq v2.4, reads not passing quality filter were removed.

Data analysis

Raw reads after demultiplexing were trimmed to remove the sequence of Illumina 3' adapter. Trimming was performed with Trimmomatic v0.32[317] with the following parameters: -phred33 /adapters/TruSeq3-SE.fa:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:15 MINLEN:8 AVGQUAL:30. Tests with a training dataset demonstrated that with stringency parameter =10 Trimmomatic removes the adapter with minimal size of 16 nt. Thus, only trimmed and adapter-free reads of <35 nt were taken for alignment. Alignment to the reference tDNA sequences was performed with bowtie 2 ver2.2.4[318] in End-to-End mode using the following parameter set: --no-1mm-upfront -D 15 -R 2 -N 0 -L 10 -i S,1,1.15. Preliminary tests demonstrated that the use of soft trimming during alignment (-- local settings) is not recommended, since 5'- and 3-end may be incorrectly determined. Selection of only uniquely mapped reads (single reported alignment position in the reference sequence) in the resulting *.sam file was done using values of NM and XS fields (grep -E "@|NM:" *.sam | grep -v "XS:"). Uniquely mapped reads were extracted from *sam file by RNA ID and converted to *.bed format using bedtools v2.25.0[319]. Positional counting of 5'-and 3'-ends of each reads was performed with awk Unix command. Further treatment steps were done in R environment (v3.0.1). In brief, 5'-end and 3'-end counts were merged together by RNA position and used for calculation of ScoreMEAN (derived from MAX Score described previously), as well as Scores A and B[63] and MethScore[64]. Scores were calculated for 2 neighboring nucleotides instead of 6 in standard RiboMethSeq procedure. Profiles of RNA cleavage at selected (candidate and previously known) positions were extracted and visually inspected.

Acknowledgments: This work was supported by an ANR HTRNAMod ANR-13-ISV8-0001 grant to Y.M. Part of this work was supported by grants of the German Research Foundation: Da592/5 to AD and by the DFG SPP1784, MH3397/12-1, MH3397/14-1 to M.H.

Author Contributions: “Y.M., A.D. and M.H. conceived and designed the experiments; V.M., I.F. and L.A. performed the experiments; F.P. and Y.M. analyzed the data; Y.M., V.M., A.D. and M.H. wrote the paper.”

Conflicts of Interest: “The authors declare no conflict of interest."

“The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results”.

Supplementary materials

Methods:

Isolation of tRNA^{Met} (*E. coli*)

E. coli tRNA^{Met} was isolated from total RNA (*E. coli* strain MC4100) by hybridization with a biotinylated DNA-oligonucleotide followed by immobilization on strepatividin-coated magnetic beads (Dynabeads® MyOne™ Streptavidin T1, Thermo Fisher Scientific, Waltham, MA, USA). The oligonucleotide was complementary to the full-length tRNA and carried a 5'-biotin, 3 additional adenosines at the 5'-end and a 3'-atto488-label for improved detection (sequence: 5'-AAATGGTGGCTACGACGGGATTCGAACCTGTGACCCCATCATTATGAGTGATGT GCTCTAACCAACTGAGCTACGTAGCC-3'). The hybridization step was performed in 5x SSC buffer (20x: 3 M sodium chloride, 300 mM trisodium citrate, pH 7.0) using 100 pmol oligonucleotide and 135 µg of total RNA per 25 µL Dynabeads®. Samples were denatured at 90 °C for 3 min and subsequently hybridized at 65 °C for 10 min followed by cooling down to room temperature. Dynabeads® MyOne Streptavidin T1 were washed 3 times using 1x Binding and Washing buffer (5 mM Tris-HCl pH 7.5, 0.5 M EDTA, 0.5 M sodium chloride) according to the manual and then equilibrated once in 5x SSC buffer before the hybridized samples were added. Immobilization of the hybrid was performed at 25 °C under shaking for 30 min. Subsequently, the supernatant containing non-target tRNAs was removed and the beads were washed once in 1x SSC buffer and 3 times in 0.1x SSC buffer. Finally, the beads were resuspended in MilliQ water and heated to 75 °C for 3 minutes in order to elute tRNA^{Met}.
In order to remove any remaining DNA-oligonucleotide, a DNase 1 (50 U/µL, Thermo Fisher Scientific) digestion step was performed at 37 °C in 1x DNase1 buffer (Thermo Fisher Scientific) for 2 hours. To exclude the presence of unspecific RNAs or degradation products, the eluted tRNA was then loaded onto a 10 % denaturing polyacrylamide gel (containing 8 M urea). For visualization of the tRNA band, SYBR®Gold (Thermo Fisher Scientific) was used for gel staining. The band corresponding to tRNA^{Met} was eluted from the gel, purified by ethanol precipitation and stored at -20 °C.

LC-MS/MS analysis of isolated tRNA^{Met} (*E. coli*)

Sample preparation: Prior to LC-MS/MS analysis, 100 ng isolated tRNA^{Met} were digested into nucleosides according to the following protocol: the RNA was incubated in presence of 1/10 volume of 10x nuclease P1 buffer (0.2 M NH_4OAc pH 5.0, $ZnCl_2$ 0.2 mM), 0.3 U nuclease P1 (Sigma Aldrich, Munich, Germany) and 0.1 U snake venom phosphodiesterase (Worthington, Lakewood, USA) at 37 °C for 2 h. Next, 1/10 volume of 10x fast alkaline phosphatase buffer (Thermo Fisher Scientific) and 1 U fast alkaline phosphatase (Thermo Fisher Schientific) were added, and the samples was incubated for additional 60 min at 37 °C. The digested RNA was diluted to 100 nM and stored at -20 °C. Calibration solutions of Gm and C were prepared in following concentrations: 1, 10 and 100 nM (Gm) and 1, 10 and 100 pmol (C).

HPLC parameters: The calibration solutions (injection volumes 2, 5 and 10 µL) and the digested tRNA sample (injection volume 5 µL) were analyzed on an Agilent 1260 HPLC series equipped with a diode array detector (DAD) and a triple quadrupole mass spectrometer (Agilent 6460). A Synergy Fusion RP column (4 µm particle size, 80 Å

pore size, 250 mm length, 2 mm inner diameter) from Phenomenex (Aschaffenburg, Germany) was used at 35 °C column temperature. The solvents consisted of 5 mM ammonium acetate buffer adjusted to pH 5.3 using acetic acid (solvent A) and pure acetonitrile (solvent B). The elution was performed at a flow rate of 0.35 mL/min using a linear gradient from 0 % to 8 % solvent B at 10 min, 40 % solvent B at 20 min and 0 % solvent B at 23 min. For additional 7 min, the column was rinsed with 100 % solvent A to restore the initial conditions.

MS/MS parameters: Prior to entering the mass spectrometer, the eluent from the column was measured photometrically at 254 nm by the DAD for quantification of C. The triple quadruple mass spectrometer, equipped with an electrospray ion source (Agilent Jet Stream), was run at the following ESI parameters: gas (N2) temperature 350°C, gas (N2) flow 8 L/min, nebulizer pressure 50 psi, sheath gas (N2) temperature 350°C, sheath gas (N2) flow 12 L/min and capillary voltage 3000 V. The MS was operated in the positive ion mode using Agilent MassHunter software. The mass transitions and QQQ parameters used for Gm analysis can be found in Table S4. Peak areas for both the UV peak of C and the mass peak of Gm were determined employing Agilent MassHunter Qualitative Analysis Software.

Table S4: MS/MS parameters for the analysis of Q

Modified nucleoside	**Precursor ion m/z**	**Product ion m/z**	**Fragmentor voltage [V]**	**Collision energy [eV]**	**Cell accelerator voltage [V]**	**Retention time RT [min]**	**ΔRT [min]**
Gm	298	152	72	5	2	12.4	1

Quantification: External calibration curves of Gm and C were obtained from analysing the calibration solutions and plotting the respective peak areas against the injected amount of nucleoside. Using the slopes of the linear fits, the amounts of Gm and C in the digested $tRNA^{Met}$ were determined, and via the number of C nucleosides in the tRNA, the amount of $tRNA^{Met}$ injected was accessible. Finally, the amount of Gm per $tRNA^{Met}$ molecule was calculated.

Stimulation of human PBMCs with *E. coli* tRNAs

PBMCs were isolated from heparinized blood of healthy donors by standard Ficoll-Hypaque density-gradient centrifugation (Bicoll 1.078 g/ml; Eberle et al., 2008). PBMCs were resuspended in complete medium prepared of RPMI 1640 (Biochrom) supplemented with heat-inactivated (1 h, 56°C) 2% autologous serum. Cells were plated at 4 × 105 cells/well in a 96-well flat-bottom plate. tRNA was encapsulated with DOTAP (N-[1-(2, 3-dioleoyloxy)propyl]-N, N, N-trimethylammonium-methylsulfate) (Carl Roth GmbH) at a ratio of 2,5 µL DOTAP per 1 µg of RNA in serum-free medium and incubated for 10 min at room temperature. PBMCs were stimulated in a humidified 5% CO2 atmosphere at 37°C for 16 h. As positive control, PBMCs were stimulated with 1 µM of the TLR9-specific stimulus CpG2216. Cell-free supernatant was analyzed for secretion of IFN-α using a sandwich ELISA (Affymetrix eBioscience).

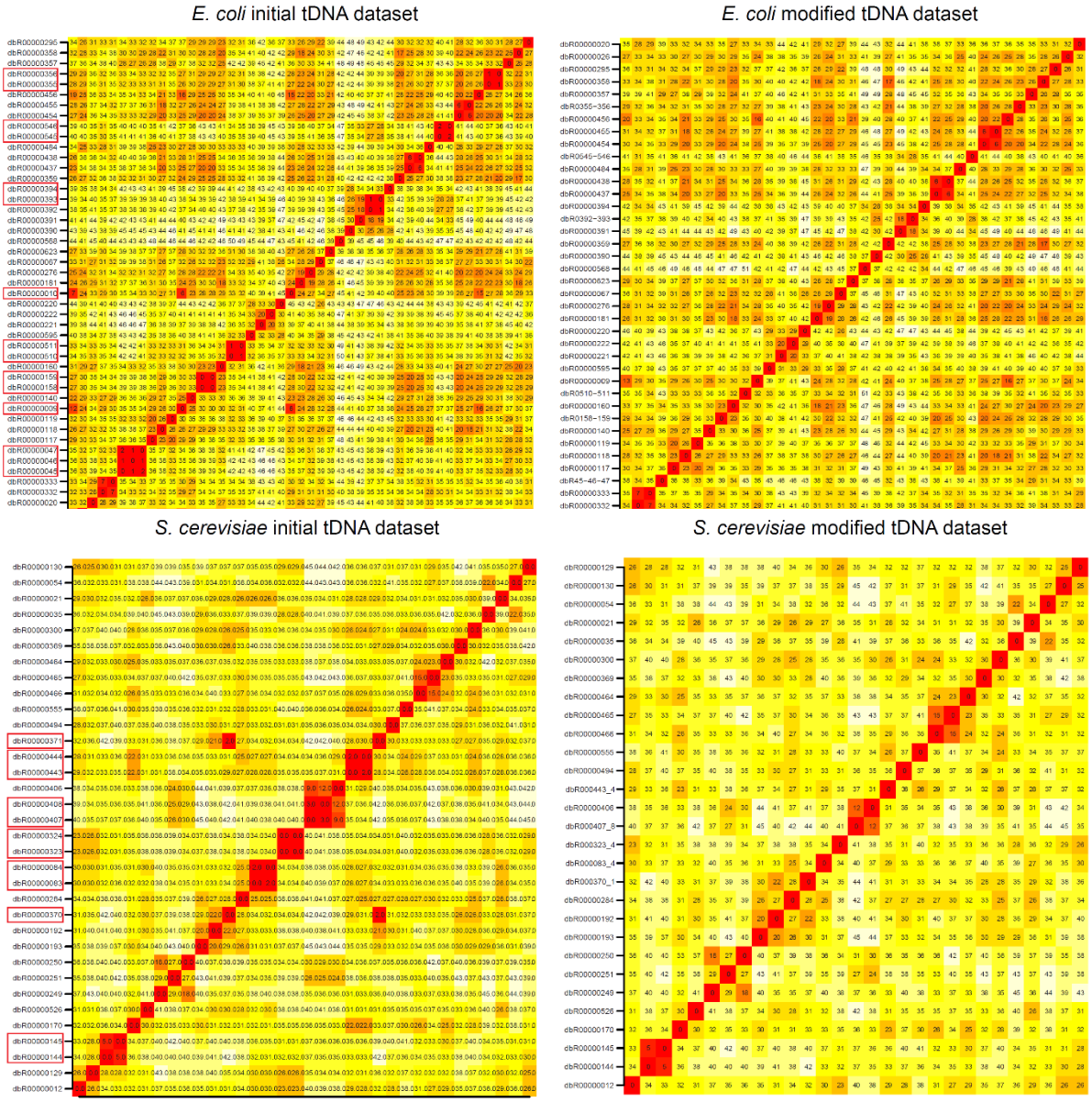


Figure S1 **Levenshtein distances between tDNA sequences from *E. coli* and *S. cerevisiae***
On the left, the initial tDNA dataset, on the right, the collapsed and corrected dataset. Collapsed sequences are boxed in red.

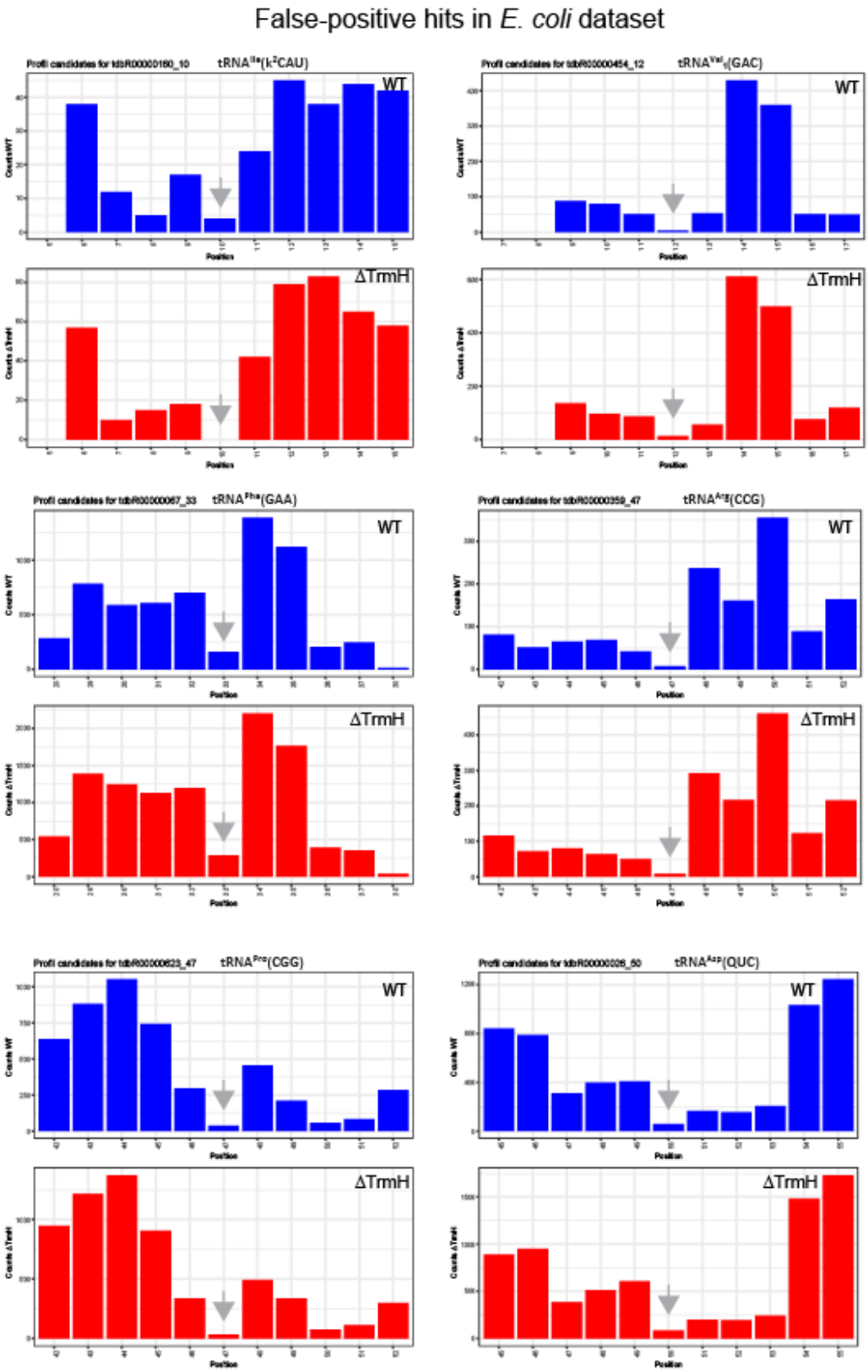


Figure S2 **False-positive hits observed in *E. coli* tRNAs,** position of the signal is centered on the graph, tRNA ID and position of false-positive hit are indicated in the header.

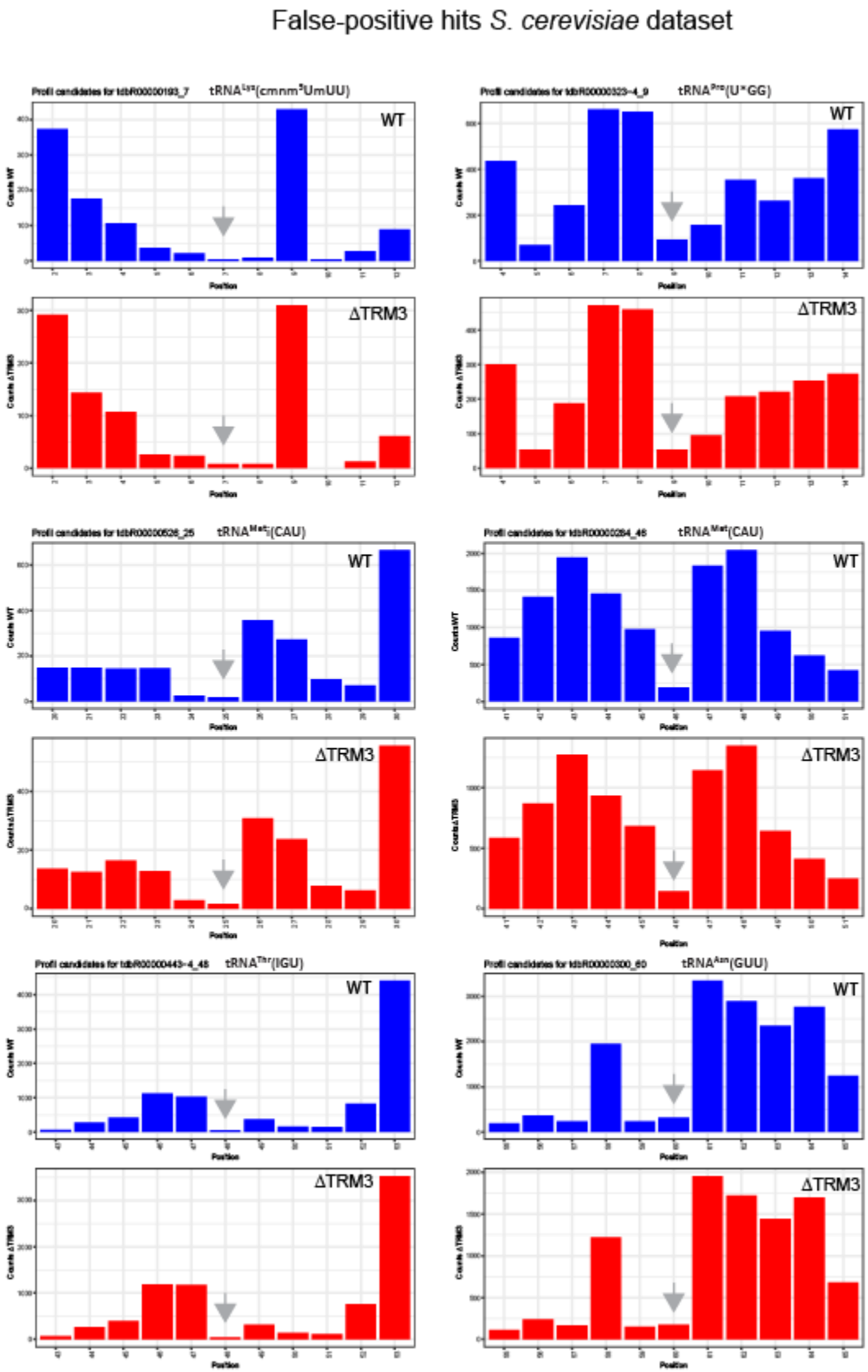


Figure S3 **False-positive hits observed in *S. cerevisiae* tRNAs**, position of the signal is centered on the graph, tRNA ID and position of false-positive hit are indicated in the header.

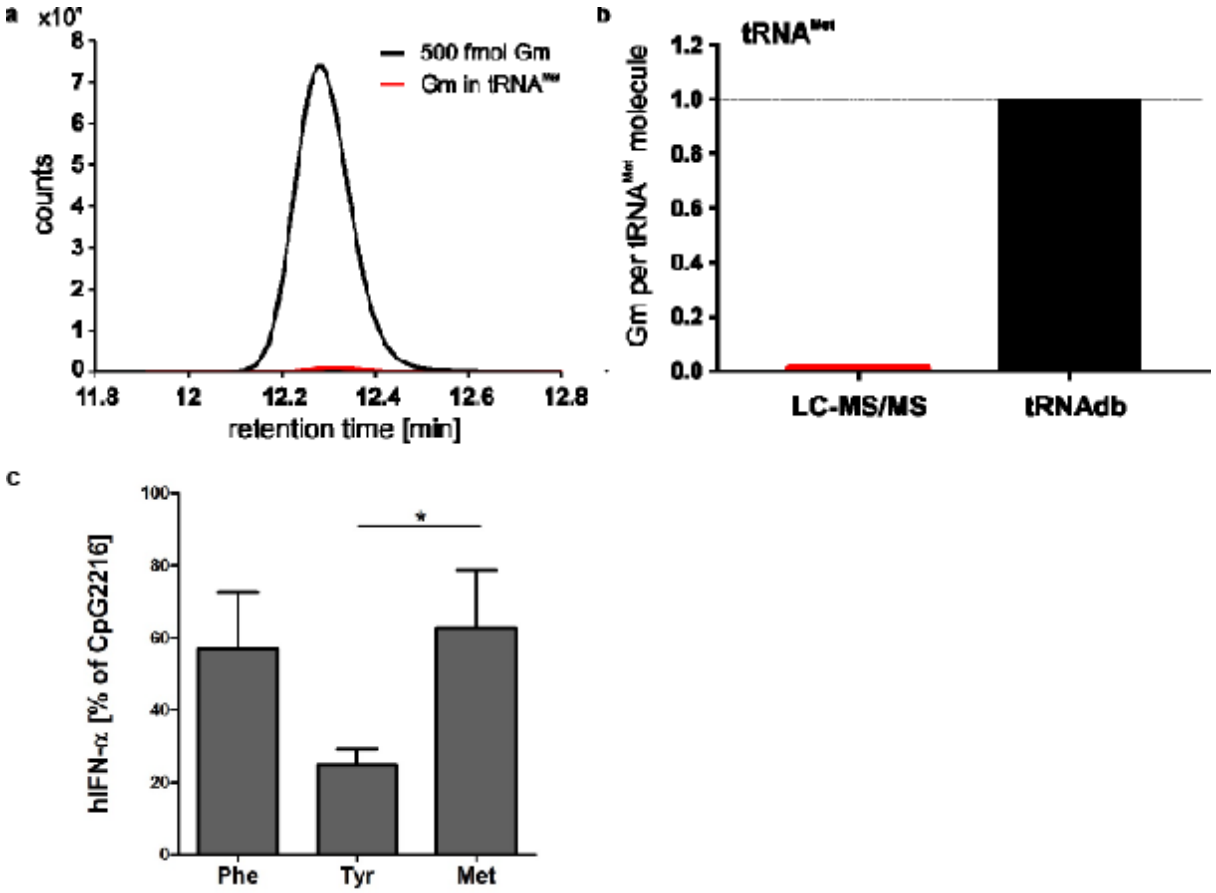


Figure S4: **Quantification of Gm in *E. coli* tRNA^{Met} by LC-MS/MS.** tRNA^{Met} was isolated from *E. coli* total RNA and analysed for its Gm content by LC-MS/MS.
(a) Mass chromatogram displaying the Gm peak corresponding to 500 fmol Gm and the Gm peak detected in approximately 500 fmol tRNA^{Met}.
(b) The quantification was performed using external calibrations of Gm and C and determining the amount of Gm per tRNA molecule. The result is shown in comparison to the expected number of Gm molecules per tRNA according to the tRNA database
(c) Human PBMCs were stimulated with *E. coli* tRNAs in a concentration of 1µg/ml and analyzed for secretion of IFN-α. Data are normalized to cytokine production induced by 1 µM CpG2216 to account for donor variation. (n = 3)

-- End of article --

4.1.2 Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2'-O-Methylations in RNA by RiboMethSeq. [DOI: 10.3389/fgene.2020.00038]

Florian PICHOT[1,2,3], Virginie MARCHAND[2], Lilia AYADI[1,2], Valérie BOURGUIGNON-IGEL[1,2], Mark HELM[3], Yuri MOTORIN[1,2]

[1] IMoPA UMR7365 CNRS-UL, BioPole Université de Lorraine, 9 avenue de la Forêt de Haye, 54505 Vandœuvre-lès-Nancy, France

[2] Epitranscriptomics and RNA Sequencing (EpiRNA-Seq) Core Facility, UMS2008 IBSLor (CNRS-UL) / US40 (INSERM), Université de Lorraine, 9 avenue de la Forêt de Haye, 54505 Vandœuvre-lès-Nancy, France

[3] Institute of Pharmacy and Biochemistry, Johannes Gutenberg-University Mainz, Staudingerweg 5, 55128 Mainz, Germany

Overview

The first RiboMethSeq protocol published for pyrosequencing technology[63] contained valuable information for its reproducibility and adaptation to others sequencing technologies, as done here[64] and within the precedent chapter. However, no justifications have been given about the different parameters value chosen for the 2'-O-methylated sites detection. A deeper look at the impact of multiple parameters, especially scores' parameters, would be beneficial to improve detection performance and reliability for projects applications.

To do so, three representative RiboMethSeq samples with substantial detection performance differences on known and verified 2'-O-sites from literature have been selected to determine the influence of multiple parameters on RiboMethSeq analysis. Metrics used here are based on Receiver Operating Characteristic curve (ROC curve) analysis using confusion matrix terminology to assess predictions performance through multiple point of view; mainly Matthews correlation coefficient (MCC) which is a general informative metric about predictions performance but also False Discovery Rate (FDR), the predictions' error rate, since false positive hits reduction is an important matter for RiboMethSeq trustworthiness.

First, we investigated the impact of sequencing depth, trimming and alignment on detection performance. Then, the different published RiboMethSeq scores have been scrutinized by decreasing or increasing the different default values included in their calculations, such as the window size for their normalisation. In parallel, quantification performance of the RiboMethSeq score "Score C" has been conducted, and an improved version of "score MAX", entitled "score Mean", has been proposed.

To summarize our findings, validation of the optimised parameters compared to the default parameters has been conducted through detection performance on additional human and yeast datasets. This displays a substantial improvement of performance with the new scores parameters introduced.

I took care of the choice of the different parameters retained for this study and created the R pipeline to interpret their impact on RiboMethSeq performance based on the ROC curve method. I have also posted this pipeline online on my GitHub page (https://github.com/FlorianPichot/RiboMethSeqPipeline). At last, I have participated in the writing process for the results about the scores' calculation window, score C quantification performance, score Max and the final validation.

Abstract

A major trend in the epitranscriptomics field over the last 5 years has been the high-throughput analysis of RNA modifications by a combination of specific chemical treatment(s), followed by library preparation and deep sequencing. Multiple protocols have been described for several important RNA modifications, such as 5-methylcytosine (m^5C), pseudouridine (Ψ), 1-methyladenosine (m^1A), and 2'-O-methylation (Nm). One commonly used method is the alkaline cleavage-based RiboMethSeq protocol, where positions of reads' 5'-ends are used to distinguish nucleotides protected by ribose methylation. This method was successfully applied to detect and quantify Nm residues in various RNA species such as rRNA, tRNA and snRNA. Such applications require adaptation of the initially published protocol(s), both at the wet bench and in the bioinformatics analysis. In this manuscript, we describe the optimization of RiboMethSeq bioinformatics at the level of initial read treatment, alignment to the reference sequence, counting the 5'- and 3'- ends, and calculation of the RiboMethSeq scores, allowing precise detection and quantification of the Nm-related signal. These improvements introduced in the original pipeline permit a more accurate detection of Nm candidates and a more precise quantification of Nm level variations. Applications of the improved RiboMethSeq treatment pipeline for different cellular RNA types are discussed.

Introduction

The precise and high-throughput mapping of modified nucleotides in RNA is a real challenge in the field of epitranscriptomics (RNA modifications). Several recent publications have demonstrated the presence of numerous RNA modifications, not only in rather well studied species such as tRNA/rRNA/sn(sno)RNA, but also in coding RNAs (mRNA), in all living organisms studied to date. Thus, several high-throughput methods for the identification of RNA modifications have been developed and successfully applied for mapping m^5C, m^6A, pseudouridines, and (more recently) 2'-O-methylations (2'-O-Me), along with m^1A and $m^7G/m^3C/D$[69, 270, 274, 314, 315, 320–322].

The major impediment to applying high-throughput screening methods resides in various experimental and bioinformatics biases, which are only partially controlled and may affect the precision of the final result or even lead to a non-negligible number of false-positive identifications. Taking into account the size of eukaryotic transcriptomes, thousands of false-positive signals may appear at the transcriptome-wide level, even with extremely strict criteria for candidate site selection (e.g., False Discovery Rate (FDR)< 1%). Thus, every step in bioinformatic data treatment, conversion and manipulation should be optimized in order to minimize the number of potential false-positive signals.

Recently, we published a high-throughput deep sequencing-based approach, named RiboMethSeq, for mapping of 2'-O-methylations in highly abundant RNAs, mostly in rRNA[64, 109], with possible extension to tRNA[65, 323]. This protocol is also suitable for low abundance RNAs[131]. The RiboMethSeq protocol is based on the enhanced protection of the phosphodiester bond in RNA from nucleolytic attack and cleavage due to the presence of 2'-O-methylation at the 5'-neighboring ribose moiety. This enhanced protection is evaluated as a normalized number of 5'- and 3'-ends of randomly cleaved fragments present in the sequencing library. If a residue is 2'-O-methylated, this reduces the cleavage efficiency and thus the relative number of fragments starting and ending at +1 nucleotide relative to the modification. In the RiboMethSeq approach, such relative protection compared to neighbors is calculated using different scoring schemes, and the presence/absence of a 2'-O-methylation is then deduced on this basis. An alternative calculated score (ScoreC) also allows precisely measuring the methylation ratio at a given nucleotide.

In this work, we report the comprehensive optimization of every step of the bioinformatic treatment used for the detection and quantification of ribose 2'-O-methylation by the RiboMethSeq protocol. We systematically evaluated the importance and the impact of 5'- and 3'-trimming strategies, parameters for alignment to the reference sequence, as well as the use of specific calculated scores for 2'-O-Me mapping and quantification. Our results demonstrate that a reduced calculation interval is favorable for the general discrimination of 2'-O-Me signals from potential false-positive hits. We propose new, optimized scores (ScoreMEAN2, ScoreA2, and MethScore2) that provide better FDR values and also improve the relative quantification of 2'-O-methylation in RNA.

Materials and Methods

3.1 Biological material

To optimize the RiboMethSeq scores, we used previously published datasets obtained for wild-type yeast *Saccharomyces cerevisiae* and human HeLa cell rRNA 2'-O-methylation analysis, as well as additional samples for hTERT immortalized human mammary epithelial cell line (HME)[64, 108, 109]; accession numbers PRJEB43738, PRJEB34951 and PRJEB35565.

3.2 RiboMethSeq protocol

The RiboMethSeq protocol[64, 66] consists of random RNA fragmentation under alkaline conditions (96°C, pH 9.3, ~12-14 min for rRNA), an end-repair step consisting of de-phosphorylation of the 3'- ends of the RNA fragments and 5'-end phosphorylation, library preparation using 3'-end and 5'-end ligation of adapters, an RT-step and PCR amplification coupled with barcoding. The resulting library is sequenced in paired-end PE 2x75 or, more commonly, in single-end mode (SR50) using an Illumina sequencing device (MiSeq or HiSeq1000).

3.3 Trimming and alignment

Adapter removal in this study was performed using the Trimmomatic utility[317]. With the default trimming parameters, the recognition of adapter sequences requires at least a 16 nt length. Shorter fragments of adapters are not recognized and thus are not removed. However, with a stringency of 7 (instead of 10), the adapter recognition requires only 10 nt. Considering this, 3'-end counting was carried out only for reads shorter than 40 nt after trimming. The alignment of raw reads was conducted by Bowtie2[318] in end-to-end mode.

3.4 Comparing the performance of 2'-O-methylated site detection

For the selected datasets, we first applied the RiboMethSeq pipeline under standard conditions, and the previously described scores (ScoreMAX6, ScoreA, B and C) were calculated[63, 64]. Score values were sorted in descending order, and the known modification status of every nucleotide (2'-O-methylated residue, pseudouridine, other modified residue or unmodified nucleotide) was attributed. Receiver operating characteristics (ROC) curves were plotted using these data, together with associated Matthews correlation coefficient (MCC) values. Other associated parameters of the ROC curves were calculated for maximal MCC value, taking into account true positive/false positive/true negative/false negative (TP/FP/TN/FN) hits. The performance of the treatment was evaluated on the basis of both the maximal MCC value and the associated FDR.

Results and Discussion

4.1 Brief overview of RiboMethSeq experiment

As described above, the high-throughput mapping of 2'-O-Me residues in RNA is based on random fragmentation of the phosphodiester bonds under mild alkaline conditions. The presence of a 2'-O-Me group protects the 3'-adjacent phosphodiester bond from nucleolytic cleavage, thus generating the characteristic gap in the 5'-end (as well as 3'-end) coverage profile of the sequencing library prepared from the

fragmented RNA (**Supp Fig. S1**). This enhanced protection is used as a signature for 2'-O-methylation and protection (and thus the gap's depth) and is supposed to be proportional to the level of 2'-O-Me at a given position.

In previously published studies[63, 64, 108, 109], we and others used rather standard parameters for read trimming and alignment, and calculations arbitrarily used cleavage efficiency for 12 neighboring nucleotides (+/- 6 from the methylation site). Scores allowing 2'-O-Me detection (ScoreMAX6 and ScoreA, called here ScoreA6), were calculated. The 2'-O-methylation level was assessed by calculating the MethScore (identical to the previously reported ScoreC, called ScoreC6 here).

4.2 Selection of representative datasets for optimization

Initial screening and optimization of the RiboMethSeq bioinformatic pipeline was performed with >40 available human rRNA RiboMethSeq datasets obtained under standard, previously described[64, 108, 109], conditions of RNA fragmentation, sequencing, trimming, alignment and score calculation. We used a cumulative list of human modified rRNA positions reported in a 3D rRNA modification database[324] and in the LBME snoRNA database[184], including two new positions that were recently reported[107]. Altogether, we considered 40 sites in 18S rRNA, 64 sites in 28S rRNA and 2 positions in 5.8S rRNA (see **Supp Table S1**). Some of these positions are probably variably modified, or even not modified at all in some human cell lines or tissues[107–109]; therefore, these incomplete modifications necessarily affect the number of FN hits and the max MCC values in the RiboMethSeq analysis. For each dataset, calculations of the RiboMethSeq scores were performed, and the performance of each dataset was evaluated for the detection of known rRNA 2'-O-methylated positions. Based on the pre-analysis of available RiboMethSeq human rRNA samples, we selected 3 representative human datasets corresponding to 2 different cell lines (HUVEC and HeLa), as well as cultured bone marrow stem cells for further, more extensive analysis and optimization of the whole treatment pipeline (respectively named Sample1 -- HUVEC, 2 -- BMSC, and 3 -- HeLa (**Figure 1 and Supp Fig. S2, Supp Table S2**).

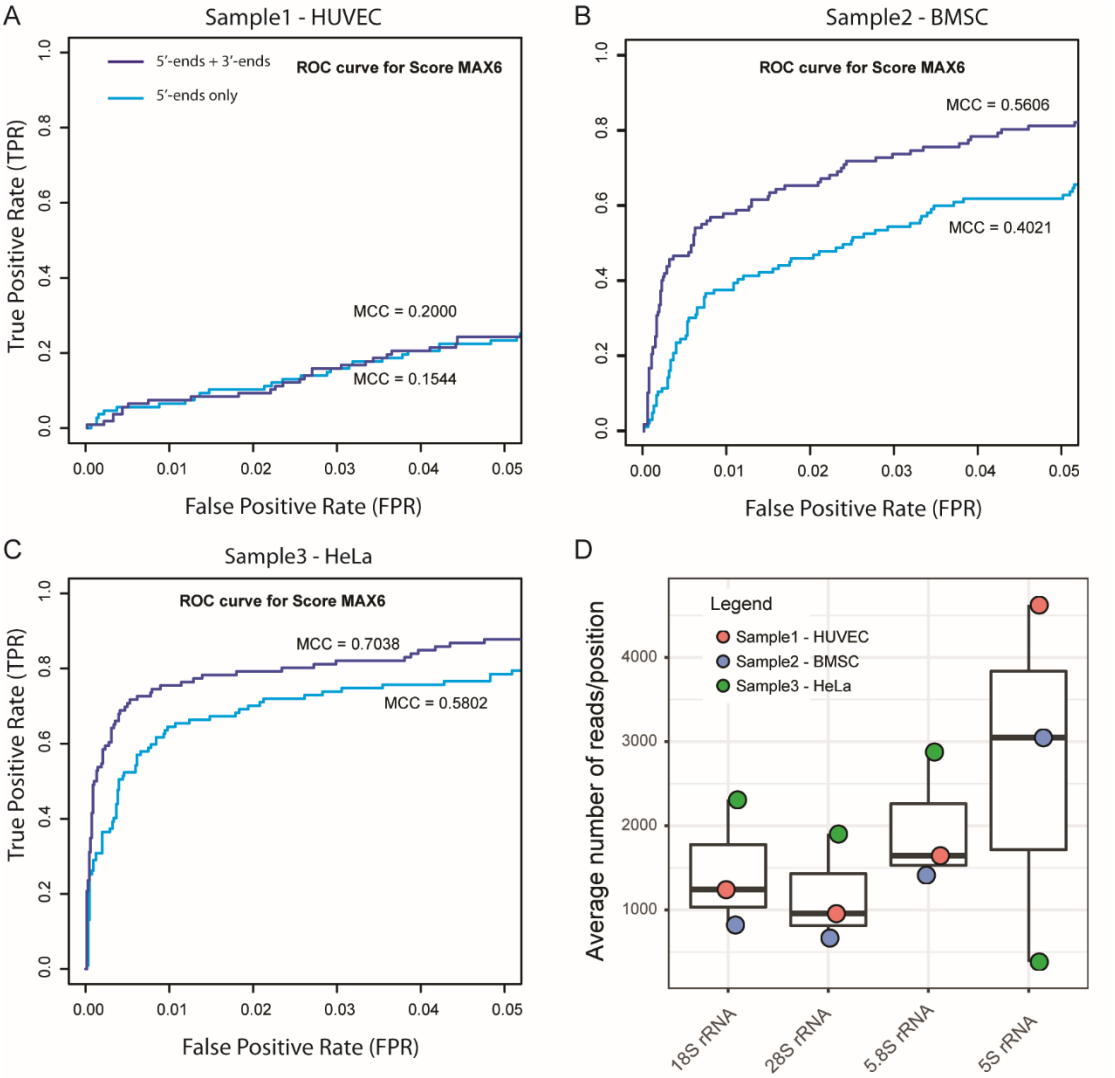


FIGURE 1 Selection of RiboMethSeq datasets for optimization. Three human datasets providing representative performance of 2'-O-Me detection (Sample1 -- HUVEC, 2 -- BMSC, and 3 -- HeLa) were selected on the basis of ROC curves and the associated max MCC values for ScoreMAX6 (Panels A, B, C). Graphs represent zoom to ROC curve 0-0.05 for FPR and 0-1 for TPR. It was previously shown[64] that 5'-end coverage (light blue curve) is sufficient for reliable construction of the RNA protection profile, but cumulated 5'- and 3'-end coverage (violet curve) provides better discrimination between methylated positions and FP hits. Panel D shows the read coverage per position for human rRNAs. 5S rRNA shows quite variable coverage, probably due to variations in 5S rRNA content in the total rRNA fraction due to biased extraction.

To allow a performance comparison between datasets, we selected samples with similar numbers of raw RiboMethSeq sequencing reads (>20 mln, see **Supp Table S2**).

4.3 Minimal number of reads required for analysis

An optimized volume of sequencing reads required for complete RNA analysis is highly important, since it allows to obtain reliable results with a minimal sequencing cost per sample. From the analysis of the yeast rRNA samples[64], sufficient coverage was evaluated to be ~ 750-1000 reads/RNA position. For a more detailed analysis, we applied human rRNA datasets. Despite a similar number of raw sequencing reads (~20 mln), these datasets behaved differently regarding the precision of 2'-O-Me detection. Notably, the fraction of 5S rRNA reads varied substantially from one sample to another (**Figure 1**), probably also reflecting the different RNA extraction protocols used. However, there was no correlation between the total coverage and the prediction quality for 2'-O-Me. To define the minimal amount of raw sequencing information required for the successful application of RiboMethSeq analysis of human rRNA, we compared the performance of the method using a variable number of input reads for the same sample.

As anticipated, human 28S rRNA was the most difficult target to get full representation for all positions in the sequence. A comparison of missing positions in 28S rRNA in relation to the sequenced population is given in **Table 1**. A low number of raw reads (4 mln) can still be used, but numerous positions of 28S rRNA have zero raw read 5'-/3'-end counts in the final set. Despite this, the analysis of known 28S rRNA 2'-O-Me was not affected, because underrepresented regions are far away from these modified positions. Increasing the read number (8-12 mln, > 1000 reads/nt) improves representativity, with only a marginal number of uncovered nucleotides, while 15-20 mln reads is recommended to achieve full coverage.

Table 1 Alignment statistics and uncovered rRNA positions in samples used for analysis

Sample	***number of raw reads used***	***4 mln***	***8 mln***	***12 mln***	***16 mln***	***20 mln***
Sample 1 HUVEC	trimmed reads	3873996	7741257	11607806	15471803	19341955
	short reads for alignment	1761248	3581209	5338558	7104102	8909438
	aligned to rRNA reference	1513287	3077770	4587105	6104305	7657111
	uncovered pos 5S rRNA	0	0	0	0	0
	uncovered pos 5.8S rRNA	0	0	0	0	0
	uncovered pos 18S rRNA	**4**	**1**	0	0	0
	uncovered pos 28S rRNA	**100**	**37**	**21**	**11**	**7**
Sample 2 BMSC	trimmed reads	3878093	7752516	11628210	15494852	19371588
	short reads for alignment	1473330	2986750	4455805	5927702	7428697
	aligned to rRNA reference	999714	2027867	3023042	4022365	5042133

	uncovered pos 5S rRNA	0	0	0	0	0
	uncovered pos 5.8S rRNA	0	0	0	0	0
	uncovered pos 18S rRNA	**2**	**1**	0	0	0
	uncovered pos 28S rRNA	**18**	**5**	0	0	0
Sample 3 HeLa	trimmed reads	3882713	7764523	11644031	15516647	19387461
	short reads for alignment	2582027	5182220	7776132	10353556	12934621
	aligned to rRNA reference	2222928	4460722	6693085	8910528	11132036
	uncovered pos 5S rRNA	0	0	0	0	0
	uncovered pos 5.8S rRNA	0	0	0	0	0
	uncovered pos 18S rRNA	0	0	0	0	0
	uncovered pos 28S rRNA	0	0	0	0	0

4.4 Minimal trimming length

The minimal trimming length used in the treatment pipeline may affect 2'-O-Me detection. Trimming parameters considerably influence the precision of 3'-end mapping for SR50 reads and the alignment quality. We thus tested variable minimal trimming lengths keeping alignment parameters constant. The calculated max MCC values for the tested human datasets showed no influence of these parameters on the final results, even if the number of ambiguously aligned short reads increased with a decreased minimal trimming length (**Supp Fig. S3**). Depending on the length and complexity of the target RNA sequence, we recommend adapting the minimal length and Bowtie2 seed length (see below); the optimal seems to be 10 or 12 for human or yeast rRNA, or even lower for shorter RNAs (e.g., tRNA).

4.5 Variation of alignment parameters

The original RiboMethSeq protocols[64, 66] used rather strict alignment parameters in Bowtie2: end-to-end mode, a minimal seed length of 22 nt and zero mismatches allowed in the seed (preset option "– sensitive"). The influence of the alignment mode (end-to-end versus local) was previously evaluated, and the soft read trimming performed in the local mode was found to be unsuitable for precise mapping of the read ends[64]. However, the seed length and number of mismatches allowed may also influence the quality of the alignment, since human rRNA has variations in nucleotide sequence and contains other modified nucleotides, which alter cDNA sequences, thus perturbing alignment to the reference.

To evaluate the importance of the alignment parameters, we proceeded with treatment using a reduced seed length and allowing (or not) mismatched nucleotides in the seed. The following parameter combinations were tested: seed lengths of 22 nt (default value for "– sensitive" preset option, and used previously), 16 nt, 12 nt and 8 nt, allowing (or not) mismatched nucleotides in the seed.

The data in **Supp Fig. S4A** show that the total proportion of aligned reads (unique or multiple alignments) vary only very slightly as a function of seed length and allowed mismatches; a seed length of 22 nt and mismatches in the seed allows only a slightly better alignment (69.41% vs 65.41% of aligned reads). Variation of the Bowtie2 seed length does not much affect the max MCC value for both scores used for optimization (**Supp Fig. S4B)**. Similarly, allowing mismatches in the seed also has no influence on the final results. This shows that the alignment to the reference sequence is quite robust and mostly depends on the quality of sequencing data. Based on these observations, we recommend the use of 8-12 nt seed length, depending on the complexity and the length of the target RNA sequence. For better performance, the seed length can be coordinated with the minimal trimming size for sequencing reads (**Supp Fig. S5**).

4.6 Importance of the calculation window with neighboring nucleotides

In the originally published RiboMethSeq protocols[63, 64], 12 neighboring nucleotides (+/- 6 nt window) were taken into account to calculate the 2'-O-Me scores (ScoreA, B and C). Since the size of this window was arbitrarily selected, we explored the influence of the window's size on the discrimination of 2'-O-Me signals from background. We compared the maximal MCC and the FDR values for the calculation interval from +/-2 nt up to +/-8 nt (window of 4 to 16 nt). The graph in **Figure 2A** shows that the ScoreMAX is nearly insensitive to the size of the calculation window, while ScoreA shows the best performance (and the lowest FDR) with the smallest window size (+/-2 nt). A larger window size has a detrimental effect for both scores. On the basis of these observations we suggest reducing the calculation window size to 4 neighboring nucleotides (Score 2 calculation scheme, +/-2 nt); the scores calculated with this window are referred to as ScoreMAX2 and ScoreA2.

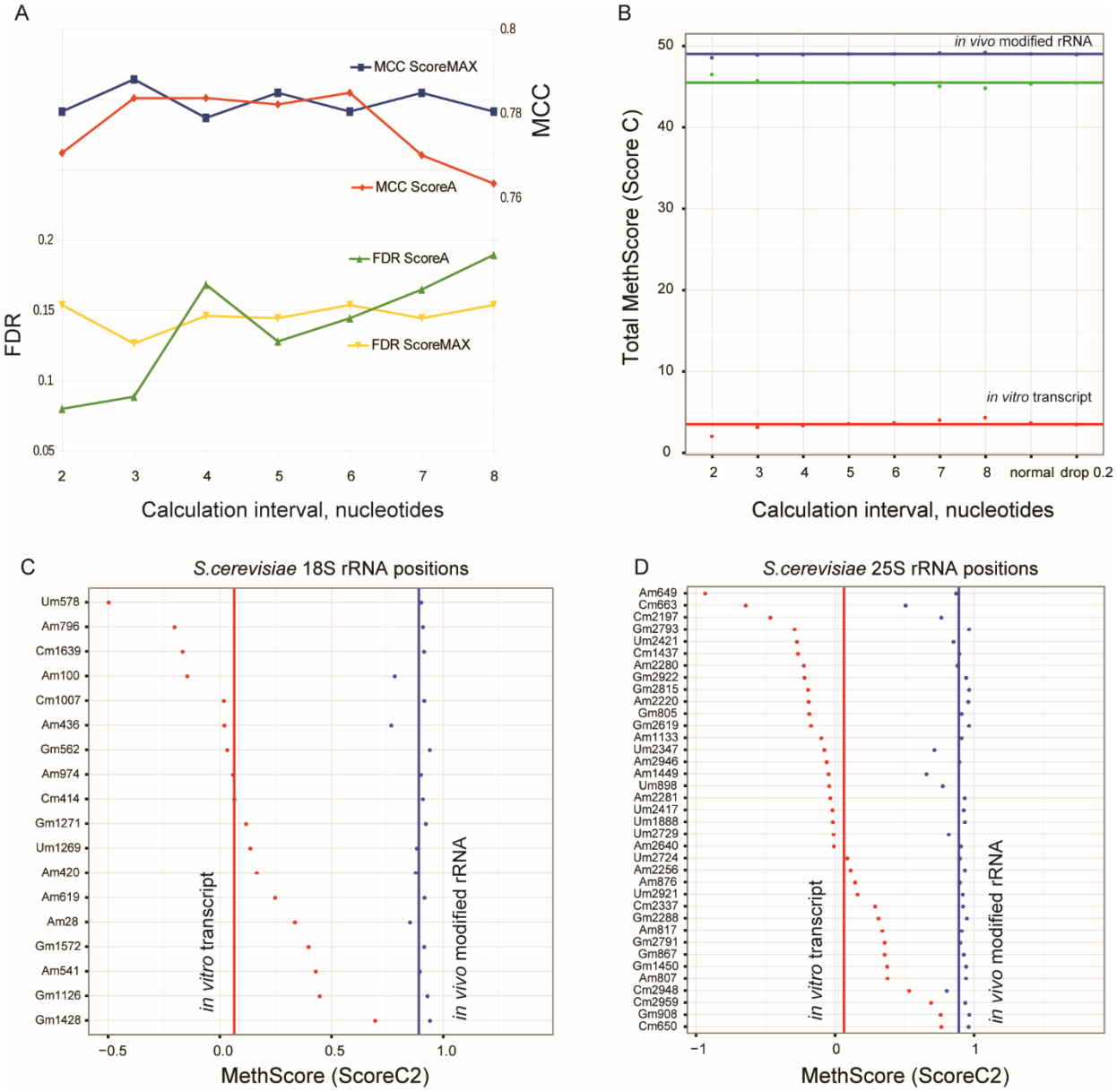


FIGURE 2 Panel A - Performance of ScoreMAX and ScoreA calculated using variable numbers of neighboring nucleotides (from +/-2 to +/-8). The standard RiboMethSeq protocol uses a +/-6 interval. Values for FDR and max MCC are given. The scale on the left corresponds to FDR, and on the right to MCC. Sample 2 -- BMSC was used here for all calculations; other datasets gave similar trends. Panel B shows global values for MethScore (Score C) calculated for modified yeast rRNA and in vitro rRNA transcripts using different neighboring intervals. The total number of "2'-O-Me groups" in rRNA is given (red - in vitro transcript, blue - modified rRNA). Panels C and D MethScores2 (ScoreC2) for individual 2'-O-methylated positions in 18S (panel C) and 25S rRNA (Panel D, red - in vitro transcript, blue - modified rRNA). Lines correspond to average values.

4.7 Quantification of 2'-O-methylation with MethScore (ScoreC)

In the original RiboMethSeq protocol[64], the MethScore (identical to ScoreC[63, 107]) was used for quantification of the 2'-O-Me level because the MethScore demonstrates a linear dependence on the depth of the gap in a cleavage profile, which is supposed to represent the protection and thus the degree of 2'-O-Me. In an ideal situation of a homogeneous cleavage profile, the MethScore varies from 0 to 1.0 and thus can be used as a measure of the degree of RNA methylation. In the case of yeast rRNA studied previously, the MethScore varied from negative values to 1.0, while conserving the linear dependence on the methylation rate[64].

We noticed that RiboMethSeq detection scores behave better with a reduced calculation interval (+/- 2 nt); therefore, we also explored variations of the MethScore. To select the best calculation interval, we compared the MethScore for intervals from 2 to 8 neighboring nucleotides for naturally modified yeast rRNA and for unmodified synthetic rRNA transcripts. The MethScore values were expected to reach maximum for naturally modified rRNA and minimum for the unmodified counterpart. **Figure 2B** shows the difference between the MethScore values calculated for variable intervals for rRNA and synthetic transcript. For MethScore6 (+/- 6 nt), we also tested different weight contributions of neighboring nucleotides.

The MethScore demonstrates the maximal cumulative difference between rRNA and synthetic transcript for the shortest interval of +/-2 nt; the other tested intervals gave roughly the same results. On the basis of this observation, we suggest calculating the MethScore for 2 neighboring nucleotides (referred to as MethScore2).

A detailed analysis, position by position, for 18S and 25S rRNA (**Figure 2CD**) shows that methylated yeast rRNA displays MethScore2 values close to 0.9-1.0 for almost all modified positions (blue dots and line for average value), while the average level for synthetic unmodified transcript (red dots and line) is rather low. However, it is notable that the MethScore2 values for unmodified transcript are extremely variable, ranging from -1 to almost 0.9. For a limited subset of sites (Gm1428 in SSU 18S-rRNA and 4 positions in LSU 25S-rRNA), the difference of MethScore2 between modified and unmodified RNA is as low as 0.1. Precise measuring of the 2'-O-Me level variations at these rRNA positions is thus extremely difficult. However, over 90% of methylation sites display considerable MethScore2 differences between the modified and unmodified state, thus validating relative quantification of the methylation rate. Absolute values of the 2'-O-methylation cannot be directly measured using the RiboMethSeq approach; however, for a limited subset of sites the values of MethScore2 are comparable with independent measurements of 2'-O-Me rates assessed by LC-MS/MS[105, 200].

4.8 Optimization of the ScoreMAX

The originally used ScoreMAX6[64] was designed to favor directional 5'->3' gap depth compared to the opposite orientation. This design was based on the assumption that directional 5'->3' drop is more informative for measuring the protection of the phosphodiester bond, while the drop in the opposite direction (3'->5') may represent nonspecific RNA structural effects. To verify the validity of such assumption, we tested different variants of ScoreMAX and their performance in the detection of 2'-O-Me. We compared the ScoreMAX2 (+/-2 nt window) with two modified versions. The first modified score retained the maximal value of normalized 5'->3' and 3'->5' drop (ScoreMAX-MAX), while the second version calculated the average value between the two (ScoreMEAN2).

Calculation of the max MCC value demonstrated that ScoreMAX-MAX is less performant than the original ScoreMAX2, while ScoreMEAN2 shows better discrimination of FP hits. For the same dataset, Sample3 -- HeLa application of the ScoreMEAN2 allows attainment of the maximal MCC value of 0.7954 and reduces the FDR from 27% to 4%. A similar tendency was observed for two other datasets (**Supp Table S3**).

An important source of FP hits in RiboMethSeq analysis is the reduced ligation efficiency observed when modified RNA nucleotides are present at the 5’-end extremity of the fragment. This was experimentally observed for pseudouridine and other rare RNA modifications[63, 64]. Since this reduced ligation efficiency also generates a “gap” in the sequencing profile, such FP signals are difficult to discriminate from undermethylated 2’-O-Me sites. Both scores (MEAN2 and A2) show a fair separation of values for 2'-O-Me nucleotides (blue) and unmodified residues (gray), but the peak for pseudouridines (red) partially overlaps with that of 2'-O-methylation. ScoreA2 shows better separation of 2'-O-Me and pseudouridine signals, while ScoreMEAN2 demonstrates a higher MCC and thus better performance for 2'-O-methylation detection (**Supp Fig. S6**).

We also attempted to cumulate values for ScoreA2 and ScoreMEAN2 together by calculating their normalized sum (ScoreD). Despite the fact that ScoreA2 and ScoreMEAN2 generally pick out different FP hits, ScoreD does not further improve the performance (max MCC and FDR) compared to ScoreMEAN2 alone. In conclusion, the best results for detecting 2'-O-methylated residues were obtained with a calculation window of 4 nt (+/-2 nt) using ScoreA2 and ScoreMEAN2.

4.9 Validation of ScoreMEAN2 and ScoreA2 with human and S. cerevisiae *rRNA datasets*

To compare improvements associated with the use of ScoreMEAN2 and ScoreA2, we used independent human RiboMethSeq datasets obtained for other HeLa samples, human mammary epithelial cells (HME), human fibroblasts and Wharton's jelly MSC (19 altogether). Comparison of max MCC and FDR values obtained for Scores MAX6/MEAN2 and Scores A6/A2 calculated with different scoring schemes is shown on **Figure 3 and Supp Fig. S7)**. In all cases, calculation scheme Score 2 improved detection of rRNA 2'-O-methylation (**Supp Table S3**). Additional validation was also performed with yeast *S. cerevisiae* rRNA RiboMethSeq dataset. **Figure 4AB** shows a side-by-side comparison of ScoreMAX6 and ScoreA6 values with their respective distributions (panel A), as well as ScoreMEAN2 with ScoreA2 (panel B). The new scoring scheme provides better separation of 2'-O-Me signals from those for pseudouridine and unmodified nucleotides, which is also confirmed by a better FDR and MCC values (**Supp Fig. S8**). These data validate the newly proposed ScoreMEAN2 and ScoreA2, which can now be used for the detection of 2'-O-methylation in various RNA types.

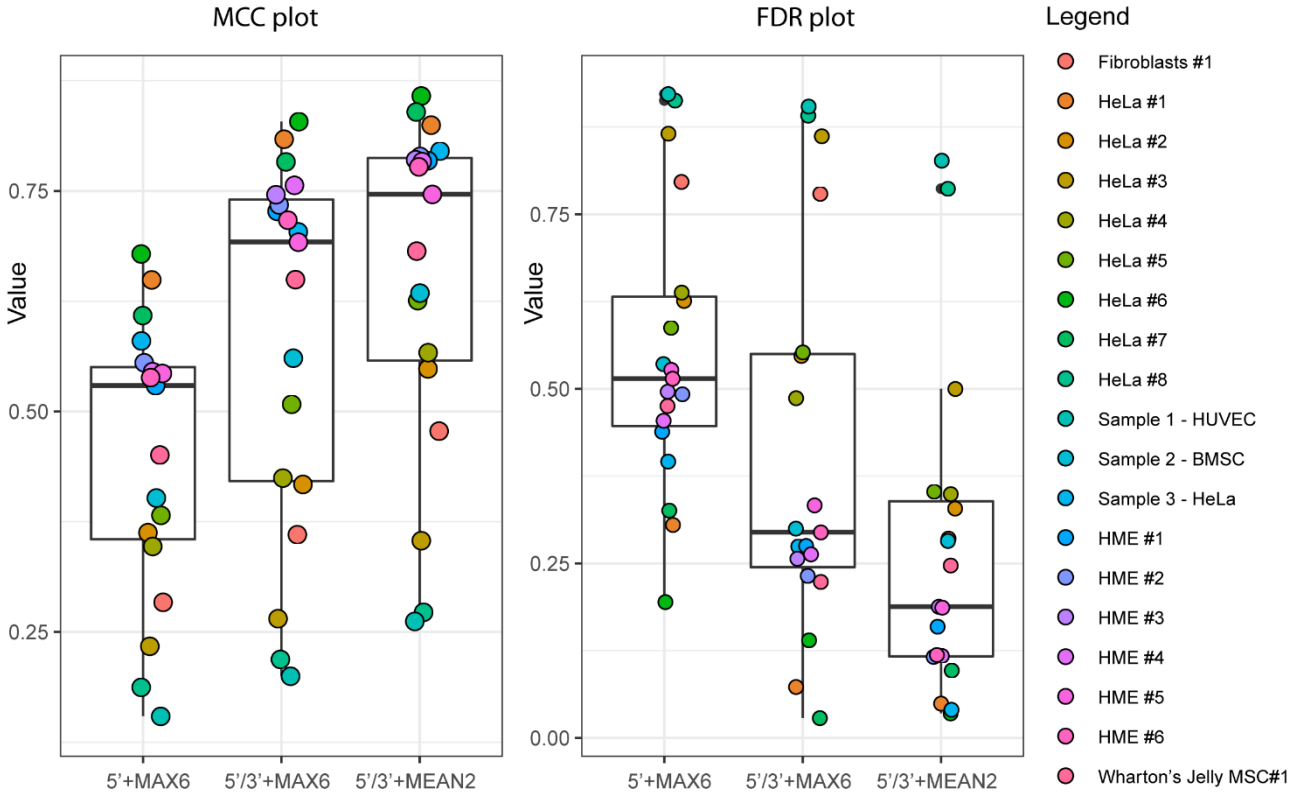


FIGURE 3 Improvement of ScoreMAX/MEAN (MAX6 and MEAN2) with 5'/3'-counts and reduced calculation window (Score 2 calculation scheme). Boxplot shows max MCC values (left) and associated FDR (right) for all 19 RiboMethSeq datasets used for validation. Identity of the RiboMethSeq datasets is given on the right.

Conclusion

The results of this optimization suggest that RiboMethSeq analysis of rRNA can be performed using short trimming lengths (10-12 nt) and adapted Bowtie2 alignment parameters, which allows a gain in sequencing information. Calculations of RiboMethSeq scores for short calculation intervals (4 neighboring nucleotides) and ScoreMEAN2 and ScoreA2 can be used for mapping of the modified positions. Quantification of 2'-O-Me is accomplished using MethScore2, also calculated for 4 neighbors. A minimal of 4-6 mln of raw reads (~400 reads/nt on average) can be used to evaluate the methylation level for known rRNA methylation sites, but at least 15 mln reads (~1500 reads/nt on average) should be used to discover new methylation candidates. For different RNA species such as tRNAs, the trimming/seed length can be further reduced (up to 8 nt), but the calculation of RiboMethSeq scores with 4 neighbors can be maintained. The optimal read coverage is somewhat similar (15-20 mln raw reads for yeast/human tRNAs, respectively).

Acknowledgements

We thank Team 05 of IMoPA UMR7365 for generous gift of HUVEC and bone marrow stem cells used for RNA extraction. This work was supported by FRCR EpiARN project from Lorraine Region to YM and SPP1784, HE 3397/13-2, HE 3397/14-1 from DFG to MH. This work was performed in the framework of EPITRAN COST initiative (CA16120).

Supplementary Figures and Tables

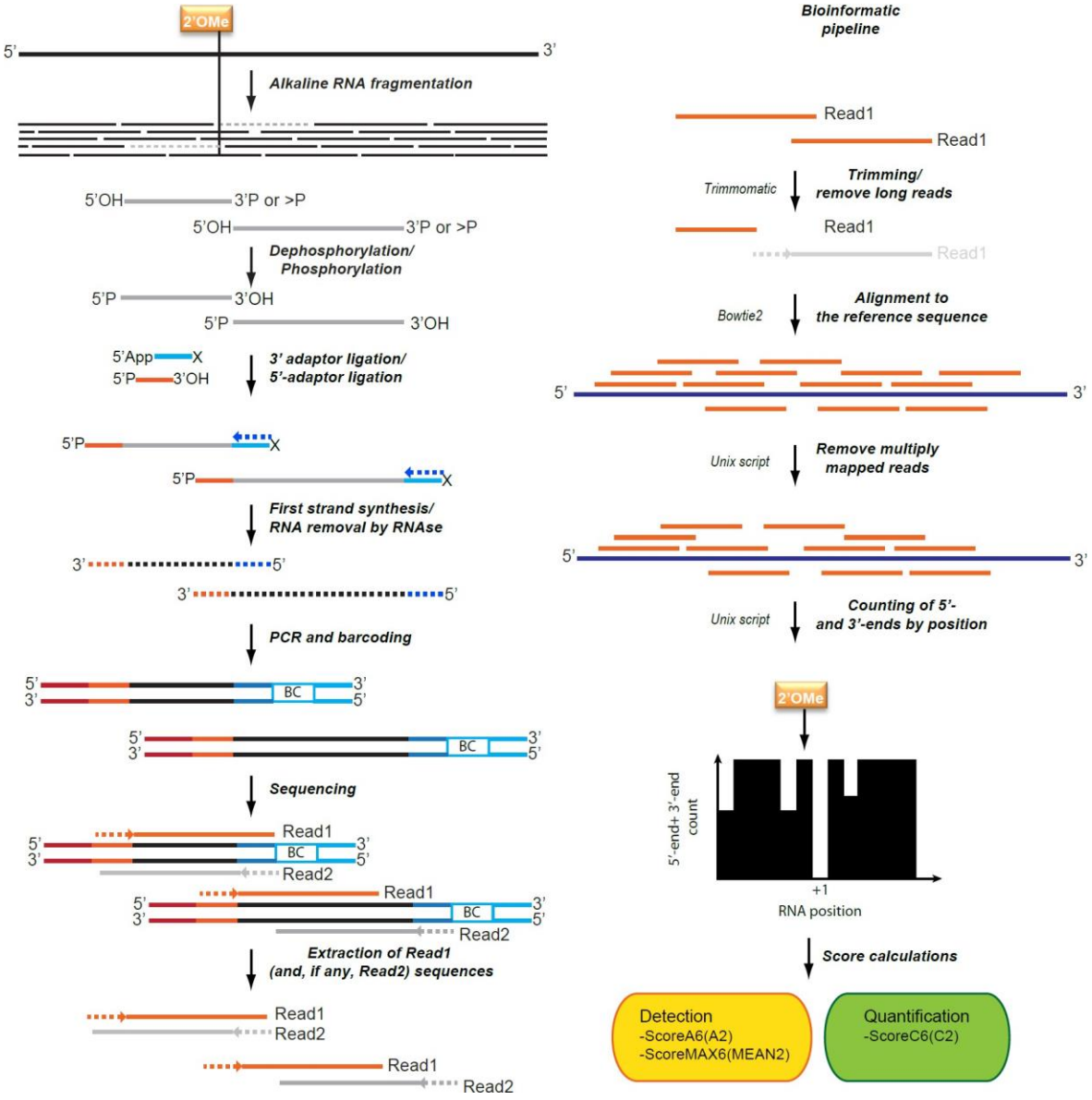


Figure S1 Outline of RiboMetHSeq protocol for 2'-O-methylation in RNAs, "wet" lab steps and major steps of the data treatment pipeline are shown.

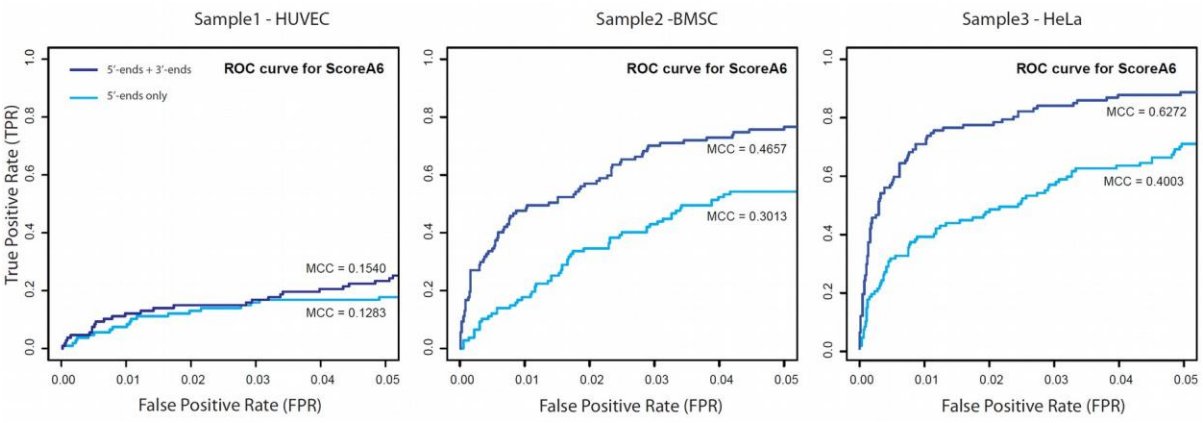


Figure S2 Selection of RiboMethSeq datasets for optimization. Three datasets providing representative performance in 2'-O-Me detection (Sample1 - HUVEC, 2 - BMSC, and 3 - HeLa, from left to right) were selected on the basis of ROC curve and associated max MCC values for ScoreA6. Graphs represent zoom to ROC curve 0-0.05 for FPR, 0-1 for TPR. It was previously shown that 5'-end coverage only (light blue curve) is sufficient for reliable construction of RNA protection/2'-O-Me profile, but cumulated 5'- and 3'-ends coverage (dark blue curve) generally provides better discrimination between methylated positions and FP hits.

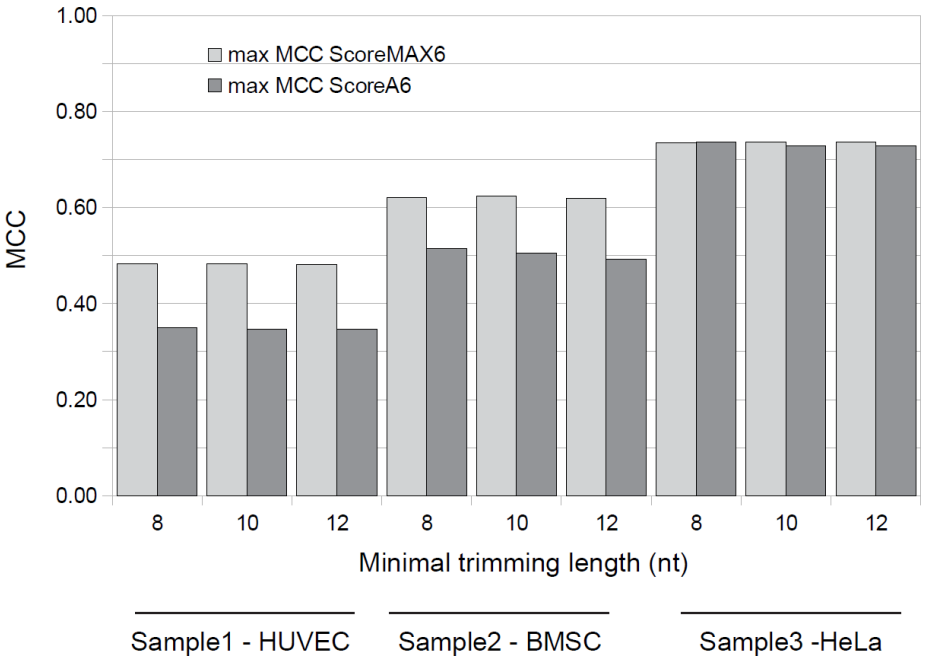


Figure S3 Influence of trimming parameters on max MCC values for ScoreMAX (MAX6) and scoreA6. Trimming of raw reads was done using three minimal length 8 nt, 10 nt and 12 nt). Barplot shows max MCC values for ScoreMAX6 and ScoreA6, for three samples (Sample1 - HUVEC, 2 - BMSC, and 3 - HeLa) retained for evaluation.

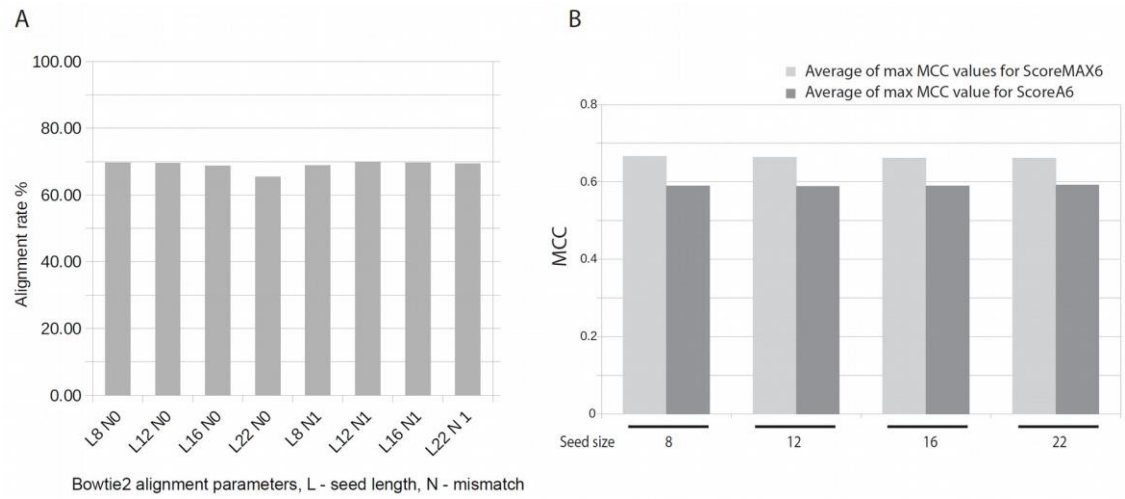


Figure S4 Influence of the Bowtie2 alignment parameters on the proportion of total aligned reads and on the performance of 2'-O-Me detection. Alignment statistics for Sample2- BMSC with different seed length and mismatch (Panel A). Panel B - ScoreMAX (MAX6) gives better performance compared to scoreA6 in detection of 2'-O-Me sites, at all seed length tested. Average values of max MCC calculated for Samples1 - HUVEC, 2 - BMSC and 3 - HeLa.

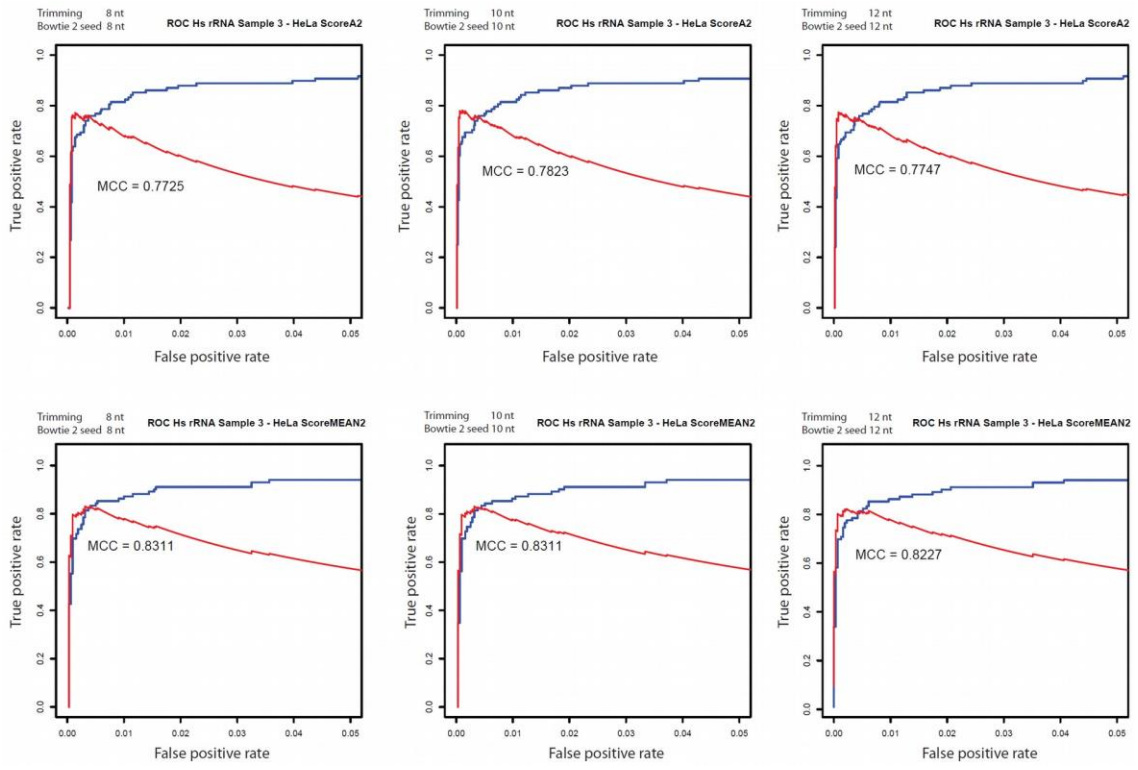


Figure S5 Measured performance for combinations trimming/Bowtie2 parameters (trimming length/Bowtie2 seed length) for ScoreA2 and ScoreMEAN2. Sample 3 - HeLa was used for calculations of ROC curves. Blue curve - TP hits, red curve - MCC. Max MCC value is given on the graph. The Same coordinates as in Figure S1 were used (FPR 0-0.05, TPR 0-1).

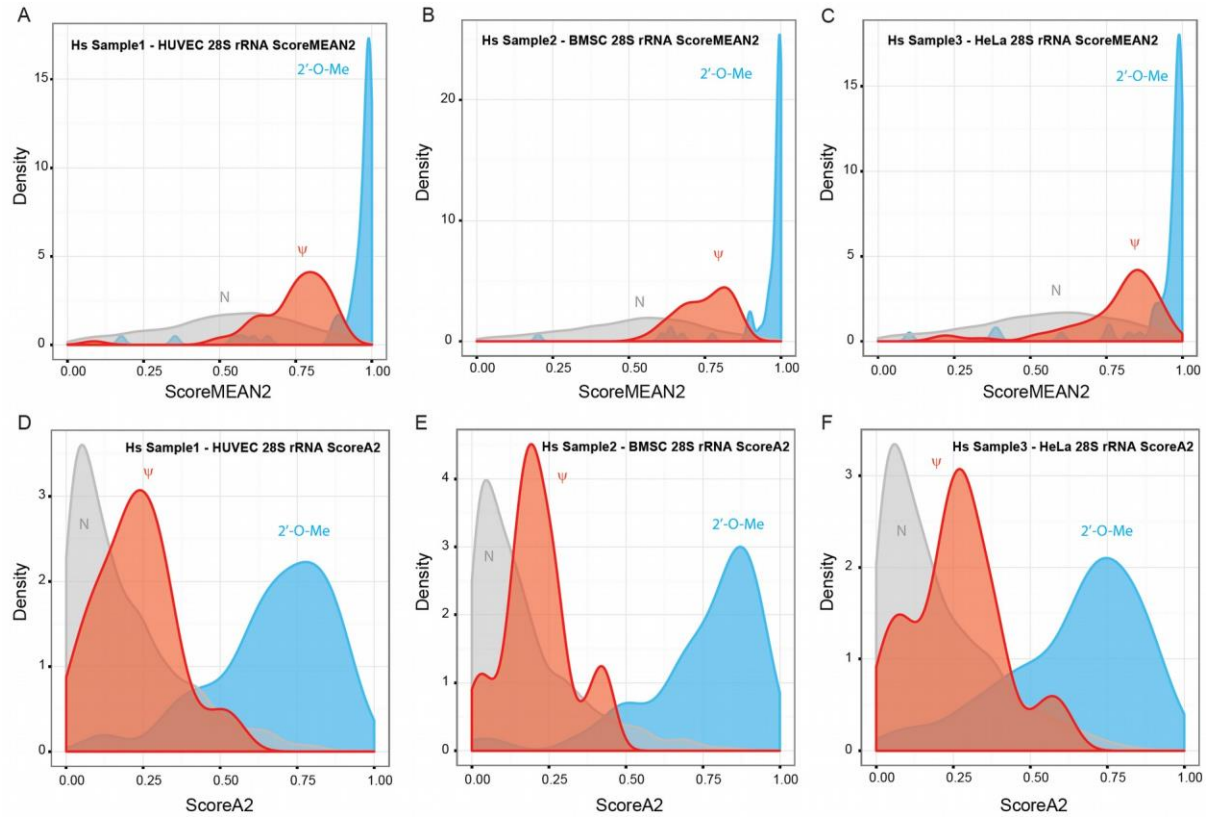


Figure S6 Density plots for singals observed at 2'-O-me positions (light blue), pseudouridines (red) and unmodified nucleotides (gray) in human 28S rRNA. Graphs were made for Sample1 - HUVEC, 2 - BMSC, and 3 - HeLa. Samples 1, 2 and 3 (Panels A/D, B/E and C/F, respectively, and for scoreMEAN2 and ScoreA2 (top and bottom, respectively).

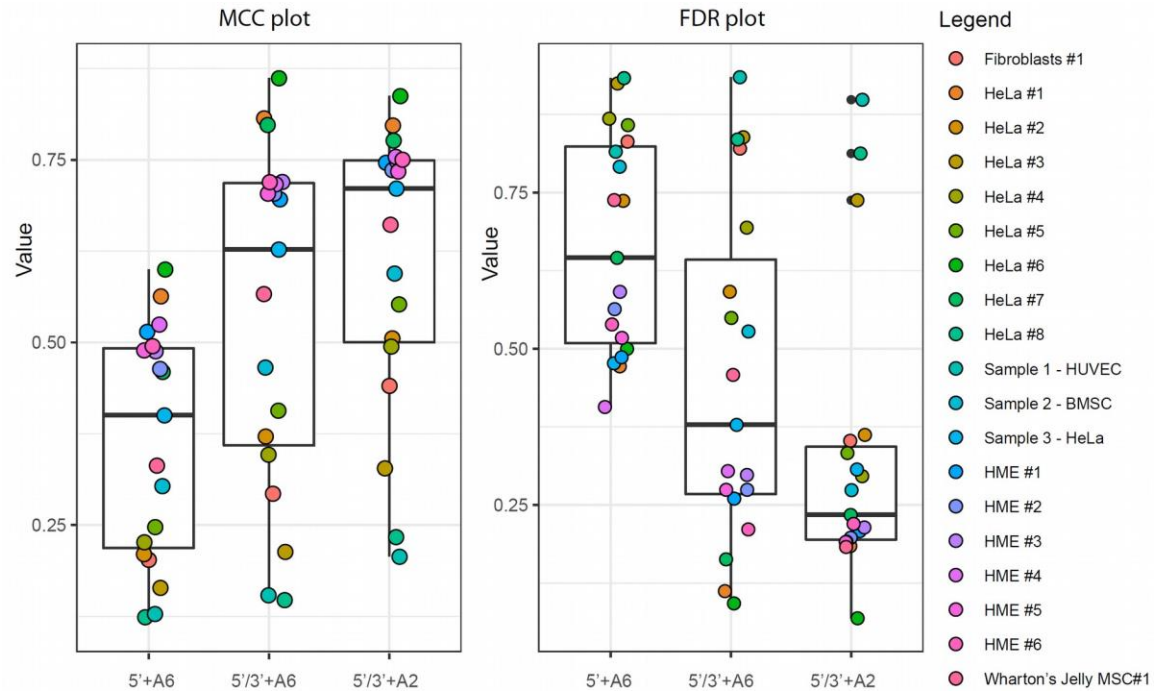


Figure S7 Improvement of ScoreA (A6 and A2) with 5'/3'-counts and reduced calculation window (Score 2 calculation scheme). Boxplot shows max MCC values (left) and associated FDR (right) for all 19 RiboMethSeq dataset used for validation. Identity of the RiboMethSeq datasets is given on the right.

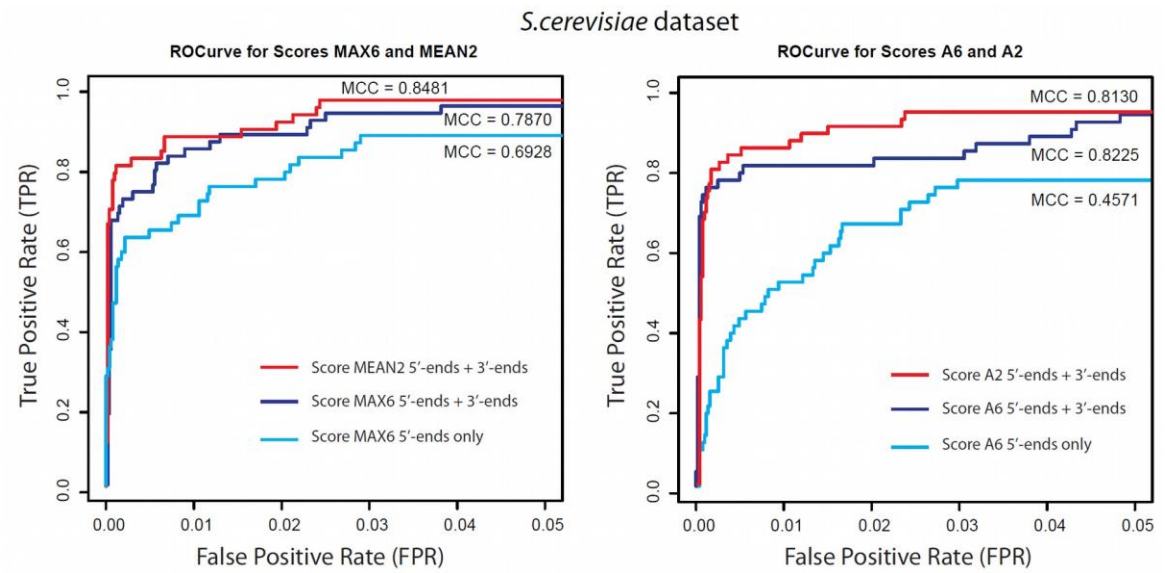


Figure S8 Validation of RiboMethSeq Score 2 calculation scheme using S. cerevisae rRNA datset. Graphs represent zoom to ROC curve 0-0.05 for FPR, 0-1 for TPR. It was previously shown that 5'-end coverage only (light blue curve) is sufficient for reliable construction of RNA protection/2'-O-Me profile, but cumulated 5'- and 3'-ends coverage (dark blue curve) generally provides better discrimination between methylated positions and FP hits. Improvement related to use of Score 2 calculation scheme are shown by red curve and associated MCC value.

Supplementary Tables

Table S1 list of human rRNA modifications considered in this study (reference for human 28S rRNA sequence NR_046235 for the position / GeneBank U13369 for the modifcations numbering)

RNA	position	modification	RNA	position	modification	RNA	position	modification	RNA	position	modification
5.8S	14	Um14	18S	966	Psi966	28S	1862	Psi1849	28S	3944	Gm3923
5.8S	55	Psi55	18S	1004	Psi1004	28S	1871	Am1858	28S	3959	Psi3938
5.8S	69	Psi69	18S	1031	Am1031	28S	2292	Cm2279	28S	4042	Gm4020
5.8S	75	Gm75	18S	1056	Psi1056	28S	2351	Cm2338	28S	4054	Cm4032
			18S	1081	Psi1081	28S	2363	Am2350	28S	4196	Gm4166
18S	27	Am27	18S	1174	Psi1174	28S	2364	Gm2351	28S	4219	m6A4189
18S	34	Psi34	18S	1238	Psi1238	28S	2365	Cm2352	28S	4227	Um4197
18S	36	Psi36	18S	1244	Psi1244	28S	2401	Am2388	28S	4228	Gm4198
18s	93	Psi93	18S	1248	m1acp3P	28S	2415	Um2402	28S	4293	Psi4263
18S	99	Am99	18S	1272	Cm1272	28S	2422	Cm2409	28S	4296	Psi4266
18S	105	Psi105	18S	1288	Um1288	28S	2424	Gm2411	28S	4299	Psi4269
18S	109	Psi109	18S	1326	Pm1326	28S	2508	Psi2495	28S	4306	Um4276
18S	116	Um116	18S	1328	Gm1328	28S	2787	Am2774	28S	4312	Psi4282
18S	119	Psi119	18S	1347	Psi1347	28S	2804	Cm2791	28S	4353	Psi4323
18S	121	Um121	18S	1367	Psi1367	28S	2815	Am2802	28S	4361	Psi4331
18S	159	Am159	18S	1383	Am1383	28S	2824	Cm2811	28S	4370	Gm4340
18S	166	Am166	18S	1391	Cm1391	28S	2837	Um2824	28S	4392	Gm4362
18S	172	Um172	18S	1442	Um1442	28S	2861	Cm2848	28S	4403	Psi4373
18S	174	Cm174	18S	1445	Psi1445	28S	2876	Gm2863	28S	4420	Psi4390
18S	210	Psi210	18S	1447	Gm1447	28S	3637	Psi3616	28S	4423	Psi4393
18S	218	Psi218	18S	1490	Gm1490	28S	3639	Psi3618	28S	4431	Psi4401
18S	406	Psi406	18S	1625	Psi1625	28S	3695	Psi3674	28S	4442	Psi4412
18S	428	Um428	18S	1639	m7G1639	28S	3701	Cm3680	28S	4443	m5C4413
18S	436	Gm436	18S	1643	Psi1643	28S	3715	Psi3694	28S	4456	Cm4426
18S	462	Cm462	18S	1668	Um1668	28S	3718	Am3697	28S	4457	Psi4427
18S	468	Am468	18S	1678	Am1678	28S	3724	Am3703	28S	4471	Psi4441
18S	484	Am484	18S	1692	Psi1692	28S	3730	Psi3709	28S	4494	Gm4464
18S	509	Gm509	18S	1703	Cm1703	28S	3734	Psi3713	28S	4498	Um4468
18S	512	Am512	18S	1804	Um1804	28S	3744	Gm3723	28S	4499	Gm4469
18S	517	Cm517	18S	1832	m6A1832	28S	3758	Psi3737	28S	4500	Psi4470
18S	572	Psi572	18S	1850	m62A1850	28S	3760	Am3739	28S	4521	Psi4491
18S	576	Am576	18S	1851	m62A1851	28S	3762	Psi3741	28S	4523	Am4493
18S	590	Am590				28S	3764	Psi3743	28S	4530	m3U4500
18S	601	Gm601	28S	398	Am389	28S	3768	Psi3749	28S	4531	Psi4501
18s	609	Psi609	28S	400	Am391	28S	3770	m5C3761	28S	4536	Cm4506
18S	627	Um627	28S	1316	Gm1303	28S	3782	Am3764	28S	4552	Psi4522
18S	644	Gm644	28S	1322	m1A1309	28S	3785	Gm3771	28S	4571	Am4541
18S	649	Psi649	28S	1326	Am1313	28S	3792	Cm3787	28S	4576	Psi4546
18S	651	Psi651	28S	1340	Cm1327	28S	3808	Pm3797	28S	4579	Psi4549
18S	668	Am668	28S	1522	Gm1509	28S	3822	Psi3801	28S	4590	Am4560
18S	681	Psi681	28S	1524	Am1511	28S	3825	Am3804	28S	4618	Gm4588
18S	683	Gm683	28S	1534	Am1521	28S	3830	Am3809	28S	4620	Um4590
18S	686	Psi686	28S	1536	Psi1523	28S	3841	Cm3820	28S	4623	Gm4593
18S	797	Cm797	28S	1582	Psi1569	28S	3844	Psi3823	28S	4628	Psi4598
18S	799	Um799	28S	1625	Gm1612	28S	3851	Psi3830	28S	4636	Psi4606
18s	801	Psi801	28S	1677	Psi1664	28S	3853	Psi3832	28S	4637	Gm4607
18S	814	Psi814	28S	1683	Psi1670	28S	3867	Am3846	28S	4673	Psi4643
18S	815	Psi815	28S	1744	Psi1731	28S	3869	Cm3848	28S	4689	Psi4659

Table S2 Human RiboMethSeq datasets used for extensive analysis and validation

	Cell type	Name	Raw reads	Ref ENA
1	Human umbilical vein endothelial cells (HUVEC)	**Sample1**	22,400,000	PRJEB34951
2	Bone marrow stem cells (BMSC)	**Sample2**	22,600,000	PRJEB34951
3	HeLa	**Sample3**	26,400,000	PRJEB43738
4	HeLa	**HeLa #1**	30,700,000	PRJEB43738
5	HeLa	**HeLa #2**	22,300,000	PRJEB35565
6	HeLa	**HeLa #3**	33,100,000	PRJEB35565
7	HeLa	**HeLa #4**	14,000,000	PRJEB35565
8	HeLa	**HeLa #5**	25,000,000	PRJEB35565
9	HeLa	**HeLa #6**	29,700,000	PRJEB43738
10	HeLa	**HeLa #7**	21,600,000	PRJEB43738
11	HeLa	**HeLa #8**	18,600,000	PRJEB35565
12	Human fibroblasts	**Fibroblasts #1**	21,900,000	PRJEB35565
13	Wharton's Jelly MSC (WJ-MSC)	**Wharton's Jelly MSC#1**	25,100,000	PRJEB35565
14	Human mammary epithelial cells (HME)	**HME #1**	13,200,000	PRJEB35565
15	Human mammary epithelial cells (HME)	**HME #2**	13,300,000	PRJEB35565
16	Human mammary epithelial cells (HME)	**HME #3**	16,000,000	PRJEB35565
17	Human mammary epithelial cells (HME)	**HME #4**	11,300,000	PRJEB35565
18	Human mammary epithelial cells (HME)	**HME #5**	14,600,000	PRJEB35565
19	Human mammary epithelial cells (HME)	**HME #6**	17,600,000	PRJEB35565

Table S3 Comparison of performance for old and new RiboMethSeq scores

	5prime A6 MAX6		*5prime3prime A6 MAX6*		*5prime3prime A2 MEAN2*	
sample	**MCC_max6**	**MCC_scoreA6**	**MCC_max6**	**MCC_scoreA6**	**MCC_mean2**	**MCC_scoreA2**
Fibroblasts #1	0.2839	0.2026	0.3605	0.2931	0.4780	0.4407
Sample 1 – HUVEC	0.1546	0.1284	0.2001	0.1540	0.2622	0.2068
Sample 3 – HeLa	0.5802	0.4003	0.7039	0.6273	0.7955	0.7107
HeLa #1	0.6494	0.5632	0.8089	0.8069	0.8249	0.7971
HeLa #2	0.3628	0.2103	0.4175	0.3714	0.5485	0.5058
HeLa #3	0.2340	0.1642	0.2653	0.2134	0.3537	0.3278
HeLa #4	0.3472	0.2266	0.4247	0.3465	0.5669	0.4945
HeLa #5	0.3825	0.2474	0.5084	0.4068	0.6259	0.5523
HeLa #6	0.6788	0.6000	0.8288	0.8618	0.8580	0.8373
HeLa #7	0.6090	0.4598	0.7832	0.7978	0.8398	0.7764
HME #1	0.5295	0.5146	0.7272	0.6960	0.7846	0.7461
HME #2	0.5555	0.4639	0.7344	0.7038	0.7894	0.7360
HME #3	0.5449	0.4880	0.7460	0.7198	0.7855	0.7482
HME #4	0.5453	0.5246	0.7566	0.7166	0.7835	0.7542
HME #5	0.5434	0.4893	0.6921	0.7038	0.7463	0.7341
HME #6	0.5386	0.4951	0.7171	0.7195	0.7776	0.7503
HeLa #8	0.1875	0.1240	0.2192	0.1475	0.2726	0.2337
Sample 2 – BMSC Wharton's Jelly MSC#1	0.4509	0.3315	0.6497	0.5664	0.6820	0.6613

sample	**FDR_max6**	**FDR_scoreA6**	**FDR_max6**	**FDR_scoreA6**	**FDR_mean2**	**FDR_scoreA2**
Fibroblasts #1	0.7966	0.8315	0.7796	0.8204	0.2857	0.3529
Sample 1 – HUVEC	0.9224	0.8154	0.9045	0.9346	0.8270	0.8985
Sample 3 – HeLa	0.3960	0.4769	0.2745	0.3784	0.0405	0.3070
HeLa #1	0.3053	0.4720	0.0732	0.1124	0.0494	0.1845
HeLa #2	0.6262	0.7368	0.5474	0.5914	0.3288	0.3623
HeLa #3	0.8657	0.9245	0.8620	0.8386	0.5000	0.7377
HeLa #4	0.6381	0.8682	0.4868	0.6939	0.3494	0.2963
HeLa #5	0.5876	0.8579	0.5524	0.5495	0.3529	0.3333
HeLa #6	0.1948	0.5000	0.1400	0.0928	0.0353	0.0690
HeLa #7	0.3258	0.6455	0.0286	0.1633	0.0968	0.2342
HME #1	0.4388	0.4865	0.2752	0.2604	0.1596	0.2083
HME #2	0.4924	0.5635	0.2323	0.2745	0.1163	0.1978
HME #3	0.4961	0.5912	0.2569	0.2982	0.1881	0.2143
HME #4	0.4545	0.4070	0.2632	0.3043	0.1176	0.1915
HME #5	0.5274	0.5175	0.3333	0.2745	0.1868	0.1910
HME #6	0.5147	0.5391	0.2946	0.2111	0.1190	0.2200
HeLa #8	0.9131	0.9331	0.8915	0.8350	0.7868	0.8125
Sample 2 – BMSC Wharton's Jelly MSC#1	0.4756	0.7380	0.2237	0.4583	0.2472	0.1831

-- End of article --

4.1.3 Non-Redundant tRNA Reference Sequences for Deep Sequencing Analysis of tRNA Abundance and Epitranscriptomic RNA Modifications. [DOI: 10.3390/genes12010081]

Florian PICHOT[1,2], Virginie MARCHAND[2], Mark HELM[1], Yuri MOTORIN[2,3]

[1] Institute of Pharmaceutical and Biomedical Sciences, Johannes Gutenberg-University Mainz, Staudingerweg 5, 55128 Mainz (Germany)

[2] Université de Lorraine, CNRS, INSERM, IBSLor (UMS2008/US40), Epitranscriptomics and RNA Sequencing Core Facility, F54000 Nancy, France

[3] Université de Lorraine, CNRS, IMoPA (UMR7365), F54000 Nancy

Overview

tRNAs are essential molecules implied in translation. Information about their sequence and modifications through sequencing methodologies is of the utmost interest to explore gene expression mechanisms. Unfortunately, little information is given in epitranscriptomics studies regarding tRNA reference used for analysis, usually designed as “reference extracted from GtRNAdb/tRNAdb 2009”[86, 325–327] without further input.

In fact, there are two kinds of tRNA sequence databases used for tRNA projects. The first one such as tRNAdb 2009 or MODOMICS database is based on experimentally verified tRNA sequences with their complete epitranscriptome annotation[328, 329]. However, they are incomplete since the amount of relevant tRNA molecules can be estimated at slightly over 40 across different species[330] and tRNAdb 2009 generally provides less sequences (around 20 for *H. sapiens* for exemple). The other kind of database such as gtRNAdb is based on *in sillico* tRNA genes predictions from genomic sequence context and computational processing them into their RNA counterparts[331]. This latter reference is used in most epitranscriptomics projects[332, 333] thanks to the high amount of tRNA sequences available. However, each research group adds different modifications to the different genomic tRNA references, making each personalised reference different from one group to another. This lack of consensual tRNA reference generates difficulties in replicating and comparing results from different studies.

In regards to that matter, we decided to generate tRNA reference in a semi-automated way using tRNA genes predictions from gtRNAdb. Then we remove duplicated sequences and regroup all similar sequences into different groups called clusters. Subsequently, a consensus sequence representative of each cluster is created and their compilation gives a tRNA sequence reference which minimizes ambiguous alignments and is therefore suitable for epitranscriptomics studies. At last, since these references are created only from sequence context, validation through alignment of *in situ* sequencing data is carried on to ensure their application and allows their optimisation from biological data.

As a result, these compilations of consensus sequences increase the overall alignment to tRNAs to at least 83% after rRNA removal, and provide more than 85% of unambiguous mapping from different model organisms, substantially increasing the amount of tRNA sequences available for further studies.

Finally, the semi-automated procedure and scripts for the creation of such references are also described and available online in GitHub for reproducibility and transparency.

In this study, I have designed the method and the R pipeline to regroup similar tRNA sequences into clusters, create consensus sequences for each cluster and validate them from experimental sequencing data. I also took care of the GitHub page for reference and R scripts disposal and finally, participated in the writing process of the results parts and produced related graphs.

Abstract: Analysis of RNA by deep-sequencing approaches has found widespread application in modern biology. In addition to measurements of RNA abundance under various physiological conditions, such techniques are now widely used for mapping and quantification of RNA modifications. Transfer RNA (tRNA) molecules are among the frequent targets of such investigation, since they contain multiple modified residues. However, the major challenge in tRNA examination is related to a large number of duplicated and point-mutated genes encoding those RNA molecules. Moreover, the existence of multiple isoacceptors / isodecoders complicates both the analysis and read mapping. Existing databases for tRNA sequencing provide near exhaustive listings of tRNA genes, but the use of such highly redundant reference sequences in RNA-seq analyses leads to a large number of ambiguously mapped sequencing reads. Here we describe a relatively simple computational strategy for semi-automatic collapsing of highly redundant tRNA datasets into a non-redundant collection of reference tRNA sequences. The relevance of the approach was validated by analysis of experimentally obtained tRNA-sequencing datasets for different prokaryotic and eukaryotic model organisms. The data demonstrate that non-redundant tRNA reference sequences allow improving unambiguous mapping of deep sequencing data.

1. Introduction

Transfer RNA molecules (tRNAs) are essential adaptors in mRNA decoding, whereby these small ncRNA species are in charge for implementation of the genetic code at the molecular level [334, 335]. These RNAs are heavily modified during post-transcriptional maturation steps [292, 336] and, in consequence, analysis of these RNA species is of particular interest both for studies of mRNA translation [337, 338], and for analysis of their epitranscriptomic modifications [339–341]. tRNAs are relatively abundant in the cell and represent ~10-15% of the total RNA content, depending on the organism [342]. Due to the presence of multiple species of similar size and even similar sequence, isolation of individual tRNA species for analysis is possible, but relatively laborious. The most efficient methods are based on specific hybridization of complementary biotinylated DNA oligonucleotide, followed by affinity separation and specific elution [343, 344]. Alternative methods of tRNA analysis, like microarrays or microscale thermophoresis, give excellent results in quantification of different species in a pool [345, 346] but are not high throughput techniques. Taking all these considerations into account, the use of deep sequencing has become the most popular way for analysis of tRNA species, their relative abundance and also their modification profiles.

Analysis of tRNAs by deep sequencing is rather straightforward and may be performed directly from total RNA. For better efficiency, the tRNA population may be fractioned either by size selection (on agarose or polyacrylamide gels) or by selective extraction from cells [344]. Depletion of rRNA by Ribo-Zero™ or equivalent “subtractive” methods to obtain tRNA enriched fraction can be used as well [344]. Library preparation is performed either by direct ligation of adaptors [347, 348] to 5’- and 3’-ends of tRNA (such

ligation protocol is also implemented in the majority of small RNA kits for library preparation) [340], or by primer extension using 3'-end ligated tRNA and tagging of the resulting cDNA [316, 321, 349]. In some protocols, preliminary fragmentation of tRNAs is used for analysis (e.g. in Hydro-tRNASeq and RiboMethSeq protocols) [64, 65, 350].

Whatever protocol is chosen for library preparation from tRNA fractions, the resulting sequencing reads have to be processed by alignment to the appropriate reference sequence, to allow meaningful biological interpretation. Thus, a major challenge in tRNA sequencing consists in construction of such precise and unambiguous alignment for further counting of tRNA reads and/or analysis of their post-transcriptionnal modifications.

Multiple approaches were proposed as alignment strategies for tRNAs. Existing tRNA-oriented databases (the most popular are gtRNAdb [331] and tRNAdb-CE 2011 [351]), collect already published information and also propose prediction of tRNA genes in multiple species, generally using tRNAScan-SE [352, 353]. These sources of information are extremely helpful, but, as a reference sequence, they propose either full genomic datasets of all tRNA genes or non-redundant collection of all possible tRNA species found (and predicted) in the genomic DNA. These reference sequences are certainly complete, but their direct use in bioinformatic analyses leads to numerous ambiguously (multiply) aligned reads. On the other hand, databases collecting experimental data on already sequenced tRNAs (MODOMICS [30], tRNAdb [328] and also T-psi-C database [354]) propose only validated sequences of existing and presumably most abundant tRNA species, however, for many reference/model organisms this information is scarce or simply not available (e.g. *A. thaliana* or *D. radiodurans*).

These limitations have so far prevented a generalized approach or common reference sequence for routine tRNA analysis by deep sequencing. Published analytical pipelines either use full genomic collection of tRNA species [339, 340] (or better, non-redundant collections where identical species are already collapsed) [338, 341] or apply rather laborious and multistep analytical pipelines involving alignment to full genomic DNA and extraction of mapped tRNA reads by coordinates, with subsequent collapsing of identical sequences at that stage [281, 347, 348, 350, 355]. These approaches give excellent results, but have been applied only to limited number of living species. Moreover, such complex multistep pipelines are not truly compatible with one another, precluding a direct comparison of the reported results. Also, they are not suitable for routine analysis of tRNA pools in biomedical projects or for analysis of tRNA modification profiles when multiple species are involved.

We propose a simple and reproducible algorithm as a tool for semi-automatic analysis of tRNA datasets starting from full and redundant genomic references identified by tRNA-ScanSE. The dataset, downloaded as fasta file (*.fa) from gtRNAdb or tRNAdb-CE 2011 was first collapsed in full automatic mode, to obtain a non-redundant tRNA

reference. This reference may be used to extract all tRNA-related reads from experimental tRNA sequencing data. Full non-redundant tRNA reference was further analyzed in a semi-automatic way, to collapse closely related tRNA sequences in a single tRNA reference. The distance for collapsing into a unique entry is proposed by the algorithm and modified or validated by the user after visual inspection of the pairwise Levenshtein distance heatmap. These optimized non-redundant reference sequences were validated using experimental data available for several model living species and can be employed for routine analysis of tRNAs and their modifications.

2. Materials and Methods

2.1. Library preparations

All deep sequencing tRNA libraries were prepared using the same protocol, implemented in the RiboMethSeq procedure [64, 356]. Total RNA (or size-selected tRNA fraction) was subjected to fragmentation under strong alkaline conditions (10-15 min in 50 mM bicarbonate buffer pH 9.2 at 96°C). Fragments were de-phosphorylated at the 3'-end and re-phosphorylated at the 5'-end to insure their compatibility with a direct adapter ligation step. Further steps were performed using the NEBNext Small RNA kit (E7330L) following the manufacturer's recommendations. Quality of the libraries was assessed by HS DNA chip (Bioanalyzer 2100 Agilent). Sequencing of 50 nucleotides of the insert was performed in single-read SR50 mode on HiSeq 1000 Illumina sequencer. Target number of raw reads was 10-25 millions, depending on the source of RNA (see supplementary Table S1 for the exact number of sequencing reads used for analysis). Both total RNA and enriched tRNA fractions can be used for tRNA analysis, however, sequencing of total RNA should be done at much higher depth, to obtain sufficient amount of reads for all tRNAs. Since tRNAs represent ~ 5-15% of total RNA, ~ 10 times higher sequencing coverage is generally required in such cases.

2.2. Computations

2.2.1. tRNA reference sequence

Analysis of tRNA reference sequences was performed in R-studio with R ver 3.5.3 and R packages seqinr and msa. Initial tRNA sequences (mature reference sequences) were downloaded from gtRNAdb (http://gtrnadb.ucsc.edu/GtRNAdb2/index.html), intron-containing sequences were not included in the analysis. No modification of the reference (except U->T conversion to get standard DNA *.fasta file) was performed prior to downstream analysis. Conserved 3'-CCA tRNA sequence was not introduced at this stage since 3'-CCA can be already encoded in genomic sequences, like in *Escherichia coli*, and thus included in the sequences from gtRNAdb. However, this is

not the case for *Homo sapiens*, and many other eukaryotes. In most complex cases like for *Bacillus subtilis*, some tRNA genes encode 3'-CCA while the others do not. This prevents the automatic addition of 3'-CCA to the reference sequence and a manual verification of tRNA cloverleaf structure and alignment is mandatory.

The first step of analysis consists in the collapsing of identical sequences by a selection according to the amino-acid's specificity, since 'unexpected' anticodons (see definition in gtRNAdb) are relatively rare and eventual existence of such anticodon-mutated tRNA species is verified at the final treatment step. Identical sequences are identified and merged together (NoDuplicated tRNA reference/Step1) in a fully automatic mode (see Figure 1A).

The second step aims to identify closely related tRNA species and collapse them into a unique consensus sequence, replacing ambiguous nucleotides by the nucleotide found in majority of sequences or by N, if the number is equivalent. We decided not to introduce IUPAC codes for ambiguous DNA residues since these codes are not systematically interpreted by popular alignment algorithms (like Bowtie/Bowtie2 [318] used in this work). Collapsing into clusters is performed in semi-automatic mode, and a heatmap showing calculated Levenshtein distances between sequences is displayed to help the user to define the appropriate number of clusters. The script suggests the optimal number of distant clusters (with default max distance of 8 substitutions), but, if required, this number can be overridden by manual entry of a more appropriate value. Such manual correction is optional for the majority of simple bacterial and lower eukaryote tRNA references, but was found to be necessary for very complex clusters found for tRNA genes in higher eukaryotes. If the number of clusters to create is erroneously selected too low, the tRNA sequences, forced to merge in the same cluster, but still distant more than 10 substitutions are removed and stored in the "Removed sequence" file.

After collapsing to clusters, a final heatmap is created for visual inspection and verification, sequences are re-annotated to use only the amino acid identity and anticodon sequence, isoacceptors/isodecoders with the same anticodon are numbered sequentially.

Calculation scripts were tested for 20 organisms listed in the Table 1, Step 1 for merging identical sequences worked in all tested cases, Step 2 failed only when the number of fully identical sequences was > ~100, due to the excessive length of fasta file header, not accepted by R package seqinr.

When the semi-automatic clustering of tRNA sequences was not sufficient to obtain <10-15% of multiply mapped reads after analysis of real tRNA datasets, tRNA reference was manually inspected and redundant isoforms creating ambiguous mapping were removed or collapsed in a single entry.

Finally, if not present in the genomic tRNA sequences, 3'-CCA end was manually added to the reference, to improve alignment quality in an End-to-End mode with sequencing reads issued from mature CCA-containing tRNA species. If present, anticodons with A34NN sequence were replaced by G34NN since A->I conversion at position 34 creates the residue behaving as G in base pairing.

2.2.2. Alignment of the experimental datasets to reference tRNA sequences

Available experimental datasets obtained for total RNA or enriched tRNA fractions were treated by the same analytical pipeline. Demultiplexing and trimming was done without option for removal of potential PCR duplicates, since this is not appropriate for large datasets generated from relatively short reference sequence (here collection of tRNAs). Raw reads were trimmed using Trimmomatic ver 0.32 [317], using the following parameters (minLen 8 nt, maxLen 50 nt, single-end mode, stringency 7 (with these parameters the sequence of adapter over 10 nt is detected and removed). Taking into account very short cumulated reference sequence (maxi 250 sequences of 75-90 nt only, so < 15 kb) minLen of 8 nt is, in principle, sufficient for unambiguous mapping. However, in practical, a proportion of very short reads may be still ambiguously aligned. Only reads <40 nt were selected for analysis to make sure that used raw reads are not contaminated by residual adapter sequence. Alignment of short reads was first performed to complete rRNAs sequence of the organism, since rRNAs are always dominating RNA species in total RNA and, very frequently, even in presumably 'enriched' tRNA fraction. Alignment was done by Bowtie2 v.2.2.4 [318], in sensitive (-D 15 -R 2 -N 0 -L 10 -i S,1,1.15) End-to-End mode (no soft-trimming). Reads, non-aligned to rRNA sequences, were retained for further analysis, aligned to the NonDuplicated tRNA reference /Step 1 sequences and all aligned reads were conserved ('tRNA-mapped reads'). Unique and multiply mapped reads for each tRNA were counted. The second tRNA alignment of all 'tRNA-mapped reads' was done with an optimized tRNA reference (Step2 sequences), with the same Bowtie2 parameters. Unique and multiply mapped reads by tRNA were counted. Reads, non-aligned to optimized reference sequence, were conserved and re-aligned to NonDuplicated tRNA reference / Step 1 sequences to identify tRNA species 'missing' in the optimized reference.

2.3. Practical guidelines for optimization of the reference tRNA sequences

For practical analysis of tRNA dataset for an organism, we suggest the following pipeline:

1. Collapse full genomic tRNA dataset in collection of non-redundant sequences (automatic mode, Step 1).

2. Verify the number of non-duplicated sequences (Step 1), numbers of <60 indicate almost non-redundant dataset, higher values are indication of ambiguous redundant sequences.

3. Use the distance of 8 nt for genomic datasets of <250 tRNA genes (<60 Step 1 sequences) and the distance of 10 nt for larger genomic references. There may be intermediate cases for organisms having between 250 and 300 tRNA genes.

4. Verify the number of optimized (Step2) sequences, values close to 40 (or less) are indication of a good quality non-redundant reference, while numbers > ~50 mean still complex and potentially redundant tRNA collection.

5. Validate the optimized (Step 2) reference with experimentally obtained tRNA dataset. If proportion of uniquely mapped reads is still <90%, repeat collapsing in Step 2 with increased distance threshold.

3. Results

3.1. A two-steps algorithm for tRNA analysis

Depending on the complexity of the organism and the level of duplication, a full genomic tRNA reference sequence may be relatively simple (<100 tRNA genes in many bacteria and archaebacteria), of intermediate complexity (~120-250 tRNA genes in many bacteria and lower eukaryotes), or extremely complex with over 550 tRNA genes in *H. sapiens*/*M. musculus*, *A. thaliana* and other higher eukaryotes (Table 1 and Supplementary Figure S1) [357]. This increased complexity generally results from duplication of the existing tRNA genes occasionally featuring point mutations in non-essential regions and not from the appearance of new and thus very different tRNA species [346, 358]. In simple and intermediate cases, elimination (or collapsing them into one sequence) of redundant (fully identical) tRNA species is generally sufficient to obtain appropriate tRNA reference and further semi-automatic step to find and merge tRNA isodecoders with a limited number of point mutations is rather straightforward (Figure 1).

3.2. Analysis of simple tRNA references (<100 tRNA genes)

For common bacterial species, the proposed approach was validated using tRNA references for *E. coli* and *B. subtilis*. The full genomic *E. coli* tRNA reference contains 89 tRNA genes (source gtRNAdb). Fully identical duplicated genes were found and collapsed, creating a non-redundant tRNA dataset of 48 tRNA sequences (‘Step1’, see Table 1). Further collapsing of closely related tRNA species with the maximal distance of 8 substitutions gave an optimized tRNA reference of 41 tRNA genes. The representativeness of tRNA anticodons was verified and compared to the initial full dataset. All anticodons were found to be correctly represented, except for

$tRNA^{Ala}$(GGC) and $tRNA^{Ala}$(TGC) which were merged in a single entry $tRNA^{Ala}$(NGC) with 7 substitutions and two $tRNA^{Thr}$(CGT) isodecoders which were maintained, since relatively distant (see Supplementary data). The optimized reference sequence for *E. coli* contains all tRNA species reported in MODOMICS and tRNAdb and is similar to the one used previously for tRNA modification analysis [65].

Similarly, collapsing of the full tRNA gene set from *B. subtilis* (86 sequences) gave 43 non-redundant tRNA sequences at the first step and 35 optimized tRNA references with a maximum of 8 allowed substitutions. The resulting optimized reference is fully representative for all used codons and adds 11 additional tRNA isoacceptors compared to MODOMICS/tRNAdb reference databases.

Additional tests with other bacterial species demonstrated that semi-automatic collapsing is sufficient to obtain highly optimized tRNA references in such cases (see Table 1). The minimal number of retained tRNA genes is 31 (which is close to anticipated minimum required for correct mRNA decoding), and the average is about 40 species in low complexity genomic reference (<100 tRNA genes).

3.3. tRNA references of intermediate complexity (<300 genes)

Collections of tRNA gene reference sequences of intermediate complexity are found in lower eukaryotes or representative insect species, like *S. cerevisiae* (275 tRNA genes in full genomic reference) or *D. melanogaster* (total of 290 tRNA genes). For *S. cerevisiae*, an automatic merging of identical tRNAs reduced the number of tRNA sequences to 54 species and further optimization by collapsing of sequences with minimal distance of 8 substitutions reduced this number to 38. Three closely related tRNA pairs showing very short distance (3-4 substitutions only, $tRNA^{Ser}$(CGA/TGA), $tRNA^{Glu}$(CTC/TTC) and $tRNA^{Gln}$(CTG/TTG)) were merged together, the other codons are represented as in the full genomic tRNA dataset. All sequences listed in MODOMICS /tRNAdb are present in the optimized reference sequence, but this last reference contains 10 supplementary tRNA species compared to MODOMICS /tRNAdb.

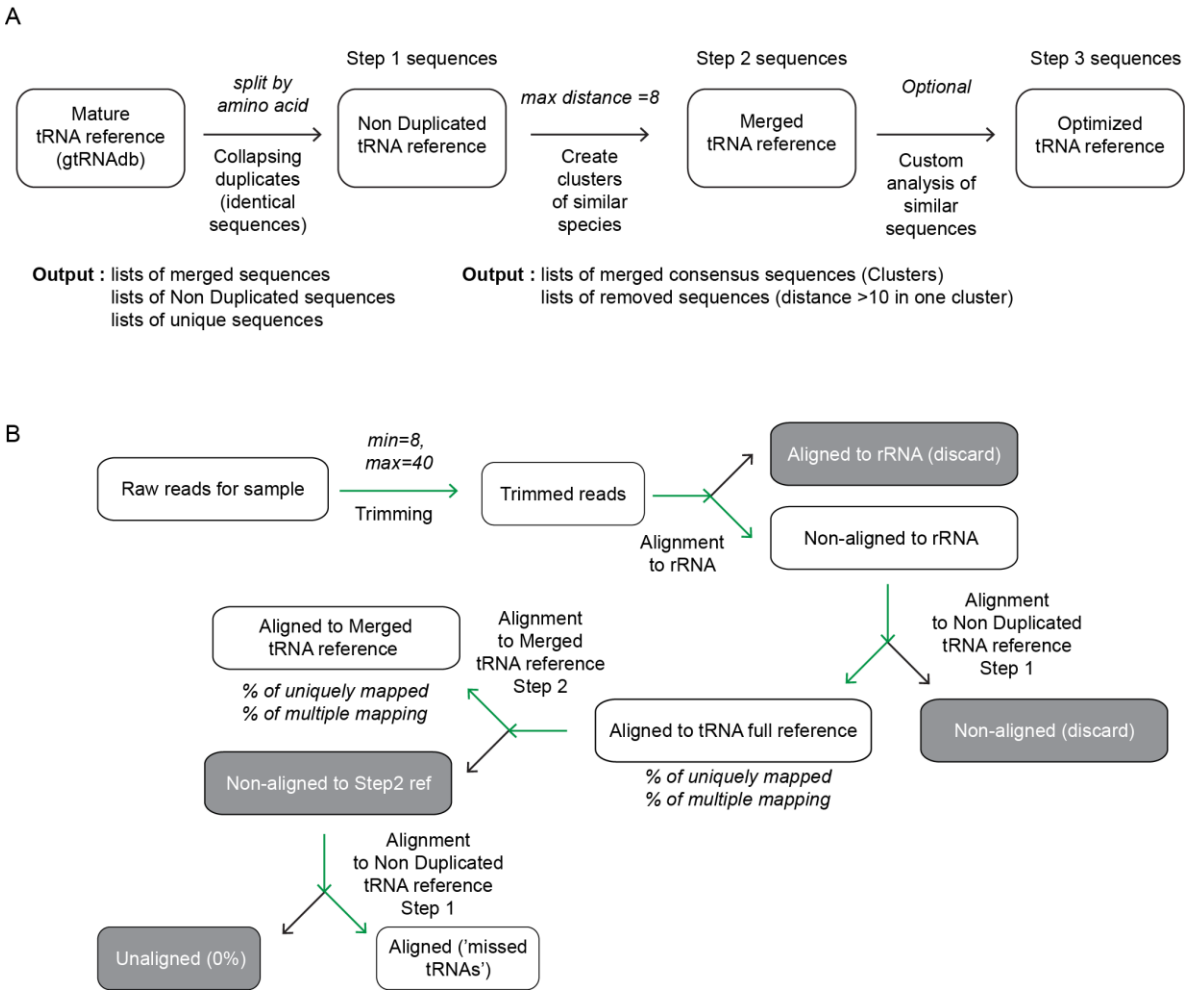


Figure 1. Strategy to create and validate a tRNA reference sequence. Panel A – Collapsing of tRNAScan-SE predicted tRNA genes in NonDuplicated (non-redundant) tRNA reference followed by further optimization (Step2) by merging of closely related sequence (distance <=8 by default). Further improvements may be introduced in a manual mode, by analysis of residual redundant mapping. Panel B – Experimental analysis and validation of proposed tRNA reference sequences. Raw sequencing reads obtained in SR50 mode are trimmed and first filtered by alignment to rRNA reference sequence of the organism. Non-mapped reads are aligned to Step1 (non-redundant) tRNA sequences. Mapped reads are re-aligned to Step2, optimized reference and % of total and unique mapping are calculated. Finally, non-mapped Step2 reads are re-analyzed by alignment to Step1 reference, to get 'missing tRNAs'.

Table 1 tRNA reference datasets for selected model organisms

Organism	Domain	"High confidence set" gtRNAdb[1]	Step1 Non-redundant tRNA set	Step2 Optimized "collapsed" tRNA set (d=8)	Step3 Validated "optimized" tRNA set
Plasmodium falciparum 3D7	E	45	45	40	
Sulfolobus acidocaldarius N8	A	50	50	33	
Haloferax volcanii DS2	A	52	46	46	
Staphylococcus aureus subsp. aureus NCTC 8325	B	59	**43**	**31**	
Leishmania major strain Friedlin (ASM272v2)	E	82	49	44	
Bacillus subtilis subsp. subtilis str. 168	B	86	50	35	33
Escherichia coli str. K-12 substr. MG1655	B	89	48	41	39
Candida albicans A20	E	129	50	41	
Candida albicans WO-1	E	146	73	58	
Schizosaccharomyces pombe 972h-	E	171	60	44	

Saccharomyces cerevisiae PW5	E	171	47	42	
Candida glabrata CBS 138	E	189	45	38	
Candida tropicalis 121	E	203	70	57	
Saccharomyces cerevisiae P301	E	226	49	37	
Saccharomyces cerevisiae S288c	E	275	54	38	38
Drosophila melanogaster (BDGP Rel. 6/dm6)	E	290	76	37 35 (d=10)	34
Homo sapiens (GRCh37/hg19)	E	416	177	61 45 (d=10)	43
Bombyx mori (Domestic silkworm ASM15162v1)	E	435	115	44	
Arabidopsis thaliana (TAIR10 Feb 2011)	E	580	139	48	43
Zea mays B73 (RefGen_v4 AGPv4)		771	191	**70**	
Strongylocentrotus purpuratus (S. purpuratus Mar. 2015 Spur_4.2)	E	931	192	61	

*Xenopus_tropicalis*_v9.1	E	3010	**245**	68

[1] Table is sorted in ascending order of predicted 'high-confidence' tRNA genes (source gtRNAdb).

For the *D. melanogaster* reference (290 tRNA genes), collapsing of fully identical sequences gave 76 tRNA genes and further optimization reduced this number to 37 representative sequences. This list will be further reduced to 34 (see point 3.5 and Supplement for more details). All sequences listed in MODOMICS/tRNAdb are also present in this reference and it contains 20 additional tRNA sequences.

Analysis of other genomes from this group showed that the minimal number of tRNA genes retained in optimized references is close to 35, while the maximum increases to about 60, indicating considerable diversity (see Table 1). Noteworthy, the numbers of tRNA species in non-redundant reference and in the optimized one are not directly proportional to a total number of tRNA genes in the genome.

3.4. Highly complex tRNA references (>400 genes)

Higher eukaryotic genomes contain substantial numbers of tRNA genes (see Table 1), mostly due to duplicated sequences and multiple point mutants. Depending on the genome build, human tRNA reference sequence lists >400 tRNA genes, even more tRNA sequences were found in the *C. elegans* worm genome (568 in "high confidence set", gtRNAdb) and the *A. thaliana* plant genome has 580 sequences in "high confidence set". Even after collapsing of identical tRNA sequences, the residual non-redundant tRNA reference remains very large (>150 non-identical tRNA genes). However, these genes form large clusters of related sequences. Here we analyzed *H. sapiens* and *A. thaliana* tRNA genes in more detail, since human tRNAs were in a large proportion directly sequenced in the past and included in MODOMICS and tRNAdb, while only very limited information exists on the tRNAs from *A. thaliana*. The human non-redundant tRNA reference has 177 non-identical tRNA genes, while collapsing with the default 8 nt distance gave 61 entries in Step2 Optimized sequences. Despite a larger genomic tRNA gene collection (590 sequences), *A. thaliana* tRNA genes were found to be less diverse, leading to only 133 entries in non-redundant Step1 collection and 48 after Step2 (see Table 1).

The algorithm was also tested for highly complex models, like plant *Z. mais* (770 tRNA genes) and frog *X. tropicalis*, where >3000 tRNA genes were found. Collapsed reference contains, respectively, 191 and 245 non-redundant tRNA species, while "optimized dataset" uses only 70 and 68 representative tRNA sequences (Table 1).

3.5. Validation of the optimized tRNA reference sequences

For an experimental validation we selected 6 representative tRNA reference sequences: two from bacterial, common model organisms *E. coli* and *B. subtilis*, respectively from Gram- and Gram+ groups, and 4 from eukaryotic species showing rather variable number of tRNA genes in the genome and optimized tRNA references: *S. cerevisiae*, *D. melanogaster*, *H. sapiens* and *A. thaliana* (see Table 1, in red). For each organism, RiboMethSeq-like datasets were prepared for 6-20 samples, from total RNA, or from size-selected tRNA fraction. Alignment statistics for analyzed datasets is shown in Table S1.

Raw trimmed reads were cleaned up from rRNA contamination and aligned first to NonDuplicated (non-redundant) full tRNA reference (Step 1, see Figure 1B). Mapped reads were extracted and used for further analysis by alignment to an optimized tRNA reference (Step 2). In both cases uniquely and multiply mapped reads were counted, globally, as well as by individual tRNA.

The results shown in Figure 2 (and Supplementary Figure S2) demonstrated that for all species tested, >92-95% of total tRNA reads were still aligned to optimized reference ('Aligned Step2'), showing that collapsing to clusters did not reduce alignment quality. However, the proportion of the uniquely mapped reads increased very drastically with Step2 reference sequences, with concomitant decrease of ambiguous multiple mapping (Figure 2, compare Uniquely and Multiply aligned for Step1 and Step2). In the best cases, the residual multiple mapping was <5%, and routinely <10% for simple tRNA references. Only minor further adaptations were necessary to remove/manually collapse residual sequences with ambiguous mapping from *E. coli, B. subtilis* and *A. thaliana* references (indicated in red on Figure 2 and described in Supplement).

However, when the standard distance of 8 substitutions was applied to human and drosophila tRNA references, residual ambiguous mapping was still quite substantial (>15%) (see supplementary Figure S2). This is certainly related to the excessive number of rather similar but non-merged tRNA sequences. Thus, we explored a semi-automatic fusion with an increased number of allowed substitutions (up to 10). Manual correction of the suggested cluster number was systematically required for such analysis. The final reference for *H. sapiens* contained only 45 sequences, instead of 61 for a maximal distance of 8 substitutions (see Table 1). Similarly, the drosophila tRNA reference sequence was reduced to 35 unique entries. Analysis of uniquely and multiply mapped reads from experimental datasets showed substantial improvement compared to the previous reference sequence with n=8 as maximum distance. The proportion of multiply mapped reads decreased to ~10-12%, which was considered acceptable (see Figure 3).

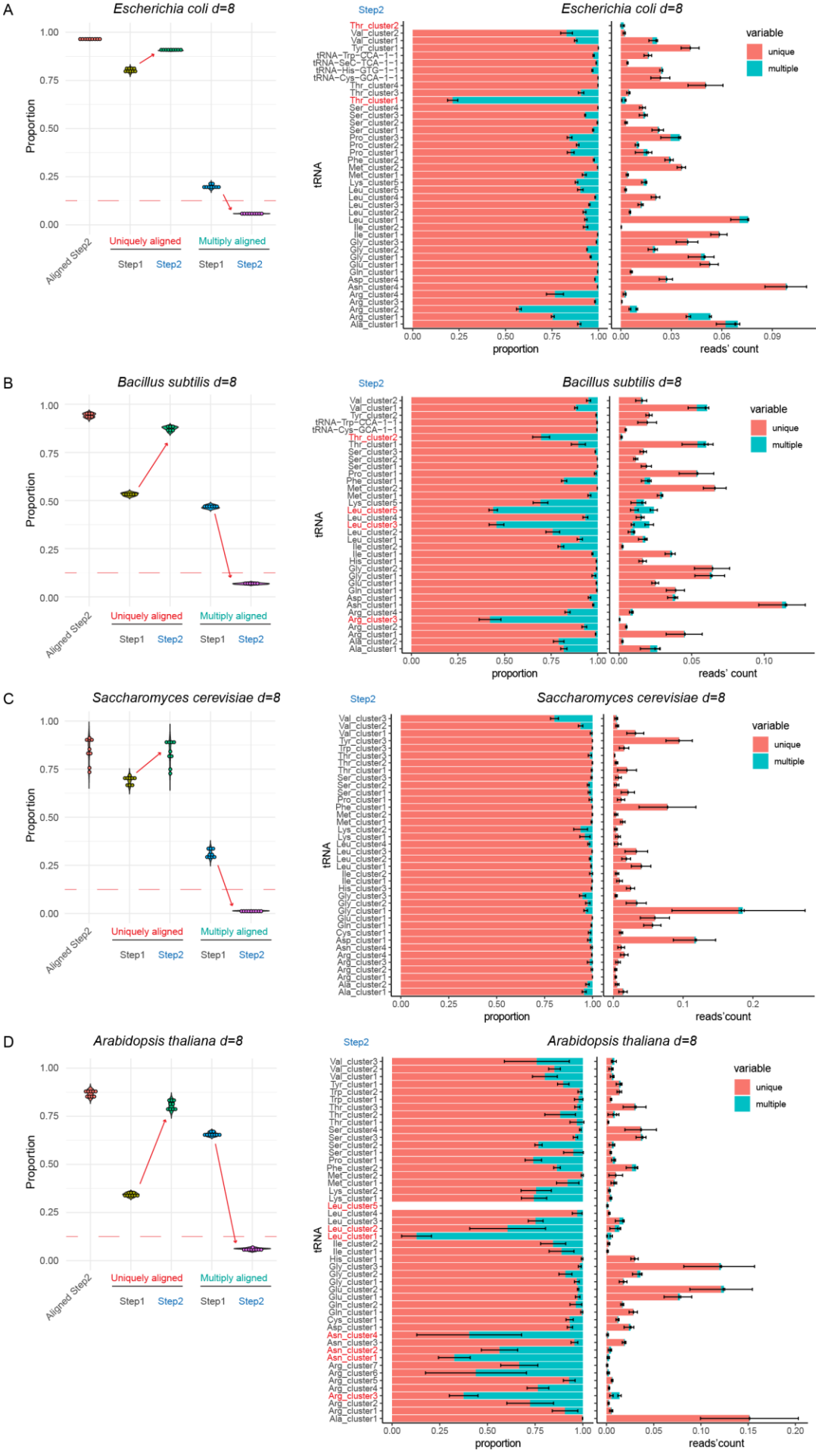

A
Escherichia coli d=8
Proportion
Aligned Step2
Uniquely aligned
Multiply aligned
Step1
Step2
Escherichia coli d=8
tRNA
proportion
reads' count
variable
unique
multiple
B
Bacillus subtilis d=8
C
Saccharomyces cerevisiae d=8
reads'count
D
Arabidopsis thaliana d=8

Figure 2. Alignment results for NonDuplicated (non-redundant) tRNA reference (Step1) and optimized tRNA set (Step2) for 4 model species, with a maximum distance of 8 substitutions. Boxplot on the left shows the proportion of tRNA sequencing reads aligned to Step2 reference ('Aligned Step2') and proportions of uniquely and multiply mapped reads at both steps. Red dashed lines indicate the 12.5% level. The increase of unique mapping events and the decrease of multiple mapping events are shown by arrows. Barplots at the right represent unique and multiple mapping by tRNA species at Step2, in proportion to total and in absolute number of sequencing reads obtained by tRNA, expressed as proportion to total number of mapped reads. tRNAs showing excessive proportion of ambiguous mapping are shown in red.

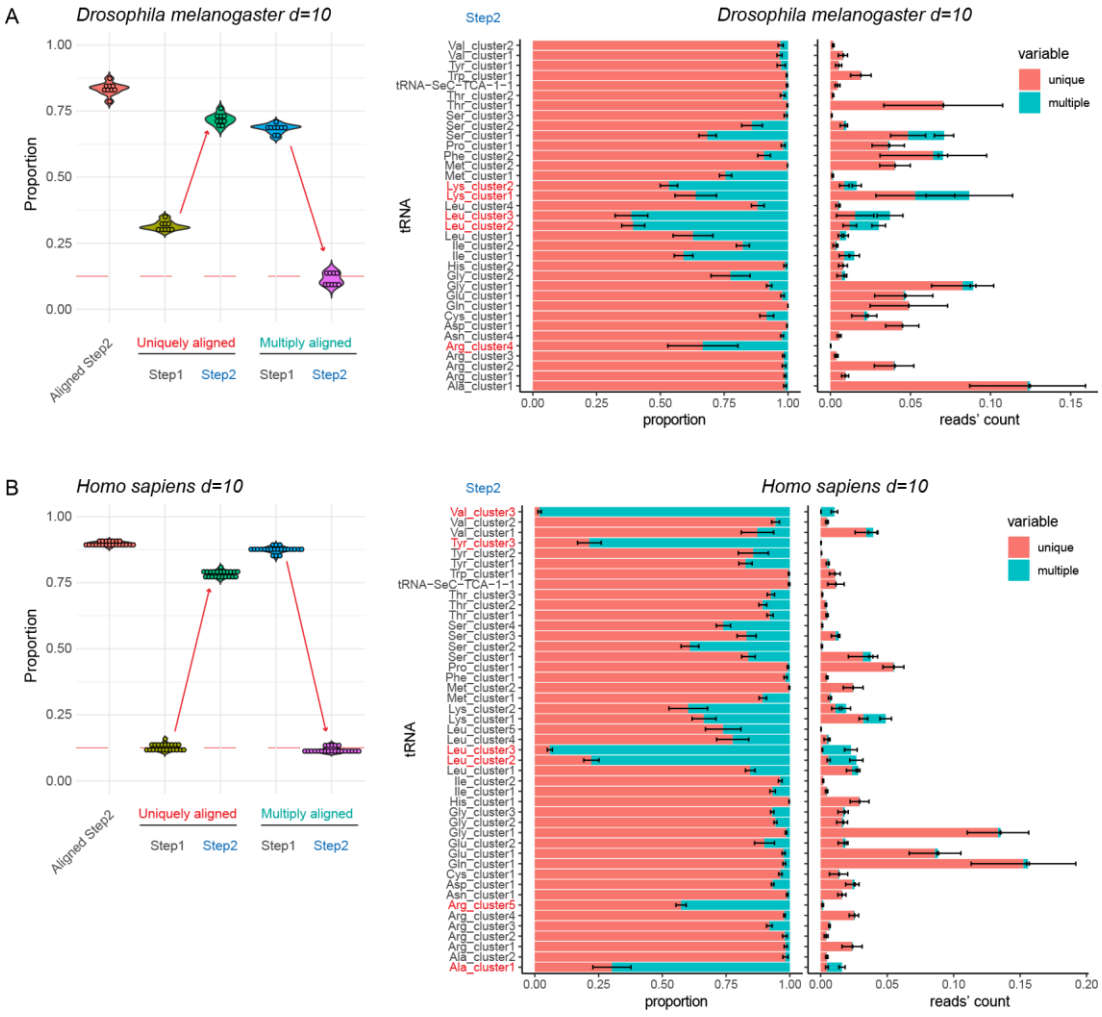


Figure 3. Alignment results for NonDuplicated (non-redundant) tRNA reference (Step1) and optimized tRNA set (Step2) for 2 complex model species, maximal distance used is 10 substitutions. Boxplot on the left shows the proportion of tRNA sequencing reads aligned to Step2 reference ('Aligned Step2') and proportions of uniquely and multiply mapped reads at both steps. Red dashed line indicates 12.5% level. Increase of unique mapping and decrease of multiple mapping is shown by arrows. Barplots at the right represent unique and multiple mapping by tRNA species

at Step2, in proportion to total and in absolute number of sequencing reads obtained by tRNA, expressed as proportion to total number of mapped reads. tRNAs showing excessive proportion of ambiguous mapping are shown in red.

Inspection of multiply mapped reads showed that only $tRNA^{Leu}$, $tRNA^{Lys}$ and $tRNA^{Arg}$ in the *D. melanogaster* dataset still showed substantial multiple mapping (shown in red, Figure 3). For the human reference, this was the case for pairs of $tRNA^{Leu}$, $tRNA^{Arg}$, one $tRNA^{Tyr}$ and $tRNA^{Ala}$/$tRNA^{Val}$ (Figure 3). These cases were manually inspected to find the origin of such ambiguous mapping. Both in *D. melanogaster* and in human, two $tRNA^{Leu}$ isoacceptors were found well represented in the RNA-seq data and showed very similar sequences at the 5'- and 3'-extremities but differed in the anticodon loop (Supplementary Figure S3 and information in the Supplementary data). Human $tRNA^{Tyr}_3$ reads were rather scarce compared to two other $tRNA^{Tyr}$ isoacceptors and thus $tRNA^{Tyr}_3$ was removed from the curated reference sequence. Finally, $tRNA^{Val}_3$ (encoded by the unique gene tRNA-Val-AAC-6-1) differs from $tRNA^{Ala}_1$ only at 10 positions and appeared to be an anticodon (missense suppressor) mutant of $tRNA^{Ala}_1$. Thus, this $tRNA^{Val}_3$ was also removed from the final reference sequence list.

Final manually curated reference sequences for six model organisms now include the 3'- CCA end, and all anticodon A34 residues were replaced by G since A34 is systematically converted to inosine (I) [359] which behaves as G in base pairing. Alignment results of full tRNA reads to curated references are shown in Table 2.

Table 2. Alignment of full tRNA reads dataset to manually curated tRNA references (Step3).

Organism	**tRNA gene number**	**% of aligned reads**	**Uniquely aligned reads**[1]	**Multiply aligned reads**[1]
Escherichia_coli_str_K-12_substr_MG1655	39	95.87+/-0.26	95.95+/-0.23	4.05+/-0.23
Bacillus_subtilis_subsp_subtilis_str_168	33	94.89+/-0.91	93.93+/-0.38	6.07+/-0.38
Saccharomyces_cerevisiae_S288c	38	84.93+/-6.61	98.52+/-0.27	1.48+/-0.27

Arabidopsis_thaliana_TAIR108feb2011	43	85.93+/-2.34	94.08+/-0.93	5.92+/-0.93
Drosophila_melanogaster_BDGP6_dm6	34	83.67+/-2.36	86.88+/-2.36	13.12+/-2.36
Homo_sapiens_GRCh37hg19	43	90.08+/-0.70	89.15+/-0.92	10.85+/-0.92

[1] Normalized to the total number of aligned reads.

4. Discussion

4.1 Merging of similar tRNA genes in a single reference sequence

Highly redundant RNA sequences issued from gene duplication and sequence mutations are difficult targets for extensive and unambiguous bioinformatic analysis of deep sequencing data, since short sequencing reads obtained in many popular protocols do not allow to distinguish between these very similar species. Transfer RNA (tRNA) populations in most organisms are well-studied examples. Most bacterial species have 60-80 tRNA encoding genes, and this number increases rapidly with genome complexity. Most complex (from a tRNA gene-centered point of view) eukaryotic species may contain > 5,000 duplicated and point-mutated tRNA genes. After removal of duplicates, over 100 (or even more) distinct sequences are still present. Analysis of such extremely complex mixtures requires unambiguous mapping, which is only possible with long sequencing reads. Only reads over 30-40 nt in length are truly unique enough to be precisely mapped to a given RNA sequence. However, such reads are relatively rare in the case of tRNA isoacceptors/isodecoders, which are of 75-90 nt in length only and heavily modified, preventing, in most cases, trouble-free read-through by RT. tRNA-demethylation protocols (e.g. DM-Seq, DRM-Seq, …) improve efficiency of primer extension, but do not remove all RT-blocking modified nucleotides. Thus, tRNA deep sequencing data are known to by heavily biased, with preferential amplification of only certain tRNA species [345]. Up to now, no general approach allowing rapid analysis of tRNA deep sequencing data was established in the community, and so far, every new organism required particular attention and almost manual inspection of numerous tRNA sequences.

In this work, we propose a generalized approach to reduce the complexity of tRNA reference sequences, intended to be used for reads mapping. The first, most evident, step consists in removal of all duplicated tRNA genes of exactly identical sequence. This step is efficient and allows reducing complexity very considerably. However, such

automatic removal of duplicates does not provide unambiguous mapping reference, since point-mutated variants persist. Such non-redundant reference is suitable for mapping of long (preferentially full size) tRNA reads, obtained without RNA fragmentation [347, 348]. Unfortunately, such non-fragmented tRNA libraries are highly biased due to certain tRNA modifications resulting in abortive (incomplete) cDNA fragments. Alternative approaches using tRNA fragmentation are more efficient, since fragments may be devoid of RT-arresting modifications; however shorter sequencing reads are obtained in such cases [350]. This is also a feature of the RiboMethSeq protocol, which is now extensively used for analysis of tRNA 2'-O-methylations [64, 65]. Other protocols, like AlkAnilineSeq for detection of $m^7G/m^3C/D/ho^5C$ [360] and HydraPsiSeq for pseudouridine mapping and quantification [361], also result in relatively short tRNA reads.

To obtain unambiguous reads mapping, which may ultimately be used for both tRNA quantification and modification analysis, we propose to group (merge) closely related species into a single entry, replacing "ambiguous" mutated nucleotides by most frequent nucleotide or by N. Even if the use of IUPAC nucleotide code seems to be more appropriate, such nucleotides are considered as N by many alignment utilities (Bowtie2, STAR, BWA) [362, 363]. The number of N nucleotides should be limited, to reduce ambiguous mapping, but not become too small, since otherwise closely related species will lead to multiple mapping events. Empirically, N=7-8 maximum (10% of the tRNA length) was found to be suitable for the majority of simple tRNA references (bacteria/archaea/lower eukaryotes), while 10-11 substitutions have to be allowed for more complex cases, like human and drosophila tRNAomes.

4.2 Representativeness of optimized tRNA datasets

In order to check if such reduced tRNA reference datasets remain representative, anticodon compositions of full and optimized datasets were compared (see Supplementary Materials). In general, when some apparent loss of rarely used tRNA anticodons was found, manual inspection revealed that those correspond to rare tRNA isoacceptors (encoded by genes with low copy number) and represent point mutant of major tRNA isoacceptor. These conversions are generally limited to tRNAs decoding the same aminoacid, with an exception of one human $tRNA^{Val}$(AAC), which is in fact derived from $tRNA^{Ala}$(NGC) by a limited number of mutations, including two in the anticodon. Interestingly, this $tRNA^{Val}$(AAC) is most likely behaving as a missense suppressor *in vivo* since it retains a G3*U70 base pair and a discriminator nucleotide A73, two well-known identity determinants for aminoacylation by AlaRS [364].

Analysis of several experimentally available tRNA datasets showed that only minor fractions (<10%) in these whole collections of tRNA reads were not aligned to the optimized reference sequence and thus the representativeness of the reduced reference was clearly maintained.

4.3 Known limitations and troubleshooting

Counting of tRNA reads and comparative studies of tRNA expression are possible with a reduced/optimized reference, however, if only one particular tRNA isoacceptor has to be quantified, the best approach would be to use a non-redundant tRNA dataset and to select only (rare) uniquely mapped reads for a given tRNA. Analysis of tRNA modification with a reduced dataset is based on the assumption that all species in the group have identical tRNA modification profiles, independently from the presence of a point mutation. This assumption is generally relevant, since mutations occur mainly in the tRNA regions, which are not known to be extensively modified, however, this may not be true for some essential tRNA anticodon positions (N34, for example). In such cases, inspection of nucleotide proportion (e.g; by IGV visualization or mismatch calculation from mpileup format) may be of interest to determine the proportion of really expressed tRNA variants. In extreme cases, preliminary evaluation of tRNA expression using non-redundant datasets may be required, followed by the analysis of only major expressed tRNA isoforms. However, this complex 2 steps methodology cannot be applied in an automatized routine analysis.

4.4. Applications in analysis of tRNA expression and modifications

Since currently used approaches for tRNA analysis by deep sequencing suffer from multiple and thus very ambiguous mapping of tRNA reads, we foresee multiple practical applications of the proposed general approach for tRNA reference sequences. Unified and straightforward analysis of tRNA deep sequencing data is still lacking in the field. Quantification of tRNA pool and modulations of its composition under stress or in disease represent one of the major applications. One other major field which will certainly benefit from these improvements is analysis of epitranscriptomic tRNA modifications (2'-O-methylation, pseudouridine, m^7G/m^3C, D and others), important for adaptation of cellular translational machinery in stress or disease. With some minor adaptations, similar computational and validation approaches can be also applied for simplification of other highly redundant collection of RNA genes, e.g. rRNA and snRNA sequences.

5. Data availability

All scripts, supplementary data as well as non-redundant and optimized datasets are available from GitHub (https://github.com/FlorianPichot/tRNA_reference_construction.)

Author Contributions: Conceptualization, FP. and Y.M.; software, F.P.; investigation, F.P and V.M.; writing—original draft preparation, Y.M.; writing—review and editing, V.M and M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by FRCR EpiARN project from Grand Est Region to YM and SPP1784, HE 3397/14-2 from DFG to MH, and JPND “RNA NEURO” / Bmbf 01ED1804 to MH. This work was performed in the framework of EPITRAN COST initiative (CA16120).

Acknowledgments: We thank M. Kristen for the critical reading of the manuscript.

Conflicts of Interest: “The authors declare no conflict of interest.”

Publisher’s Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



© 2020 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).

Supplementary material

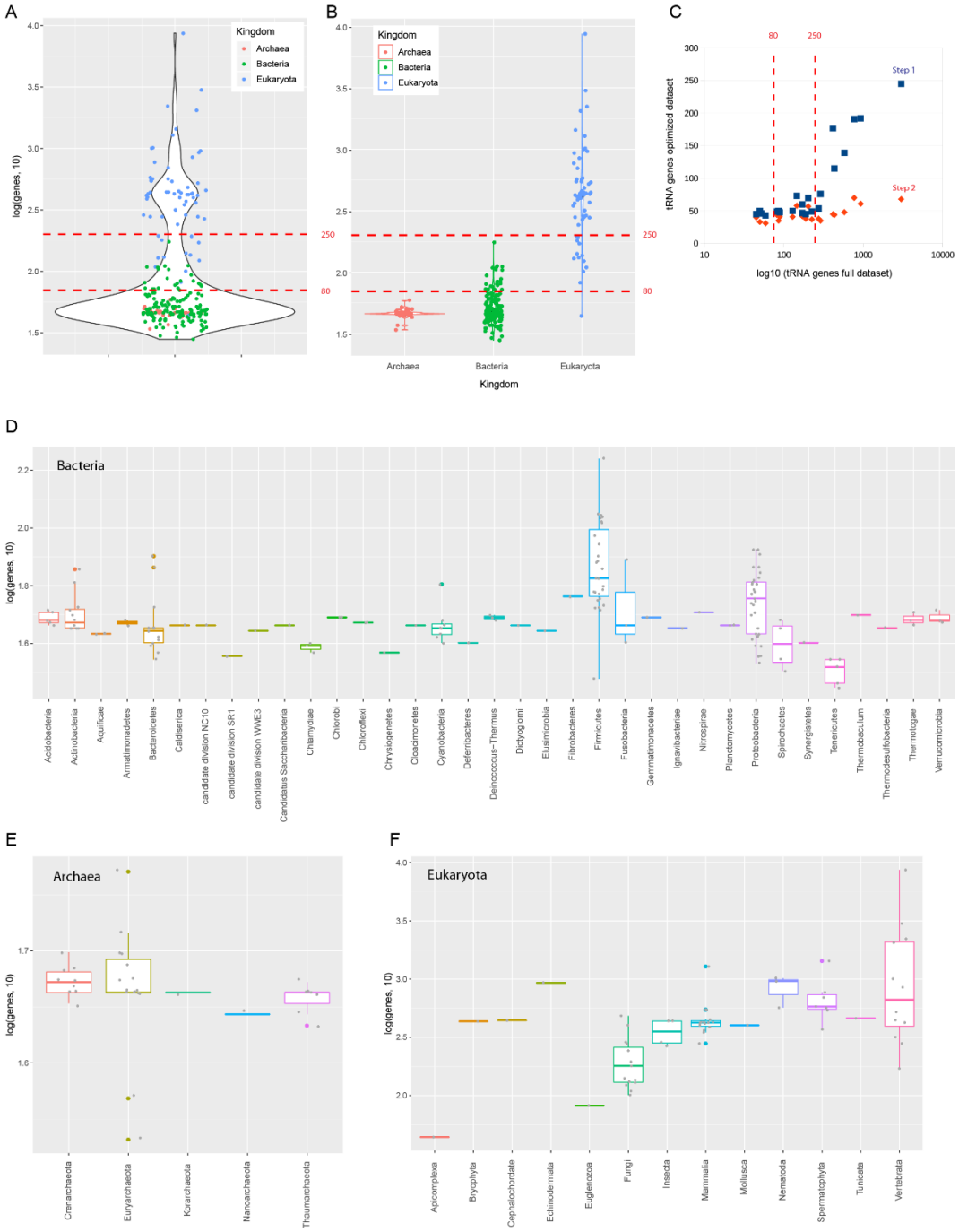


Figure S1 Number of genomic tRNA genes detected by tRNAScanSE (source gtRNAdb). Genomes (~250 in total) were randomly chosen to provide representative selection for Archaea (35 genomes), Bacteria (150 genomes) and Eukaryota (65 genomes). Only tRNA genes corresponding to 20 standard amino acids were considered. Panel A represents global distribution (number of tRNA genes in $\log_{10}$ scale), panel B – same data sorted by Kingdom. Panel C shows the number of tRNA genes in non-redundant (blue) and in optimized (Step2) references (red), in function of the total number of tRNA genes in genomic reference. Panel D, E and F show distribution by Kingdom and phyla/groups.

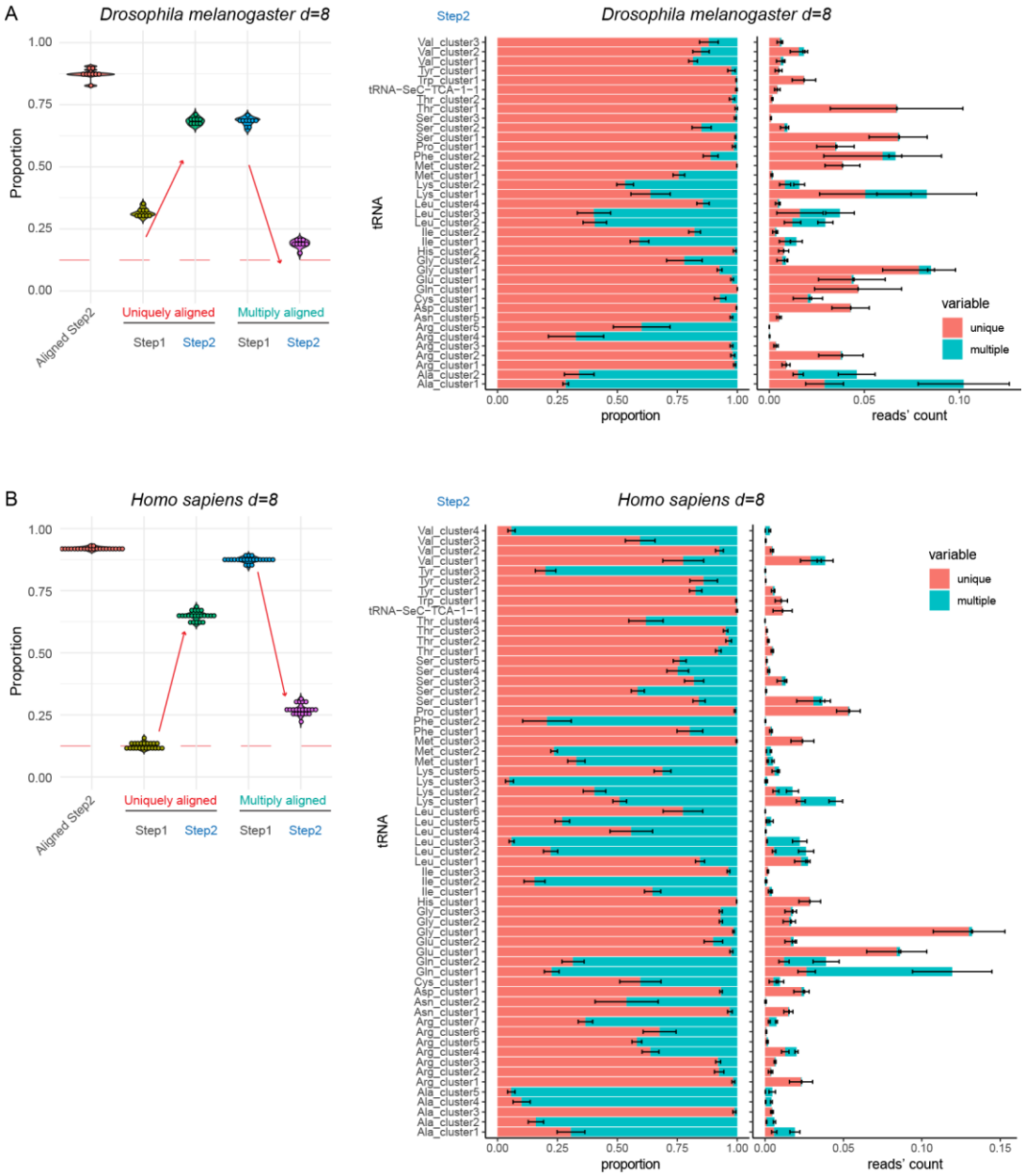


Figure S2 Alignment results for Non Duplicated (non redundant) tRNA reference (Step1) and optimized tRNA set (Step2) for *D. melanogaster* and *H. sapiens* references, maximal distance used is 8 substitutions. Boxplot on the left shows the proportion of tRNA sequencing reads aligned to Step2 reference ('Aligned Step2') and proportions of uniquely and multiply mapped reads at both steps. Red dashed line indicate 12.5% level. Increase of unique mapping and decrease of multiple mapping is shown by arrows. Barplots at the right represent unique and multiple mapping by tRNA species at Step2, in proportion to total and in absolute number of sequencing reads obtained by tRNA, expressed as proportion to total number of mapped reads. tRNAs showing excessive proportion of ambiguous mapping are shown in red.

```
D.melanogaster tRNALeu2 and tRNALeu3
Leu3 1>GTCAGGATGGCCGAGTGGTCTAAGGCGCTGCGTTCAGGTCGCAGTCTACTCTGTAGGCGTGGGTTCGAATCCCACTTCTGACA>83
Leu2 1>GTCAGGATGGCCGAGCGGTCTAAGGCGCCAGACTCAAGTTCTGGTCCTCTCTGAGGGCGTGGGTTCGAATCCCACTTCTGACA>83

H.sapiens tRNALeu2 and tRNALeu3
Leu3 1>GTCAGGATGGCCGAGCNGTCTNAAGGCGCTGCGTTCANNTCGCANNCTCC-NCTGGAGGCGTGGGTTCGAATCCCACTNNTGACA>84
Leu2 1>GTCAGGATGGCCGAGTGGTCT-AAGGCGCCAGACTCAAGTTCTGGTCTCCNNATGGAGGCGTGGGTTCGAATCCCACTTCTGACA>84

H.sapiens tRNATyr1, Tyr2 and and Tyr3
Tyr2 1>CCTTCAATAGTTCAGCTGGTAGAGCAGAGGACTATAGGTCCTTAGGTT-GCTGGTTCGATTCCAGCTTGAAGGA>73
Tyr3 1>TCTTCAATAGCTCAGCTGGTAGAGCGGAGGACTGTAGATTCTTAGG-T-GCTGGTTTGATTCCGACTTGGAGAG>72 minor

Tyr3 1>TCTTCAATAGCTCAGCTGGTAGAGCGGAGGACTGTAGATTCTTAGG-T-GCTGGTTTGATTCCGACTTGGAGAG>72 minor
Tyr1 1>CCTTCGATAGCTCAGNTGGTAGAGCGGAGGACTGTAGATCCTTAGG-TCGCTGGTTCGANTCCGGCTCGAAGGA>73

H.sapiens tRNAAla1 and tRNAVal3
Val3 1>GGGGGTGTAGCTCAGTGGTAGAGCGTATGCTTAACATTCATGAGGCTCTGGGTTCGATCCCCAGCACTTCCA>72
Ala1 1>GGGGGTGTAGCTCAGTGGTAGAGCGCNTGCTTNGCATGTATGAGGCCCCGGGTTCGATCCCCGGCACCTCCA>72
```

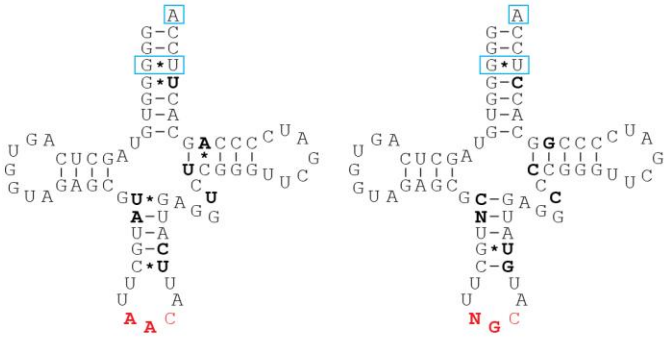


Figure S3 Sequences of tRNA species showing excessive ambiguous mapping for *D. melanogaster* and *H. sapiens* references. Non-identical nucleotides are in bold case, anticodon is in red and underlined.
Bottom – cloverleaf structures of *H. sapiens* $tRNA^{Val}_3$(AAC) and $tRNA^{Ala}_1$(NGC). Substitutions are in bold case. Anticodon is in red. Excessive number of mismatched nucleotides in $tRNA^{Val}_3$(AAC) stems may drive its instability (degradation) in vivo. Conserved identity elements for AlaRS aminoacylation (A73 and G3*U70) are boxed n bleu. 3'-CCA sequence not shown.

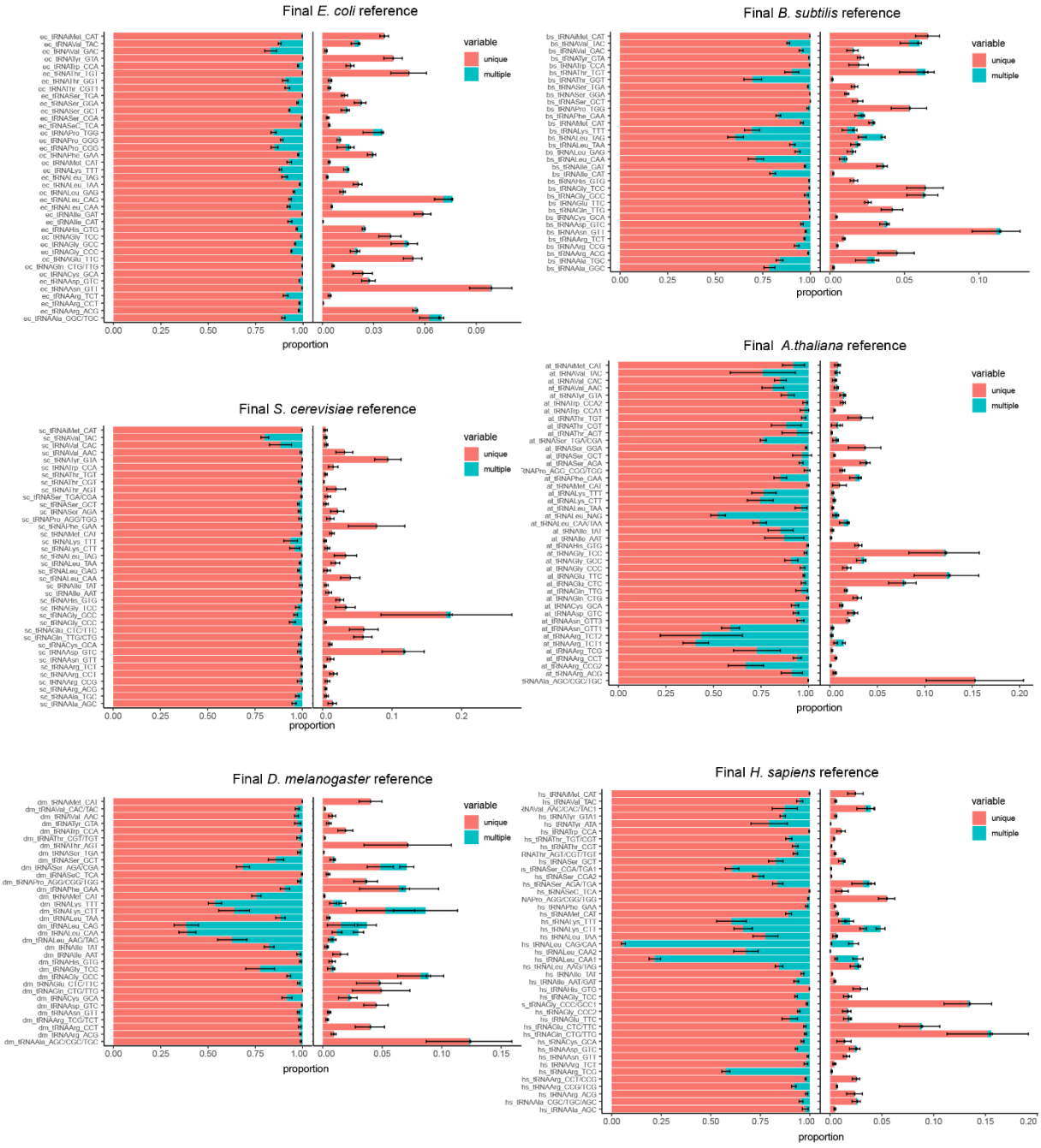


Figure S4 Barplots representing unique and multiple mapping by tRNA species for final manually curated tRNA references (Step 3), in proportion to total and in absolute number of sequencing reads obtained by tRNA, expressed as proportion to total number of mapped reads.

Detailed description of modifications in final tRNA references
Problematic cases

Escherichia coli

>Thr_cluster1 consensus sequence tRNA-Thr-CGT-1-1

GCCGATATAGCTCAGTTGGTAGAGCAGCGCATTCGTAATGCGAAGGTCGTAGGTTCGACTCCTATTATCGGCACCA

>Thr_cluster2 consensus sequence tRNA-Thr-CGT-2-1

GTAGTTAAAAATGCATTAACATCGCATTCGTAATGCGAAGGTCGTAGGTTCGACTCCTATTATCGGCACCA

>Thr_cluster3 consensus sequence tRNA-Thr-GGT-1-1|tRNA-Thr-GGT-2-1

GCTGATATNGCTCAGTTGGTAGAGCGCACCCTTGGTAAGGGTGAGGTCNNCAGTTCGANTCTGNNTATCAGCACCA

>Thr_cluster4 consensus sequence tRNA-Thr-TGT-1-1

GCCGACTTAGCTCAGTAGGTAGAGCAACTGACTTGTAATCAGTAGGTCACCAGTTCGATTCCGGTAGTCGGCACCA

Thr_cluster1 -- Matches:76; Mismatches:0; Gaps:0; Unattempted:0

Thr_cluster2 -- Matches:60; Mismatches:11; **Gaps:5**; Unattempted:0

Thr_cluster3 -- Matches:54; Mismatches:22; Gaps:0; Unattempted:0

Thr_cluster4 -- Matches:56; Mismatches:20; Gaps:0; Unattempted:0

* * * * * * *

Thr_cluster1 1>GCCGATATAGCTCAGTTGGTAGAGCAGCGCATTCGTAATGCGAAGGTCGTAGGTTCGACTCCTATTATCGGCACCA>76

Thr_cluster2 1>GTAGTTAAAAATGCATT-----AACATCGCATTCGTAATGCGAAGGTCGTAGGTTCGACTCCTATTATCGGCACCA>71

Thr_cluster3 1>GCTGATATNGCTCAGTTGGTAGAGCGCACCCTTGGTAAGGGTGAGGTCNNCAGTTCGANTCTGNNTATCAGCACCA>76

Thr_cluster4 1>GCCGACTTAGCTCAGTAGGTAGAGCAACTGACTTGTAATCAGTAGGTCACCAGTTCGATTCCGGTAGTCGGCACCA>76

Conclusion: Thr_cluster2 CGT=ec_tRNAThr_CGT2 is probable pseudogene, no U8, no G18/G19 in the D-loop (removed from the reference)

>Arg_cluster1 consensus sequence tRNA-Arg-ACG-1-1=tRNA-Arg-ACG-1-2=tRNA-Arg-ACG-1-3=tRNA-Arg-ACG-1-4

GCATCCGTAGCTCAGCTGGATAGAGTACTCGGCTACGAACCGAGCGGTCGGAGGTTCGAATCCTCCCGGATGCACCA

>Arg_cluster2 consensus sequence tRNA-Arg-CCG-1-1

GCGCCCGTAGCTCAGCTGGATAGAGCGCTGCCCTCCGGAGGCAGAGGTCTCAGGTTCGAATCCTGTCGGGCGCGCCA

>Arg_cluster3 consensus sequence tRNA-Arg-CCT-1-1

GTCCTCTTAGTTAAATGGATATAACGAGCCCCTCCTAAGGGCTAATTGCAGGTTCGATTCCTGCAGGGGACACCA

>Arg_cluster4 consensus sequence tRNA-Arg-TCT-1-1

GCGCCCTTAGCTCAGTTGGATAGAGCAACGACCTTCTAAGTCGTGGGCCGCAGGTTCGAATCCTGCAGGGCGCGCCA

Arg_cluster1 -- Matches:77; Mismatches:0; Gaps:0; Unattempted:0

Arg_cluster2 -- Matches:57; Mismatches:20; Gaps:0; Unattempted:0

Arg_cluster3 -- Matches:45; Mismatches:29; Gaps:4; Unattempted:0

Arg_cluster4 -- Matches:55; Mismatches:21; Gaps:2; Unattempted:0

```
* * * * * * *

>Arg_cluster1 1>GCATCCGTAGCTCAGCTGGATAGAGTACTCGGCTACGAA-CCGAGC-GGTCGGAGGTTCGAATCCTCCCGGATGCACCA>77

>Arg_cluster2 1>GCGCCCGTAGCTCAGCTGGATAGAGCGCTGCCCTCCGGA-GGCAGA-GGTCTCAGGTTCGAATCCTGTCGGGCGCGCCA>77

>Arg_cluster3 1>GTCCTCTTAG-TTAAATGGATATAACGAGCCCCTCCTAA---GGGCTAATTGCAGGTTCGATTCCTGCAGGGGACACCA>75

>Arg_cluster4 1>GCGCCCTTAGCTCAGTTGGATAGAGCAACGACCTTCTAAGTCGTG—GGCCGCAGGTTCGAATCCTGCAGGGCGCGCCA>77
```

Conclusion: Low-expressed Arg_cluster2=ec_tRNAArg_CCG shows identical regions with highly expressed Arg_cluster1 (thus Arg_cluster2 is removed from the reference)

```
Inosine-34 is known to be present in:
>ec_tRNAArg_ACG gcatccgtagctcagctggatagagtactcggctGcgaaccgagcggtcggaggttcgaatcctcccggatgcacca
Final E.coli reference 39 sequences : Final_Step3_Escherichia_coli_str_K-12_substr_MG1655__39seq_2020-11-27
```

Bacillus subtilis

```
>Arg_cluster1 consensus sequence tRNA-Arg-ACG-1-1=tRNA-Arg-ACG-1-2=tRNA-Arg-ACG-1-3=tRNA-Arg-ACG-1-4
GCGCCCGTAGCTCAATTGGATAGAGCGTTTGACTACGGATCAAAAGGTTAGGGGTTCGACTCCTCTCGGGCGCGCCA
>Arg_cluster2 consensus sequence tRNA-Arg-CCG-1-1
GCGCTCGTAGCTCAGTTGGATAGAGCGGTGGTTTCCGGTACCACGTCTGTCGGGGGTTCGAATCCCTCCGAGCGCG
>Arg_cluster3 consensus sequence tRNA-Arg-CCT-1-1
GCTCTAGTAGCACAGCGGATAGTGCAGCAGTTTCCTAAACTGCAGGTCGGGAGTTCGAATCTCTCCTAGAGCG
>Arg_cluster4 consensus sequence tRNA-Arg-TCT-1-1
GTCCCAGTAGCTCAGCTGGATAGAGCAACGGCCTTCTAAGCCGTCGGTCGGGAGTTCGAATCTCTCCTGGGACG

>Arg_cluster1 -- Matches:77; Mismatches:0; Gaps:0; Unattempted:0
>Arg_cluster2 -- Matches:55; Mismatches:19; Gaps:5; Unattempted:0
>Arg_cluster3 -- Matches:45; Mismatches:27; Gaps:6; Unattempted:0
>Arg_cluster4 -- Matches:47; Mismatches:27; Gaps:3; Unattempted:0

* * * * * * *
>Arg_cluster1 1>GCGCCCGTAGCTCAATTGGATAGAGCGTTTGACTACGG-ATCA-AAAGGTTAGGGGTTCGACTCCTCT-CGGGCGCGCCA>77
>Arg_cluster2 1>GCGCTCGTAGCTCAGTTGGATAGAGCGGTGGTTTCCGGTACCACGTCTGTCGGGGGTTCGAATCCCTC-CGAGCGCG~~~>76
>Arg_cluster3 1>GCTCTAGTAGCAC-AGCGGATAGTGCAGCAGTTTCCTA-AACT-GCAGGTCGGGAGTTCGAAT-CTCTCCTAGAGCG~~~>73
>Arg_cluster4 1>GTCCCAGTAGCTCAGCTGGATAGAGCAACGGCCTTCTA-AGCC-GTCGGTCGGGAGTTCGAAT-CTCTCCTGGGACG>74
```

Conclusion: Low-expressed Arg_cluster3=bs_tRNAArg_CCT shows identical regions with highly expressed Arg_cluster4 (thus Arg_cluster3 is removed from the reference)

```
>Leu_cluster3 consensus sequence tRNA-Leu-CAG-1-1
GCGGATGTGGCGGAATTGGCAGACGCGCTAGAATCAGGCTCTAGTGTCTTTACAGACGTGGGGGTTCAAGTCCCTTCATCCGCACCA
>Leu_cluster5 consensus sequence tRNA-Leu-TAG-1-1|tRNA-Leu-TAG-2-1
GCGGGTGTGGCGGAATTGGCAGACGCGCTAGACTTAGGATCTAGTGTCTTNANGACGTGGGGGTTCNAGTCCCTTCACCCGCA

>Leu_cluster3 -- Matches:75; Mismatches:8; Gaps:4; Unattempted:0
>Leu_cluster5 -- Matches:83; Mismatches:0; Gaps:0; Unattempted:0

* * * * * * * *
>Leu_cluster3 1>GCGGATGTGGCGGAATTGGCAGACGCGCTAGAATCAGGCTCTAGTGTCTTTACAGACGTGGGGGTTCAAGTCCCTTCATCCGCACCA
>Leu_cluster5 1>GCGGGTGTGGCGGAATTGGCAGACGCGCTAGACTTAGGATCTAGTGTCTTNA-NGACGTGGGGGTTCNAGTCCCTTCACCCGCA
```

Conclusion: Very high level of identity between two tRNAs Leu_cluster3=bs_tRNALeu_CAG is removed from the reference

```
Inosine-34 is known to be present in:
>bs_tRNAArg_ACG  gcgcccgtagctcaattggatagagcgtttgactGcggatcaaaaggttaggggttcgactcctctcgggcgcgcca
```

Saccharomyces cerevisiae

```
No manual adaptations required

Inosine-34 is known to be present in:
>sc_tRNAAla_AGC gggcgtgtggcgtagtcggtagcgcgctcccttGgcatgggagaggtctccggttcgattccggactcgtccacca
>sc_tRNAArg_ACG ttcctcgtggcccaatggtcacggcgtctggctGcgaaccagnagattccaggttcnagtcctggcggggaagcca
>sc_tRNAIle_AAT ggtctcttggcccagttggttaaggcaccgtgctGataacgcggggatcagcggttcgatcccgctagagaccacca
>sc_tRNASer_AGA ggcaacttggccgagtggttaaggcgaaagattGgaaatcttttgggctttgncccgcgcaggttcgagtcctgcagttgtcgcca
>sc_tRNAThr_AGT gcttctatggccaagttggtaaggcgccacactGgtaatgtggagatcatcggttcaaatccgattggaagcacca
>sc_tRNAVal_AAC ggtttcgtggtctagtcggttatggcatctgcttGacacgcagaacgtccccagttcgatcctgggcgaaatcncca
```

Arabidopsis thaliana

```
>Arg_cluster1 consensus sequence tRNA-Arg-ACG-1-1=tRNA-Arg-ACG-5-1=tRNA-Arg-ACG-5-2|tRNA-Arg-ACG-2-1=tRNA-
Arg-ACG-3-1=tRNA-Arg-ACG-3-2|tRNA-Arg-ACG-6-1=tRNA-Arg-ACG-6-2|tRNA-Arg-ACG-4-1
GACTCCATGGCCCAATGGATAAGGCGCTGGTCTACGAAACCAGAGATTCTGGGTTCGATCCCCAGTGGAGTCG
>Arg_cluster2 consensus sequence tRNA-Arg-TCG-1-1=tRNA-Arg-TCG-1-2=tRNA-Arg-TCG-1-3|tRNA-Arg-TCG-2-1=tRNA-
Arg-TCG-2-2=tRNA-Arg-TCG-2-3
```

GACCGCATAGCGCAGTGGATTAGCGCGTNTGACTTCGGATCANAAGGTCGTGGGTTCGACTCCCACTGTGGTCG
>Arg_cluster3 consensus sequence tRNA-Arg-TCT-1-1=tRNA-Arg-TCT-5-1=tRNA-Arg-TCT-5-2|tRNA-Arg-TCT-2-1=tRNA-Arg-TCT-7-1=tRNA-Arg-TCT-8-1|tRNA-Arg-TCT-3-1=tRNA-Arg-TCT-4-1
GCACCCGTGGCCTAATGGATAAGGCGTTTGACTTCTAATCAAACGATTGTGGGTTCGAGTCCCACCGGGTGTG
>Arg_cluster4 consensus sequence tRNA-Arg-CCG-3-1=tRNA-Arg-CCG-3-2|tRNA-Arg-CCG-2-1
GNNNGCGTGGCCTAATGGATAAGGCGCTCGCCTCCGNAGCGGGAGATTGTGGGTTCGANTCCCANCGNGNNCG
>Arg_cluster5 consensus sequence tRNA-Arg-CCT-1-1=tRNA-Arg-CCT-1-2=tRNA-Arg-CCT-1-3=tRNA-Arg-CCT-1-4=tRNA-Arg-CCT-1-5=tRNA-Arg-CCT-2-1|tRNA-Arg-CCT-3-1
GCGCCTGTAGCTCAGTGGATAGAGCGTCTGTTTCCTAAGCAGAANGTCGNAGGTTCGACCCCTNCCTGGCGCG
>Arg_cluster6 consensus sequence tRNA-Arg-TCT-6-1
GCGCTCGTGGCCCAATGGATAAGGCGTCTGACTTCTAATCAGACGATTGTGGGTTCGATCCCCACCGAGCGTG
>Arg_cluster7 consensus sequence tRNA-Arg-CCG-1-1
GATCCCATAGCGGAGTGGATATCGCGTTAGACTCCGAATCTAAAGGTCGTGGGTTCGATTCCCACTGGGATCA

>Arg_cluster1 1>GACTCCATGGCCCAATGGATAAG-GCGCTGGTCTACGAAACCAGAGATTCTGGGTTCGATCCC-CAGTG-GAGTCG>73
>Arg_cluster2 1>GACCGCATAGCGCAGTGGATTAGCGCGTNTGACTTCGGATCANAAGGTCGTGGGTTCGACTCC-CACTG-TGGTCG>74
>Arg_cluster3 1>GCACCCGTGGCCTAATGGATAAG-GCGTTTGACTTCTAATCAAACGATTGTGGGTTCGAGTCC-CACCG-GGTGTG>73
>Arg_cluster4 1>GNNNGCGTGGCCTAATGGATAAG-GCGCTCGCCTCCGNAGCGGGAGATTGTGGGTTCGANTCC-CANCG-NGNNCG>73
>Arg_cluster5 1>GCGCCTGTAGCTCAGTGGATAGA-GCGTCTGTTTCCTAAGCAGAANGTCGNAGGTTCGACCCCTNCCTG-GCG-CG>73
>Arg_cluster6 1>GCGCTCGTGGCCCAATGGATAAG-GCGTCTGACTTCTAATCAGACGATTGTGGGTTCGATCCC-CACCGAGCGT-G>73
>Arg_cluster7 1>GATCCCATAGCGGAGTGGATATC-GCGTTAGACTCCGAATCTAAAGGTCGTGGGTTCGATTCC-CACTG-GGATCA>73

Conclusion: Low expressed Arg_cluster4=at_tRNAArg_CCG1 is highly similar to Arg_cluster3 and thus can be removed

>Asn_cluster1 consensus sequence tRNA-Asn-GTT-3-1=tRNA-Asn-GTT-3-2=tRNA-Asn-GTT-3-3=tRNA-Asn-GTT-3-4=tRNA-Asn-GTT-3-5=tRNA-Asn-GTT-3-6|tRNA-Asn-GTT-1-1|tRNA-Asn-GTT-2-1|tRNA-Asn-GTT-8-1
GCTGGAATAGCTCAGTTGGTTAGAGCGTGTGGCTGTTAACCACAAGGTCGGAGGTTCGANCCCTCCTTCTAGCG
>Asn_cluster2 consensus sequence tRNA-Asn-GTT-4-1=tRNA-Asn-GTT-4-2=tRNA-Asn-GTT-4-3=tRNA-Asn-GTT-4-4|tRNA-Asn-GTT-5-1
GCTGGAGTAGCTCAGTTGGTTAGAGCGTGTGGCTGTTAACCACAAGGTCAGAGGTTCGACCCCTNTCTCTAGCG
>Asn_cluster3 consensus sequence tRNA-Asn-GTT-7-1
TCCTCAGTAGCTCAGTGGTAGAGCGGTCGGCTGTTAACTGATTGGTCGTAGGTTCAAATCCTACTTGGGGAG
>Asn_cluster4 consensus sequence tRNA-Asn-GTT-9-1
GCTGGAATAGCTCAGTAGGTTTAGAGCGTGTGGCTGTTAACCTCAAGGTCGGAGGTTCGACCCCTCCTTTCAGCG

1>GCTGGAATAGCTCAGTTGG-TTAGAGCGTGTGGCTGTTAACCACAAGGTCGGAGGTTCGANCCCTCCTTCTAGCG>74
1>GCTGGAGTAGCTCAGTTGG-TTAGAGCGTGTGGCTGTTAACCACAAGGTCAGAGGTTCGACCCCTNTCTCTAGCG>74
1>GCTGGAATAGCTCAGTAGGTTTAGAGCGTGTGGCTGTTAACCTCAAGGTCGGAGGTTCGACCCCTCCTTTCAGCG>75

Conclusion: Merge Asn_cluster1=at_tRNAAsn_GTT1 and Asn_cluster2=at_tRNAAsn_GTT2, Asn_cluster4=at_tRNAAsn_GTT4 removed

1>GCTGGAATAGCTCAGTTGG-TTAGAGCGTGTGGCTGTTAACCACAAGGTCGGAGGTTCGANCCCTCCTTCTAGCG>74
1>GCTGGAGTAGCTCAGTTGG-TTAGAGCGTGTGGCTGTTAACCACAAGGTCAGAGGTTCGACCCCTNTCTCTAGCG>74
Consensus 1>gctggantagctcagttgg-ttagagcgtgtggctgttaaccacaaggtcngaggttcgancccttnnntctagcg>74

>Leu_cluster1 consensus sequence tRNA-Leu-AAG-1-1=tRNA-Leu-AAG-1-10=tRNA-Leu-AAG-1-2=tRNA-Leu-AAG-1-3=tRNA-Leu-AAG-1-4=tRNA-Leu-AAG-1-5=tRNA-Leu-AAG-1-6=tRNA-Leu-AAG-1-7=tRNA-Leu-AAG-1-8=tRNA-Leu-AAG-1-9=tRNA-Leu-AAG-2-1|tRNA-Leu-CAG-1-1=tRNA-Leu-CAG-1-2=tRNA-Leu-CAG-1-3
GTNNANATGGCCGAGTTGGTCTAAGGCGCCAGNTTNAGGTNCTGGTCCGAAAGGGCGTGGGTTCAAATCCCACTNTNNACA
>Leu_cluster2 consensus sequence tRNA-Leu-TAG-1-1=tRNA-Leu-TAG-1-2=tRNA-Leu-TAG-1-3=tRNA-Leu-TAG-1-4|tRNA-Leu-TAG-2-1=tRNA-Leu-TAG-2-2=tRNA-Leu-TAG-2-3|tRNA-Leu-TAG-3-1=tRNA-Leu-TAG-3-2
GACAGTTTGGCCGAGTGGTCTAAGGCGCCAGATTTAGGCTCTGGTCCGAAAGGGCGTGGGTTCAAATCCCACAGCTGTCA
>Leu_cluster3 consensus sequence tRNA-Leu-CAA-1-1=tRNA-Leu-CAA-1-2=tRNA-Leu-CAA-1-3=tRNA-Leu-CAA-3-1|tRNA-Leu-CAA-2-1=tRNA-Leu-CAA-2-2=tRNA-Leu-CAA-2-3|tRNA-Leu-CAA-5-1=tRNA-Leu-CAA-5-2|tRNA-Leu-CAA-4-1|tRNA-Leu-TAA-1-1
GTCAGGATGGCCGAGTGGTCTAAGGCGCCAGACTCAAGTTCTGGTCCTCGTAAGAGGGCGTGGGTTCAAATCCCACTTCTGACA
>Leu_cluster4 consensus sequence tRNA-Leu-TAA-2-1=tRNA-Leu-TAA-2-2=tRNA-Leu-TAA-2-3=tRNA-Leu-TAA-3-1=tRNA-Leu-TAA-3-2
gcaggtttgcccgagtggttaaggggaagacttaagttcttctgcacataagtgcgcgtgggttcgaaccccacagcctgca
>Leu_cluster5 consensus sequence tRNA-Leu-AAG-4-1
gggcatttggtctagtggtttgattttcgcttaaggtgcgagaggtcccgagttcaattctcagaatgcccc

Leu_cluster1 GTNNANATGGCCGAGTTGGTCTAAGGCGCCAGNTTNAGGTNCTGGTCC----GAAAGGGCGTGGGTTCAAATCCCACTNTNNACA
Leu_cluster2 GACAGTTTGGCCGAGT-GGTCTAAGGCGCCAGATTTAGGCTCTGGTCC----GAAAGGGCGTGGGTTCAAATCCCACAGCTGTCA
Leu_cluster3 GTCAGGATGGCCGAGT-GGTCTAAGGCGCCAGACTCAAGTTCTGGTCCTCGTAAGAGGGCGTGGGTTCAAATCCCACTTCTGACA
Leu_cluster4 gcaggtttgcccgagtggttaaggggaagacttaagttcttctgcacataagtgcgcgtgggttcgaaccccacagcctgca
Leu_cluster5 gggcatttggtctagtggtttgattttcgcttaaggtgcgagaggtcccgagttcaattctcagaatgcccc

Conclusion: Sequence Leu_cluster5=at_tRNALeu_AAG2 too short, unique gene tRNA-Leu-AAG-4-1 (pseudogene?), to remove

Leu_cluster1 GTNNANATGGCCGAGTTGGTCTAAGGCGCCAGNTTNAGGTNCTGGTCC----GAAAGGGCGTGGGTTCAAATCCCACTNTNNACA
Leu_cluster2 GACAGTTTGGCCGAGT-GGTCTAAGGCGCCAGATTTAGGCTCTGGTCC----GAAAGGGCGTGGGTTCAAATCCCACAGCTGTCA
Leu_cluster3 GTCAGGATGGCCGAGT-GGTCTAAGGCGCCAGACTCAAGTTCTGGTCCTCGTAAGAGGGCGTGGGTTCAAATCCCACTTCTGACA

Conclusion: Three tRNAs are very similar, so keep Leu_cluster3 (seems to be major) and collaps Leu_cluster 1 and 2 in one
Leu_cluster1 GTNNANATGGCCGAGTTGGTCTAAGGCGCCAGNTTNAGGTNCTGGTCC----GAAAGGGCGTGGGTTCAAATCCCACTNTNNACA
Leu_cluster2 GACAGTTTGGCCGAGT-GGTCTAAGGCGCCAGATTTAGGCTCTGGTCC----GAAAGGGCGTGGGTTCAAATCCCACAGCTGTCA
consensus gnnnnnntggccgagttggtctaaggcgccagnttnaggnnctggtcc----gaaagggcgtgggttcaaatcccacnnnnnca

Inosine-34 is most likely present in (no data in tRNAdb or MODOMICS):

>at_tRNAAla_AGC_CGC_TGC ggggatgtagctcatatggtagagcgctcgcttNgcatgcgagaggcacggggttcgattccccgcatctccacca
>at_tRNAArg_ACG gactccatggcccaatggataaggcgctggtctGcgaaaccagagattctgggttcgatccccagtggagtcgcca

>at_tRNAIle_AAT ggccnattagctcagttggttagagcgtcgtgctGataacgcgaaggtcncaggttcgannncctgnatnggccacca
>at_tRNAPro_AGG_CGG_TGG gggcatttggtctagtggtatgattctcgcttNgggtgcgagaggtcccgagttcgattctcggaatgcccccca
>at_tRNASer_AGA gtggacgtgccggagtggtttatcgggcatgactGgaaatcatgtgggctttgcccgcgcaggttcgaatcctgccgttnacgcca
>at_tRNAThr_AGT gctntcntagctcagttggttagagcacccgtttGgtaagcgggaggtcttgagttcaactctcaanganagcacca
>at_tRNAVal_AAC ggtttcgtggtgtagttggtttatcacgtcagtctGacacactnaaggtctccggttcgaacccgggcgaagccacca
N are NOT replaced

Drosophila melanogaster

>Lys_cluster2 consensus sequence tRNA-Lys-TTT-2-1=tRNA-Lys-TTT-2-2=tRNA-Lys-TTT-2-3=tRNA-Lys-TTT-2-4=tRNA-Lys-TTT-2-5|tRNA-Lys-TTT-1-1
GCCCGGNTAGCTCAGTCGGTAGAGCATTGGACTTTTAATCCAAGGGTCCAGGGTTCAAGTCCCTGNTCGGGCGcca
>Lys_cluster1 consensus sequence tRNA-Lys-CTT-1-1=tRNA-Lys-CTT-1-10=tRNA-Lys-CTT-1-11=tRNA-Lys-CTT-1-12=tRNA-Lys-CTT-1-13=tRNA-Lys-CTT-1-2=tRNA-Lys-CTT-1-3=tRNA-Lys-CTT-1-4=tRNA-Lys-CTT-1-5=tRNA-Lys-CTT-1-6=tRNA-Lys-CTT-1-7=tRNA-Lys-CTT-1-8=tRNA-Lys-CTT-1-9
gcccggctagctcagtcggtagagcatgagactcttaatctcagggtcgtgggttcgagccccacgttgggcgCCA

>Lys_cluster1 1>gcccggctagctcagtcggtagagcatgagactcttaatctcagggtcgtgggttcgagccccacgttgggcgCCA>76
>Lys_cluster2 1>GCCCGGNTAGCTCAGTCGGTAGAGCATTGGACTTTTAATCCAAGGGTCCAGGGTTCAAGTCCCTGNTCGGGCGcca>76
Conclusion: 14 different nucleotides, but 3’-end is common, keep both sequences, Lys_TTT seems to be minor tRNA

>Leu_cluster2 consensus sequence tRNA-Leu-CAA-1-1=tRNA-Leu-CAA-2-1=tRNA-Leu-CAA-2-2=tRNA-Leu-CAA-2-3
gtcaggatggccgagcggtctaaggcgccagactcaagttctggtcctctctgagggcgtgggttcgaatcccacttctgacaCCA
>Leu_cluster3 consensus sequence tRNA-Leu-CAG-1-1=tRNA-Leu-CAG-1-2=tRNA-Leu-CAG-1-3=tRNA-Leu-CAG-1-4=tRNA-Leu-CAG-1-5=tRNA-Leu-CAG-1-6=tRNA-Leu-CAG-1-7=tRNA-Leu-CAG-1-8
gtcaggatggccgagtggtctaaggcgctgcgttcaggtcgcagtctactctgtaggcgtgggttcgaatcccacttctgacaCCA

>Leu_cluster2 1>gtcaggatggccgagcggtctaaggcgccagactcaagttctggtcctctctgaggcgtgggttcgaatcccacttctgacaCCA>86
>Leu_cluster3 1>gtcaggatggccgagtggtctaaggcgctgcgttcaggtcgcagtctactctgtaggcgtgggttcgaatcccacttctgacaCCA>86

Conclusion: Two tRNAs are too similar at the 5’- and 3’-ends, keep both

>Arg_cluster4 consensus sequence tRNA-Arg-TCT-1-1|tRNA-Arg-TCT-2-1
GNCCCTTTNGCGCANNGGATAGCGCGTTGGACTTCTAATCCAAAGGTNGCGGGTTCGATNCCCGCAAGGGNTGcca
Conclusion: Arg_cluster4=dm_tRNAArg_TCT is low expressed, contains too many N, may be similar with Arg_cluster3, thus removed

Inosine-34 is known to be present in:
>dm_tRNASer_AGA_CGA gcagtcgtggccgagcggttaaggcgtctgactGgaaatcagattccctctgggagcgtaggttcgaatcctaccggctgcgcca
>dm_tRNAVal_AAC gtttccgtggtgtagnggttatcacatcngcctGacacgcngaaggcccccggttcnatcccgggcggaaacacca

And is most likely present in (no data in tRNAdb or MODOMICS):

>dm_tRNAAla_AGC_CGC_TGC ggggatgtagctcagtggtagagcgctcgcttNgcatgtgagaggtcccgggttcgatccccggcatctccacca
>dm_tRNAArg_ACG ggtcctgtggcgcaatggataacgcgtctgactGcggatcagaagattccaggttcgactcctggcaggatcgcca
>dm_tRNAIle_AAT ggcccattagctcagttggttagagcgtcgtgctGataacgcgaaggtcgcgggttcgatcccctcatgggccacca
>dm_tRNALeu_AAG_TAG ggnagcgtggccgagcggtctaaggcgctggtttNaggcaccagtcnncnnnnnggcgtgggttcgaatcccaccgctgncacca
>dm_tRNAPro_AGG_CGG_TGG ggctcgttggtctaggggtatgattctcgcttcgggtgcgagaggtcccgggttcaaatcccggacgagcccccca
>dm_tRNAThr_AGT ggcgccgtggcttagttggttaaagcgcctgtctGgtaaacaggagatcgtgagttcgaatntcgccggggcctcca

Homo sapiens

>Leu_cluster2 consensus sequence tRNA-Leu-CAA-1-1=tRNA-Leu-CAA-1-2|tRNA-Leu-CAA-2-1=tRNA-Leu-CAA-3-1=tRNA-Leu-CAA-4-1
GTCAGGATGGCCGAGTGGTCTAAGGCGCCAGACTCAAGTTCTGGTCTCCNNATGGAGGCGTGGGTTCGAATCCCACTTCTGACAcca
>Leu_cluster3 consensus sequence tRNA-Leu-CAG-1-1=tRNA-Leu-CAG-1-2=tRNA-Leu-CAG-1-3=tRNA-Leu-CAG-1-4=tRNA-Leu-CAG-1-5=tRNA-Leu-CAG-1-6=tRNA-Leu-CAG-1-7=tRNA-Leu-CAG-2-1=tRNA-Leu-CAG-2-2|tRNA-Leu-CAA-6-1
GTCAGGATGGCCGAGCNGTCTNAAGGCGCTGCGTTCANNTCGCANNCTCCNCTGGAGGCGTGGGTTCGAATCCCACTNNTGACAcca

>Leu_cluster2 1>GTCAGGATGGCCGAGTGGTCT-AAGGCGCCAGACTCAAGTTCTGGTCTCCNNATGGAGGCGTGGGTTCGAATCCCACTTCTGACA>
>Leu_cluster3 1>GTCAGGATGGCCGAGCNGTCTNAAGGCGCTGCGTTCANNTCGCANNCTCC-NCTGGAGGCGTGGGTTCGAATCCCACTNNTGACA>84
Conclusion: Too similar at the 5’- and 3’-ends, keep both

>Tyr_cluster1 consensus sequence tRNA-Tyr-GTA-1-1=tRNA-Tyr-GTA-2-1|tRNA-Tyr-GTA-3-1=tRNA-Tyr-GTA-6-1|tRNA-Tyr-GTA-4-1=tRNA-Tyr-GTA-5-1=tRNA-Tyr-GTA-5-2=tRNA-Tyr-GTA-5-3=tRNA-Tyr-GTA-5-4=tRNA-Tyr-GTA-5-5|tRNA-Tyr-GTA-7-1|tRNA-Tyr-GTA-8-1
CCTTCGATAGCTCAGNTGGTAGAGCGGAGGACTGTAGATCCTTAGGTCGCTGGTTCGANTCCGGCTCGAAGGAcca
>Tyr_cluster2 consensus sequence tRNA-Tyr-ATA-1-1
ccttcaatagttcagctggtagagcagaggactataggtccttaggttgctggttcgattccagcttgaaggaCCA
>Tyr_cluster3 consensus sequence tRNA-Tyr-GTA-9-1
tcttcaatagctcagctggtagagcggaggactgtagattcttaggtgctggtttgattccgacttggagagCCA

>Tyr_cluster2 1>CCTTCAATAGTTCAGCTGGTAGAGCAGAGGACTATAGGTCCTTAGGTT-GCTGGTTCGATTCCAGCTTGAAGGA>73
>Tyr_cluster3 1>TCTTCAATAGCTCAGCTGGTAGAGCGGAGGACTGTAGATTCTTAGG-T-GCTGGTTTGATTCCGACTTGGAGAG>72 minor

>Tyr_cluster3 1>TCTTCAATAGCTCAGCTGGTAGAGCGGAGGACTGTAGATTCTTAGG-T-GCTGGTTTGATTCCGACTTGGAGAG>72 minor
>Tyr_cluster1 1>CCTTCGATAGCTCAGNTGGTAGAGCGGAGGACTGTAGATCCTTAGG-TCGCTGGTTCGANTCCGGCTCGAAGGA>73
Conclusion: Low expressed Tyr_cluster3 is removed

>Ala_cluster1 consensus sequence tRNA-Ala-CGC-1-1=tRNA-Ala-CGC-2-1=tRNA-Ala-TGC-2-1=tRNA-Ala-TGC-3-1=tRNA-Ala-TGC-3-2|tRNA-Ala-TGC-5-1=tRNA-Ala-TGC-7-1|tRNA-Ala-AGC-2-1=tRNA-Ala-AGC-2-2=tRNA-Ala-AGC-3-1=tRNA-Ala-AGC-7-1|tRNA-Ala-AGC-1-1|tRNA-Ala-CGC-3-1|tRNA-Ala-CGC-4-1|tRNA-Ala-TGC-1-1|tRNA-Ala-TGC-4-1|tRNA-Ala-TGC-6-1|tRNA-Ala-AGC-4-1|tRNA-Ala-AGC-5-1|tRNA-Ala-AGC-6-1
GGGGGTGTAGCTCAGTGGTAGAGCGCNTGCTTNGCATGTATGAGGCCCCGGGTTCGATCCCCGGCACCTCCAcca

>Val_cluster3 consensus sequence tRNA-Val-AAC-6-1
gggggtgtagctcagtggtagagcgtatgcttaacattcatgaggctctgggttcgatccccagcacttccaCCA

>Ala_cluster1 1>GGGGGTGTAGCTCAGTGGTAGAGCGCNTGCTTNGCATGTATGAGGCCCCGGGTTCGATCCCCGGCACCTCCA>72
>Val_cluster3 1>GGGGGTGTAGCTCAGTGGTAGAGCGTATGCTTAACATTCATGAGGCTCTGGGTTCGATCCCCAGCACTTCCA>72
Conclusion: Val_cluster3 is removed

Inosine-34 is known to be present in:

>hs_tRNAAla_CGC_TGC_AGC gggggtgtagctcagtggtagagcgcntgcttNgcatgtatgaggccccgggttcgatccccggcacctccacca
>hs_tRNAAla_AGC ggggaattagctcaagtggtagagcgctcgcttGgcatgcgagaggtagcgggatcgatgcccgcattctccacca

And is most likely present in (no data in tRNAdb or MODOMICS):

>hs_tRNAArg_ACG gggccagtggcgcaatggataacgcgtctgactGcggatcagaagattctaggttcgactcctggctggctcgcca
>hs_tRNAIle_AAT_GAT ggccggttagctcagttggttagagcgtggtgctGataacgccaaggtcgcgggttcgatccccgtacnggccacca
>hs_tRNALeu_AAG_TAG ggtagcgtggccgagcggtctaaggcgctggattGaggctccagtctcttcgggggcgtgggttcgaatcccaccgctgccacca
>hs_tRNAPro_AGG_CGG_TGG ggctcgttggtctaggggtatgattctcgcttNgggtgcgagaggtcccgggttcaaatcccggacgagccccca
>hs_tRNASer_AGA_TGA gtagtcgtggccgagtggttaaggcgatggactGgaaatccattggggtntccccgcgcaggttcgaatcctgccgactacgcca
>hs_tRNAThr_AGT_CGT_TGT ggctcngtggcttagttggttaaagcgcctgtctGgtaaacaggagatcctgggttcgaatcccagcggggcctcca
>hs_tRNATyr_ATA ccttcaatagttcagctggtagagcagaggactGtaggtccttaggttgctggttcgattccagcttgaaggacca
>hs_tRNAVal_AAC_CAC_TAC1 gtttccgtagtgtagtggttatcacgttcgcctCacacgcgaaaggtccccggttcgaaaccgggcggaaacacca

-- End of article --

4.1.4 2'-O-methylation within prokaryotic and eukaryotic tRNA inhibits innate immune activation by endosomal Toll-like receptors but does not affect recognition of whole organisms. [DOI: 10.1261/rna.070243.118]

Isabel FREUND[1], Daniel BUHL[1], Sébastien BOUTIN[1], Annika KOTTER[2], Florian PICHOT[2,3], Virginie MARCHAND[4], Tim VIERBUCHEN[5], Holger HEINE[5], Yuri MOTORIN[3], Mark HELM[2], Alexander H. DALPKE[1,6], Tatjana EIGENBROD[1]

[1] Department of Infectious Diseases, Medical Microbiology and Hygiene, Heidelberg University Hospital, 69120 Heidelberg, Germany

[2] Institute of Pharmacy and Biochemistry, Johannes Gutenberg-University Mainz, 55128 Mainz, Germany

[3] IMoPA UMR7365 CNRS-Lorraine University, BioPole Lorraine University, 9 avenue de la Forêt de Haye, Vandoeuvre-les-Nancy, France

[4] UMS2008 IBSLor, CNRS-Lorraine University-INSERM, BioPole Lorraine University, 9 avenue de la Forêt de Haye, Vandoeuvre-les-Nancy, France

[5] Division of Innate Immunity, Research Center Borstel, Borstel, Germany

[6] Institute of Medical Microbiology and Hygiene, Technical University Dresden, 01307 Dresden, Germany

4.1.4 2'-O-methylation within prokaryotic and eukaryotic tRNA inhibits innate immune activation by endosomal Toll-like receptors but does not affect recognition of whole organisms. [DOI: 10.1261/rna.070243.118]

Overview

This paper is the direct continuation of the previous work done on RiboMethSeq analysis of tRNA data (Section 4.1.1) after the optimisation of the RiboMethSeq pipeline for trimming, alignment and scores settings (Section 4.1.2) and creation of adapted tRNA references for epitranscriptomics studies (Section 4.1.3). Our goal was to further investigate the role of innate immunity impact of tRNA 2'-O-methylation Gm18 in prokaryotic and eukaryotic organisms: *E. coli*, *S. cerevisae* and *H. sapiens*.

LC-MS/MS experiments on tRNA fractions and IFN-α quantification after tRNA transfection on human cells show a negative correlation between Gm content and IFN-α secretion, suggesting Gm has an immune inhibitory role in human cells.

To confirm this, depletion of Gm18, the most prevalent Gm modification in tRNAs, has been conducted in *E. coli*, *S. cerevisiae* and *H. sapiens* by knocking out their respective genes responsible for Gm18 2'-O-methylation catalyzation. Then, RiboMethSeq experiments have been carried out to detect presence or absence of 2'-O-methylation of Gm18 sites within tRNAs. Data comparison between wild type cell lines and mutant ones allowed validation of Gm18 sites for the three species. Those validations have been deemed possible thanks to the optimisation of the RiboMethSeq pipeline and creation of adapted tRNA references, respectively reducing False Positive hits to ensure reliable Gm18 detection and quantification and increasing the amount of tRNA sequences available for analysis.

Once the Gm18 depletion has been confirmed and mapped by RiboMethSeq in mutant strains, further immunostimulatory experiments have validated and given insights about the role of Gm18 as an immune response inhibitor for prokaryotic and eukaryotic species, displaying Gm18 as a potential therapeutic target for autoimmune diseases.

On a personal scale, I have carried on the RiboMethSeq analysis using the optimised RiboMethSeq pipeline and non-redundant tRNA reference I have created to validate the different Gm18 sites for this study

.

Abstract

Bacterial RNA has emerged as important activator of innate immune responses by stimulating Toll-like receptors TLR7 and TLR8 in humans. Guanosine 2'-O-methylation at position 18 (Gm18) in bacterial tRNA was shown to antagonize tRNA induced TLR7/8 activation, suggesting a potential role of Gm18 as immune escape mechanism. This modification also occurs in eukaryotic tRNA, yet a physiological immune function remained to be tested. We therefore set out to investigate the immune modulatory role of Gm18 in both prokaryotic and eukaryotic microorganisms, *Escherichia coli* and *Saccharomyces cerevisiae,* and in human cells. Using RiboMethSeq analysis we show that mutation of *trmH* in *E. coli*, *trm3* in *S. cereviase* and CRISPR/Cas9 induced knockout of TARBP1 in *H. sapiens* results in loss of Gm18 within tRNA. Lack of Gm18 across the kingdoms resulted in increased immunostimulation of peripheral blood mononuclear cells when activated by tRNA preparations. In *E. coli*, lack of 2'-O-methyltransferase trmH also enhanced immune stimulatory properties by whole cellular RNA. In contrast, lack of Gm18 in yeasts and human cells did not affect immunostimulation by whole RNA preparations. When using live *E. coli* bacteria, lack of *trmH* did not affect overall immune stimulation although we detected a defined TLR8/RNA-dependent gene expression signature upon *E. coli* infection. Together, these results demonstrate that Gm18 is a global immune inhibitory tRNA modification across the kingdoms, contributes to tRNA recognition by innate immune cells but as an individual modification has insufficient potency to modulate recognition of the investigated microorganisms.

Introduction

To ensure a rapid immune response against a broad range of invading pathogens, cells of the host's innate immune system are equipped with a limited set of pattern recognition receptors (PRRs) that sense highly conserved microbial structures, so-called pathogen-associated molecular patterns (PAMPs). Bacterial RNA has been identified as a potent activator of innate immune responses and was shown to be sensed by Toll-like receptors (TLRs) localized in the endolysosome[365–367]. Within human peripheral blood mononuclear cells (PBMCs), monocytes recognize microbial RNA in a TLR8 dependent manner, triggering the release of pro-inflammatory mediators including TNF, IL-12p40 and IFNβ. By contrast, plasmacytoid dendritic cells (DCs) sense bacterial RNA via endosomal TLR7, culminating in the production of IFN-a[368–372].

As both, bacterial and host RNA, are composed of the same building blocks, a reliable discrimination of self and non-self by the innate immune system is critical to ensure induction of an adequate anti-microbial defense in case of pathogen encounter while at the same time avoiding aberrant immune stimulation by host ('self') RNA. Indeed, several auto-immune disorders including systemic lupus erythematosus have been associated with abnormal recognition of self nucleic acids[373, 374]. Apart from compartmentalization - i.e. localization of nucleic acid sensing TLRs in the endolysosome to which self-RNA has limited access - and differences in the RNA

sequence composition, another determinant for self/non-self discrimination has been attributed to RNA modifications profiles. As such, both the extent and type of post-transcriptional RNA modifications differ between self and microbial RNA. In this regard, various studies have highlighted the importance of ribose 2'-O-methylation as a potent modification impeding immune stimulation by endosomal TLRs[235, 237, 304, 375, 376]. 2'-O-methylation not only impairs TLR7 and TLR8 activation but also acts as a TLR7/8 antagonist, thus inhibiting immune cell stimulation by otherwise immunogenic RNA species. Subsequent studies identified a DmDN [D=all but C; N=all] and a DmRC [D=all but C; R = G, A] motif as most efficient in antagonizing TLR7 and TLR8, respectively[377]. Of note, 2'-O-methylation of cytosine consistently failed to impede TLR7/8 activation in previous studies[237, 238, 378].

Although the highest abundance of 2'-O-methylation has been described in eukaryotic rRNA, this modification is also found in certain bacterial tRNAs[30]. Specifically, tRNA 2'-O-methylation of guanosine at position 18 (Gm18) which is present in five tRNA species in *Escherichia coli* (*E. coli*) has been demonstrated to be necessary and sufficient to suppress immune stimulation via TLR7 by whole *E. coli* tRNA preparations. Total tRNA preparations derived from an *E. coli* mutant lacking Gm18 2'-O-methyltransferase activity (*trmH* knock-out) elicited a stronger TLR7 response than the respective wildtype samples[235, 304]. These observations raised the hypothesis that Gm18, apart from its structural function, may also serve as an immune escape mechanism for *E. coli* or other microorganisms carrying this specific modification. Yet, this assumption is, thus far, exclusively based on experiments involving the transfection of purified tRNA preparations into immune cells. How *E. coli* tRNA 2'-O-methylation indeed shapes bacterial immunogenicity under physiological conditions, i.e. during infection with the whole, living microorganism, remains elusive. Moreover, Gm18 is present in the tRNA of lower and higher eukaryotes, including *Saccharomyces cerevisiae* (*S. cerevisiae*) as well as human cells. Its physiological significance for immune recognition via endosomal TLRs across the different kingdoms has not been studied so far. Given that eukaryotic RNA generally shows a much higher degree of post-transcriptional RNA modifications than prokaryotic RNA[287, 379] Gm18 might play a more redundant role in eukaryotes as compared to bacteria but experimental evidence is lacking.

The aim of the present study was therefore to determine the role of Gm18 within tRNA in pro- and eukaryotic organisms, specifically *E. coli*, *S. cerevisiae* and human cells, for modulating innate immune recognition. It should be tested whether Gm18 exerts an immune modulatory function across kingdoms and how this modification affects immune stimulation in general. We show that Gm18 deficiency augmented PBMC activation by purified total tRNA preparations from all investigated organisms. We also show for the first time an impact of Gm18 on the immune stimulatory properties of human tRNA. However, an impact of Gm18 in tRNA on immune activation by whole cellular RNA was only visible for *E. coli* but not for eukaryotes, presumably due to dominant differences in 2'-O-methylation in ribosomal RNA. In neither microbe did a knockout of the Gm18 methyltransferase impact the overall immune reaction, although

we identified an RNA-TLR8 regulated gene signature in response to *E. coli* in general. Thus, Gm18 is a global immune modulatory tRNA modification but it does not seem to play a role as immune escape mechanism during infection with the investigated microorganisms. The results suggest the existence of a threshold for Gm content in an invading pathogen's cellular RNA that governs TLR7/8 mediated immune responses.

Results

Yeast tRNA fractions show inverse correlation of Gm content with type I interferon secretion by human peripheral blood mononuclear cells

First we intended to test whether 2'-O-methylation might affect the immune stimulatory potency of *S. cerevisiae* tRNA as had been described before for *E. coli*. As a starting point we investigated fractions of tRNA isoacceptors that had been obtained in counter current distribution experiments[380], a technique which separates tRNA according to their relative lipophilicity. Fractions were analysed for their relative Gm, Am, Um and Cm content by LC-MS/MS and compared to their IFN-a inducing properties upon stimulation of human PBMCs. Due to highly cell-type specific receptor and cytokine expression profiles, IFN-a secretion within PBMCs is commonly used as a read-out for TLR7-dependent plasmacytoid DC activation. Gm-rich tRNAs were predominantly eluted in fractions 1-6 and 13-15, while relative Am levels showed a peak in fractions 13-15 (but were absent in the first ten fractions). Um was found prominently in fractions 3-6 and Cm in fractions 1-8 and 18-21 (Fig. 1 A). Introduction of Gm, Am and Um into synthetic RNA previously had been described to affect immune recognition whereas Cm consistently failed to impede TLR7/8 activation[237, 377, 378]. When the individual tRNA fractions were tested for induction of type I IFN from PBMCs, they varied strongly in their IFN-a inducing capacity (Fig. 1 B), suggesting a negative correlation between 2'-O-methylation and IFN induction also in eukaryotic RNA. In a first approximation a correlation matrix showed a significant negative correlation between Gm content and IFN secretion (Fig. 1 C), whereas Am, Cm and Um content showed no relation. A closer examination confirmed a general negative impact of Gm on IFN secretion (Fig. 1D). We do not exclude that other modifications might contribute to immune modulation, e.g. Am and Um may affect the lowered IFN-a levels observed in fractions 13-15 and fractions 3-6, respectively.

4.1.4 2'-O-methylation within prokaryotic and eukaryotic tRNA inhibits innate immune activation by endosomal Toll-like receptors but does not affect recognition of whole organisms. [DOI: 10.1261/rna.070243.118]

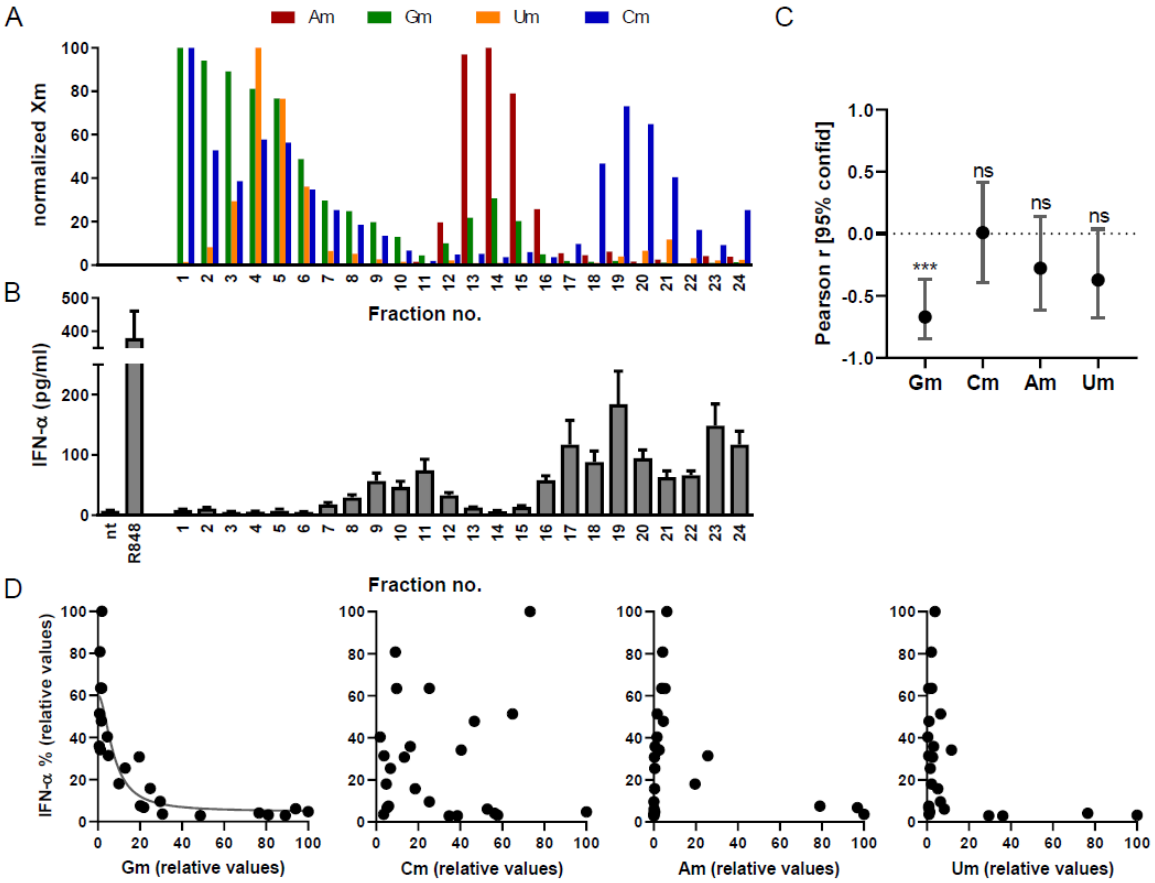


Figure 1: IFN-a production by PBMCs upon stimulation with fractionated tRNAs from *S. cerevisiae* inversely correlates with the Gm content

(A) Relative levels of 2'-O-methylated nucleotides quantified by LC-MS/MS in *S. cerevisiae* fractions that were generated as described in the materials and methods section. Values were normalized to highest relative amount of each modified nucleotide. (B) Human PBMCs (400,000/well in a total volume of 200 µl) were stimulated overnight in duplicate wells with fractionated and DOTAP-encapsulated tRNAs (c= 1 µg/ml). IFN-α release was measured in cell-free supernatants by ELISA. Bars represent pooled data from 3 individual donors +SEM. (C) A correlation matrix of IFN-α secretion and relative amount of Gm, Cm, Am, Um was calculated for Pearson's r including 95% confidence interval. (***) p<0.001 (D) Individual correlations of IFN-α and Xm, normalized to the highest respective value are shown. For Gm goodness of fit was calculated using an appropriated nonlinear dose response model (GraphPad Prism).

To confirm these starting observations of a negative regulatory function of Gm in *S. cerevisiae* we used a targeted, independent experimental approach: an isogenic mutant strain deficient in *trm3*, the *trmH* homolog in yeast, was analyzed. Total tRNA and whole cellular RNA preparations were purified from WT and mutant *S. cerevisiae*. We used unbiased RiboMethSeq analysis[356] to study the impact of loss of *trm3* on the yeast tRNA modification pattern: Thereby we confirm the presence of Gm18 in eight tRNA isoacceptors corresponding to five tRNA species in WT samples[30]. *trm3* deficient cells showed a lack of Gm18 in total tRNA extracts (Fig. 2 A). We also used a complementary LC-MS/MS approach that demonstrated a residual Gm content in these preparations; this can be explained by the presence of Gm at position 34 in $tRNA^{Phe}$ (Fig. 2 B and [30]). In line with the results obtained with tRNA fractions, specific lack of Gm18 in *trm3* mutant *S. cerevisiae* significantly enhanced TLR7 stimulatory properties of total tRNA preparations, although IFN-a secretion was generally lower as compared to stimulation with bacterial RNA, the latter serving as positive control (Fig. 2 C). Thus, a physiological role of Gm18 modified tRNA in eukaryotes could be demonstrated. However, whole cellular RNA from both WT and *Dtrm3 S. cerevisiae* failed to induce any TLR7 activation (Fig. 2 C). We also tested DtrmH *E. coli* lacking the corresponding Gm tRNA methyltransferase. Gm18 deficiency was also confirmed by RiboMethSeq and LC-MS/MS analysis. For DtrmH *E. coli* not only enhanced IFN-a release by total tRNA, as published previously[235, 304], but also by whole bacterial RNA preparations (Fig. 2 D-F) was observed. Thus, in both microorganisms the presence of a few Gm18 modified tRNA isoacceptors was sufficient to manipulate recognition of whole tRNA preparations confirming a dominant-inhibitory effect of this modification. Yet whole RNA preparations showed differing influence by loss of Gm18 in tRNA.

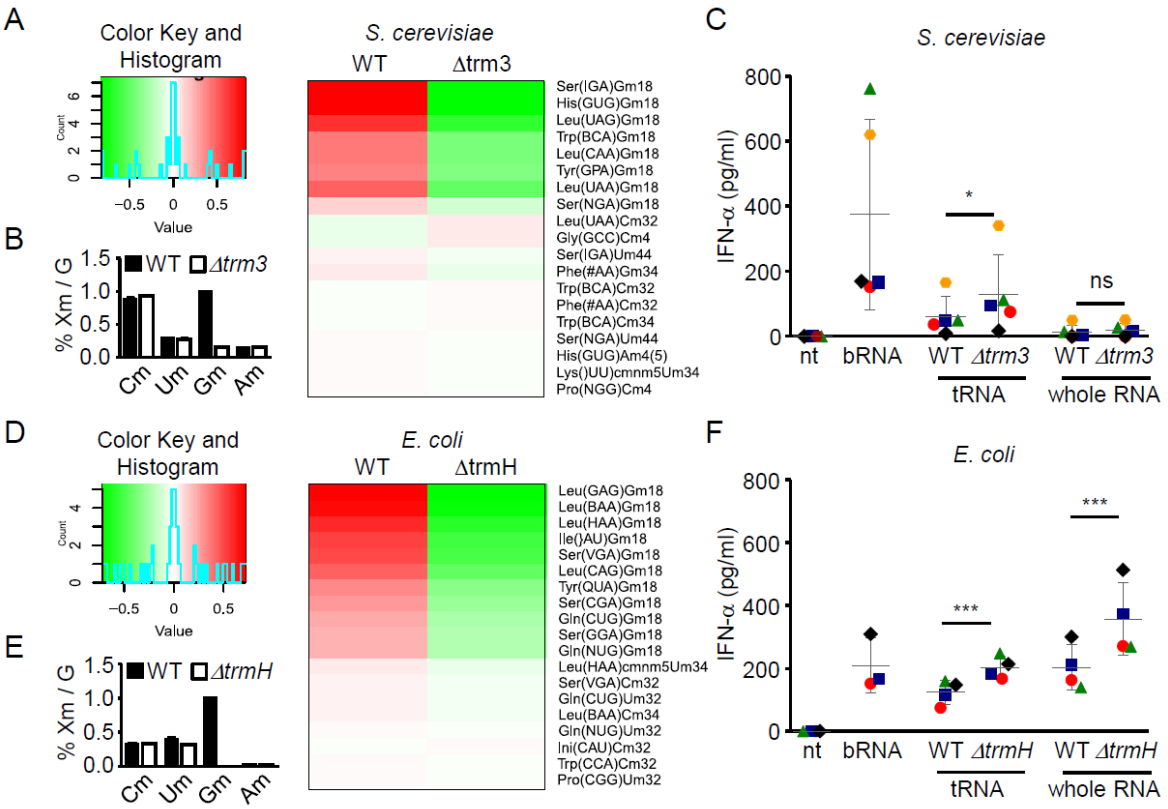


Figure 2: Gm18 deficiency differentially regulates immune stimulation by isolated RNA species derived from *S. cerevisiae* and *E. coli*

(A, D) RiboMethSeq and (B, E) quantitative LC/MS-MS analysis of isolated *S. cerevisiae* (A-B) and *E. coli* (D-E) total tRNA preparations derived from WT strains and 2'-O-methyltransferase deficient mutants (*Dtrm3, DtrmH*). (A, D) Heatmaps display normalized mean methylation scores. tRNA isoacceptors and modified nucleotide positions are indicated on the right. (B) Percentage of Xm modified nucleotides normalized to total guanosine content. (C, F) Human PBMCs were transfected overnight with DOTAP-encapsulated total tRNAs and whole cellular RNA preparations (c=0.5 µg/ml) from WT and 2'-O-methyltransferase deficient *S. cerevisiae* (*Dtrm3*) (C) or *E. coli* (*DtrmH)* (F). Stimulation with bacterial RNA derived from *S. aureus* served as positive control. IFN-α release was quantified in cell-free supernatants by ELISA. Each data point represents the mean value of one donor stimulated in duplicate wells. Error bars indicate mean ±SD of 3-5 individual donors. (*) p<0.05; (***) p<0.001

Gm18 deficiency does not alter cytokine production in PBMCs upon infection with living *E. coli* and *S. cerevisiae*

In order to investigate whether Gm18 not only affects immune stimulation by purified RNA species but also by infection with the whole organism which is composed of a variety of PAMPs activating potentially a large set of PRRs, PBMCs were infected with *E. coli* and *S. cerevisiae* WT and Gm18 deficient mutant strains at various multiplicities of infection (MOIs) (Fig 3 A-D). Although total cellular RNA preparations isolated from *E. coli DtrmH* induced higher levels of IFN-a than their WT counterparts (Fig. 2 F), no differences in IFN-a or TNF release were observed upon infection with the whole, living organism (Fig. 3 A, B). Similar results were obtained for infection with WT and *Dtrm3 S. cerevisiae* (Fig. 3 C, D). Of note, infection with *S. cerevisiae* entirely failed to induce IFN-a, as already observed for stimulation with total cellular RNA (Fig. 2 C and Fig. 3 C). While upon transfection of contaminant-free RNA into PBMCs TNF is commonly used as read-out for TLR8-dependent monocyte activation, this cytokine can also be induced by a variety of other PAMPs during infection with whole pathogens, for example via TLR4 by *E. coli* LPS or by yeast cell wall mannans[381–383]. To better appreciate potential RNA-dependent effects, analysis was therefore expanded to the production of IFN-b, a mediator which is more prominently induced by RNA than by microbial cell wall components in both a TLR7- and TLR8-dependent manner. Yet, neither for *E. coli* nor for S*. cerevisiae*, lack of Gm18 affected IFN-b induction (Fig. 3 E, F). Thus, the data do not provide evidence for a role of Gm18 as immune escape mechanism under physiological conditions of infection within the context of the here analysed living whole microorganisms.

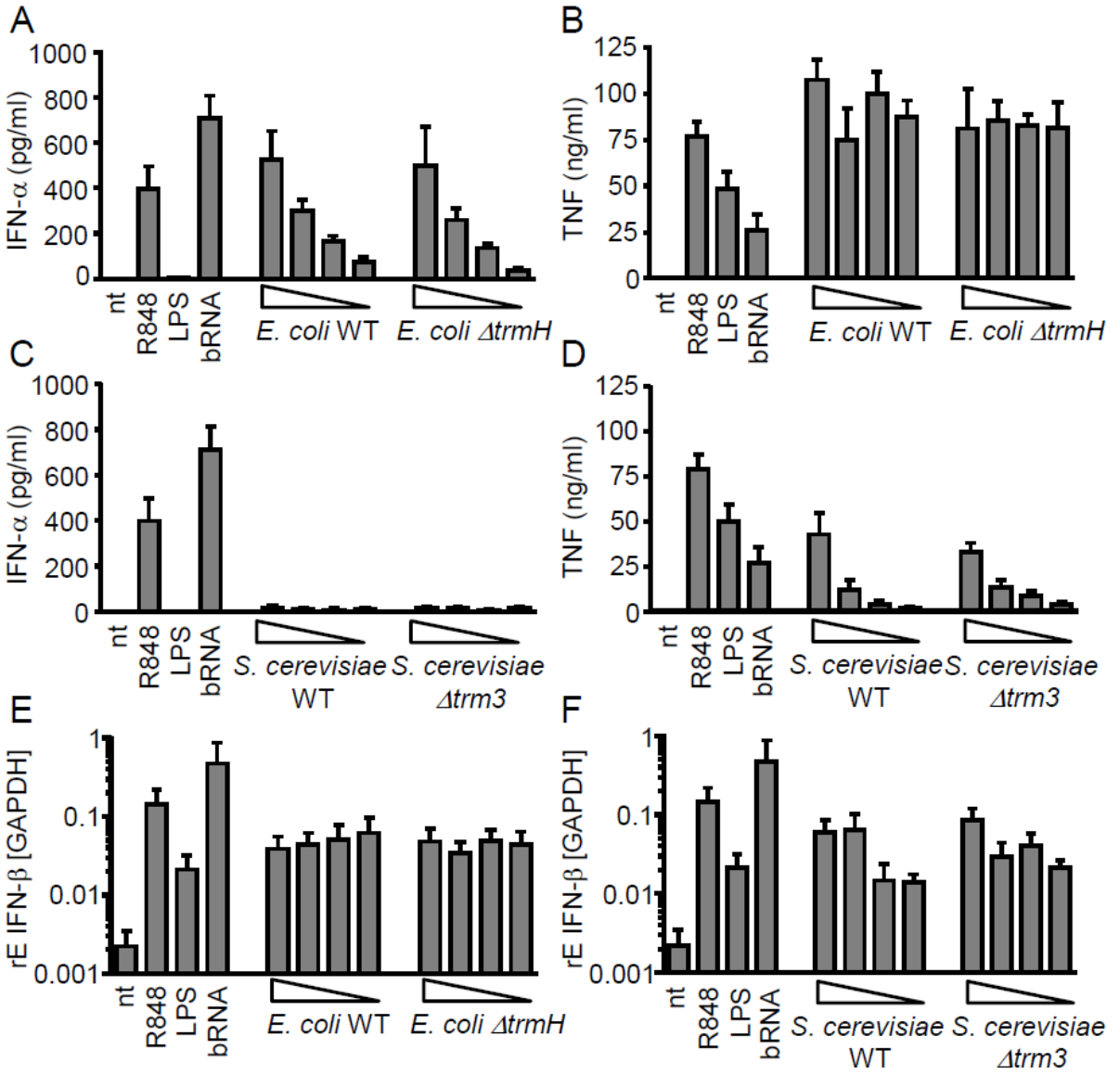


Figure 3: Lack of Gm18 does not affect innate immune sensing of *E. coli* and *S. cerevisiae* upon infection of PBMCs

(A-D) Human PBMCs were infected in duplicate wells with *E. coli* (A, B) or *S. cerevisiae* (C, D) WT strain or methyltransferase deficient mutant at different MOIs (wedges indicate MOI 1, 0.3, 0.1 and 0.03) and supernatants were collected after overnight incubation. Stimulation with LPS (TLR4 ligand), R848 (TLR7/8 ligand) and bacterial RNA (bRNA, c=0.5 µg/ml) derived from *S. aureus* served as controls. Levels of IFN-α (A, C) and TNF (B, D) were measured in cell-free supernatants by ELISA. (E, F) PBMCs were infected as in (A-D) and relative expression of IFN-β mRNA was assessed by qPCR 5 h post infection with the indicated microorganisms. All results represent mean data (+SEM) of 3 independent experiments with different donors. Statistical analysis did not show any significant difference between infection with the WT strain and the respective Gm18 deficient mutant.

Gene expression profiling reveals TLR8-dependent pathways in *E. coli* sensing

Next, we aimed to more broadly explore a potential impact of RNA sensing in general and of Gm18 in particular on *E. coli* mediated immune activation, as D*trmH E. coli* gave differing results for whole RNA vs. whole bacterial cell stimulation (Figs. 2, 3). To this end, microarray-based gene expression profiling was performed in trans-differentiated BLaER1 cells which recognize bacterial and archaeal RNA in a TLR8-dependent manner (Fig. 4 A and [384]) and are used to model primary human monocyte responses[385, 386]. In a first experiment, WT BLaER1 cells were infected with *E. coli* WT and *trmH* deficient mutant for 5 h to specifically detect Gm18 regulated pathways. This time point was evaluated as optimal for differential cytokine induction in preliminary time-course experiments (data not shown). Gene expression profiling revealed a total of 2066 and 2070 significantly regulated genes (defined as log2 fold-change ≥ +1 or ≤ -1 and p<0.05) upon infection with WT or *DtrmH E. coli*, respectively, when compared to the untreated control (Fig. 4 B). In both groups, the top regulated genes predominantly contained a variety of well-characterized inflammatory mediators involved in innate immunity, including IL-6, TNF and IFN-b (Suppl. Table 1 A, B). However, analysis demonstrated that none of these genes were significantly different between WT and *DtrmH E. coli* (Fig 4 C), indicating that Gm18 does not influence overall immune stimulation by *E. coli*. Yet, it is well known that *E. coli* possesses a very bioactive LPS triggering TLR4 and indeed the gene signature induced by *E. coli* showed a large overlap with LPS which was included as positive control (Fig.4 D).

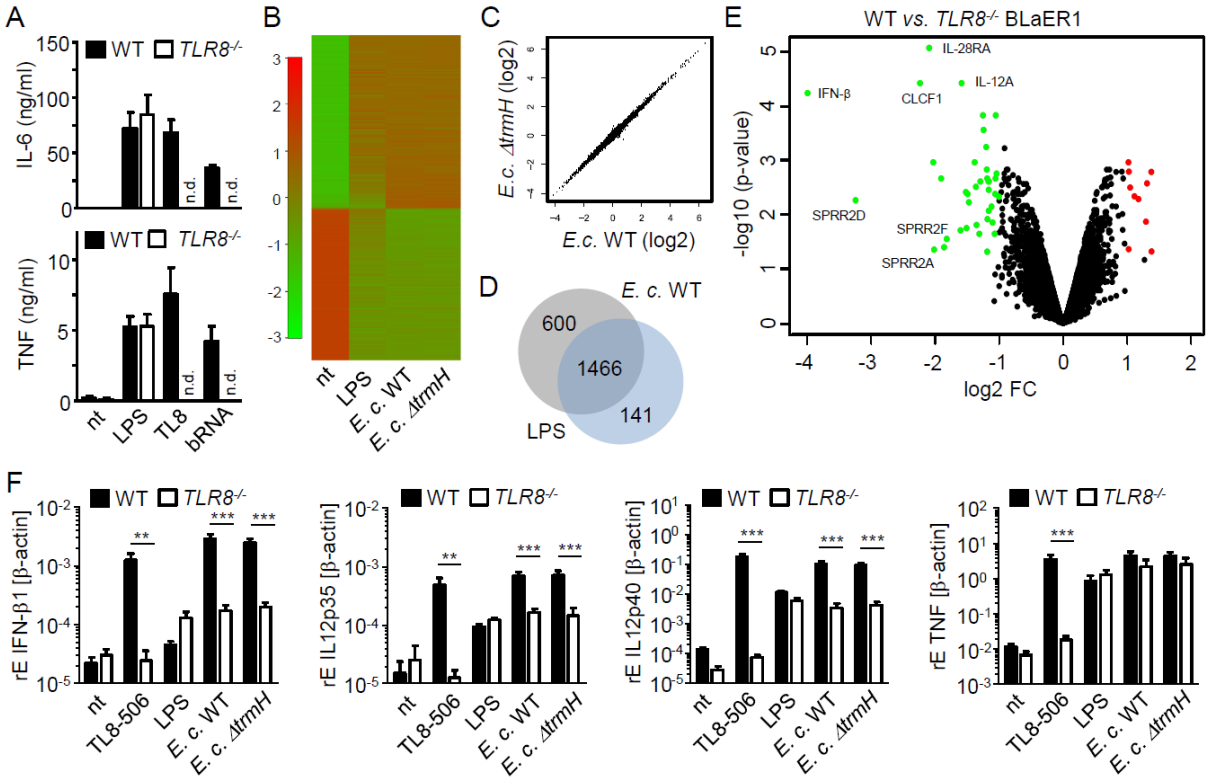


Figure 4: *E. coli* induces an RNA/TLR8-dependent gene signature which is not affected by Gm18

(A) BLaER1 cells were stimulated overnight with LPS, TLR8-specific ligand TL8-506 (TL8) and bacterial RNA from *S. aureus*. Levels of IL-6 and TNF were measured in cell-free supernatants by ELISA. (B-D) BLaER1 cell were stimulated in with whole, live *E. coli* WT, *E. coli DtrmH* (MOI 10) or LPS for 5 h and gene expression profiling was performed using an Illumina HT12 microarray. (B) Heatmap depicts 2304 significantly differentially regulated genes (defined as log2 fold-change >1 or <-1 and p<0.05; mean of 3 independent experiments) for each condition as compared to the non-treated control. For each condition, data of three independent biological replicates were pooled. (C) Scatter Plot displaying log2 intensity values of genes regulated by *E. coli* WT *versus E. coli DtrmH* relative to the unstimulated control. (D) Venn diagram of *E. coli* WT and LPS regulated genes. (E) Volcano plot of 482 regulated genes of BLaER1 WT compared to BLaER1 *TLR8*$^{-/-}$ upon stimulation with *E. coli* WT. Significantly differentially regulated genes (defined as log2 fold-change >1 or <-1 and p<0.05) are highlighted in green and red. All microarray data were analyzed using Chipster software. (F) WT and *TLR8*$^{-/-}$ BLaER1 cells were stimulated for 5 h with TLR8-specific ligand TL8-506, LPS, *E. coli* WT and *E. coli DtrmH* (MOI 10) and relative cytokine expression was determined by qPCR.Values represent mean data (+SEM) of 3 independent experiments. (***) p<0.001

We next asked the question to what extent an RNA-TLR8 axis contributed to *E. coli* induced cell activation in general. To this end, a gene expression analysis was performed in WT and *TLR8*$^{-/-}$ BLaER1 cells upon infection with WT *E. coli*. Interestingly, 50 genes were significantly differentially regulated (40 down and 10 up) in *TLR8*$^{-/-}$ as compared to WT cells, including IFN-b, IL-12p35 and three members of the small proline-rich protein family (SPRRs) whose function in innate immunity is as yet insufficiently defined (Fig. 4 E and Suppl. Table 1 C). These results thus clearly point towards an RNA-TLR8 dependent gene signature in *E. coli* sensing even in the presence of a strong LPS response. Microarray data were further validated by qPCR, confirming a strong reduction of IFN-b and partial impairment of IL-12p35 and IL-12p40 expression in *TLR8*$^{-/-}$ cells upon *E. coli* exposure, while induction of TNF and IL-6 were not altered (Fig. 4 F and data not shown). Together, these data demonstrate that despite the existence of TLR8-dependent pathways in *E. coli* recognition, Gm18 does not affect innate immune activation by this pathogen and thus does not serve as an immune escape mechanism under the here tested physiological conditions.

Gm18 deficiency in human cells enhances immune cell activation by total tRNA

In a next step, we further wanted to analyse the influence of Gm18 on immune stimulation in the human system (higher eukaryote). Therefore, *TARPB1*, the proposed human homolog of *trmH*, was targeted by CRISPR/Cas9 in HEK293 cells, a cell line that is easily to transfect and to genetically manipulate. Successful knock-out of *TARBP1* was confirmed by sequencing, demonstrating bi-allelic frame-shift inducing insertions resulting in altered amino acid sequences (Fig. 5 A), as well as on the

protein level (Fig. 5 B). Lack of Gm18 in tRNA derived from *TARBP1*-deficient cells was then analyzed by RiboMethSeq analysis (Fig. 5 C). Indeed we can show that *TARBP1* in human cells has the proposed 2'-O-methyltransferase activity. However, surprisingly, Gm18 was detected in only 3 isoacceptors corresponding to 2 tRNA species (Ser, Gln) in HEK WT cells but not in further tRNAs. Yet, despite the relatively low abundance of Gm18 in human tRNA, preparations derived from *TARBP1*$^{-/-}$ cells still caused a slight but statistically significant increase in immune stimulation by total tRNA preparations (Fig. 5 D). By contrast, whole cellular RNA was in general non-stimulatory and no differences were observed between RNA from WT and *TARBP1*$^{-/-}$ cells. This pattern was similar to the results obtained with *S. cerevisae* (Fig. 2 C) but different to *E. coli* (Fig. 2 F). We speculated that within whole cellular RNA the high content of ribosomal RNA might explain the differences. Indeed, when testing rRNA preparations we observed a clear discrimination of prokaryotic rRNA, which was stimulatory, from eukaryotic rRNA, that was inactive (Fig. 5 E), indicating a dominant contribution of immune silencing modifications present in rRNA.

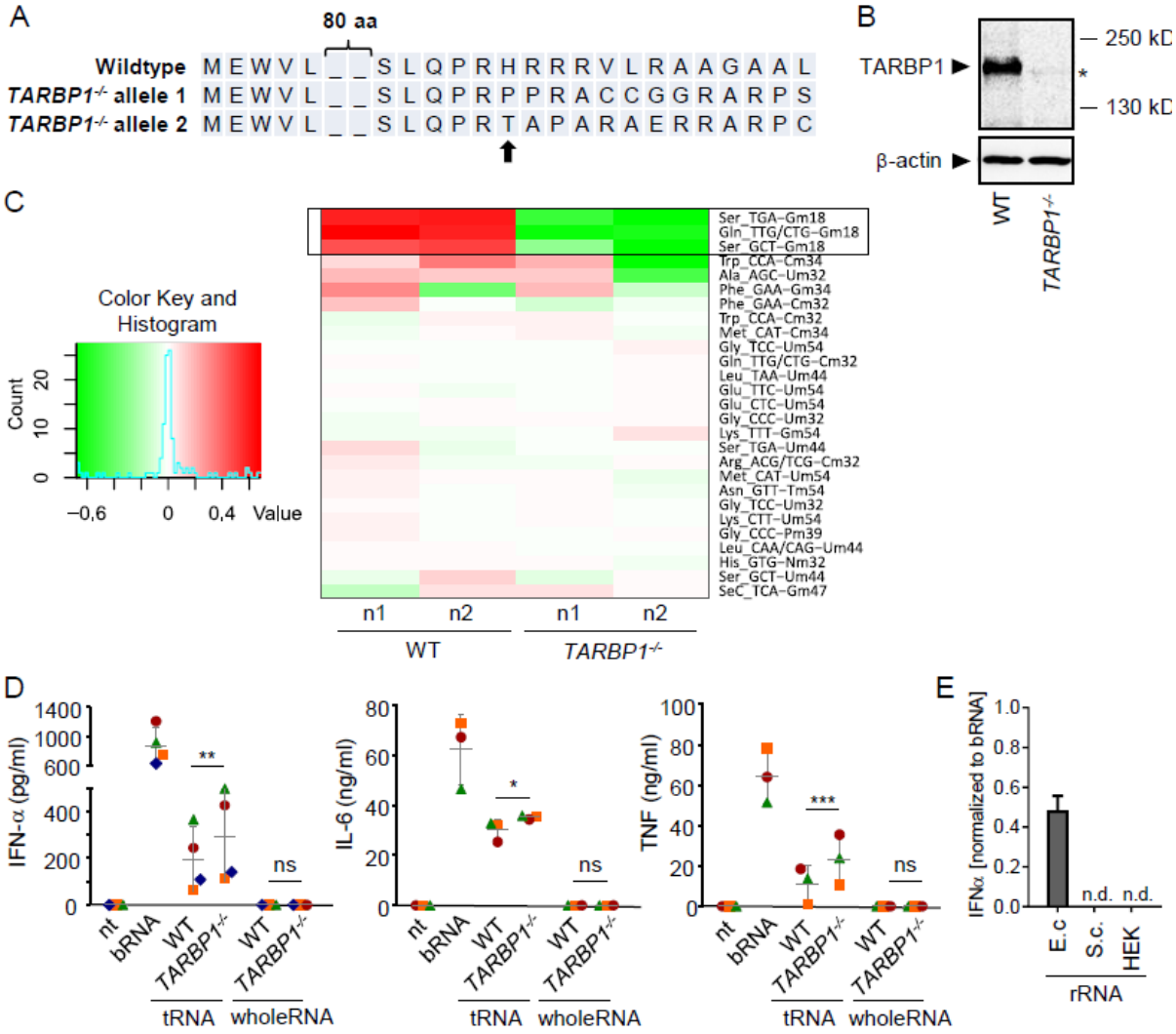


Figure 5: Gm18 deficiency in *TARBP1*$^{-/-}$ HEK cells enhances immune stimulation by isolated tRNAs but not by whole cellular RNA

(A) HEK cells in which TARBP1 was targeted by CRSIPR/Cas9 were analyzed by Sanger sequencing confirming bi-allelic frame-shift inducing indels resulting in an altered amino acid sequence. Arrow indicates cleavage site. (B) Western blot analysis of WT and *TARBP1*$^{-/-}$ HEK cells. (*) indicates unspecific band. β-actin was used as loading control. (C) RiboMethSeq analysis of total tRNA preparations from WT and *TARBP1*$^{-/-}$ HEK cells. Heatmap indicates 2'-O-methylation sites. Analysis was performed from two independent biological samples (n1 and n2). (D) Human PBMCs were stimulated overnight with 0.5 µg/ml DOTAP-encapsulated RNA samples. IFN-α, IL-6 and TNF release were measured in cell-free supernatants by ELISA. Each data point represents the mean value of one donor stimulated in duplicate wells. Error bars indicate mean ±SD of 3-4 individual donors. (*) $p<0.05$; (**) $p<0.01$; (***) $p<0.001$ (E) rRNA fractions were prepared from *E. coli*, *S. cerevisae* and HEK WT cells and used for stimulation of PBMCs. RNA was used at 0.5 µg/ml and secreted IFN-a was normalized to stimulation by 1 µg/ml whole bacterial RNA (n=3, mean+SD; n.d., not detectable).

Discussion

RNA modifications have been recognized as critical modulators of RNA-mediated innate immune responses and a number of studies have highlighted the TLR7/8 antagonistic properties of ribose 2'-O-methylation in this context[233, 237, 377, 387]. Yet, it was unknown whether Gm in its natural context plays a comparable immune-modifying role across different kingdoms. We now show that deletion of Gm18 by knock-out of the modifying enzymes trmH, trm3 and TARBP1 in prokaryotes, lower and higher eukaryotes results in an increased immune stimulation by tRNA. Using the unbiased RiboMethSeq approach to study frequencies of Gm, this immune modulatory effect could be observed although only few isoacceptors had a Gm modification, thus demonstrating a potent antagonistic activity of naturally occurring Gm18.

Apart from its role in microorganisms, we also investigated the significance of Gm18 in the human system. Surprisingly, RiboMethSeq analysis identified Gm18 in only three isoacceptors corresponding to two tRNA species (Ser and Gln). This finding was rather unexpected as human tRNA is generally believed to show the highest degree of post-translational RNA modifications. Yet, despite its low abundance, Gm18 deficiency still enhanced cytokine release by total tRNA, pointing towards a potential role of this modification for immune modulation. Mitochondrial tRNA, which accounts for approximately 13 +/-5% of total tRNA in our study, may at least partially contribute to the overall immune stimulation observed by tRNA preparations[375]. Despite an effect of TARBP1 that could only be observed for tRNA but not whole RNA, links to a physiological contribution of Gm18 to self/foreign discrimination have been published: Rare single-nucleotide variants were found in TARBP1 in exome sequencing studies associated with autoimmune psoriasis[388] and TARBP1 is differentially regulated between healthy patients and those with autoimmune SLE or ANCA vasculitis[389].

Moreover, HIV has been recently shown to use TARBP1 to recruit 2'-O-methyltransferase activity to its RNA to hide the RNA from immune recognition[255].
Human and *S. cerevisae* rRNA and total cellular RNA preparations lacked immune stimulatory properties entirely, suggesting the presence of relevant immune silencing modifications within rRNA. Of note, rRNA from *S. cerevisiae* and especially from *Homo sapiens* has been reported to be heavily 2'-O-methylated on both guanosine and adenosine residues by the 2'-O-methyltransferases NOP1 in yeast and its homolog fibrillarin in humans[64, 109, 390, 391]. Studying fibrillarin and rRNA 2'-O-methylation would be interesting but is challenging as NOP1 was described to be crucial for cell viability in yeast[391]. Interestingly, anti-fibrillarin antibodies are detected in the serum of patients suffering from autoimmune diseases[373, 392, 393].
The fact that 2'-O-methylation is also naturally occurring in a subset of tRNA isoacceptors at position 18 (Gm18) in *E. coli* gave rise to the hypothesis that this modification, beyond its structural role, may have acquired a secondary function as immune escape mechanism in organisms bearing this modification[235, 304, 394]. Yet, this idea was based on the transfection of purified tRNAs into PBMCs, a situation not adequately reflecting the physiological conditions of infection with the live microorganism. In the present study we demonstrate that Gm18 deficiency in *E. coli* indeed enhances immune stimulation by isolated total tRNA and even whole cellular RNA preparations, yet does not interfere with innate immune sensing of live *E. coli* by both PBMCs and BLaER1 cells. Similar results were obtained for *S. cerevisiae.* These results therefore argue against a role of Gm18 as a relevant immune escape mechanism under physiological conditions at least in the microorganisms tested. As tRNA constitutes only 10-20% of total *E. coli* RNA with only five tRNA species being Gm18 methylated, it is conceivable that the ratio of Gm18 methylated to unmethylated RNA might not be sufficient to suppress immune activation during infection when also rRNA is present. Due to the dominant action of LPS when studying *E. coli* (Fig. 4 D), a role of Gm18 might be more prominent in other bacteria. Another influencing factor might be accessibility of different RNA species.
The results are compatible with a model of a certain threshold for Gm content in an organism's cellular RNA to modulate TLR7/8 mediated immune responses: In *S. cerevisae* and human cells, the overall Gm content is high, especially due to its abundance in rRNA, and exceeds a certain threshold, thus resulting in lack of immune recognition. Deletion of Gm only in tRNA does not increase sensing of the whole organism as the overall Gm content is still sufficiently high (rRNA). In contrast in *E. coli*, Gm levels are generally low due to the lack of rRNA 2'-O-methylation, resulting in immune stimulation by all RNA species. However, when only tRNA is analysed, a situation of Gm enrichment is achieved which now reveals the antagonistic function of this modification.
E. coli is known to be recognized in a predominantly TLR4-dependent manner[395, 396], in line with a large overlap of the *E. coli* and LPS induced gene expression profiles observed in our study. Yet, microarray-based analysis of WT and *TLR8*$^{-/-}$ BLaER1 cells clearly identified an additional TLR8- and thus RNA-dependent gene signature

triggered by *E. coli* infection, including regulation of classical pro-inflammatory mediators like IFN-b and IL-12 as well as proteins with less well-defined functions in innate immunity (e.g. small proline-rich proteins, SPRRs). Interestingly, an RNA-dependent induction of specific cytokines by *E. coli* was also described by the groups of L. Sander and J. Blander who raised the concept of bacterial RNA as so-called 'vita-PAMP': Due to the rapid degradation of RNA after bacterial cell death, RNA was proposed to signify microbial viability and thus a persistent infectious threat to the host, culminating in more robust innate and adaptive immune responses[397–399]. Indeed, the here observed genes that are strongly dependent on TLR8 (e.g. IL-12 and IFN-b) have also be described by the Blander and Sander groups. While initial investigations emphasized the relevance of *E. coli* RNA for the induction of IFN-b and activation of the NLRP3 inflammasome with subsequent production of bioactive IL-1b[397], recent studies highlighted the contribution of microbial viability sensing via TLR8 for T follicular helper cell differentiation and antibody responses in a manner dependent on IL-12 and on an IFN-b/IL-1b axis[398, 399]. It might therefore be interesting to analyze T helper cell differentiation instead of global gene expression of monocytes by stimulation with *trmH* mutants as those might affect only a small, yet important part of the gene signature.
In summary, the findings of this study demonstrate that Gm18 is an immune inhibitory RNA modification across organisms of different kingdoms, yet it does not serve as an immune evasion mechanism in the studied organisms. The results also identify an impact of Gm18 on immune stimulatory properties within human tRNA; in conjunction with published data on differential TARBP1 regulation and SNPs in autoimmune diseases a potential contribution in tolerance to self-RNA can be envisioned.

Materials and Methods

Reagents

RPMI 1640 containing stable glutamine was purchased from Anprotec (Bruckberg, Germany), FCS was obtained from Gibco (Darmstadt, Germany), Ficoll (1.078 g/ml) and DPBS w/o Ca^{2+} and Mg^{2+} from Pan Biotech (Aidenbach, Germany), TRIzol from Thermo Fisher Scientific (Darmstadt, Germany), R848 and TL8-506 from Invivogen (San Diego, CA, USA), DOTAP (N-[1-(2, 3-dioleoyloxy)propyl]-N, N, N-trimethylammonium methylsulfate) from Roth (Karlsruhe, Germany), peptone (Sigma-Aldrich, Steinheim, Germany), yeast extract (AppliChem, Darmstadt, Germany) Nanosept MF, bio-inert membrane, purple, 0.45 µm from Pall Laboratory (Dreieich, Germany) and ammonium acetate from J.B Baker (Deventer, Netherlands).

Preparation of RNA samples from E. coli, S. cerevisiae and HEK cells

WT (parental strain: BW25113) and *trmH* deficient (JW3626-1) *Escherichia coli* (provided from *E. coli* Genetic Stock Center CGSC, Yale University) were grown in Luria-Bertani (LB) medium (Merck, Darmstadt, Germany) and harvested within the mid log phase growth, pelleted and cell wall was digested by lysozyme (c=40 mg/ml, 20 min at 37°C). WT (parental strain: Y00000) and *trm3* deficient (Y03809) *S. cerevisiae*

(purchased from Euroscarf, Johann Wolfgang Goethe-University, Frankfurt, Germany) were grown in YPD-Medium (20 g peptone; 10 g yeast extract; ad 975 ml dH_2O; pH 6,5), harvested in the mid log phase growth, pelleted and cell wall was digested by Lyticase (100 U, 20 min at 37°C). HEK cells were harvested using Trypsin-solution and pelleted at 1,300 rpm for 15 min. Total RNA of all species was isolated using TRIzol reagent according to the manufacturer's protocol. The obtained RNA was further separated by agarose gel electrophoresis and RNA bands corresponding to tRNA and/or rRNA were excised. Gel slices were transferred to a 2 ml reaction tube, mashed and 500 µl ammonium acetate 0.5 M was added. Samples were stored for 1 h at -80°C and shaken overnight at room temperature at 500 rpm to extract RNA. The next day, samples were transferred to a Nanosept filter and flow through was mixed with 1250 µl ethanol to precipitate RNA. RNA preparations were stored for 1-2 h at -80°C and centrifuged for 60 min at 10,000 g, 4°C. Pellets were washed in 70% ethanol and RNA was dissolved in nuclease free water. Purity of the RNA preparations was validated by determining the 260/230 nm and 260/280 nm extinction ratio by NanoDrop (Thermo Scientific). Only RNA preparations with ratios >1.8 were used for stimulation experiment.

Fractionation of S. cerevisiae tRNA

The fractions of tRNA yeast used in experiments shown in Figure 1 where generated several decades ago using fractions counter current distribution, an old separation technique which separated tRNA according to their relative lipophilicity. Despite longterm storage, analysis of the fractions by high-resolution capillary electrophoresis (QIAxcel, Qiagen, Hilden Germany) did not indicate significant degradation. According to a mapping of the tRNA distribution[380], the relevant fractions 1-6 contain Gm-containing tRNAs accepting the aminoacids Phe (Gm34), Ser, Tyr, Trp, and Leu (all Gm18) and fractions 13-15 contain His (Gm18, Am4) and Val (no Gm). Fractions 4-5 contained Ser (Um44)[30]. All analyses on the 2'-O-methylation status of the preexisting tRNA fractions using LC-MS and RiboMethSeq analysis, as described below, were performed during the course of the present study.

LC-MS measurements of Gm, Am, Um and Cm levels

tRNA samples were digested into nucleotides using 0.6 U nuclease P1 from *P. citrinum* (Sigma-Aldrich), 0.2 U snake venom phosphodiesterase from *C. adamanteus* (Worthington), 2 U FastAP (Thermo Scientific), 10 U benzonase (Sigma-Aldrich), 200 ng Pentostatin (Sigma-Aldrich) and 500 ng Tetrahydrouridine (Merck-Millipore) in 25 mM ammonium acetate (pH 7.5; Sigma-Aldrich) over night at 37°C. The nucleosides were then spiked with internal standard (^{13}C stable isotope-labeled nucleosides from *S. cerevisiae*, SIL-IS) and subjected to analysis. Technical triplicates with 100 ng digested RNA and 50 ng internal standard were analyzed via LC-MS (Agilent 1260 series and Agilent 6460 Triple Quadrupole mass spectrometer equipped with an electrospray ion source (ESI)). The solvents consisted of 5 mM ammonium acetate buffer (pH 5.3; solvent A) and LC-MS grade acetonitrile (solvent B; Honeywell). The

elution started with 100% solvent A with a flow rate of 0.35 ml/min, followed by a linear gradient to 8% solvent B at 10 min and 40% solvent B after 20 min. Initial conditions were regenerated with 100% solvent A for 10 min. The column used was a Synergi Fusion (4 µM particle size, 80 Å pore size, 250 × 2.0 mm; Phenomenex). The UV signal at 254 nm was recorded via a diode array detector (DAD) to monitor the main nucleosides. ESI parameters were as follows: gas temperature 350°C, gas flow 8 l/min, nebulizer pressure 50 psi, sheath gas temperature 350°C, sheath gas flow 12 l/min, capillary voltage 3000 V. The MS was operated in the positive ion mode using Agilent MassHunter software in the dynamic MRM (multiple reaction monitoring) mode. For quantification, a combination of external and internal calibration was applied as described previously[400].

RiboMethSeq Analysis

Total tRNA RiboMethseq analysis was performed essentially as described previously[66]. Briefly, about 100 ng of total tRNA either from *E. coli*, *S. cerevisiae* or human HEK cells were randomly fragmented in bicarbonate buffer for different times (6-10 min for *E. coli*), (8-12 min for *S. cerevisiae*) and (14 min for human HEK) at 95°C. The reaction was stopped by ethanol precipitation as previously described in [64]. RNA fragments were then subjected to end-repair and converted to library using NEBNext® Small RNA Library kit (ref#E7330S, New England Biolabs, UK) following the manufacturer's instructions. Library sequencing was performed using Illumina HiSeq 1000 sequencer in a single read mode for 50 nt. Analysis of human tRNA 2'-O-methylation was performed by trimming of raw sequencing reads using Trimmomatic v0.32, followed by end-to-end alignment by Bowtie 2.0 to non-redundant collection of human intronless tRNA genes. Reference sequence was created by collapsing of identical tRNA sequences and deduction of the consensus sequences from closely related species. The resulting reference allowed unambiguous alignment of > 80% of tRNA reads. Further data analysis was done by previously described bioinformatic pipeline[356]. Briefly, RiboMethSeq scores were calculated from the cleavage profiles derived from cumulated 5'- and 3'-end counts. Candidate 2'-O-methylated sites in tRNA species were identified by values of Score A >0.5 and Score Mean >0.90, followed by visual inspection of the coverage profiles to exclude irregularly cleaved regions. MethScore (Score C2) was used to quantify the 2'O-methylation level. Differential heatmaps for figures were constructed by normalization of MethScore levels for every given position to the average value observed in the series of samples. Positive (shown in red) values indicate higher methylation compared to the average (overmethylation), while negative (shown in green) positions have MethScore below the average (undermethylation).

Generation of TARBP-1 deficient HEK cells using CRISPR-Cas9

TARBP1 deficient HEK cells were created via a *S. pyogenes* derived CRISPR/Cas9 system using a NGG PAM sequence. gRNAs were designed by using Blue Heron webpage (www.blueheronbio.com) with a GC content between 60 and 80% and

blasted to control for off target effects. gRNAs were custom synthesized by Eurofins Genomics (Ebersberg, Germany) and cloned into a GFP-encoding vector backbone using the restriction enzyme BbsI[401]. Plasmids were kindly provided by Prof. Dr. Dirk Grimm, BioQuant, University of Heidelberg, Heidelberg, Germany. The *TARBP1*$^{-/-}$ clone used in the current study was generated using following gRNA fw: 5'-CACCGCGGCGCCAGGCGCGCGCCAC-3'; rev: 5'-AAACGTGGCGCGCGCCTGGCGCCGC-3'. Nucleotides highlighted in grey indicate the overhang required for cloning into the BbsI digested vector. Resulting plasmids were transfected into HEK293 cells using Lipofectamine2000 and GFP-positive cells were sorted into 96-well plates 24 h post transfection. After 1-2 weeks, single clones were analyzed for successful targeting of TARBP1 by both Sanger sequencing (primer sequences for PCR and sequencing: fw 5'- ACTCCTGGCCTTACGTCTAAATC-3', rev 5'- ACGGCTGCTAGCACTTCCAC-3') and Western Blot analysis (anti-TARBP1 antibody [EPR8727] Abcam, Cambridge, United Kingdom). Sequences were analyzed using CRISP-ID software[402].

Isolation and stimulation of human peripheral blood mononuclear cells (PBMCs)

Human PBMCs were isolated from heparinized blood of healthy volunteers upon informed consent and approval by the local ethical committee (admission number: S-716/2017) by standard Ficoll-Hypaque density gradient centrifugation (Ficoll 1.078 g/ml). PBMCs were resuspended in RPMI 1640 supplemented with 2% heat-inactivated human serum. For transfection experiments, all RNAs were encapsulated with DOTAP at a ratio of 3 µl DOTAP per 1 µg of RNA in serum-free medium according to the manufacturer's protocol. Unless otherwise indicated, cells were stimulated with RNA preparations at a final concentration of 500 ng/ml and transfection was performed at a density of 2×10^5 PBMCs per well in a 96-well flat bottom plate at a final volume of 100 µl. Cells were incubated overnight in a humidified 5% CO_2 atmosphere at 37°C. For cytokine measurement, levels of TNF, IL-6 (BD, Heidelberg, Germany) and IFN-a (eBioscience, Frankfurt, Germany) were determined in cell-free supernatants by ELISA.

Trans-differentiation and stimulation of BLaER1 cells

WT and *TLR8*$^{-/-}$ B-cell like BLaER1 cells, the latter generated as previously described[384], were kindly provided by Holger Heine, Research Center Borstel, Germany. BlaER1 cells were trans-differentiated into monocyte/macrophage-like cells by seeding 1 million cells/well into a 6-well in 3 ml/well RPMI 1640 medium supplemented with 10% FCS, 1% Pen/Strep, 10 ng/ml IL-3, 10 ng/ml M-CSF and 150 nM b-Estradiol. Plates were incubated at 37 °C and 5% CO_2. At day 2 and 5 of trans-differentiation, cells were collected by pipetting up and down, centrifuged at 1,300 rpm for 10 min and fresh medium with afore mentioned supplements was added. After one week, BLaER1 cells were sorted using AutoMACS magnetic cell sorter system and the human CD19 MicroBeads kit according to manufacturer's instructions (Miltenyi Biotech, Bergisch Gladbach, Germany) to deplete residual $CD19^+$ BLaER1 cells.

CD19 negative, trans-differentiated monocyte/macrophage like BlaER1 cells were resuspended in RPMI 1640 supplemented with 10% FCS. For experiments involving measurement of IL-6 and TNF in cell-free supernatants by ELISA (BD, Heidelberg, Germany), 1 x 10^5 cells were transferred into a 96-well flat bottom plate and stimulated overnight in duplicate wells with LPS (c=100 ng/ml), TL8-506 (c=100 ng/ml) and bacterial RNA derived from *S. aureus* complexed with DOTAP at ratio of 3 µl DOTAP per 1 µg RNA (c=500 ng/ml, *S. aureus* RNA was isolated as described previously[377]. For microarray analysis and for assessment of cytokine induction by quantitative real-time PCR, 2 x 10^5 cells were seeded in a 48-well plate in a final volume of 250 µl. To avoid pre-activation of BLaER1 cells due to MACS sorting, cells were rested overnight at 37 °C, 5% CO_2 prior to stimulation with LPS, TL8-506 or infection with live WT and *ΔtrmH E. coli* (MOI 10) for 5 h in duplicate wells. 1.5 h post infection, Penicillin/Streptomycin was added to kill extracellular bacteria.

Microarray analysis and quantitative real-time PCR

For microarray analysis and quantitative real-time PCR total RNA was extracted 5 h post stimulation or infection using the RNeasy mini kit including an on-column DNase treatment according to the manufacturer's instructions (Qiagen; Hilden, Germany). For microarray based gene expression profiling, biotin-labelled cRNA of three independent experiments was hybridized to a whole-genome HumanHT-12 array from Illumina (Santa Clara, CA, USA) according to the manufacturer's instructions and Microarray scanning was carried out using an iScan array scanner. As test for significance, a Bayes test was used on the bead expression values of the groups of interest. The average expression value was calculated as mean of the measured expressions of beads together with the standard deviation of the beads. Genes with a p-value < 0.05 were selected and log2 fold expression changes of differentially regulated genes were calculated. All analyses were performed using Chipster Software. For quantitative real-time PCR total RNA was reverse transcribed using 10 µl isolated RNA in a total reaction volume of 20 µl with High Capacity cDNA RT Kit (Applied Biosystems; Darmstadt, Germany) according to the manufacturer's instructions. For quantitative real-time PCR analysis, SYBR®Green PCR Master Mix Fast (Applied Biosystems; Darmstadt, Germany) was used. noRT and non-template controls as well as melt curve analysis served as specify control. Relative expression of target gene mRNA (for Primers see Table 1) was calculated by ΔCT compared to the indicated housekeeping gene.

Table 1: List of primer sequences used for quantitative real-time PCR

Target	Forward (5' – 3')	Reverse (5' – 3')
β-actin	AGAGCTACGAGCTGCCTGAC	AGCACTGTGTTGGCGTACAG
GAPDH	ACGGATTTGGTCGTATTGGGC	TTGACGGTGCCATGGAATTTG

IFN-β	ATGACCAACAAGTGTCTCCTCC	GGAATCCAAGCAAGTTGTAGCTC
IL-12p35	CTCCTGGACCACCTCAGTTTG	GGTGAAGGCATGGGAACATT
IL-12p40	GCGGAGCTGCTACACTCTC	CCATGACCTCCATGGGCAGAC
TNF	GCCCAGGCAGTCAGATCATCTTC	TGAGGTACAGGCCCTCTGATGG

Statistical analysis

Statistical significance between groups was determined by the two-tailed Student's t-test or 2-way ANOVA with Bonferroni post-test. For experiments comparing immune stimulation by RNA species derived from WT cells as compared to the respective Gm18 deficient mutant, significance was calculated using the linear mixed-effects model with Tukey's rank test to account for donor variation. Differences were considered significant for (*) $p<0.05$, (**) $p<0.01$, (***) $p<0.001$. Analyses were performed using Graph Pad Prism 7 (Graph Pad Software, San Diego, USA) or using R version 3.3.0 with the package nlme.

Abbreviations

MOI, multiplicity of infection; PBMC, peripheral blood mononuclear cells; *E. coli, Escherichia coli; S. cerevisiae, Saccharomyces cerevisiae*, *S. aureus; Staphylococcus aureus*; PAMP, pathogen-associated molecular pattern; PRR, pattern recognition receptor; TLR, Toll-like receptor

Acknowledgements

We thank Suzan Leccese and Selina Hassel for excellent technical support. We are grateful to Gérard Keith (IBMC Strasbourg, France) for yeast tRNA fractions. This work was supported by the German Research Foundation (DFG) grant DA592/5 and /7 (to A.D.) and He 3397/18-1 (to M.H.).

Supplemental Table 1

A) Top 20 up-regulated genes for stimulation with WT *E. coli* in WT BLaER1 cells as compared to the untreated control

Name	Description	p-value	log2 fold change
LTA	lymphotoxin alpha	$< 1 \times 10^{-6}$	6.43
CCL20	chemokine (C-C motif) ligand 20	1×10^{-6}	5.89
LOC728830	similar to C-C motif chemokine 3-like 1 precursor	$< 1 \times 10^{-6}$	5.72
IL6	interleukin 6	4×10^{-6}	5.61
LOC387763	hypothetical LOC387763	$< 1 \times 10^{-6}$	5.38
HES4	hes family bHLH transcription factor 4	$< 1 \times 10^{-6}$	5.27
MMP10	matrix metallopeptidase 10	$< 1 \times 10^{-6}$	5.19
IL1RN	interleukin 1 receptor antagonist	$< 1 \times 10^{-6}$	5.07
CD40	CD40 molecule, TNF receptor superfamily member 5	$< 1 \times 10^{-6}$	5.01
TNF	tumor necrosis factor	$< 1 \times 10^{-6}$	5.01
SERPINE2	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2	8×10^{-6}	4.99
MIR155HG	MIR155 host gene, non-coding RNA	1×10^{-6}	4.98
LOC730249	similar to Immune-responsive protein 1	$< 1 \times 10^{-6}$	4.84
CCL5	chemokine (C-C motif) ligand 5	$< 1 \times 10^{-6}$	4.84
IFNB1	interferon, beta 1, fibroblast	$< 1 \times 10^{-6}$	4.83
TRAF1	TNF receptor-associated factor 1	$< 1 \times 10^{-6}$	4.73
CD40	CD40 molecule, TNF receptor superfamily member 5	$< 1 \times 10^{-6}$	4.72
CCL14	chemokine (C-C motif) ligand 14, transcription variant 2	$< 1 \times 10^{-6}$	4.41
CCL4L1	chemokine (C-C motif) ligand 4-like 1	$< 1 \times 10^{-6}$	4.37
LRRC50	dynein, axonemal, assembly factor 1	$< 1 \times 10^{-6}$	4.37

B) Top 20 up-regulated genes for stimulation with *E. coli* ΔtrmH in WT BLaER1 cells as compared to the untreated control

Name	Description	p-value	log2 fold change
LTA	lymphotoxin alpha	$< 1 \times 10^{-6}$	6.45
CCL20	chemokine (C-C motif) ligand 20	1×10^{-6}	5.84
LOC728830	similar to C-C motif chemokine 3-like 1 precursor	$< 1 \times 10^{-6}$	5.77
IL6	interleukin 6	3×10^{-6}	5.55
LOC387763	hypothetical LOC387763	$< 1 \times 10^{-6}$	5.42
MMP10	matrix metallopeptidase 10	$< 1 \times 10^{-6}$	5.25
IL1RN	interleukin 1 receptor antagonist	$< 1 \times 10^{-6}$	5.25
CD40	CD40 molecule, TNF receptor superfamily member 5	$< 1 \times 10^{-6}$	5.16
HES4	hes family bHLH transcription factor 4	$< 1 \times 10^{-6}$	5.11
CCL5	chemokine (C-C motif) ligand 5	$< 1 \times 10^{-6}$	5.02
SERPINE2	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 2	3×10^{-6}	5.01
MIR155HG	MIR155 host gene, non-coding RNA	$< 1 \times 10^{-6}$	5.00
TNF	tumor necrosis factor	$< 1 \times 10^{-6}$	4.88
TRAF1	TNF receptor-associated factor 1	$< 1 \times 10^{-6}$	4.86
IFNB1	interferon, beta 1, fibroblast	$< 1 \times 10^{-6}$	4.80
LOC730249	similar to Immune-responsive protein 1	$< 1 \times 10^{-6}$	4.79
CD40	CD40 molecule, TNF receptor superfamily member 5	$< 1 \times 10^{-6}$	4.77
CCL4L1	chemokine (C-C motif) ligand 4-like 1	$< 1 \times 10^{-6}$	4.39
CCL14	chemokine (C-C motif) ligand 14, transcription variant 2	$< 1 \times 10^{-6}$	4.33
LRRC50	dynein, axonemal, assembly factor 1	1×10^{-6}	4.30

C) Differentially expressed genes for stimulation with WT *E. coli* in WT *vs*. *TLR8*$^{-/-}$ BLaER1 cells as compared to the untreated control

Name	Description	p-value	log2 fold change
Down-regulated in *TLR8*$^{-/-}$ BLaER1 cells			
IFNB1	interferon, beta 1, fibroblast	5.5×10^{-5}	-3.99
SPRR2D	small proline-rich protein 2D	5.3×10^{-3}	-3.24
CLCF1	cardiotrophin-like cytokine factor 1	3.6×10^{-5}	-2.24
IL28RA	interferon, lambda receptor 1	8.0×10^{-6}	-2.09
KCNJ2	potassium channel, inwardly rectifying subfamily J, member 2	1.1×10^{-3}	-2.03
SPRR2A	small proline-rich protein 2A	4.4×10^{-2}	-2.02
RGS1	regulator of G-protein signaling 1	2.1×10^{-3}	-1.91

SLAMF1	signaling lymphocytic activation molecule family member 1	4.0×10^{-2}	-1.86
SPRR2F	small proline-rich protein 2F	2.8×10^{-2}	-1.82
EDN1	endothelin 1	1.9×10^{-2}	-1.60
IL12A	interleukin 12A	3.6×10^{-5}	-1.59
PRAGMIN	homolog of rat pragma of Rnd2	3.8×10^{-3}	-1.52
CD40	CD40 molecule, TNF receptor superfamily member 5	1.8×10^{-2}	-1.51
CD40	CD40 molecule, TNF receptor superfamily member 5	4.0×10^{-3}	-1.49
CTSC	cathepsin C	5.8×10^{-3}	-1.47
RGS2	regulator of G-protein signaling 2	1.1×10^{-3}	-1.38
EXT1	exostosin glycosyltransferase 1	3.0×10^{-3}	-1.36
LILRA5	leukocyte immunoglobulin-like receptor, subfamily A (with TM domain), member 5	1.5×10^{-2}	-1.35
TNFSF9	tumor necrosis factor (ligand) superfamily, member 9	2.2×10^{-2}	-1.31
LILRB2	leukocyte immunoglobulin-like receptor, subfamily B (with TM and ITIM domains), member 2	2.4×10^{-3}	-1.29
CCL2	chemokine (C-C motif) ligand 2	1.4×10^{-4}	-1.25
RAB3IP	RAB3A interacting protein	2.7×10^{-4}	-1.24
RAB11FIP1	RAB11 family interacting protein 1 (class I)	5.5×10^{-4}	-1.20
CRIP1	cysteine-rich protein 1 (intestinal)	1.4×10^{-3}	-1.19
NEU4	sialidase 4	2.1×10^{-3}	-1.19
LILRA5	leukocyte immunoglobulin-like receptor, subfamily A (with TM domain), member 5	4.8×10^{-2}	-1.19
CCL7	chemokine (C-C motif) ligand 7	1.2×10^{-2}	-1.18
ADORA2B	adenosine A2b receptor	2.4×10^{-3}	-1.17
TMEM217	transmembrane protein 217	3.4×10^{-3}	-1.16
ID3	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein	8.4×10^{-3}	-1.16
EPHA4	EPH receptor A4	2.4×10^{-3}	-1.16

DKFZp761P0423	homolog of rat pragma of Rnd2	7.0×10^{-3}	-1.12
LOC728253	hypothetical LOC728253	1.4×10^{-2}	-1.10
NCOA7	nuclear receptor coactivator 7	2.2×10^{-3}	-1.07
PGBD5	piggyBac transposable element derived 5	4.1×10^{-3}	-1.06
RND3	Rho family GTPase 3	2.1×10^{-3}	-1.06
PDCD1	programmed cell death 1	1.4×10^{-4}	-1.05
GPR132	G protein-coupled receptor 132	1.7×10^{-3}	-1.04
GLA	galactosidase, alpha	4.6×10^{-3}	-1.00
IL1RN	interleukin 1 receptor antagonist	5.4×10^{-3}	-1.00
Up-regulated in *TLR8*$^{-/-}$ BLaER1 cells			
SERTAD2	SERTA domain containing 2	1.6×10^{-3}	1.38
HEY1	hes-related family bHLH transcription factor with YRPW motif 1	4.6×10^{-2}	1.38
MMP1	matrix metallopeptidase 1	2.6×10^{-3}	1.31
SBSN	suprabasin	1.3×10^{-2}	1.29
LFNG	O-fucosylpeptide 3-beta-N-acetylglucosaminyltransferase	5.1×10^{-3}	1.18
HS.374257	mRNA; cDNA DKFZp779K2051	4.5×10^{-3}	1.11
HAVCR2	hepatitis A virus cellular receptor 2	3.1×10^{-3}	1.05
CCR7	chemokine (C-C motif) receptor 7	1.6×10^{-3}	1.03
CNKSR3	CNKSR family member 3	1.1×10^{-3}	1.02
CSF3	colony stimulating factor 3 (granulocyte)	4.3×10^{-2}	1.02

-- End of article --

4.1.5 Machine learning algorithm for precise prediction of 2'-O-methylation (Nm) sites from experimental RiboMethSeq datasets. [DOI: 10.1016/j.ymeth.2022.03.007]

Florian PICHOT[1,2], Virginie MARCHAND[2], Mark HELM[1] and Yuri MOTORIN[2,3]

[1] Institute of Pharmacy of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Mainz, Germany.

[2] Université de Lorraine, CNRS, INSERM, UAR2008/US40 IBSLor, EpiRNA-Seq Core facility, F-54000 Nancy, France.

[3] Université de Lorraine, CNRS, UMR7365 IMoPA, F-54000 Nancy, France.

Overview

After scores optimisation discussed earlier inside the Section 4.1.2, we decided to assess the potentiality of using machine learning algorithms to further automatise and improve modified site predictions. Machine learning algorithms are powerful tools to solve complex data problems if used correctly[403]. Unsupervised machine learning algorithms are recurrently used in medical fields for pathogenicity studies[404, 405] while supervised ones such as Random Forest (RF) and Support Vector Machines (SVM) are used in epitranscriptomic to do *in silico* modified sites predictions[406]. In our case, we wanted to find if these algorithms could be applied *in situ* to RiboMethSeq data.

Prior to manuscript writing, preliminary tests on SVM and RF algorithms have been carried on to choose the right tool for our needs. RF models gave better results while providing some valuable insights on score importance and thus has been chosen for further experimentations.

After defining the different training and testing datasets, multiple RF models have been designed with different sets of features for model construction (adding or removing sequence context and/or auxiliary metrics such as coverage). We found that giving auxiliary information leads to major overfitting of results which should be avoided. The best models provide a good compromise for detection of yet undiscovered 2'-O-methylated sites by reducing the amount of wrong predictions as much as possible while keeping the number of sites detectable as high as possible.

In addition, some models have been designed to detect additional modifications which should not be detected by RiboMethSeq to see if this can improve or reduce models' performance. In fact, while using RF models for guided sites detection, specificity assessment towards multiple modifications can also be tested that way. This approach seems to be an unprecedented application of Random Forests algorithms. Specifying location of other (theoretically) unseen modifications to display potential high-throughput sequencing methods' biases induced by other epitranscriptomics markers could be beneficial to gauge detection methods' robustness. Other deep learning algorithms such as neural networks or unsupervised machine learning algorithms have not been evaluated since Random Forest fits to our different objectives for this work, but this can still be done in the future.

In this work, I took care of the preliminary results to see the potential application of Random Forest algorithms to RiboMethSeq *in situ* data. Then, I created the R pipeline for models creation and assessment and carried on the comparative analysis of each model. Finally, I wrote most of the results alongside the different graphs and the Material and Methods part of this manuscript.

Abstract

Analysis of epitranscriptomic RNA modifications by deep sequencing-based approaches brings an essential contribution to the general knowledge on their precise locations and relative stoichiometry in cellular RNAs. To reveal RNA modifications, several analytical approaches have been proposed, including antibody-driven enrichment, analysis of RT-signatures and specific chemical treatments. However, analysis and interpretation of these massive datasets, especially for low abundant cellular RNAs (e.g. mRNA and lncRNA) is not easy nor straightforward, since the insufficient specificity and selectivity are leading to massive false-positive and false-negative identifications. The main issue in the application of these methods relies on a subjective classification of potentially modified positions, mostly based on arbitrarily defined threshold values for different scores. Such approach using pre-defined scores' values was revealed to be appropriate for limited complexity datasets (for tRNA and/or rRNA analysis), but application to longer reference sequences requires much better classification algorithms. In this work we applied a machine learning algorithm (Random Forest, RF) to create a predictive model for analysis of 2'-O-methylated sites in RNA using RiboMethSeq datasets. Model's training was performed on a large collection of human rRNA datasets with well-known modification profiles and the performance of the prediction was assessed using experimentally defined profiles for other eukaryotic rRNAs (*S. cerevisiae* and *A. thaliana*). Application of this Random Forest prediction model for detection of other RNA modifications and to more complex datasets is discussed.

Introduction

Mapping and precise quantification of epitranscriptomic RNA modifications are essential to better understand their functions in different steps of RNA metabolism. RNA modifications were shown to regulate RNA stability, RNA processing and translation[407–417]. Amongst approaches, capable of detecting RNA modifications, the deep sequencing-based protocols now play a major role[65, 408, 418–421]. Only these techniques do not need purification of totally homogeneous RNA prior to analysis and, with optimized protocols of library preparation, require only a tiny amount of input RNA. Due to these essential features, deep sequencing is now widely used as a gold standard method for detection, mapping and quantification of RNA modified residues[420, 421].

However, such high-throughput analysis of complex reference sequences precludes site-by-site validation of the potential modification candidates, and thus such approaches should have extraordinary specificity and precision, to provide reliable RNA modification profiles in an automatic bioinformatic pipeline[422–426]. Even a relatively low False Discovery Rate (FDR) will generate hundreds or thousands of false positive identifications, even when applied to transcriptomes of 10^6-10^7 nucleotides.

Considering this, application of simple thresholds for detection of modifications in highly abundant and extensively modified RNA (like rRNAs) is still possible but becomes of limited value when analysis should be performed for sparsely (and, in addition, only partially) modified RNA species.

Alkaline hydrolysis-based RiboMethSeq protocol is now extensively used for quantification of RNA modifications (mostly in rRNA and tRNAs)[64, 65, 109, 117, 255, 427], and, in some instances, was also applied to establish *de novo* modification profiles. Such *de novo* mapping was recently performed for *A. thaliana* and *X. laevis* rRNA[428, 429], and for FTSJ3-dependent Nm modifications in HIV-1 RNA[255]. In all these cases, multiple false positive hits were detected, and additional validation by complementary approaches was necessary to establish a reliable Nm modification pattern.

In the current model used for prediction of Nm modification sites in unknown RNA, only two major RiboMethSeq scores[430–432] define classification of a potential candidate site. ScoreMean provides sensitivity, but has somehow a limited specificity, while ScoreA displays the opposite behavior (**Figure 1**). Thus, in practice, the best compromise in detection of modified positions is defined as a combination of two thresholds (e.g. ScoreMean > 0.92 and ScoreA > 0.5), this combination was determined through the previous analysis of ROC curves[430]. In most cases, this default combination of scores ensures efficient detection of highly modified Nm sites (observed MethScore=ScoreC > 0.85) but shows only a limited performance in prediction of partially modified positions (with MethScore < 0.75). Additional RiboMethSeq scores[433, 434] (ScoreB and MethScore) are rarely used for *de novo* mapping since they show only limited discrimination of modified and unmodified positions. Thus, to improve both specificity and sensitivity of Nm mapping from experimental RiboMethSeq datasets[435], more advanced approaches are required, ideally including combination of different scores and other features of the analyzed RNA sequence. In this perspective, the use of machine learning algorithms may fulfill criteria for more accurate approaches.

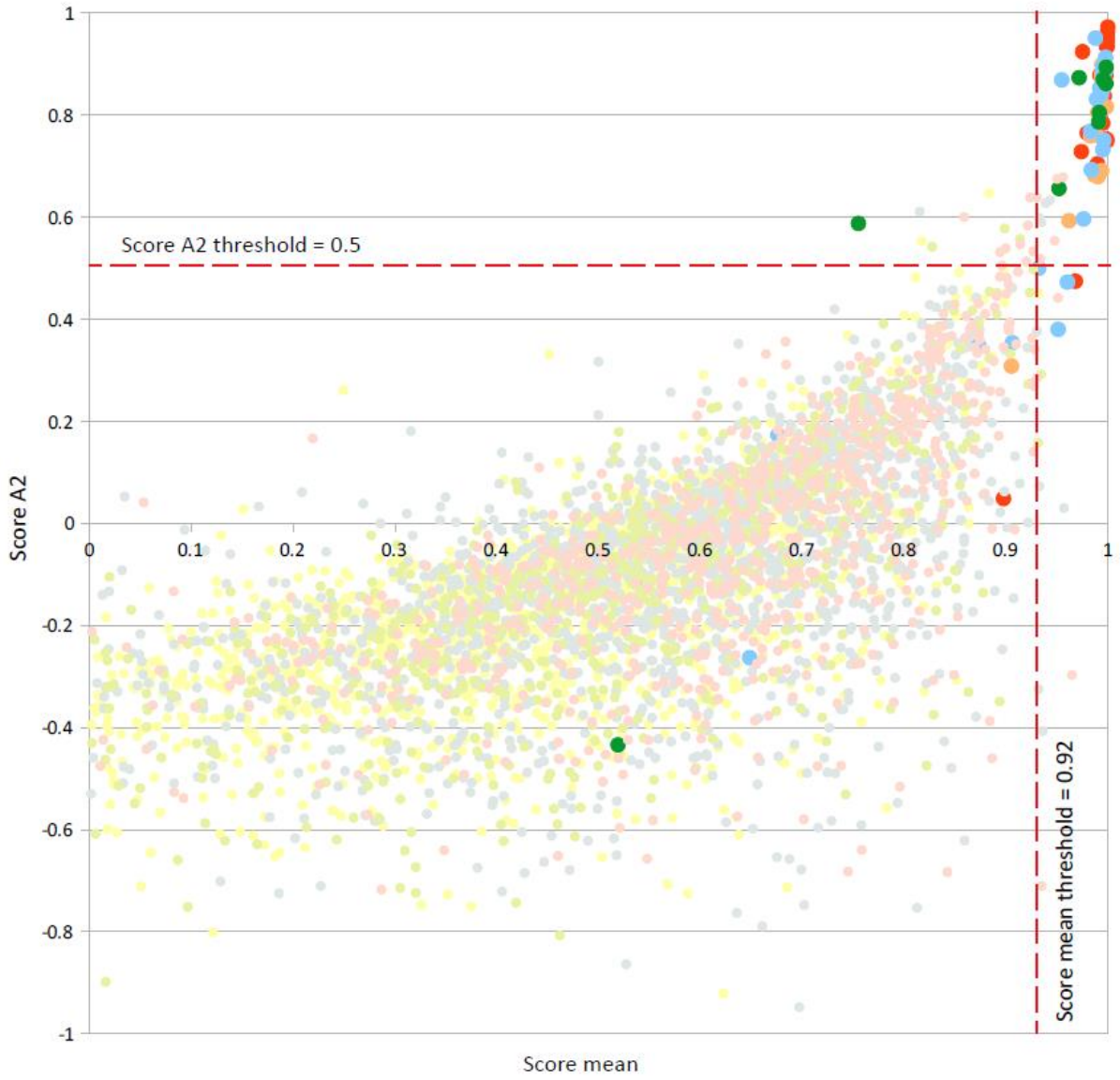


Figure 1 Scatter plot for scores mean and A2 for randomly selected RiboMethSeq sample of human 28S rRNA. Unmodified residues (A, C, G and U) are shown as dots with pastel color, known modified sites are in solid color. Commonly used combination of thresholds for score A2 (> 0.5) and score mean (> 0.92) is shown as dashed red lines. Positions holding exceeding values for both scores are considered to be 2'-O-methylated candidates.

Machine learning algorithms' applications for data treatment are recurrent, especially in biomedical/clinic research fields, due to very high global volume of data to be analyzed. Depending on the goal behind data analysis, multiple machine learning algorithms offering different approaches can be used. Unsupervised learning algorithms have proved to be able to discover new mechanisms of pathology through studious analyses of genomic and transcriptomic data[436–438]. Supervised learning algorithms, like PSI-Sigma or DeepM6ASeq[439, 440], are used in epitranscriptomics for modification detection, on the basis of the sequence context and secondary structure

predictions. Such algorithms are also used in Nanopore sequencing for base calling and modification predictions[441, 442]. Such tools could also be applied to chemically treated or pre-enriched RNAs, like in the case of Bisulfite sequencing, MeRIP-seq or in this work, to RiboMethSeq.
Kernel machines such as support-vector machines (SVM), decision trees with Random Forest (RF) and artificial neural networks seem to be suitable for prediction of modified RNA sites from experimental datasets. Artificial neural networks is the main model used in Nanopore sequencing technology[443–445] due to its capacity to mimic brain behavior for decision making. However, complexity of such models makes model's prediction traceback difficult. SVM allow a clear view on how the predictions are done, and tuning the prediction performances can be done easily, but the kernel trick approach shows weaknesses for noisy data and overlapping classes. RF, based on predictions by multiple decision trees, makes the prediction procedure clear enough to understand how it is achieved, shows better accuracy than most of the other classification algorithms while providing probabilistic estimates through trees classification votes' rate. This tool has already been used for modification predictions by analysis of RT signatures[268, 321]. However, RF is also known to be highly biased if incorrectly configured. Out of these three approaches, RF seems to be the best algorithm for analysis of RiboMethSeq data and prediction of the potentially modified Nm sites.
In this manuscript we describe application of RF prediction model for a comprehensive analysis of RiboMethSeq datasets.

Materials and Methods

RiboMethSeq datasets used in the study

RF models have been created and tested on 62 complete wild-type human rRNA samples (446 214 nucleotide positions) from different cell cultures (Table S1, [435], ENA accession n° PRJEB46620). To reduce biases of sample selection, 80% of the data (356 970 of randomly selected sites) has been used as a training dataset, while the last 20% (89 243 sites) served as an internal testing dataset for initial assessment of models' performance.
RiboMethSeq datasets for yeast *S. cerevisiae* (6 datasets) and plant *A. thaliana* (3 datasets) have been described previously[64, 428] and available in public databases (ENA accession n° PRJEB35565).
Data storage and analyses have been handled by a Unix server Dell R740 under Ubuntu LTS 20.4. Raw data have been trimmed by Trimmomatic version 0.32/0.39 and aligned by Bowtie2 version 2.2.2/2.2.4 as described[64, 428, 435].
All calculations for RF application and model analysis has been done on RStudio Server Version "Juliet Rose" 1.4.1717 with R version "Bird Hippie" 4.1.2. Specific data transformation required R package "reshape2" version 1.4.4, RF algorithm has been

provided by the R package “randomForest” version 4.6-14 and graphics have been produced by R packages “ggplot2” version 3.3.5 and “RColorBrewer” version version 1.1-2.

Results

Brief outline of RiboMethSeq analysis

Analysis of Nm sites in RNA by RiboMethSeq is essentially based on a relative protection against alkaline hydrolysis, which can be provided by different factors, including adjacent 2’-O-methylation. The replacement of 2’-OH by 2’-O-CH_3 very substantially reduces the nucleophilicity of the 2'-oxygen atom in ribose, thus reducing the cleavage of the 3’-adjacent phosphodiester bond in the RNA sequence. This protection is not fully quantitative, but 95-99% reduction of the cleavage rate is generally observed. Partial methylation provides only partial protection, therefore the modification rate can be rather precisely defined[446]. In a randomly cleaved RNA, fragments starting at N+1 position and ending exactly at an Nm residue are substantially under-represented. This under-representation is detected and quantified by deep sequencing, after conversion of the RNA fragments to sequencing library and sequencing in single read (generally SR50) or paired-end mode. Both count profiles for 5’-reads’ extremities at N+1 and 3’-ends at N positions show a characteristic coverage gap at the position of modification. Initial versions of RiboMethSeq analysis mainly used 5’-end count showing typical N+1 nt gap[64, 446], but cumulated 5’/3’-ends’ count provides a better discrimination of modified positions and more precise quantification of the modification rate[430]. During merging of 5’-end and 3’-end counts, the -1 backshift for 5’-ends is introduced, to align the gap exactly to the position of Nm residue in RNA (**Figure 2**).

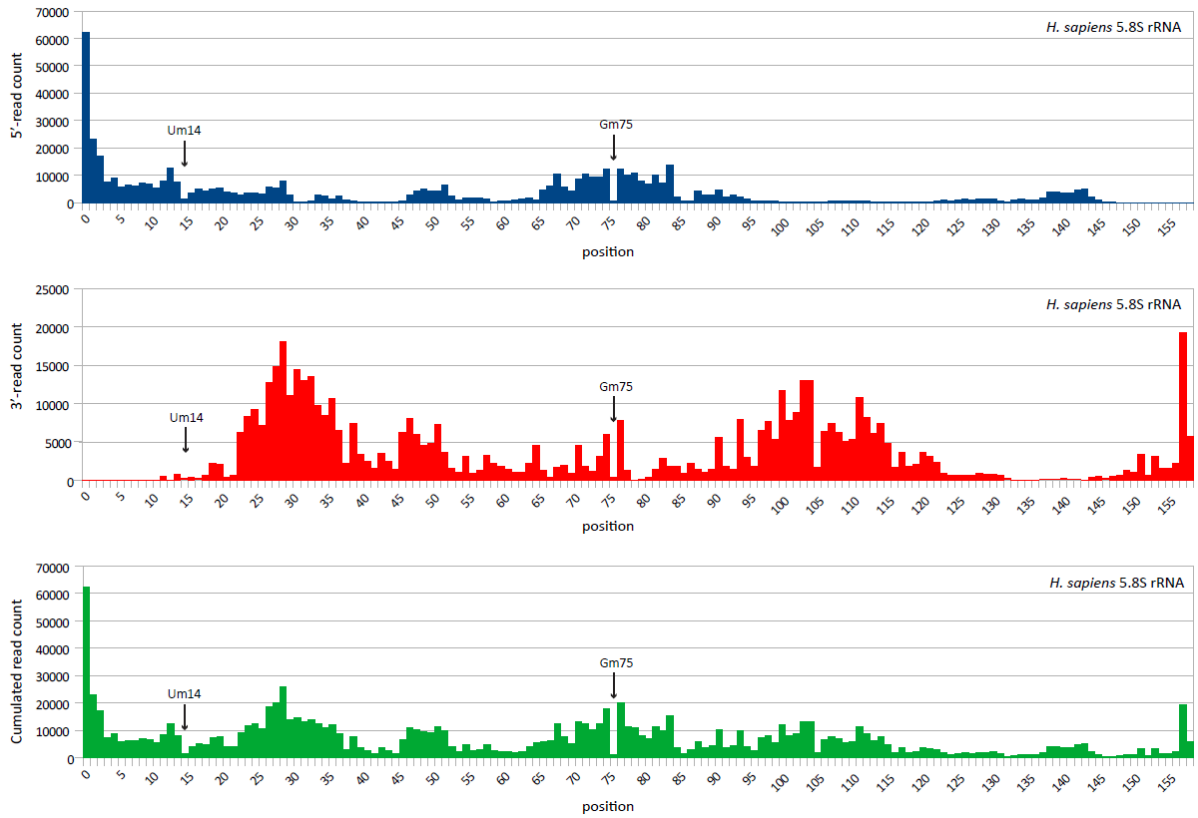


Figure 2 Typical RiboMethSeq profiles for human 5.8S rRNA. Top panel - 5'-end count (blue), middle panel - 3'-end count (red) and merged profile (bottom, green). Positions of two known Nm residues (Um14 and Gm75) in 5.8S rRNA are shown. Numbering is in 0->n-1 format (from *.bed format), so Um14 and Gm75 signals appear at the nucleotide positions N and not at N+1.

Parameters and scores extracted from RiboMethSeq dataset

The depth and other characteristic features of the Nm-related gap in cumulated profile is measured using several RiboMethSeq specific scores. Construction of this scoring system was previously described in details[64, 430, 433] and outlined in **Supp Figure S1**. Since the random alkaline cleavage profile for real RNA is not perfectly regular, some nucleotide positions may show partial protection, related to multiple factors, such as residual RNA structure under alkaline hydrolysis, unequal cleavage at different nucleotides or nucleotide's environments, as well as to biases related to preferential adapter ligation to RNA fragments and PCR amplification during library preparation. At the clustering and sequencing steps, some sequence contexts may be under- or over-represented, and exclusion of ambiguously mapped reads also alters the observed ends' coverage. All these factors collectively contribute to irregular cleavage profiles obtained for different RNA species.

The easiest and the most straightforward parameter indicating a potential Nm residue in RNA sequence might be global protection from cleavage, compared to other neighboring positions. However, in practice, this global (raw) protection is not a

sufficiently robust criteria for precise discrimination between irregular cleavage and the presence of Nm residue.

Considering various factors contributing to irregularities of RNA cleavage, RiboMethSeq scores are designed to include the cleavage profile at the neighboring positions in the value calculated for a given nucleotide. Initial protocol arbitrarily suggested inclusion of positions from -6 to +6, albeit with reduced weight coefficients, varying from 1.0 to 0.5, depending on their relative distance to the Nm site[433]. Later, thoughtful analysis of different calculation intervals demonstrated that the optimal distance, providing the best discrimination of Nm positions and their quantification is +/-2 nt (instead of +/-6 nt used previously)[430]. Thus, all scores used in this work were calculated for +/-2 nt interval with weight coefficients 1 (pos -1/+1) and 0.9 (pos -2/+2). In addition to the classical RiboMethSeq scoring system[430, 433], an additional ScoreAngle was suggested[117]. This score represents only the depth of the gap, without considering other neighboring positions in the profile. Thus, ScoreAngle provides very high sensitivity, but relatively low specificity, since any accidental drop in the coverage profile may be considered as a valid signal. Other RiboMethSeq scores (ScoreA, ScoreB and ScoreC=MethScore) are less sensitive to position-to-position variations, since they consider global variations of the coverage in a given region.

Discrimination of Nm from other modified nucleotides

An additional layer of complexity in the analysis of RiboMethSeq data, especially for eukaryotic rRNA, is related to the false positive (FP) signals raised by other modified nucleotides, mostly pseudouridines (ψ), which are also rather frequent in these RNA species. The exact origin of such non-Nm signals observed in RiboMethSeq remains unclear, it was suggested that they may appear due to adapter ligation biases[64, 433]. Similar behavior was also noticed for m^7G (modification common in tRNAs)[65] and, in part, for $m^6{}_2A$ (a highly conserved double methylation found at the 3'-end of 16S/18S rRNA). The signal for $m^6{}_2A$ may be related to both ligation bias and to strong RT-arrest signal during primer extension. However, other RT-arresting RNA modifications (like m^1A, m^1G, m^3U, $m^1acp^3ψ$, …) were not shown to generate sharp RiboMethSeq signals, but rather progressively reduce global coverage in the region (~20 nt) upstream of such modification. RT-silent modified residues (m^5C, m^2G, ac^4C, …) are not visible in RiboMethSeq protection profiles.

Considering these points, the discrimination model for prediction of Nm sites should include also other potential RNA modifications (at least pseudouridines), especially if prediction is done for eukaryotic rRNA, rich in such modified residues.

Construction of the predictive Random Forest model

In order to include in the Random Forest (RF) model the maximum of information about the RNA sequence as well as on other features of the cleavage profile, the parameters

of different nature were retained for model construction. We considered sequence features (nature of the base included as “base” parameter and identity of -6/+6 surrounding bases), calculated RiboMethSeq scores (in -2/+2 neighbourhood) and score-related values for a given position (ratioL/R, aroundL/R, protection, scores, etc, see **Supp Figure S1** for more details on all these scores and related values), scores calculated for neighboring positions (called prevScores) and auxiliary information to annotate low coverage regions (see **Table 1** for a full list). Three main RF models were created, and their performance evaluated: “Full model” including all possible features, “no_Surrounding_base_info” model retaining all scores but omitting sequence information (except the identity of the modified residue), and “no_Surrounding_info” model where both sequence and scores for neighboring positions were omitted.

Importance of each parameter was evaluated using different RF model trained on a large collection of experimentally obtained RiboMethSeq datasets for human rRNA, and considering possibility to predict Nm and pseudouridines only, as well as other modified nucleotides found in human rRNA (separated in two groups: RT-arresting and RT-silent RNA modifications).

Overall inspection of models’ performance in prediction of Nm and ψ residues, as well as RT-arresting and RT-silent residues (see **Supp Figures S2-S6**) pointed out the importance of both RiboMethSeq scores and sequence information for prediction of modifications. However, for prediction of Nm and ψ residues the nature of the modified nucleotide (base) plays a major role in discrimination (**Supp Figure S2**), followed by values of RiboMethSeq scores, while the sequence context brings only a minor contribution. In contrast, the prediction of RT-arresting and RT-silent nucleotides is mostly based on surrounding sequence features, since these modifications are rare in rRNA and found at conserved sequences in all samples in the dataset.

The performance of different models was evaluated by “precision” and “sensitivity” parameters (**Supp Figure S3**). Precision was calculated as ratio of true positives to total predicted positives, while sensitivity represents the ratio of true positives to total (actual) positives in the data. “Full model” and “no_Surrounding_base_info” models show very similar precision, while sensitivity for low and variably modified positions decreases without surrounding sequence information, since these modifications with low RiboMethSeq scores are mostly predicted in a sequence context, like RT-silent and RT-arresting residues.

The same conclusion can be drawn from inspection of prediction frequency (**Supp Figure S4**); the “Full model” model reliably predicts both highly modified and invariable residues (Am/Cm/Gm/Um/ψm) and also variably modified and low modified/almost unmodified positions (vAm/vCm/vGm/vUm/_Cm/_Um/_Gm). However, when surrounding sequence information is not included (model “no_Surrounding_base_info”), prediction of highly modified positions is only minorly affected, while for others classification into unmodified group substantially increases. This clearly indicates that those variable and low-modified residues (as well as RT-silent and most of RT-arresting modifications) are classified mostly relying on the sequence context, and not according to the RiboMethSeq signals. Such prediction

using the conserved cleavage profile around the modified site as prediction criteria also explains very good results in prediction of Ψm residues in the internal human dataset.

Taking into account these observations, the model excluding sequence related information for the neighboring positions ("no_Surrounding_base_info") shows the best results in prediction of Nm and Ψ residues and thus was investigated in more details. The model excluding all sequence and score-related information on the neighbors of the modified nucleotide ("no_Surrounding_info") showed considerably degraded performance compared the two others.

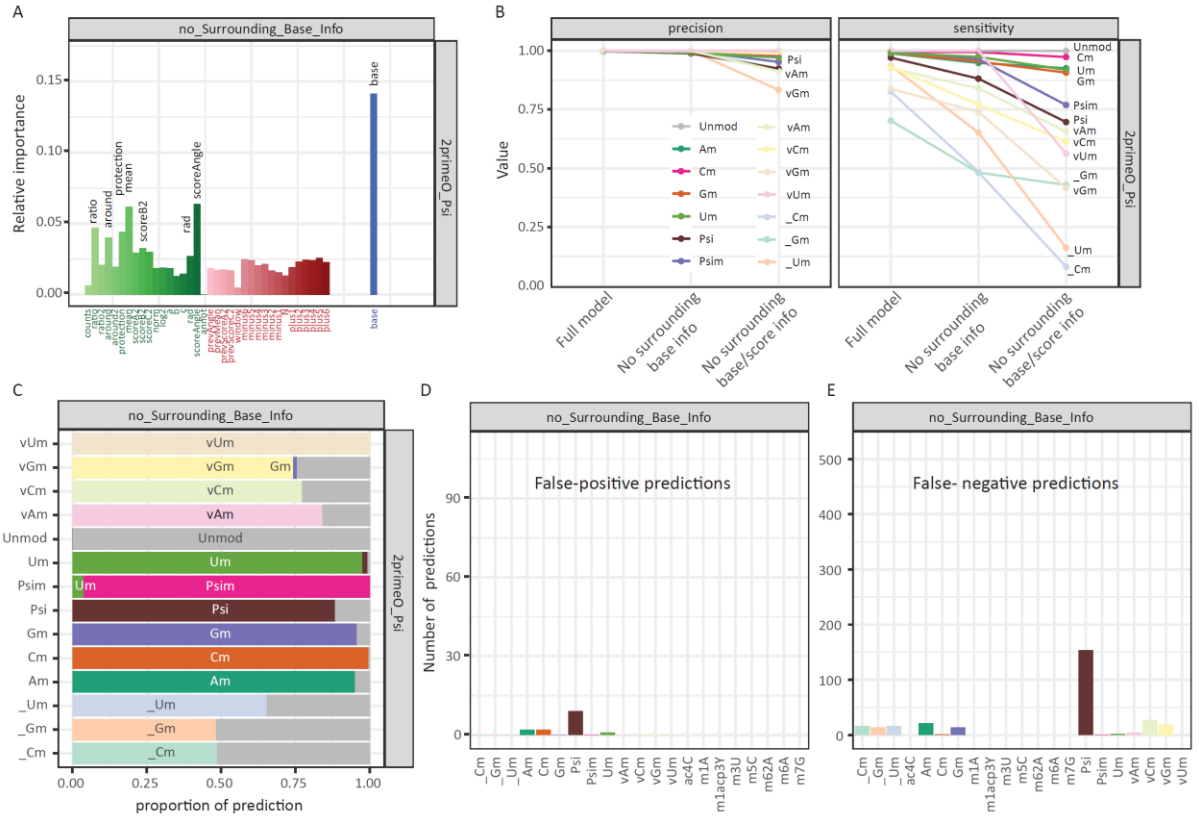


Figure 3 Importance of RF model's parameters and performance of "no_Surrounding_Base_info" model in detection of Nm and pseudouridine (Psi/ψ) residues (other models and their performance are illustrated in **Supp Figures S2-S6**). **Panel A** – Relative importance for each feature has been calculated by normalizing Gini importance (or mean decrease impurity) across all features. **Panel B** shows the influence of the RF model on the precision (left) and sensitivity (right) of Nm and ψ detection. Identity of the models is shown on the bottom (Full model, no surrounding base info, no surrounding info). **Panel C** shows the results of predictions for confirmed and invariable Nm residues (Am, Cm, Gm, Um), variable positions Nm (vAm, vCm, vGm, vUm) and some low modified positions (_Um, _Gm, _Cm). No low modified _Am residues were present in the training dataset. Unmod stands for unmodified nucleotide in the sequence. **Panels D** and **E** show distribution of False-positive (FP) and False-

negative (FN) hits for predictions using internal training dataset (20% of human RiboMethSeq datasets kept aside for validation, see Materials&Methods).

The behavior of the “no_Surrounding_base_info” model is illustrated in **Figure 3**. As mentioned above, the nature of the base, the ratioL/aroundL values as well as protection, ScoreMean and ScoreAngle play major roles in discrimination (**Figure 3A**). Precision of the model is very high, the value is very close to 1.0 and does not differ much from the “Full model” for Nm prediction, while sensitivity remained high for constitutively modified sites (**Figure 3B**). Proportion of correctly predicted Nm and Ψ positions was high for constitutive rRNA modifications and moderate for variably and low modified sites (annotated as vNm and _Nm, respectively) (**Figure 3C**). Analysis of FP hits showed a very low number of incorrectly predicted locations (~10 unmodified positions were predicted as Ψ residues for the whole dataset) **(Figure 3D**). Moreover, FN predictions were also moderate, with <10 Nm which were not predicted, however this proportion is much higher for Ψ residues (**Figure 3E**). Overall, considering the dataset used for validation (20% of total dataset, so ~ 13 samples), the “no_Surrounding_base_info” model showed almost exceptional performance in the Nm detection and discrimination from Ψ residues.

Validation of the predictive RF model using S. cerevisiae and A. thaliana RiboMethSeq datasets

Since the use of identical sequences (here *H. sapiens* rRNA datasets) for both training and performance measurements is highly biased by predictions based on “known sequence” and “known profile”, the performance of the RF models have been also assessed using additional RiboMethSeq datasets, for relatively distant rRNAs from *S. cerevisiae* and from *A. thaliana*.

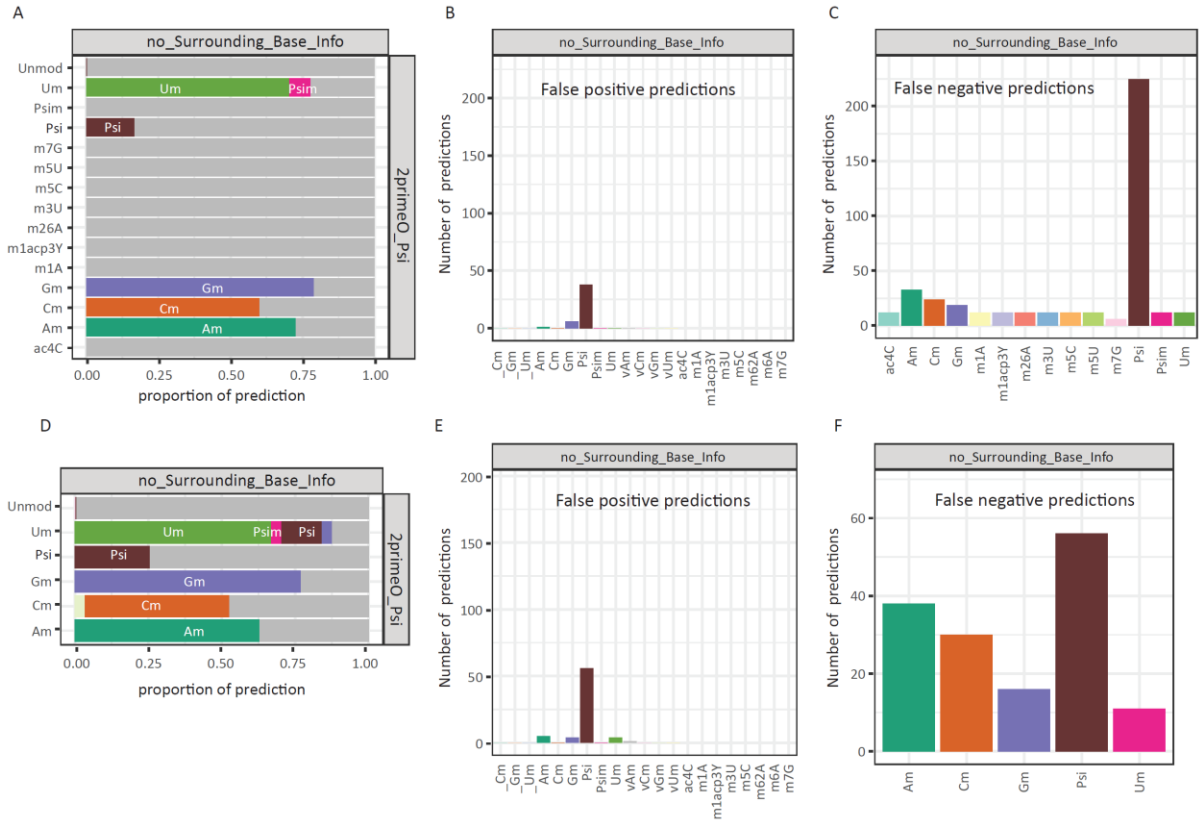


Figure 4 Validation of RF models using RiboMethSeq datasets for *S. cerevisiae* and *A. thaliana* rRNAs. Only Nm and Ψ residues were reliably mapped in *A. thaliana* rRNAs and this limited modification profile was used as a reference. Results of prediction for Nm, ψ and other RNA modifications are shown for "no_Surrounding_Base_info" model applied for *S. cerevisiae* (**A**) and *A. thaliana* (**D**) RiboMethSeq datasets. Identity of prediction is shown by color. **Panels B** and **F** show FP hits for *S. cerevisiae* (**B**) and *A. thaliana* (**E**) datasets, while **panels C** and **F** list FN predictions for both.

Figure 4 demonstrates that analysis of *S. cerevisiae* and *A. thaliana* RiboMethSeq datasets gives very similar results for "no_Surrounding_Base_info" RF model, even with relatively different rRNA sequences. Substantial proportion of Nm residues was correctly predicted (**Figure 4A-D**), moreover, the FP/FN predictions remain very limited (**Figure 4B-F**), again as in the case of internal human dataset, most FP/FN hits are ψ, but not Nm residues.

Overall, Nm sites in both datasets were predicted with a high precision albeit with somehow reduced sensitivity, which is also related to incomplete modification at some *S. cerevisiae* and *A. thaliana* rRNA sites.

More detailed analysis of performance and predictions obtained using other RF models (**Supp Figures S2-S6**) demonstrated that the results obtained with "Full model" are rather similar to "no_Surrounding_Base_info" model, except detection of almost universal modification 18S-m$^6{}_2$A, which is present in highly conserved

sequence context and thus efficiently detected by “Full model” in yeast rRNA. Other RT-arresting or RT-silent modifications were not detected, since no RiboMethSeq signal was generated for those residues. Overall performance of the selected RF model is much better than combination of scores’ thresholds (**Supp Figure S7**).

Performance of reduced RF model for Nm and Ψ prediction

Since the importance of different RF models’ features is not equal, to reduce computational efforts and complexity of the scripts, we decided to test the possibility to use the reduced RF prediction model including only 8 most important parameters, namely nature of nucleotide (base), and different RiboMethSeq scores (ratioL, ratioR, around L, aroundR, protection, ScoreMean and ScoreAngle, see **Table 1** and **Figure 5A**). Performance of this reduced model was evaluated as three others, using first internal human dataset and validation datasets from *S. cerevisiae* and *A. thaliana*. As shown in **Figure 5B**, the proportion of correctly predicted human rRNA modifications remained correct, even if overall prediction was less precise compared to “no_Surrounding_Base_info” model. However, the validation with *S. cerevisiae* and *A. thaliana* RiboMethSeq data pointed out a very poor performance of the reduced model for these different eukaryotic species. Only Um and Am residues were correctly predicted (**Figure 5C**) and even if FP identifications remained limited (**Figure 5D**), the number of FN predictions becomes very high. Precision of the reduced model was still acceptable, but the sensitivity was clearly degraded (**Supp Figure S3-S4**). This model has degraded performance and is not suitable for sensitive prediction of Nm and ψ residues.

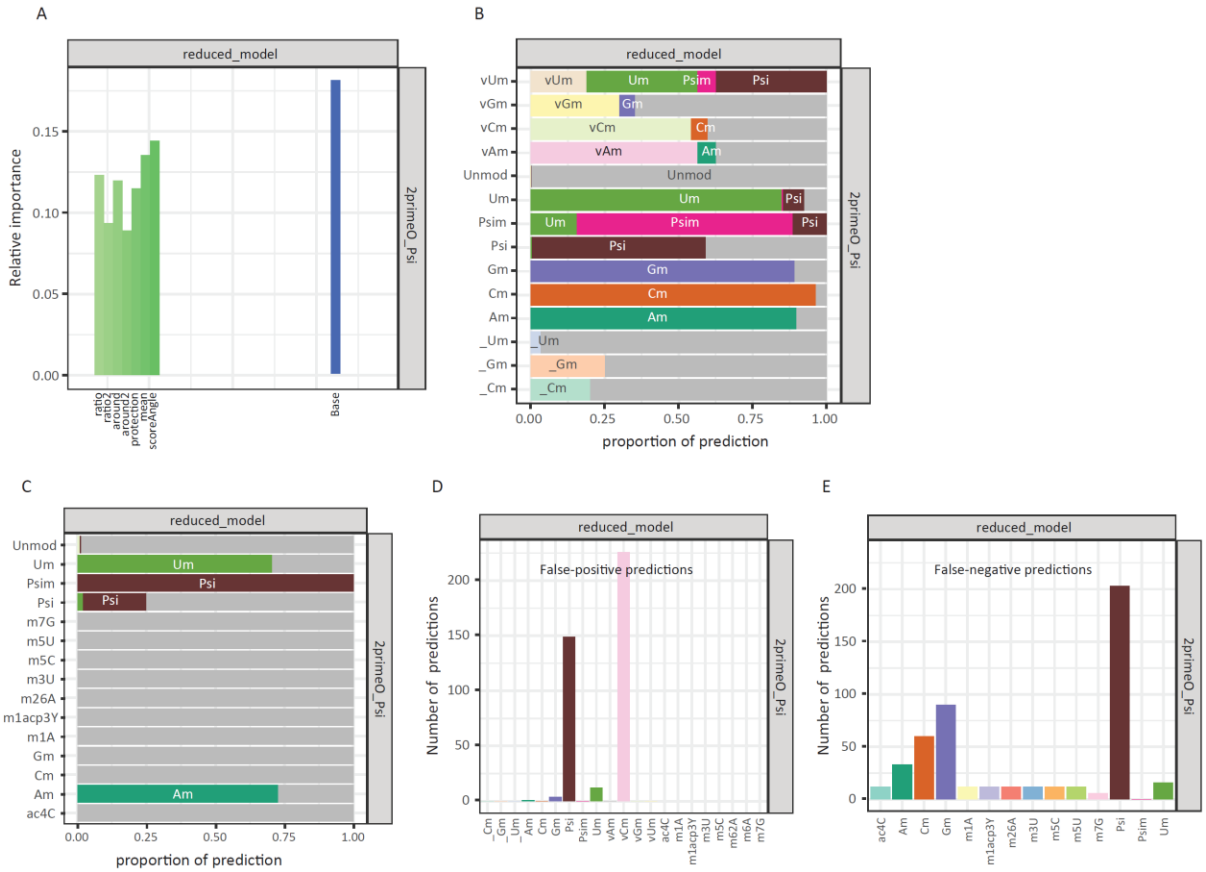


Figure 5 Performance of the reduced RF model for prediction of Nm and ψ residues. Relative importance of different features included in the reduced model (**A**). Prediction of human rRNA modifications using the internal dataset (**B**). Prediction of yeast S. cerevisiae rRNA modifications (**C**). Number and distribution of FP (**D**) and FN (**E**) hits. Notice that Cm and Gm residues are not at all predicted to be modified by the reduced RF model.

Discussion

The choice of the predictive model

In this study (overall design is outlined in **Figure 6**) we compared behavior and performance of four RF models ("Full model", "no_Surrounding_base_info", "no_Surrounding_info" and "Reduced model"), differing in the number and nature of parameters taken for the training and prediction. Performance of these models was evaluated for prediction of all possible modifications, for the target group (Nm and ψ residues), as well as for detection of RT-arresting/RT-silent RNA modifications. As anticipated, the "Full model" demonstrates a better performance for internal dataset of the same nature and for RNA modifications appearing in a highly conserved sequence. However, in case of external validation datasets (*S. cerevisiae* and *A. thaliana*), the best performance is obtained for "no_Surrounding_base_info" model, which does not use sequence consensus for prediction, but essentially based on the values of RiboMethSeq scores in the considered region. This model is well adapted for prediction of Nm and ψ residues, but, as anticipated, is not suitable for other RNA modifications, since the latter do not generate any distinct RiboMethSeq signal. Even if some non-Nm residues (e.g. $m^6{}_2A$) generate RT-stop and sudden drop in 5'-coverage, other approaches and additional features (like mutational RT signatures) have to be used in such cases of RT-arresting RNA modifications[268, 316, 321]. The attempt to reduce the number of RF model's parameters to the minimum ("Reduced model") clearly demonstrated that this approach does not provide a sufficient sensitivity in detection of true modified sites. This is likely related to the cumulative character of the prediction, every parameter brings incremental, but still important contribution into final decision, and thus reduction of the model leads to a substantial decrease of the performance.

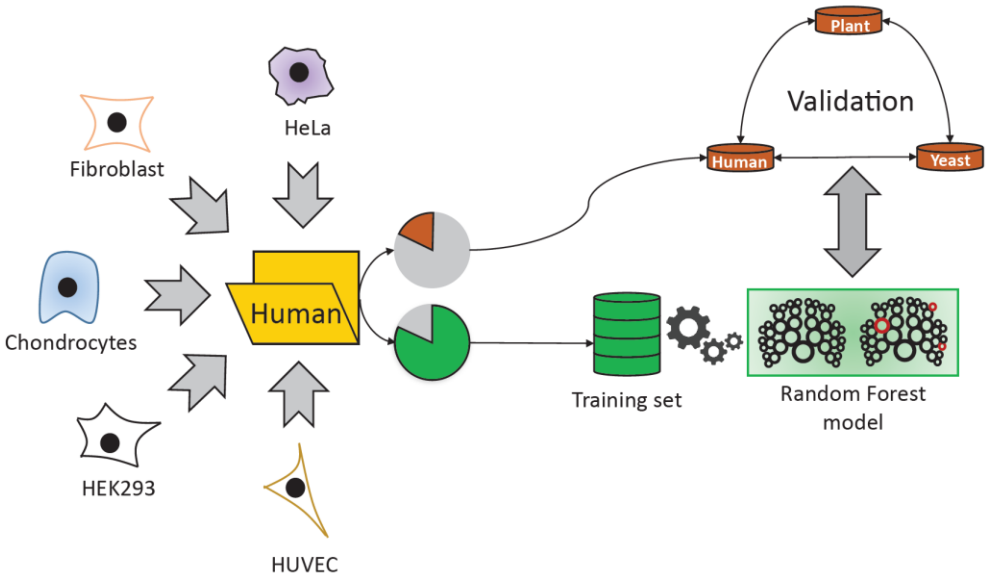


Figure 6. Outline of the RF model construction and validation. RiboMethSeq datasets from various human cell lines have been used as training and internal validation. To

reduce sample selection biases, 80% of the data (356 970 of sites randomly selected from 62 samples) has been used as a training dataset, while the last 20% (89 243 sites) served as an internal testing dataset for initial assessment of models' performance. Further validation of the model and evaluation of its performance was done using yeast *S. cerevisiae* and plant *A. thaliana* RiboMethSeq datasets.

In order to reduce the number of FP and FN hits, the models used in this work have been trained on a validated collection of constitutively modified human 2'-O-methylated sites. We introduced the specific annotations for variably modified (vNm) or low modified (_Nm) positions. This annotation scheme showed to be efficient, since the models' precision is relatively high compared to preliminary models (not shown), where simple annotation even variable or low modified sites as Nm was used. However, RF is known to have multiple biases[447] on variable importance and/or classification and such biases can be anticipated to influence the results of sequencing detection methods. These biases may be related to variations between sequencing/analysis runs (for example, due to different behavior of the Reverse Transcriptase, used library preparation kit and parameters for reads alignment/trimming). However, these sources of variability were not explored in this study, since all samples were prepared using the same library preparation kit with the same RT enzyme.
In conclusion, considering degraded performance of other tested RF models, we recommend using "no_Surrounding_base_info" model for training and prediction of Nm and ψ residues.
In contrast to purely computational machine learning methods essentially based on primary sequencing data(like NmSEER V2.0[448], NmRF[449], or DeepOMe[450]), the approach proposed in this work is intended to be applied to deep analysis experimental RiboMethSeq datasets, and thus demonstrates superior performance in the absence of directly entered sequencing information.

Precision and sensitivity of RF model

Type I (FP hits) and type II (FN hits) errors are the main parameters assessed for predictive model performance. Although model optimization tends to minimize both error types, applications of such predictors are not really suitable for biological projects, since even relatively low False Discovery Rate (FDR) can lead to tremendous numbers of FP hits, since certain modifications are rare or even very rare in RNA. This is particularly true for mRNA analysis where overlap in modification mapping between different working groups varies from 30% to 60% only, even when the same detection method was used[423]. This is the reason to prioritize minimization of type I errors over type II in mapping of potential modified sites in unknown RNA.
Analysis of the essential performance parameters (mainly precision and sensitivity) demonstrates that all tested models show very low FP rate, albeit at the expense of some loss in sensitivity, and thus an increasing number of FN hits. Since the main

challenge in the analysis of epitranscriptomics data, especially for large transcriptomics datasets of millions of nucleotides, is the highest precision/low FP rate (ratio of true positives to total predicted positives) some limited loss of sensitivity is quite acceptable. The observed precision close to 1 (0.99695282 for prediction of Nm and ψ) is largely sufficient for rRNA analysis (~10,000 nt) and similarly sized RNA species but is still inappropriate for very large and sparsely modified transcriptomics data, since the proportion of FP hits may be still very high. Combination of different approaches (e.g. RiboMethSeq and Nm-Seq at low [dNTP][67, 428]) may bring an acceptable solution to this problem.

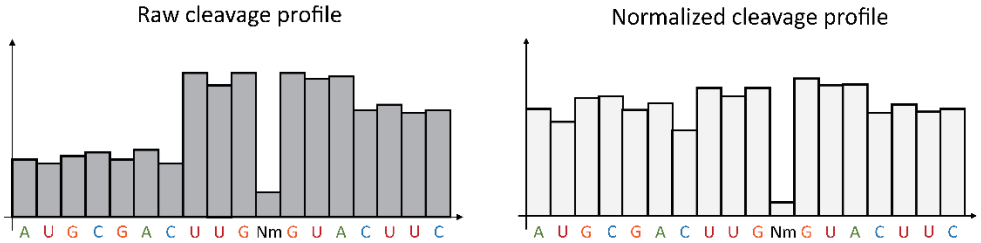
Analysis of the observed False Positive (FP) and False Negative (FN) hits

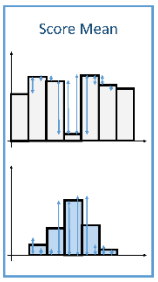
Inspection of FP and FN hits obtained for the best-performing model "no_Surrounding_base_info", clearly shows that most frequent FP hits are predictions of unmodified nucleotide as ψ, followed by some low frequency of FP Gm/Am/Um predictions. This is anticipated from the way of model's construction and is an assumed attempt to make a distinction between Nm and potential ψ residue. In practice, the main goal of RiboMethSeq analysis is prediction of Nm candidates and if FP ψ predictions are sorted out for further validation by HydraPsiSeq[361, 451], the total number of FP hits remains very low. Similarly, the most frequent FN hits also correspond to non-prediction of ψ, since many of those residues give only weak RiboMethSeq signal, undistinguishable from background. These minor limitations of the proposed RF predictive model must be considered in practical applications. Alternative methods for ψ mapping, like HydraPsiSeq[361] or CMCT-based protocols[314, 320] are better adapted to the precise mapping of those modified residues.
Applications to other RNA modifications or different mapping protocols
The approach described in this work may be further extended for analysis of other RNA modification generating any distinguishable signal in RiboMethSeq or other appropriate deep sequencing-based protocol for RNA analysis, like HydraPsiSeq[361], AlkAnilineSeq[360] or current BisulfiteSeq protocols[452]. With some minor adaptations, the RF model could be also applied to other sequencing detection methods such as MeRIP-seq, and to Nanopore sequencing data. Analysis of observed False Positive and False Negative hits for these two last methods would assess their detection robustness against different epitranscriptomics markers. The main difficulty remains the availability of appropriate and validated experimental datasets required for model's training and evaluation of performance.

Acknowledgements

This work was supported by ANR PRC grant "MethRibo" and Epi-ARN2018 and ViroMOD2020 FRCR grants from Grand Est Region, France to YM and Deutsche Forschungsgemeinschaft (DFG) grants to M.H. [HE3397/17-1, SPP1784, and TRR319 RMaP, TP C01].

Supplementary material





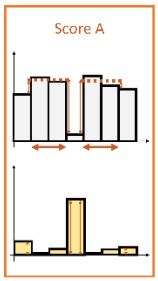
ratioL = Ratio of reads at n+1 to reads at n

ratioR = Ratio of reads at n to reads at n-1

aroundL = 1- ratioLeft(n)/(mean(n-2;n-1)+mean(n+1:n+2))

aroundR = 1- ratioRight(n)/(mean(n-2;n-1)+mean(n+1:n+2))

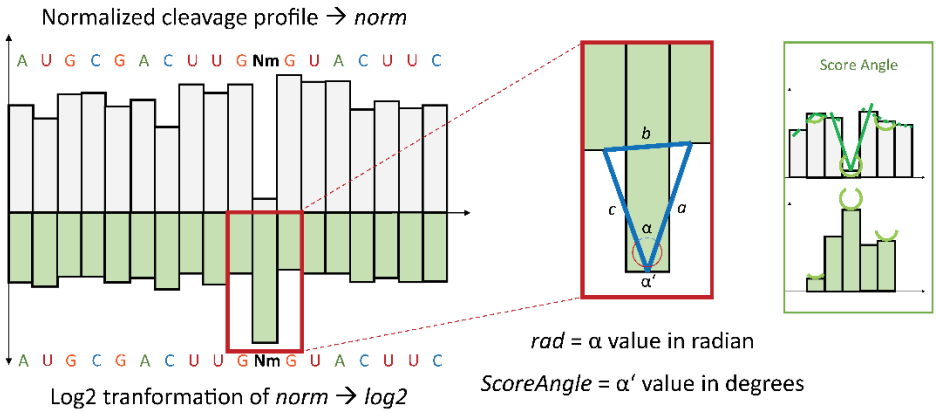
ScoreMean = mean of aroundLeft(n-1) and aroundRight(n+1)



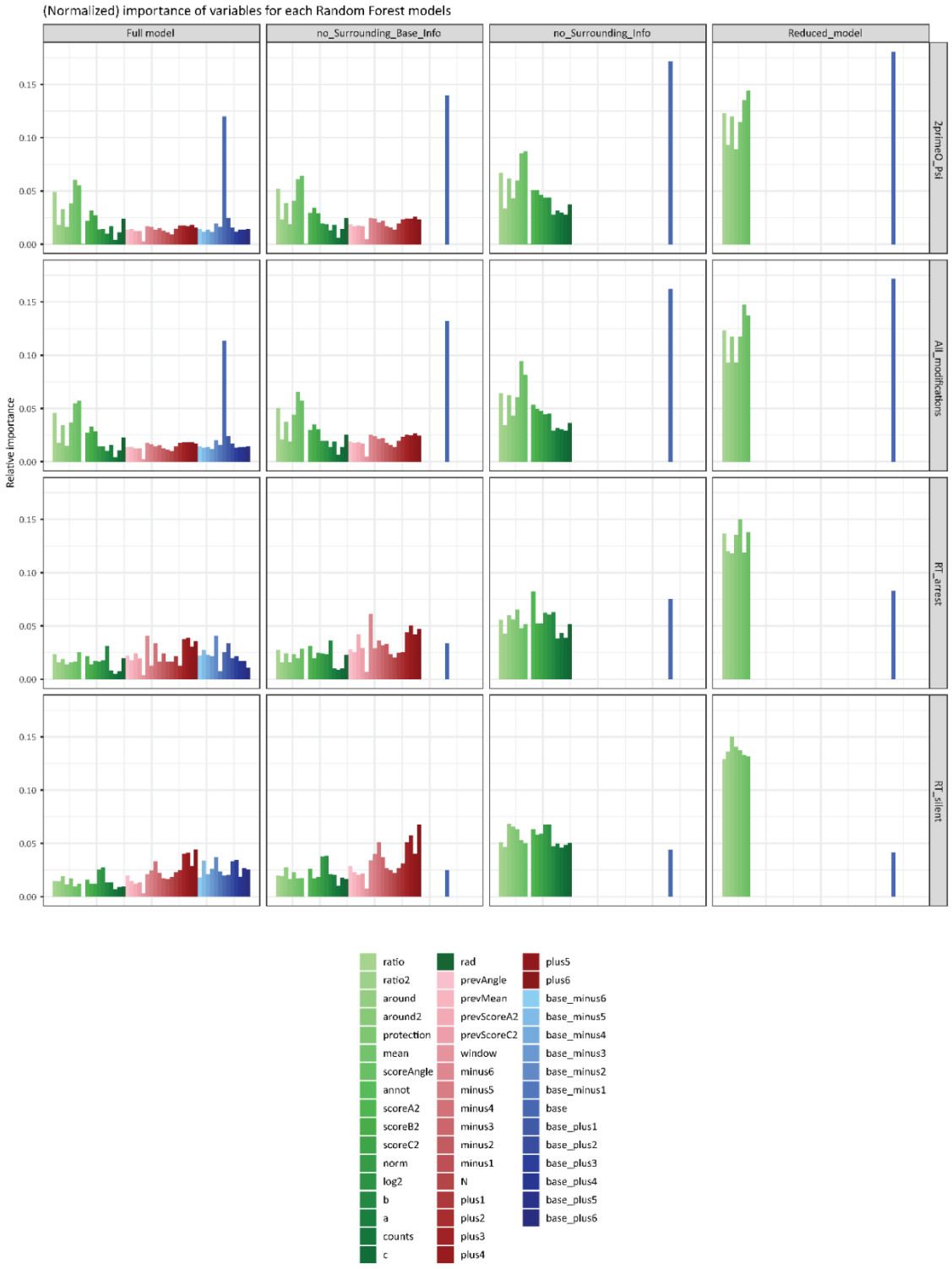
$$Score\ A2 = 1 - \frac{2n_i + 1}{\frac{1}{2}|\mu_l - \sigma_l| + n_i + \frac{1}{2}|\mu_r - \sigma_r| + 1}$$

$$Score\ B2 = \frac{n_i - \frac{1}{2}\left(\frac{\sum_{j=i-2}^{i-1}\omega_j n_j}{\sum_{j=i-2}^{i-1}\omega_j} + \frac{\sum_{j=i+1}^{i+2}\omega_j n_j}{\sum_{j=i+1}^{i-2}\omega_j}\right)}{n_i + 1}$$

$$Score\ C2 = \frac{n_i}{\frac{1}{2}\left(\frac{\sum_{j=i-2}^{i-1}\omega_j n_j}{\sum_{j=i-2}^{i-1}\omega_j} + \frac{\sum_{j=i+1}^{i+2}\omega_j n_j}{\sum_{j=i+1}^{i-2}\omega_j}\right)}$$

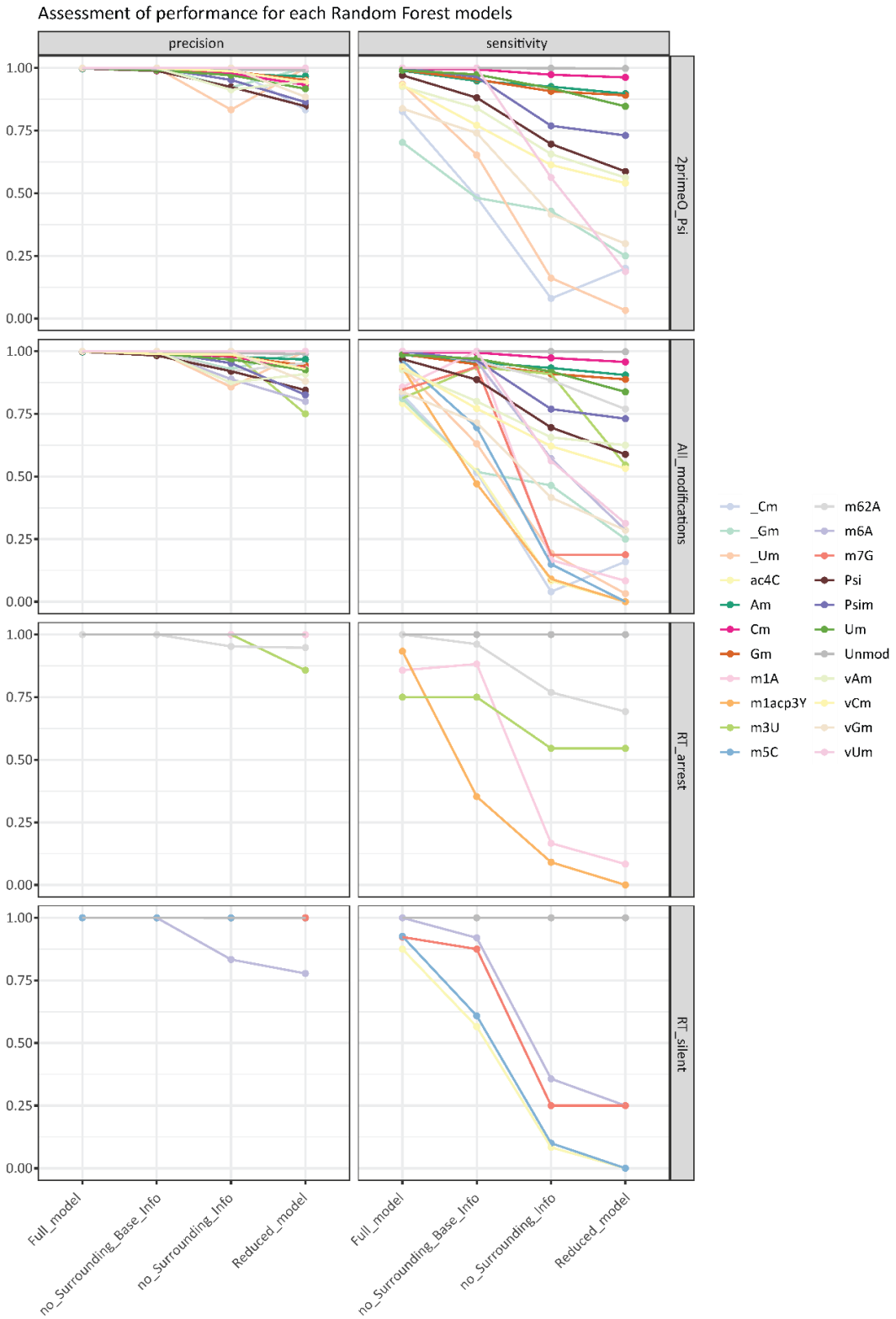


Supplementary Figure S1 Illustration and calculation formula for RiboMethSeq scores. RiboMethSeq scores are calculated either from raw cleavage profiles or from normalized cleavage profiles, obtained from raw cleavage by normalization to local cleavage efficiency in a rolling window of 11 nt. ScoreMean is calculated from RatioL/R and AroundL/R values, as indicated. Formula for ScoreA, ScoreB and ScoreC=MethScore are given. ScoreAngle and distances a, b and c are calculated from normalized and log2-transformed profile.



Supplementary Figure S2 Relative importance of RF model parameters.

For each feature, importance was calculated as Gini importance (or mean decrease impurity) using the average for all RF decision trees. The color code indicates the Score-related values (green), scores and profile for neighboring positions (rose) and sequence around of the analyzed position (blue). Identity of the model is indicated on the top of each column, the training set of modifications on the right.

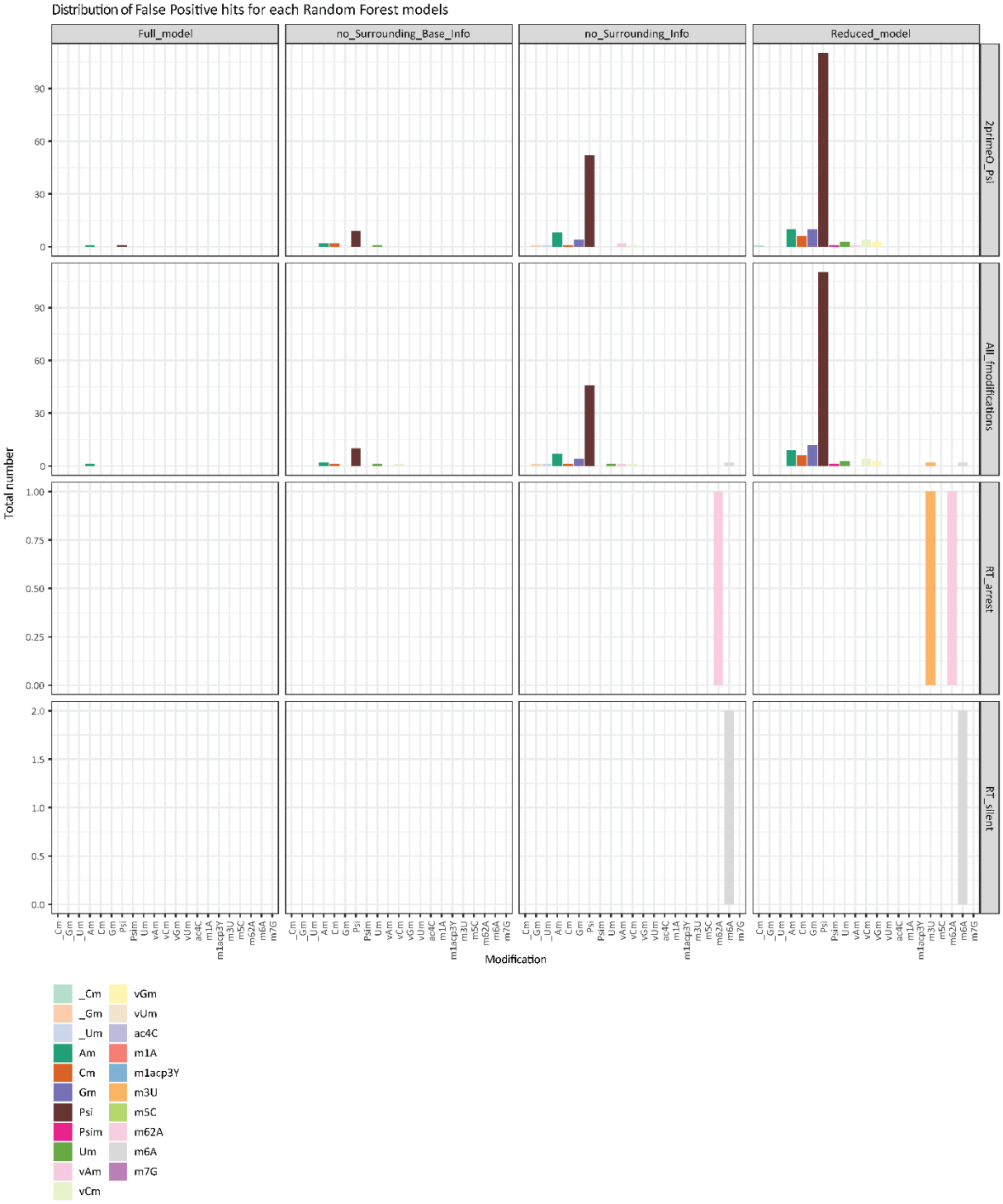


Supplementary Figure S3 Precision (left) and sensitivity (right) of each tested RF model. Identity of the model is indicated at the bottom, the training set of modifications on the right. Conventional abbreviations are used for all modified nucleotides, except hypermodified $m^1acp^3\psi$ which is shown as m1acp3Y. vNm and _Nm designate variable and low modified positions, respectively.



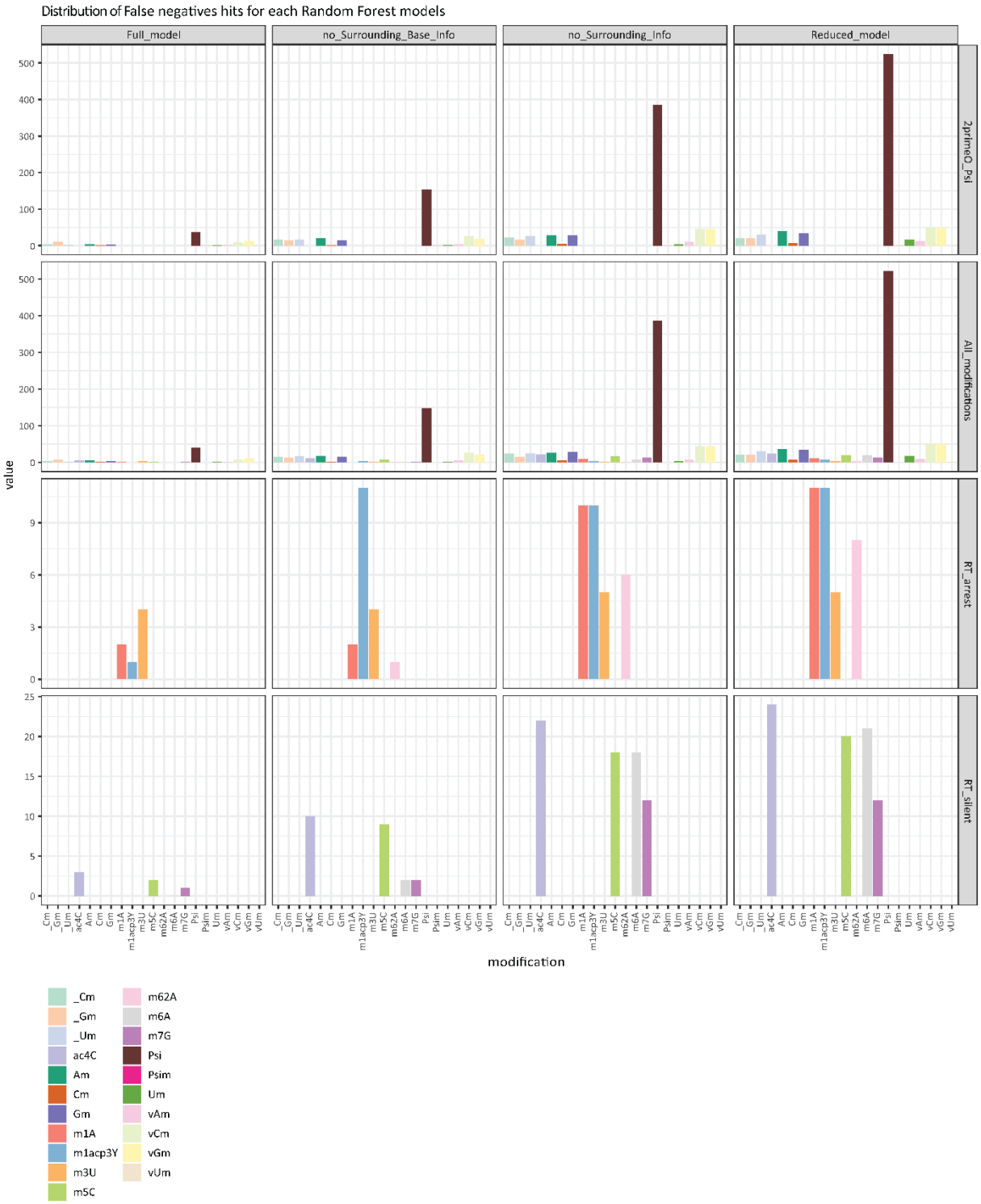
Supplementary Figure S4 Predictions of modified nucleotides in the internal human dataset by each tested RF model.

Real identity of the modification is given on the left, horizontal barplot corresponds to predictions given for each model. Color code for predictions is given at the bottom. Identity of the model is indicated on the top of each column, the training set of modifications on the right. Conventional abbreviations are used for all modified nucleotides, except hypermodified $m^1acp^3\psi$ which is shown as m1acp3Y. vNm and _Nm designate variable and low modified positions, respectively.



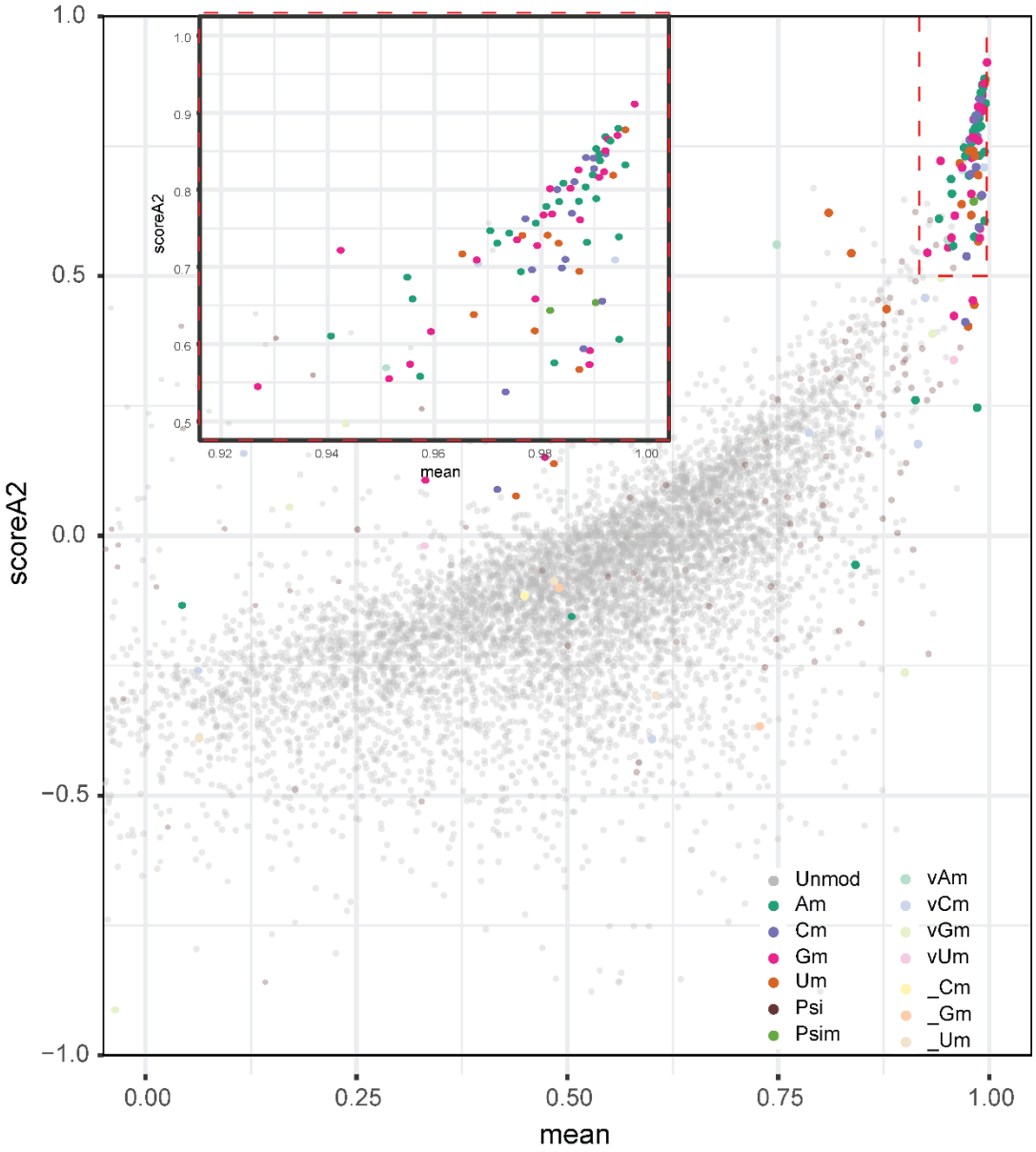
Supplementary Figure S5 Distribution of False Positive hits for the internal human dataset by each tested RF model.

Color code for predictions is given at the bottom. Identity of the model is indicated on the top of each column, the training set of modifications on the right. Conventional abbreviations are used for all modified nucleotides, except hypermodified $m^1acp^3\psi$ which is shown as m1acp3Y. vNm and _Nm designate variable and low modified positions, respectively.



Supplementary Figure S6 Distribution of False Negative hits for the internal human dataset by each tested RF model.

Color code for predictions is given at the bottom. Identity of the model is indicated on the top of each column, the training set of modifications on the right. Conventional abbreviations are used for all modified nucleotides, except hypermodified $m^1acp^3\psi$ which is shown as m1acp3Y. vNm and _Nm designate variable and low modified positions, respectively.



Supplementary Figure S7 Prediction by RF model outperforms score-based prediction. Predictions for the same human rRNA dataset was done by combination of scores thresholds (ScoreMean >0.92 AND ScoreA2 >0.5) and by RF model. Both numbers of FP hits, grey dots inside of dash squared region, and FN hits, colored dots predicted as modified positions outside of the dash squared region, are decreased.

Table S1 List of highly confident rRNA modification sites detected in human *H. sapiens* rRNA

Median values for Scores and Votes for internal *H. sapiens* validation datasets are shown, rRNA positions numbering is according to NR_046235.

rRNA and position	modification	ScoreMean	scoreA2	scoreAngle	Max Vote Mod	Predicted as
hs_rRNA_28S_2363	Am	0,77426759	0,60429956	202,957093	0,928	vAm
hs_rRNA_28S_4498	Um	0,87446365	0,66693227	228,569903	0,9245	Um
hs_rRNA_18S_119	Psi	0,84659487	-0,04323755	217,681019	0,9095	Psi
hs_rRNA_28S_1536	Psi	0,5846821	-0,49015679	172,527345	0,9075	Psi
hs_rRNA_28S_4227	Um	0,88032965	0,60134173	233,342609	0,8725	Um
hs_rRNA_28S_1326	Am	0,9474001	0,49240366	285,96953	0,839	Am
hs_rRNA_28S_2448	_Gm	0,72160408	-0,2381226	179,683622	0,828	_Gm
hs_rRNA_18S_590	Am	0,96828788	0,73650713	317,582541	0,821	Am
hs_rRNA_18S_1678	Am	0,96947601	0,74083355	316,982829	0,81	Am
hs_rRNA_18S_99	Am	0,99137502	0,77326645	328,566425	0,8085	Am
hs_rRNA_28S_3762	Psi	0,6302388	-0,36295912	216,804231	0,801	Psi
hs_rRNA_28S_2815	Am	0,98515977	0,71621255	325,200863	0,7925	Am
hs_rRNA_28S_4493	Psi	0,28250547	0,30102856	178,439615	0,7905	Psi
hs_rRNA_18S_468	Am	0,96939152	0,67823368	317,766997	0,7855	Am
hs_rRNA_28S_4523	Am	0,99338862	0,53823786	330,812392	0,784	Am
hs_rRNA_28S_4571	Am	0,95642391	0,69504152	309,83561	0,784	Am
hs_rRNA_28S_2787	Am	0,96715943	0,68096733	317,559372	0,7795	Am
hs_rRNA_28S_4636	Psi	0,41266317	0,07527384	173,079322	0,777	Psi
hs_rRNA_28S_2364	Gm	0,95734933	0,46349172	296,997659	0,7765	Gm
hs_rRNA_28S_2839	Psi	0,85670572	0,05180749	274,222816	0,774	Psi
hs_rRNA_28S_3758	Psi	0,71742029	-0,09296818	222,080667	0,773	Psi
hs_rRNA_18S_27	Am	0,9815479	0,71403327	324,567673	0,761	Am
hs_rRNA_18S_681	Psi	0,74294146	-0,17739863	209,514778	0,76	Psi
hs_rRNA_18S_484	Am	0,97783125	0,76198699	322,875016	0,756	Am
hs_rRNA_28S_4590	Am	0,96285046	0,65934875	316,417037	0,7535	Am
hs_rRNA_28S_1871	Am	0,98550278	0,81766868	326,415983	0,752	Am
hs_rRNA_28S_3785	Am	0,97883554	0,76433898	323,472914	0,752	Am
hs_rRNA_18S_1238	Psi	0,93637025	0,41711425	302,352116	0,748	Psi
hs_rRNA_28S_4532	Psi	0,49881569	-0,03420037	179,276971	0,7475	Psi
hs_rRNA_18S_512	Am	0,97196831	0,64880584	316,926526	0,7455	Am
hs_rRNA_28S_3844	Psi	0,92045894	0,22032367	280,848069	0,7445	Psi
hs_rRNA_18S_1643	Psi	0,95834893	0,4813325	315,107307	0,742	Psi
hs_rRNA_18S_668	Am	0,98766709	0,80785978	327,924086	0,741	Am
hs_rRNA_28S_3851	Psi	0,68045695	-0,08303826	228,681029	0,74	Psi
hs_rRNA_18S_159	Am	0,98970244	0,84819438	329,0629	0,7335	Am
hs_rRNA_18S_866	Psi	0,90675454	0,39158324	290,064479	0,7305	Psi

hs_rRNA_28S_2843	Psi	0,48974807	-0,25737895	178,383063	0,726	Psi
hs_rRNA_28S_1323	Am	0,83191004	-0,0896389	215,479076	0,725	Am
hs_rRNA_28S_3867	Am	0,8753757	0,22165072	290,687986	0,72	Am
hs_rRNA_18S_1232	Psi	0,88793542	0,34777272	287,975304	0,719	Psi
hs_rRNA_18S_1328	Gm	0,98419103	0,47547686	323,321669	0,713	Gm
hs_rRNA_18S_814	Psi	0,87982692	0,28369308	285,951401	0,7125	Psi
hs_rRNA_28S_4403	Psi	0,89966671	0,23300878	294,70042	0,703	Psi
hs_rRNA_18S_572	Psi	0,96795606	0,36033503	316,558177	0,6975	Psi
hs_rRNA_28S_1862	Psi	0,62713406	-0,0182754	207,198783	0,6925	Psi
hs_rRNA_18S_897	Psi	0,58305637	-0,08029769	181,382975	0,692	Psi
hs_rRNA_5,8S_55	Psi	0,77366806	0,02158327	250,123678	0,6885	Psi
hs_rRNA_28S_4521	Psi	0,92797498	-0,18083114	283,215801	0,6855	Psi
hs_rRNA_18S_121	Um	0,96961521	0,73483752	320,255146	0,6815	Um
hs_rRNA_28S_3920	Psi	0,83659082	0,11927031	268,773513	0,6805	Psi
hs_rRNA_28S_398	Am	0,98367333	0,20696552	323,015702	0,68	Am
hs_rRNA_28S_3637	Psi	0,75735035	-0,27005857	252,163754	0,679	Psi
hs_rRNA_18S_1490	Gm	0,97164647	0,67505504	318,874292	0,6765	Gm
hs_rRNA_28S_4689	Psi	0,85171726	-0,28386707	274,832364	0,6765	Psi
hs_rRNA_28S_3830	Am	0,98298189	0,74827348	325,666996	0,676	Am
hs_rRNA_18S_36	Psi	0,85429191	-0,09679698	278,09075	0,673	Psi
hs_rRNA_18S_649	Psi	0,76881982	-0,03602094	259,460996	0,673	Psi
hs_rRNA_28S_4312	Psi	0,52363763	-0,06896687	178,680796	0,668	Psi
hs_rRNA_28S_400	Am	0,98833685	0,73530904	326,292593	0,6665	Am
hs_rRNA_5,8S_14	Um	0,8839982	0,45151743	287,267057	0,666	Um
hs_rRNA_28S_2365	Cm	0,42264141	0,57071252	184,357769	0,6635	vCm
hs_rRNA_18S_651	Psi	0,74021714	0,02821044	251,92138	0,6635	Psi
hs_rRNA_18S_1445	Psi	0,86876409	0,05462921	262,879892	0,658	Psi
hs_rRNA_18S_1004	Psi	0,6327213	-0,06662794	225,415025	0,657	Psi
hs_rRNA_28S_4618	Gm	0,88058416	0,18146527	291,342687	0,6525	vGm
hs_rRNA_28S_2292	_Cm	0,42303141	-0,13957403	170,61753	0,6495	_Cm
hs_rRNA_18S_918	Psi	-0,08428131	-0,60667765	103,109508	0,646	Psi
hs_rRNA_28S_1779	Psi	0,73961253	-0,06110438	207,56362	0,6415	Psi
hs_rRNA_18S_683	Gm	0,97642586	0,75178899	322,438306	0,6385	Gm
hs_rRNA_18S_799	Um	0,97763958	0,41541567	321,087962	0,635	Um
hs_rRNA_18S_1440	Cm	0,47993773	-0,43447617	174,387549	0,633	vCm
hs_rRNA_28S_3639	Psi	0,89991703	0,33803025	298,528581	0,63	Psi
hs_rRNA_28S_3770	Psi	0,89563465	0,42785135	297,80772	0,63	Psi
hs_rRNA_28S_4456	Cm	0,9796844	0,54711835	323,351247	0,629	Cm
hs_rRNA_28S_3724	Am	0,99218497	0,85575964	330,763367	0,628	Am
hs_rRNA_28S_4499	Gm	0,93927505	0,57420056	277,648126	0,628	Gm

hs_rRNA_18S_1031	Am	0,99010921	0,82164935	329,210651	0,624	Am
hs_rRNA_18S_1045	Psi	0,87407572	0,28885616	274,303776	0,6225	Psi
hs_rRNA_18S_801	Psi	0,84864924	-0,23415399	278,272161	0,622	Psi
hs_rRNA_18S_166	Am	0,98613476	0,78729493	326,496444	0,6215	Am
hs_rRNA_18S_1668	_Um	0,51014161	-0,0792842	163,653572	0,6185	_Um
hs_rRNA_18S_1244	Psi	0,92558275	0,45482051	304,610839	0,614	Psi
hs_rRNA_18S_1692	Psi	0,87686428	0,42184268	285,722388	0,613	Psi
hs_rRNA_28S_1524	Am	0,9823989	0,59102239	323,881917	0,6105	Am
hs_rRNA_28S_3825	Am	0,98725062	0,81851501	325,286624	0,6045	Am
hs_rRNA_18S_34	Psi	0,8371426	0,2081261	280,508632	0,603	Psi
hs_rRNA_28S_3959	Psi	0,86918499	0,39668823	291,498149	0,6	Psi
hs_rRNA_18S_1602	_Um	0,19934197	-0,3842433	115,802915	0,5945	_Um
hs_rRNA_18S_576	Am	0,96201988	0,7186203	309,095641	0,594	Am
hs_rRNA_28S_1522	Gm	0,98987158	0,59668026	327,996806	0,592	Gm
hs_rRNA_18S_1136	Psi	0,55815848	-0,00338208	194,242893	0,5915	Psi
hs_rRNA_28S_1625	Gm	0,98051274	0,76633245	324,381686	0,591	Gm
hs_rRNA_28S_4296	Psi	0,93361123	0,2819715	286,019823	0,5895	Psi
hs_rRNA_18S_428	Um	0,94870023	0,56734371	312,677501	0,589	Um
hs_rRNA_28S_2632	Psi	0,86221354	0,37136395	279,263981	0,589	Psi
hs_rRNA_18S_1174	Psi	0,91284704	0,52188053	300,07791	0,5885	Psi
hs_rRNA_18S_174	Cm	0,97646467	0,38500565	319,768403	0,587	vCm
hs_rRNA_28S_4576	Psi	0,89650677	0,40684257	296,955102	0,5865	Psi
hs_rRNA_28S_4361	Psi	0,75903114	0,25430363	259,235091	0,5865	Psi
hs_rRNA_18S_172	Um	0,97166487	0,40428544	319,262786	0,5855	Um
hs_rRNA_28S_4306	Um	0,9832872	0,70455751	323,59294	0,585	Um
hs_rRNA_28S_1781	Psi	0,70871624	0,14745792	241,574997	0,584	Psi
hs_rRNA_18S_601	Gm	0,97447886	0,70058617	320,637095	0,5835	Gm
hs_rRNA_28S_4228	Gm	0,9465077	0,75503007	272,391031	0,5835	Gm
hs_rRNA_18S_627	Um	0,9777862	0,42815862	320,532409	0,583	Um
hs_rRNA_28S_3715	Psi	0,90925334	0,12571608	287,129907	0,583	Psi
hs_rRNA_28S_2424	Gm	0,99431963	0,70157478	331,38119	0,5815	Gm
hs_rRNA_18S_797	Cm	0,89596228	0,12186844	288,408316	0,5815	vCm
hs_rRNA_28S_3768	Psi	0,63973582	-0,25859065	220,615089	0,5815	Psi
hs_rRNA_28S_1316	Gm	0,91287247	0,42101424	299,05871	0,5765	Gm
hs_rRNA_28S_3884	Psi	0,869611	0,20501559	277,573498	0,57	Psi
hs_rRNA_18S_1536	_Gm	0,43350912	-0,12586412	166,269603	0,5695	_Gm
hs_rRNA_28S_1760	Gm	0,97811775	0,74924313	323,06079	0,569	Gm
hs_rRNA_28S_3718	Am	0,9766153	0,7397475	320,359175	0,569	Am
hs_rRNA_18S_966	Psi	0,80202498	0,24866155	253,031531	0,569	Psi
hs_rRNA_18S_116	Um	0,98409065	0,79211601	325,126524	0,5675	Um

hs_rRNA_28S_4552	Psi	0,84303971	0,41664674	279,022678	0,567	Psi
hs_rRNA_28S_1534	Am	0,98874385	0,81828243	325,545693	0,5665	Am
hs_rRNA_28S_4392	Gm	0,98830781	0,81326161	326,42321	0,5655	Gm
hs_rRNA_28S_4196	Gm	0,99067073	0,79229922	329,783184	0,5645	Gm
hs_rRNA_18S_815	Psi	0,31886244	-0,45841411	156,779983	0,5635	Psi
hs_rRNA_5,8S_75	Gm	0,98972025	0,77063	329,261588	0,562	Gm
hs_rRNA_28S_3792	Gm	0,99521332	0,87926284	333,120606	0,5615	Gm
hs_rRNA_28S_4353	Psi	0,87575701	0,47040041	288,396362	0,5615	Psi
hs_rRNA_18S_1383	Am	0,99119046	0,68329371	329,550201	0,5605	Am
hs_rRNA_28S_1773	_Um	0,69365122	-0,30487107	226,194262	0,559	_Um
hs_rRNA_28S_1683	Psi	0,80586676	0,13060556	266,17724	0,556	Psi
hs_rRNA_18S_686	Psi	0,93018962	0,23302709	269,132853	0,5545	Psi
hs_rRNA_28S_4623	Gm	0,99212962	0,84951649	329,934859	0,551	Gm
hs_rRNA_18S_1347	Psi	0,89545331	0,50974708	292,044817	0,55	Psi
hs_rRNA_28S_4628	Psi	0,71519529	0,15191999	235,265318	0,544	Psi
hs_rRNA_28S_5010	Psi	0,57224047	0,04712546	202,542169	0,542	Psi
hs_rRNA_18S_644	Gm	0,98719384	0,80113618	324,414998	0,5415	Gm
hs_rRNA_28S_4370	Gm	0,96686924	0,68676868	318,520683	0,5405	Gm
hs_rRNA_28S_3627	Gm	0,97519446	0,72352452	322,106972	0,538	Gm
hs_rRNA_28S_2861	Cm	0,93509239	0,3899272	307,546673	0,538	vCm
hs_rRNA_28S_3944	Gm	0,96504957	0,56995726	318,011973	0,5375	Gm
hs_rRNA_28S_3695	Psi	0,83280707	0,32499515	274,773291	0,535	Psi
hs_rRNA_28S_3853	Psi	0,9153173	0,5840738	296,228948	0,53	Psi
hs_rRNA_18S_621	_Cm	-0,04743662	-0,45545292	106,201889	0,529	_Cm
hs_rRNA_28S_4299	Psi	0,88805291	0,30078444	274,909356	0,5245	Psi
hs_rRNA_28S_4431	Psi	0,79576077	0,30325145	257,921108	0,5195	Psi
hs_rRNA_28S_1782	Psi	0,63995442	0,09893354	208,555807	0,518	Psi
hs_rRNA_18S_436	Gm	0,93539448	0,54476969	305,324047	0,516	Gm
hs_rRNA_18S_822	Psi	0,80358512	0,09229153	256,844829	0,5145	Psi
hs_rRNA_28S_4620	Um	0,96298273	0,49557784	311,751899	0,514	vUm
hs_rRNA_18S_1447	Gm	0,82693746	-0,34066121	265,558077	0,514	vGm
hs_rRNA_28S_2401	Am	0,93675489	0,55626867	308,142946	0,5125	Am
hs_rRNA_18S_1056	Psi	0,9011867	0,27528841	295,844703	0,5125	Psi
hs_rRNA_18S_406	Psi	0,83276323	0,17122815	276,132935	0,512	Psi
hs_rRNA_18S_1177	Psi	0,79057053	0,06374014	249,333254	0,5105	Psi
hs_rRNA_18S_1326	Psim	0,99092776	0,68746688	328,869075	0,5085	Psim
hs_rRNA_18S_517	Cm	0,98913784	0,8258877	328,513285	0,5065	Cm
hs_rRNA_18S_1442	Um	0,98610738	0,7845168	326,879734	0,5025	Um
hs_rRNA_28S_4536	Cm	0,9877144	0,81776104	325,553756	0,501	Cm
hs_rRNA_28S_1677	Psi	0,83158916	-0,17132486	263,982255	0,4985	Psi

hs_rRNA_28S_1340	Cm	0,96086031	0,59921181	316,247016	0,497	vCm
hs_rRNA_28S_2422	Cm	0,98467411	0,54163464	323,281738	0,4955	Cm
hs_rRNA_28S_3899	Gm	0,98852172	0,81864837	328,573255	0,495	Gm
hs_rRNA_28S_4972	Psi	0,84013472	0,16989184	270,384208	0,494	Psi
hs_rRNA_28S_4579	Psi	0,89739128	-0,05585102	269,533859	0,4925	Psi
hs_rRNA_18S_509	Gm	0,98699272	0,76712582	326,207422	0,492	Gm
hs_rRNA_18S_218	Psi	0,81538669	0,3128936	266,104559	0,489	Psi
hs_rRNA_18S_867	Gm	0,45704474	0,14513732	187,404783	0,4855	Unmod
hs_rRNA_18S_609	Psi	0,90799054	0,60627097	243,182661	0,4775	Psi
hs_rRNA_28S_3887	Cm	0,98187148	0,6960043	324,086904	0,472	Cm
hs_rRNA_28S_4494	Gm	0,9667341	0,65270162	317,169937	0,4675	Gm
hs_rRNA_18S_1804	Um	0,97269427	0,66210967	320,039318	0,4665	Um
hs_rRNA_28S_2351	Cm	0,9863401	0,82104316	327,077804	0,4655	Cm
hs_rRNA_28S_2876	Gm	0,90558356	0,41382534	296,005905	0,4595	Unmod
hs_rRNA_18S_210	Psi	0,79680991	0,19379994	259,897487	0,4595	Unmod
hs_rRNA_28S_1881	Cm	0,91179323	0,40853625	299,389061	0,4575	Unmod
hs_rRNA_18S_1288	Um	0,98912362	0,75456655	328,058074	0,456	Um
hs_rRNA_28S_3869	Cm	0,96063168	0,49381644	315,346316	0,453	Cm
hs_rRNA_18S_1703	Cm	0,99067882	0,83458274	329,443389	0,4505	Cm
hs_rRNA_28S_2824	Cm	0,9570382	0,57481776	315,209693	0,448	Cm
hs_rRNA_28S_4457	Psi	-0,1303916	-0,026803	162,507894	0,4445	Unmod
hs_rRNA_28S_2837	Um	0,97279992	0,55390796	315,427126	0,4425	Um
hs_rRNA_28S_4054	Cm	0,98445861	0,80358787	326,172666	0,43	Cm
hs_rRNA_28S_2804	Cm	0,98204003	0,72995815	311,050934	0,4295	vCm
hs_rRNA_28S_3760	Am	0,95023662	0,51209897	312,908146	0,428	vAm
hs_rRNA_18S_462	Cm	0,98980912	0,75877804	328,478948	0,426	Cm
hs_rRNA_28S_3808	Cm	0,97937552	0,77842862	323,845692	0,425	Cm
hs_rRNA_28S_3744	Gm	0,93056128	0,33625026	302,852503	0,422	vGm
hs_rRNA_18S_354	Um	0,90124173	0,38910415	275,951002	0,418	Unmod
hs_rRNA_28S_3925	Um	0,98827807	0,68853205	328,08521	0,416	Um
hs_rRNA_28S_3701	Cm	0,97826692	0,63007695	322,833249	0,412	Cm
hs_rRNA_28S_3841	Cm	0,98862659	0,58802956	325,134839	0,39	Cm
hs_rRNA_18S_1391	Cm	0,98861812	0,81987953	326,963296	0,387	Cm
hs_rRNA_28S_2415	Um	0,90968907	0,47774437	300,45006	0,3755	Um
hs_rRNA_28S_3818	Psim	0,97334818	0,61361664	315,935724	0,3225	Psim

Column “modification“ contains known identity of rRNA residue, Nm, Psi or Unmodified (Unmod)

Columns “Vote Mod” holds information on RF prediction to be Modified

Table S2 List of highly confident rRNA modification sites detected in yeast *S. cerevisiae* rRNA

Median values for Scores and Votes for six *S. cerevisiae* validation datasets are shown

rRNA and position	modification	ScoreMean	scoreA2	scoreAngle	Max Vote Mod	Predicted as
sc_rRNA_18S_541	Am	0,98428331	0,79178051	325,0480264	0,7815	Am
sc_rRNA_18S_619	Am	0,9887353	0,8303337	328,468759	0,7705	Am
sc_rRNA_25S_1133	Am	0,98850177	0,83634938	328,4711909	0,736	Am
sc_rRNA_25S_2946	Am	0,99420054	0,73534393	332,065589	0,7335	Am
sc_rRNA_18S_28	Am	0,98574515	0,78664016	326,8433722	0,7285	Am
sc_rRNA_25S_2640	Am	0,98558639	0,75448355	325,8186171	0,7105	Am
sc_rRNA_18S_796	Am	0,9809671	0,73194225	324,4859038	0,6645	Am
sc_rRNA_18S_1271	Gm	0,99014425	0,63264518	327,0258661	0,663	Gm
sc_rRNA_18S_974	Am	0,98912302	0,73284214	327,6216994	0,6535	Am
sc_rRNA_25S_2220	Am	0,99615403	0,88054671	334,2906997	0,6395	Am
sc_rRNA_25S_876	Am	0,98559841	0,78363359	326,6618347	0,6185	Am
sc_rRNA_18S_1187	Psi	0,94156811	0,45174693	307,9898662	0,6065	Psi
sc_rRNA_18S_1181	Psi	0,91796198	0,22502491	288,5188739	0,6055	Psi
sc_rRNA_18S_100	Am	0,96048696	0,59831634	314,0061244	0,5605	Am
sc_rRNA_25S_1450	Gm	0,97511493	0,84046354	294,6426708	0,559	Gm
sc_rRNA_25S_1888	Um	0,97934576	0,74356858	323,4773475	0,559	Um
sc_rRNA_25S_981	Unmod	0,86998682	0,19693897	283,7286683	0,5525	Psi
sc_rRNA_25S_2314	Psi	0,86330612	0,30947093	282,2949118	0,552	Psi
sc_rRNA_25S_805	Gm	0,98938034	0,65834707	327,9843681	0,5415	Gm
sc_rRNA_25S_2256	Am	0,97500712	0,64633751	322,0909426	0,5335	Am
sc_rRNA_25S_2266	Psi	0,8842604	0,30756658	292,6645789	0,5335	Psi
sc_rRNA_25S_2921	Um	0,98879576	0,91466607	309,2882416	0,5325	Um
sc_rRNA_18S_1428	Gm	0,98076209	0,71549736	323,7885532	0,516	Gm
sc_rRNA_25S_2258	Psi	0,79139992	-0,53930975	252,3031168	0,513	Psi
sc_rRNA_18S_1290	Psi	0,8429452	0,32157124	272,656757	0,5115	Psi
sc_rRNA_18S_1126	Gm	0,96484624	0,62173175	316,4955371	0,5105	Gm
sc_rRNA_18S_227	Unmod	0,90018654	0,52535734	294,8941432	0,509	Psi
sc_rRNA_25S_817	Am	0,9831328	0,78070467	321,8621998	0,508	Am
sc_rRNA_18S_1117	Unmod	0,85177884	0,30272621	281,2440598	0,5075	Psi
sc_rRNA_25S_2619	Gm	0,99163887	0,80658856	330,5437163	0,5065	Gm
sc_rRNA_25S_2153	Unmod	0,8386365	-0,04368898	265,1724854	0,5015	Psi
sc_rRNA_25S_1437	Cm	0,98659633	0,78822633	326,5352178	0,4905	Cm
sc_rRNA_25S_2815	Gm	0,98541178	0,77505625	326,3823024	0,489	Gm
sc_rRNA_25S_3078	Unmod	0,89848299	0,09983209	289,2418465	0,483	Psi
sc_rRNA_25S_2944	Psi	0,86604828	-0,19859108	270,4106952	0,481	Psi
sc_rRNA_25S_2826	Psi	0,93764822	0,31483231	305,6752028	0,471	Psi
sc_rRNA_25S_908	Gm	0,97562744	0,68203287	320,436976	0,4635	Gm

sc_rRNA_25S_2260	Psi	0,92089407	0,33329778	298,0591517	0,461	Unmod
sc_rRNA_25S_1829	Unmod	0,96361963	0,5881075	315,1402955	0,4505	Gm
sc_rRNA_25S_2959	Cm	0,98739579	0,82188612	324,4271798	0,4445	Cm
sc_rRNA_25S_2288	Gm	0,99233227	0,85342886	330,8934263	0,4415	Gm
sc_rRNA_25S_867	Gm	0,97963168	0,76821612	324,0227086	0,437	Gm
sc_rRNA_18S_578	Um	0,98019204	0,56463517	323,2478755	0,4365	Um
sc_rRNA_18S_420	Am	0,94439003	0,58530901	309,8133588	0,4345	Am
sc_rRNA_25S_2197	Cm	0,95570209	0,45929764	313,0136135	0,434	Unmod
sc_rRNA_25S_1449	Am	0,85102849	0,65837545	216,4853247	0,432	Unmod
sc_rRNA_25S_2417	Um	0,98549695	0,72771185	324,7057174	0,43	Um
sc_rRNA_18S_1639	Cm	0,98649088	0,8031294	327,1170404	0,419	Cm
sc_rRNA_25S_2729	Um	0,96239695	0,67883574	317,7305657	0,4035	
sc_rRNA_25S_2922	Gm	0,84374525	0,61713284	203,7669678	0,4025	Gm
sc_rRNA_25S_2421	Um	0,99349533	0,85115137	331,8534751	0,39	Um
sc_rRNA_18S_1572	Gm	0,99165853	-0,02360179	326,6465383	0,388	Unmod
sc_rRNA_25S_807	Am	0,97330467	0,47491464	319,6011964	0,3825	Am
sc_rRNA_18S_414	Cm	0,97781437	0,68168954	322,528177	0,367	Cm
sc_rRNA_25S_2337	Cm	0,99177844	0,76090885	328,8308954	0,3665	Cm
sc_rRNA_18S_1007	Cm	0,98205044	0,75211757	324,6214159	0,3525	Cm
sc_rRNA_25S_898	Um	0,95082852	0,64500419	314,1005211	0,3465	Um
sc_rRNA_25S_2724	Um	0,99804871	0,89153566	336,9633954	0,3415	Unmod
sc_rRNA_18S_1269	Um	0,99222113	0,6532909	328,9112021	0,329	Psim
sc_rRNA_18S_562	Gm	0,9789443	0,46916691	323,1002118	0,3065	Gm
sc_rRNA_25S_2793	Gm	0,99430267	0,68203318	331,9954347	0,2925	Unmod
sc_rRNA_25S_2791	Gm	0,96678343	0,22310727	312,7613178	0,261	Unmod

Column “modification“ contains known identity of rRNA residue, Nm, Psi or Unmodified (Unmod)

Columns “Vote Mod” and “Vote Unmod“ hold information on RF prediction to be Modified or Unmodified

Table S3 List of highly confident rRNA modification sites detected in plant *A. thaliana* rRNA

Median values for Scores and Votes for three *A. thaliana* validation datasets are shown

rRNA and position	modification	ScoreMean	scoreA2	scoreAngle	Max Vote Mod	Predicted as
at_rRNA_28S_1377	Am	0,97361881	0,7306559	321,4587979	0,746	Am
at_rRNA_28S_2947	Am	0,99467671	0,71566025	332,5698293	0,729	Am
at_rRNA_28S_2327	Am	0,98228162	0,77051275	325,3353277	0,704	Am
at_rRNA_28S_2641	Am	0,97663763	0,7408018	322,4645646	0,682	Am
at_rRNA_18S_440	Am	0,97672521	0,72140782	320,3926822	0,672	Am
at_rRNA_18S_801	Am	0,98245937	0,64146504	322,374694	0,668	Am
at_rRNA_28S_1143	Am	0,96289319	0,66913425	315,6823511	0,667	Am
at_rRNA_28S_885	Am	0,99000999	0,79862746	327,7950813	0,653	Am
at_rRNA_5.8S_47	Am	0,9652334	0,64491358	317,9325972	0,646	Am
at_rRNA_18S_796	Am	0,98731509	0,65138938	327,0406011	0,639	Am
at_rRNA_18S_623	Am	0,96031047	0,70448885	315,8454935	0,605	Am
at_rRNA_28S_2597	Unmod	0,92499166	0,44052423	303,975264	0,599	Psi
at_rRNA_28S_1481	Unmod	0,93876062	0,265622	306,6172551	0,577	Psi
at_rRNA_28S_1263	Am	0,97290622	0,7411224	321,2947464	0,573	Am
at_rRNA_28S_601	Unmod	0,85074047	0,03284147	283,5536872	0,561	Psi
at_rRNA_28S_743	Unmod	0,8280069	-0,10555418	272,0861051	0,561	Psi
at_rRNA_28S_2614	Unmod	0,85124962	0,40535885	286,2621546	0,551	Psi
at_rRNA_28S_2322	Am	0,98095073	0,69340751	321,6665797	0,55	Am
at_rRNA_18S_1314	Unmod	0,88491534	0,37199117	289,0731499	0,547	Psi
at_rRNA_28S_1528	Unmod	0,8754249	0,36445532	283,1015393	0,547	Psi
at_rRNA_28S_1249	Unmod	0,90101812	0,23646784	278,5254533	0,542	Psi
at_rRNA_28S_2652	Gm	0,96727241	0,75030526	308,1798766	0,54	Gm
at_rRNA_28S_2221	Am	0,97318534	0,719021	320,7966213	0,529	Am
at_rRNA_28S_2319	Unmod	0,90250984	0,1958924	290,0354626	0,529	Psi
at_rRNA_28S_2651	Um	0,57996963	0,28187904	187,7035172	0,528	Um
at_rRNA_28S_2127	Am	0,98374344	0,600458	324,6921278	0,516	Am
at_rRNA_28S_2884	Um	0,91586385	0,50883974	296,7457346	0,51	Psi
at_rRNA_28S_3286	Unmod	0,84783095	0,06173101	279,3350663	0,51	Psi
at_rRNA_18S_1275	Gm	0,96765263	0,2388956	315,6725782	0,504	Gm
at_rRNA_28S_2396	Gm	0,98497886	0,80568863	326,7711104	0,503	Gm
at_rRNA_28S_1877	Unmod	0,83655735	0,30431623	277,7719452	0,503	Psi
at_rRNA_18S_1434	Gm	0,95447839	0,56162255	313,0579889	0,495	Gm
at_rRNA_28S_2456	Um	0,90231169	0,46998598	299,1164514	0,495	Psi
at_rRNA_28S_2212	Psi	0,90588029	0,10501106	289,9661908	0,492	Psi
at_rRNA_28S_2137	Psi	0,90566652	0,43636094	301,5020982	0,487	Psi
at_rRNA_18S_763	Psi	0,86302234	0,30156819	287,2072665	0,487	Psi
at_rRNA_18S_791	Unmod	0,857046	0,41575163	286,0216125	0,487	Psi

at_rRNA_28S_2935	Am	0,96839327	0,648149	317,6749822	0,486	Am
at_rRNA_28S_1479	Cm	0,9494252	-0,02384479	292,5618821	0,485	Unmod
at_rRNA_28S_1598	Unmod	0,8175477	0,36099503	277,480419	0,485	Psi
at_rRNA_28S_1587	Unmod	0,91730006	0,32838821	293,0929534	0,485	Psi
at_rRNA_28S_1001	Psi	0,92406198	0,40936571	304,5006946	0,483	Psi
at_rRNA_18S_1384	Um	0,97888247	0,77397352	323,447965	0,483	Um
at_rRNA_18S_1554	Um	0,92099857	0,15714963	285,3308746	0,482	Psi
at_rRNA_28S_2620	Gm	0,98211278	0,75929847	325,2689296	0,477	Gm
at_rRNA_18S_780	Am	0,96783998	0,57952173	317,1079283	0,47	Am
at_rRNA_28S_2362	Unmod	0,95801372	0,53517068	314,2045222	0,467	Am
at_rRNA_18S_1305	Unmod	0,86164169	0,23619267	286,9448381	0,466	Psi
at_rRNA_28S_2725	Unmod	0,89469522	0,33876568	293,7534478	0,464	Psi
at_rRNA_28S_661	Am	0,94752364	0,54450684	302,6845002	0,455	Unmod
at_rRNA_28S_803	Um	0,9783353	0,77042643	322,9016428	0,455	Um
at_rRNA_18S_1579	Am	0,9563641	0,6057188	311,9380787	0,452	Unmod
at_rRNA_28S_828	Psi	0,90852711	0,00525179	293,466748	0,451	Psi
at_rRNA_28S_1447	Cm	0,98179856	0,7982211	324,5018755	0,449	Cm
at_rRNA_28S_2816	Gm	0,96826466	0,66461932	318,0134687	0,449	Gm
at_rRNA_28S_3100	Unmod	0,89872303	0,45583358	299,0985976	0,448	Psi
at_rRNA_28S_1518	Cm	0,94147559	0,61631282	309,6954091	0,447	Unmod
at_rRNA_28S_826	Am	0,96197659	0,2660352	312,5710896	0,444	Unmod
at_rRNA_28S_2114	Um	0,97888341	0,65550098	323,3639298	0,439	Um
at_rRNA_18S_468	Am	0,9438491	0,60600615	310,8735044	0,436	Unmod
at_rRNA_18S_1185	Unmod	0,95149281	0,46520109	307,4529841	0,436	Psi
at_rRNA_28S_399	Gm	0,97928407	0,7136452	323,2433052	0,435	Gm
at_rRNA_28S_2215	Am	0,95616935	0,61707858	311,6119797	0,434	Unmod
at_rRNA_18S_28	Am	0,94939312	0,61563018	314,3932127	0,433	Am
at_rRNA_28S_945	Am	0,94680972	0,54066572	312,0769638	0,428	Unmod
at_rRNA_28S_917	Gm	0,95756119	0,67263488	315,439053	0,428	Gm
at_rRNA_28S_2392	Gm	0,98091619	0,71514726	323,2809038	0,425	Gm
at_rRNA_28S_2736	Um	0,96331472	0,63079981	318,069896	0,422	Um
at_rRNA_28S_2198	Cm	0,95619382	0,5095454	313,9505568	0,419	Unmod
at_rRNA_28S_3292	Gm	0,96900799	0,69203078	319,560779	0,419	Gm
at_rRNA_28S_1892	Um	0,96737631	0,62774963	316,9084764	0,419	Um
at_rRNA_28S_1855	Gm	0,95119378	0,48963203	307,7601364	0,416	Unmod
at_rRNA_28S_3301	Um	0,98522802	0,75351938	326,8505976	0,412	Um
at_rRNA_5.8S_155	Unmod	0,97626771	0,63381748	321,5303896	0,411	Unmod
at_rRNA_28S_2289	Gm	0,98895674	0,81246309	328,1262377	0,41	Unmod
at_rRNA_18S_1309	Unmod	0,91852848	0,60815827	300,7543287	0,409	Psi
at_rRNA_18S_1273	Um	0,98488554	0,56587028	325,3159842	0,402	Um

at_rRNA_18S_978	Am	0,97551966	0,53205797	316,3678778	0,401	Am
at_rRNA_28S_814	Gm	0,97526237	0,40911002	320,683132	0,395	Gm
at_rRNA_18S_812	Unmod	0,96353258	0,48440894	316,7115804	0,393	Unmod
at_rRNA_18S_418	Cm	0,98958785	0,76198089	328,7057882	0,392	Unmod
at_rRNA_28S_2125	Gm	0,98581054	0,63764315	325,6228162	0,392	Gm
at_rRNA_18S_246	Gm	0,95970367	0,61171781	316,315909	0,389	Gm
at_rRNA_28S_1460	Gm	0,94885629	0,66153182	252,6748561	0,387	Gm
at_rRNA_18S_615	Um	0,9480666	0,65460236	312,0639482	0,386	Um
at_rRNA_28S_816	Am	0,96943595	0,46514042	318,871161	0,384	Am
at_rRNA_18S_1235	Um	0,98022625	0,718265	323,7275031	0,384	Um
at_rRNA_28S_378	Unmod	0,96496442	0,40424032	313,3370286	0,383	Unmod
at_rRNA_18S_582	Um	0,95725174	0,25723859	311,7781462	0,382	Unmod
at_rRNA_28S_2792	Gm	0,96842938	0,20444538	310,8406892	0,375	Unmod
at_rRNA_18S_1121	Psi	0,91807084	0,3967756	301,7121952	0,374	Psi
at_rRNA_28S_2960	Cm	0,97134578	0,73832099	318,2111799	0,373	Cm
at_rRNA_18S_1645	Cm	0,98194546	0,77470263	324,3748581	0,369	Cm
at_rRNA_18S_599	Gm	0,97484067	0,6687857	317,9818681	0,368	Gm
at_rRNA_18S_392	Gm	0,9508924	0,52656058	312,9686216	0,359	Gm
at_rRNA_28S_2366	Cm	0,96581005	0,60436735	316,7388003	0,353	Cm
at_rRNA_5.8S_79	Gm	0,95385435	0,57208257	314,24389	0,351	Gm
at_rRNA_28S_676	Psi	0,88722505	0,34024808	249,6255305	0,349	Psi
at_rRNA_18S_1264	Um	0,97859451	0,44340393	322,4055087	0,349	Um
at_rRNA_18S_1219	Cm	0,96997556	0,75301724	317,6021711	0,347	Cm
at_rRNA_28S_2422	Um	0,98426424	0,67325023	326,011715	0,345	Um
at_rRNA_18S_1330	Am	0,97499392	0,37316384	317,530588	0,344	Unmod
at_rRNA_28S_2794	Gm	0,99202208	0,47723456	327,267988	0,339	Gm
at_rRNA_18S_1448	Um	0,95625008	0,67347868	313,2029851	0,336	Um
at_rRNA_18S_213	Unmod	0,96667285	0,70772751	318,2820924	0,332	Um
at_rRNA_28S_2237	Unmod	0,99220026	0,84216727	330,8528968	0,331	Um
at_rRNA_28S_2410	Unmod	0,89295005	0,69871262	231,6090017	0,33	vAm
at_rRNA_28S_1860	Cm	0,9885725	0,81658458	325,8191186	0,327	Cm
at_rRNA_28S_2683	Cm	0,97983375	0,77798187	323,5885872	0,326	Cm
at_rRNA_28S_1067	Um	0,96863794	0,72102231	319,3149901	0,326	Um
at_rRNA_28S_144	Um	0,95240098	0,59729341	312,6247848	0,326	Unmod
at_rRNA_18S_604	Um	0,98204824	0,70977825	324,9947006	0,321	Um
at_rRNA_28S_2880	Cm	0,95289932	0,23023818	310,5031621	0,318	Cm
at_rRNA_28S_1850	Cm	0,97386312	0,65867397	316,018962	0,318	Unmod
at_rRNA_28S_2294	Cm	0,98091338	0,7759619	324,6125055	0,315	Cm
at_rRNA_18S_473	Cm	0,96762198	0,57284949	316,7984351	0,314	Cm
at_rRNA_28S_2338	Cm	0,98118126	0,66722202	318,1323349	0,313	Cm

at_rRNA_28S_1278	Um	0,9512535	0,67083282	313,3146086	0,313	Um
at_rRNA_18S_1013	Um	0,9890782	0,814281	328,1506878	0,306	Um
at_rRNA_18S_38	Cm	0,9461602	0,47907064	311,6287974	0,305	Unmod
at_rRNA_28S_48	Um	0,96697378	0,70849712	318,5042711	0,294	Um
at_rRNA_28S_2494	Um	0,97526522	0,73781943	322,1712466	0,294	Um
at_rRNA_18S_123	Um	0,98055305	0,73189081	324,5109966	0,271	Um
at_rRNA_28S_2411	Um	0,92463032	0,65779686	274,0541063	0,262	Um
at_rRNA_18S_1266	Um	0,97760244	0,4826133	320,7803686	0,246	Psi
at_rRNA_28S_44	Um	0,98392963	0,72558924	325,2239929	0,238	Gm

Column “modification“ contains known identity of rRNA residue, Nm, Psi or Unmodified (Unmod)

Columns “Vote Mod” and “Vote Unmod“ hold information on RF prediction to be Modified or Unmodified

-- End of article --

Chapter 2: Creation of alternative detection methods

4.2.1 HydraPsiSeq: a method for systematic and quantitative mapping of pseudouridines in RNA. [DOI: 10.1093/nar/gkaa769]

Virginie MARCHAND[1], Florian PICHOT[1,2], Paul NEYBECKER[3], Lilia AYADI[1,3], Valérie BOURGUIGNON-IGEL[1,3], Ludivine WACHEUL[4], Denis L.J. LAFONTAINE[4], Astrid PINZANO[3], Mark HELM[2], Yuri MOTORIN[1,3]

[1] Université de Lorraine, CNRS, INSERM, IBSLor (UMS2008/US40), Epitranscriptomics and RNA Sequencing Core Facility, F54000 Nancy, France

[2] Institute of Pharmaceutical and Biomedical Science, Johannes Gutenberg-University Mainz, 55128 Mainz, Germany

[3] Université de Lorraine, CNRS, IMoPA (UMR7365), F54000 Nancy, France

[4] RNA Molecular Biology, ULB-Cancer Research Center (U-CRC), Center for Microscopy and Molecular Imaging (CMMI), Fonds de la Recherche Scientifique (F.R.S./FNRS), and Université Libre de Bruxelles (ULB), BioPark campus, B-6041 Gosselies, Belgium

Overview

During the development of this method (2018), only methods using CMC-adducts existed: Psi-Seq (or ψ-seq)[314], PSI-Seq[453], Pseudo-seq[269] and CeU-seq[454]. Even if they are based on the same methodology, their results did not overlap[455], creating confusion about the trustworthiness of each article. Moreover, the limitations of such an “RT-arrest” approach about detection and quantification are well-known (see Section 1.3). This is why in order to have a CMC-independant, quantitative detection method for pseudouridine, an attempt based on chemical signature of pseudouridine named “HydraPsiSeq” has been conducted.

In this approach, hydrazine treatment occurs in ring-opening of the pyrimidine base of uridine resulting in an abasic site but not for pseudouridine, allowing a chemical distinction between those residues. The abasic sites are then removed through subsequent aniline treatment, resulting in specific RNA cleavages at uridine sites detectable through high-throughput sequencing. Then, the pseudouridine detection resides in the relative absence of cleavage compared to the unmodified uridines.

Validation of this method has been done on *S. cerevisiae* strains with known rRNA pseudouridine sites. Mutant strains for two H/ACA box snoRNA, snR44 and snR81 responsible for the enzymatic catalyzation of respectively two and one known specific rRNA pseudouridine site served as negative control. In that case, the HydraPsiSeq signal is shown to be negative at specified sites, proving the sensitivity of the method. Similar results have been obtained with inactive Cbf5 yeast strain, Cbf5 being a catalytical subunit of H/ACA snoRNP responsible for isomerisation of uridines to pseudourindines.

Afterwards, mixes of this mutant strain and wild-type strain have been carried on to evaluate the quantification reliability of the method, which shows indeed high correlation with pseudouridine content. HydraPsiSeq has also been tested on *S. cerevisiae* mRNAs and found similar results with Carlile et al.[269] and Schwartz and al.[314] works.

Finally, HydraPsiSeq rRNA pseudouridine profiling has been conducted to a study project on human cells which shows pseudouridylation variation for a subset of sites during cell differentiation. This implies a dynamic adaptation role of rRNA pseudouridine throughout cell differentiation.

For this paper, I took care of the bioinformatic pipeline creation for the alignment, data importation in R, data normalization and detection scores set-up.

ABSTRACT

Developing methods for accurate detection of RNA modifications remains a major challenge in epitranscriptomics. Next-generation sequencing-based mapping approaches have recently emerged but, often, they are not quantitative and lack specificity. Pseudouridine (Ψ), produced by uridine isomerization, is one of the most abundant RNA modification. Ψ mapping classically involves derivatization with soluble carbodiimide (CMCT), which is prone to variation making this approach only semi-quantitative. Here we developed "HydraPsiSeq", a novel quantitative Ψ mapping technique relying on specific protection from hydrazine/aniline cleavage. HydraPsiSeq is quantitative because the obtained signal directly reflects pseudouridine level. Furthermore, normalization to natural unmodified RNA and/or to synthetic *in vitro* transcripts allows absolute measurements of modification levels. HydraPsiSeq requires minute amounts of RNA (as low as 10-50 ng), making it compatible with high-throughput profiling of diverse biological and clinical samples. Exploring the potential of HydraPsiSeq, we profiled human rRNAs, revealing strong variations in pseudouridylation levels at ~20-25 positions out of total 104 sites. We also observed the dynamics of rRNA pseudouridylation throughout chondrogenic differentiation of human bone marrow stem cells. In conclusion, HydraPsiSeq is a robust approach for the systematic mapping and accurate quantification of pseudouridines in RNAs with applications in disease, aging, development, differentiation and/or stress response.

INTRODUCTION

Single nucleotide-resolution mapping and precise quantification of RNA modified nucleotides are key challenges in epitranscriptomics [32, 31, 422, 420]. This information is of crucial importance for analysis of biological functions of these modified RNA residues [456, 416, 457, 415], since only precise and reliable knowledge of RNA modification profiles allows their correlation with physiological state of the cell and other factors influencing RNA maturation and functions. Quantification of RNA modification is possible by conventional and low-throughput methods such as HPLC, or combination of LC with MS analysis [458–461], but since those techniques rarely provide unambiguous mapping information, profiling of biological samples mostly relies on high-throughput approaches based on deep sequencing of second [420], and now third generation [462, 463]. Despite a great success in mapping different RNA modified residues in full transcriptomes, only very few available deep sequencing-based methods for RNA modification mapping combine both single-nucleotide resolution and precise quantification, since multiple enrichment steps and/or various biases during library preparation generally make quantification very complicated, or even impossible.

Deep sequencing-based methods relying on antibody capture (e.g. MeRIP,[464, 465]) provide limited quantitative information on the modification level due to variable enrichment [466], but become quantitative upon inclusion of spike-in controls and parallel sequencing of both bound and unbound fractions [467]. Protocols using chemical modification strategies without further selection of modified residues

generally provide quantitative information on the modification level, in particular if the readout signal is not a full reverse transcriptase (RT)-arrest of primer extension, which is generally considered as being noisy. For instance, quantitative assessment of m^5C levels is readily achieved in bisulfite-sequencing (BS) and derived protocols, under the assumption that deamination level and C->U conversion is close to 100%, and thus residual non-deaminated signal is essentially generated by modified m^5C [468–470]. Another example of quantification for a common RNA modification is the alkaline hydrolysis-based RiboMethSeq, where protection of the phosphodiester bond between nucleotides N and N+1 correlates with 2'-O-methylation level of the first nucleotide (N). RiboMethSeq thus allows precise profiling of Nm methylation in various RNAs and under varying growth/stress conditions [64, 434, 446]. Finally, several protocols, such as m^1A detection by RT misincorporation signatures [321, 322, 471], or the AlkAniline-Seq method for m^7G, m^3C and D detection [360], provide quantitative information, but require careful calibration with artificial mixes of differentially modified RNA or corresponding synthetic oligonucleotides.

Pseudouridine, an abundant isomer of uridine, is a commonly modified nucleotide in RNA, mostly prevalent in tRNAs and rRNAs, but also present in snRNAs, snoRNAs and eukaryotic mRNAs [472, 473, 320, 241]. The broad distribution of pseudouridine residues stimulated the development of methods for their mapping in RNAs, first in low-throughput implementations using RT primer extension after chemical treatment, followed by analysis by high-resolution sequencing gels [282, 474, 475, 307, 260], and, more recently, by coupling with deep sequencing to increase throughput and sensitivity [320, 314, 454, 476]. However, despite notable differences in output data volume, both low-and high-throughput approaches described up to now were based on the use of the same chemistry, namely CMC-modification of pseudouridine residues under strongly denaturing conditions, followed by resolution of CMC-conjugates to canonical U and G residues by long incubations in alkaline bicarbonate buffer [282, 307]. Inevitably, due to the instability of RNA under these relatively "harsh" conditions, the noise is high. Among other factors, this results from accelerated unspecific phosphodiester bond cleavage, thus causing numerous CMC-independent RT-stops and corresponding hits in deep sequencing. Inclusion of CMC-untreated controls [477, 307, 476] and/or analysis of time courses for bicarbonate CMC removal, allow to alleviate these limitations to some extent, but makes the already sophisticated experimental procedure even more complicated and relatively expensive in terms of sequencing costs. In classical RT primer extension versions, CMC-Ψ adducts were not quantified, since only partial linearity was observed at low modification levels, and the intensity of the signal was highly variable [282, 474, 478], while deep sequencing CMC-based protocols were made semi-quantitative by inclusion of synthetic spike-ins and calibrations [34, 314]. Other chemical cleavage/modification approaches, namely using hydrazine [282, 475], acrylonitrile [479, 480] and methylvinylsulfone [481] were also explored in the past, but did not become popular in practice [307].

Here we developed a sensitive, reliable, and quantitative approach for pseudouridine mapping in RNAs based on a chemistry orthogonal to classical CMC derivatization.

To achieve this goal, we explored and optimized random RNA cleavage at uridine residues by hydrazine, followed by aniline treatment for RNA chain scission at abasic sites. Protected residues (negative hits) reveal the presence of pseudouridines based on their resistance to hydrazine-dependent scission. These protection signals were compared to efficiency of neighboring cleavages at unmodified uridines to obtain a quantitative information on the pseudouridylation levels. The method termed "HydraPsiSeq" was successfully applied for precise mapping and quantification of pseudouridine residues in yeast *S. cerevisiae* and *H. sapiens* rRNAs and tRNAs, as well as *S. cerevisiae* mRNAs. Application of HydraPsiSeq revealed differential rRNA pseudouridylation profiles in different human cell lines and highlighted a dynamic regulation of rRNA pseudouridylation levels during chondrogenic stem cells differentiation.

MATERIALS AND METHODS

Yeast strains and cultures

Yeast strains used in this study were obtained either from the lab of Denis L. J. Lafontaine or from the EUROSCARF collection (Germany, see Table S1 in Supporting information). Cells were grown in standard Yeast Extract/Peptone/Dextrose (YPD) to 0.6-0.7 OD_{600} for exponential phase.

The box H/ACA snoRNA-associated pseudouridine synthase Cbf5 is encoded by an essential gene [482]. To modulate pseudouridylation levels in RNA, we made use either of genetic depletion (pGAL-cbf5, strain YDL521-1, [482], or expression of a catalytically-deficient allele (Cbf5-D95A allele [483, 484], strain YDL2932, and its isogenic control YDL2931).

To produce rRNAs with reduced box H/ACA-mediated pseudouridylation, we depleted Cbf5 in a haploid yeast strain expressing as its sole copy of CBF5 a GAL-regulated allele [482]. This allele drives the expression of CBF5 in galactose-based but not in glucose-based medium. YDL521-1 cells were grown in permissive conditions (2% galactose, 2% sucrose, 2% raffinose synthetic medium lacking histidine) to mid-log phase, washed in pre-warmed water, and resuspended in pre-warmed non-permissive conditions (2% glucose synthetic medium lacking histidine) at an OD_{600} of 0.3. Cells were maintained in log phase by continuous dilution in fresh 2% glucose synthetic medium lacking histidine and collected before transfer (time point 0), and after 24- and 48-hr of transfer to glucose.

To produce rRNA deprived of box H/ACA mediated pseudouridylation, we used cells expressing as sole source of Cbf5 a catalytically deficient allele (CBF5cata, Cbf5-D95A, strain YDL2932). YDL2932 is a haploid yeast strain in which the endogenous CBF5 gene was deleted by insertion of a TRP1 cassette and rescued by expression of Cbf5-D95A from a low copy ARS/CEN HIS3 plasmid. Cells were grown to mid-log phase in YPD rich medium (1% yeast extract, 2% peptone, 2% glucose).

RNA extraction
Total RNA from yeast cells was isolated using hot acid phenol [485, 486]. RNA concentration was measured on a Nanodrop One and RNA quality was assessed by capillary electrophoresis using a PicoRNA chip on Bioanalyzer 2100 (Agilent technologies, USA). Yeast PolyA+ mRNA fraction was prepared from total RNA using NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB), according to manufacturer's instructions.

RNA fragmentation conditions
RNA (50-300 ng) was subjected to hydrazine treatment (50% final concentration) for 30-60 min on ice. The reaction was stopped by ethanol precipitation using 0.3M NaOAc, pH5.2 and Glycoblue and incubated 30 min at -80°C. After centrifugation, the pellet was washed twice with 80% ethanol and resuspended in 1M aniline, pH4.5. The reaction was incubated in dark for 15 min at 60°C and processed as described above, by ethanol precipitation.

RNA 3'-end dephosphorylation
RNA was dephosphorylated at the 3'-end using 10 U of T4 PNK in 100 mM Tris-HCl pH6.5, 100 mM MgOAc and 5 mM □-mercaptoethanol and incubated for 6h at 37°C. T4 PNK was inactivated by incubation for 20 min at 65°C. RNA was extracted by phenol:chloroform and ethanol precipitated. The pellet was resuspended in RNase free water.

Library preparation
RNA was converted to library using NEBNext Small RNA Library kit (NEB, UK) using the manufacturer's instructions. Library quality was assessed using a High Sensitivity DNA chip on a Bioanalyzer 2100. Library quantification was done using a fluorometer (Qubit 3.0 fluorometer, Invitrogen, USA).

Sequencing
Libraries were multiplexed and subjected for high-throughput sequencing using an Illumina HiSeq 1000 instrument with 50 bp single-read runs. Libraries were loaded onto flow-cell at 10-12 pM final concentration.

Bioinformatic pipeline
Adapter removal was performed using the Trimmomatic utility v32.0 [317], with a stringency parameter of 7, very short reads < 8 nt were excluded. The alignment of raw reads was conducted by Bowtie2 [318] in end-to-end mode. Mapped and sorted *.bam file was transformed to *.bed format. Locations of 5'-extremities of mapped unique reads were counted from *.bed file, giving raw cleavage profile. Reads' 5'-end counts were normalized to local background in rolling window of 10 nucleotides, values for U residues were excluded to calculate the median. If RNA reference contains multiple U-rich stretches > 10 in length, normalization window can be extended to 16

nt or more. Locally normalized U profiles (NormUcount, full profiles shown in figures) were used as protection Uscore value and further transformed to U cleavage profiles only, by omitting values for other nucleotides. Resulting U profiles were used for calculation of ”RiboMethSeq-like” scores (ScoreMEAN, A, B and PsiScore, equivalent to ScoreC in RiboMethSeq). Window of 4 neighboring nucleotides (+/-2 nt) was used for score calculation [430].

ScoreMEAN for each position is calculated in two steps, as follows: first, a ratio of number of cumulated 5’/3’-reads ends between preceding and following position is defined and, second, ScoreMEAN is calculated as a ratio of a drop for a given position compared to the average and variation for 4 neighboring positions (-2/+2). PsiScore(ScoreC2 in RiboMethSeq), is calculated using the following formula : $RiboMethScore = 1- n_i/(0.5*(SUM(n_j*W_j)/SUM(W_j)+SUM(n_k*W_k)/SUM(W_k))$, where n_i – cumulated 5’/3’-end count for a given position, j – varies from i-2 to i-1, k varies from i+1 to i+2, Weight parameters are defined as 1.0 for -1 /+1 and 0.9 for -2/+2 positions. ScoreA and ScoreB were previously reported [433] and are described in more details in the Supplementary Material.

Mapping of yeast mRNA reads was done to reference collection of yeast CDS (ENSEMBL release R64-1-1 Saccharomyces_cerevisiae.R64-1-1.cdna.all.fa). Only uniquely mapped reads were taken into account for further analysis and calculation of protection scores.

Isolation and expansion of human bone marrow mesenchymal stem cells (BMMSCs)

MSCs were isolated from human bone marrow biopsy from patient during total hip arthroplasty after informed consent. This study was approved by our local ethics committee (Registration number DC-2014-2148). Bone marrow biopsy was heparined, diluted in Phosphate-Buffered Saline solution (PBS) and centrifuged at 1500 rpm for 5 min. The pellet was resuspended, seeded in Petri dishes at $4x10^6$ cells/dish and cultured at 37°C in a humidified atmosphere containing 5% (v/v) CO_2. The expansion phase was performed until passage P2. The expansion medium is composed of Dulbecco’s modified Eagle’s medium with low glucose (DMEM-LG, Gibco) supplemented with 10% fetal bovine serum (FBS, Sigma), 1 ng/ml basic fibroblast growth factor (bFGF, Miltenyi Biotec), 1% glutamine (Gibco) and 1% penicillin-streptomycin (Gibco).

Predifferentiation of BMMSCs in monolayer and production of cartilaginous TE substitutes

To favor chondrogenic differentiation, BMMSCs were cultured at passage 3 with predifferentiation medium until confluence before seeding to biomaterials (collagen sponges). This medium was composed of Dulbecco’s modified Eagle’s medium with high glucose (DMEM-HG, Gibco) supplemented with 10% fetal bovine serum (FBS, Sigma), sodium pyruvate (110 µg/ml, Gibco), bFGF (1 ng/ml, Miltenyi Biotech), 1% penicillin-streptomycin (Gibco), chondrogenic supplements (PAD: proline (40 µg/ml,

Sigma), L-ascorbic acid-2-phosphate (50 µg/ml, Sigma) and dexamethasone (0.1 µM, Sigma).

To study MSCs differentiation during cartilaginous TE substitutes production, BMMSCs were trypsinized at the end of the predifferentiation at passage 3 (trypsin-EDTA 0.05%, Gibco) and seeded into collagen sponges (Symatèse Biomatériaux, Chapanost, France) at a density of 0.5 million BMMSCs/sponge. These sponges measured 5 mm in diameter and 2 mm in thickness and were composed of 95% type I collagen and 5% type III collagen. MSCs seeded in collagen sponges were cultured at 37°C in a humidified atmosphere containing 5% CO2 (v/v). Half of the sponges were cultured with a control medium (ITS) that allowed MSCs survival but not chondrogenic or other differentiation. The other half of the sponges were cultured with chondrogenic medium (TGF-ß1). The control medium was composed of DMEM-HG supplemented with 1% ITS (ITS + premix, BD Biosciences), 1% glutamine (Gibco), sodium pyruvate (110 µg/ml, Gibco), 1% penicillin streptomycin (Gibco), and PAD. The chondrogenic medium was prepared with control medium supplemented with TGF-ß1 at 10 ng/ml (Miltenyi Biotech). At D7, D14, D21 and D28, analysis of cartilaginous matrix synthesis by BMMSCs inside collagen sponges was performed.

Analysis of cartilaginous matrix synthesis of TE substitutes

At D7, D14, D21 and D28, cartilaginous TE substitutes seeded with human MSCs were fixed (n=3 substitutes/time) with 4% paraformaldehyde during 24 hours and embedded in paraffin. 5 µm sections were stained with Alcian blue staining to visualize proteoglycan content. To visualize collagen II content, immunohistochemical analysis was performed as described previously [487]. The stained slides were observed and recorded by light microscopy (DMD108, Leica) at 4X original magnification. The scale bars represent 500 µm.

RESULTS

General overview of the HydraPsiSeq mapping technique

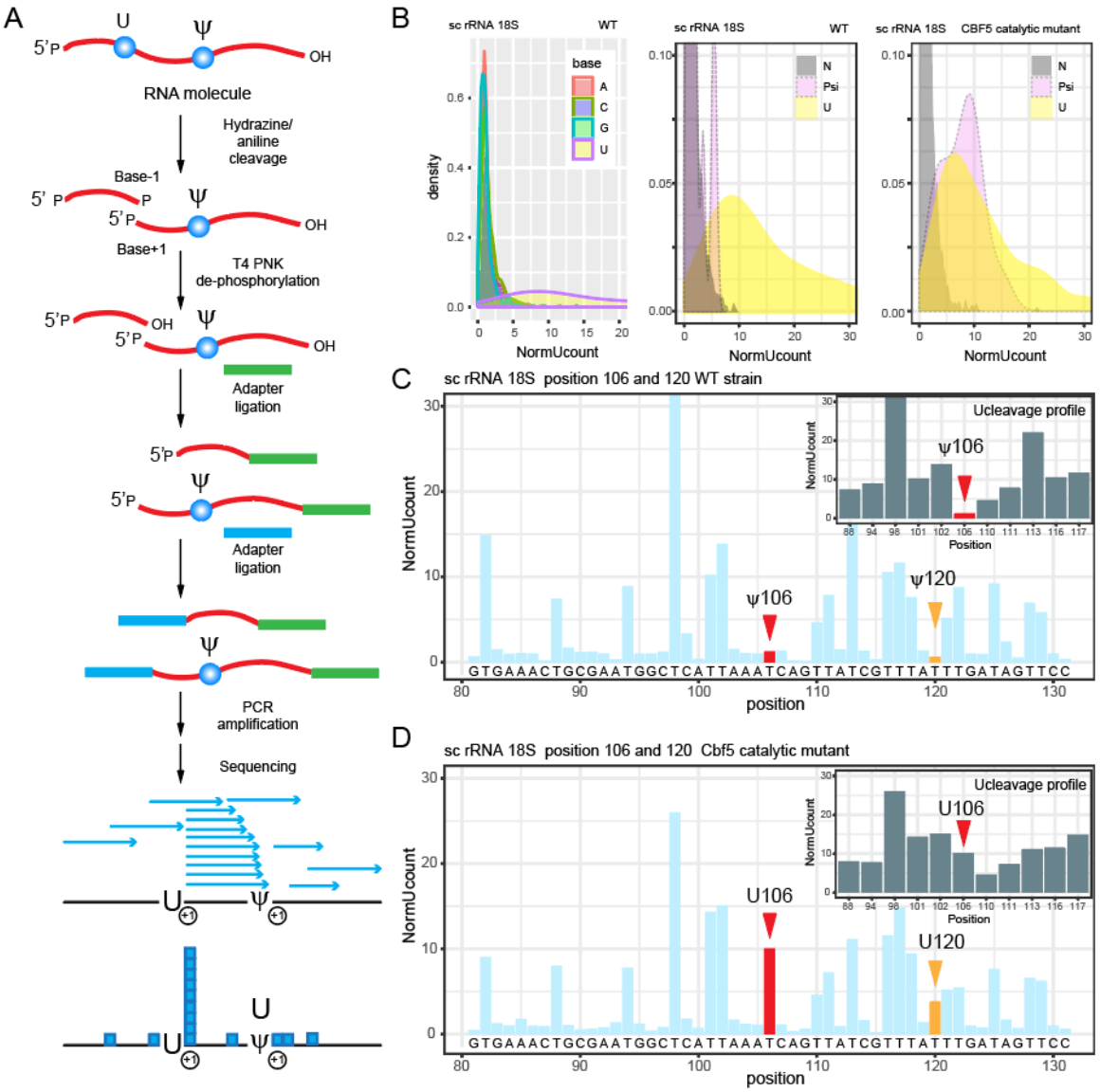


Figure 1: HydraPsiSeq protocol for Ψ mapping and quantification.

A – General overview of the protocol. RNA is treated by hydrazine and subjected to aniline cleavage. 3'-phosphates are removed by T4 PNK treatment and adapters are ligated to 3'-ends and 5'-ends of RNA fragments. After sequencing and alignment to a reference sequence, 5'-ends of all fragments are counted to generate Ucleavage profiles. U residues are sensitive to hydrazine and thus efficiently cleaved, while Ψ residues are resistant and provide only background signals. B – normalized U count for all 4 RNA nucleotides (A, C, G, U). Middle and right panels show NormUcount values for A/C/G nucleotides (N), U residues and Ψ residues, in WT and strain expressing Cbf5 catalytic mutant (no Cbf5 activity, rRNA is not pseudouridylated). C and D – cleavage profiles for position Ψ106/Ψ120 in 18S rRNA in WT yeast strain and in the strain expressing Cbf5 catalytic mutant. Modified positions are shown in red and orange. U-only profiles are shown in inserts.

The quantitative HydraPsiSeq protocol for pseudouridine mapping and quantification is based on random (statistical) cleavage at all uridine residues in RNA by combination of hydrazine treatment (ring-opening of the pyrimidine base in uridines) and aniline (cleavage of the resulting RNA abasic site and release of 5'-phosphorylated RNA fragments directly competent for adapter ligation at the N+1 nucleotide in the sequence, Figure 1A, Supplementary Figure S1). The canonical RNA bases (A, C, G), as well as pseudouridines, are known to be resistant to hydrazine cleavage and thus generate only background signals [488]. Decomposition of RNA abasic site [489] by consecutive □- and □-elimination generates two RNA fragments ending at the N-1 and starting at N+1 nucleotide to the U residue. A 3'-end dephosphorylation step using T4 PNK serves to remove all eventual 3'-phosphates [490] from upstream RNA fragment, and thus to ensure its compatibility with direct 3'-adapter ligation. Fragments produced by random cleavage at U residues are then selectively converted to a library and sequenced. Mapping to the reference sequence allows counting of such reads that start at the N+1 nucleotide position relative to U residues, while Ψ residues (as well as A, C and G), due to their relative resistance to hydrazine cleavage, give only background signals. These counts for A, C and G are used for local normalization of the U cleavages to background (NormUcount, see Materials and Methods), to account for local accessibility of different RNA regions. Further bioinformatic steps of analysis are similar to the well-established RiboMethSeq procedure [430], and allow calculation of scores termed ScoreMEAN, ScoreA and ScoreC for Ucleavage profiles, where Ψ residues appear as protected positions (similarly to Nm residues in RiboMethSeq).

Inspection of cleavage efficiencies in yeast rRNA (used here as initial model and containing 47 known Ψ sites, one each in 5S and 5.8S rRNAs, 14 in 18S rRNA and 31 in 25S rRNA [491]), demonstrated that, as expected, cleavage at A, C and G is relatively marginal (NormUcount <5), while cleavage at U residues is very efficient (NormUcount ranging from Zero to 50-60), see Figure 1B (left panel) for 18S rRNA data, similar profiles were also obtained for 25S rRNA. Technical and biological replicates demonstrated an excellent correlation of Ucleavages in different samples (Supplementary Figure S2). Deconvoluting the U's into positions which are known to be pseudouridylated and those which are not, the pseudouridines appear to display a profile similar to that of A, C and G (i.e. 'uncleaved/protected', Figure 1B, middle panel) and well distinct from U's (the pink and yellow peaks do not overlap). The peaks of Ψ cleavages in the area of 6-7 NormUcount units correspond to incompletely modified pseudouridylation sites. Further evidences that pseudouridylated positions are protected from RNA cleavage was obtained by analysis of rRNA extracted from yeast cells expressing a catalytically inactive form of Cbf5. Cbf5 is responsible for all rRNA pseudouridylations [483], except one in 5S rRNA catalyzed by Pus7 [492]. In this case, positions normally pseudouridylated behaved like unmodified Us (i.e. 'cleaved/unprotected', Figure 1B, right panel, the pink and yellow peaks largely overlap).

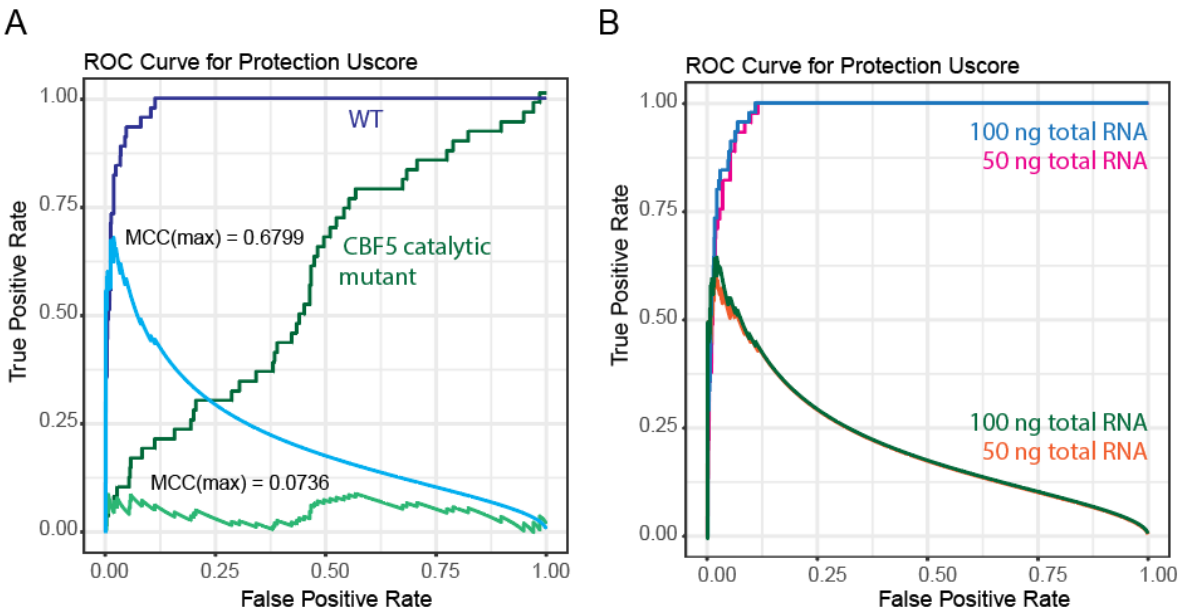
A typical cleavage profile observed for yeast 18S rRNA at position 106 and the corresponding Uprofiles (insert) are shown in Figure 1C and D. Cleavage of

pseudouridine at position 106 was very low in WT rRNA (Figure 1C), while in the absence of modification (in strain expressing Cbf5 catalytic mutant, Figure 1D), the signal was roughly equivalent to the neighboring unmodified Us in the sequence. Similar observation was done for other sites (see also position 120, Figure 1C and D).

Optimization of the hydrazine cleavage conditions

Conditions of hydrazine cleavage were optimized with respect to incubation temperature and time to achieve a fragment size compatible with sequencing and analysis (from ~10 to 30 nt) and to maximize signal-to-noise ratio. We found that the best results were obtained for hydrazine cleavage at 0°C for 45 min as longer incubation times were detrimental to read mapping during the alignment step. This is likely because of excessive amount of short reads mapped at multiple locations (Supplementary Figure S3). Under optimized hydrazine treatment conditions, cleavage of U residues remains rather heterogeneous, with NormUcount varying from 10-15 to >200. This may be related to particular sequence context or RNA structural effects. Even if RNA 2-D and 3-D structures have been described to be totally compromised in 50% w/w hydrazine solution [488], the use of additives, known to destabilize RNA 3-D and 2-D structures, such as urea/DMF or DMSO, led to accelerated hydrazine cleavage. However, these additives did not allow to improve the U-to-Ψ detection specificity or make cleavage more homogeneous (data not shown).

Pseudouridine detection and mapping



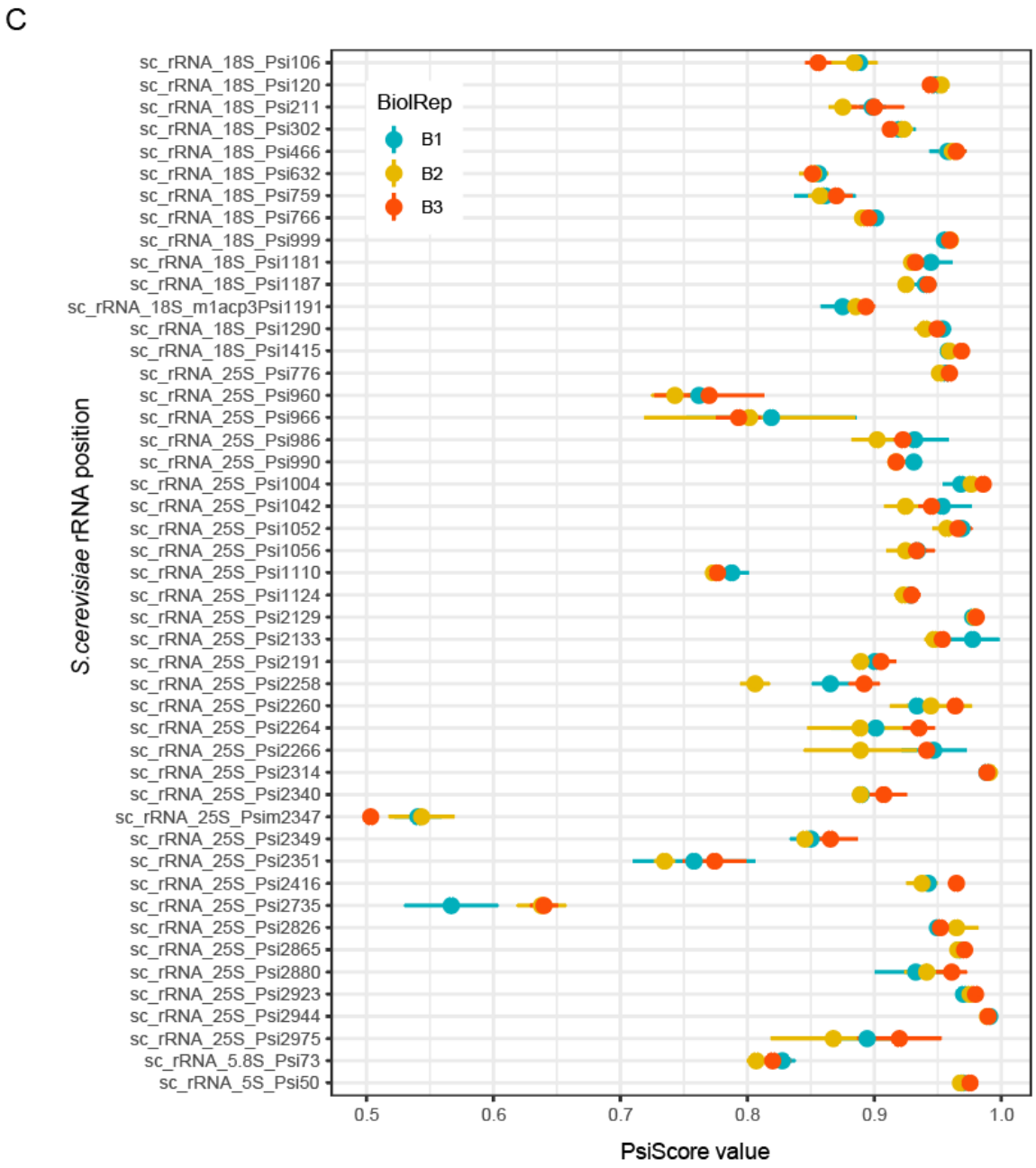


Figure 2: Pseudouridine detection in WT yeast rRNA.

A – Receiver Operating Characteristics (ROC) curve for Ψ detection using U protection score (dark blue). Maximal Matthews correlation coefficient (MCCmax) value is shown (light blue). The corresponding trace for unmodified rRNA from Cbf5 catalytic mutant is shown in dark green (MCC trace in light green). B – ROC curves for Ψ detection for samples with variable total RNA input. As low as 50 ng of input RNA (~40 ng of rRNA) is sufficient for good representativity and quantification of all rRNA Ψ sites. C – Dispersion of PsiScore values for technical and biological replicates. Analysis of S. cerevisiae rRNA pseudouridylation was done in three biological replicates (indicated in colors) with color bar representing dispersion of technical replicates.

The reliability of pseudouridine detection by HydraPsiSeq was evaluated by Receiver Operating Characteristic (ROC) curves and associated Matthews correlation coefficient (MCC) values for WT yeast rRNA containing 47 Ψ residues altogether [491], of which 46 are installed by Cbf5 as part of H/ACA snoRNPs [482], while formation of Ψ50 in 5S rRNA is catalyzed by the stand-alone enzyme Pus7 [492]. ROC curve for protection Uscore (defined as normalized protection of U residues from cleavage, see Materials and Methods) shows a reliable detection for the majority of pseudouridylation sites (Figure 2A). As anticipated, in the absence of pseudouridines in rRNA (isolated from a Cbf5 catalytic mutant), there is no distinction from U signals anymore.

Since the amount of biological material required for reliable detection and quantification is of great importance for analysis of precious biological samples, we attempted to reduce the amount of input rRNA from 300 ng, used in our initial tests, to 50 ng. Figure 2B demonstrates that even with such reduced amount of input RNA the detection of pseudouridines is highly reliable. Analysis of performance for different Ψ prediction/discrimination scores (protection Uscore, ScoreMEAN, ScoreA, ScoreB and PsiScore, equivalent to ScoreC) demonstrates that Uscore and PsiScore (equivalent to ScoreC in RiboMethSeq) provide the best discrimination power between true-positive and false-positive hits (Supplementary Figure S4). Analysis of global ROC curve for all yeast rRNAs shows that 25 out of the 29 most protected U residues correspond to known Ψ sites, and at the optimal MCCmax value, 35 sites are detected, with 13 false postitve (FP) hits. The majority of false negative (FN) identifications correspond to partially modified positions. Measurements of PsiScore demonstrated an excellent reproducibility between technical and biological replicates and showed that molar Ψ level varies from 0.8 to 1 for the majority of tested positions, only few sites were incompletely modified in yeast rRNA under normal growth conditions (complete glucose-based medium, Supplementary Figure S5). Correlation graphs (Supplementary Figures S2 and S5) also attest that hydrazine cleavage profiles were highly reproducible for all technical and biological replicates analyzed in this study. We found the average standard deviations to be of less than 5%, and the most variable positions in technical replicates only showed a maximum variation of 10-15%. Similar data were obtained for biological replicates as well. Thus, variations of PsiScore of > 10% (0.1 in molar Ψ amount per site) can be considered as significant and reproducible, opening the way for precise profiling of pseudouridine content.

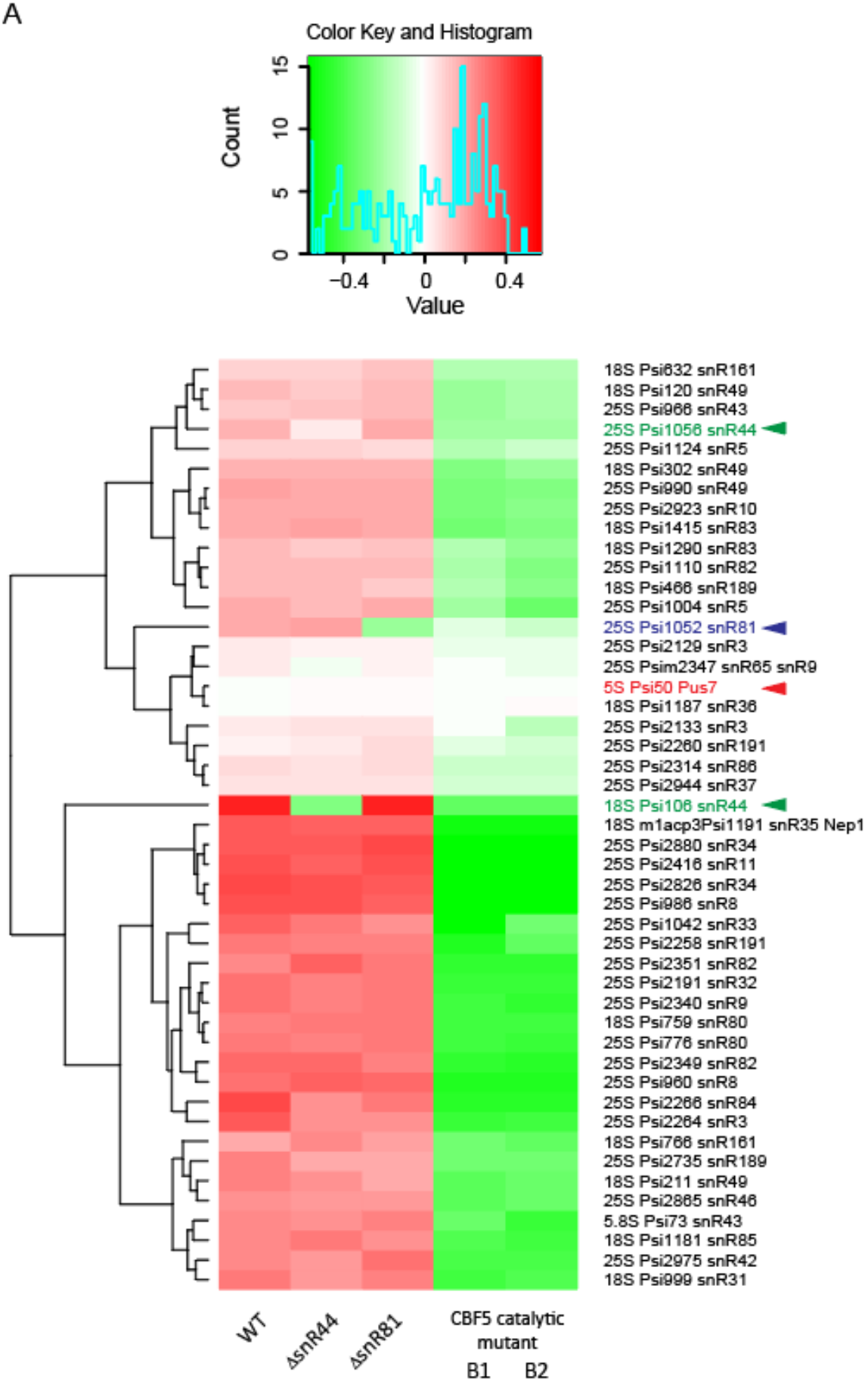
Comparison of the results obtained for different sequencing libraries demonstrated that 20-25 millions of reads were sufficient for an almost perfect coverage of all *S. cerevisiae* rRNA positions, thus this number of raw reads was used in further experiments.

Analysis of HydraPsiSeq signals in snoRNA-deleted S. cerevisiae strains and in yeast cells expressing inactive Cbf5

To further prove the reliability of HydraPsiSeq in Ψ detection and quantification, we performed analysis of two yeast strains bearing deletions of H/ACA snoRNAs snR44

or snR81. The snoRNA snR44 guides rRNA modifications at two positions (18S-Ψ106 and 25S-Ψ1056), while snR81 is responsible for modification of 25S-Ψ1052 [493](Supplementary Figure S6). We also included two biological replicates for a strain expressing an inactive Cbf5 catalytic mutant, and a time course for depletion of WT Cbf5 expressed under the control of a GAL promoter (See Supplementary figure S7,[482]). As shown in Figure 3, the only differences in PsiScore values between WT and KO snoRNA strains were precisely those located at positions (labelled in green and blue) guided by the deleted snoRNAs, and were similar to values obtained for unmodified rRNA observed in the Cbf5 catalytic mutant. It is noteworthy that the level of the Pus7-catalyzed 5S rRNA Ψ50 was not affected upon Cbf5 inactivation or depletion (labelled in red). The Cbf5 depletion experiment highlights that for the majority of Ψ sites, 48h-depletion gave roughly the same residual level of Ψ as observed with the Cbf5 catalytic mutant (Supplementary Figure S7).

A


Color Key and Histogram
Count
0
5
10
15
−0.4
0
0.4
Value
18S Psi632 snR161
18S Psi120 snR49
25S Psi966 snR43
25S Psi1056 snR44
25S Psi1124 snR5
18S Psi302 snR49
25S Psi990 snR49
25S Psi2923 snR10
18S Psi1415 snR83
18S Psi1290 snR83
25S Psi1110 snR82
18S Psi466 snR189
25S Psi1004 snR5
25S Psi1052 snR81
25S Psi2129 snR3
25S Psim2347 snR65 snR9
5S Psi50 Pus7
18S Psi1187 snR36
25S Psi2133 snR3
25S Psi2260 snR191
25S Psi2314 snR86
25S Psi2944 snR37
18S Psi106 snR44
18S m1acp3Psi1191 snR35 Nep1
25S Psi2880 snR34
25S Psi2416 snR11
25S Psi2826 snR34
25S Psi986 snR8
25S Psi1042 snR33
25S Psi2258 snR191
25S Psi2351 snR82
25S Psi2191 snR32
25S Psi2340 snR9
18S Psi759 snR80
25S Psi776 snR80
25S Psi2349 snR82
25S Psi960 snR8
25S Psi2266 snR84
25S Psi2264 snR3
18S Psi766 snR161
25S Psi2735 snR189
18S Psi211 snR49
25S Psi2865 snR46
5.8S Psi73 snR43
18S Psi1181 snR85
25S Psi2975 snR42
18S Psi999 snR31
WT
ΔsnR44
ΔsnR81
CBF5 catalytic mutant
B1
B2

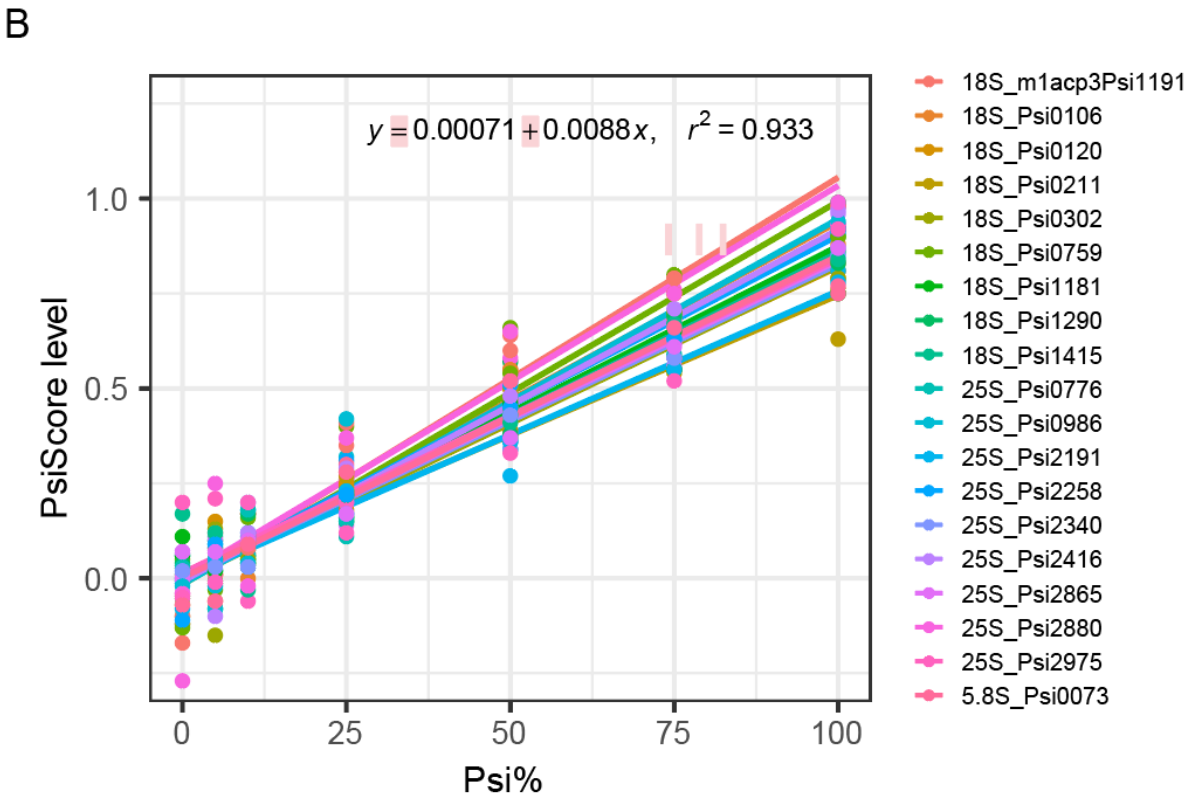


Figure 3

A – Differential heatmap for variations of PsiScore level for all known sited in S. cerevisiae rRNA. Normalization was done by the average score for a given position. SnoRNA independent Ψ50 in 5S rRNA is highlighted in red. Ψ106 and Ψ1056 guided by snR44 are highlighted in green. The identity of the S. cerevisiae strains used is provided at the bottom. Strain expressing Cbf5 catalytic mutant was analyzed in biological replicate. Identity of H/ACA snoRNA potentially guiding pseudouridylation is indicated together with the position. Clustering was done by hclust R function with ward.D2 algorithm. Color key for differential PsiScore is given in the insert. Visible variation of the signal for Ψ1187 is very minor upon CBF5 inactivation, but is related to drastic change of protection at the neighboring $m^1acp^3Ψ1191$, and thus perturbation in the shape of Uprofile, used to calculate PsiScore. B – Linear correlation of PsiScore with Ψ content of yeast rRNA. Calibration curve was produced using mixes of Cbf5cata and WT rRNAs, at ratio of 0, 5, 10, 25, 50, 75 and 100% of WT rRNA. HydraPsiSeq analysis was done in technical replicate, calibration curves for 19 non-clustered Ψ sites are shown.

Analysis of unmodified rRNA extracted from the catalytically inactive Cbf5 yeast strain (Cbf5cata) also provided a calibration of PsiScore signals to “a zero modification level”, thus allowing precise quantification of rRNA pseudouridylation. In order to demonstrate that PsiScore can be used for precise quantification of Ψ content, we performed calibration of PsiScore signals with a mixes of rRNA samples extracted from WT (modified) and Cbf5cata (unmodified) yeast strains. Analysis was done in technical duplicates using 0, 5, 10, 25, 50, 75 and 100% of modified WT rRNA in the mix. Calibration curves shown in Figure 3B and Supplementary Figure S8 demonstrates a very good linear correlation between Ψ content and observed

PsiScore for majority of yeast rRNA sites. Precise quantification is certainly one the most valuable feature of HydraPsiSeq protocol, since, by comparison, other deep sequencing-based methods cannot afford such analysis.

The applicability of HydraPsiSeq was also tested on yeast *S. cerevisiae* tRNAs. These molecules are highly pseudouridylated, and at least 7 tRNA:Ψ-synthases catalyze U to Ψ conversion in tRNAs [473, 494]. Although hydrazine cleavage profiles in tRNAs appeared less uniform than in rRNA, pseudouridinylated sites could be clearly identified by drastic changes in protection against hydrazine/aniline cleavage in null-mutants of the corresponding Pus-enzyme (see results in cells lacking PUS1, PUS4, and PUS7 in Supplementary Figure S9). Thus, while d*e novo* mapping of the pseudouridylated sites in tRNAs from WT strain was somewhat limited, combination with data from the corresponding knock-out strain allowed not only clear identification/confirmation of modified sites, but also to appreciate changes in protection and thus to quantify a modulation of individual Ψ levels.

In order to extend HydraPsiSeq applications to low abundant RNAs (such as mRNAs), we performed analysis of yeast *S. cerevisiae* mRNAs from WT, ΔPUS3 and ΔPUS7 strains. Only 150 ng of mRNA polyA+ preparation was engaged in this analysis (this amount could even be reduced to as low as 50 ng), demonstrating low requirements for input material. With moderate sequencing depth (100 mln reads/sample) sufficient coverage for analysis was obtained for >1200 yeast mRNAs. Yeast mRNA was reported to contain ~200-250 potential Ψ modification sites [314, 320, 495]. We were able to extract NormUcount values for ~70 of them and evaluate protection level for these residues. Compilation of these data allowed to confirm moderate to high protection score (UScore 0.5-10) for ~35 sites reported by two previous studies, and out of 14 sites reported to be common between two published datasets, HydraPsiSeq confirmed high protection level at 2 positions (Figure 4A).

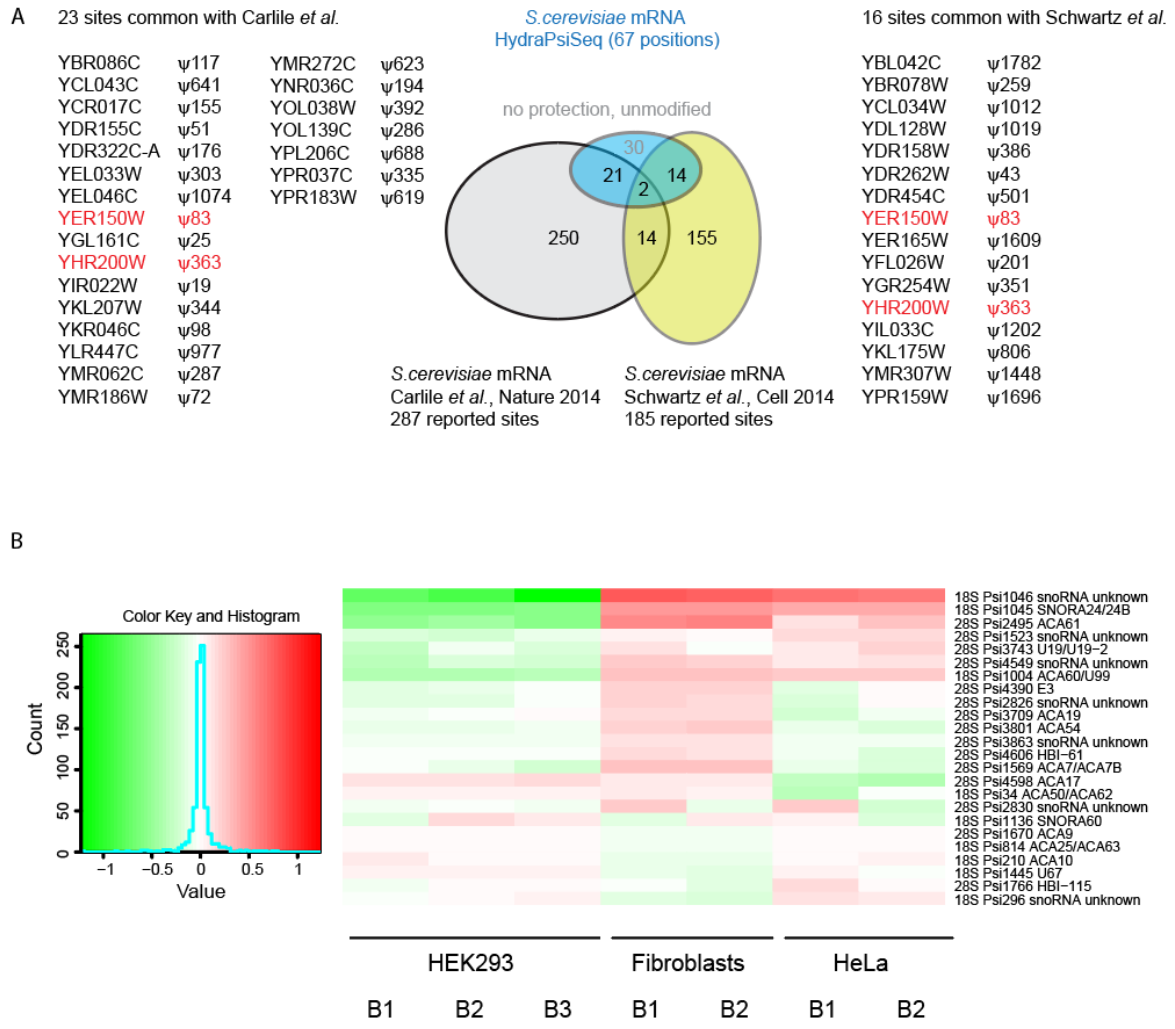


Figure 4: Profiling of yeast mRNA and human rRNA pseudouridylation by HydraPsiSeq.

A – Profiling of yeast S. cerevisiae mRNA modifications by HydraPsiSeq. Protection profiles with sufficient coverage were obtained for 67 positions previously reported to be pseudouridylated [314, 320]. Out of those, 23 showing moderate to high protection (UScores from 0.5 to 10) are common with the list from Carlile et al (2014) and 16 with Schwartz et al (2014), two sites (indicated in red) are common for all three datasets. B – Human cell lines (HEK293, fibroblasts and HeLa cells) were used in biological duplicate or triplicate (labeled B1/B2/B3). Profiling was done for all previously known human rRNA modifications (104 positions). Only 24 most variable positions are shown in Figure, full heatmap is given in the Supplementary Figure S10. For reference, the identity of the H/ACA snoRNA potentially guiding pseudouridylation is indicated together with the position. Clustering was done using hclust R function with ward.D2 algorithm. Color key for differential PsiScore is provided in the insert.

Profiling rRNA pseudouridylation in different human cell lines

Having successfully analyzed pseudouridine sites in yeast rRNAs, tRNAs, and mRNAs, we next applied HydraPsiSeq profiling to human rRNAs, containing altogether 104 reported Ψ sites [110].

The HydraPsiSeq profiling was performed in biological duplicate (or triplicate) for three commonly used human cell lines, namely: fibroblasts, kidney cells (HEK293 cells), and

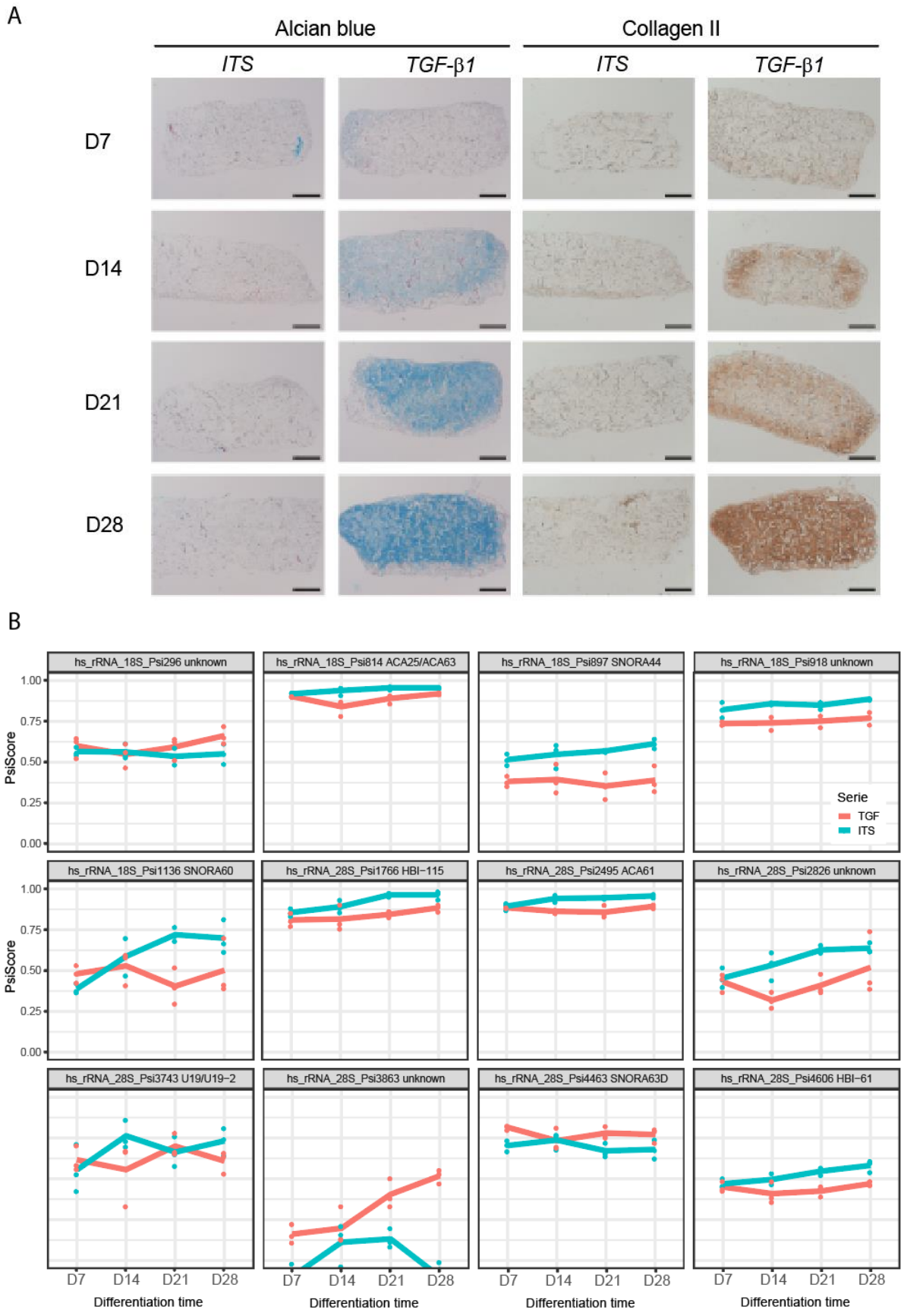
cervix cancer cells (HeLa). As shown in Figure 4B, HydraPsiSeq profiling revealed that about half of measured pseudouridylation sites (top part of the heatmap presented in Supplementary Figure S10) displayed distinctly different PsiScore levels in the three cell lines tested.

The finding that rRNA pseudouridylation levels differ from one cell line to another, underscores the presence of differentially modified populations of ribosomes. Such differential rRNA modification was previously observed for other modifications, such as 2'-O methylation [496, 497]. Most affected cell type-specific pseudouridylation sites (24 sites, see Figure 4B) were located both on 18S (9 sites) and 28S rRNAs (15 sites), and mostly clustered in the region 800-1150 for 18S rRNA and were relatively evenly distributed throughout 28S rRNA domains (4 in Domain I, 1 in Domain II, 6 in Domain III, 1 in Domain V –PTC and 3 in Domain VI). Differential pseudouridylation may reflect variable efficiency of the rRNA pseudouridylation machinery (as suggested previously for 2'-O methylation [497]), and/or a novel layer of regulation at the rRNA modification level.

Profiling rRNA pseudouridylation during human chondrogenic differentiation

To demonstrate further applications of HydraPsiSeq, we performed profiling of pseudouridine content in rRNA extracted from human bone marrow mesenchymal stem cells (BMMSCs) during their chondrogenic differentiation. Non-differentiated BMMSCs are capable to engage selective differentiation program, to become osteoblasts, chondrocytes or adipocytes and each of these programs requires specific expression of particular proteomes [498, 499]. Bone marrow stem cells were first extracted from clinical sample and chondrogenic differentiation was performed in type I/III collagen 3-D sponges. A comparison between differentiated and non-differentiated cells was embodied in a 3-D setting where the cells were either grown in the presence of TGF-□□, allowing the differentiation into chondrocytes, or in non-differentiating ITS medium, sustaining their growth but impeding differentiation [500, 501]. Chondrogenic differentiation of BMMSCs was assessed by staining with alcian blue, for detection of glycosaminoglycans, and by immunostaining for collagen II, which is a reliable chondrocyte marker (Figure 5A). In parallel, samples were collected at 4 time points (7-, 14-, 21- and 28 -days) and subjected to HydraPsiSeq analysis.

The majority of pseudouridylation sites remained constitutively modified during the time course of differentiation, both in non-differentiating ITS medium and during differentiation in TGF-□□-containing medium (Supplementary Figure S11), but specific positions displayed altered pseudouridylation levels between these two states (Figure 5B). Remarkably, this subset of variable pseudouridylation sites largely overlap with the set of cell type-specific modifications (see above and Figure 5C), and some of those residues have been shown to play an important role in ribosome functions (e.g. 28S-Ψ3743 [502]).


A
Alcian blue
Collagen II
ITS
TGF-β1
ITS
TGF-β1
D7
D14
D21
D28
B
hs_rRNA_18S_Psi296 unknown
hs_rRNA_18S_Psi814 ACA25/ACA63
hs_rRNA_18S_Psi897 SNORA44
hs_rRNA_18S_Psi918 unknown
hs_rRNA_18S_Psi1136 SNORA60
hs_rRNA_28S_Psi1766 HBI−115
hs_rRNA_28S_Psi2495 ACA61
hs_rRNA_28S_Psi2826 unknown
hs_rRNA_28S_Psi3743 U19/U19−2
hs_rRNA_28S_Psi3863 unknown
hs_rRNA_28S_Psi4463 SNORA63D
hs_rRNA_28S_Psi4606 HBI−61
PsiScore
1.00
0.75
0.50
0.25
0.00
Serie
TGF
ITS
D7
D14
D21
D28
Differentiation time

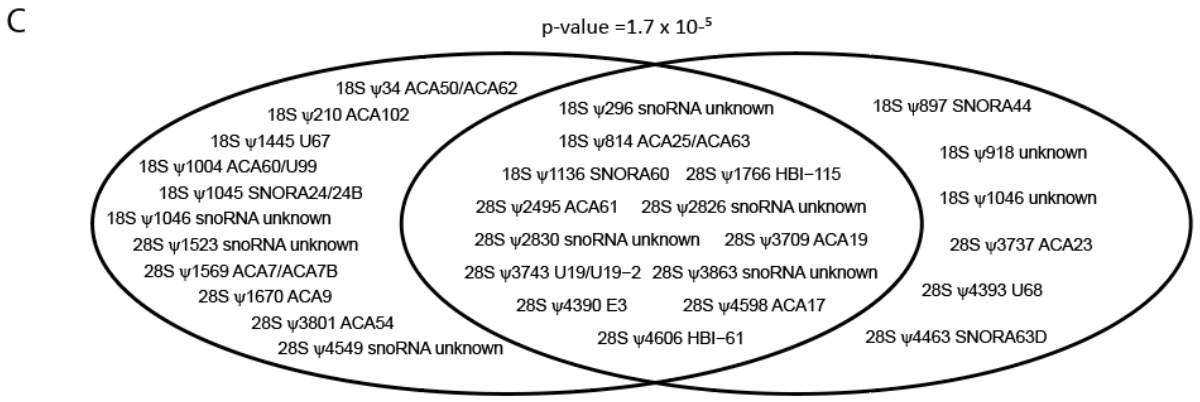


Figure 5: Modulation of human rRNA pseudouridylation profile during TGF-□1 stimulated differentiation of human bone marrow stem cell into chondrocyte-like cells.

A – accumulation of glycosaminoglycans revealed by alcian blue staining (blue, left) and immunostaining for collagen type II (brown, right). Samples were analyzed at days 7 (D7), 14 (D14), 21 (D21) and 28 (D28). The type of medium used is indicated at the top of each column. For simplicity, only one biological replicate is shown on Figure. B – Time courses of variations of PsiScore levels at selected pseudouridylated sites in human rRNA. Only highly variable positions are shown, other are presented in Supplementary Figure S11. Differentiation was followed for three independent biological replicates for the BMMSCs issued from the same patient. C – Human rRNA pseudouridylation sites showing variability between different cell lines and during BMMSCs differentiation. Venn diagram illustrates the overlap of two different datasets.

In conclusion, our analysis demonstrates that global pseudouridylation profile of human rRNAs may vary considerably between different cell lines, and throughout cell differentiation programs, supporting the emerging concept that cells produce different types of ribosomes, which may carry specialized function in order to adapt to specific translation requirements [503, 504].

DISCUSSION

Advantages and limitations of HydraPsiSeq

A unique feature of HydraPsiSeq, by comparison to previously described high throughput methods, is that it provides for the first time quantitative information about the amounts of pseudouridylation at any particular position of an RNA molecule with nucleotide precision. HydraPsiSeq does not involve the selective enrichment of Ψ-modified RNAs, and as such it is particularly efficient on abundant cellular RNAs, such as rRNAs and tRNAs. Nonetheless, we have shown that it is also efficient on lower abundant RNAs, such as yeast mRNAs. In that respect, HydraPsiSeq is similar to RiboMethSeq, which allows the systematic mapping of 2'-O methylated sites [64]. By comparison, AlkAniline-Seq [360], for detection of m^7G, and other base modifications, includes the selective enrichment of the modified sites [31]. The absence of an enrichment step in HydraPsiSeq implies that a substantial coverage at all U positions in a given RNA sequence must be achieved (sequencing depth of 20-25 millions of reads for eukaryotic rRNA, see below), but, in return, HydraPsiSeq allows highly

sensitive, reliable and precise quantification of Ψ residues in RNA and, with inclusion of a calibration of the PsiScore signal to natural unmodified RNA or synthetic transcript, even to reach an absolute measurement of the modification level. Since the main advantage of this new protocol over previously published CMC-based techniques [314, 320, 454] resides in quantification of pseudouridines, we foresee quantitative profiling of eukaryotic rRNA as the main application, together with analysis of tRNA modification dynamics under different cell growth and physiological or pathological conditions.
Application for analysis of rare RNAs, such as mRNA or lncRNA is attainable, but mostly for validation purposes, and presumably only for the most abundant molecules, since otherwise quite substantial sequencing depth would be required for significant analysis at the transcriptome-wide level which would increase costs. This limitation of HydraPsiSeq in analysis of moderately abundant mRNAs can be overcome, providing that an appropriate enrichment step is included, e.g. through the development of a Ψ-specific antibody for IP-like enrichment [466].

Required amount of input material and sequencing depth

In our optimized protocol, HydraPsiSeq requires about 50 ng of total RNA for analysis of rRNA modifications, or equivalent amount of enriched RNA fraction for analysis of low abundant RNAs. With some, but still acceptable, loss in efficiency in the library preparation step, this amount can be even further reduced to only ~10-20 ng. This exceptional sensitivity makes HydraPsiSeq compatible with the great majority of fundamental studies on RNA modifications, and also for large-scale screening of clinical sample collections. It is noteworthy that, except for the Illumina-based RiboMethSeq protocol, which requires 10 ng of RNA input at the minimum [64], all other reported deep-sequencing methods for analysis of RNA modifications require at least µg (or even mg) amounts of input RNA [65]. Also of note is that, in principle, the detection of protected Ψ residues by comparison to cleaved U residues (negative detection) requires much higher sequencing depth than "positive" detection methods (e.g. CMC-based protocols). Our optimized protocol provides robust coverage of all U residues in human rRNA and thus reliable detection and quantification of all pseudouridines in these species, with 20-25 millions raw reads available for analysis (2,000-3,000 reads/position). This sequencing depth is a bit higher than that required in other common protocols (e.g., RiboMethSeq), but remains relatively reasonable, since no costly selection steps are necessary, and thus the overall cost of library preparation remains rather limited.

Reproducibility and precision of pseudouridine quantification

Analysis of technical and biological replicates for yeast (Supplementary Figure S5) and human rRNA (Figure 4B) demonstrated a very low dispersion of measured protection Uscores and derived PsiScores used for quantification. Technical variability was found to be higher for partially modified rRNA positions and inversely correlated with sequencing coverage in the region used for calculation of scores. In conclusion, HydraPsiSeq now provides quantification of RNA pseudouridylation with precision at

least equivalent to mass spectrometry approaches [458, 459], but the use of deep sequencing allows accommodation of larger series of samples even when only minute amounts of RNA are available.

Known false-positive/negative identifications

Random and uniform cleavage of U residues by hydrazine is the pre-requisite for successful analysis of hydrazine-resistant pseudouridines by HydraPsiSeq. Analysis of yeast and human rRNAs demonstrated that, most likely due to residual 2-D and/or 3-D RNA structures under hydrazine treatment at low temperature (0°C), a minor subpopulation of inaccessible U residues pertains and displays only very limited cleavage (Uscore). Thus such sites may be interpreted as false-positive hits. Another known source of false-positive hits is related to resistance of 5-modified U residues (mostly m^5U, but also mo^5U, etc.) to hydrazine cleavage, thus such residues are also detected as "protected" U's and can give false-positive identification. However, compared to pseudouridines, which are abundantly present in many different RNAs, these modifications are relatively rare (they are mostly found in tRNAs). Dihydrouridines (D) are most probably resistant to hydrazine themselves, but finally cleaved by the combination of alkaline conditions under hydrazine treatment and aniline scission, and thus are not revealed as false-positive protected hits in HydraPsiSeq (data not shown). Reciprocally, some rare non-U modified residues in rRNAs and tRNAs displayed an unexpectedly high sensitivity to cleavage under the conditions used in HydraPsiSeq (e.g. m^7G and m^3C), but this did not affect Ψ detection or quantification.

Constitutive and variable pseudouridylation sites in human rRNAs in cancer and cell differentiation programs

Profiling of different human cell lines and stem cells during chondrogenic differentiation demonstrated that the great majority of rRNA pseudouridines is constitutively modified and hardly show any variation. These represent over 80% of all known modified sites. Remarkably, however, a subset of rRNA positions display variable modification levels, and the exact profile appears to be specific for a given cell line or cell differentiation step (Figure 5C). Similarly, only few positions vary during stem cells differentiation, and show dependence towards growth medium. The exact reasons and molecular mechanisms underlying such variability are not yet known, but one can speculate that these differences reflect adaptation of eukaryotic ribosome to translation of given mRNA species required to produce essential proteins, or protein isoforms, in particular conditions. The existence of variable pseudouridylated positions is to be interpreted in light of the observation that snoRNA H/ACA-dependent rRNA pseudouridylation [505] is strongly connected to different pathologies and development [506–508]. Notably, it was reported that rRNA pseudouridylation affects the balance between cap-dependent and IRES-driven initiation of mRNA translation, and that it affects translation fidelity [484]. Profiling of rRNA (and tRNA) pseudouridylation under normal and stress conditions

using HydraPsiSeq will undoubtedly help addressing these and other fundamental biological questions in the future.

SUPPLEMENTARY DATA
Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT
We thank members of IMoPA Team RNA-RNP, Florence SCHLOTTER, Charbel ALFEGHALY and Lisa HETTAL for kind gift of human cells (HeLa, HEK293 and skin fibroblasts) for RNA extraction.

FUNDING
Y.M. is supported by Epi-ARN FRCR project from Grand Est Region, M.H. by Deutsche Forschungsgemeinschaft (DFG) grant HE3397/17-1 and RO 4681/6-1, as well as HE 3397/13-2 in the SPP1784. Research in the Lab of D.L.J.L. is supported by the Belgian Fonds de la Recherche Scientifique (F.R.S./FNRS) [“RiboEurope” European Joint Programme on Rare Diseases (EJP RD/JTC2019/PINT-MULTI) grant n°R.8015.19 and PDR grant n°T.0144.20], the Université Libre de Bruxelles (ULB), the Région Wallonne (SPW EER) [“RIBOcancer” FSO grant n°1810070], the Fonds Jean Brachet, the Internationale Brachet Stiftung. Work was perfumed frame of the EPITRAN COST initiative (CA16120).

AUTHOR CONTRIBUTION
V.M., M.H. and Y.M. designed HydraPsiSeq experiments, performed analysis and interpreted the data. F.P contributed to the development of bioinformatic analysis pipeline. P.N. and A.P. provided BMMSCs differentiation model, L.A. and V.BI. performed total RNA extraction, library preparation and deep sequencing, L.W. and D.LJL. supplied yeast Cbf5 depleted strains and performed depletion, V.M., D.L.J.L., M.H. and Y.M. analyzed data and wrote the paper.

COMPETING FINANCIAL INTERESTS
The authors declare no conflict of interests.

Supplementary material

Description of scores used in HydraPsiSeq

Since HydraPsiSeq measures protection of modified Ψ residues to chemical cleavage, scoring system used in the method is conceptually similar to scores used in RiboMethSeq protocol [433]. However, instead of cleavage profile for all RNA nucleotides used in RiboMethSeq, only U-cleavage profiles are used to measure protection. Thus, all described scores are calculated for U-cleavage profiles as raw data. For all scores, normalization window is 4 nt (+/-2 nt to the position) [509].

Score A = $$1-\frac{2n_i+1}{½\left|\mu_l-\sigma_l\right|+n_i+½\left|\mu_r-\sigma_r\right|+1}$$

Where n_i is cumulated 5'/3'-end read count at position i, μ_l and μ_r are mean values for left and right neighboring nucleotides (+/-2 nt) and σ_l and σ_r are corresponding standard deviations.

Score B = $$\frac{\left|n_i-½\left(\frac{\sum_{j=i-\delta}^{i-1}\omega_j n_j}{\sum_{j=i-\delta}^{i-1}\omega_j}+\frac{\sum_{j=i+1}^{i+\delta}\omega_j n_j}{\sum_{j=i+1}^{i+\delta}\omega_j}\right)\right|}{n_i+1}$$

Where n_i is cumulated 5'/3'-end read count at position i, δ =2, ω is 1 for pos -1 and +1, and 0.9 for pos -2 and +2

PsiScore(Score C) = $$1-\frac{n_i}{½\left(\frac{\sum_{j=i-\delta}^{i-1}\omega_j n_j}{\sum_{j=i-\delta}^{i-1}\omega_j}+\frac{\sum_{j=i+1}^{i+\delta}\omega_j n_j}{\sum_{j=i-1}^{i-\delta}\omega_j}\right)}$$

Where n_i is cumulated 5'/3'-end read count at position i, δ =2, ω is 1 for pos -1 and +1, and 0.9 for pos -2 and +2

ScoreMEAN

$$r_i=\frac{n_i}{n_{i-1}} \qquad f_i=\frac{n_i}{n_{i+1}}$$

$$ScoreMEAN=\frac{\left(1-\frac{r_i}{\frac{\sum_{j=i-\delta}^{i-1}r_j}{\delta}+\frac{\sum_{j=i+1}^{i+\delta}r_j}{\delta}}\right)+\left(1-\frac{f_i}{\frac{\sum_{j=i-\delta}^{i-1}f_j}{\delta}+\frac{\sum_{j=i+1}^{i+\delta}f_j}{\delta}}\right)}{2}$$

Where n_i is cumulated 5'/3'-end read count at position i, δ =2

All scripts used in this work are available at GitHub https://github.com/FlorianPichot/HydraPsiSeqPipeline

Supplementary Table S1 Yeast strains used in this study:

n°	Name	Genotype	Source
1	BY4741(WT)	MAT a; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0	EUROSCARF
2	ΔPUS1	MAT a; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; YPL212c::kanMX4	EUROSCARF
3	ΔPUS3	MAT a; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; YFL001w::kanMX4	EUROSCARF
4	ΔPUS4	MAT a; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; YNL292w::kanMX4	EUROSCARF
5	ΔPUS7	MAT a; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; YOR243c::kanMX4	EUROSCARF
6	ΔsnR44	MAT a; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; YLR367w::kanMX4	EUROSCARF
7	ΔsnR81	MAT a; ura3Δ0; leu2Δ0; his3Δ1; met15Δ0; SNR81::kanMX4	This study
8	WT to Cbf5-cata (YDL2931)	ade2 his3 leu2 ura3 cbf5::TRP1 + p[CBF5-ARS/CEN-HIS3]	D. Lafontaine's Lab
9	Cbf5-cata mutant (YDL2932)	ade2 his3 leu2 ura3 cbf5::TRP1 + p[cbf5-D95A-ARS/CEN-HIS3]	D. Lafontaine's Lab
10	GAL::cbf5 (YDL521-1)	Mat a; ura3-52; trp- (trp1Δ63 or trp1-289); leu2Δ1; gal2; galΔ108; HIS3-pGAL10::cbf5	D. Lafontaine's Lab

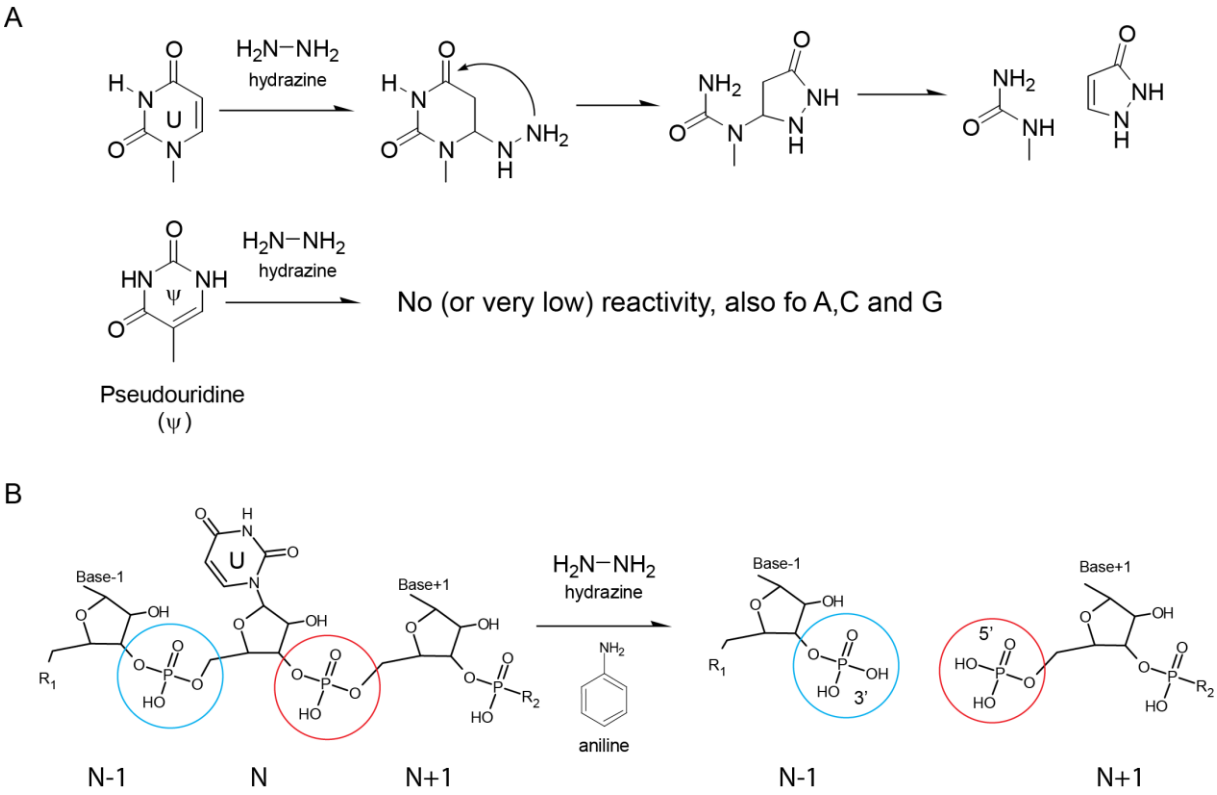


Figure S1 Chemistry of hydrazine/aniline treatment for random cleavage of U residues in RNA.

A: Uridine (U) residues are readily cleaved by hydrazine (H_2N-NH_2) with subsequent formation of RNA abasic site, while pseudouridine (Ψ) is relatively resistant to this cleavage.

B: RNA strand scission by consecutive hydrazine and aniline cleavage. 5'-phosphate at N+1 nucleotide, resulting from aniline-driven decomposition of RNA abasic site and used for 5'-adapter ligation, is circled in red. The 3'-phosphate at N-1 nucleotide (circled in blue) is removed by T4 PNK 3'-phosphatase activity (in the absence of ATP), without simultaneous removal of 5'-phosphate.

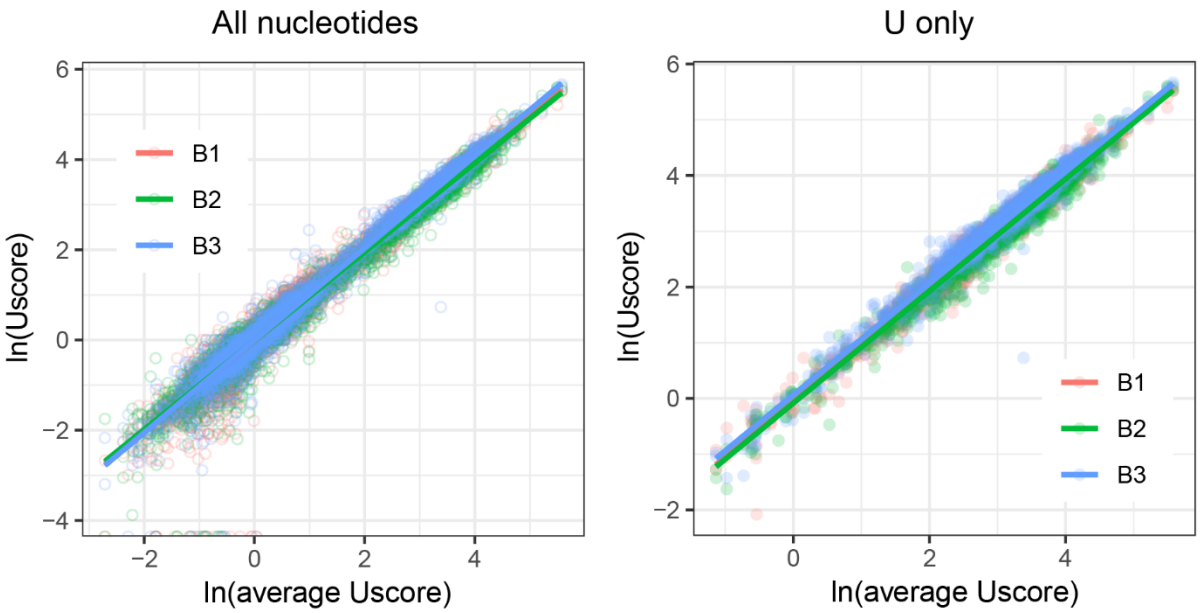


Figure S2 Reproducibility of technical and biological replicates.
Correlation of Uscores (protection values) for all nucleotides in yeast rRNA by biological replicate (**left**) and for U residues only (**right**), traced in ln scale to the average. Colors represent biological replicates (B1, B2 and B3).

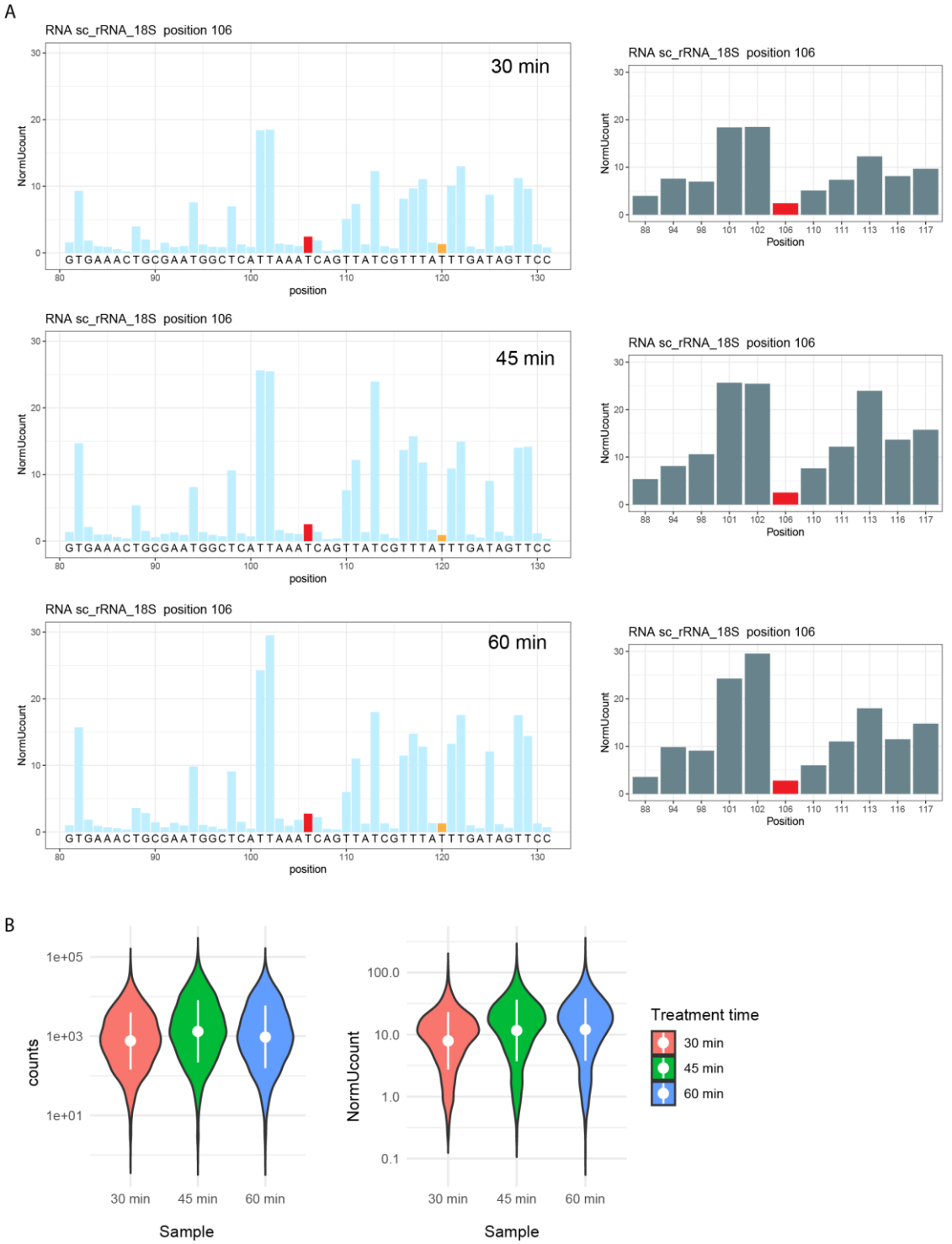
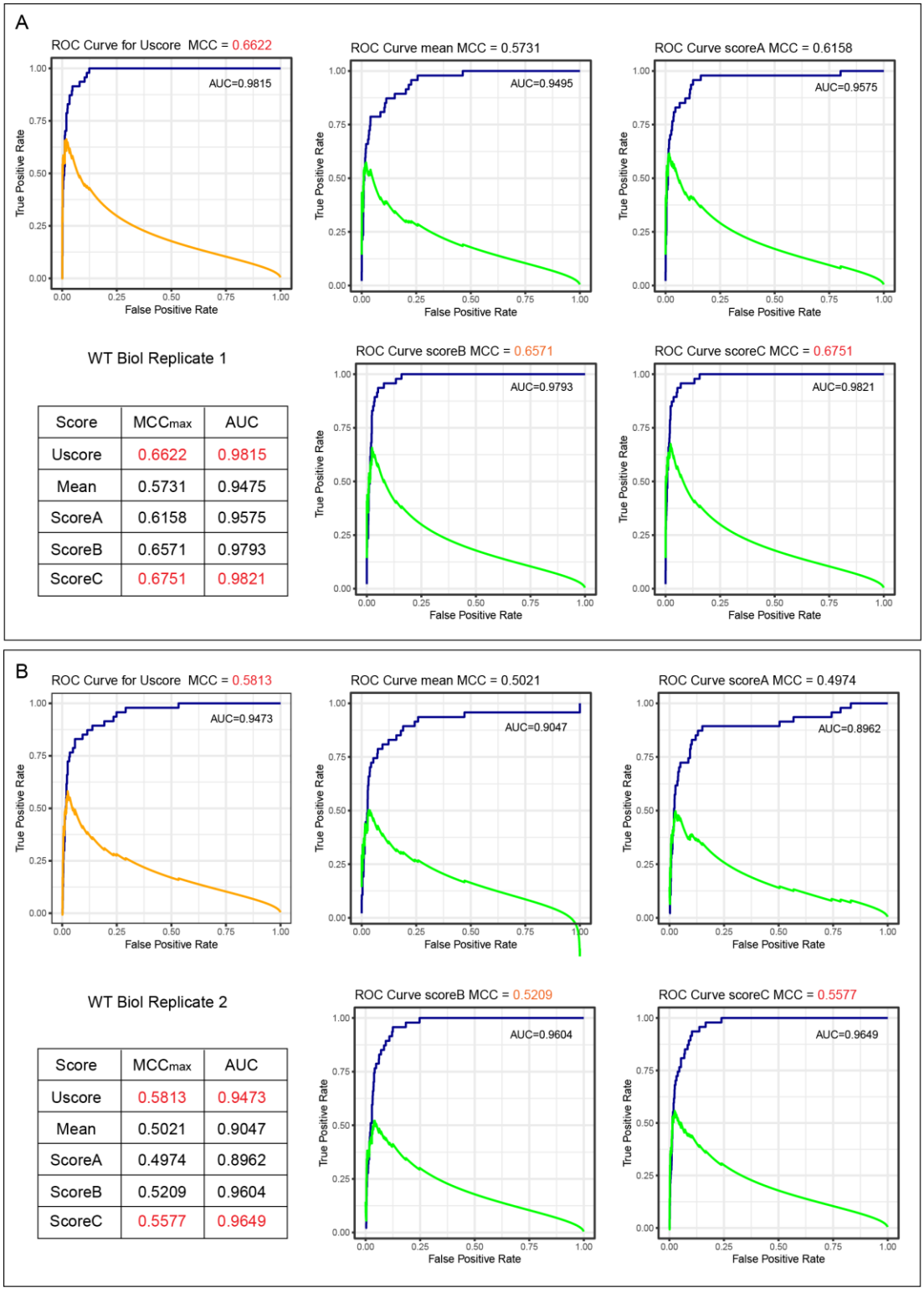


Figure S3 Optimization of hydrazine treatment to get the best compromise between mappability of RNA reads and specificity of pseudouridine detection.
A: Total and NormUcount profiles for the region spanning positions 106 and 120 in yeast 18S rRNA are shown. Position of Ψ106 is in red, neighboring position Ψ120 is in orange.
B: Variation of total number for mapped counts (left) and values of NormUcount for different treatment times. 45 min provides the best compromise and thus was used for further experiments.



Score	MCC_{max}	AUC
Uscore	0.6622	0.9815
Mean	0.5731	0.9475
ScoreA	0.6158	0.9575
ScoreB	0.6571	0.9793
ScoreC	0.6751	0.9821

Score	MCC_{max}	AUC
Uscore	0.5813	0.9473
Mean	0.5021	0.9047
ScoreA	0.4974	0.8962
ScoreB	0.5209	0.9604
ScoreC	0.5577	0.9649

Figure S4 Performance of HydraPsiSeq scores for detection of pseudouridine residues.

Yeast *S. cerevisiae* rRNA in two biological replicates (panels **A** and **B**) was used as a model. ROC curves are in dark blue, the corresponding MCC traces in orange and green. Table shows maximum values of MCC and AUC for each scoring scheme. The best discrimination values are given in red (Uscore and scoreC=PsiScore)

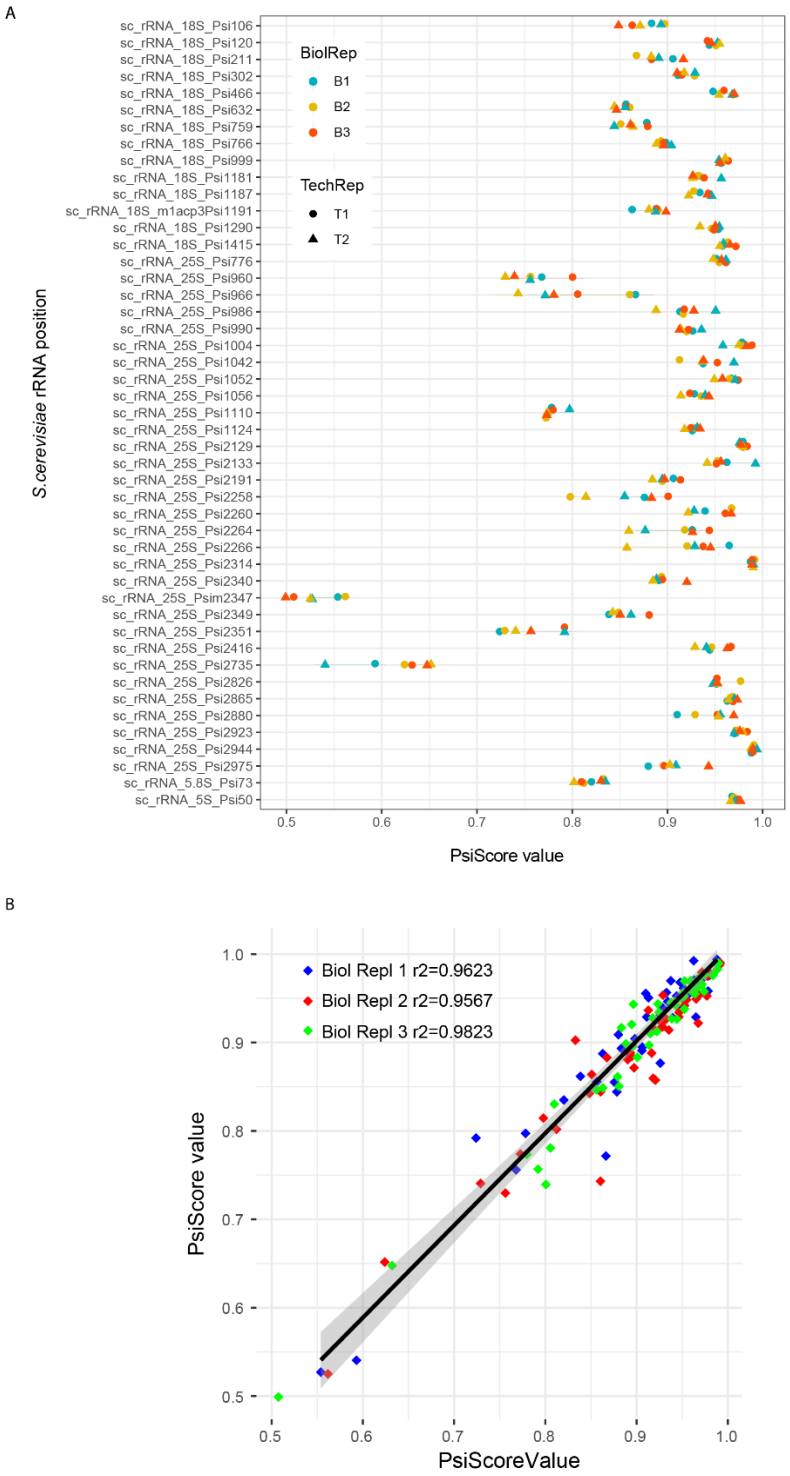


Figure S5 Analysis of technical and biological replicates for known pseudouridylated sites in *S. cerevisiae* rRNA.

A: PsiScore values are given for each site (identity is shown on the left). Colors indicate biological replicates (independent yeast cell cultures), shape of the point – technical replicates (independent library preparation/sequencing/analysis for the same total RNA sample).

B: PsiScore values for modified positions only are traced in function of average value for all replicates. Colors of the points represent biological replicates.

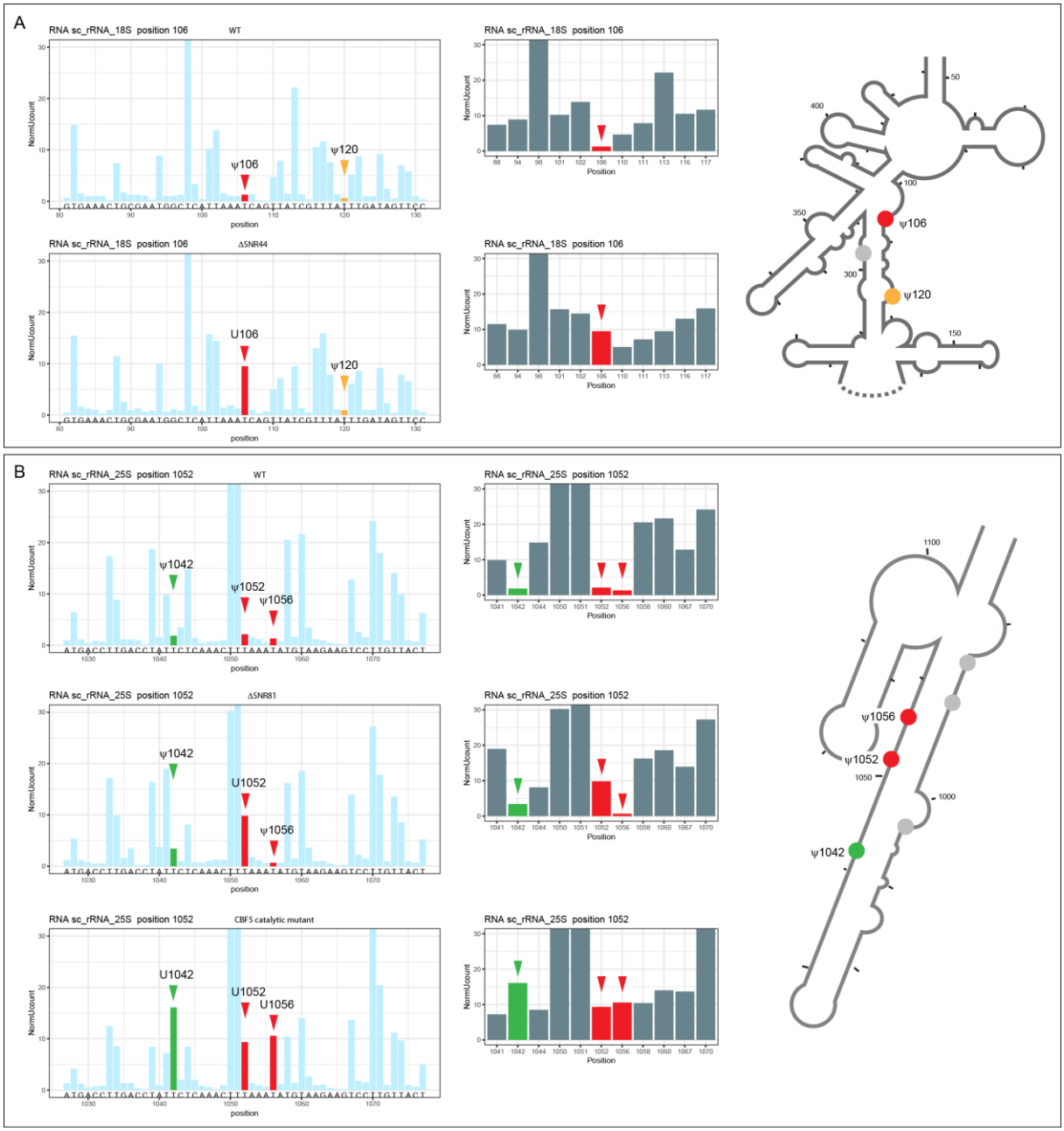


Figure S6 Analysis of pseudouridylation in yeast H/ACA snoRNA mutants.
A: Full and U protection profiles for position 18S-Ψ106 in WT (top) and snr44Δ strains. Position of Ψ106 is indicated in red, position 120 in orange, other pseudouridylated positions in grey.
B: Full and U protection profiles for positions 25S-Ψ1042, 25S-Ψ1052 and 25S-Ψ1056 in WT (top), snr81Δ strains (middle) and CBF5-cata mutant strains (bottom). Positions of Ψ1052 and Ψ1056 are indicated in red, Ψ1042 is in green, other pseudouridylated positions in grey. Only Ψ1052 is guided by SNR81 and all three residues are missing in CBF5-cata mutant strain.

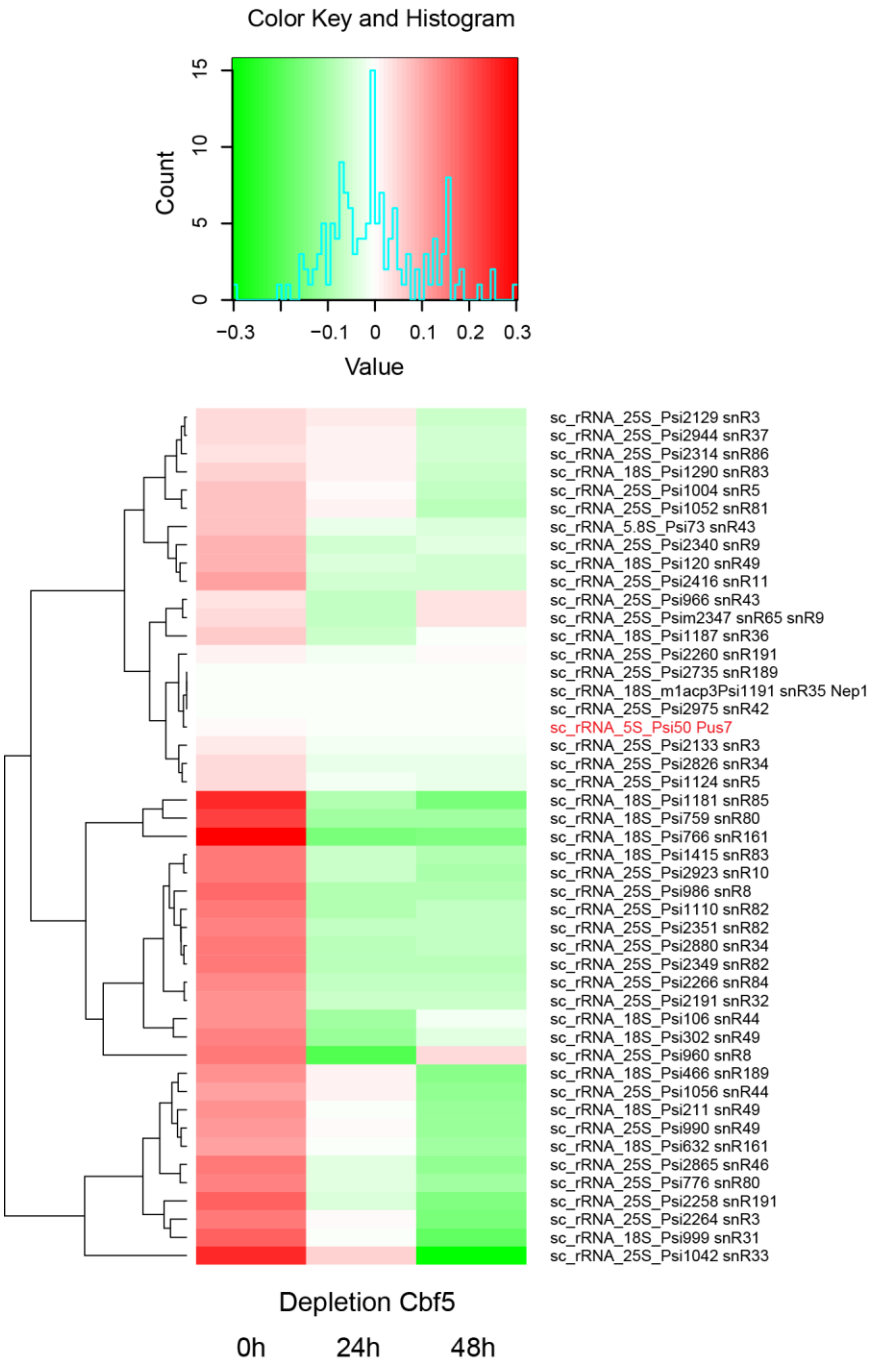
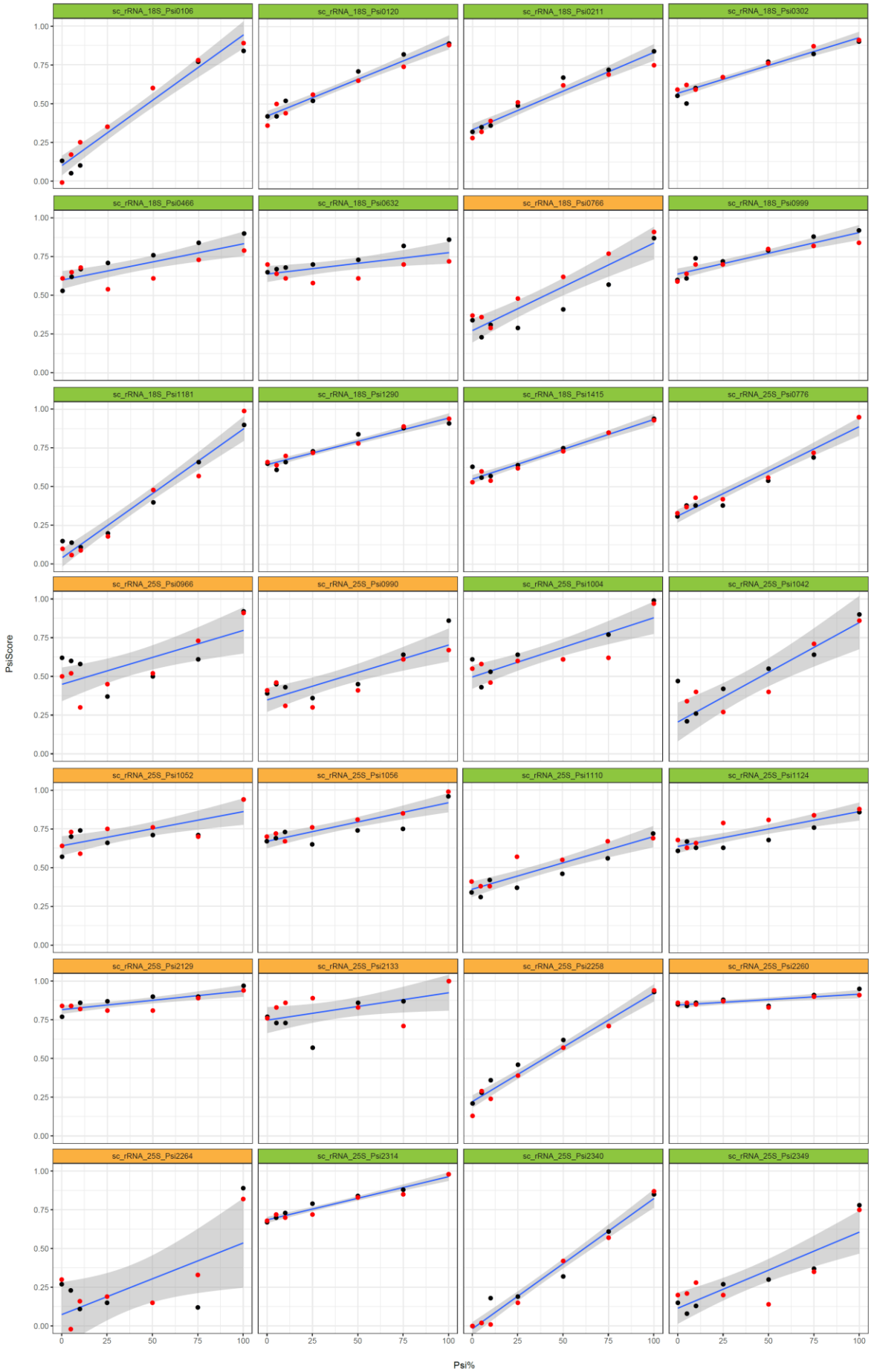


Figure S7 Progressive reduction of the rRNA pseudouridylation level in Cbf5-depleted yeast *S. cerevisiae* strain. In pGAL::cbf5 cells, CBF5 is expressed on galactose-based medium but not on glucose based medium. Depletion was performed by shifting cells to glucose-containing medium (YPD) and analysis was performed at 0h, 24h and 48h after shift. Cbf5-independent position in 5S rRNA (Ψ50) is shown in red.


sc_rRNA_18S_Psi0106
sc_rRNA_18S_Psi0120
sc_rRNA_18S_Psi0211
sc_rRNA_18S_Psi0302
sc_rRNA_18S_Psi0466
sc_rRNA_18S_Psi0632
sc_rRNA_18S_Psi0766
sc_rRNA_18S_Psi0999
sc_rRNA_18S_Psi1181
sc_rRNA_18S_Psi1290
sc_rRNA_18S_Psi1415
sc_rRNA_25S_Psi0776
sc_rRNA_25S_Psi0966
sc_rRNA_25S_Psi0990
sc_rRNA_25S_Psi1004
sc_rRNA_25S_Psi1042
sc_rRNA_25S_Psi1052
sc_rRNA_25S_Psi1056
sc_rRNA_25S_Psi1110
sc_rRNA_25S_Psi1124
sc_rRNA_25S_Psi2129
sc_rRNA_25S_Psi2133
sc_rRNA_25S_Psi2258
sc_rRNA_25S_Psi2260
sc_rRNA_25S_Psi2264
sc_rRNA_25S_Psi2314
sc_rRNA_25S_Psi2340
sc_rRNA_25S_Psi2349
PsiScore
Psi%
0.00
0.25
0.50
0.75
1.00
0
25
50
75
100

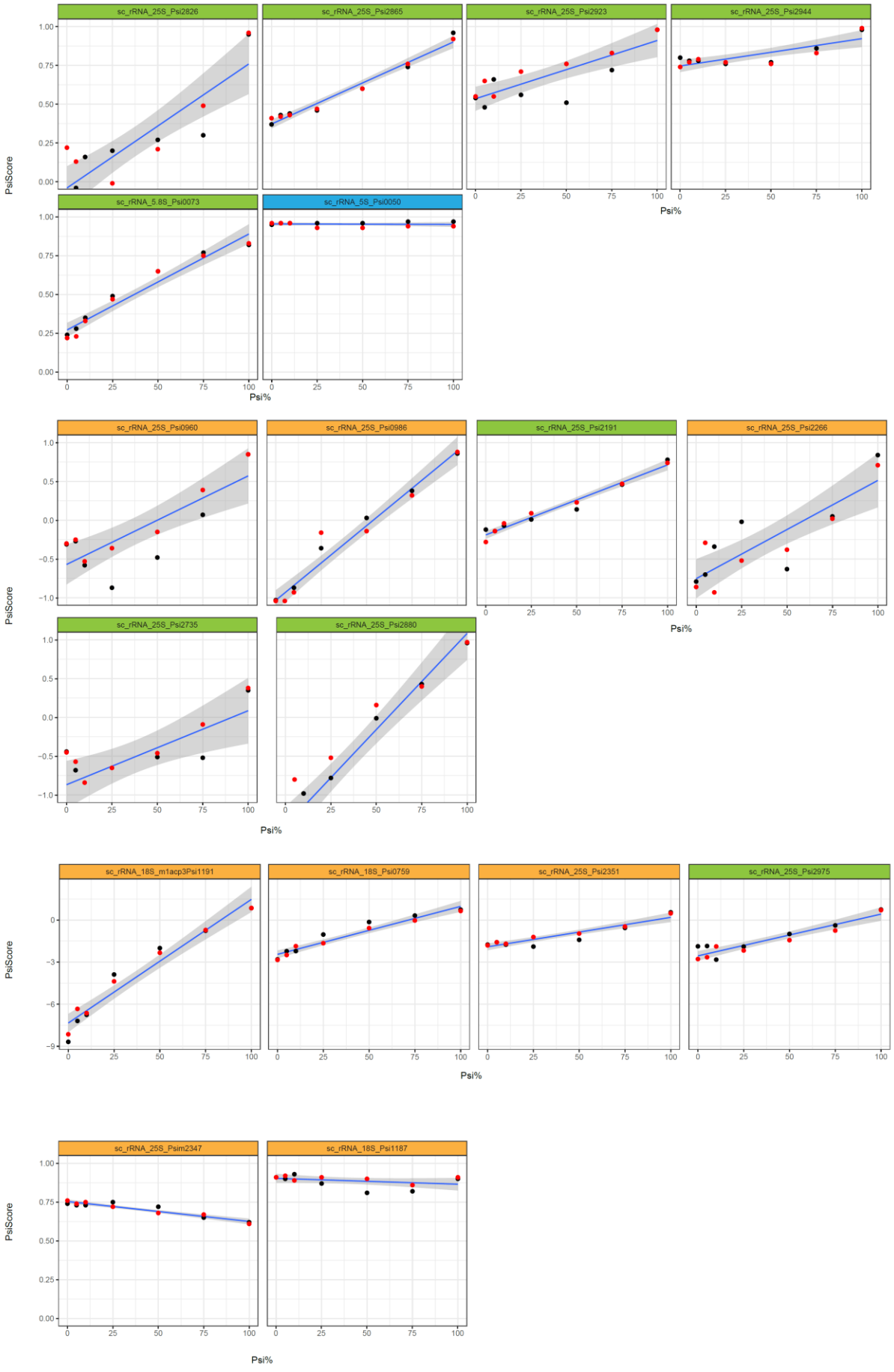

sc_rRNA_25S_Psi2826
sc_rRNA_25S_Psi2865
sc_rRNA_25S_Psi2923
sc_rRNA_25S_Psi2944
sc_rRNA_5.8S_Psi0073
sc_rRNA_5S_Psi0050
sc_rRNA_25S_Psi0960
sc_rRNA_25S_Psi0986
sc_rRNA_25S_Psi2191
sc_rRNA_25S_Psi2266
sc_rRNA_25S_Psi2735
sc_rRNA_25S_Psi2880
sc_rRNA_18S_m1acp3Psi1191
sc_rRNA_18S_Psi0759
sc_rRNA_25S_Psi2351
sc_rRNA_25S_Psi2975
sc_rRNA_25S_Psim2347
sc_rRNA_18S_Psi1187
PsiScore
Psi%

Figure S8 Calibration curves for PsiScore with *S. cerevisiae* rRNA pseudouridylated sites. Mixes of modified and unmodified rRNA from WT and CBF5-cata mutant strains were subjected to HydraPsiSeq and scatter plots show correlation of PsiScore with rRNA modification level. Non-clustered (isolated) rRNA sites are in green, clustered (at least one other Ψ site in proximity +/- 2 U nucleotides) are shown in orange, Cbf5-independent Ψ50 in 5S rRNA is shown in blue.

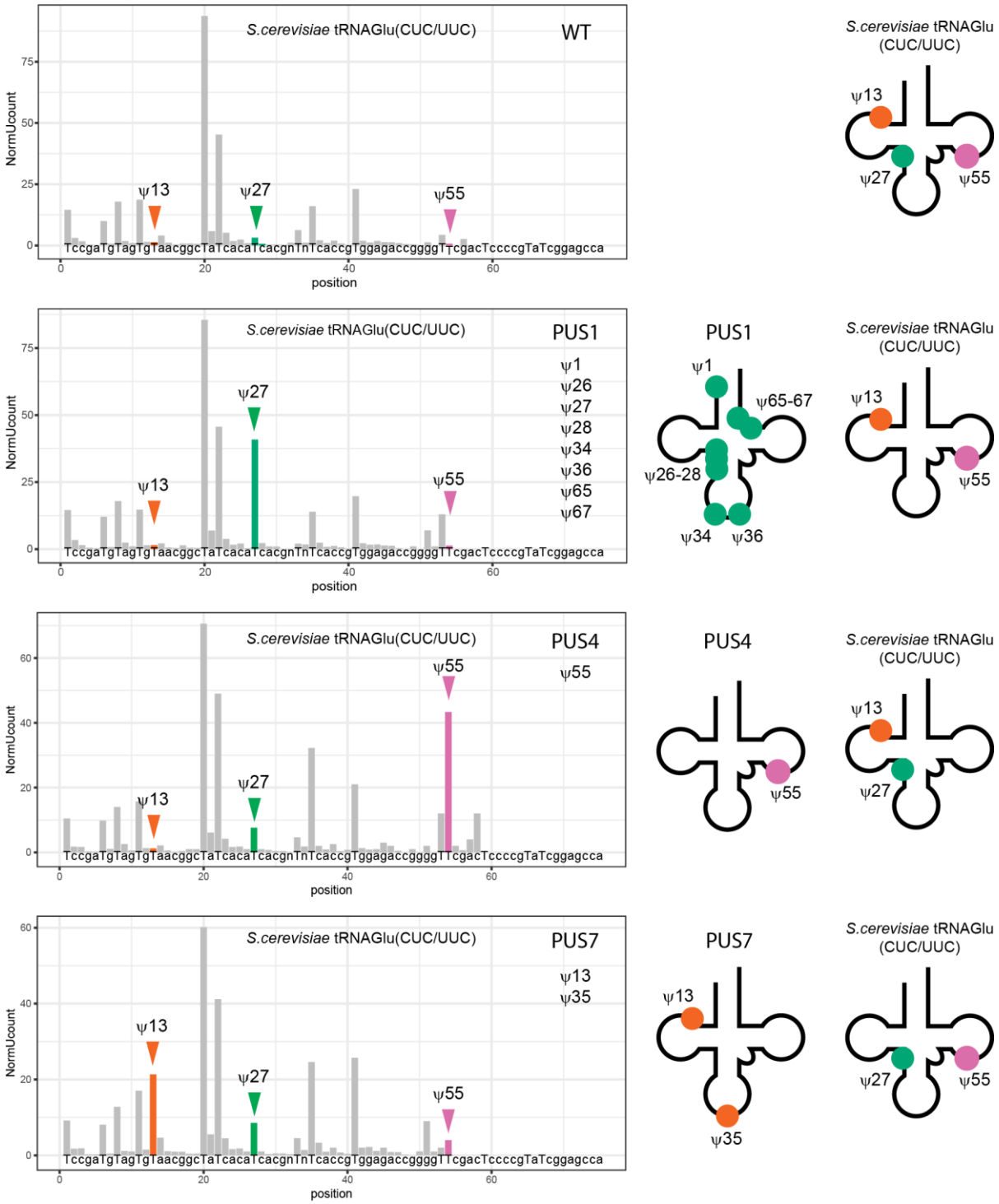


Figure S9 Analysis of tRNA pseudouridylation by HydraPsiSeq.
Full protection profiles are shown for *S. cerevisiae* $tRNA^{Glu}$(CUC/UUC), which has Ψ13 (orange), Ψ27 (green) and Ψ55 (violet), catalyzed by Pus7, Pus1 and Pus4, respectively. Cloverleaves on the right show all possible tRNA pseudouridylation sites attributed for each enzyme in *S. cerevisiae*. Position 27 shows some residual cleavage even in control strain, so it is most probably only partially modified.

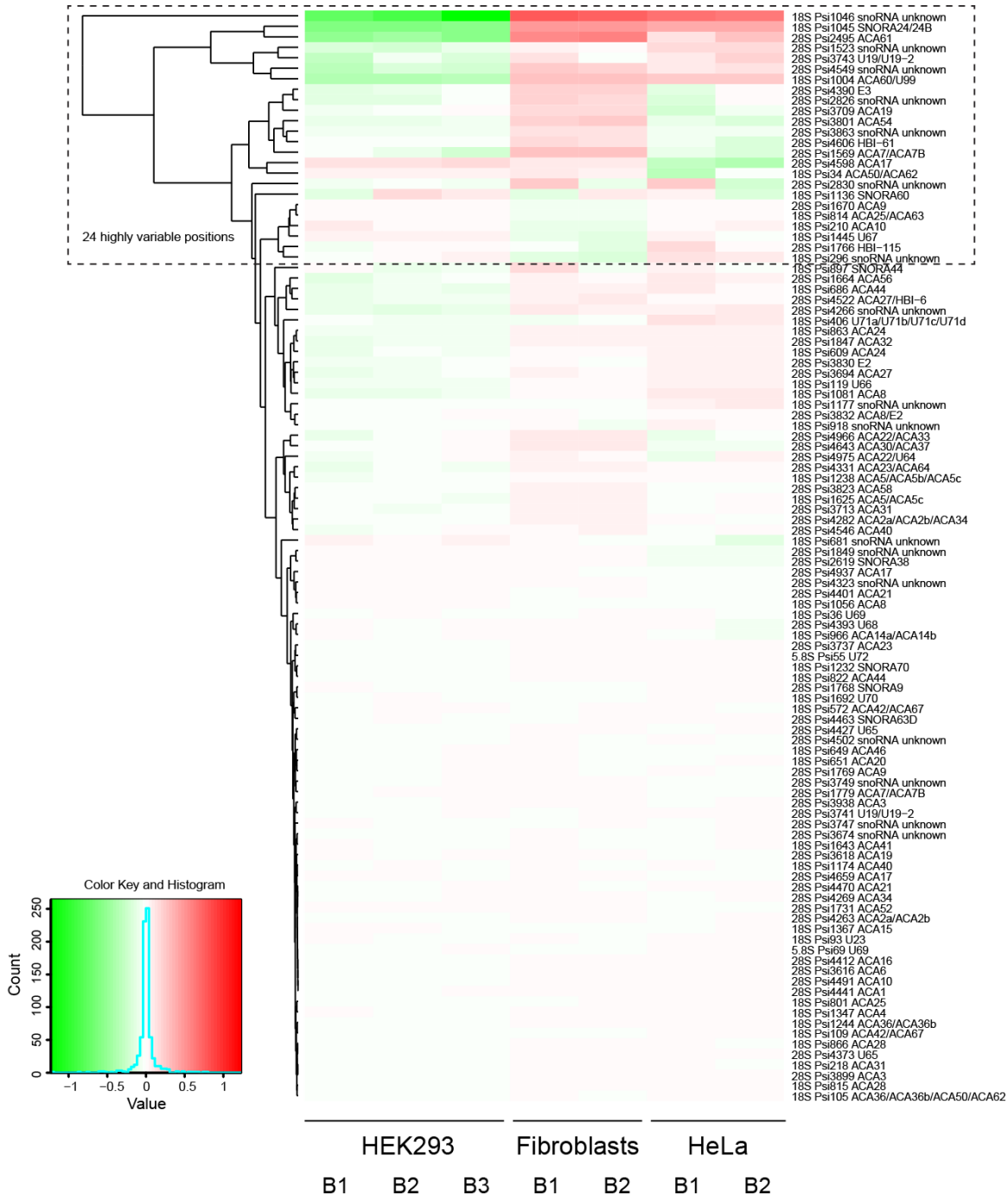
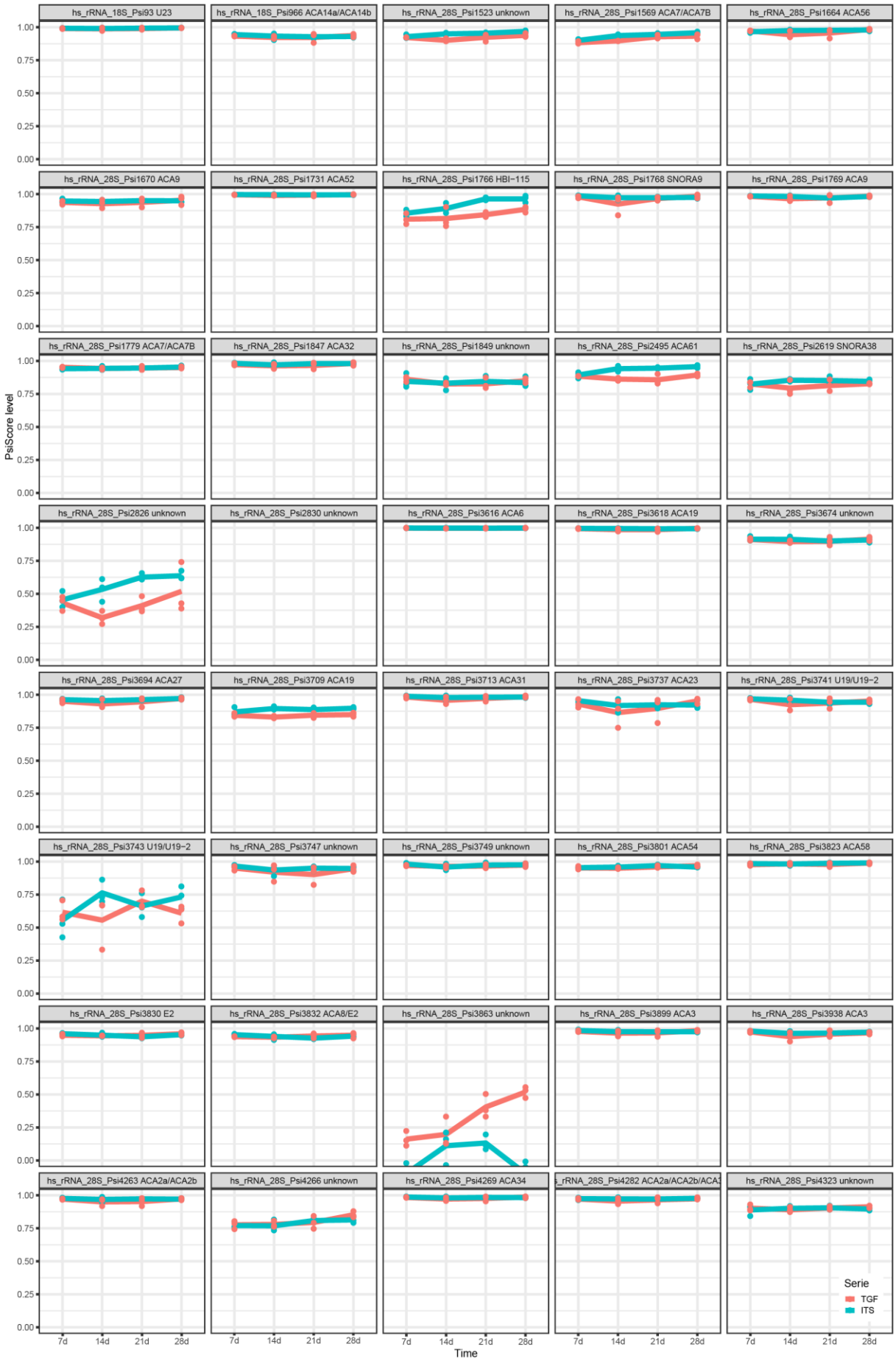


Figure S10 Profiling of human rRNA pseudouridylation by HydraPsiSeq. Human cell lines (HEK293, fibroblasts and HeLa cells) were used in biological duplicate or triplicate (labeled B1/B2/B3). Profiling was done for all previously known human rRNA modifications (104 positions), data are shown in differential row-normalized heatmap. The identity of the H/ACA snoRNA potentially guiding pseudouridylation is indicated for reference. Clustering was done using hclust R function with ward.D2 algorithm. Color key for differential PsiScore is provided in the insert.


hs_rRNA_18S_Psi1004 ACA60/U99
hs_rRNA_18S_Psi1045 SNORA24/24B
hs_rRNA_18S_Psi1046 unknown
NA_18S_Psi105 ACA36/ACA36b/ACA50/A
hs_rRNA_18S_Psi1056 ACA8
hs_rRNA_18S_Psi1081 ACA8
hs_rRNA_18S_Psi109 ACA42/ACA67
hs_rRNA_18S_Psi1136 SNORA60
hs_rRNA_18S_Psi1174 ACA40
hs_rRNA_18S_Psi1177 unknown
hs_rRNA_18S_Psi119 U66
hs_rRNA_18S_Psi1232 SNORA70
s_rRNA_18S_Psi1238 ACA5/ACA5b/ACA5
hs_rRNA_18S_Psi1244 ACA36/ACA36b
hs_rRNA_18S_Psi1347 ACA4
hs_rRNA_18S_Psi1367 ACA15
hs_rRNA_18S_Psi1445 U67
hs_rRNA_18S_Psi1625 ACA5/ACA5c
hs_rRNA_18S_Psi1643 ACA41
hs_rRNA_18S_Psi1692 U70
hs_rRNA_18S_Psi210 ACA10
hs_rRNA_18S_Psi218 ACA31
hs_rRNA_18S_Psi296 unknown
hs_rRNA_18S_Psi34 ACA50/ACA62
hs_rRNA_18S_Psi36 U69
s_rRNA_18S_Psi406 U71a/U71b/U71c/U71
hs_rRNA_18S_Psi572 ACA42/ACA67
hs_rRNA_18S_Psi609 ACA24
hs_rRNA_18S_Psi649 ACA46
hs_rRNA_18S_Psi651 ACA20
hs_rRNA_18S_Psi681 unknown
hs_rRNA_18S_Psi686 ACA44
hs_rRNA_18S_Psi801 ACA25
hs_rRNA_18S_Psi814 ACA25/ACA63
hs_rRNA_18S_Psi815 ACA28
hs_rRNA_18S_Psi822 ACA44
hs_rRNA_18S_Psi863 ACA24
hs_rRNA_18S_Psi866 ACA28
hs_rRNA_18S_Psi897 SNORA44
hs_rRNA_18S_Psi918 unknown
1.00
0.75
0.50
0.25
0.00
7d
14d
21d
28d
Time
Serie
TGF
ITS


hs_rRNA_18S_Psi93 U23
hs_rRNA_18S_Psi966 ACA14a/ACA14b
hs_rRNA_28S_Psi1523 unknown
hs_rRNA_28S_Psi1569 ACA7/ACA7B
hs_rRNA_28S_Psi1664 ACA56
hs_rRNA_28S_Psi1670 ACA9
hs_rRNA_28S_Psi1731 ACA52
hs_rRNA_28S_Psi1766 HBI−115
hs_rRNA_28S_Psi1768 SNORA9
hs_rRNA_28S_Psi1769 ACA9
hs_rRNA_28S_Psi1779 ACA7/ACA7B
hs_rRNA_28S_Psi1847 ACA32
hs_rRNA_28S_Psi1849 unknown
hs_rRNA_28S_Psi2495 ACA61
hs_rRNA_28S_Psi2619 SNORA38
hs_rRNA_28S_Psi2826 unknown
hs_rRNA_28S_Psi2830 unknown
hs_rRNA_28S_Psi3616 ACA6
hs_rRNA_28S_Psi3618 ACA19
hs_rRNA_28S_Psi3674 unknown
hs_rRNA_28S_Psi3694 ACA27
hs_rRNA_28S_Psi3709 ACA19
hs_rRNA_28S_Psi3713 ACA31
hs_rRNA_28S_Psi3737 ACA23
hs_rRNA_28S_Psi3741 U19/U19−2
hs_rRNA_28S_Psi3743 U19/U19−2
hs_rRNA_28S_Psi3747 unknown
hs_rRNA_28S_Psi3749 unknown
hs_rRNA_28S_Psi3801 ACA54
hs_rRNA_28S_Psi3823 ACA58
hs_rRNA_28S_Psi3830 E2
hs_rRNA_28S_Psi3832 ACA8/E2
hs_rRNA_28S_Psi3863 unknown
hs_rRNA_28S_Psi3899 ACA3
hs_rRNA_28S_Psi3938 ACA3
hs_rRNA_28S_Psi4263 ACA2a/ACA2b
hs_rRNA_28S_Psi4266 unknown
hs_rRNA_28S_Psi4269 ACA34
rRNA_28S_Psi4282 ACA2a/ACA2b/ACA
hs_rRNA_28S_Psi4323 unknown
PsiScore level
1.00
0.75
0.50
0.25
0.00
7d
14d
21d
28d
Time
Serie
TGF
ITS

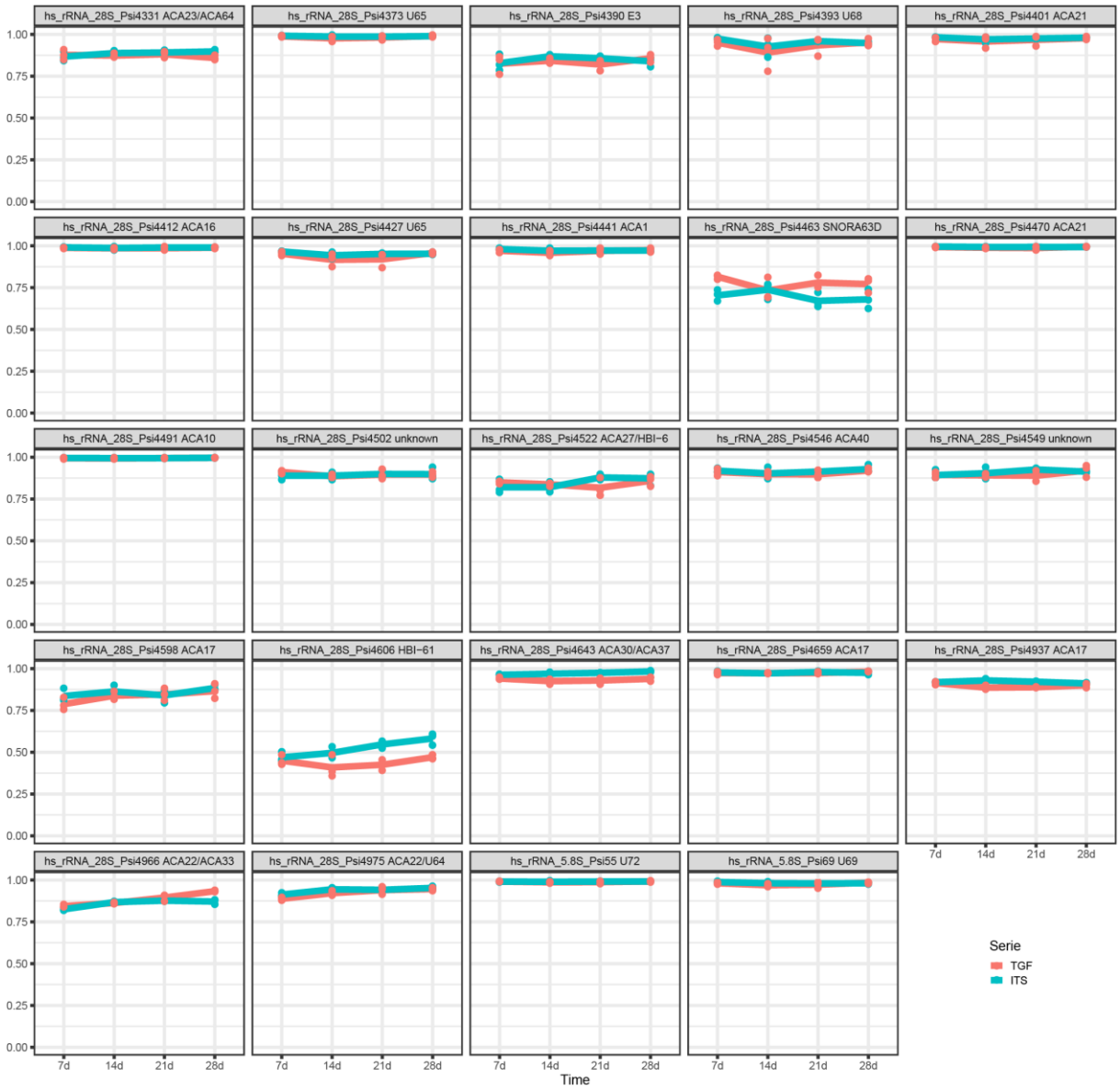


Figure S11 Variations of PsiScore during chondrogenic differentiation of BMMSCs in ITS and TGF-□1 media. Identity of positions and corresponding H/ACA snoRNAs are indicated for each panel, the differentiation was conducted for 7-, 14-, 21- and 28-days. Blue, samples in ITS medium; Red, in TGF-□1.

-- End of article --

4.2.2 Data Analysis Pipeline for Detection and Quantification of Pseudouridine (ψ) in RNA by HydraPsiSeq. [DOI: 10.1007/978-1-0716-2962-8_14]

Florian PICHOT[1,2], Virginie MARCHAND[2], Mark HELM[1] and Yuri MOTORIN[2,3]

[1] Institute of Pharmacy of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Mainz, Germany.

[2] Université de Lorraine, CNRS, INSERM, UAR2008/US40 IBSLor, EpiRNA-Seq Core facility, F-54000 Nancy, France.

[3] Université de Lorraine, CNRS, UMR7365 IMoPA, F-54000 Nancy, France.

Overview

As said in the previous section, HydraPsiSeq detection is based on absence (or low) hydrazine cleavage at uridines sites, but little details have been given about the data treatment process behind it. To ensure reproducibility and transparency about HydraPsiSeq analysis, the bioinformatics pipeline needs to be explicitly detailed.

For this, we detailed the different softwares and respective parameters used across the different steps for pseudouridine detection: Trimming, alignment, data import and data treatment alongside the different quality controls through the routine analysis. Recommendations are also given regarding detection scores threshold, different set of parameters and possible pipeline adaptation to provide better insights about the methodology itself.

Finally, the last data processing steps (known as tertiary analysis) depend on the nature and goal behind the detection of pseudouridines. It is difficult to be explicit on this part since every HydraPsiSeq project is different. Nonetheless, we provide standardized protocols based on two recurrent scenarios which are analysis of known pseudouridine sites and mapping for yet undiscovered pseudouridylated sites.

In this work, I have already created the HydraPsiSeq pipeline and wrote the technical part of the “Data treatment” and “Data analysis” section while participating in the graphs production.

Abstract

Pseudouridine, a modified RNA residue formed by the isomerization of its parental U nucleotide is prevalent in a majority of cellular RNAs; its presence was reported in tRNA, rRNA, sn/snoRNA as well as in mRNA/lncRNA. Multiple analytical deep sequencing-based approaches have been proposed for pseudouridine detection and quantification, among which the most popular relies on the use of soluble carbodiimide (termed CMCT). Recently, we developed an alternative protocol for pseudouridine mapping and quantification. The principle is based on protection of pseudouridine against random RNA cleavage by hydrazine/aniline treatment (HydraPsiSeq protocol). This "negative" detection mode requires higher sequencing depth, but provides a precise quantification of the pseudouridine content. All "wet-lab" technical details of the HydraPsiSeq protocol have been described in recent publications. Here, we describe all bioinformatics analysis steps required for data processing from raw reads to the pseudouridylation profile of known or unknown RNA.

1. Introduction

Pseudouridine (ψ), formed by the enzymatic isomerization of its parental uridine residue, is one of the most widespread RNA nucleotide modification, and is commonly observed in tRNAs, rRNAs, snRNAs, snoRNAs and mRNAs/lncRNAs [414, 510–512]. The presence of ψ alters both local RNA conformation, as well as H-bonding capacity, providing additional functional features to the RNA chain [421, 473, 513]. The prevalence and importance of ψ residues in various cellular RNAs have promoted the development of specific detection protocols [514, 515], since this nucleoside is "mass-silent" for mass-spectrometry [516–518].

Chemical methods for ψ detection stem from fundamental analysis of normal and modified nucleobases' reactivity towards various chemical reagents (reviewed in [519]). Two types of ψ reactivity were explored, the modification of nitrogens in the pyrimidine cycle [520] and Michael addition to the C5=C6 double bond, leading to the ring opening. The first approach led to a "positive" detection mode allowing distinction between U and ψ [521], while the reduced reactivity of the C=C double bond in ψ distinguishes it from uridine in a "negative" mode, in this case ψ is protected from U-specific cleavage reagents (Fig. 1A).

The "positive" detection mode using ψ-specific N-cyclohexyl-N′-β-(4-methylmorpholinium) ethylcarbodiimide p-tosylate (CMCT) was first established in low-throughput mode [520, 521] and very extensively used in the early 80's for ψ mapping in rRNAs from different species. Further developments and optimization of the protocol for deep sequencing analysis gave rise to PseudoU-Seq and its variants (termed "CMCT-Seq" here), applied for ψ mapping in yeast and human mRNAs [314, 320, 522]. Due to the "positive" detection mode, "CMCT-Seq" is relatively sensitive and, in principle, requires lower sequencing depth for analysis, but quantification of the modification content remains laborious and requires internal calibration standards [523]. A version of "CMCT-Seq" coupled to qRT-PCR was also proposed for quantification of the stoichiometry of single ψ sites in defined amplicons [524].

The “negative” detection mode relies on random cleavages of all U residues in RNA by hydrazine, while ψ residues are resistant under these conditions as they do not undergo ring’s opening (Fig. 1A). This feature was also explored for ψ detection in low-throughput mode [282, 475], but was not generalized as a widely used protocol, since the absence of a signal in primer extension analyses yields a less reliable interpretation than a distinct CMCT-related RT stop corresponding to a modified residue.

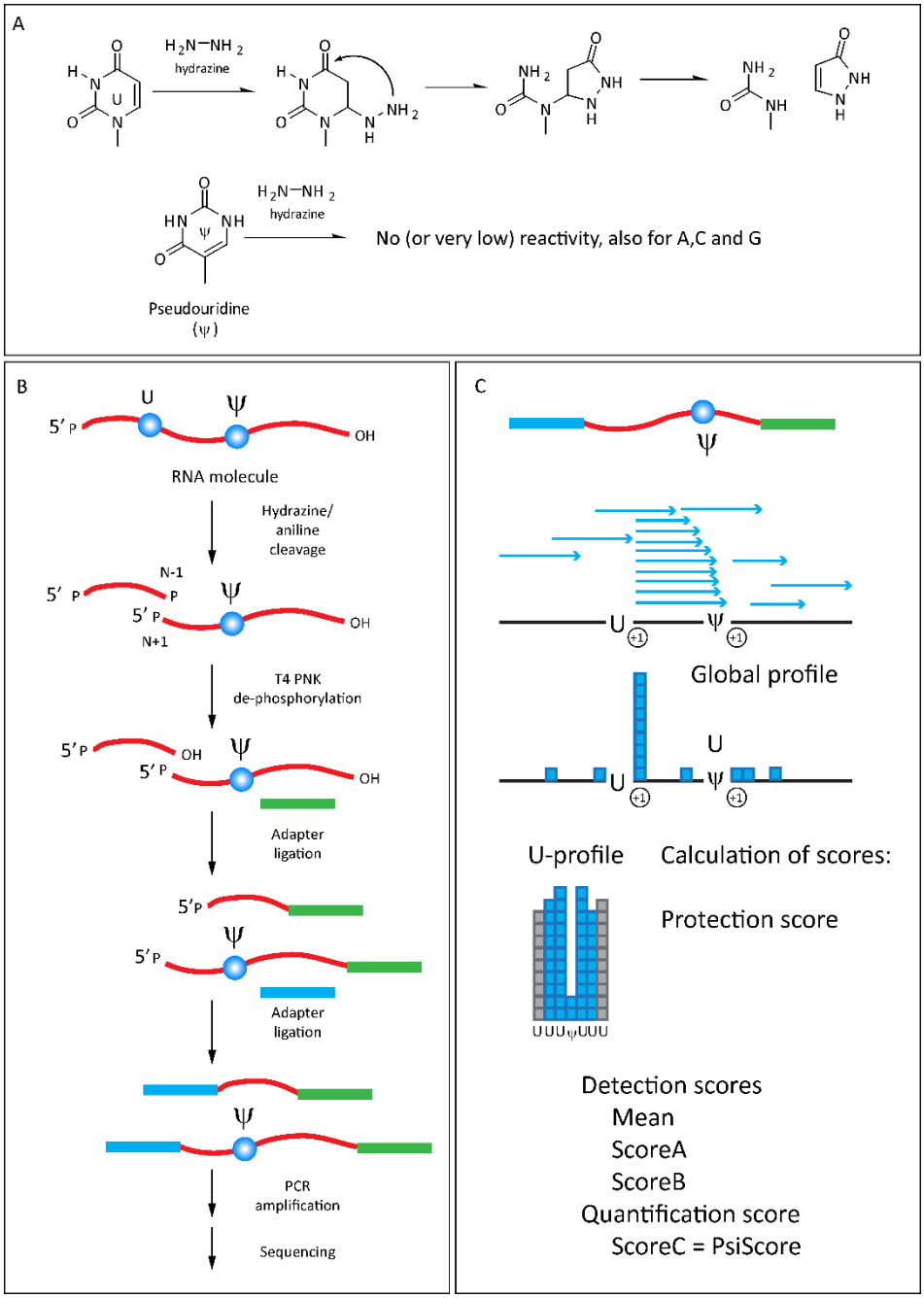


Fig. 1 Chemical basis for and experimental steps used in the HydraPsiSeq protocol.

*(**A**) - cleavage of U residues in RNA by hydrazine treatment. This is followed by aniline scission of the phosphodiester bond. (**B**) - main “wet lab” steps of the HydraPsiSeq protocol. As a key feature, only RNA fragments having adapters ligated to both extremities are amplified and sequenced, thus RT-arresting RNA modifications are not giving any sharp signal. (**C**) - outline of profiles and scores used for detection and quantification of the ψ residues.*

However, a protection at ψ residues from hydrazine cleavage allows a straightforward quantification of the modification rate and this advantage was explored in a high-throughput analytical approach, termed HydraPsiSeq [361, 451]. In this method, RNA is randomly cleaved at all unmodified U residues, yielding a 5'-P residue at the N+1 nucleotide. This feature of the resulting RNA fragments is used for specific adapter ligation, leading to strongly reduced background signal from non-U residues (Fig. 1B). In practice, the efficiency of random hydrazine cleavage varies somewhat, depending on the RNA sequence context and structure. It therefore provides a rather irregular cleavage profile even in relatively denaturing conditions (50% hydrazine in water). Since ψ is not sensitive to hydrazine, cleavages at pseudouridylation sites stay at the background (or low) levels, depending on the modification stoichiometry. In the analysis of RNA with unknown modification pattern these protected U residues in general represent potential ψ, but may also originate from other modified Us that are partially or completely resistant to hydrazine, such as m^5U, dihydrouridine (D) or hypermodified 5-substituted Us in tRNA wobble position [451].

Compared to "positive" mode detection protocols, HydraPsiSeq requires a higher coverage depth, i.e. at least >50-100 reads/RNA position. However, such sequencing depth is readily achieved with the last generation of Illumina deep sequencing machines (NextSeq/NovaSeq series) typically providing ~400-1000 million raw reads per flow cell.

Practical applications of HydraPsiSeq to the analysis of rRNA and tRNA pseudouridylation have been discussed in two recent publications [361, 451]. In this method chapter we will address the computational aspects of data analysis and interpretation.

Our data analysis pipeline for HydraPsiSeq protocol includes a quality control of the raw reads, a trimming to remove adapter's contamination, an alignment to the reference RNA sequence, a quantification of the 5'- reads' extremities, followed by calculation of window-normalized cleavage at all U residues in RNA (NormalizedUcleavage) and specific HydraPsiSeq scores. The latter allows detection of potentially modified positions in unknown RNAs and a quantification of the modification stoichiometry for RNA species of already established pseudouridylation profiles.

2. Materials

All calculations and data manipulation/storage are done using a Unix (Linux) server (we are currently using Illumina Compute Dell R815 server (2012) running Linux RedHat 6.0 and Illumina Dell R740 (2021) under Ubuntu LTS20.4). Appropriate computer configurations (Linux server and personal workstation) are described in more detail in **Notes 1 and 2**. All software and packages listed below should be installed and configured on a Linux server (see points 2.2, 2.3, 2.4 and 2.5) and PC under appropriate Windows version (Windows v.10 is currently used, see points 2.2, 2.5, 2.6 and 2.7).

2.1 Library sequencing

1. Quantified and characterized HydraPsiSeq libraries
2. Illumina sequencer (any type can be used since libraries are compatible with a full range, but throughput or flow cell capacity (nominal number of clusters) should be adapted to the required depth of sequencing). We successfully used MiSeq, HiSeq1000 and NextSeq2000 for HydraPsiSeq analysis.

2.2 Analysis of raw reads' quality by FastQC

FastQC software for Windows or Linux (current version 0.11.8, https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

2.3 Trimming

Adapter trimming software Trimmomatic (current version 0.39, http://www.usadellab.org/cms/?page=trimmomatic).

2.4 Alignment

Alignment software Bowtie 2.0 (current version 2.2.4, http://bowtie-bio.sourceforge.net/bowtie2/index.shtml).

2.5 Data processing

1. Samtools (current version 1.12, http://www.htslib.org/doc/samtools.html).
2. Bedtools (current version 2.18, https://bedtools.readthedocs.io/en/latest/).
3. (Optional) IGV (current version 2.6.2, https://software.broadinstitute.org/software/igv/).

2.6 Data treatment

R Studio server (current version "Juliet Rose" 1.4.1717, https://www.rstudio.com/products/rstudio/download-server/) running R environment (current version "Camp Pontanezen" 4.1.0, https://cran.r-project.org/bin/windows/base/). Alternatively, standalone R Studio desktop version can be installed and be used locally (current version 2021.09.0+351, https://www.rstudio.com/products/rstudio/download/)

2.7 Data analysis

Final data analysis including generating appropriate graphs and data tables can be done either in the R environment (recommended) with the ggplot2 package or in Microsoft Excel/Open/Libre office Calc.

3. Methods

"Wet-lab" library preparation steps for HydraPsiSeq (see Fig. 1B) were described in detail elsewhere [361, 451]. Here we present the data analysis pipeline, outlined in Fig. 2.

3.1 Library sequencing

1. HydraPsiSeq libraries are sequenced at a moderate depth, the typical target number of raw sequencing reads for reliable rRNA analysis is 20 million, thus 10-12 HydraPsiSeq libraries can be loaded to one lane of the HiSeq1000 flow-cell (nominal capacity of 250 million/lane, SBS v3 kits), or up to 48 libraries to one NextSeq2000 flow-cell (capacity of 1100 million/lane, P3 kits).
2. Sequencing mode is single read 50 nucleotides (SR50), this sequencing mode is well suitable for downstream analysis of HydraPsiSeq data.
3. Optimal loading of HydraPsiSeq libraries depends on the sequencer and should be optimized empirically (generally 10–12 pM for HiSeq1000), considering that clustering is relatively efficient for very short fragments obtained by hydrazine/aniline cleavage.
4. When the sequencing run is finished, raw reads are sorted by sample using Illumina i7 indexes (and, optionally i5, not used in the current library design), using *bcf2fastq* utility or onboard Dragen demultiplex utility for NextSeq2000.

3.2 Analysis of raw reads' quality by FastQC

1. FastQC analysis of raw and trimmed reads (see below) can be performed either in batch (for all FASTQ files present in the folder) or only for selected files in the interactive mode. Interactive mode is described here:
2. Open a FASTQ file using the "File->Open" option of the FastQC menu, processing starts automatically and may take time (10 sec-5 min), depending on the file size and computing power available.
3. Check raw reads statistics. The sequence length of sequencing reads should be the same for the majority of them. Base sequence quality of reads should be higher than 20 (>30 is strongly recommended) and N content should be as low as possible (< 5% is recommended). Untrimmed reads might have a high amount of adapter content and/or overrepresented sequences due to the presence of adapters and complexity of RNA samples used for analysis.

3.3 Trimming

1. For each folder containing FASTQ files with raw sequencing reads, perform trimming of the adapter sequences using Trimmomatic, using the following command *java -jar ../trimmomatic-0.39.jar SE -phred33* <INPUT_FILE> <OUPUT_FILE> ILLUMINACLIP:TruSeq3-SE.fa:2:30:N LEADING:30 TRAILING:30 SLIDINGWINDOW:4:15 MINLEN:N AVGQUAL:30. Adapter sequences are defined in TruSeq3-SE.fa file. Recommended parameters (*see* **Note 3**) ILLUMINACLIP: seed mismatches 2, palindrome clip threshold 30, sample clip threshold 7 ; SLIDINGWINDOW: window size 4, required quality 15 ; LEADING:30 ; TRAILING:30 ; MINLENGTH:8 ; AVGQUAL:30. If performance issues are encountered with Trimmomatic, reduce the number of fastq files treated in parallel, issue found with RedHat 6.0 linux server (*see* **Note 4).**

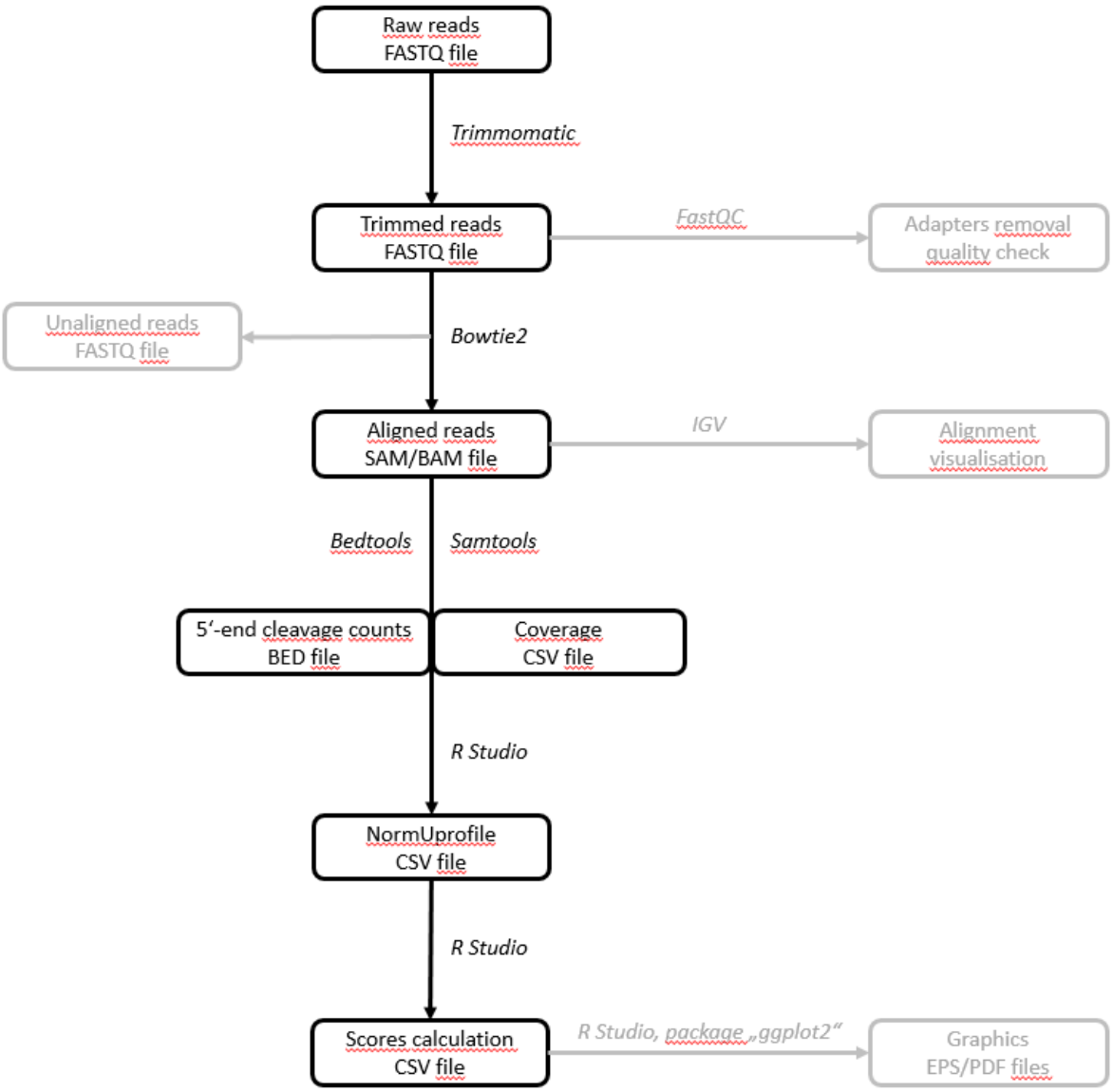


Fig. 2 Major bioinformatics steps for HydraPsiSeq data processing.

Optional conversions are shown in grey. File formats used in different steps, as well as computer environments are indicated.

2. Assess the trimming quality with FastQC (as described above). Check if adapter sequences were efficiently removed and not present anymore in trimmed reads sequences.

3.4 Alignment

1. Select the *.bt2 file containing the reference sequence(s) for *bowtie2* alignment tool (*see* **Note 5**).
2. Align trimmed reads from step 3.3 using bowtie2 in "end-to-end" mode (default *bowtie2* mode) to the appropriate reference sequence.
   Recommended parameters -p (number of computer processors/cores - 2); --no-unal; --no-1mm-upfront; -D 15; -R 2; -N 0; -L 10; -i S,1,1.15

3. (Optional) Save unaligned reads in *.fastq.gz for further analysis and *bowtie2* error message (STDERR) in a *.txt file for subsequent quality control of the alignment statistics.

3.5 Data processing

1. Keep only uniquely aligned reads (for unambiguous mapping) from the *bowtie2* output SAM file. To do this, use the command *grep -E "@|NM:"* <*SAM*> | *grep -v "XS:"*.
2. Extract reads coverage information from a uniquely aligned SAM file by *samtools mpileup* utility (*samtools mpileup* -f <FASTA> -aBQ0 -d10000000 <SAM> | cut -f1-4 > <CSV>) to generate a CSV file. Only the first four columns (the reference sequence name, the position, the reference nucleotide base and the reads coverage) are required for the downstream steps of the analysis pipeline.
3. Using *samtools view* utility convert the SAM file containing uniquely aligned reads to BAM format (*samtools view* -Sb <SAM> > <BAM>).
4. Using *samtools index* utility index the BAM file from step 3.5.3 (*samtools index* <BAM>). This step is also required for *.bam file visualization by IGV and subsequent quality check for alignment to the reference.
5. Convert the BAM file to a BED format using *bedtools bamtobed* utility (*bedtools bamtobed* -i <BAM> > <BED>). Counting of only 5'-reads' ends can be done directly from *.sam formatted file, but this step ensures compatibility with other pipelines using both 5'- and 3'-ends counting.
6. For each reference RNA sequence and BED file from step 3.5.5, generate 5'-ends counting (information available in the 2nd column of BED file) and 3'-ends counting (information available in the 3rd column of BED file) *awk '{print $2}'* <BED> | *sort* | *uniq -c* | *awk '{print $3,$2,$1,$4}'* | *sort -n* > <5primeCount.csv> (*see* **Note 6**). These files by RNA species are used to generate raw cleavage profiles. At the same time, generate coverage files from the mpileup.csv file (from step 3.5.2). Save the different output as distinct .csv files.
7. (Optional) BED and SAM files can be removed to save storage space.

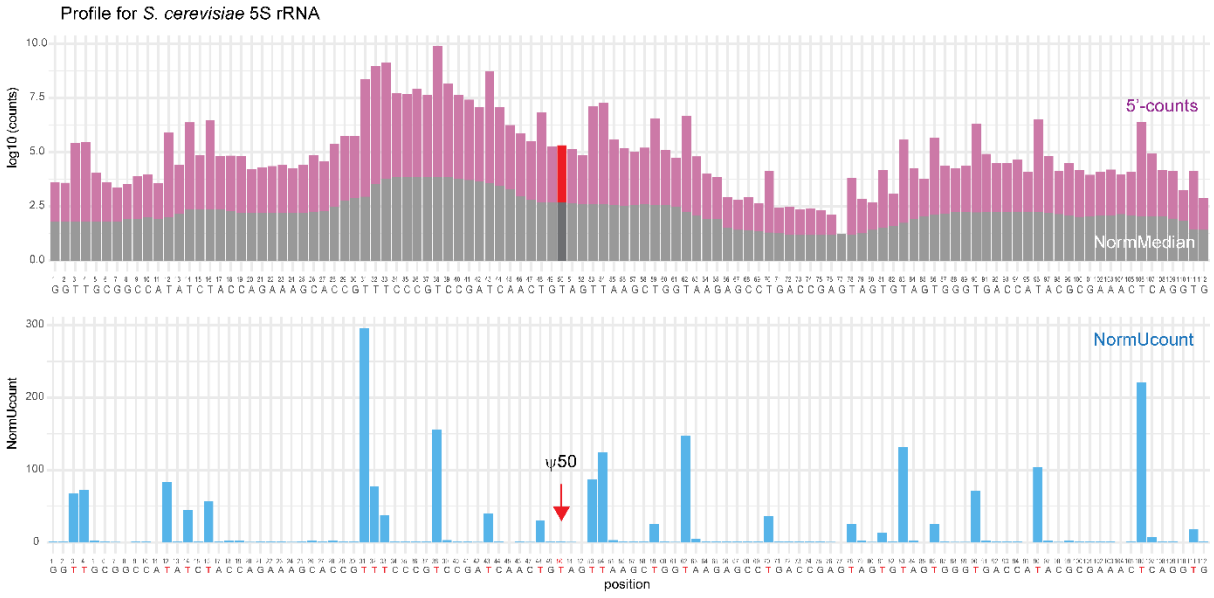


Fig. 3 Experimentally observed HydraPsiSeq profile for S. cerevisiae 5S rRNA bearing a single ψ residue at position 50 (ψ50).

*Since real HydraPsiSeq cleavage signals appears at N+1 nucleotides, to align the signal with the corresponding positions, the -1 backshift is already introduced in the graphs (see **Note 7**). Top panel shows the intensity of 5'-count signals in log10 scale (pink) as well as normalization background (in grey) to obtain NormUcounts. The bottom panel shows NormUcount profile (in blue) in a linear scale, position of protected ψ50 is highlighted and shown by a red arrow.*

3.6 Data treatment

1. List all sample folders containing 5'-end counts and coverage files.
2. For each folder and each reference sequence of interest, merge the coverage and 5'-end count files. Notice that since the original HydraPsiSeq cleavage signal appears at N+1 nucleotide, so -1 backshift is introduced during the merge step to align the cleavage signal with the nucleotide at N position (*see* **Note 7**).
3. Normalize the 5'-end counts (cleavage profile) to the local background in a sliding window for surrounding nucleotides (10 in total, 5 upstream and 5 downstream positions) to obtain NormUcount. For this, the median for cleavage at the non-U nucleotides is calculated and used as normalizing background value (see Fig. 3). (*see* **Note 8**)
4. From NormUcount generate NormUprofile by keeping only uridine positions. NormUprofile shows a protection of a given U residue compared to other Us in the neighborhood and is further used for calculation of HydraPsiScores (*see* **Note 9**). Save NormUprofile in a *.csv file for further use.
5. In order to perform detection and quantification of known and potential ψ sites, different HydraPsiSeq scores are calculated. These scores are somehow similar to RiboMethSeq scoring strategy, but, instead of using values for all nucleotides in the sequence, only NormUcounts (NormUprofile) are used. Scores A, B and C (see formulas below) were proposed by Nielsen's lab [433] for +/- 6 nucleotide interval, but further modified to reduce the neighboring

region to +/-2 nucleotides [430]. This was shown to improve detection selectivity. Examples of profiles for scores obtained for yeast *S. cerevisiae* 5S rRNA containing a single ψ residue at position 50 are shown in Fig. 4.

a. ScoreA2 is used for the detection of potential modified sites (A2 means ScoreA calculated for +/-2 neighboring nucleotides) and is calculated using the following formula:

$$Score\ A2 = 1 - \frac{2n_i + 1}{\frac{1}{2}|\mu_l - \sigma_l| + n_i + \frac{1}{2}|\mu_r - \sigma_r| + 1}$$

b. where n is the read count at position i, μl and μr are the mean values for left and right flanking regions, and σl / σr are standard deviations. In contrast to the initial formula proposed by [433] the updated version may also take into account negative values, those are not systematically rounded to zero. ScoreA2 varies from negative values to 1.0 (maximal value).

c. ScoreB2 is also used for detection of unknown modified sites (also for +/-2 neighbouring nucleotides). This score is highly sensitive and used for confirmation of ScoreA2 detection. Values of ScoreB2 vary from negative values and the maximum value is not limited.

$$Score\ B2 = \frac{n_i - \frac{1}{2}\left(\frac{\sum_{j=i-2}^{i-1}\omega_j n_j}{\sum_{j=i-2}^{i-1}\omega_j} + \frac{\sum_{j=i+1}^{i+2}\omega_j n_j}{\sum_{j=i+1}^{i-2}\omega_j}\right)}{n_i + 1}$$

where n is the read count at position i, and weight coefficients ω are 0.9 for -2 and +2 positions and 1.0 for -1 and +1 nucleotides.

d. ScoreC2 (also named PsiScore) is not used for the detection of modification candidates, but rather for quantification of the modification level, since this score shows a linear dependency between score value and protection of a given nucleotide against cleavage. Like ScoreA2, ScoreC2 is not rounded to zero and takes negative values, maximal value is 1.0. Weight coefficients are the same for ScoreB2 and ScoreC2.

$$Score\ C2 = \frac{n_i}{\frac{1}{2}\left(\frac{\sum_{j=i-2}^{i-1}\omega_j n_j}{\sum_{j=i-2}^{i-1}\omega_j} + \frac{\sum_{j=i+1}^{i+2}\omega_j n_j}{\sum_{j=i+1}^{i-2}\omega_j}\right)}$$

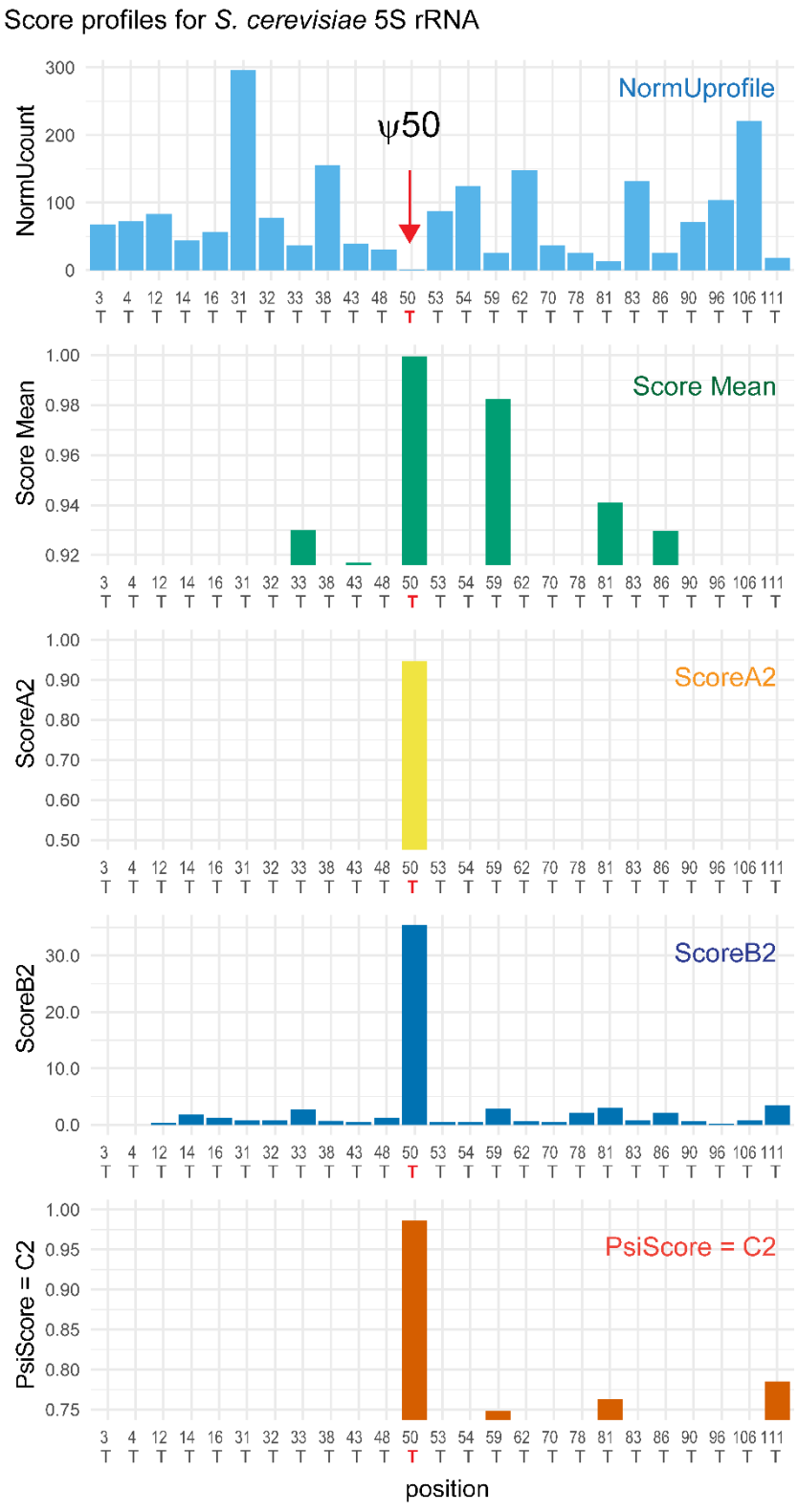


Fig. 4 NormUcount profile and the corresponding HydraPsiSeq scores for U (shown here as T in the sequence).

Only U residues are taken for Uprofile (top panel) and subsequent calculations of the HydraPsiSeq scores. ScoreMean, ScoreA2, ScoreB2 and ScoreC2/PsiScore are shown. Depending on the score, a different scale at the Y axis is used to exclude background values for unmodified residues. Position of U to ψ conversion, T(U)50 is shown by an arrow.

e. ScoreMean is an additional score which was proposed for the detection of Nm candidates [64] and applied here for ψ detection. Calculation of the ScoreMean is illustrated in Fig. 5 for simulated reads and real profile for yeast 5S rRNA. First, ratioLeft and ratioRight are calculated for neighboring nucleotides. At a second step, variation of the ratio at nucleotide n to neighboring positions is used. The ScoreMean is calculated as the mean value for aroundLeft and aroundRight (highlighted in more intense colors in Fig. 5).

6. Save the table containing the scores in a .csv file for later use.

3.7 Data analysis

Final data analysis pipeline depends on the nature of analyzed RNA and if the modification sites are already identified and/or validated. If so, the detection of candidates becomes optional, and analysis can proceed directly to extraction of the HydraPsiSeq scores (almost exclusively of PsiScore/ScoreC2) used for quantification of the modification level. This extraction is done using the reference list for RNA modifications in the species to be considered.

Major steps of this analysis are described below:

1. Extraction and analysis of known modification sites
    a. Load the list (in *.csv format) containing the known modified positions in the RNA of interest.
    b. Select only type of modifications to be extracted and further analyzed (here positions of known Psi residues)
    c. Merge the scores table with the list by RNA type and position.
    d. Generate PsiScore heatmaps for raw and row-normalized values, by the *ggplot2* R package. Save graphs as EPS or PDF.
2. Mapping of yet unknown modification candidates.
    a. When RNA modification profile for a given modification (here for ψ residues) is unknown and has to be established; mapping of potential candidate sites is performed on the basis of “detection” scores, mainly ScoreMean and ScoreA2. Score B2 can also be used as an additional parameter to decrease the number of potential false positives.
    b. Analysis of rRNA from different sources and construction of the corresponding ROC curves pointed out that threshold levels of ScoreMean >0.92 and ScoreA2 >0.5 generally provide a good compromise between selectivity and specificity (*see* **Note 10**).
    c. If necessary, threshold levels for ScoreB2 and PsiScore can be introduced, empirical values to begin with are ScoreB2 >2.5 and PsiScore >0.75.
    d. These recommended combinations of scores allow detection of highly modified positions; partial modification at a given site generally gives only low scores which do not exceed above mentioned levels. Thus,

partial RNA modification with PsiScore< ~ 0.7 escapes identification (*see* **Note 11**).

e. Detected candidates always require validation by orthogonal approaches, therefore CMCT-based protocol for ψ mapping can be recommended.

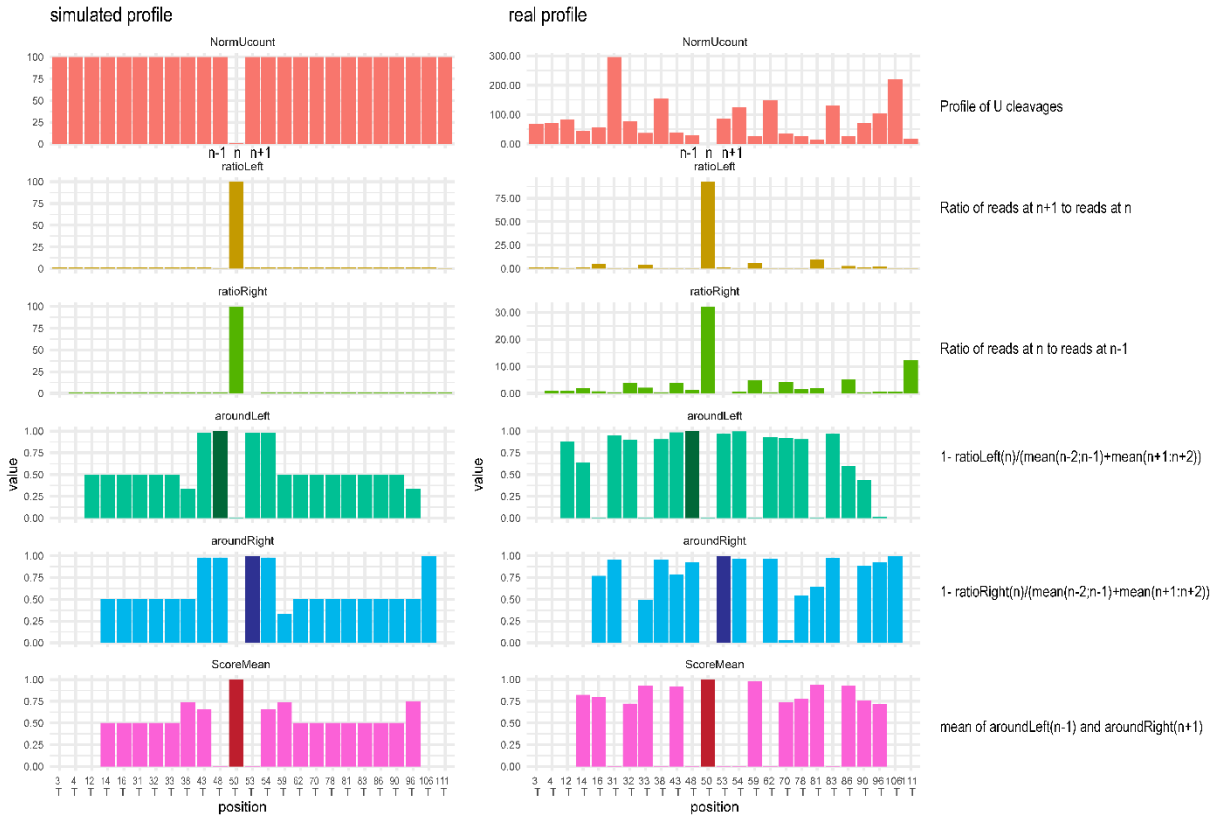


Fig. 5 Intermediate calculation to obtain ScoreMEAN used for mapping of ψ candidates.

Behavior of the different metrics is shown for ideal simulated U profile (left) and for the S. cerevisiae 5S rRNA. RNAs have the same length and the same positions for T(U) residues. Formulas used to obtain different profiles are shown on the right. Values of aroundLeft and aroundRight used for calculation of ScoreMEAN (pink) are highlighted by more intense color.

4. Notes

1. Recommended minimal configuration for Linux server used for data storage, trimming, alignment and reads' counting. Computing power is not the most crucial limiting factor for HydraPsiSeq data analysis pipeline. Raw reads trimming and alignment can be successfully performed on mid-range Intel-based servers (e.g. Xeon E5-2630, 6 cores/12 threads, 2.6 GHz, 32 Gb RAM, ~15000 CPU marks, see https://www.cpubenchmark.net/ for CPU marks for different processors) or higher (e.g. 2x Xeon Gold 6226R, 16 cores/32 threads, 2.9 GHz, 128 Gb RAM, ~50000 CPU marks), depending on the number of sequencing projects to be treated. Recommended HDD storage capacity is 10-20 Tb, if possible, at least a part of the storage capacity should be in SSD format to accelerate exchange of data. Depending on the computing power, trimming of typical raw reads fastq files takes 2-3 minutes for each, but may be done in parallel, depending on the number of available cores.

Alignment and counting are the most time-consuming steps, about 10-15 min/sample, depending on the length of the RNA reference. Estimations are given here for rRNA analysis. Of note, the R Studio environment can also be installed on the Linux server and used via Web interface (R Studio Server), this greatly improves performance in data processing at the second step. This can be used as an alternative to Win10 PC running desktop R/R Studio for data processing.

2. Configuration of a personal computer running R/R Studio is relatively flexible, almost ordinary mid-range office PC can be successfully used for data analysis. For optimal performance we recommend Intel i5/i7 based configurations with at least 16Gb (better 32 Gb) of RAM and SSD storage.

3. Depending on the RNA origin and the reads length, the sample clip threshold and MINLENGTH may be adjusted. Here are some recommended alternatives:

- combination of simple clip threshold 10; MINLENGTH:17 may be used for very complex RNA references
- combination of simple clip threshold 7; MINLENGTH:12 intermediate settings, allowing to reduce the number of ambiguously mapped short reads

4. (Optional) If performance issues are found for the Trimmomatic step, reduce the number of java threads running in parallel. This can be done either by treating fastq files one by one (with degraded performance) or by starting successive Trimmomatic threads with delay. The time required is computer-dependent but should be in a range of seconds (10-15 sec for example).

5. Reference sequence(s) for *bowtie2* are stored in a FASTA file, but only *bowtie2* index files *.bt2 are used for alignment. Original FASTA file containing the reference sequence needs to be indexed by the *bowtie2-build* utility.

6. Only 5’-end counts will be used in the HydraPsiSeq data analysis pipeline. 3’-end counting is conserved for eventual quality control by manual inspection. Generation of these 3’-count files may be omitted, but gain in storage volume and processing time is negligible.

7. A coverage file contains sequence information for all positions in RNA, while certain locations may be missing in the 5’-end counts file due to insufficient coverage, thus the coverage file is used as a backbone for merge by default. Coverage file is generated from a SAM file and thus has 1 to n coordinate space, while 5’-end counts are created from a BED format and positions are numbered from 0 to n-1. This backshift is taken into account during the merge step and also allows alignment of HydraPsiSeq signals to the cleaved nucleotides (and not to N+1 positions).

8. This normalization algorithm is appropriate for most rRNA (or tRNA) species since they do not contain over 9 U residues in a row. However, this may be not true for any randomly taken RNA, like mRNA, sn/snoRNA or lncRNA. In such cases, the window can either be extended to the 14-20 nucleotides, or the last non-zero background value obtained before entering the U-rich region is used for background normalization. Upon our experience, these solutions seem to be relatively equivalent. Similarly, values for the first 5 and last 5 positions in RNA are calculated using the adapted calculation window (1:5) and (n-4:n), respectively.

9. From this point, protected sites, corresponding to potential pseudouridylation positions, can be detected by manual inspection of the NormUcount values and global shape of the NormUprofile obtained from NormUcount by omission of non-U nucleotides. Naturally occurring pseudouridylation sites in RNA display a very low cleavage rate due to pseudouridine protection against hydrazine/aniline cleavage, and thus can be distinguished from unmodified uridines by low NormUcount value (< 2). Graphs for NormUprofile can be also generated here for visual inspection of cleavage protection. However, for more rigorous detection and quantification of putative pseudouridylated sites we recommend the use of HydraPsiSeq scores described below.

10. Combination of scores' threshold values ScoreMean >0.92 and ScoreA2 >0.5 is recommended for initial inspection and rather stringent mapping of potential pseudouridylation sites in unknown RNA. For more relaxed detection allowing to capture partially modified candidates, the combination of ScoreMean>0.90 and ScoreA2>0.45 can be used. In case of a very important number of candidate sites in long RNA references the thresholds for ScoreMean and ScoreA2 can be combined with ScoreB2>2.5 and PsiScore >0.9. Such a combination gives the most highly confident candidate sites.

11. Validation of the modification candidates in unknown RNA can be done either by alternative orthogonal technologies [31], or, better by using unmodified RNA obtained by *in vitro* transcription or, if available, extracted from KO cells having deleted or inactivated RNA modification enzyme. We found that the use of T7 transcripts may give only partial validation, since the cleavage profiles of native and *in vitro* transcribed RNA may be not identical and thus both false positives and false negatives hits are still possible.

Acknowledgement

This work was supported by ANR grants (MetRibo, D-erase) and Région Grand Est FRCR grants (EpiARN, ViroMOD) to the V.M./Y.M., and Deutsche Forschungsgemeinschaft (DFG) grants to M.H. [HE3397/17-1, SPP1784, and TRR319 RMaP, TP C01].

-- End of article --

4.2.3 NOseq: amplicon sequencing evaluation method for RNA m^{6}A sites after chemical deamination. [DOI: 10.1093/nar/gkaa1173]

Stephan WERNER[1,†], Aurellia GALLIOT[1,†], Florian PICHOT[1], Thomas KEMMER[2], Virginie MARCHAND[3], Maksim V. SEDNEV[4], Tina LENCE[5], Jean-Yves ROIGNANT[5,6], Julian KÖNIG[5], Claudia HÖBARTNER[4], Yuri MOTORIN[7], Andreas HILDEBRANDT[2], and Mark HELM[1]

[1] Institute of Pharmaceutical and Biomedical Sciences, Johannes Gutenberg-University Mainz, Staudingerweg 5, 55128 Mainz, Germany

[2] Institute of Computer Science, Johannes Gutenberg-University Mainz, Staudingerweg 9, 55128 Mainz, Germany

[3] Epitranscriptomics and Sequencing (EpiRNA-Seq) Core Facility, UMS2008 IBSLor CNRS-Université de Lorraine-INSERM, Biopôle, 9 Avenue de la Forêt de Haye, 54505 Vandœuvre-lès-Nancy, France

[4] Institute of Organic Chemistry, Julius Maximilian University Würzburg, Am Hubland, 97074 Würzburg, Germany

[5] Institute of Molecular Biology, Ackermannweg 4, 55128 Mainz, Germany

[6] Génopode - Center for Integrative Genomics, Université de Lausanne, 1015 Lausanne, Switzerland

[7] IMoPA, UMR7365 CNRS-Université de Lorraine, Biopôle, 9 Avenue de la Forêt de Haye, 54505 Vandœuvre-lès-Nancy, France

† These authors contributed equally to the paper as first authors.

Overview

To overcome antibody-enrichment detection methods limitations for m^6A, an innovative m^6A detection method based on chemical approach has been developed in our laboratory. Named “NOseq” and inspired by bisulfite sequencing, this method relies on RNA nucleobase deamination, converting all bases with primary amine to another residue. But m^6A methyl group is located on the amine group, hindering then the deamination thus providing a clear distinction with unmodified adenosines. However, other unmodified nucleobases are also sensitive to this treatment, resulting in chemical conversion of almost all nucleobases. This treatment changes the base pairing of most of the nucleobases, making it impossible to assign deaminated reads locations with conventional alignment tools.

Due to this particular combination of pairing, a personalised alignment tool adapted to this protocol needed to be developed to ensure the correct assignment of reads identity. Also, a subsequent script for detection of m^6A residues by assessing the absence of deamination conversion needed also to be written. The distinction between unmodified adenosines and m^6A sites is based of the non-conversion rate of adenosines sites to guanosines (inosines in reality) in a similar manner as bisulfite non-conversion rate used in bisulfite sequencing[468].

To validate this method, sequencing tests have been done on synthesized oligos bearing unmodified adenosines or m^6A at a known position. The validation has also been tested on *in situ* m^6A sites, namely in human rRNA 18S, human lncRNA MALAT1 and *D. melanogaster* Hairless mRNA. The latter proves the strength of NOseq for m^6A detection

In order to align the deaminated reads, I created a R script for suitable alignment of NOseq data and took care of the algorithms and data processing from the trimming to the m^6A detection. This pipeline was then adapted in Python by my colleague Aurellia Galliot to improve alignment script’s performance.

ABSTRACT

Methods for the detection of m^{6}A by RNA-Seq technologies are increasingly sought after. We here present NOseq, a method to detect m^{6}A residues in defined amplicons by virtue of their resistance to chemical deamination, effected by nitrous acid. Partial deamination in NOseq affects all exocyclic amino groups present in nucleobases and thus also changes sequence information. The method uses a mapping algorithm specifically adapted to the sequence degeneration caused by deamination events. Thus, m^{6}A sites with partial modification levels of ~50% were detected in defined amplicons, and this threshold can be lowered to ~10% by combination with m^{6}A immunoprecipitation. NOseq faithfully detected known m^{6}A sites in human rRNA, and the long non-coding RNA MALAT1, and positively validated several m^{6}A candidate sites, drawn from miCLIP data with an m^{6}A antibody, in the transcriptome of *Drosophila melanogaster*. Conceptually related to bisulfite sequencing, NOseq presents a novel amplicon-based sequencing approach for the validation of m^{6}A sites in defined sequences.

INTRODUCTION

The occurrence of m^{6}A in eukaryotic polyadenylated RNA is among the most investigated phenomena in recent RNA research. As we now know, m^{6}A deposition on nascent RNA is an early and important event in maturation of mRNA, which is orchestrated by a sophisticated catalytic complex, whose components are subject to continuous updates [525–527]. Methylation of adenosines has been shown to be important for regulation of downstream events including splicing [528], transport [529, 530], degradation [531] and translation [532]. Variegated amounts or lack of m^{6}A have been associated with numerous pathologies [533–536], and impact in infection biology is a much-investigated aspect [537, 538]. The presence of m^{6}A in mRNA has been known for decades [539–541], pioneered about 50 years ago by analytical techniques that are still widely used, namely isolation of mRNA by hybridization to poly(dT)-cellulose [542, 543] or poly(U)-sepharose [544] and enrichment of m^{6}A-containing RNA using antibodies [545, 546]. Breakthrough papers in 2012 [547, 548] combined this approach with deep sequencing to create the first m^{6}A maps of what is now termed the “epitranscriptome” [549], thus creating a boost of the field that is still unabated [550, 551]. However, the community is in agreement, that the use of antibodies alone, or in refining combinations with other techniques [539–541], still does not provide quantitative data at single nucleotide resolution [283]. Numerous tools from chemical biology have been applied to the task, including catalytic DNA [552], chemically altered dNTPs [553], engineered SAM analogues [554], or engineered polymerases [555] for the generation of a reverse transcription signature, or derivatives of SAM or metabolic precursor methionine [556] for click chemistry-based enrichment. Current developments focus on the use of nanopore technology [557], m^{6}A-discriminating nucleases [558–560] and fusion proteins for RNA-editing adjacent to m^{6}A sites [561, 562]. Despite these developments, a technology that resembles the bisulfite sequencing of m^{5}C in terms of resolution and quantification is not yet in sight. Taking a clue from the chemistry of bisulfite

sequencing [563–565], we examined the possibility of a reagent that would deaminate adenosines to inosine (I), and thus change its sequencing properties, while differentially reacting with m^{6}A such as to not alter its reverse transcription signal. While adenosines are largely inert to nucleophiles like bisulfite, we found reports on adenosine deamination by nitrous acid [566], which led to the characterization of *N*^{6}-methyl-*N*^{6}-nitrosoadenosine (NOm^{6}A) [567] (*Figure 1a*), with unknown base-pairing properties. For a targeted site-specific investigation of modified nucleotides in biological templates, a common methodology is the amplification of sequence regions. By applying primers to define sequence contexts, amplicon libraries for deep sequencing can be prepared and subsequently analysed [568].

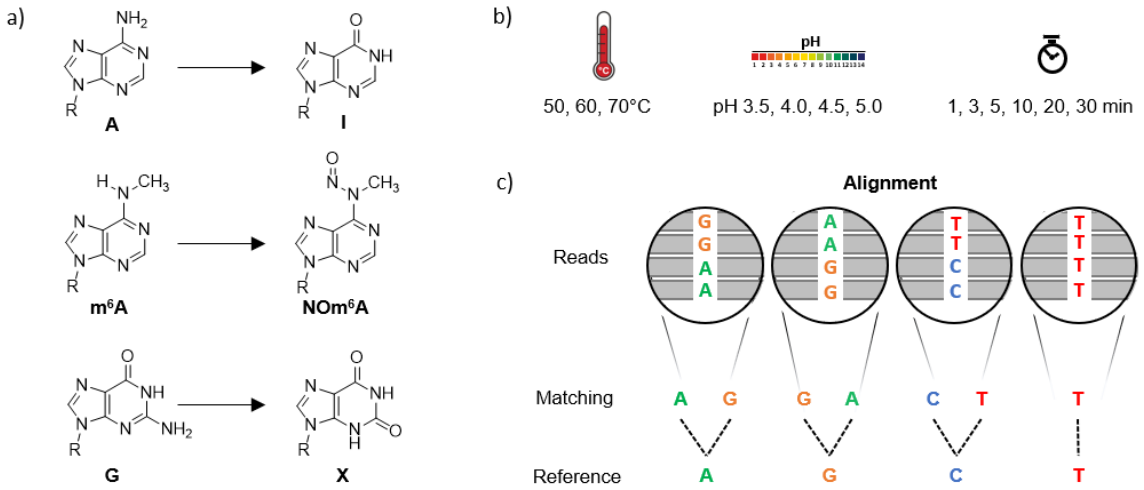


Figure 1: *Deamination treatment and deamination sequencing (NOseq) output. a) Reaction of adenosine, N^6-methyladenosine (m^{6}A) and guanosine under deamination treatment and their corresponding reaction products inosine, N^6-methyl-N^6-nitrosoadenosine (NOm^{6}A) and xanthosine. (Supplement Figure S1 for cytidine and uridine). b) Deamination treatment of a 53mer, containing m^{6}A at position 33, with 72 different conditions (composed of different reaction temperatures, pH and reaction times). (Supplement Figure S4 for additional information). c) Partial deamination sequencing output and required alignment adaptations (Example: native adenosine residues gave rise to either A or G signals, the latter deriving from the fraction having undergone deamination to I).*

Here we present NOseq, an Illumina sequencing-based analysis pipeline for deaminated RNA sequences and m^{6}A site detection by amplicon sequencing. The workflow combines an optimized deamination protocol, target-specific library preparation and a deamination-related alignment algorithm. The latter is based on the well-known seed-and-extend mapping strategy, first introduced in Blast [569] and mostly used by short read analysis tools [570] as SOAP [571] and BOWTIE [572]. The algorithm principle stands on the extension and selection of matching k-mers (short sequence of k bases) between reads and references. Basically, an alignment score is attributed according to a substitution matrix, taking mismatches and gaps into account. Here, we

implemented an asymmetric substitution matrix considering deamination-induced nucleotide conversion, an alignment strategy known from bisulfite sequencing (e.g. BSMAP [573]). We successfully recapitulated known m^6A sites in human MALAT1 lncRNA, human 18S rRNA and validated new m^6A sites in mRNA *Hairless* (H) and *female-lethal-2-d* (fl(2)d) from *Drosophila melanogaster*.

MATERIAL AND METHODS

RNA preparation

Total RNA extraction from HEK and HeLa cells. HEK293 or HeLa cells were cultured at 37°C with 5% CO_2 in T-160 cell culture flasks containing growth medium (90% D-MEM, 10% FBS and 1% Pen/Strep). Cultures at around 80% confluency were routinely splitted i.e. every 24 to 48 hours in new growth medium until harvesting of the cells and up to one month (10 passages). For total RNA extraction, cells were detached from the flasks for 5 min with 3 mL of Trypsin/EDTA (Sigma-Aldrich) and pelleted by low speed centrifugation in a 50 mL tube. The cells at confluency were evaluated (seeding) at around 18 million, resuspended and vortexed in 2 mL TriReagent (Sigma Aldrich) before adding 800 µL of chloroform allowing phase separation under centrifugation (12,000x g, 4°C, 15 min). The upper aqueous phase containing RNA was precipitated with 250 µL of 100% nuclease-free isopropanol. After vortexing, incubation and centrifugation (15,000x g, 4°C, 10 min), the pellet was washed twice with 80% ethanol, centrifuged again and the air-dried pellet was finally dissolved in MilliQ water.

PolyA RNA extraction from Drosophila melanogaster. *Drosophila melanogaster* Oregon-R wild-type strain and *Ime4^{null}* knockout strain [574] were maintained at 25°C, at standard conditions. For total RNA isolation, 20 flies were smashed together per 1.5 mL tube with 100 µL TriReagent (Sigma Aldrich) before adding another 400 µL and incubating the mixture at 4°C for 5 min. After adding 100 µL chloroform and vortexing for 15 s, an incubation step at 4°C for 10 min was followed by centrifugation (12,000x g, 4°C, 15 min), permitting phase separation. The upper aqueous phase was kept for total RNA precipitation by adding 250 µL of 100% nuclease-free isopropanol. After centrifugation (15,000x g, 4°C, 10 min), the RNA pellet was washed twice with 750 µL 80% ethanol and centrifuged again. Ethanol was then removed, and the pellet air-dried for 5 min at room temperature. The pellet was finally dissolved in MilliQ water and quantified. For polyA RNA extraction, 60 µg of total RNA were first treated with DNase I (Thermo Scientific) to avoid DNA contamination before incubation with 100 µL washed oligo d(T)25 magnetic beads (New England Biolabs), to isolate mRNA from total RNA according to a modified protocol of Dynabeads (Thermo Scientific), including a second round of purification and elution in MilliQ water.

RNA quantification and quality control. RNA was quantified using the UV-Vis spectrophotometer Nanodrop 2000 (Thermo Scientific) and the purity was assessed according to both absorption ratios A260/A280 and A260/A230 respectively in the range of 2 and 1.8–2.2 reflecting contaminant free samples. Isolated total RNA and mRNA were analysed, and quality evaluation by automated electrophoresis on the Agilent 4200 TapeStation system, following RNA ScreenTape Assay instructions, was performed.

Methylated RNA immunoprecipitation (MeRIP). The anti-m^{6}A monoclonal mouse antibody (purified IgG) (Synaptic Systems) was first incubated overnight at 4°C in 1x PBS with Protein G Sepharose 4 Fast Flow antibody purification resin (GE Healthcare). Around 8 µg of total RNA, mRNA or synthetic oligonucleotide (*Table 1*)

were added to the binding buffer after several antibody-beads washing steps and kept for 2 hours incubation at 4°C and under shaking. This incubation step was repeated, including the washing steps, together with the addition of 0.5 mg/mL adenosine (Sigma-Aldrich) in order to induce a competitive elution of non-targeted RNA. Targeted m^6A RNA fragments were then isolated from the antibodies by TriReagent (Sigma Aldrich) and following ethanol precipitation.

Table 1. RNA oligonucleotides (purchased from IBA Lifesciences, Germany)

Name	Sequence
53mer (m^6A_{33})	5′-AUAGGGGAAUGGGCCGUUCAUCUGCUAAAAGG(m^6A)CUGCUUUUGGGGCUUGUAGU-3′
53mer (A_{33})	5′-AUAGGGGAAUGGGCCGUUCAUCUGCUAAAAGG(A)CUGCUUUUGGGGCUUGUAGU-3′

Deamination

Deamination with sodium nitrite. RNA (synthetic/isolated/immunoprecipitated) was deaminated under different conditions. Reactions were performed in a total volume of 100 µL, including RNA, 1 M sodium acetate/acetic acid buffer (pH 3.5, 4.0, 4.5 or 5.0) and 1 M sodium nitrite solution (final concentrations) at 50, 60 or 70°C for 1, 3, 5, 10, 20 or 30 min. The reaction was stopped by adding 0.5 M ammonium acetate solution (final concentration) and 1 µL glycogen (Thermo Scientific) to a final volume of 250 µL. By adding 750 µL ethanol (-80°C) the treated RNA was ethanol precipitated for 1 h at -80°C. Samples were centrifuged for 45 min at 5° and 15,000x g. Pellets were washed with 80% ethanol and centrifuged again. Then, the ethanol precipitation step was repeated. Pellets were resuspended with 10 µL MilliQ water.

***Deamination with diethylene glycol dinitrite*.** Diethylene glycol dinitrite was synthesized in analogy to a known procedure [574] as follows. A mixture of diethylene glycol (54 g; 0.5 mol), sodium nitrite (69 g; 1.0 mol) and ice (100 g) was placed into a one-liter glass beaker. Conc. hydrochloric acid (101 g; 0.5 mol) was added dropwise under vigorous stirring and cooling with an ice bath so that the reaction temperature did not exceed 5 °C. Afterwards, the reaction mixture was stirred for 5 min and then transferred to a separating funnel. The organic layer was separated, washed with diluted aq. potassium carbonate and dried with potassium carbonate to give 34.4 g (42%) of the title product as a yellow oil. 1H NMR (400 MHz, $CDCl_3$): δ = 3.77 (m, 4H, 2×CH_2), 4.86 (m, 4H, 2×CH_2) ppm. 13C NMR (125 MHz, $CDCl_3$): δ = 67.2, 69.3 ppm. Synthetic RNA was deaminated by incubation in an aqueous solution containing 1 M diethylene glycol dinitrite, 2 M pyridine and 1 M potassium thiocyanate (final concentrations) for 2 h, 5 h or 10 h at 37°C. The solution was diluted with buffer (10 mM Tris pH 7.4, 300 mM sodium chloride, 1 mM ethylenediaminetetraacetic acid) and the RNA was precipitated by addition of three volumes of cold ethanol and incubation in liquid nitrogen for 15 min. After centrifugation and washing with 75% ethanol, the pellet was resuspended in MilliQ water, and the integrity of the RNA was checked by analytical PAGE (stained by SYBR gold).

LC-MS

LC-MS/MS analysis was used to monitor the deamination levels of treated RNA samples. Prior to LC-MS/MS analysis, RNA samples (were digested into nucleosides according to the following protocol: samples were incubated in a buffer of 25 mM ammonium acetate (pH 7.5), 0.3 U nuclease P1, 0.1 U snake venom phosphodiesterase, 1 U fast alkaline phosphatase and 10 U Benzonase overnight at 37 °C. The digested samples were analysed on an Agilent 1100 HPLC series equipped with a diode array detector (DAD) and an Agilent 1100 mass selective detector (LC/MSD-Trap). A YMC Triart C18 column (S-3 µm/12 nm, 120 Å, column size: 150 × 3.0 mm I.D.) from YMC Europe (Germany) was used at 30°C column temperature. Mobile phase A (MPA) consisted of 5 mM ammonium acetate buffer adjusted to pH 5.4 using acetic acid and mobile phase B (MPB) of pure acetonitrile. The elution was performed at a flow rate of 0.4 mL/min using the following gradient: 3% MBP from 0 to 3 min, 3-15% MPB from 3 to 30 min, 15-90% MPB from 30 to 31 min, 90% MBP from 31 to 35 min, 90-3% MPB from 35 to 36 min and 3% MPB from 36 to 45 min. The effluent from the column was measured photometrically at 254 nm by the DAD for the detection of the nucleosides. In addition to the UV detection, a mass spectrometer was used for the identification of nucleosides, e.g. N^6-methyl-N^6-nitrosoadenosine (NOm^6A). Mass spectrometer settings used included ion polarity: positive; ion Source: ESI; dry temp: 350 °C; nebulizer: 15 psi; dry gas: 12 l/min; trap drive: 43.0; Octapole RF amplitude: 140.3 Vpp; capillary exit: 96.6 V; skimmer: 40.0 V; Oct 1 DC: 12.0 V; Oct 2 DC: 1.7 V; scan range: 105–600 m/z; averages: 2 spectra; maximum accumulation time: 200 ms; and ICC target: 30 000. Data was analysed with Bruker LC-MS data reading software (DataAnalysis) and nucleosides were monitored by multiple reaction monitoring (dynamic MRM mode).

Library preparation and sequencing

***Synthetic oligonucleotide*.** Around 500 ng of (treated) synthetic RNA oligonucleotide (*Table 1*) were used as starting material for library preparation. This includes all synthetic oligonucleotide samples, used directly as delivered, mixed or immunoprecipitated before library preparation. First, a two-step end-repair was performed, combining 5'- and 3'-dephosphorylation (Antarctic Phosphatase) and subsequent 5'-phosphorylation (Polynucleotide Kinase PNK), according to the manufacturer's instructions (New England Biolabs) and with addition of an RNase inhibitor (RNaseOUT, Invitrogen). The end-repaired RNA was purified with RNeasy MinElute Cleanup kit (Qiagen), using a larger quantity of ethanol to recover smaller RNA fragments. Library preparation was then prepared following the NEBNext Small RNA Library kit for Illumina (New England Biolabs): after ligation of the first adapter at the 3'-end of the end-repaired RNA, the reverse transcription primer was hybridized, followed by the second adapter ligation at the 5'-end of the RNA. The reverse transcription (RT) was performed, using ProtoScript II (in general) or SuperScript IV (for RT test - *Supplement Figure S11*), during 1 h at 50°C, according to the manufacturer's manual, directly followed by polymerase chain reaction (PCR) employing LongAmp Taq polymerase during 15 cycles at a primer annealing

temperature of 62°C. The indexing of the samples takes place at this step: a P7 primer was chosen among the 48 differently barcoded primers from the NEBNext Multiplex Small RNA Library Prep Set for Illumina (New England Biolabs). The P7 and P5 primer allow amplification of the templates and implementation of the necessary sequences for Illumina sequencing. A final clean-up and size selection step were performed to purify the library, applying the GeneJet PCR kit, according to the supplier's instructions.

***Amplicon sequencing with biological samples*.** Around 500 ng of (treated) RNA (messenger or total RNA) were incubated at 75°C for 5 minutes with 2.5 µM of targeted RT primers (*Table 2*). Reverse transcription was performed, using 10 U ProtoScript II (New England Biolabs) in addition with 0.5 mM dNTP mix, 10 mM DTT, 0.4 U RNase inhibitor, 1x ProtoScript II Buffer (final concentrations) and MilliQ water for 2 hours at 50°C. Then, polymerase chain reaction (PCR) was performed by mixing 5 U of LongAmp Taq DNA Polymerase (New England Biolabs) with 0.4 µM of targeted P5 primer (*Table 2*) and 0.4 µM P7 primer from the NEBNext Multiplex Small RNA Library Prep Set for Illumina (New England Biolabs) in addition to 300 mM dNTP mix, 1x LongAmp Taq Reaction Buffer (final concentrations) and MilliQ water. PCR was performed for 25 cycles at a primer annealing temperature of 50°C. The amplicon obtained was analysed *via* 10% denaturing polyacrylamide gel electrophoresis for 40 min at 10 W. The corresponding gel area (amplicon size), determined by GeneRuler Low Range DNA Ladder (Thermo Scientific), was cut out and eluted. The gel elution was performed overnight at 25°C in 0.5 M ammonium acetate solution, followed by Nanosep filtering (0.45 µm, VWR) and subsequent ethanol precipitation.

Table 2. DNA oligonucleotides for amplicon sequencing (purchased from IBA Lifesciences, Germany and Biomers, Germany)

Name	Sequence
Malat1 RT primer	5′-AGACGTGTGCTCTTCCGATCTNNNNNNNTGCTAGTCCTCAGGA-3′
Malat1 P5 primer	5′-AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCC GATCTAAGATCAAGAGTAAT-3′
18S RT primer	5′-AGACGTGTGCTCTTCCGATCTNNNNNNNCCTTCCGCAGGTTCA-3′
18S P5 primer	5′-AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCC GATCTGACGGTCGAACTTGA-3′
H RT primer	5′-AGACGTGTGCTCTTCCGATCTNNNNNNNCATATTCTTATTGCA-3′
H P5 primer	5′-AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCC GATCTCGGCCCGCCGTGTGT-3′
fl(2)d RT primer	5′-AGACGTGTGCTCTTCCGATCTNNNNNNNGTTTTGCTCGTATTT-3′
fl(2)d P5 primer	5′-AATGATACGGCGACCACCGAGATCTACACACACTCTTTCCCTACACGACGCTCTTCC GATCTGTCGCTGCAATGACT-3′

DNA library quantification and quality control. DNA libraries were quantified using the Qubit dsDNA HS (High Sensitivity) Assay Kit with the Invitrogen benchtop fluorometer Qubit 2.0. Each library (amplicon and synthetic oligonucleotide) was analysed, and quality evaluation by automated electrophoresis on the Agilent 2100 Bioanalyzer system was performed, following the Bioanalyzer High Sensitivity DNA Kit instructions.

Sequencing method. Libraries were subjected to high-throughput sequencing using an Illumina MiSeq instrument with single-read (SR50) or paired-end (PE75) mode. Prior to loading on the sequencing chip, the multiplexed libraries were diluted to 6-8 pM final concentration.

miCLIP. miCLIP was performed as described previously [540] using 10 µg of purified mRNA from *Drosophila melanogaster* S2R+ cells and 5 µg of anti-m^{6}A antibody

(Synaptic Systems). Immunoprecipitations were performed in quadruplicates and as a control one immunoprecipitation was performed where UV-crosslinking was omitted. Of note, this sample produced a library of limited complexity, reflecting a low amount of background mRNA binding. Briefly, total RNA was isolated using Trizol reagent (Invitrogen) and DNA was removed with DNase-I treatment (NEB). Polyadenylated RNA was purified by two rounds of binding to Oligo (dT)25 magnetic beads (NEB) and fragmented with RNA fragmentation solution (Ambion) using 1 µL of solution per 2 µg of mRNA and with 7 min incubation at 70 °C. Immunoprecipitation was performed at 4°C in 500 µL of binding buffer (BB) (50 mM Tris HCl pH 7,4, 150 mM sodium chloride, 0,5 % NP-40). First, isolated mRNA and antibody were incubated for 2h. Samples were then transferred to individual well of a 12-well cell culture plate and crosslinked on ice (two-times at 150 mJ/cm^2). Next, 60 µL of magnetic ProteinG beads (Invitrogen) were resuspended in 500 µL of BB and added to the IP sample. Samples were then incubated for additional 2h at 4 °C, before washing with ice-cold solutions was performed: 1x with BB, 2x with high salt buffer (50 mM Tris HCl pH 7,4, 1 M sodium chloride, 1 % NP-40, 0,1 % sodium dodecyl sulfate), 1x BB , 2x with PNK buffer (20 mM Tris HCl pH 7,4, 10 mM magnesium chloride, 0,2 % Tween). All washes were performed by gentle pipetting and with 1 min incubation on ice. Washes with HSB were additionally rotated for 2 min at 4 °C. Finally, beads were resuspended in 900 µL of PNK buffer. 40 µL were used for WB analysis to evaluate immunoprecipitation efficiency. Remaining 860 mL were used for library preparation. All steps of library preparation were performed as previously described in Sutandy *et al*. [575]. Libraries were amplified with 17 PCR cycles and sequenced on a NextSeq500 with a read length of 120 bp, single end (*Table 3*).

Table 3. DNA oligonucleotides for miCLIP (purchased from Integrated DNA Technologies, Inc., USA)

Name	Sequence
L3-Linker	5′-rApp-AGATCGGAAGAGCGGTTCAG-ddC-3′
miCLIP_IP1	5′-P-NNCTCGNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC-3′
miCLIP_IP2	5′-P-NNTGTGNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC-3′
miCLIP_IP3	5′-P-NNTTTCNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC-3′
miCLIP_IP4	5′-P-NNCGATNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC-3′
miCLIP_ctrl	5′-P-NNGACCNNNAGATCGGAAGAGCGTCGTGGATCCTGAACCGC-3′

Bioinformatic analysis

NOseq alignment. NOseq Illumina output demultiplexed FastQ files were unzipped. The nucleotide sequences (FASTA format), read information (identity number) and their corresponding quality scores (encoded in ASCII) were extracted, compiled and analysed using Python. The bioinformatic pipeline is single file dependent and each file is associated to a reference sequence. Prior to the alignment of the reads, the whole reference sequence was divided into smaller overlapping segments with identical size (k-mer of size k = 11 per default) and indexed by starting position. In order to increase the algorithm speed, k-mers were alphabetically ordered. The alignment process is relying on seed-and-extend strategy. Sequenced reads were individually segmented into k-mers of size k, scanned against the reference sequence k-mers and scored according to the asymmetric substitution matrix (*Supplement*

Figure S5). In case of matching seeds, the alignment was then extended as well as the score associated. Regarding the +1/-1 reward/penalty system and by allowing mismatches according to partial deamination (*Figure 1*), only the reads with a score equal to their length were kept. Furthermore, in order to avoid shifting of the reading frame, sequencing errors and to have a homogeneous coverage, only reads starting at the first nucleotide of the reference sequence were selected for further evaluation. In addition, this strict selection process is applied in order to limit the occurrence of NGS errors within the short sequencing reads [576]. Finally, the totality of perfectly aligned reads was used for alignment analysis.

NOseq data analysis. From the aligned reads, the number of adenosine (A), cytidine (C), guanosine (G) and thymidine (T) was summed for each position and corresponding frequencies were plotted using matplotlib python library. In addition, only A positions in the reference sequence were selected and associated A and/or G rates were plotted. From this analysis, the cumulative density function of the normal distribution was calculated for each A position using the python package scipy.stats.norm and plotted together with the detection threshold at 0.95 (red line) (see *Figure 2* as example). Above this limit, 5% of the most distant A rates to the overall distribution was observed, highlighting A sites with a high probability to behave differently to the other positions, i.e. presenting m^6A candidate sites under deamination conditions.

NOseq UMI assessment. NOseq data were analysed in order to determine the read redundancy by assessment of the unique molecular identifiers (UMI). The UMI sequence (7 nt) of every aligned read was determined and the dataset was checked for duplicates (reads with same UMI sequences).

***miCLIP data analysis*.** miCLIP Data analysis was performed as previously described in Linder *et al*. [540].

RESULTS

Screening of deamination conditions by LC-MS and sequencing

In order to distinguish m^6A from unmethylated adenosines in sequencing data, the major step was the conversion of adenosines into inosines to change the base-pairing properties in reverse transcription. According to reports about adenosine deamination by nitrous acid [566], we set up a deamination protocol for RNA and performed a condition screening with a synthetic 53mer containing m^6A at position 33 (*Table 1*). After optimizing a range of temperature, incubation time, and pH values, we identified conditions leading to complete deamination of RNA (70°C, pH 3.5, 60 min), as analysed by RP18 chromatography. Deamination included all exocyclic, primary amines, thus converting adenosines to inosines, cytidines to uridines, and guanosines to xanthosines (X) [576] in the process (*Figure 1a* and *Supplement Figure S1*). Using LC-MS, we detected, in addition to the above, a peak with an m/z=311, indicating partial formation of N^6-methyl-N^6-nitrosoadenosine (NOm^6A) from m^6A [567] to a maximum of 50-70% (*Supplement Figure S2*). The deaminated synthetic RNA template was then subjected to library preparation and first sequencing attempts. Thereby we expected a change in base-pairing properties for all deaminated and converted nucleotides, with unknown properties for NOm^6A (*Supplement Figure S3*). In the process of optimizing cDNA synthesis by reverse transcription and library preparation, it turned out that RNA degradation under the acidic deamination conditions rendered subsequent Illumina sequencing impracticable. However, by screening a set of in total 72 deamination conditions, resulting from permutation of different reaction temperatures, incubation times and pH values (*Table 1, Figure 1b)*, we identified a deamination treatment that allowed successful library preparation and sequencing of the synthetic RNA template (*Supplement Figure S4*). A range of conditions leading to partial deamination of RNA could be determined (with A-to-I conversion between 10 to 50%), within which the sequencing output was optimized in terms of aligned reads and m^6A signal strength, i.e. the data were evaluated for the fraction of aligned reads and maximum m^6A signal intensity. Importantly, partial deamination prevented the use of conventional alignment algorithms. As shown in *Figure 1c* (and *Supplement Figure S3*), native adenosine residues gave rise to either adenosine or guanosine signals, the latter deriving from the fraction having undergone deamination to inosine. Similarly, cytidines yielded cytidine or thymidine signals, respectively. At guanosine positions in the reference sequence, the analysis yielded a vast majority of guanosine signals, with a small remainder of adenosine. Of note, only native m^6A produced unaltered adenosine signals. These observations and the associated need for deliberately ambiguous alignments due to partial deamination required a specific alignment algorithm, since conventional mapping tools did not offer this capability.

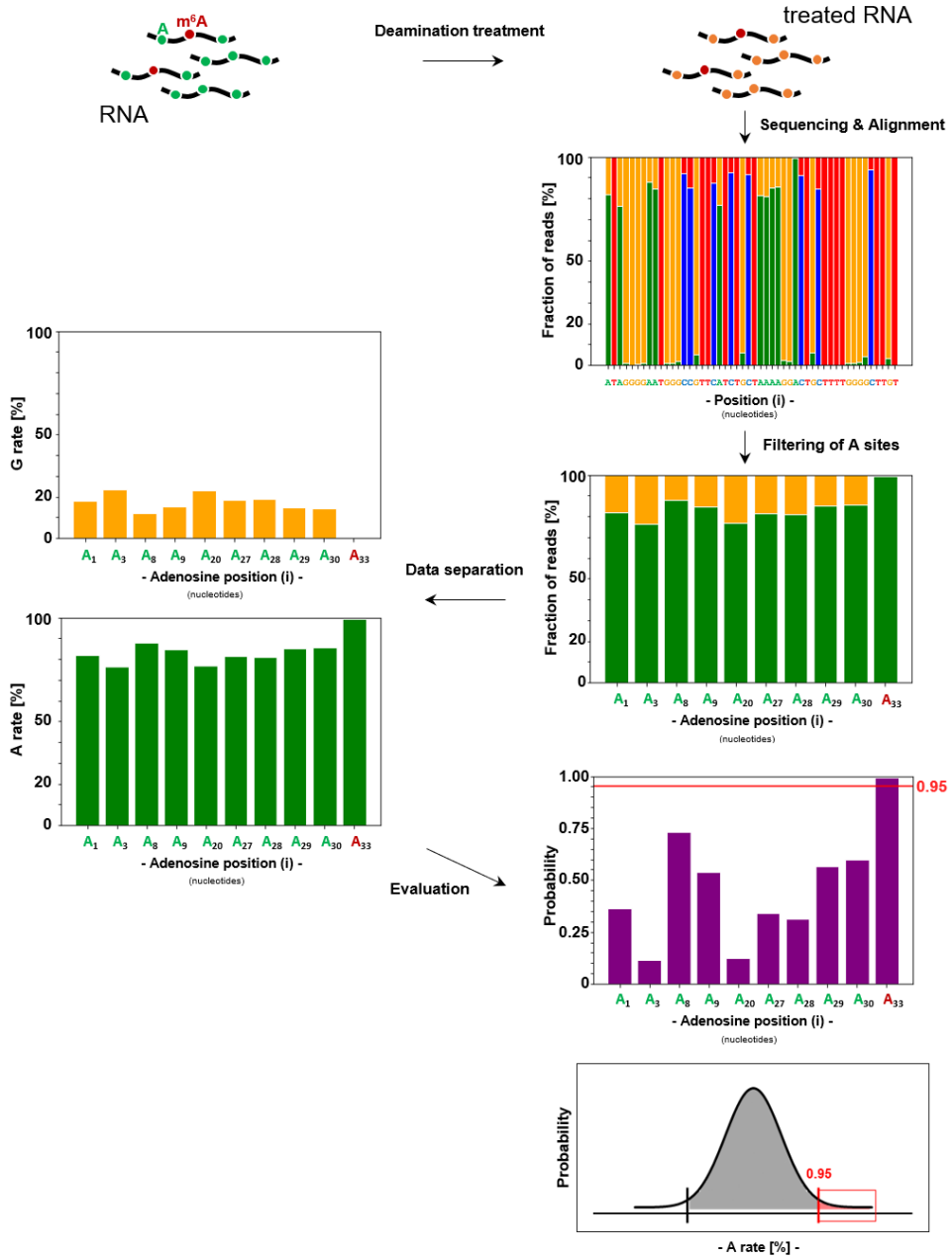


Figure 2: *General analysis pipeline of NOseq data (exemplary for deamination of the 53mer with a treatment of 50°C, pH 4.0 and 20 min). After deamination treatment the treated RNA is used for library preparation and sequenced. (for comparison to two-step alignment - see Supplement Figure S17). a) Reads are aligned to the reference with an asymmetrical substitution matrix (see Figure S5) to address partial deamination and the corresponding fraction of reads in % is plotted for each position in an alignment plot (A in green, G in orange, C in blue and T in red). b) Adenosine sites are filtered from the alignment plot. c) The fraction of reads in % is separated to generate adenosine and guanosine rate plots (separated adenosine and guanosine fraction of reads in % at A sites). d) The A rate is then used for evaluation of the m^{6}A site in a probability plot, which illustrates the probability of an A site being an m^{6}A, calculated by the distance of the respective A rate to the average A rate, with a red line at 0.95 showing the detection threshold.*

Alignment and analysis of NOseq data

To enable effective mapping of reads, from what we termed NOseq, an adapted mapping algorithm using a seed-and-extend strategy based on local alignment [569] was developed. At the core of this algorithm was a substitution matrix composed of score values that reward or penalize alignment options. We adapted these values to account for the possibility of mapping guanosine and adenosine onto any purine in the reference genome, and of thymidine (or uridine, respectively) onto any pyrimidine, as depicted in *Supplement Figure S5.* We obtained an alignment plot, showing the fraction of respective nucleotides for each position mapped along the reference sequence (*Figure 2a*). The samples were analysed by the degree of A-to-I conversion, expressed as the fraction of decreased adenosine and increased guanosine signals at adenosine (and prospective m^6A) sites (*Figure 2b*). In all but the weakest and strongest deamination conditions, the m^6A_{33} showed the highest adenosine and lowest guanosine (A-to-I deamination) signals. For further analysis we separated the adenosine and guanosine fractions into two individual plots, showing the adenosine and the guanosine rate respectively (*Figure 2c*). Using only reads that cover the entire sequence under investigation, the distance of the A rate (at each position) from the mean deamination yield (mean A rate) was used to identify potential m^6A candidates. Thereby a conventional probability plot was applied to highlight positions with a signal, significantly different from other adenosine sites (*Figure 2d*). The m^6A signal strength was then used to identify optimal parameters. In general, the weakest deamination conditions were not sufficient to achieve a distinction between A and m^6A, and the strongest conditions led to non-informative data output, similar to non-deamination of the templates, although LC-MS analysis clearly documented near-complete deamination (*Supplement Figure S6*). We hypothesized that the product of the slowest deamination reaction, namely the guanosine-to-xanthosine conversion, might interfere with the RT-based amplification step, thus selecting exclusively those few RNA molecules that have escaped deamination. This hypothesis was confirmed by primer extension experiments on a synthetic RNA fragment, which showed, that xanthosine residues indeed significantly slow down and partially block reverse transcription (*Supplement Figure S7*). Given that each RNA contains an individual number of guanosine residues, we considered the possibility that optimal deamination strength might vary among different sequences under investigation. However, our follow-up experiments for target-specific evaluation of m^6A sites by amplicon sequencing to investigate biological templates (*vide infra*), revealed that m^6A sites can be detected in different sequence contexts with the same deamination treatment and analysis pipeline, without additional optimization steps. A probable explanation is that under the milder conditions used, the xanthosine blockage plays a minor role, since the guanosine conversion is slower in comparison to those of cytidine and adenosine (*Supplement Figure S6*). Therefore, the probability for multiple guanosine-to-xanthosine deamination events are indeed very low, even in G-rich stretches, and such species do not significantly contribute to the overall sequencing output. This

assumption was also supported by our data on the synthetic oligo with multiple G-rich stretches (*Supplement Table S1*).

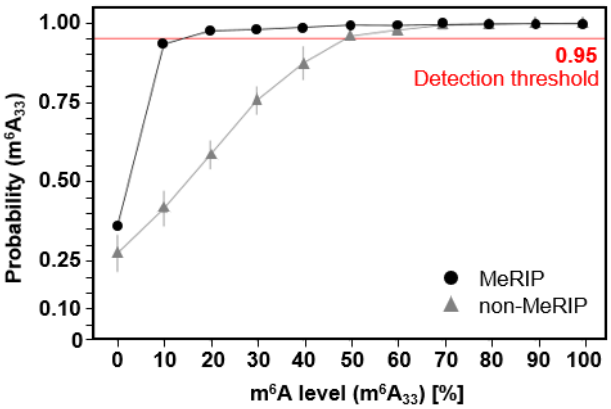


Figure 3: *NOseq data comparison (MeRIP enriched vs. non-MeRIP) to address its detection threshold for sub-stoichiometric m^{6}A with an m^{6}A content at position 33 (m^{6}A/A) of the 53mer that was varied from 0 to 100%. (black dots MeRIP, grey triangles non-MeRIP). Data was averaged from duplicates (error bars show standard deviations). Red line shows detection threshold at 0.95.*

m^{6}A detection and threshold determination

To evaluate the sensitivity of our method and address its detection threshold for sub-stoichiometric m^{6}A levels at biological m^{6}A sites, we performed a calibration by analysing the m^{6}A signal in the same synthetic 53mer, now with an m^{6}A content at position 33 (m^{6}A/A) that was varied from 0 to 100%. Partial deamination parameters (50°C, pH 4.0, 20 min), representing a compromise featuring optimized mapping yield versus adenosine deamination, were chosen for this test (*Figure 2*). Furthermore, decreased formation of NOm^{6}A (not shown) under these milder conditions is thought to avoid potential RT-arrest at m^{6}A sites. NOseq data showed a statistically valid m^{6}A signal in samples containing down to ~50% m^{6}A. This *de novo* detection limit could be lowered to ~10% by conducting a methylated RNA immunoprecipitation (MeRIP) enrichment [577] prior to NOseq (*Figure 3*). Such detection on the basis of antibody-enriched RNA fragments was tested with regard to application of NOseq on low-abundant RNA species as mRNA. Expectedly, the linear dependence of m^{6}A signal to m^{6}A content is lost as a consequence of this enrichment step. The m^{6}A signal in MeRIP-enriched RNA sequences therefore does not allow any conclusions to be drawn about the methylation level. However, once detected, a quantification of the m^{6}A site in a subsequent step by using a series of mixtures of synthetic oligos with defined amounts of m^{6}A seems practicable. Of such a calibration series, we determined the A signal at the m^{6}A site in the deaminated calibration samples and put it into relation to the respective m^{6}A content. Given the clearly linear correlation (*Supplement Figure S8*), we conclude that quantification of m^{6}A sites might, in principle, be possible in a two-step procedure, where the first step must identify the

position in a detection step, and the second step would require sequence-specific calibration for quantitative analysis. However, in contrast to straightforward validation of m^{6}A in amplicons (*vide infra*), such a calibration would have to rely on synthetic modified RNA, and require considerable effort in terms of time and resources.

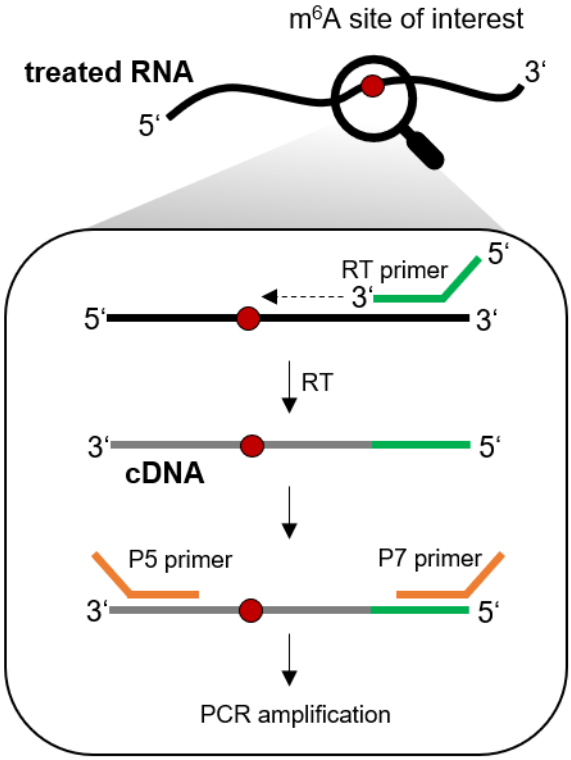


Figure 4: Library preparation scheme for amplicon sequencing of targeted m^{6}A sites. The target region is amplified in a two-step process. First, reverse transcription of the deaminated RNA is performed with a target-specific RT primer. The latter contains the UMI and the binding sequence for the universal P7 primer including the barcode for Illumina sequencing. Afterwards, a target specific P5 primer and the P7 primer are used for PCR amplification.

Investigation and validation of biological m^{6}A sites by amplicon sequencing

We applied NOseq to five biological m^{6}A sites to demonstrate the applicability of the method. To this end, we developed an amplicon sequencing method for target-specific evaluation of m^{6}A sites (*Figure 4*). Of note, library preparation for the above calibration with synthetic RNA was based on an adapter ligation, while biological samples required the design of targeted primers, compatible with the Illumina sequencing technology, for each m^{6}A site under investigation. In a two-step process, including reverse transcription and PCR amplification, sequencing libraries were prepared. To address partially deaminated RNA templates, we also tested primers with complementary degenerated target sequences. The extended binding possibilities yielded unspecific sequence amplification (not shown) and this strategy was therefore not further pursued. Rather, as a compromise of signal-to-noise *versus* sequencing

output, we turned to conventional primers and used milder deamination conditions (60°C, pH 5.0, 10 min, *Supplement Figure S9*), thus obtaining sufficient PCR product for RNAseq. Assessment of unique molecular identifiers (UMIs) in typical amplicons (*Supplement Figure S13 and Supplement Table S1*) revealed an average read redundancy around 10% (*Supplement Table S2*), showing the samples to be representative and essentially free of significant PCR-amplification bias.

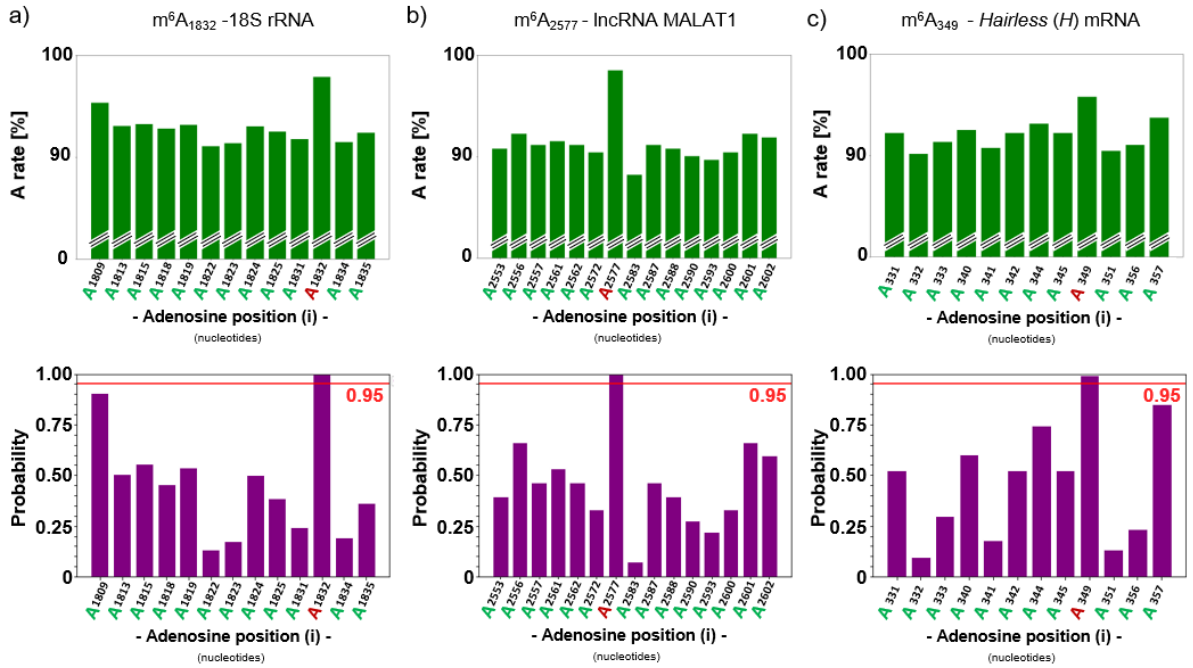


Figure 5: *Adenosine rate (A fraction of reads in % at A sites) and probability plots of three biological amplicons, showing the progressed A-to-I conversion and corresponding m⁶A site evaluation (red line shows detection threshold at 0.95). a) m⁶A at position 1832 in human 18S rRNA (HEK293T cells). b) m⁶A at position 2577 in human lncRNA MALAT1 (HeLa cells). c) m⁶A at position 349 in Drosophila melanogaster H mRNA after MeRIP enrichment. (Supplement Figure S14 & 15 for additional information, including non-treated references).*

An initial experiment was targeted at the highly abundant and well-described m^{6}A site at position 1832 in human 18S rRNA [578]. Under the selected conditions we were able to clearly distinguish and detect the m^{6}A site from the surrounding A sites (*Figure 5a*). Of note, plotting C-to-U (T-rate) and A-to-I (G-rate) changes across the amplicon sequence revealed a uniform distribution of deamination events, including the direct proximity to the m^{6}A target (*Supplement Figure S15*). This demonstrated that the use of conventional primers did not incur deamination biases in the amplicon.
Without changes, these conditions for target-specific evaluation of m^{6}A sites were subsequently applied to other m^{6}A positions in different sequence contexts. As another m^{6}A site for evaluation, the m^{6}A at position 2577 in the human lncRNA MALAT1 was analysed. Again, the same conditions led to clear distinction between m^{6}A and the surrounding A sites (*Figure 5b*). From the clear signals without prior MeRIP enrichment, the occupancy of the sites in rRNA and lncRNA was gauged to

exceed 50% (determined detection threshold – *Figure 3*), which is in keeping with values obtained by other methods [578]. As third validation, we applied NOseq in combination with MeRIP [577] to scrutinize m^{6}A candidate sites in mRNA from *Drosophila melanogaster*. The first investigated m^{6}A candidate site was selected at position 349 in the 5'-untranslated region (UTR) of the protein-coding mRNA *Hairless* (*H*) in *Drosophila melanogaster*, which plays a role in Notch signalling [579]. This candidate site produced a strong signal in miCLIP experiments (*Supplement Figure S10a*) and was thus validated by NOseq as genuine m^{6}A site (*Figure 5c*). Finally, a further validation was conducted on a sequence region comprising two m^{6}A candidate sites drawn from miCLIP experiments (*Supplement Figure S10b*) in the coding sequence of the mRNA *female-lethal-2-d* (*fl(2)d*) in *Drosophila melanogaster,* an associated component of the methyltransferase complex mediating m^{6}A methylation [579]. Both sites were detected *via* NOseq. As an ultimate validation, NOseq signals from these sites were compared to those obtained with RNA from an *Ime4* knockout mutant. We had previously characterized this knockout of the homolog of mammalian METTL3 protein, and found mRNA from this mutant to contain drastically reduced levels of m^{6}A in polyA-RNA [579]. Correspondingly, neither of the two m^{6}A candidate sites were detectable by NOseq anymore in the respective amplicon (*Supplement Figure S16)*, validating the detection of these m^{6}A sites in mRNA.

DISCUSSION

In summary, we developed an m^{6}A detection approach based on chemical deamination of RNA. The pre-treatment with sodium nitrite in an acidic milieu and under heat influence enables discrimination of adenosines and m^{6}A in RNA-Seq data. Chemical deamination of adenosines leads to a conversion to inosines and therefore to a certain number of guanosine signals in sequencing due to changes in reverse transcription properties. However, treatment does not lead to altered sequence information at m^{6}A sites, which can hence be identified by lack of deamination products, corresponding to a high A rate, at a given position. Because xanthosine as a deamination product of guanosine effectively blocks reverse transcription, we had to resort to partial instead of complete deamination, as is known from bisulfite sequencing. Therefore, the method is not yet suitable for transcriptome-wide studies, but might be thus developed, if the xanthosine block can be overcome. One option could be a polymerase variant that bypasses xanthosine with higher efficiency. However, given that we already tested other RT enzymes, specific engineering or *in vitro* evolution might be necessary [555]. Our method allows exchange of the RT enzyme without further adaptations (*Supplement Figure S11*). In this perspective, we have also successfully applied an organic nitrite as less acidic nitrosylation reagent (diethylene glycol dinitrite), with results similar to nitrous acid, but with decreased degradation (*Supplement Figure S12*). The detection of m^{6}A sites with different methylation levels is possible, and the detection limit of the method can be further reduced by enrichment of m^{6}A fragments with m^{6}A-specific antibodies prior to the

treatment, whereby even adenosines with very low methylation level (about 10%) can be differentiated.

A quantification of methylation stoichiometry in m^6A sites subsequent to their validation by NOseq is made plausible by our proof-of-principle data on synthetic RNA. This would comprise synthetic template mixtures with ascending m^6A content in the sequence context in question. The sequence-specific calibration differentiates the procedure from known methods [555]. While plausible, it would admittedly require major efforts and an individual calibration for every m^6A site in its particular sequence context. However, if corresponding calibration repetitions in different sequence contexts were producing consistently linear correlations, that might plausibly allow a transfer and application of linear calibrations (such as shown in Figure S8) from *bona fide* sites to new sites in question, and thus eliminate the need for subsequent sequence-tailored validations. However, the number of such analyses required for sufficient confidence is well outside the scope of this work.

Importantly, we here prove the applicability of our amplicon-based detection in as many as five biological m^6A sites, two of them previously known, namely in human 18S rRNA and lncRNA MALAT1. Importantly, the other three led to validations of new sites in mRNA from *Drosophila melanogaster.* We anticipate that this approach will find application in the RNA research field by allowing site-specific m^6A evaluation in various organisms and RNA species. In our experience, the method is easy to apply, and the specific primer target sequences can be easily adapted for other m^6A sites under investigation (*Supplement Figure S13*). The required sequencing depth is rather low, with verified m^6A detectability in our amplicon experiments based on a few hundred reads. Given a designed amplicon, an assay with modest requirements for experimental skills is readily established and enables straightforward comparative assessment of RNA samples in large numbers.

ABBREVIATIONS

A selected list of abbreviations is available in the Supplementary Data.

AVAILABILITY

The bioinformatic script and test datasets used for this study are available in the Supplementary Data (zip file).

Test data: Sample1.fastq – 53mer (m^6A_{33}) - 50°C, pH 4.0, 20 min
Sample2.fastq – 18S rRNA amplicon - 60°C, pH 5.0, 10 min

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We are grateful to Dr. Lilia Ayadi and Valérie Igel-Bourguignon (EpiRNA-Seq Core Facility, UMS2008 IBSLor, CNRS-Université de Lorraine-INSERM, Nancy, France) for excellent technical assistance in MiSeq sequencing, the IMB Core Facility Genomics for their helpful support and use of its Illumina NextSeq machine (INST 247/870-1 FUGG), and Anke Busch from the IMB Core Facility Bioinformatics for analyses of

miCLIP data. We want to further thank Dr. Lyudmil Tserovski for his contributions during early stages of this project.

FUNDING

This work was supported by the Deutsche Forschungsgemeinschaft (DFG) to M.H./J.-Y.R. [HE3397/17-1, RO4681/6-1 & RE4193/1-1], C.H. [HO 4436/2-2] and J.K. [KO4566/3-2] as well as the SPP1784. This work was also supported by the FRCR EpiARN project from Lorraine Region to Y.M. and was performed in the framework of the EPITRAN COST initiative (CA16120).

AUTHOR CONTRIBUTIONS

M.H. designed the study. S.W., A.G. and designed and performed experiments and interpreted the data. A.G. created and performed the bioinformatic analysis pipeline. F.P. and T.K. contributed to the bioinformatic analysis. V.M. and Y.M. performed RNA-Seq. T.L., J.K. and J.-Y.R. performed miCLIP and delivered *Drosophila melanogaster* stocks. M.V.S. and C.H. delivered data on xanthosine and performed experiments with diethylene glycol dinitrite. A.H., J.-Y.R., T.L., J.K., Y.M., V.M. and C.H. contributed to the paper with helpful discussions and proofreading of the paper. S.W., A.G. and M.H. wrote the article. M.H. supervised the work.

CONFLICT OF INTEREST

The authors declare no conflict of interests.

Supplementary Material

List of abbreviations (selected)

lncRNA	long non-coding RNA
m5C	5-methylcytosine
m6A	*N*6-methyladenosine
MeRIP	Methylated RNA Immunoprecipitation
miCLIP	m6A Individual-nucleotide-resolution Cross-Linking and Immuno-Precipitation
mRNA	messenger RNA
NOm6A	*N*6-methyl-*N*6-nitrosoadenosine
NOseq	m6A deamination sequencing
PCR	Polymerase Chain Reaction
rRNA	ribosomal RNA
RT	Reverse Transcription
SAM	S-adenosylmethionine
SCARLET	Site-specific Cleavage And Radioactive-labelling followed by Ligation-assisted Extraction and Thin-layer chromatography

Table S1. *Investigated sequences and sequence regions (amplicon sequencing) for NOseq. m⁶A candidate sites are shown in brackets.*

Name	Sequence
Oligo (m⁶A33)	5'-AUAGGGGAAUGGGCCGUUCAUCUGCUAAAAGG (m⁶A) CUGCUUUUGGGGCUUGUAGU-3'
MALAT1 (m⁶A2577)	5'-UACCAA**C**UUAAUGUUUUUGCAUUGG (m⁶A**)** CUUUGAGUUAAGAUUAUUUUUUAAA-3'
18S rRNA (m⁶A1832)	5'-CUAUCUAGAGGAAGUAAAAGUCGUA (m⁶A) CAAGGUUUCCGTAGG-3'
H (m⁶A342)	5'-GCCUGUGAAAUUGGUUAAAUAACGU (m⁶A) CAGUGGAAUGGGCCAAUGGCCAAAA-3'
fl(2)d (m⁶A451 & m⁶A472)	5'-AUGG (m⁶A**)** CGAUCAAAGACCCUGUAUGA (m⁶A**)** CAGUUACGACAAAAUGCCCCCGACC-3'

Table S2. *UMI analysis of NOseq amplicon sequencing data.*

Name	Read redundancy (%)
MALAT1 (m⁶A2577)	8 %
18S rRNA (m⁶A1832)	10 %
H (m⁶A342)	17 %
fl(2)d (m⁶A451 & m⁶A472)	5 %
Average	10 %

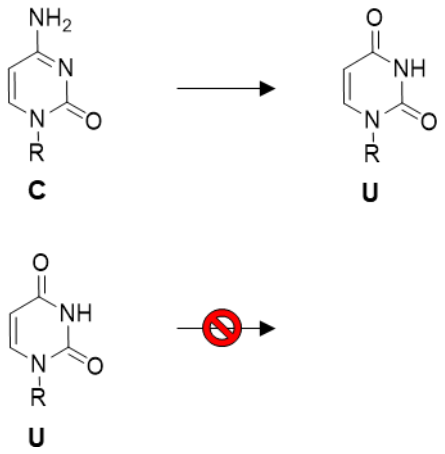


***Figure S1**. Reaction of cytidine and uridine under deamination treatment and their reaction products. (See Figure 1 for adenosine, m^{6}A and guanosine).*

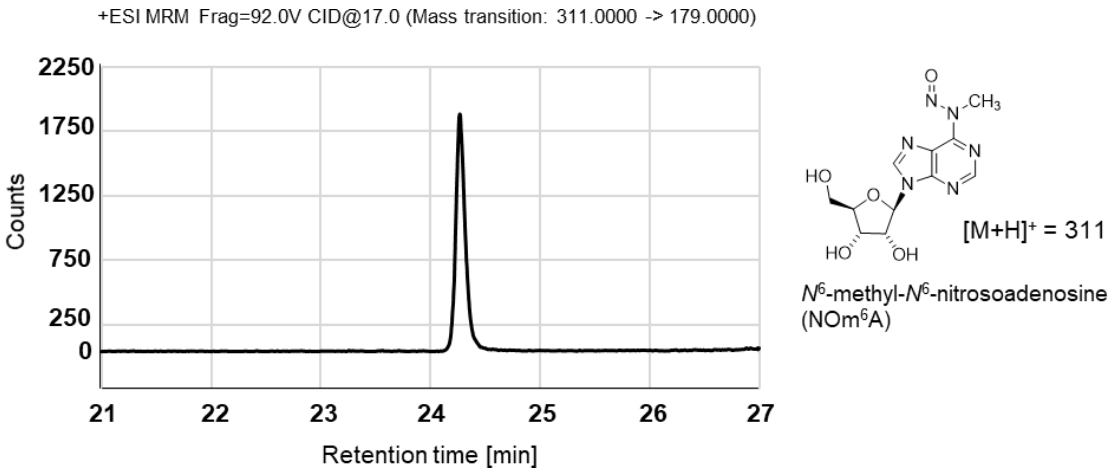


***Figure S2.** Extracted ion chromatogram (EIC) from LC-MS analysis of N^6-methyl-N^6-nitrosoadenosine (NOm^{6}A), showing a peak at 24.3 min with a mass transition of 311 to 179.*

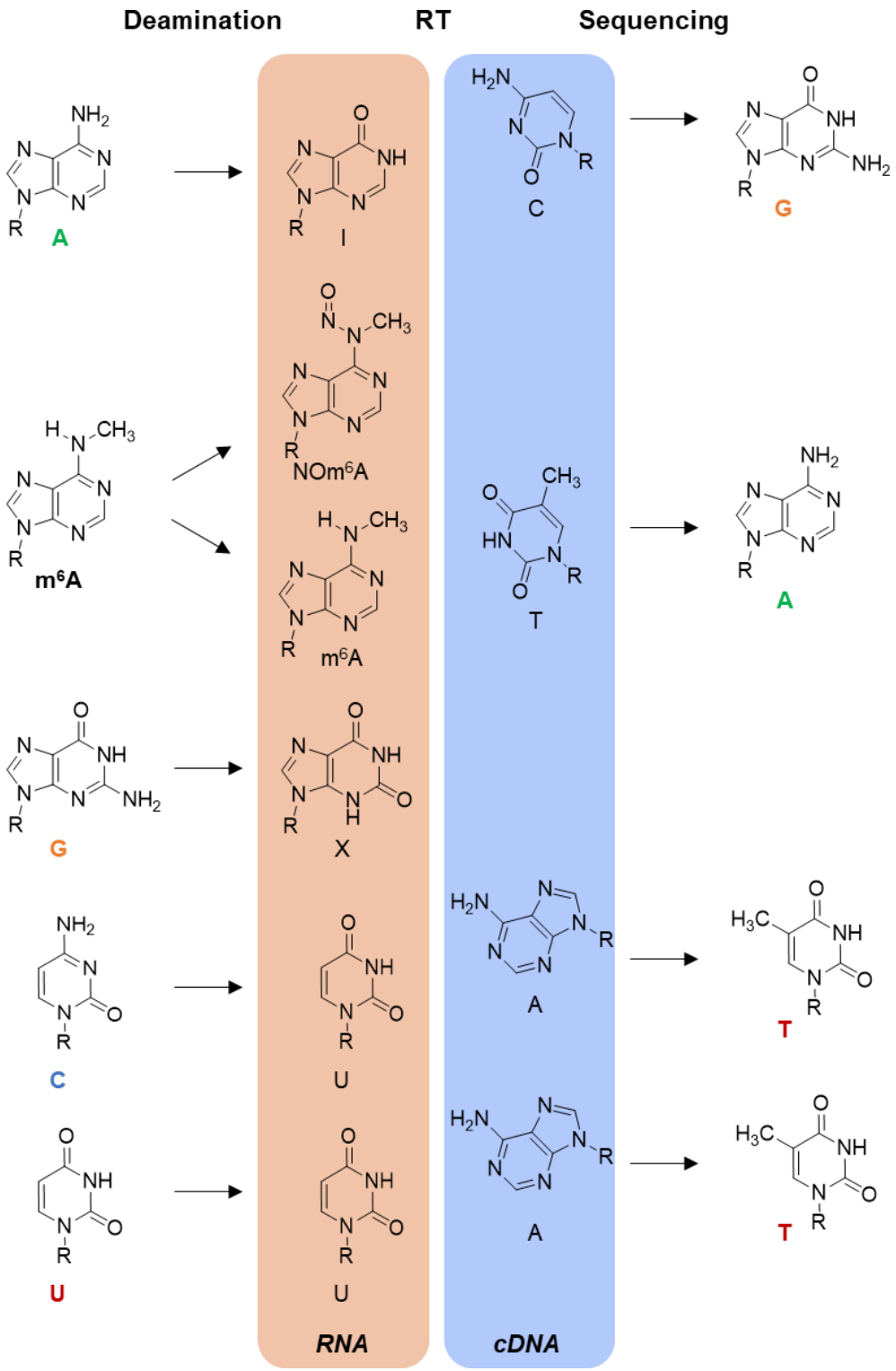
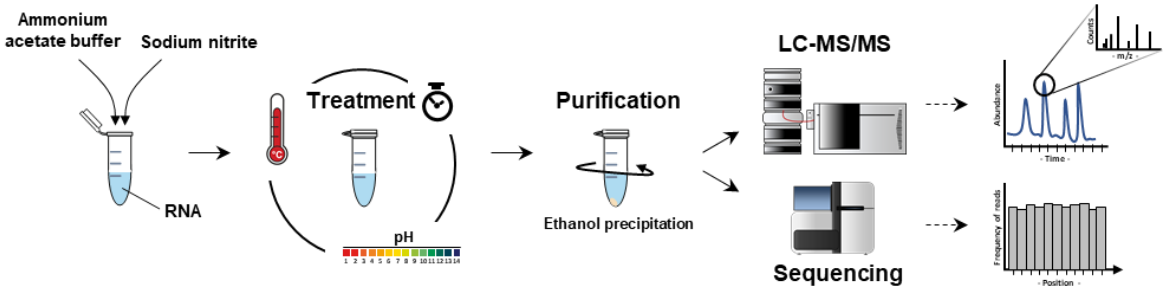


***Figure S3**. Base conversion after deamination treatment and following influence on reverse transcription (RT), base-pairing properties and corresponding sequencing output.*

a)



b)

5'-ATAGGGGAATGGGCCGTTCATCTGCTAAAAGGACTGCTTTTGGGGCTTGTAGT-3'

Figure S4. *Deamination treatment and following screening workflow. a) Deamination screening workflow. b) Sequence of synthetic 53mer with m^6A at position 33 (red underline).*

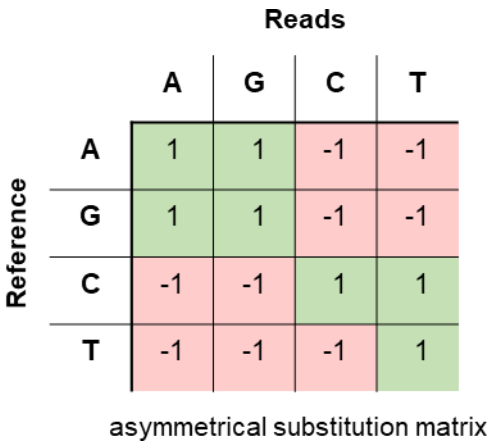


Figure S5*. Asymmetrical substitution matrix for alignment of partially deaminated reads (green boxes show rewards of +1 for the alignment score, red boxes show penalties of -1 correspondingly).*

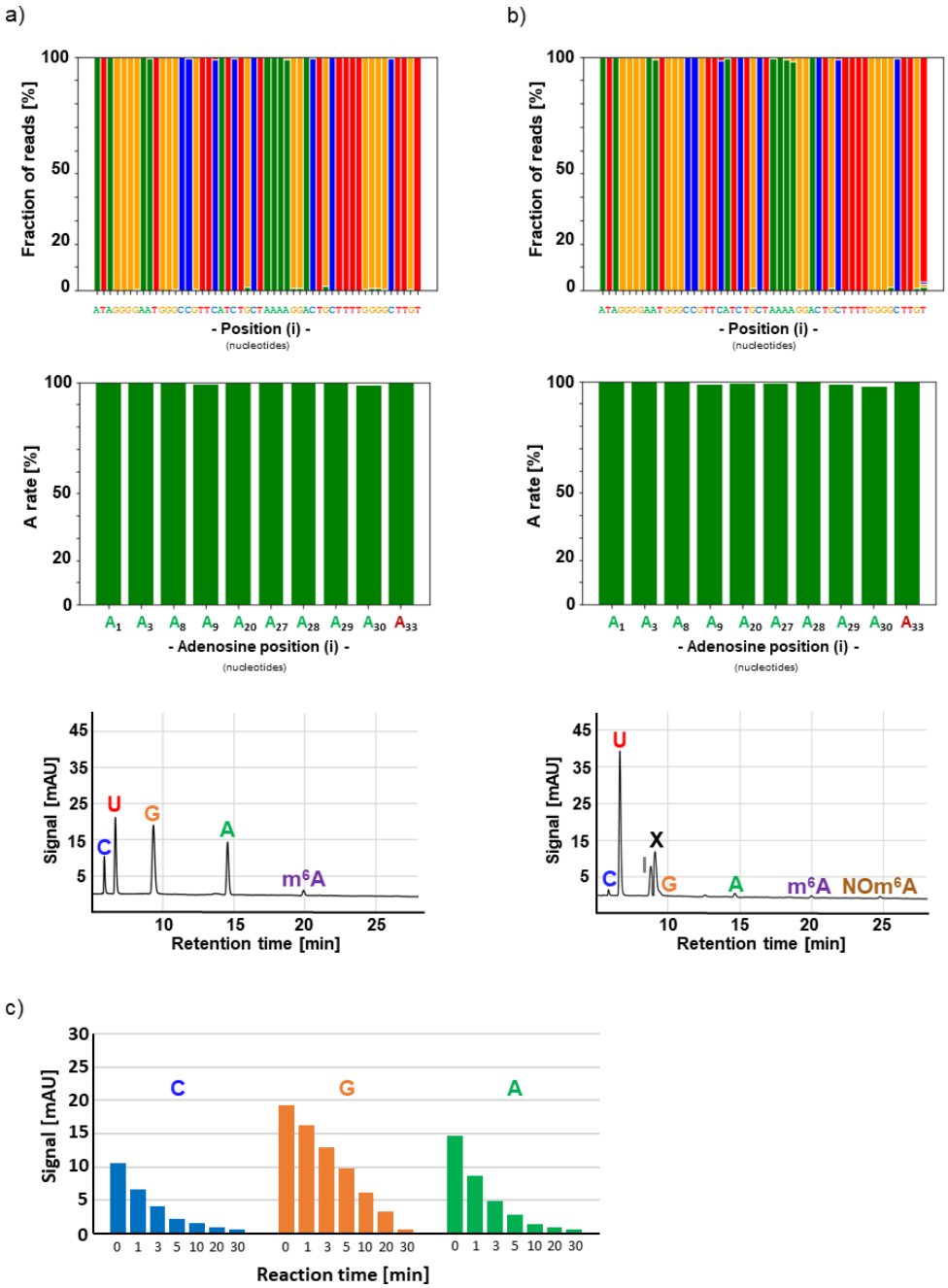


Figure S6. *Strong deamination of the 53mer (70°C, pH 3.5 and 30 min treatment) and the corresponding NOseq and LC-MS data. a) Non-treated reference. Shown are the alignment plot (with fraction of reads, plotted for each position (A in green, G in orange, C in blue and T in red), the adenosine rate plot and the LC-MS chromatogram. b) Strongly deaminated sample. Shown are the alignment plot (with fraction of reads, plotted for each position (A in green, G in orange, C in blue and T in red), the adenosine rate plot and the LC-MS chromatogram. c) LC-MS graph (exemplary for one reaction condition set) showing the course of the deamination reaction of C, G and A under strong deamination conditions (70°C, pH 3.5) at different time points.*

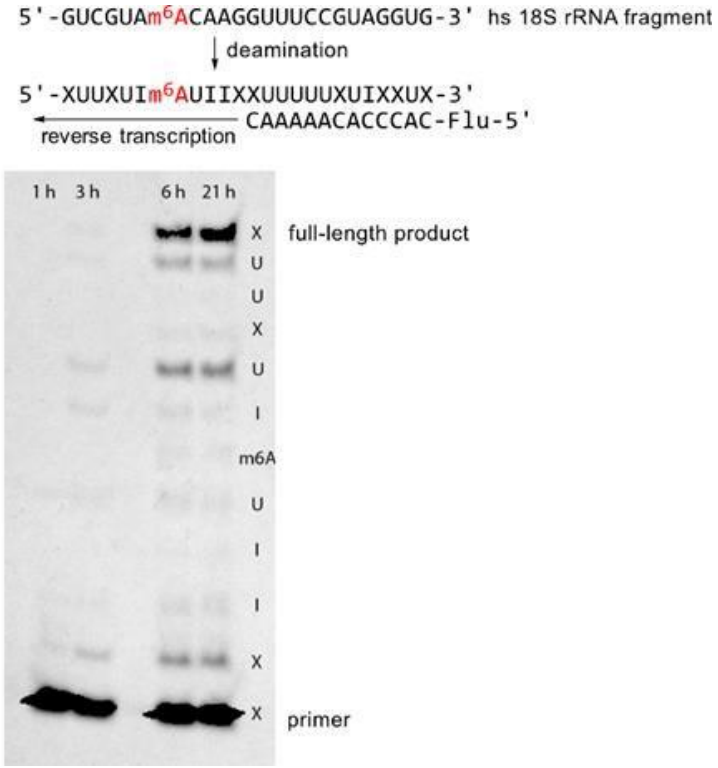


***Figure S7**. Primer extension on a deaminated 24mer fragment of human 18S rRNA containing m^{6}A1832. The RNA was treated with 1 M diethylene glycol dinitrite in 2 M pyridine, 1 M KSCN, at 37°C for 5 h and isolated by ethanol precipitation. Digestion and analysis by RP-HPLC revealed that all adenosines and cytidines were deaminated to inosine and uridine, respectively, and 50% of guanosines were converted to xanthosines. A 5'-fluorescein-labelled primer complementary to the deaminated RNA was annealed, and primer extension reaction with 0.5 mM dNTPs and 0.1 U/µL M-MuLV RT was incubated at room temperature for up to 21 hours. Aliquots were analyzed on a 20% denaturing PAGE (7 M urea) gel at 35 W. The full-length extension product was produced only after prolonged incubation time and significant abort bands were visible before xanthosines.*

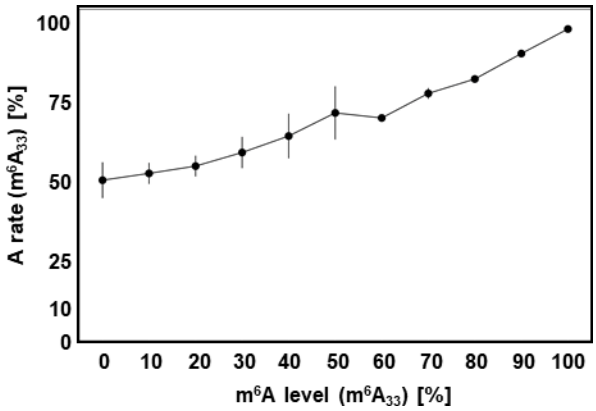


***Figure S8**. NOseq data comparison to investigate the quantification of previously detected m^{6}A sites by calibration with deaminated synthetic oligonucleotide mixtures (50°C, pH 4.0 and 20 min), containing ascending m^{6}A contents at position 33 (m^{6}A/A) of the 53mer that was varied from 0 to 100%. The A rate [%] was plotted. Data was averaged from triplicates (error bars show standard deviations; data points for 60, 80 and 90% m^{6}A content derive from one replicate).*

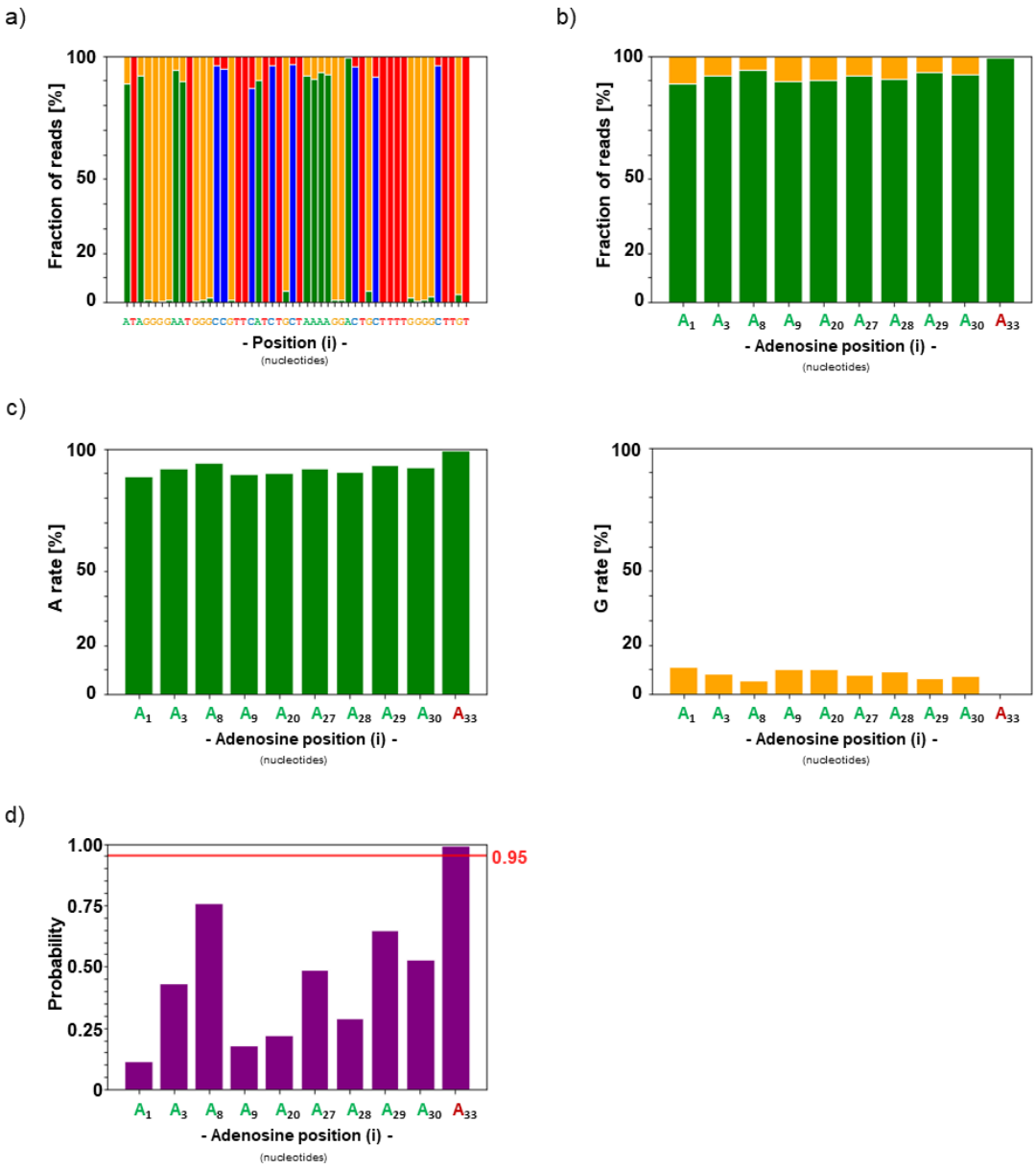


***Figure S9**. Plots for deamination condition screening (60°C, pH 5.0 and 10 min treatment) for the 53mer. a) Alignment plot (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red). b) Filtered adenosine site alignment plot. c) Adenosine rate and guanosine rate plots (separated adenosine and guanosine fraction of reads in % at A sites).*
d) Probability plot (red line at 0.95 showing the detection threshold).

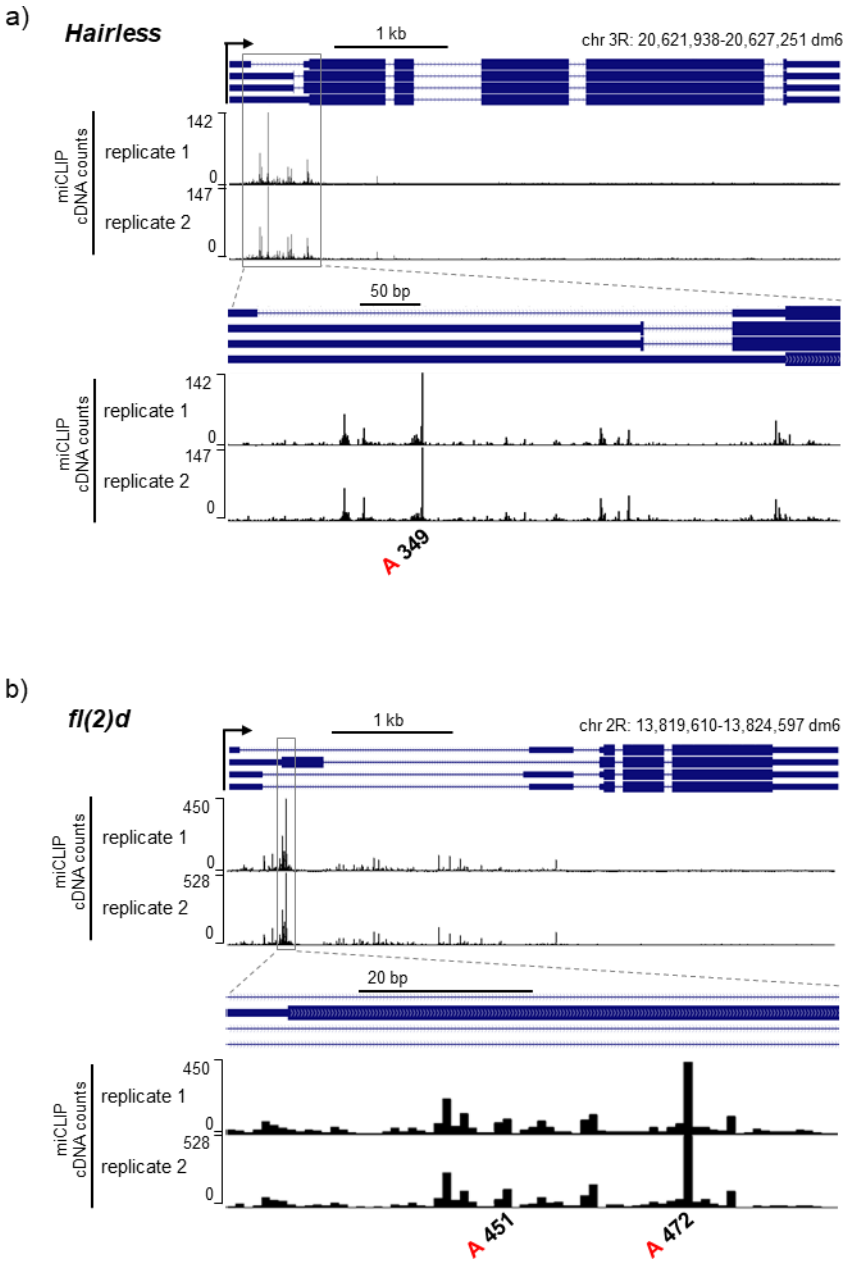


Figure S10. *UCSC Genome Browser screenshots of Hairless and fl(2)d transcript showing miCLIP tracks with cDNA counts of predicted m^{6}A positions. a) Gene architecture of Hairless is shown on top, with thin blue boxes representing the 5' and 3' UTRs, thick blue boxes representing the CDS, and thin lines representing introns. Zoom-in view of selected region is shown below with indicated miCLIP cDNA count peak signal of putative m^{6}A site (A349). b) Gene architecture of fl(2)d is shown on top, with thin blue boxes representing the 5' and 3' UTRs, thick blue boxes representing the CDS, and thin lines representing introns. Zoom-in view of selected region is shown below with indicated miCLIP cDNA count peak signal of two putative m^{6}A sites (A451 and A472).*

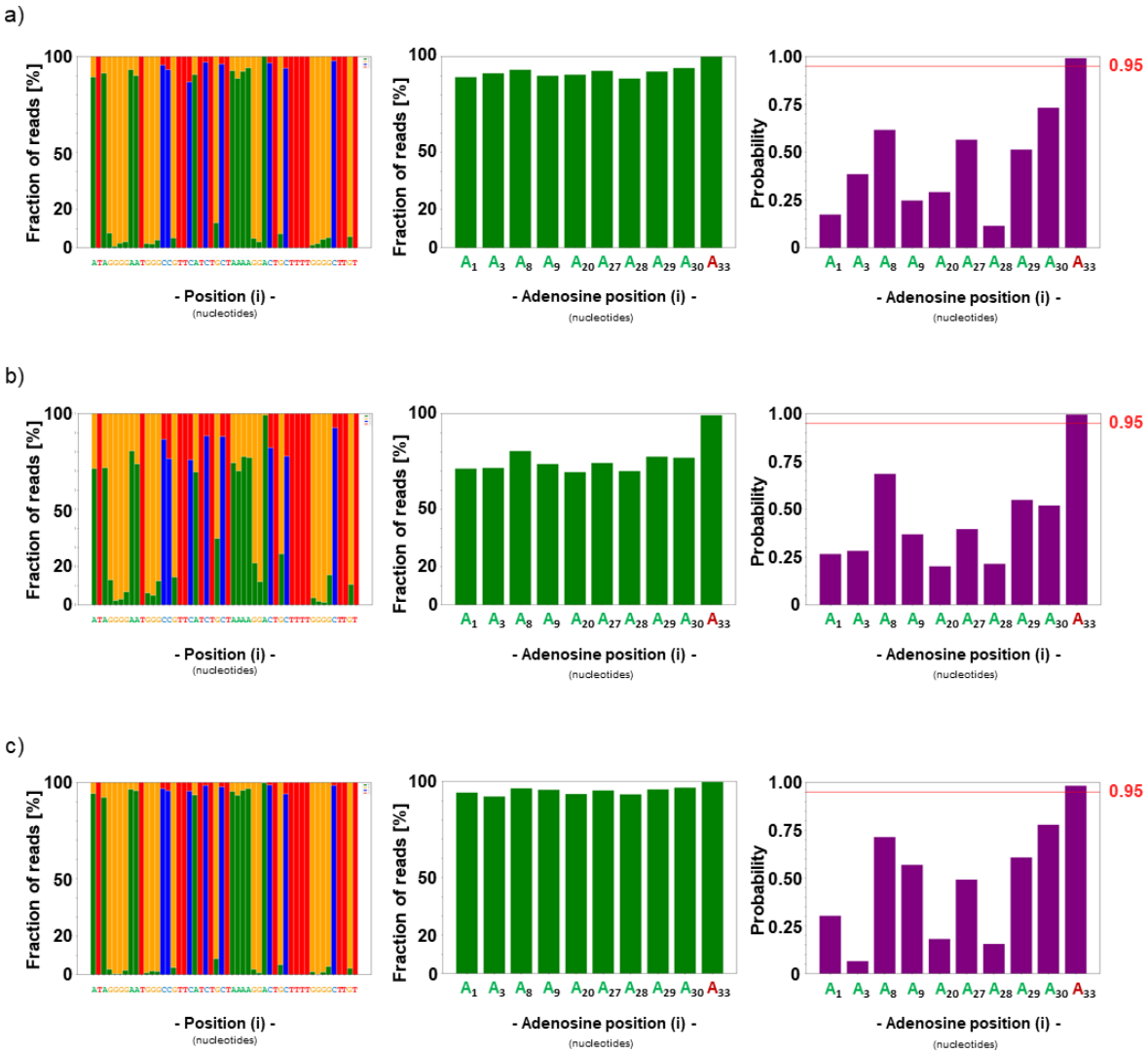


Figure S11. *Plots for deamination condition screening (50°C, pH 4.0) for the 53mer with SuperScript IV (Thermo Fisher Scientific).*

a) Alignment plot (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red), Adenosine rate plot (separated adenosine fraction of reads in % at A sites) and probability plot (red line at 0.95 showing the detection threshold) for 3 min treatment. b) Alignment plot (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red), Adenosine rate plot (separated adenosine fraction of reads in % at A sites) and probability plot (red line at 0.95 showing the detection threshold) for 10 min treatment. c) Alignment plot (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red), Adenosine rate plot (separated adenosine fraction of reads in % at A sites) and probability plot (red line at 0.95 showing the detection threshold) for 30 min treatment. RT was performed according to the manufacturer's manual.

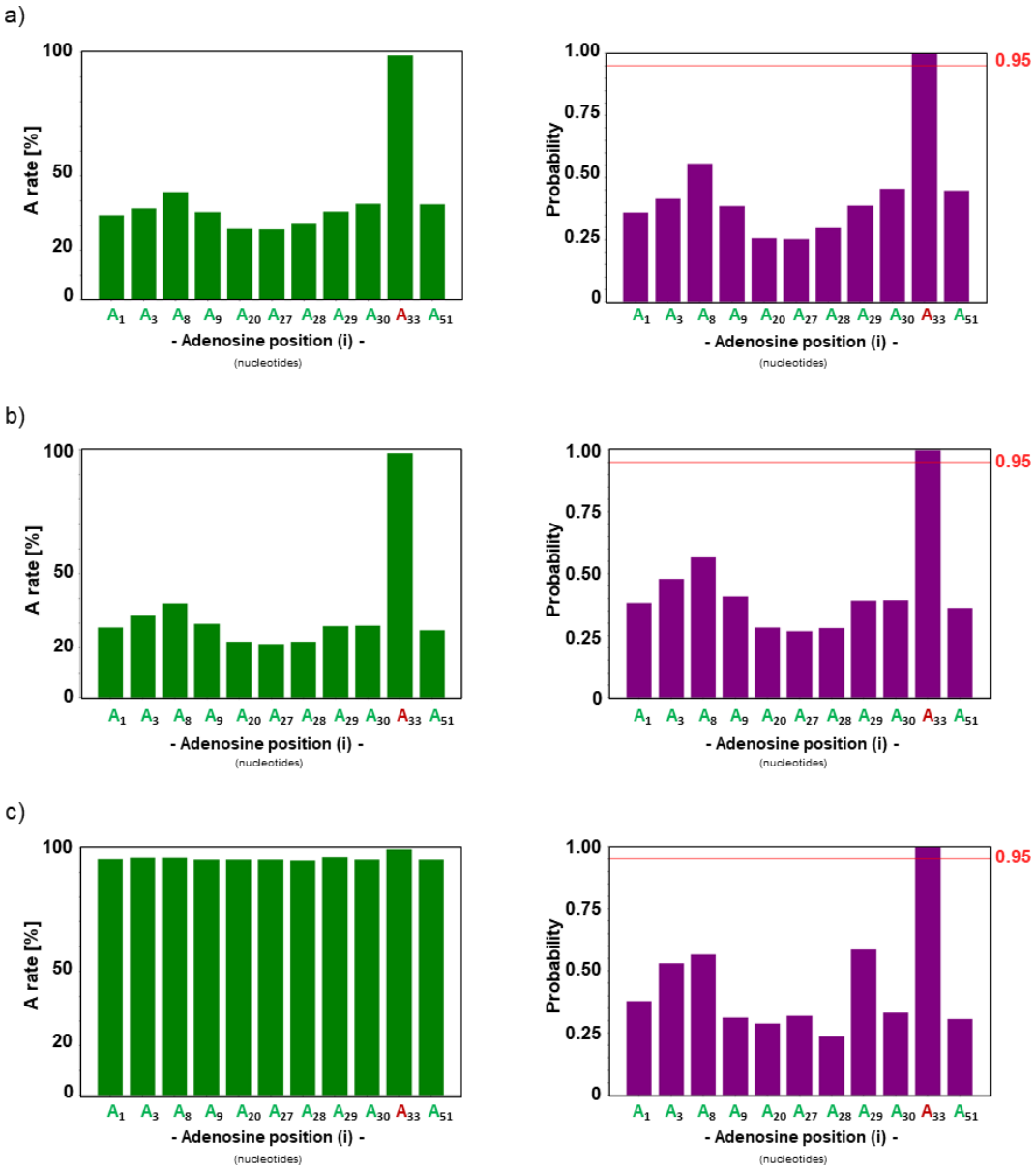


Figure S12. *Deamination of 53mer with diethylene glycol dinitrite. Shown are the adenosine rate (separated adenosine fraction of reads in % at A sites) and probability plots. a) 2 h treatment. b) 5 h treatment. c) 10 h treatment.*

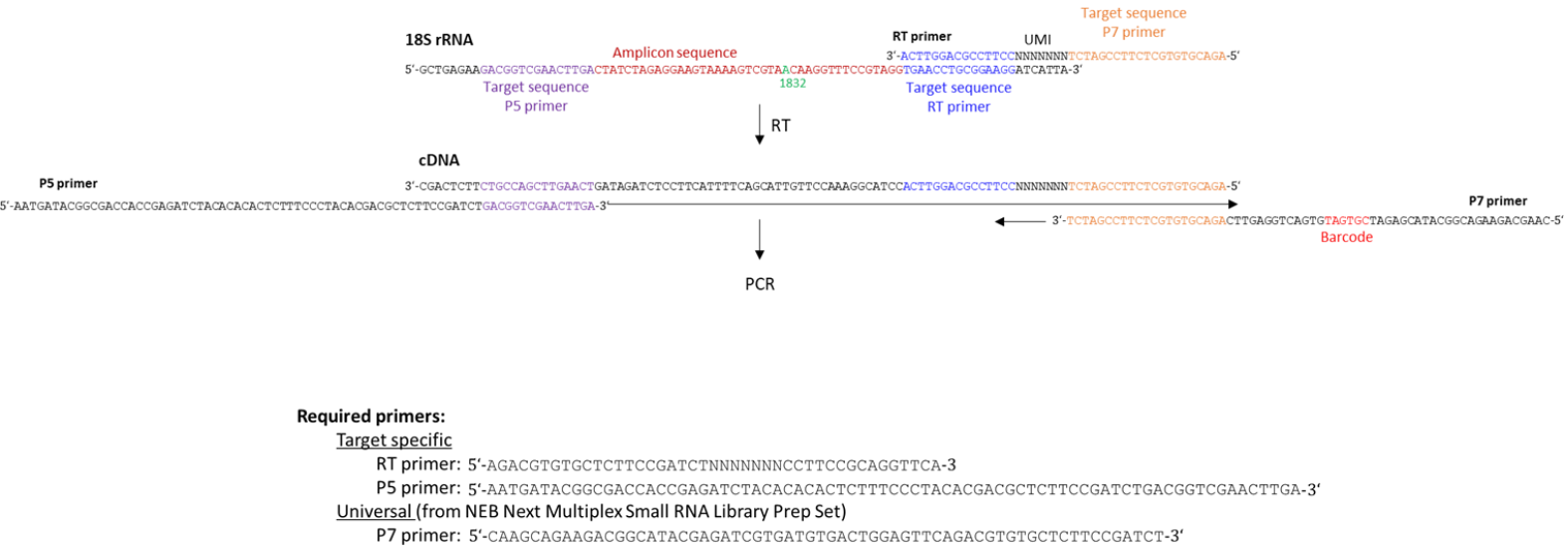


Figure S13. *Primer design for a typical amplicon (red) for amplicon sequencing (exemplary for m^{6}A1832 in human 18S rRNA). Shown are the target-specific sequences (purple and blue) around the m^{6}A site (green) in the amplicon sequence (red). Shown in blue is the target sequence for the RT primer, containing the unique molecular identifier and the universal target sequence for the P7 primer (orange) from the NEBNext Multiplex Small RNA Library Prep Set. After RT, the target specific P5 primer and the universal P7 primer (including the barcode sequence) are added and the amplicon is amplified via PCR. For other m^{6}A sites, the target-specific sequences for the RT primer (blue) and the P5 primer (purple) need to be adapted (see primer sequences in Table 2). The amplicon sequences containing the other investigated m^{6}A candidate sites can be found in Supplement Table S1).*

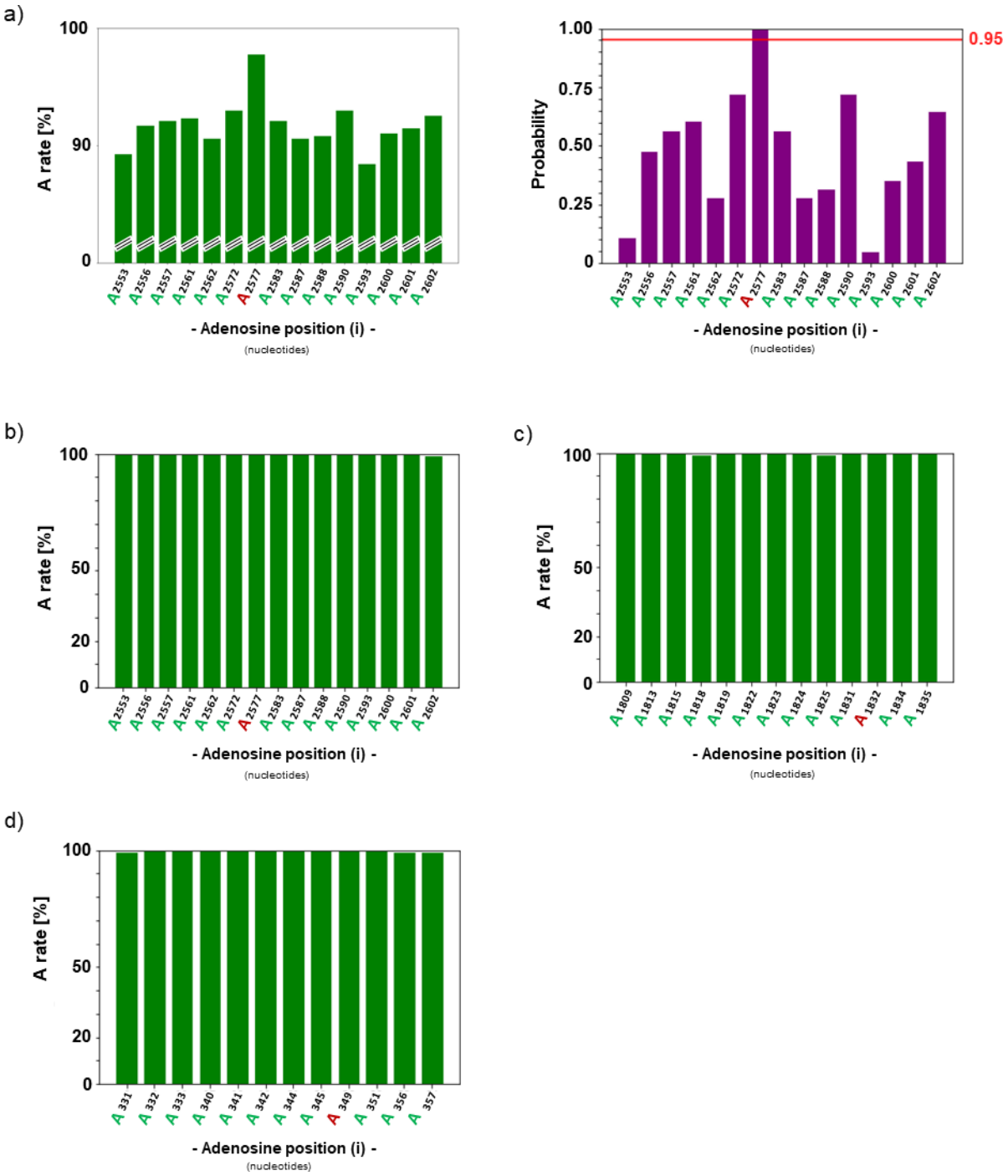


Figure S14. *Additional data for biological samples. a) Adenosine rate (separated adenosine fraction of reads in % at A sites) and probability plot of human lncRNA MALAT1 (HEK293T cells). b) Adenosine rate plot of the non-treated reference of human lncRNA MALAT1 (HeLa cells). c) Adenosine rate plot of the non-treated reference of human 18S rRNA (HEK293T cells). d) Adenosine rate plot of the Drosophila melanogaster H non-treated reference after MeRIP enrichment.*

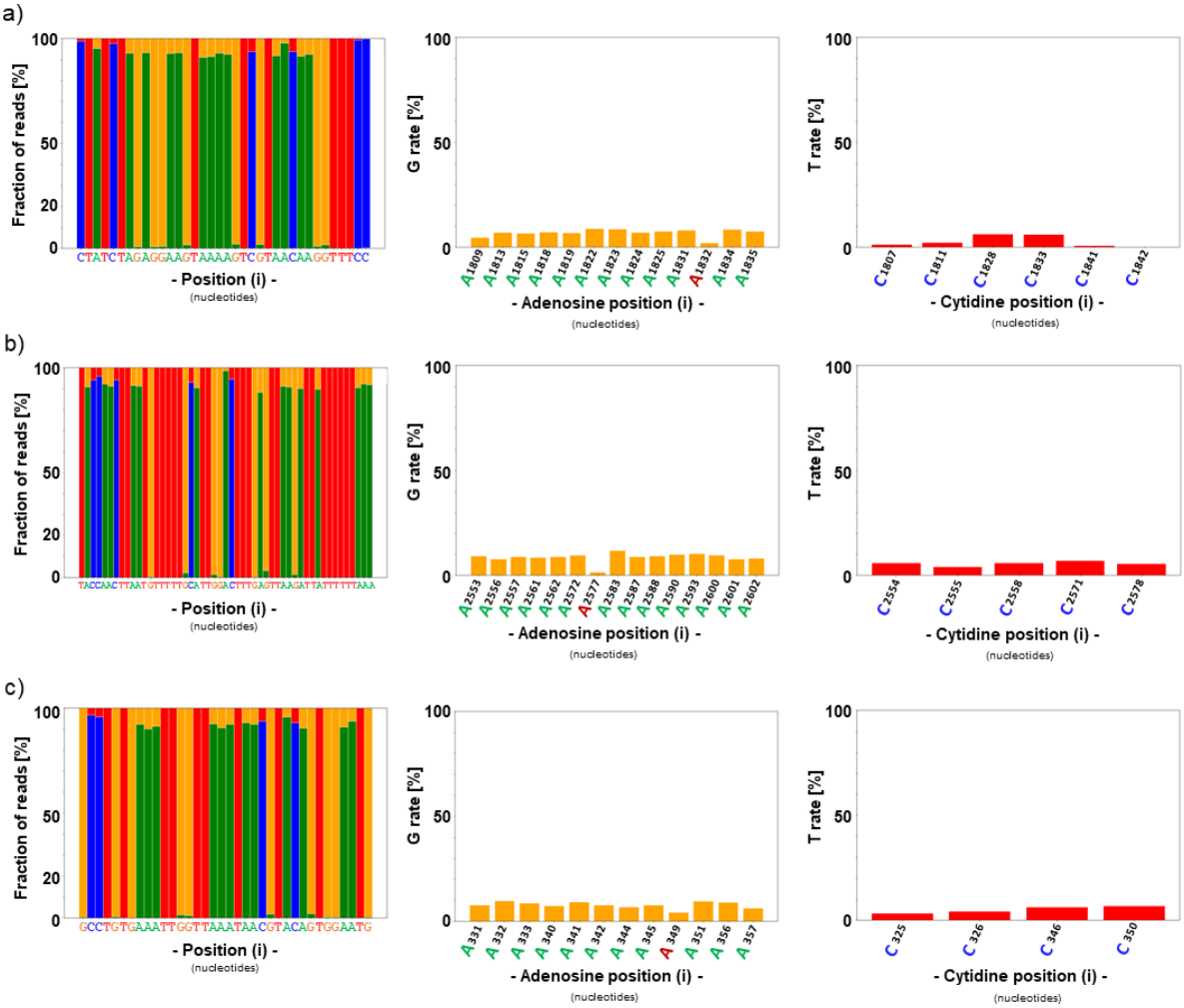


Figure S15. *NOseq data for biological samples. a) 18S rRNA. Alignment (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red), Guanosine rate (separated Guanosine fraction of reads in % at A sites) and T rate plot (separated Thymidine fraction of reads in % at C sites). b) MALAT1 lncRNA. Alignment (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red), Guanosine rate (separated Guanosine fraction of reads in % at A sites) and T rate plot (separated Thymidine fraction of reads in % at C sites). c) H mRNA. Alignment (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red), Guanosine rate (separated Guanosine fraction of reads in % at A sites) and T rate plot (separated Thymidine fraction of reads in % at C sites).*

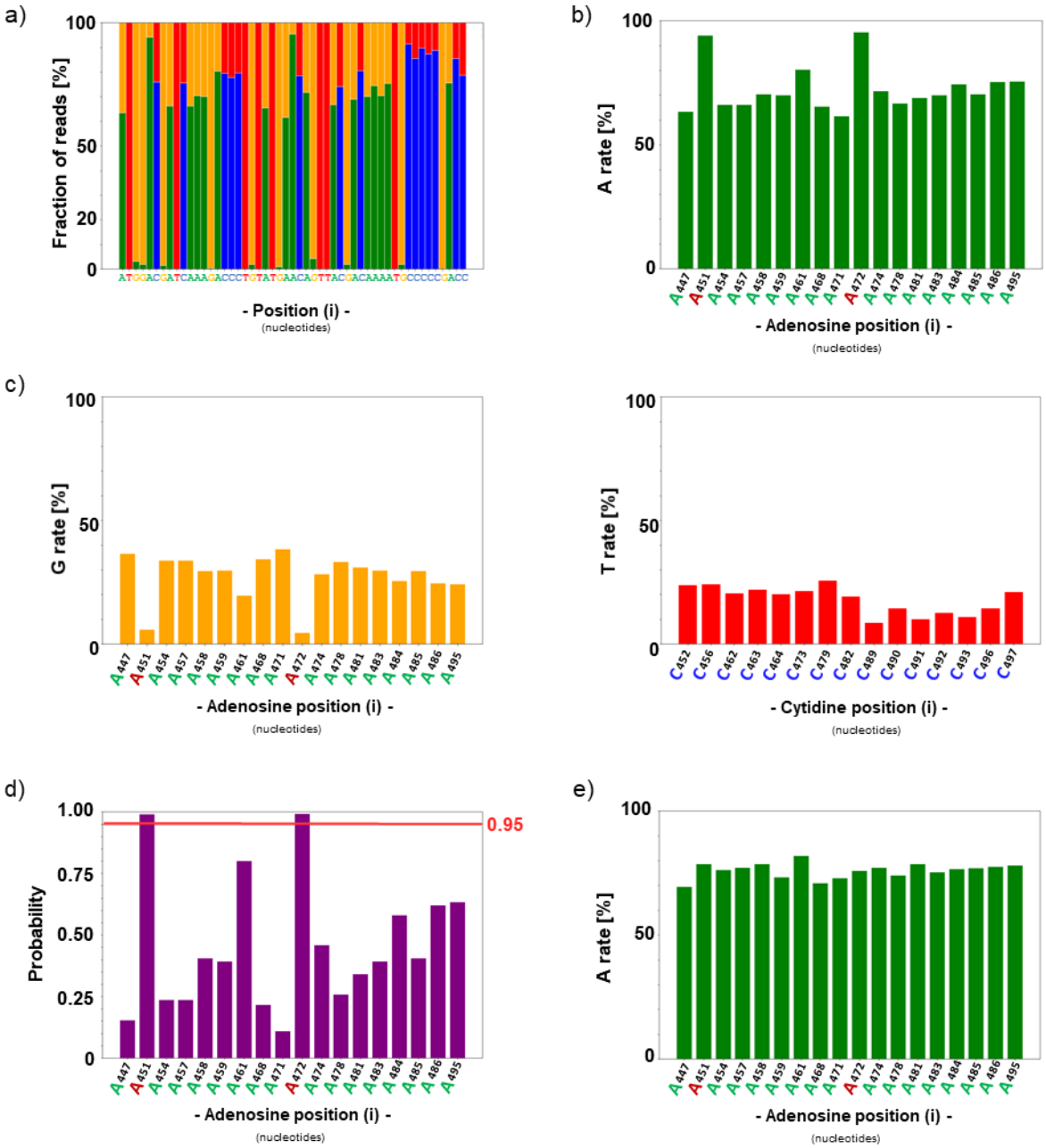


***Figure S16**. NOseq data for Drosophila fl(2)d mRNA (from wildtype flies for a) - d), and IME4- knockout flies for e)). a) Alignment plot (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red). b) Adenosine rate (separated Adenosine fraction of reads in % at A sites). c) Guanosine rate (separated Guanosine fraction of reads in % at A sites) and T rate plot (separated Thymidine fraction of reads in % at C sites). d) Probability plot (red line at 0.95 showing the detection threshold). e) Adenosine rate plot (separated Adenosine fraction of reads in % at A sites) of Drosophila fl(2)d mRNA (from IME4- knockout flies).*

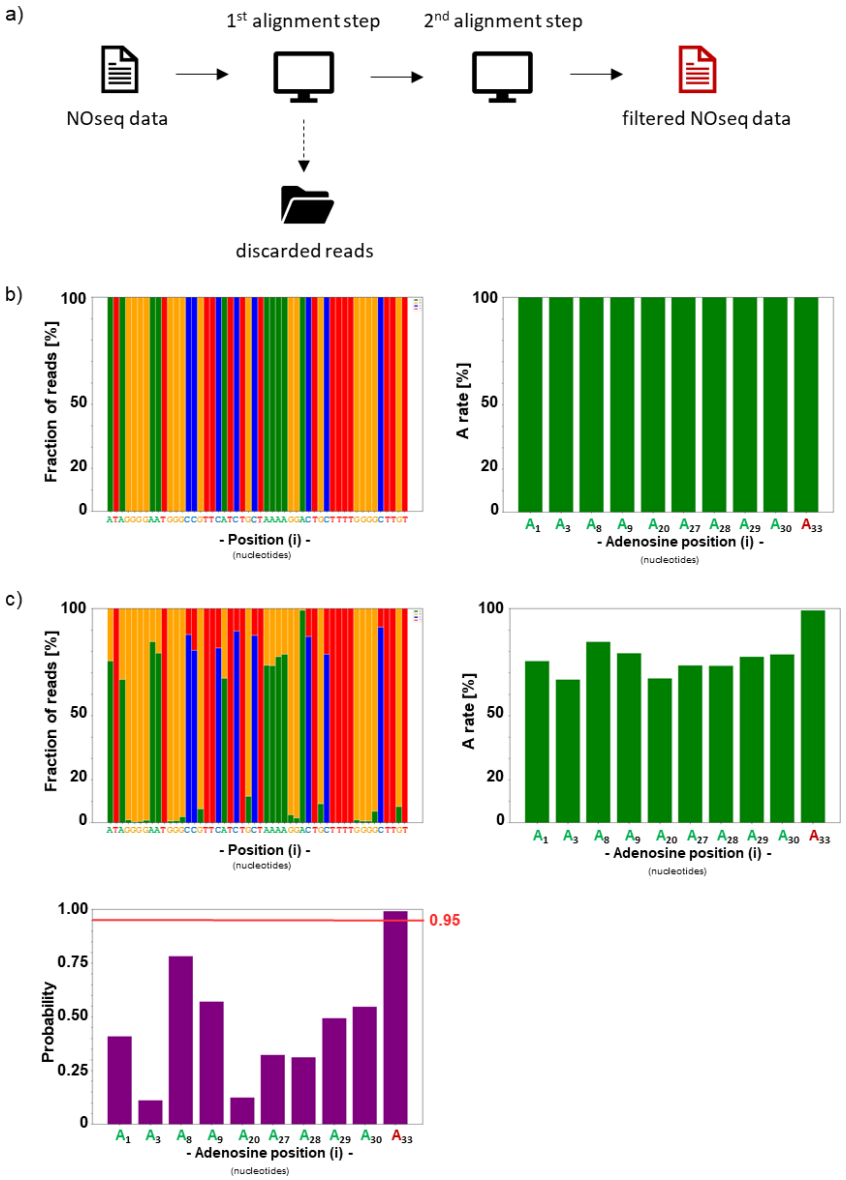


Figure S17. *NOseq alignment test (two-step alignment), exemplary for deamination of the 53mer with a treatment of 50°C, pH 4.0 and 20 min (for comparison to standard alignment - see Figure 2). a) Analytical pipeline. NOseq data is aligned in a 1st alignment step to the reference sequence (allowing no mismatches and no multiple alignment). Perfectly aligned reads are filtered out and discarded from the dataset. The reduced dataset is then aligned in a 2nd alignment step with the NOseq alignment, according to the asymmetric substitution matrix and the custom alignment pipeline (Supplement Figure S5). b) Filtered, discarded reads from 1st alignment step. Alignment (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red) and Adenosine rate plot (separated Adenosine fraction of reads in % at A sites). c) Reduced dataset used in 2nd alignment step. Alignment (with fraction of reads in %, plotted for each position (A in green, G in orange, C in blue and T in red), Adenosine rate (separated Adenosine fraction of reads in % at A sites) and Probability plot (red line at 0.95 showing the detection threshold).*

-- End of article --

Final chapter: Compiled application of the developed methods

4.3 Quantification of substoichiometric modification reveals global tsRNA hypomodification, preferences for angiogenin-mediated tRNA cleavage, and idiosyncratic epitranscriptomes of human neuronal cell-lines. [DOI: 10.1016/j.csbj.2022.12.020]

Florian PICHOT[1,2], Marion C. HOGG[3,#], Virginie MARCHAND[2], Valérie BOURGUIGNON[2,4], Elisabeth JIRSTRÖM[3], Cliona FARRELL[3], Hesham A. GIBRIEL[3], Jochen H. M. PREHN[3], Yuri MOTORIN[2,4], and Mark HELM[1]

[1] Institute of Pharmaceutical and Biomedical Sciences, Johannes Gutenberg University Mainz, Staudingerweg 5, 55128 Mainz, Germany.

[2] Université de Lorraine, CNRS, INSERM, IBSLor (UAR2008/US40), Epitranscriptomics and RNA Sequencing Core Facility, F54000 Nancy, France.

[3] Department of Physiology and Medical Physics and SFI FutureNeuro Research Centre, Royal College of Surgeons in Ireland, St. Stephen's Green, Dublin, D02 YN77, Ireland.

[4] Université de Lorraine, CNRS, IMoPA (UMR7365), F54000 Nancy, France.

[#] Present address: School of Science and Technology, Nottingham Trent University, Nottingham, NG1 4FQ, United Kingdom

Overview

This last manuscript is based on a collaborative work with Pr. Prehn's laboratory which studies the molecular mecanisms behind neurodegenerative diseases and more specifically, the impact of tRNA fragments (tsRNAs) induced by angiogenin on brain cells. Angiogenin is an RNA endonuclease which cleaves tRNAs at the anticodon-loop generating tRNA halves (tiRNAs) during cell stress. Link between tRNA fragmentation and neurodegenerative diseases has already been identified[580] but remains elusive. However, previous studies have shown that angiogenin cleavage is influenced by the presence of tRNA modifications, highlighting the fact RNA modifications have a role towards neurodegenerative diseases pathogenicity. In order to better understand the role of modifications in tRNA fragmentation, an epitranscriptomic profile of tRNAs and their relative tsRNAs in human brain cells has been carried on here.

In this regard two tumor cell lines derived from brain cells have been used as models for the study which has been treated (or not for control) with angiogenin. To establish a throughput profiling of tRNA and tsRNA epitranscriptome of these cells, different next-generation sequencing detection methods have been used. 5-methylcytidine (m^5C) detection has been carried out by bisulfite sequencing, 2'-O-methylation by RiboMethSeq and 7-methylguanosine (m^7G), 3-methylcytidine (m^3C) and dihydrouridine (D) by AlkAnilineSeq. An attempt for pseudouridine detection by HydraPsiSeq has unfortunately failed due to insufficient reads' coverage.

To ensure trustworthiness of the results, only putative sites located within known enzymatic targets by transcriptional modification machinery have been chosen as valid hits. This leads to studying the different reported sites with high confidence. Also, cartography of tRNA modifications has been conceived using epitranscriptomics-adapted tRNA reference discussed in previous chapter (Section 4.1.3) to study poorly reported tRNA sequences and so indicative unreported sites.

The stoichiometry assessment of the different reported sites through quantification scores have displayed clear tRNAs/tsRNAs modification level differences as previously seen in other papers, but mainly on a subset of sites and not on a global scale. Although tRNA and tsRNA stoichiometry differences is the major effect observed within the provided data, we also observed stoichiometry differences between the two cell lines and between angiogenin treated and untreated samples. These potential markers, especially the angiogenin ones, would need to be validated and studied by orthogonal methods to further investigate their roles in neurodegenerative diseases. Of note dihydrouridine on position 47 seems to promote tRNA cleavage, while other modifications are known to protect tRNA from cleavage. Validation of such observation would be truly beneficial for neurodegenerative diseases pathogenicity and so their cure.

For this work, I carried on the treatment, analysis, and interpretation of RiboMethSeq, BisulfiteSeq, AlkAnilineSeq (and HydraPsiSeq) data. I have also designed most of the graphs and wrote the results part of this manuscript.

Abstract

Modification of tRNA is an integral part of the epitranscriptome with a particularly pronounced potential to generate diversity in RNA expression. Eukaryotic tRNA contains modifications in up to 20% of their nucleotides, but not all sites are always fully modified. Combinations and permutations of partially modified sites in tRNAs can generate a plethora of tRNA isoforms, termed modivariants. Here, we investigate the stoichiometry of incompletely modified sites in tRNAs from human cell lines for their information content. Using a panel of RNA modification mapping methods, we assess the stoichiometry of sites that contain the modifications 5-methylcytidine (m^5C), 2'-O-ribose methylation (Nm), 3-methylcytidine (m^3C), 7-methylguanosine (m^7G), and Dihydrouridine (D). We discovered that up to 75% of sites can be incompletely modified and that the differential modification status of a cellular tRNA population holds information that allows to discriminate e.g. different cell lines. As a further aspect, we investigated potential causal connectivity between tRNA modification and its processing into tRNA fragments (tiRNAs and tRFs). Upon exposure of cultured living cells to cell-penetrating angiogenin, the modification patterns of the corresponding RNA populations was changed. Importantly, we also found that tsRNAs were significantly less modified than their parent tRNAs at numerous sites, suggesting that tsRNAs might derive chiefly from hypomodified tRNAs.

1. Introduction

Transfer RNA (tRNA) molecules are key players in implementation of the genetic code in the living cell. These relatively short RNAs (75-90 nt in average) are responsible for mRNA decoding by specific codon-anticodon interactions and also provide activated amino acids for protein synthesis. tRNAs are well known to contain numerous modified nucleotides, generally clustered in structurally and functionally important tRNA regions (e.g. anticodon). Due to redundancy in the genetic code, several tRNAs (called isoacceptors) may correspond to the same amino acid, encoded by different codons, moreover, tRNAs decoding the same codon may have slightly different sequence (isodecoders).

The role of tRNA and tRNA modifications in neurological disorders such as Amyotrophic lateral sclerosis (ALS), frontotemporal dementia (FTD) and Parkinson's disease (PD), was driven to the spotlight with the demonstration of an impact of RNA modifications on tRNA processing into tRNA-derived fragments of different size, generally referred as tsRNAs, or, for miR-like fragments as tRFs[581]. Such processing is a complex process, involving a number of different nucleases, including angiogenin (ANG), with the latter cleaving tRNA in the anticodon loop and thus generating so-called tiRNAs[582, 583] while other processing steps that give rise to tiRNA-type fragments are ANG-independent[584]. For the remainder of this manuscript, we will use

the term tsRNA for simplicity. A central finding in this field was that the tRNase activity of ANG, otherwise known as an important stimulator of the growth of new blood vessels, was sensitive to tRNA modifications, in particular 5-methylcytidine (m^5C), generated by the RNA methyltransferase (MTase) activity of Dnmt2[585] and NSUN2[586]. tRNA cleavage in Ang-dependent and independent manners was reportedly influenced by the presence of 1-methyladenosine (m^1A)[584, 587], subsequent work identified ribose methylations (Nm) as protective against RNase activities[427, 588]. In contrast, some tRNA modifications in the anticodon loop, like Inosine and hypermodified U*34 may also promote accelerated tRNA cleavage by specific tRNA nucleases, like endonuclease V and zymocin, respectively[589–592]. While some of the chemical structures of the modifications offer plausible biochemical mechanisms of protection against RNA cleavage[407, 593], a systematic evaluation of protective m^5C and Nm residues is missing, as is an exploration of the effect of other modifications.

tRNAs represent 8-10% of the cellular total RNA population, making those molecules in general, as well as each individual tRNA isoacceptor, highly abundant RNA species. In general, the amount of tRNAs being processed into tsRNAs is way too low to significantly change the abundance of the cellular tRNA pool, and therefore unlikely to produce a pronounced phenotype rooted essentially in tRNA abundance. In contrast, the tsRNAs resulting from tRNA processing make up a significant fraction of the pool of small noncoding RNAs in the size range of 15-40 nucleotides or so. They have been reported to exhibit numerous downstream activities, including impact on protein biosynthesis[594]. An important finding relating this context to neurological disorders was the identification of ANG loss-of-function mutations in ALS[595]. Subsequently, ANG was shown to display neuroprotective activities[596], raising the possibility that its RNase activity might also have beneficial properties, and that RNA modification enzymes, in turn, might counter such beneficial effects. Importantly, ANG can induce tRNA cleavage upon its penetration into target cells[597], allowing a re-enactment of this scenario in tissue culture via treatment with exogenous ANG added *in cellulo*.

An open question in neuroscience is hence, which tRNA modifications modulate tRNA processing into tsRNAs in general, and by ANG in particular[598–600]. We have recently shown, by quantitative LC-MS of tRNA isolates from different brain tissues, that the tRNA modification levels of cerebellum hippocampus and cortex are sufficiently distinct to discriminate these tissues[601]. This shows that the task of identifying ANG-modulating modifications requires analysis of defined tRNA populations of pure tissue or of isolated cell lines. A further consideration is that a mere quantification of the modification content of tRNA preparations from nerve cells, as e.g. by LC-MS, does not provide any sequence context or position of modifications. Based on the above findings concerning the protective effects of m^5C and Nm, knowledge not only of the amount, but of the specific positions of ANG-mediating modifications is highly desirable. A promising approach to this goal resides in a number of high throughput sequencing techniques, which have been recently developed to map positions of

certain modifications (reviewed in [31, 275]). These methods typically make use of specific chemical, biochemical, or biophysical properties of the corresponding modification type[519]. Among these are methods tailored to the mapping of m^5C[468] , Nm[66, 602], as well as for the simultaneous mapping of 7-methylguanosine (m^7G), 3-methylcytidine (m^3C), and dihydrouridine (D)[360], that were applied here.

For reasons laid out above, we chose to assess ANG-modulating modifications in two tumor cell lines derived from different cellular backgrounds, namely from glial and nerve cells. Modification profile differences between these two populations will also be discussed.

2. Material and methods

2.1 Cell culture and angiogenin treatment

Human glioblastoma MZ-294 cells (RRID: CVCL_M410) derived from a primary glioblastoma[603] were grown as a monolayer in DMEM (Lonza) supplemented with 10% heat-inactivated fetal bovine serum (FBS; Sigma), 2 mM L-Glutamine (Sigma), 100 U/ml Penicillin and 0.1 mg/ml Streptomycin (P/S: Sigma) and maintained in a humidified incubator at 37 °C and 5 % CO2.

The SH-SY5Y cell line (RRID: CVCL_0019), is a cloned subline of the human neuroblastoma cell line SK-N-SH derived from a metastatic bone tumour[604]. SH-SY5Y cells were obtained from the American Type Culture Collection (ATCC, CRL-2266) and grown as a monolayer in DMEM/F12 (Gibco) supplemented as above and maintained in a humidified incubator at 37 °C and 5 % CO2.

Both cell lines were confirmed to be mycoplasma-free.

Lyophilized human recombinant ANG (rhANG; R&D Systems, #265-AN/CF) was reconstituted at 10 µg/ml in sterile PBS containing 1% BSA and stored at −20°C until use. MZ-294 cells and SH-SY5Y cells were treated with either rhANG (500 ng/ml) or vehicle (PBS with 1% BSA) in serum-free Neurobasal plus medium (Gibco) for 3h in a humidified incubator at 37 °C and 5 % CO2.

2.2 Immunocytochemistry

To assess ANG cellular localisation, immunocytochemistry (ICC) was performed on MZ-294 cells and SH-SY5Y treated with rhANG or vehicle. For ICC, MZ-294 cells and SH-SY5Y cells were seeded on glass coverslips in standard 24-well plates at densities of $2.5x10^4$ cells/well and $1x10^5$ cells/well, respectively. 24h following seeding, the cells were treated with rhANG or vehicle and then subjected to standard ICC techniques. Briefly, cells were fixed in 4% paraformaldehyde for 15 min at RT, permeabilized in ice-cold 95% ethanol with 5% glacial acetic acid for 10 min at RT and blocked in 5% BSA/PBS with 0.1% Triton X-100 for 1h at RT. Cells were incubated with the following

antibodies in blocking buffer at 4°C overnight: goat anti-human ANG (1:100, PC317L, Calbiochem, RRID: AB_213593) and mouse anti-tubulin (1:500, T6199, Sigma, RRID: AB_477583). Following primary antibody incubation, cells were incubated with the appropriate Alexa-dye conjugated secondary antibodies (1:500, Invitrogen) and Hoechst nuclear counterstain (1 µg/ml, Invitrogen) for 2h at RT. Coverslips were mounted on glass microscope slides (Thermo Scientific) using ProLong Gold Antifade Mountant with DAPI (Invitrogen). Images were acquired using a confocal Zeiss LSM 710 microscope with a 40X oil immersion objective (EC Plan-Neofluar 40x/1.30 Oil DIC M27, Zeiss) and processed using ImageJ (National Institutes of Health, USA).

2.3 Total RNA extraction and tRNA/tsRNAs purification

Total RNA was extracted and processed into tRNA/tsRNAs fractions as described in Richter et al.[601]. In brief, separation of tRNA and tsRNA fragments was achieved by 10% PAGE and bands, corresponding to tRNAs (60-75 nt) and tsRNAs (~15-50 nt) were excised and eluted using the ZR small-RNA™ PAGE Recovery Kit according to the manufacturers specifications.

2.4 Library preparation for tRNAs and tsRNAs fractions

Gel-extracted tRNA and tsRNAs fractions were quantified and ~150 ng of input RNA material was directly used for specific chemical treatment, allowing detection of Nm (RiboMethSeq), m5C (BisulfiteSeq) and m7G/m3C/D residues (AlkAnilineSeq), see below for description of those methods. In all cases chemical and enzymatic treatment used was aimed to create RNA fragments having 5'-P and 3'-OH extremities. Following steps in library preparation included the use of commercial NEB Small RNA kit for Illumine, comprising ligation of adapters at the 3'-end of RNA, annealing of the RT primer, ligation et the 5'-extremity and cDNA synthesis by RT. cDNA was amplified and barcoded for Illumina sequencing.

2.5 tRNA reference for alignment

Construction of reduced tRNA reference dataset for alignment was described previously[602]. In brief, predicted tRNA genes from gtRNAdb database (http://gtrnadb.ucsc.edu/) were clustered based on Euclidean distance between tRNA sequences. From this clustering, consensus sequences representative of each of the gene were compiled into a single reference sequence used for reads' alignment.

2.6 Sequencing detection methods

All tRNAs and tsRNAs samples have been sequenced with the HiSeq 1000 system from Illumina. Sequenced reads have been demultiplexed with bcl2fastq software, trimmed with Trimmomatic version 0.32, and aligned with bowtie2 version 2.2.2 on a

Linux Server Power Edge R815 from Dell. Their subsequent analyses from this point depends on the method used (RiboMethSeq, AlkAnilineSeq and Bisulfite Sequencing) on R software version “Bird Hippie” 4.1.2 with RStudio server integrated development environment (IDE) version “Juliet Rose” 1.4.1717. Data treatment and scores calculations have been done through R basic functions, heatmaps have been created from the “heatmap.2” function from the *gplot* package and the other graphs have been designed through the *ggplot2* package.

The demultiplexed unprocessed reads are available on ENA platform with accession number PRJEB57983

2.7 RiboMethSeq

RiboMethSeq protocol has already been described in detail previously[64, 66, 602]. RiboMethSeq is based on alkaline hydrolysis protection resulting from 2’-O-methylation. This modification located in the vicinity of the phosphodiester bond of RNA molecules prevents the deprotonation of the 2’-OH moiety which can interact with the neighboring phosphate group resulting in a 2’-3’-cyclic phosphate formation occurring a RNA cleavage. An alkaline treatment on RNA sequences of interest allows the creation of a cleavage profile created from 5’ and 3’ reads’ extremities counting through sequencing data manipulation in order to analyze such profile for cleavage protection detection. The detection and quantification of this alkaline protection is done through analysis of RiboMethSeq Scores, namely Score A (also known as Score A2), Score C (also known as Score C2 or MethScore), Score mean (or mean2) and Score Angle. Score C is used as a quantitative score for stoichiometry assessment. Here is the different formula for the RiboMethScores :

$$ScoreA = 1 - \frac{2n_i + 1}{\frac{1}{2}|\mu_l - \sigma_l| + n_i + \frac{1}{2}|\mu_r - \sigma_r| + 1}$$

with n_i, the number of 5’/3’-end reads counts at given position i.

μ_l and μ_r, the 5’/3’-end reads’ counts average from the preceding (left region) or subsequent (right region) neighboring nucleotides of position i, respectively. The range of these regions is 2 positions in this study.

The same method applies for σl and σr, which are respectively the 5’/3’-end reads’ counts standard deviation of the left and right region.

$$ScoreC = \frac{n_i}{\frac{1}{2}\left(\frac{\sum_{j=i-\delta}^{i-1}\omega_j n_j}{\sum_{j=i-\delta}^{i-1}\omega_j} + \frac{\sum_{j=i+1}^{i+\delta}\omega_j n_j}{\sum_{j=i+1}^{i-\delta}\omega_j}\right)}$$

with n_i and n_j, the number of 5'/3'-end reads counts at given position i and j, respectively

δ is the range of the neighboring positions taken in account (2 in this study).

ω_j is the incremented weight attributed to each neighboring positions, ranging from 1 when j = i +/- 1 (subsequent positions) to 0.5 when j = i +/- 6. Further is j from i, lower is the weight.

$$r_i = \frac{n_i}{n_{i-1}} \qquad f_i = \frac{n_i}{n_{i+1}}$$

$$ScoreMEAN = \frac{\left(1 - \frac{r_i}{\frac{\sum_{j=i-\delta}^{i-1} r_j}{\delta} + \frac{\sum_{j=i+1}^{i+\delta} r_j}{\delta}}\right) + \left(1 - \frac{f_i}{\frac{\sum_{j=i-\delta}^{i-1} f_j}{\delta} + \frac{\sum_{j=i+1}^{i+\delta} f_j}{\delta}}\right)}{2}$$

with n_i, the number of 5'/3'-end reads counts at given position I and δ the range of the neighboring positions taken in account (2 in this study).

2.8 AlkAnilineSeq

As described in previous publications[360, 605], the AlkAnilineSeq method for the detection of m^7G is based on RNA mild alkaline hydrolysisleading to a ring opening of m^7G and form an abasic site which can be treated by aniline to create a 5'-P suitable for detection by high-throughput sequencing. The method was found to also react to the presence of other modifications such as m^3C and D, presumably leading to an abasic site as a common denominator. One important aspect of this method is the selection of RNA abasic sites through the direct adapter ligation to the 5'-phosphate generated by aniline treatment. The main score used for m^7G detection is Normalized Cleavage score but this score is not suitable for tRNA analyses. Therefore, stop ratio, a ratio between the number of reads' 5'-extremities occurring at a position of interest (pinpointing direct 5'-adapter ligation due to the presence of an abasic site) and the total number of reads passing through the same position gives insight as a detection of quantification of m^7G sites.

$$\text{stop} = \frac{\text{Reads starts}}{\text{Total number of reads}}$$

2.9 Bisulfite Sequencing

Bisulfite sequencing procedure has been described in numerous papers in the literature[273, 468]. This method is based on the deamination of cytidine by the bisulfite ion (HSO_3^-) which results in the formation of a uridine. However, the methyl group of m^5C prevents this deamination to occur, allowing a distinction between unmodified cytidine converted to uridine and m^5C which remains unconverted cytidine. In order to pinpoint m^5C sites, a bisulfite non-conversion rate (NonConv) is calculated. It is a ratio which is calculated for each C residue and represents the proportion of reads harboring a C at this given position and the cumulated numbers of reads harboring either a C or a U. Since m^5C is not converted to uridine, the number of reads having C at those positions can be served as a detection tool but also as a quantitative score since this rate is directly proportional to m^5C presence.

$$\text{NonConv} = \frac{\text{Unconverted cytidines}}{\text{Cytidines converted to uracils} + \text{ Unconverted cytidines}}$$

2.10 Principal Component Analysis (PCA)

All PCAs analyses has been conducted through the function *prcomp()* from the R package *stats* using the correlation table of our data containing the median of 3 samples of quantitative score respective of each methods ; Bisulfite non-conversion rate for BisulfiteSeq, Score C for RiboMethSeq and stop ratio for AlkAnilineSeq. All m^7G sites due to their constant high stoichiometry have been excluded prior to analysis for noise reduction.

3.1 Results

Our previous demonstration that different brain tissues can be distinguished by their tRNA modification patterns[601] led us to investigate if such was also the case for human brain-derived cell lines. We therefore isolated tRNA and small RNA fragments from glioblastoma cell line MZ-294 (MZ) and neuroblastoma cell line SH-SY5Y (SH) (see Material and methods). Total RNA was extracted from cultured cells and size-fractionated by denaturing PAGE. Bands corresponding in size to tRNAs and tsRNAs, respectively, were excised and eluted using standard methods described previously[601].

3.2 Mapping of known and new putative m^5C, 2'O-methylated (Nm), m^3C, D and m^7G sites in human tRNAs

While our previous investigations assessed the overall modification content of tRNA fraction and small RNA fragments by LC-MS[601], we here chose a more sophisticated approach, namely pinpointing the location of tRNA modifications by RNA-seq based methods specifically tailored to mapping of (i) m^5C residues by bisulfite sequencing[468], (ii) ribose-methylated nucleotides Nm by RiboMethSeq[64], and (iii) alkali-sensitive modifications D, m^3C, and m^7G, by AlkAnilineSeq[360]. In all three methods, reads were aligned to a synthetic "tRNA genome", the tRNA reference sequences specifically designed for this type of analysis[606]. This reduced tRNA genome accounts for the occurrence of 39 cytoplasmic tRNA sequence families, each one consolidating sequences of high similarity, such as isoacceptors/isodecoders that differ only by a few SNPs. Selection of putative modified sites was based on threshold application on "detection scores", which were developed along with the methodology itself[356, 360, 607].

3.2.1 General considerations

For all three methods, we first performed an "unbiased" evaluation of our current data according to threshold values and considerations on noise, whose details are developed below. An overarching consideration of particular importance is rooted in the objective, which differs from many previous studies. Importantly, our objective was not to detect as many new modification sites as possible by setting low detection thresholds, but rather to assess variations of sub-stoichiometrically modified sites. Specifically, we wanted to detect variations of modifications between two human cell lines, which can only be achieved at tRNA sites that are incompletely modified, and therefore remain plausible substrates for modification enzymes. We posit that, for this to be relevant, a signal above noise level would have to be present in at least half of the tRNA samples investigated. Therefore, a criterion for our positive list of to-be investigated sites was occurrence of a signal in at least 50% of the biological samples. Furthermore, to be able to quantify the extent of substoichiometric modification, sites with low read coverage (below 30 reads per position) were not considered.

In a further selection step for appropriate modification sites, we applied plausibility considerations based on known RNA modification enzyme specificities, as well as known tRNA modification sites reported in the literature. This entailed removal of candidates at positions within the tRNA sequence and structure, which had not been reported as substrate of enzymes conferring the particular modification under scrutiny. For example, m^5C was never found at positions 11, 12, or 17 in tRNA, nor was any corresponding enzymatic activity in any organism reported in the literature. Residual non-deaminated C residues at those locations likely result from lower chemical reactivity to bisulfite due to local RNA structure, rather than from real presence of m^5C. Corresponding candidates were therefore deemed implausible and discarded.

Inversely, modifications established by published experimental evidence, which our unbiased assessment had failed to detect based on the scoring system, were added back in. One example detailed below concerns Gm18, a well-established modification in the D-loop of several human tRNAs, which fell through our scoring system (*vide infra*), potentially due to substoichiometric modification.

3.2.2 Bisulfite mapping of m^5C residues

RNA bisulfite sequencing is derived from well-established DNA bisulfite sequencing and is based on conversion of all C residues to U upon chemical deamination. A subset of modified C residues, e.g. m^5C, m^4C are resistant to deamination and can be detected by deep sequencing as non-converted C residues. In bisulfite sequencing, cytidines were considered modified when their signal, *i.e.* the non-conversion rate (NonConv) was at least 3 times the random non-deamination noise, a widespread custom in analytical chemistry[608, 609]. To assess this noise, NonConv distribution on reads aligned to rRNAs from the same samples was analyzed. The NonConv background from these *bona fide* unmodified cytidines (all rRNA C sites except 2 known m^5C positions in 28S rRNA) showed an average of 0.0252 and a standard deviation of 0.0677 resulting in a threshold value set to 0.228 (average + 3*standard deviation) (see Figure S1).

Hence the threshold is well in line with standard bisulfite sequencing analyses in the literature[470, 610]. In keeping with the above working hypothesis for analyzing fractional occupancy, *i.e.* substoichiometric modification of a given site in different biological contexts, we pursued our analysis only for those C sites which exceed the threshold NonConv > 0.228 in at least half of the analyzed tRNAs samples. Furthermore, the plausibility of an m^5C modification site was assessed based on known specificities of m^5C-MTases, To this end, each site was localized according to the tRNA numbering system[611] in order to homogenize site attribution. 4 sites, namely (Ser(GCT)11, Glu(CTC/TTC)12, Leu(TAA)12 and Lys(CTT)17) were excluded, since their localization did not correspond to any known m^5C modification site or tRNA:m^5C-MTase targeted sites[612]. The latter included positions 38 (modified by DNMT2); 48, 49 and 50 (NSUN2) and 72 (NSUN6). Figure 1 shows the final 47 curated sites (including 2 referenced ones which have not been seen through our detection parameters) in ascending order of tRNA numbering on the x-axis, and a plot of NonConv median values on the Y-axis, where the threshold of 0.228 is highlighted by a green horizontal line. The red dots represent the m^5C tRNA sites known from the literature “documented sites”[328, 613] while the blue ones are the positions identified as highly plausible additional m^5C sites (“retained candidates”).

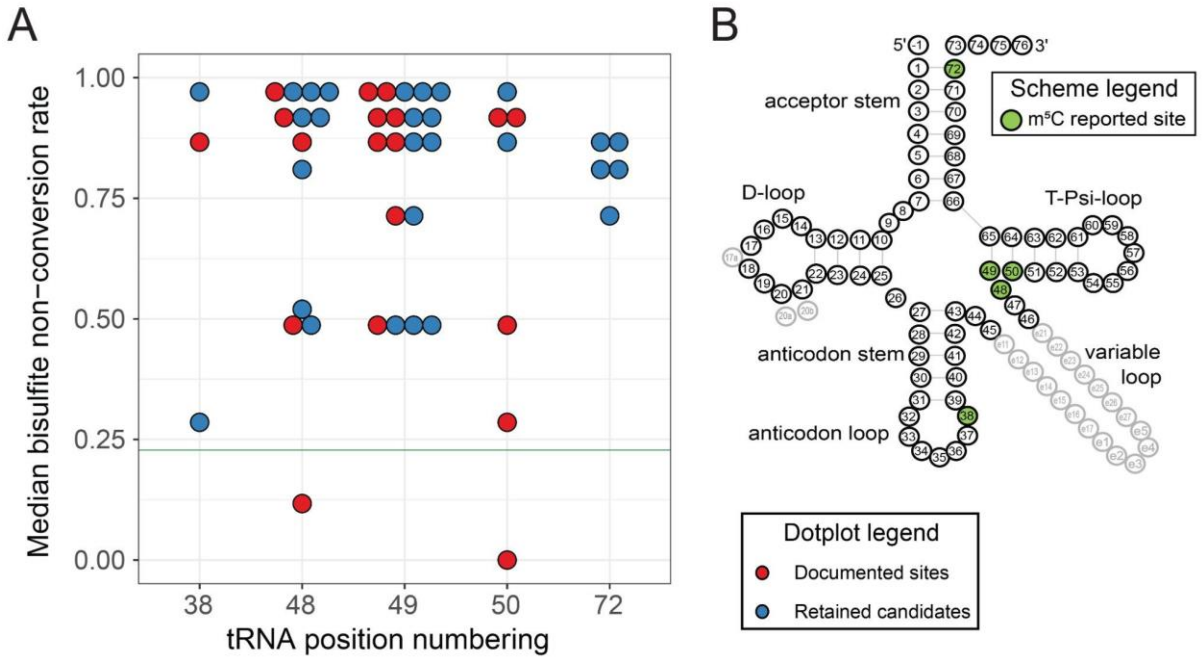


Figure 1: BisulfiteSeq mapping of m^{5}C sites in tRNAs. m^{5}C modification level was measured here as bisulfite NonConversion level C-to-U (NonConv) and plotted on the Y-axis for tRNA positions (x-axis) reported to be modified in at least one isoacceptor in any species. The threshold of average + 3*standard deviation = 0.228 is highlighted by a green horizontal line. Each dot represents a site in a different tRNA sequence. The red dots represent the m^{5}C sites known from the literature[328, 613], blue dots are plausible additional m^{5}C sites identified in other tRNAs according to their NonConv score. The cloverleaf to the right shows the canonical secondary tRNA structure along with tRNA-specific nucleotide position numbering, including extra (optional) residues 17a, 20a/20b in the D-loop as well as nucleotides e1-e27 in the variable tRNA loop/region. Positions relevant for m^{5}C are shown as green dots (pos 38, 48_49_50 and 72).

The above findings indicate that the identified modification sites are highly plausible, and that the threshold used here is meaningful for sensitive and specific m^{5}C detection by bisulfite sequencing, even at sub-stoichiometric modification level.

3.2.3 RiboMethSeq mapping of Nm residues

The RiboMethSeq protocol, developed for the detection of Nm residues in RNA assesses protection of RNA phosphodiester bonds against alkaline hydrolysis. In 2'-O-methylated residues, the 2'-OH nucleophile is capped by a methyl (-CH_3) group, thus abolishing cleavage of the 3'-adjacent phosphodiester bond almost completely. Therein RiboMethSeq analysis faces noise from unrelated, non-alkaline RNA hydrolysis on one hand, and spurious protection against hydrolysis resulting from

factors unrelated to modification, e.g., RNA secondary and tertiary structure. Other biases, namely at the adapter ligation steps are also known to affect the RiboMethSeq protection profile. Different score types developed for this protocol are applied to an ensemble of reads, without possibility to draw conclusion from a single-molecule digital conversion *versus* non-conversion event, as is the case for BisulfiteSeq. For several scores developed in previous publications, all signals were contained within three standard deviations of the noise, meaning that we needed to develop other threshold values. An analysis showed that the distribution of values Score A followed a Gaussian distribution, and the same applied to Score Angel (See Figure S2). Because Score A is rather specific, while Score Angle is sensitive, we decided to work with Nm candidates that were identified by both scores. In keeping with a conservative approach, for each of both scores, candidate sites were retained that scored above the mean noise plus 2 standard deviation, thus excluding 97.7% of noise and correspondingly accepting the remaining 2.3% as a baseline for false positives. The overlap of candidates from both scores was then scrutinized for biological plausibility according to known enzymatic activities that catalyzed the formation of Nm residues in tRNA. For reasons explained above, we added back-in candidates, such as Gm18 and U*m34 (U* stands for various wobble modifications), for which published experimental evidence was available, despite their scores being below noise + 2* standard deviation in one (or both) of the scores. An exception concerns Nm residues at position 4, whose placement near the 5'-end of tRNA prevents reliable quantification by RiboMethSeq. The results shown in Figure 2 feature the Score A and Score Angle on the Y-axis in panels A and B, respectively. The threshold values are indicated by a blue line, and dots are color-coded with respect to their validation from literature (in red) and meeting the threshold minimum. The latter are divided into two categories, namely plausible sites (in blue) and false positives, known from literature to be unmodified or carrying another known modification comprising no ribose methylation.

Note that, as a consequence of its specificity, a relatively large number of literature-known sites (red color) candidates are below threshold in Score A. These include positions 32 and 34, whose signals are frequently part of highly irregular RiboMethSeq protection profiles in the tRNA anticodon loop, in part due to hypermodified (and also RT-arresting) residues at position 37. Panels C and D show two additional scores, including score "mean", which is another sensitive detection score such as Score Angle. The bottom right panel shows Score C, which is linear with respect to modification stoichiometry, and therefore the score of choice for quantification in what follows (v*ide infra*). Further validation of Gm18 and Nm32/34 sites was obtained by independent analysis of datasets for respective human KO cell lines[65, 427]. In summary, 29 sites were retained from this analysis, which we quantitatively analyzed

to characterize differences between the two cell lines (*vide infra*).

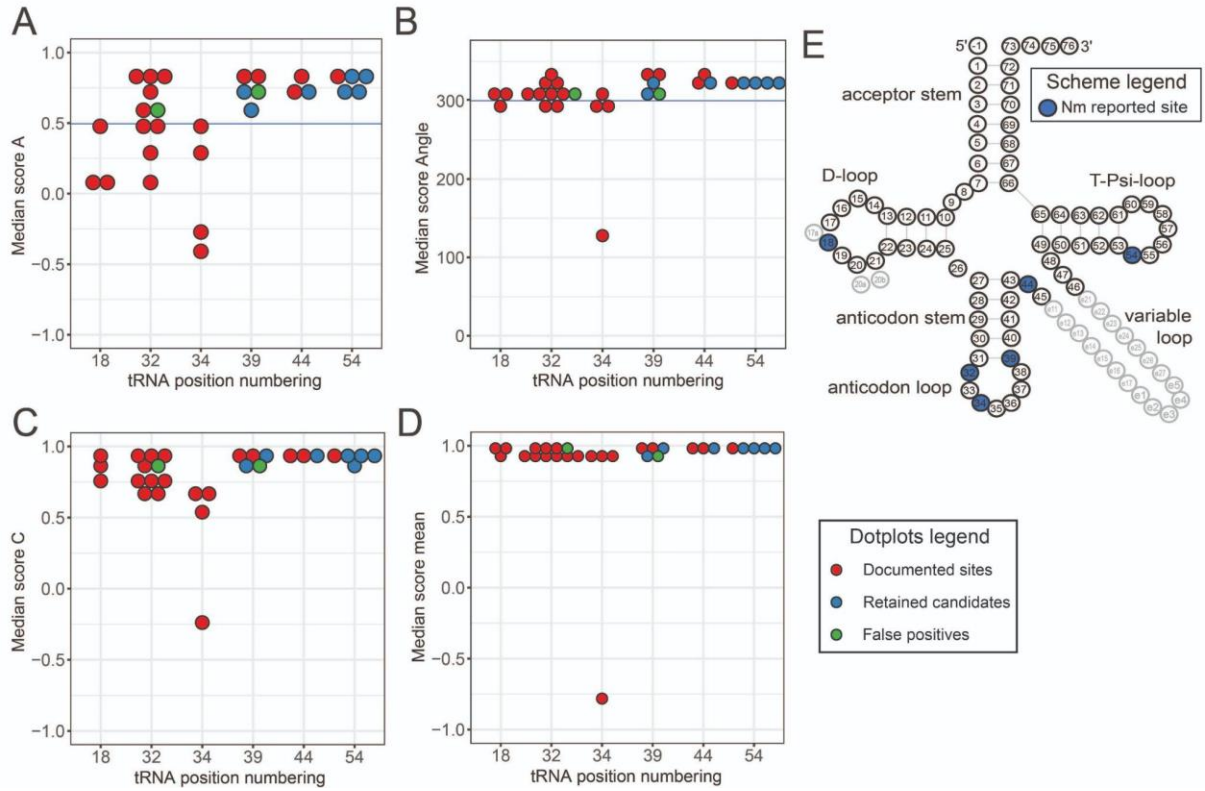


Figure 2. RiboMethSeq mapping of Nm sites in tRNAs. The upper panels A and B show results according to Score A and Score Angle, both of which were used for selection based on thresholds (blue lines). The lower panels show plots of Score mean and Score C, which were not applied for selection, but allow multiple types of assessment for any given residue. Red dots indicate known modification sites from the literature. Green dots indicate positions featuring high RiboMethSeq scores, despite being described as unmodified in the literature. These were therefore discarded. Blue dots indicate positions with scores above threshold values that represent plausible Nm modification sites. The cloverleaf to the right shows the canonical secondary tRNA structure along with tRNA-specific nucleotide position numbering, including extra (optional) residues 17a, 20a/20b in the D-loop as well as nucleotides e1-e27 in the variable tRNA loop/region. Positions relevant for Nm modification are shown as blue dots (pos 18, 32, 34, 39, 44 and 54).

3.2.4 AlkAnilineSeq mapping of m^7G, m^3C, and D residues

A subset of RNA modifications was shown to be sensitive to the combined action of heat and of alkaline pH, leading to scission of the N-glycosidic bond with subsequent formation of an RNA abasic site. Aniline treatment opens the RNA abasic site, inducing cleavage of the RNA phosphodiester bond, and thereby releasing 5'-phosphate-containing RNA fragments. Because the 5'-phosphates present a unique "molecular entry" to library preparation, the AlkAnilineSeq protocol features very low noise levels. For each modification and, thus, the parent nucleotide, the background is different, and signal strength differs between the different modifications.

The score applied for mapping of all nucleotide modifications was median stop ratio, with thresholds defined as three times the standard deviation (Figure S3) ranging from 0.263-0.418, as indicated by horizontal red lines in Figure 3. The color code of the candidates conforms to that used in previous Figures 1 and 2. Red dots label sites that were already reported, blue stand for sites that are plausible, but so far unknown, and green denote sites known to be unmodified, or containing other types of modification, respectively.

The strongest signals in AlkAnilineSeq originated from m^7G residues, whose signals were all close to the maximum value of 1 (Figure 3A) with a threshold of 0.418. All candidates derive from the only known m^7G modification site in tRNA at position 46. Signals derived from m^3C sites as shown in Figure 3B, were generally well separated from noise, with one exception. Said exception concerns position 32 in $tRNA^{SeC}$ (TCA) (Figure 3B), green dot just above 0.26 threshold line. According to the Modomics database, this C32 is not modified,and therefore was discarded. Known m^3C sites include positions 32 and position e2 in the variable loop in several tRNAs, as well as one reported m^3C at position 20[614]. Two AlkAnilineSeq signals at position 26 for $tRNA^{Trp}$(CCA) and $tRNA^{Tyr}$(GUA) could not be plausibly assigned. (supplement "Supp_Table_all_discarded_candidates.csv"). Both these tRNAs feature a guanosine at the relevant position 26, where the only known natural tRNA modification is $m^{22}G$, (rather than m^7G). The chemical properties of this doubly methylated nucleotide are not compatible with cleavage induced under conditions of the AlkAnilineSeq protocol, and other tRNAs with known $m^{22}G26$ do not show such a signal. Since, in addition, these sites are not a known target of m^7G-MTases, they were discarded. Since the AlkAnilineSeq protocol is intrinsically sensitive to RNA abasic sites, these unexpected G signals may be related to spontaneous loss of purine residues in RNA .

For D residues, which are uridine derivatives, the stop ratio noise average was 0.0329 with a standard deviation of 0.0964, setting the threshold for D sites at 0.322 (Figure 3C). At typical tRNA D-sites at positions 16, 17, 20 and 20a and 47, known to be targeted by Dus enzymes, the signal strength was rather variable. As shown by red dots in Figure 3C, many known sites of modification showed signals well below our detection threshold. This might plausibly be explained by substoichiometric

modification, which is a hallmark of dihydrouridine[615]. One 20a site from $tRNA^{Asn}$(GTT) was also discarded since this position was reported to be modified to acp^3U.

The adenosine nucleosides serve as a control, since there are no known chemical structures of adenosine derivatives that respond to the chemistry of the AlkAnilineSeq protocol. With a noise average of 0.0473 and a standard deviation of 0.117, the stop ratio threshold for A sites is 0.398, which is rather high in comparison to the other nucleosides. Interestingly, there are a number of signals observed above threshold (Figure 3D). However, all but one occur at, or next to, known bulky anticodon modifications, which are thought to interfere with an adapter ligation step in the library preparation protocol.

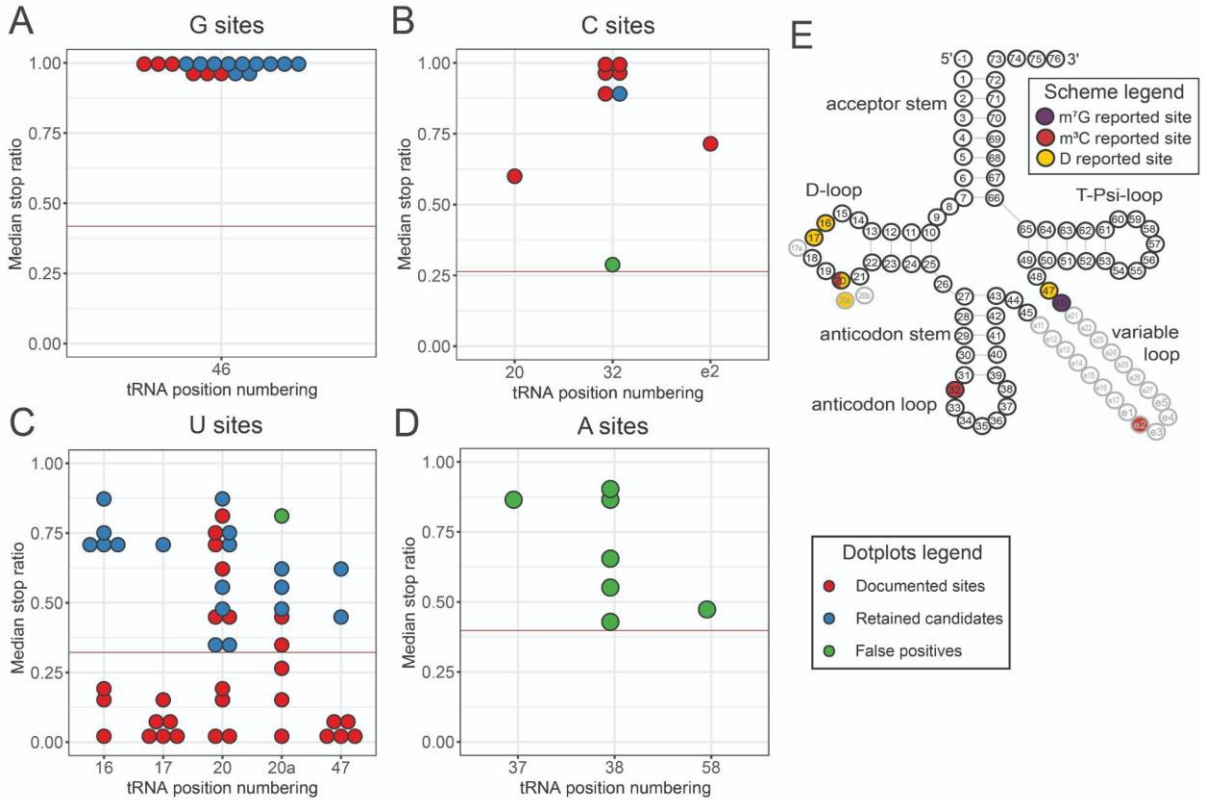


Figure 3. AlkAnilineSeq mapping of m^7G, m^3C, and D sites in tRNAs . Panels A-D show plots of stop ratio scores with corresponding thresholds indicated by horizontal pink lines. Red and green dots indicate, known modified and unmodified sites, respectively, from the literature. Known unmodified sites were therefore discarded. Blue dots indicate sites having scores above threshold values that represent plausible m7G/m3C/D modification sites. The cloverleaf in panel E shows the canonical secondary tRNA structure along with tRNA-specific nucleotide position numbering, including extra (optional) residues 17a, 20a/20b in the D-loop as well as nucleotides e1-e27 in the variable tRNA loop/region. Positions relevant to m7G-modification violet dots (pos 46), m3C sites are in red (pos 20, 32 and e2), D sites are in yellow (pos 16, 17, 20, 20a and 47).

3.2.5 A selection of tRNA modification sites for assessment of stimulus-dependent modification dynamics.

In our objective to find tRNA modification sites suitable to gauge the impact of physiological conditions such as cell type and exposure to angiogenin, the above workflow combined with data from literature identified 149 positions in human tRNAs as being robustly modified. While there are certainly additional sites to be found, those summarized in Figure 4 and Table S1 are amenable to our analytics. Sites include 47 m^5C, 29 Nm, 47 D, 9 m^3C, and 17 m^7G modifications (summarized in Table S1 and Figure S4). As shown in Figure 4A-C, the dynamic range of the different scoring systems is relatively heterogeneous. Still, summarized over the different scoring in Figures 1-4, a total of 35 of these modification sites apparently feature complete modification stoichiometry across tRNAs and tsRNAs (> 80% estimated modification level) with 10 m^5C, 7 Nm, 1 m^3C and all 17 m^7G sites. All 149 sites might change their modification status, but especially the substoichiometrically modified sites are prime candidates for comparative analysis of tRNA modification profiles, most liable to change said stoichiometry under alternate cellular physiology.

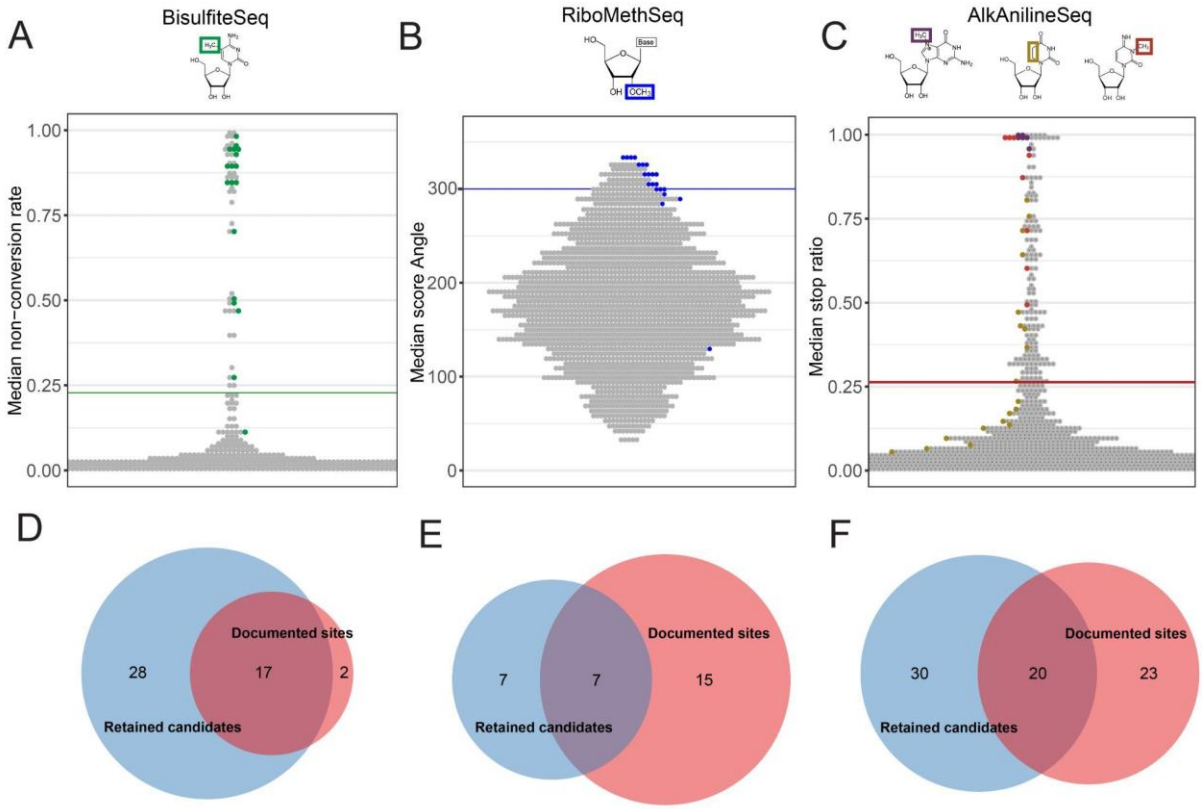


Figure 4: Experimentally mapped sites for m^5C, 2'-O-methylation and $m^7G/m^3C/D$ compared to data from the literature. (**A-C**) Relevance of applied threshold for detection of putatively modified sites (each dot represents one site) compared to known modified sites represented by green dots for BisulfiteSeq (**A**), blue dots for

RiboMethSeq (**B**) and purple (m^7G), gold (D) and orange (m^3C) dots for AlkAnilineSeq (**C**). The respective thresholds are shown by colored horizontal lines: 0.228 for bisulfite non-conversion rate, 300 for the score Angle and 0.263 for stop ratio. Values correspond to score medians from 12 tRNA samples, 3 untreated and 3 treated with angiogenin for both cell lines. (**D - F**) Overlapping and exclusivity between tRNA sites known from the literature ("documented sites", red circle) and tRNA sites detected through our workflow ("retained candidates", blue circle). Panels represent data from BisulfiteSeq (**D**), RiboMethSeq (**E**) and AlkAnilineSeq (**F**), respectively.

3.3 Modification patterns of tRNA versus tsRNAs show significant differences

From the above list of modifications deemed amenable to robust quantification, an assessment of their modification stoichiometry was done in biological triplicate. This analysis was extended from full length tRNA to tRNA fragments. The results are shown as boxplots in Figure 5, with separate plots for each modification. Plots in the first column (panels A,C, E,G) show tRNA modification profiles of the MZ and SH cell lines, whereas plots in the second column (B,D,F,H, Note that an alternative arrangement of the panels is available in the supplement in Figure S5) show the corresponding results obtained for respective tsRNAs. Results for p-values obtained from a comprehensive comparison according to Mann–Whitney–Wilcoxon test[616] are given in Tables S2-S6 in the Supplement. Differences between tRNAs of the two cell lines used were observed for m^3C and m^7G levels. More significant differences (p-value < 0.01) between tRNAs and tsRNAs were found for m^5C and Nm, and significant differences (p-value threshold of 0.05) also exist for m^3C. Globally, the tsRNAs appear less modified compared to intact tRNAs.

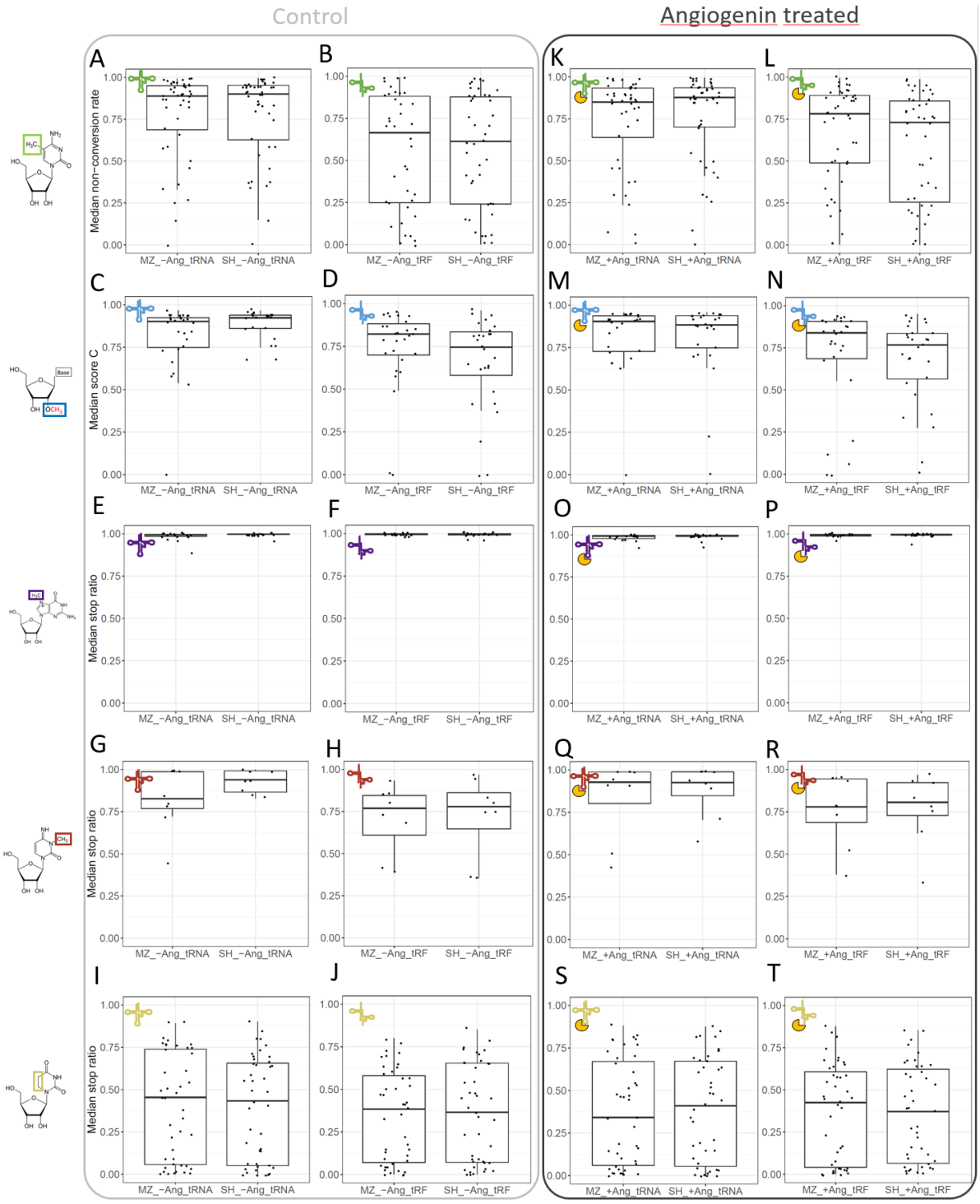


Figure 5: **Box plots showing modification levels for tRNAs and tsRNAs extracted from untreated and ANG-treated samples.** Panels (A-J) correspond to untreated samples, while panels (K-T) show the results for ANG-treated cells. Modification levels for m^5C (A, B, K, L), Nm (2'-O-Me) (C, D, M, N), m^7G (E, F, O, P), m^3C (G, H, Q, R) and dihydrouridine D (I, J, S, T) are shown. The values shown here are modification level medians of the respective detection scores obtained for three biological replicates.

3.4 Treatment with angiogenin *in cellulo* induces moderate changes in the modification patterns of tRNAs and tsRNAs

Because ANG has emerged as an important player in the processing of tRNAs into tsRNAs on one hand, and in CNS related diseases on the other hand, we decided to test its impact on tRNA/tsRNA modification patterns *in cellulo*. This was made possible by ANG's capacity to penetrate cells, documented by several groups[597, 617, 618]. Following a previously published regimen[619], both cell lines were treated with ANG, and ANG uptake was verified by fluorescence microscopy as detailed in the Materials and Methods section. Figure 6 shows the intracellular distribution of ANG (green stain, left column) being mainly cytosolic at the time of analysis, while entering the nucleus at a later time point[619]. Activity of Angiogenin has been verified by evaluation of RNA fragmentation in MZ and SH cells (Figure S6).

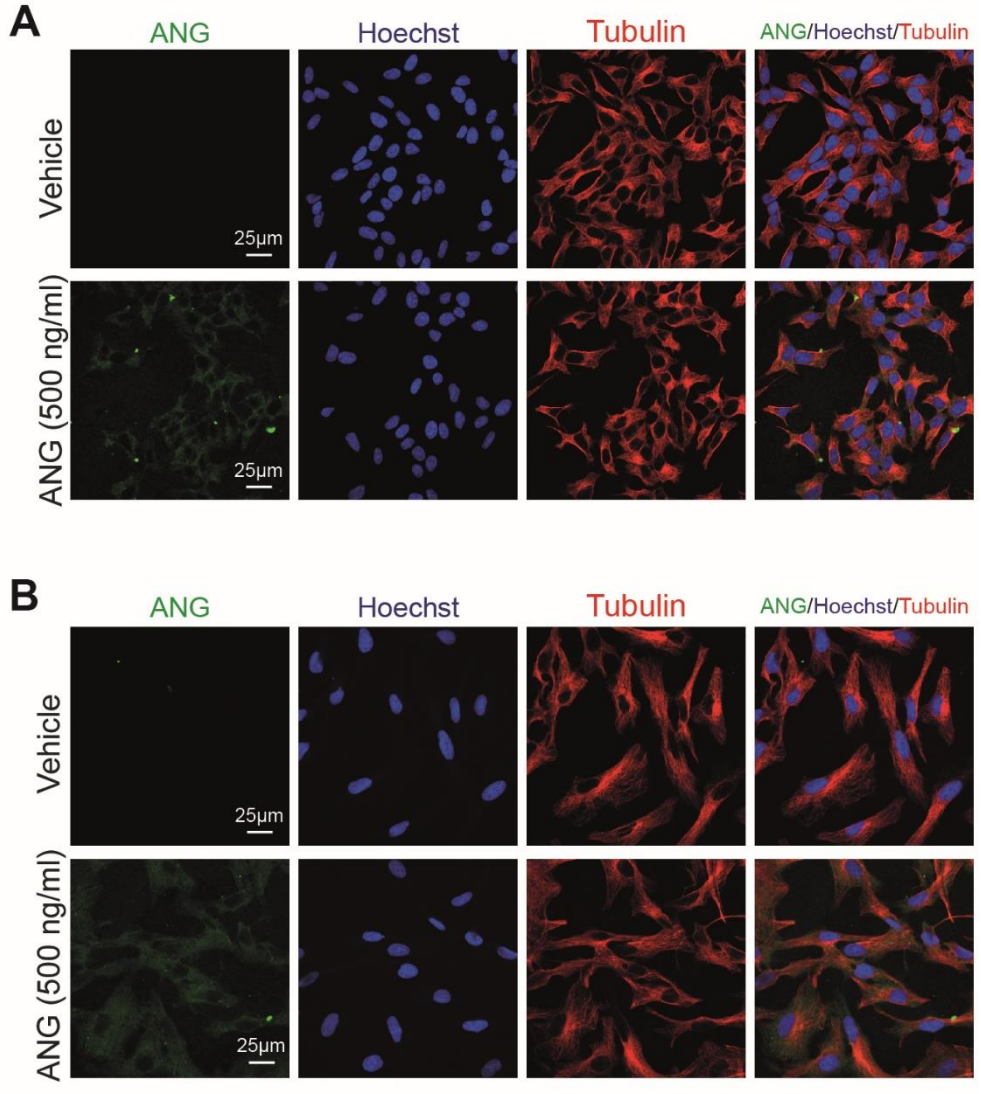


Figure 6: Uptake and subcellular localization of ANG in MZ and SH cell lines. (A-B) Confocal images of the intracellular distribution of ANG in **(A)** SH-SY5Y cells and **(B)** MZ-294 cells following 3h treatment with recombinant human ANG (500 ng/ml) or

vehicle. The cytoplasm is stained with tubulin (red) and nucleus is stained with Hoechst (blue). Scale bars, 25 µm. Images are representative of n=3 experiments.

After treatment, cells were harvested, and RNAs were isolated and analyzed as described above. Panels K-R in Figure 5 show the corresponding data in boxplots, with tRNA data in the third column (panels K,M,O,Q,S), and tsRNAs data in the fourth column (panels L,N,P,R,T. Note that an alternative arrangement of the panels is available in the supplement : Figure S5). Overall, when perusing the ensemble of tRNA modification sites, angiogenin did have only mild effects, which can be appreciated by comparing the boxplots in Figure 5 in a given left panel with the associated plot in the right panel. While shifts of the median value and changes in standard deviations occur in several instances, none of the angiogenin-induced changes are significant with respect to the ensemble of sites. For example, m^5C value plots tend to display lower dispersion after ANG treatment (compare Figure 5A with 5K, and 5B with 5L). No general direction could be identified by boxplots for Nm, m^7G, m^3C and D (Figure 5C-J, and Figure 5M-5T). As above, the p-values in tables S1-S5 indicate significant differences between modification levels of tRNAs and tsRNAs as a general rule, but no other significant differences are in evidence. Since the RNase activity of ANG is known to be impeded by selected tRNA modifications, this first result on the impact of ANG treatment is not particularly surprising. It also warrants a more in-depth scrutiny on the bases of individual tRNA sequences and modification sites.

3.5 tRF hypomethylation is not global, but site-specific

Following up on the global conclusions drawn from the data representation in Figure 5, namely significant differences between tRNAs and tsRNAs, and moderate impact of ANG treatment, data were scrutinized at the level of individual tRNA species. The corresponding heatmaps are displayed in Figure 7. Interestingly, clustering of the m^5C sites (Figure 7A) readily yields three categories that characteristically differ with regard to modification status and its variation. These comprise (i) sites with high methylation stoichiometry and low variation, (ii) low methylation levels with low variation, and (iii) average methylation levels with substantial variation within the different cell lines, RNA species, and angiogenin treatment status. Category (i) has stable modification levels even after processing or degradation into tsRNAs, suggesting little effect of the modification on the processing event. Similarly, a lack of variation in stoichiometry as in category (ii) would also be taken as a lack of impact of the modification on tRF generation. Therefore, of highest interest among these are data from category (iii), because they comprise data from a dynamic range that responds to either native processing of tRNAs into tsRNAs or to exogenous angiogenin.

Accordingly, in Figure 7A, the heatmap for m^5C levels shows category (i) species on the top on mostly dark green (high methylation) and category (ii) on the bottom, featuring mostly light green/purple (low methylation). Sandwiched between these is category (iii) in the middle, its dynamic stoichiometry visualized by the presence of

distinct shades of green and purple in each row. For m^5C, the 10 members of category (iii) include tRNA isoacceptors for Ser (2 species), Lys (2), Thr (2) and one each for Leu, Ile, Phe and Tyr. Interestingly, the sites concerned are exclusively at positions 48 and 49, and these sites are typical substrates of NSUN2. A tentative interpretation, to be discussed later in more detail, posits that these tRNAs are moderately good substrates of NSUN2, i.e. better substrates than are tRNAs from category (ii), but weaker substrates than tRNAs from category (i), and that this property intervenes in regulation of their stability.

A similar division of Nm sites is illustrated in a section of the heatmap in Figure 7B, where category (iii) comprises a total of 7 sites in tRNAs which are coding for Ser (4 sites), Lys (2) and one for Trp. In contrast to m^5C, the distribution of sites is heterogeneous throughout the tRNA, including sites at positions 18, 39, 44 and 54, thus all sites depicted in Figure 3 except for those in the anticodon loop.

A division into three categories for the mapping from AlkAnilineSeq turned out to be meaningful only for D residues. The heatmap for D residues is shown in Figure 7C. Here, we have identified 3 tRNA sequences that show high modification stoichiometry in tRNAs and a clear drop in tsRNAs, namely isoacceptors coding for Lys, Arg, and Gly. The results for the only 8 m^3C sites are shown in Figure 7D. Given the low number, categorization was deemed not meaningful and sequences were assessed on an individual basis instead. Here a Thr isoacceptor showed a clear decrease of m^3C stoichiometry between tRNAs and tsRNAs, and a Ser isoacceptor showed high modification in tRNAs, and a response to ANG in the tsRNAs. Furthermore, essentially all m^7G residues were completely modified, rendering a comparative analysis useless in the objective of this work. The corresponding heatmap for m^7G is therefore shown in the supplement (Figure S7).

Altogether the above analyses identified 20 variable sites in category (iii), as listed in Table 1. Interestingly, in all these cases, the tRNAs were more highly modified than were the tsRNAs. This has induced us to re-examine the heatmaps for cases, in which the tsRNAs were more highly modified than the tRNAs. One could indeed make two weak cases along this line (namely m^5C and m^3C), but the tsRNAs modification stoichiometry is never consistently high. Indeed, the only cases where tsRNAs reach maximum modification stoichiometry are such, where the same is true for the corresponding full length tRNAs.

Among the above cited 20 sites that are in category (iii), Isoacceptors decoding Ser (6 sites) and Lys (5) make up more than 50% of the cases. Upon inclusion of Thr (2) and Gly (1) over ⅔ of the cases are covered. Values for category (iii) are given in Figure 8, which also allows a visual verification of the claims above.

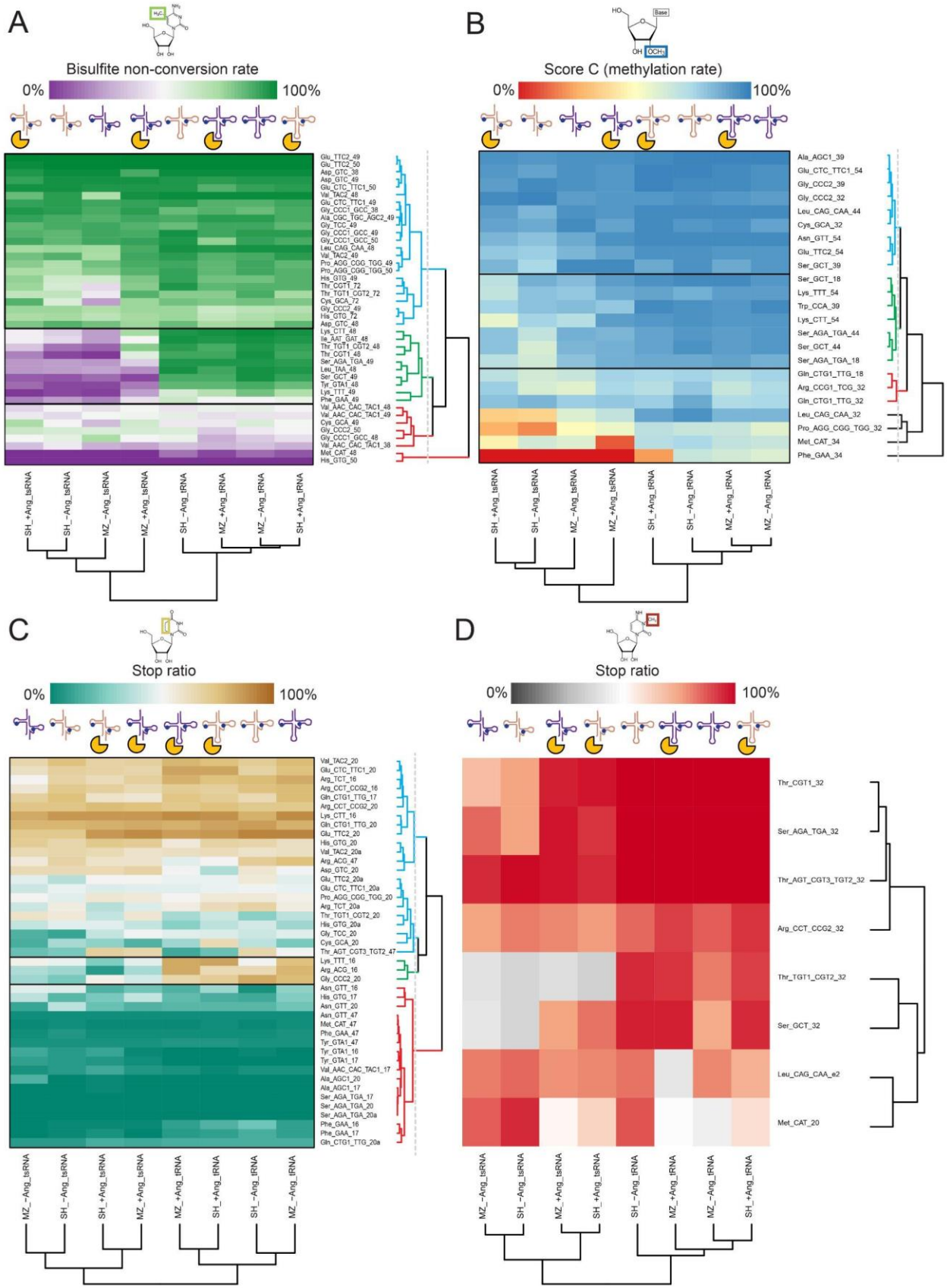

A
Bisulfite non-conversion rate
0%
100%
B
Score C (methylation rate)
0%
100%
C
Stop ratio
0%
100%
D
Stop ratio
0%
100%
SH_+Ang_tsRNA
SH_−Ang_tsRNA
MZ_−Ang_tsRNA
MZ_+Ang_tsRNA
SH_−Ang_tRNA
MZ_+Ang_tRNA
MZ_−Ang_tRNA
SH_+Ang_tRNA

Figure 7: Heatmaps for the assessment of modification changes in individual modified sites in tRNAs/tsRNAs. Each heatmap displays one specific modification's stoichiometry across the different samples (in X-axis) and the different sites retained for analysis (in Y-axis). Panel A displays m5C stoichiometry through bisulfite non-conversion rate (NonConv) from BisulfiteSeq, Panel B is for 2'-O-methylation through MethScore (also known as Score C) from RiboMethSeq and Panels C, D, and E are respectively for m7G, m3C and D though stop ratio from AlkAnilineSeq (see Material and Methods for further details). This stoichiometry is color-coded depending on the detection method used. Each sample and each site are hierarchically clustered based on the Euclidean distances between each subset, represented by dendrograms at the bottom for samples and on the right for sites. Dendrograms have been colored based on 3 main cluster properties: blue for sites with high stoichiometry and low variation, red for sites with low stoichiometry and low variation and green for sites with high stoichiometry variability. 4 sites remained unclassified for RiboMethSeq due to their unique behavior.

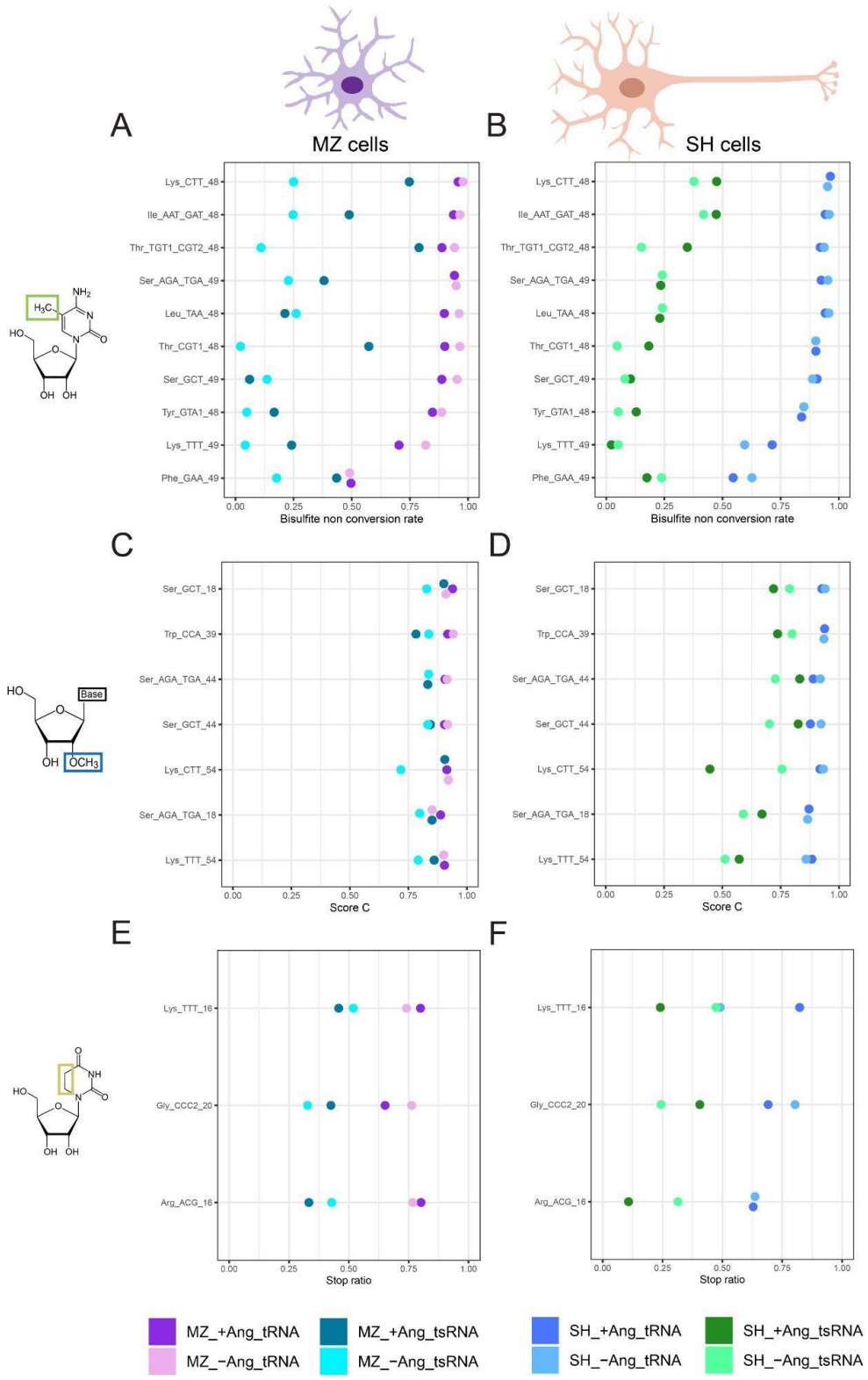


Figure 8: Dot plots of the various sites labeled as category (iii) from the heatmaps. Left panels (A, C, E) for MZ cells, right panels (B, D, F) for SH cells. The nature of modified nucleotide is shown on the left, and the identity of tRNA/tsRNA site on the left of each panel. Color code indicates the identity of the sample (tRNA/tsRNA) and treatment with ANG.

3.6 Identification of variable sites by principal component analysis

Following our rationale, that differences are to be found at sites of maximum variability, we set out to identify these using a principal component analysis (PCA). Figure 9 shows the positions of 8 datasets (MZ vs. SH, tRNA vs tsRNA, -ANG vs. ANG+) along four orthogonal vectors resulting from a PCA from all datapoints used in Figure 7, i.e. including all modification mapping methods. The PC1 vector, representing the majority of variability in the data (48.5%, Figure 9A) allows a very clear distinction between tRNAs (negative coordinate to the left) and tsRNAs (to the right), thereby confirming our previous conclusions from statistical analysis (Figures 5 & 7). We extracted the principal contributors to PC1 according to their correlation/anticorrelation coefficients with the 4 PCs (see table S7). Table 1 lists category (iii) sites, as obtained from clustering in heatmaps, along with their correlation coefficients. Clearly visible, all category (iii) sites also feature high absolute values for correlation factors for PC1. A number of these sites, especially Thr CGT1 m^5C48, as highlighted in Table 1, feature low correlation with the other 3 PCs, which are orthogonal to PC1, identifying them as potential markers for the distinction of tsRNAs from tRNAs. Taken together, this effectively validates our approach, to identify variable sites of biological significance by PC vector analysis.

Table 1. Category (iii) sites show high absolute values for correlation factors for PC1.

	PC1 loadings	PC2 loadings	PC3 loadings	PC4 loadings	Modification
Thr_CGT1_48	-0.99	-0.09	0.07	0.02	m^5C
Ser_AGA_TGA_49	-0.98	0.10	0.09	-0.02	m^5C
Lys_TTT_49	-0.98	0.14	0.05	0.06	m^5C
Tyr_GTA1_48	-0.97	0.09	0.12	0.04	m^5C
Lys_CTT_48	-0.97	-0.12	0.16	0.04	m^5C
Ile_AAT_GAT_48	-0.96	0.03	0.27	0.05	m^5C

Thr_TGT1_CGT2_48	-0.95	-0.23	0.07	0.00	m^5C
Leu_TAA_48	-0.94	0.17	0.11	-0.03	m^5C
Ser_GCT_49	-0.94	0.18	0.09	0.02	m^5C
Phe_GAA_49	-0.94	-0.08	0.11	-0.28	m^5C
Trp_CCA_39	-0.91	0.29	0.02	-0.13	Nm
Arg_ACG_16	-0.90	0.38	-0.09	0.00	D
Gly_CCC2_20	-0.89	-0.27	0.14	0.06	D
Lys_CTT_54	-0.86	0.15	-0.08	-0.31	Nm
Ser_AGA_TGA_18	-0.85	-0.15	-0.49	-0.05	Nm
Ser_AGA_TGA_44	-0.84	-0.24	-0.41	0.06	Nm
Ser_GCT_44	-0.81	-0.35	-0.42	0.09	Nm
Lys_TTT_54	-0.80	0.39	-0.14	-0.19	Nm
Ser_GCT_18	-0.79	0.32	-0.14	-0.38	Nm
Lys_TTT_16	-0.75	0.57	-0.17	0.12	D

From further inspection of the four PC vectors in Figure 9, PC3 separates the two cell lines under investigation: MZ to the left and SH to the right (Figure 9C), with the separation of tsRNAs being more pronounced. Principal contributors of PC3, as listed in Table S7 mark differences in tRNA modification stoichiometry characteristic between these cell lines. PC2 emerges as the vector harboring some information on

the effect of ANG treatment on tsRNAs (Figure 9B), which are well separated in the PC2 dimension, but this does not apply to tRNAs. PC4 again shows a uniform response to ANG treatment for each sample, in that the treatment leads to a positive shift along the PC4 vector for each tRNA MZ, tRNA SH, tRF MZ and tRF SH. Indeed, a combination of PC2 and PC4 is suited to somewhat separate samples according to their exposure to ANG (Figure S8), although there is no strong clustering clearly visible. The latter indicates that the response to ANG treatment is not very pronounced, and not characteristic.

Barplots with of modification stoichiometries of the main contributors (eigenvectors) strongly correlating (absolute correlation coefficient > 0.70) with principal components 2, 3 and 4, are shown in Figure 9 E,F, and G. Given that PC1 features 60 such contributors, the corresponding barplots were moved to the supplement (Figure S9). Also, category (iii) sites are a subset of the latter, and some values are provided in Figure 8. Using a correlation cutoff of 0.7 leads to a clean separation, meaning that no strong contributors to PC1 feature among the strong contributors to the other PCs, and vice versa. This orthogonality is visualized in the Venn diagram in Figure 9H, showing no overlap between what we herewith identify as marker sets/signatures for the three biological input parameters: RNA type (tRNA/tsRNA), exposure to angiogenin (-ANG/+ANG) and biological origin (MZ/SH cells).

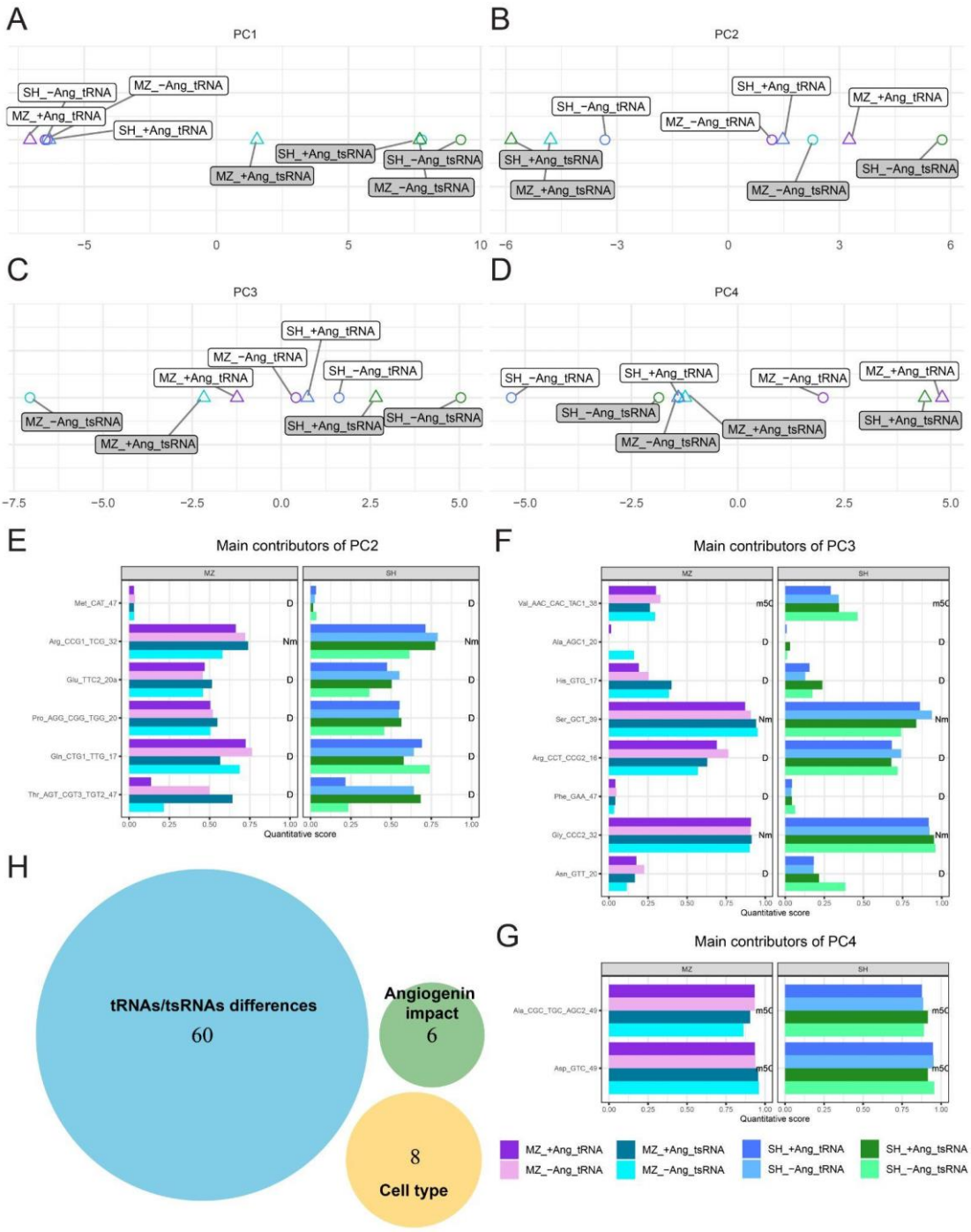


Figure 9. (**A-D**) Positions of datasets in 4 orthogonal vectors resulting from a PCA of all data points. Each principal component consists of linear combinations of modification data. The numbering PC1 through PC4 is according to the share of data that accounts for data variability: PC1 48.5%, PC2 15.4%, PC3 11.7% PC4 10.7%. (**E-G**) Barplots of the main contributors (eigenvectors) strongly correlating (absolute correlation coefficient > 0.70) with principal components 2, 3 and 4, respectively. (**H**) Venn diagram of the different contributors of the principal components and their interpretation. All PC1 main contributors have been labeled as “tRNAs/tsRNAs

differences", PC2 as "Angiogenin impact" and PC3 as "Cell type". Main contributors to PC4 show negligible data variability and their information content is therefore negligible for the case at hand.

4 Discussion

4.1 Towards a complete map of modifications in human tRNAs

There are several reasons why the topic of substoichiometric modification has come under more intensive investigation in recent years. Foremost among these, is probably the understanding that modification levels in tRNAs and other RNAs, including mRNAs, are subject to dynamic changes, associated with a new level of regulation of gene expression. The view of continuous, stimulus-dependent fine-tuning of RNA properties on the level of individual molecules replaced the picture of modification as an obligate maturation step to be invariantly undergone by every single RNA molecule of a given RNA species. Another strong contribution came from detection and sequencing technology, which enabled quantification of more than single modification sites, and which made investigations with large data sets feasible. While applications of mass spectrometry in RNA modification research have also surged, this applies even more to modification mapping techniques based on high throughput sequencing. Given the plethora of such mapping methods, we had to limit the number applied here for practical as well as financial reasons. The joint use of BS-Seq, RiboMethSeq and AlkAnilineSeq here was governed by the following rationale. Firstly, m^5C and Nm residues both occur abundantly and at numerous positions in tRNA, and moreover have both been reported to affect tRNA cleavage into tsRNAs[325, 604], hence it was decided to cover both of these by methods well established in our labs. Secondly, AlkAnilineSeq offered an attractive combination of features, such as low requirement for RNA material (~50 ng per sample), very low background, and the simultaneous mapping of m^3C, m^7G and D residues. The latter in particular are known to occur sub-stoichiometrically and are therefore prime candidates for yielding information to address our key question. Indeed, our data show a majority of D-sites to be substoichiometric as well as dynamic, and a number of them also featured among the top contributors of PC vectors.

Despite decades of tRNA studies, the complete human tRNA dataset was never entirely explored and only the global location of modified sites was firmly established (tRNAdb[328]/Modomics[30]/gtRNAdb[331]). Comprehensive analysis by the Suzuki lab established the full profile of human mitochondrial tRNAs modifications[620, 621], but almost ½ of cytoplasmic tRNA species have never been systematically studied.

4.2 Limitations of the current study and its methodology

Our combined mapping efforts identified a total of 149 sites in human tRNAs, with more than half of these not previously reported. This is not surprising, since less than

half of human cytosolic tRNAs have been exhaustively analyzed with respect to their modification content. Literature on similarly conservative approaches reports False Discovery Rates (FDR) in the single digit percentile and False Negative Rates (FNR) that are much higher. FDR examples are about 7% for Ribomethseq (RMS, [430, 602]) with FNR about 40%, FDR for Alkaniline seq better than 4% ([360], FNR ~53%), and FDR for bisulfite seq also about 7% ([315], FNR~4%). This data is in keeping with our unpublished datasets e.g. on mammalian rRNA. In comparison, when regarding the flowchart in Figure S4, we find that plausibility considerations applied to the present dataset suggest FDR rates of 13% (RMS), 4% (AAS) and 8% (BS), while FNRs are again higher at 54%, 53% and 10%, respectively. This data argues for 149 being a number that is likely still incomplete due to our deliberately high thresholds. Yet at present it is still an expansion of the previous census, and the most complete compilation of m^5C, Nm, m^3C, m^7G and D residues in this tRNA population.

4.3 Information content and biological idiosyncrasy of tRNA modification patterns

The principle research question driving our investigations was whether tRNA modification patterns contain information reflecting the physiological state of a cell, to a degree that different cell types could be distinguished from one another. Given that even differentiated cells have essentially the same set of tRNA sequences and genes for modification enzymes, such differences would likely be reflected, not in the presence or absence of a specific (marker) modification, but in the degree to which a certain position within the tRNA was modified. Information on the identity of a cell would therefore have to be extracted from the entirety of substoichiometric modifications present in tRNAs and their processing products, i.e. tRNA fragments of a cell. Inspired by the finding that tRNA modification patterns could be discriminated in different mouse brain tissues[601], we reasoned that this might be possible for different neuronal cell lines as well. We do indeed see distinct modification signatures that allow a separation of the two cell lines investigated in a PCA analysis. Interestingly, that separation was less pronounced than the difference between tRNAs and tsRNAs. To gauge the degree to which our data allowed to distinguish the influence of the principle parameters RNA species (tRNA vs. tRF, cell type, ANG treatment), we conducted a PCA. Roughly 50% of data variability was represented in PC1, whose faithful separation of tRNAs from tsRNAs could be verified by inspection of clustering in heatmap data, and visual inspection of the most variable sites, i.e. those belonging to category (iii). Distinction of ANG treatment resides in PC2 (vide infra) and of cell types in PC3, both containing a much smaller fraction of data variability. Finally, PC4 did not provide much additional distinction, and a look at its principal contributor in Figure 9G reveals that their modification stoichiometry barely shows variation. From this we conclude that parameters contained in PC4 are too close to background noise to be meaningfully interpreted. We will thus focus on PC1, PC2 and PC3 here.

4.4 tRNA fragments are hypomodified relative to tRNAs

tRNA fragments are generated from full-size mature tRNA molecules by cleavage in the loop regions, the most frequently observed are 5'-tiRNAs and 3'-tiRNAs, representing tRNA halves, but other even shorter tRNA derived species have been reported (referred to as tRFs)[588, 622, 623]. These processing pathways are mediated by different endonucleases with broad specificity. ANG cleavage is mainly discussed here, but other concomitant pathways have been described. For instance, Dicer cleaves tRNAs into miR-like fragments (tRFs)[624, 625] and another Dicer- and ANG-independent pathways have been reported[582, 587, 626]. These multiple and overlapping tRNA cleavage pathways create an impressive number of tRNA-derived small RNAs, thus appealing for more detailed characterization.

Moreover, a number of modifications enzymes and their corresponding modifications have been reported to affect cleavage of tRNAs into tsRNAs. We[565, 585, 586] and others[325] had previously contributed to work on the effect of m^5C modifications on tRNA processing into tsRNAs by ANG. Comparable protective effects were also reported of Nm modifications in the anticodon loop[604]. Similarly, m^1A modification and de-modification[627] were reportedly involved.

A question to which this study contributed significantly is how strongly modification profiles of mature tRNAs and their tsRNA processing products differ. The clearest statement to be drawn from all our statistical analyses is that there is a very substantial difference in the stoichiometry of modification between tRNAs and tsRNAs. Visual inspection of category (iii) sites in Figure 8 consistently finds tsRNAs hypomodified relative to the corresponding tRNAs. This effect is therefore dominantly reflected in PC1, and accounts for most of the variability in the dataset, which encompasses the vast majority of category (iii). The correlation of differences between tRNA and tsRNAs to PC1 makes this effect the strongest of the observed ones, while smaller shares of data variability account for PC2, PC3 and PC4.

An important aspect of data interpretation is orthogonality between the four PCs, and the correlation coefficients of a modification site's Eigenvector with any of them. An instructive example is found at the top of Table 1: The m^5C48 site in tRNAThr_CGT1 shows near perfect anti-correlation with PC1 (-0.99), and low correlation to all its orthogonal vectors. Its near complete alignment with the PC1 vector means that changes of m^5C48 occur principally between tRNAs and tsRNAs. This interpretation is readily underscored by Figure 8, which shows that tRNAs consistently show high amounts of m^5C48, whereas tsRNAs are consistently low. This confers to this modification a diagnostic value for the discrimination of tRNAs vs. tsRNAs.

Of note, the measured levels of modifications and tsRNAs are observed in living cell, where they are part of a flow that is not only governed by tsRNA generation through tRNA cleavage, but also by depletion of the tsRNA pool during further processing or degradation, making predictions or conclusions as to the effect of modifications difficult. Thus, our finding of tsRNAs being hypomodified relative to the mature tRNAs is not trivial or even easily anticipated, since the tsRNA likely have a short lifetime, possibly shorter than even hypomodified tRNAs.

4.5 Processing of modified tRNA into fragments by ANG

In the context of neuronal cells and tRNAs, their processing by ANG into tsRNAs is a matter of high interest given the involvement of ANG in various neuronal diseases[588, 628]. Most studies that originally identified connections between tRNA modification and ANG cleavage were conducted in an all-or-nothing setting, i.e. in knockout model organisms, where more subtle effects of ANG would go unnoticed[565, 585, 586]. Other studies aimed at defining the cleavage preferences of ANG in complex tRNA mixtures were performed *in vitro*[588, 629], therefore not accounting for the steady state arising from *de novo* tRNA synthesis, its modification, and its degradation and processing into tsRNAs. In the present setting, we saw the opportunity to implement exactly this, by applying exogenous ANG *in cellulo*, followed by analysis of tRNA modification stoichiometry.

Visual inspection revealed many of our 149 modification sites as dynamically affected by ANG treatment. For example, data in Figure 8A confirm that m^5C levels vary upon ANG treatment in a number of sites in category (iii), and the same applies to Nm in Figure 8B and D residues in Figure 8C. However, since many of these sites strongly correlate with PC1 and thus the tRNAs/tsRNAs differences, meaning that their modification stoichiometry is defined not only by ANG but also by other parameters, meaning that these sites hold limited information of diagnostic value.

Interestingly, a perusal of main contributors of PC2 implies that several sites of D modification are significantly affected by ANG treatment. These sites show limited to no difference between tRNA and tsRNAs, which is in line with PC2 being orthogonal to PC1. While variation of most sites is moderate, D47 in $tRNA^{Thr}$(HGT) shows a consistent response to ANG treatment, in that tRNA modification level decreases while tsRNAs show concomitant increase of D level in both cell lines. Accordingly, these characteristics are independent of the tRNAs/tsRNAs discrimination of PC1, and hold diagnostic value for identification of ANG treatment. Dihydrouridine has so far not been implied in tRNA cleavage, presumably because methods for high throughput mapping of D have only recently become available[630, 631].

While the case of D47 is to be verified by further experiments, it features highly interesting aspects of ANG mediated cleavage as follows. Pending experimental verification by orthogonal measurement, this might be a case of stimulation of ANG-mediated cleavage by modifications in the structural core of the tRNA, where other modifications have been shown to have protective effects against cleavage. This is plausible with regard to the known biophysical and biochemical effects of modifications on tRNA structure: m^5C and Nm are known to stabilize tRNA structure overall, as well as locally against cleavage. In contrast, D is thought to render tRNA structure more flexible, making it more prone to cleavage[407, 593]. Further study, beyond the scope of this paper, is required to tease out the mechanistic aspects of this intriguing observation.

5. Conflict of interest

Mark Helm is a consultant for Moderna Inc.

6. Acknowledgements

We thank Florian Richter for technical assistance.

JHMP was supported by a research grant from Science Foundation Ireland (SFI) under Grant Number 16/RC/3948 and co-funded under the European Regional Development Fund and by FutureNeuro industry partners and by the Joint Programme in Neurodegeneration Research (JPND) project RNA-NEURO (17/JPND/3455). EJ is supported by a StAR International PhD Scholarship from the Royal College of Surgeons in Ireland. MH was supported by the Joint Programme in Neurodegeneration Research (JPND) project RNA-NEURO (Bmbf FKZ: 01ED1804), and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 439669440 – TRR 319 (C01)“ and by Project HE 3397/21-1.

Table S1. List of 149 tRNA modification sites for assessment of stimulus-dependent modification dynamics. A list of all discarded candidates is available in a separated file called "Supp_Table_all_discarded_candidates.csv"

tRNA isoacceptor	Absolute position from 5'-end	Relative position according to tRNA structure	Modification	Sites attribution
Ala_AGC1	17	17	D	False negative
Ala_AGC1	20	20	D	False negative
Ala_AGC1	32	32	Nm	False negative (tRNAdb2009 + TARBP1/FTSJ3)
Ala_AGC1	39	39	Nm	True positive (tRNAdb2009)
Ala_AGC1	46	46	m7G	True positive
Ala_CGC_TGC _AGC2	45	46	m7G	Putative site
Ala_CGC_TGC _AGC2	48	49	m5C	Putative site
Arg_CCG1_TC G	32	32	Nm	False negative (TARBP1/FTSJ3)
Arg_ACG	16	16	D	Putative site
Arg_ACG	32	32	Nm	False negative (TARBP1/FTSJ3)
Arg_ACG	47	47	D	Putative site
Arg_CCT_CCG 2	16	16	D	Putative site
Arg_CCT_CCG 2	20	20	D	Putative site

Arg_CCT_CCG 2	32	32	m3C	True positive
Arg_TCT	16	16	D	Putative site
Arg_TCT	20	20a	D	Putative site
Arg_TCT	46	46	m7G	Putative site
Asn_GTT	16	16	D	False negative
Asn_GTT	20	20	D	False negative
Asn_GTT	47	46	m7G	True positive
Asn_GTT	48	47	D	False negative
Asn_GTT	55	54	Nm	Putative site
Asp_GTC	19	20	D	Putative site
Asp_GTC	38	38	m5C	Putative site
Asp_GTC	47	48	m5C	Putative site
Asp_GTC	48	49	m5C	Putative site
Cys_GCA	19	20	D	Putative site
Cys_GCA	31	32	Nm	False negative (TARBP1/FTSJ3)
Cys_GCA	45	46	m7G	Putative site
Cys_GCA	48	49	m5C	Putative site
Cys_GCA	71	72	m5C	Putative site
Gln_CTG2	48	49	m5C	Putative site

Gln_CTG1_TTG	16	17	D	Putative site
Gln_CTG1_TTG	17	18	Nm	False negative (tRNAdb2009 + TARBP1/FTSJ3)
Gln_CTG1_TTG	19	20	D	True positive
Gln_CTG1_TTG	20	20a	D	False negative
Gln_CTG1_TTG	32	32	Nm	False negative (tRNAdb2009 + TARBP1/FTSJ3)
Gln_CTG1_TTG	48	49	m5C	True positive
Gln_CTG1_TTG	49	50	m5C	True positive
Glu_CTC_TTC 1	19	20	D	True positive
Glu_CTC_TTC 1	20	20a	D	True positive
Glu_CTC_TTC 1	48	49	m5C	True positive
Glu_CTC_TTC 1	49	50	m5C	True positive
Glu_CTC_TTC 1	53	54	Nm	True positive (tRNAdb2009)
Glu_TTC2	19	20	D	Putative site
Glu_TTC2	20	20a	D	Putative site
Glu_TTC2	48	49	m5C	Putative site
Glu_TTC2	49	50	m5C	Putative site

Glu_TTC2	53	54	Nm	Putative site
Gly_TCC	19	20	D	True positive
Gly_TCC	48	49	m5C	Putative site
Gly_CCC2	19	20	D	True positive
Gly_CCC2	31	32	Nm	True positive (TARBP1/FTSJ3)
Gly_CCC2	38	39	Nm	True positive (tRNAdb2009)
Gly_CCC2	47	49	m5C	True positive
Gly_CCC2	48	50	m5C	True positive
Gly_CCC1_GC C	37	38	m5C	True positive
Gly_CCC1_GC C	46	48	m5C	True positive
Gly_CCC1_GC C	47	49	m5C	True positive
Gly_CCC1_GC C	48	50	m5C	True positive
His_GTG	16	17	D	False negative
His_GTG	19	20	D	True positive
His_GTG	20	20a	D	True positive
His_GTG	48	49	m5C	True positive
His_GTG	49	50	m5C	False negative
His_GTG	71	72	m5C	Putative site

Ile_AAT_GAT	47	46	m7G	Putative site
Ile_AAT_GAT	49	48	m5C	Putative site
Ile_TAT	47	46	m7G	Putative site
Leu_TAA	20	20	D	False negative
Leu_TAA	34	34	Nm	False negative (TARBP1/FTSJ3)
Leu_TAA	58	48	m5C	True positive
Leu_CAG_CAA	33	32	Nm	True positive (TARBP1/FTSJ3)
Leu_CAG_CAA	45	44	Nm	True positive (tRNAdb2009)
Leu_CAG_CAA	51	e2	m3C	True positive
Leu_CAG_CAA	58	48	m5C	Putative site
Lys_CTT	16	16	D	Putative site
Lys_CTT	46	46	m7G	Putative site
Lys_CTT	48	48	m5C	Putative site
Lys_CTT	54	54	Nm	Putative site
Lys_TTT	16	16	D	Putative site
Lys_TTT	46	46	m7G	Putative site
Lys_TTT	49	49	m5C	Putative site
Lys_TTT	54	54	Nm	Putative site
Met_CAT	20	20	m3C	True positive

Met_CAT	34	34	Nm	False negative (tRNAdb2009 + TARBP1/FTSJ3)
Met_CAT	46	46	m7G	True positive
Met_CAT	47	47	D	False negative
Met_CAT	48	48	m5C	False negative
Phe_GAA	16	16	D	False negative
Phe_GAA	17	17	D	False negative
Phe_GAA	32	32	Nm	False negative (tRNAdb2009 + TARBP1/FTSJ3)
Phe_GAA	34	34	Nm	False negative (tRNAdb2009 + TARBP1/FTSJ3)
Phe_GAA	46	46	m7G	True positive
Phe_GAA	47	47	D	False negative
Phe_GAA	49	49	m5C	True positive
Pro_AGG_CG G_TGG	19	20	D	Putative site
Pro_AGG_CG G_TGG	31	32	Nm	False negative (TARBP1/FTSJ3)
Pro_AGG_CG G_TGG	45	46	m7G	Putative site
Pro_AGG_CG G_TGG	48	49	m5C	Putative site
Pro_AGG_CG G_TGG	49	50	m5C	Putative site

Ser_AGA_TGA	16	17	D	False negative
Ser_AGA_TGA	17	18	Nm	False negative (tRNAdb2009 + TARBP1/FTSJ3)
Ser_AGA_TGA	19	20	D	False negative
Ser_AGA_TGA	20	20a	D	False negative
Ser_AGA_TGA	32	32	m3C	True positive
Ser_AGA_TGA	44	44	Nm	True positive (tRNAdb2009)
Ser_AGA_TGA	50	e2	m3C	False negative
Ser_AGA_TGA	57	49	m5C	True positive
Ser_GCT	17	18	Nm	False negative (TARBP1/FTSJ3)
Ser_GCT	32	32	m3C	True positive
Ser_GCT	39	39	Nm	Putative site
Ser_GCT	44	44	Nm	Putative site
Ser_GCT	57	49	m5C	Putative site
Thr_CGT1	31	32	m3C	True positive
Thr_CGT1	47	48	m5C	Putative site
Thr_CGT1	71	72	m5C	Putative site
Thr_TGT1_CGT2	19	20	D	Putative site
Thr_TGT1_CGT2	32	32	m3C	Putative site

Thr_TGT1_CGT2	46	46	m7G	Putative site
Thr_TGT1_CGT2	48	48	m5C	Putative site
Thr_TGT1_CGT2	72	72	m5C	Putative site
Thr_AGT_CGT3_TGT2	33	32	m3C	True positive
Thr_AGT_CGT3_TGT2	48	47	D	Putative site
Thr_AGT_CGT3_TGT2	49	48	m5C	Putative site
Thr_AGT_CGT3_TGT2	50	49	m5C	Putative site
Thr_AGT_CGT3_TGT2	73	72	m5C	Putative site
Trp_CCA	31	32	Nm	False negative (TARBP1/FTSJ3)
Trp_CCA	33	34	Nm	False negative (TARBP1/FTSJ3)
Trp_CCA	38	39	Nm	Putative site
Trp_CCA	45	46	m7G	Putative site
Tyr_GTA1	16	16	D	False negative
Tyr_GTA1	17	17	D	False negative
Tyr_GTA1	46	46	m7G	True positive
Tyr_GTA1	47	47	D	False negative
Tyr_GTA1	48	48	m5C	True positive

Val_AAC_CAC_TAC1	16	17	D	False negative
Val_AAC_CAC_TAC1	19	20	D	False negative
Val_AAC_CAC_TAC1	20	20a	D	False negative
Val_AAC_CAC_TAC1	38	38	m5C	Putative site
Val_AAC_CAC_TAC1	46	46	m7G	True positive
Val_AAC_CAC_TAC1	47	47	D	False negative
Val_AAC_CAC_TAC1	48	48	m5C	Putative site
Val_AAC_CAC_TAC1	49	49	m5C	Putative site
Val_TAC2	19	20	D	Putative site
Val_TAC2	20	20a	D	Putative site
Val_TAC2	46	46	m7G	Putative site
Val_TAC2	48	48	m5C	True positive
Val_TAC2	49	49	m5C	True positive

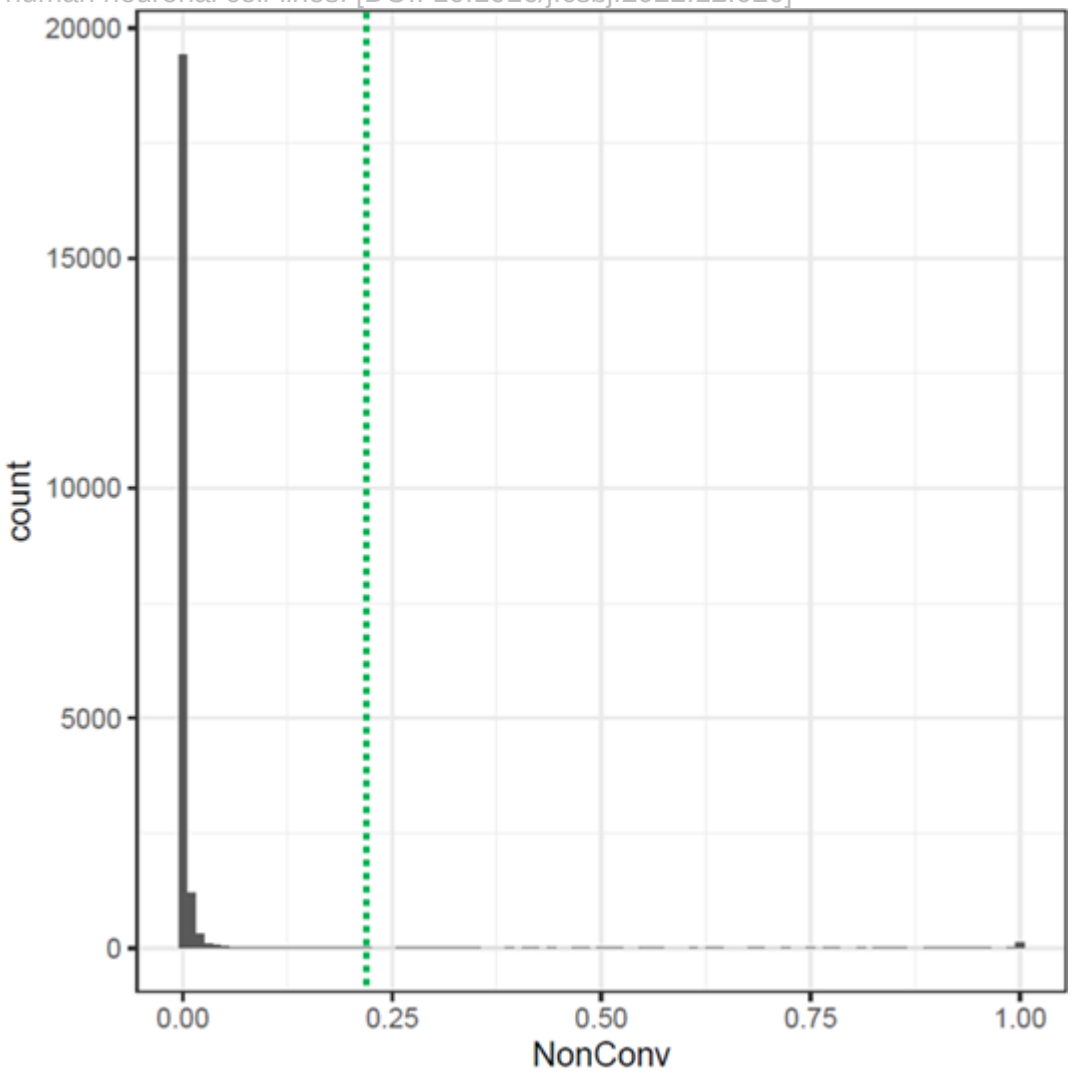


Figure S1 : Noise estimation of Bisulfite Non-conversion rate from Bisulfite sequencing. Each bar represents the counts of non-m5C sites (all rRNA C sites except 2 known positions in rRNA 28S) according to their bisulfite non conversion rate (NonConv). The distribution of these non-m5C sites, considered as noise, helped setting up a detection threshold at 0.228, corresponding to the average plus 3 times the standard deviation of the noise distribution, highlighted here by a green vertical bar.

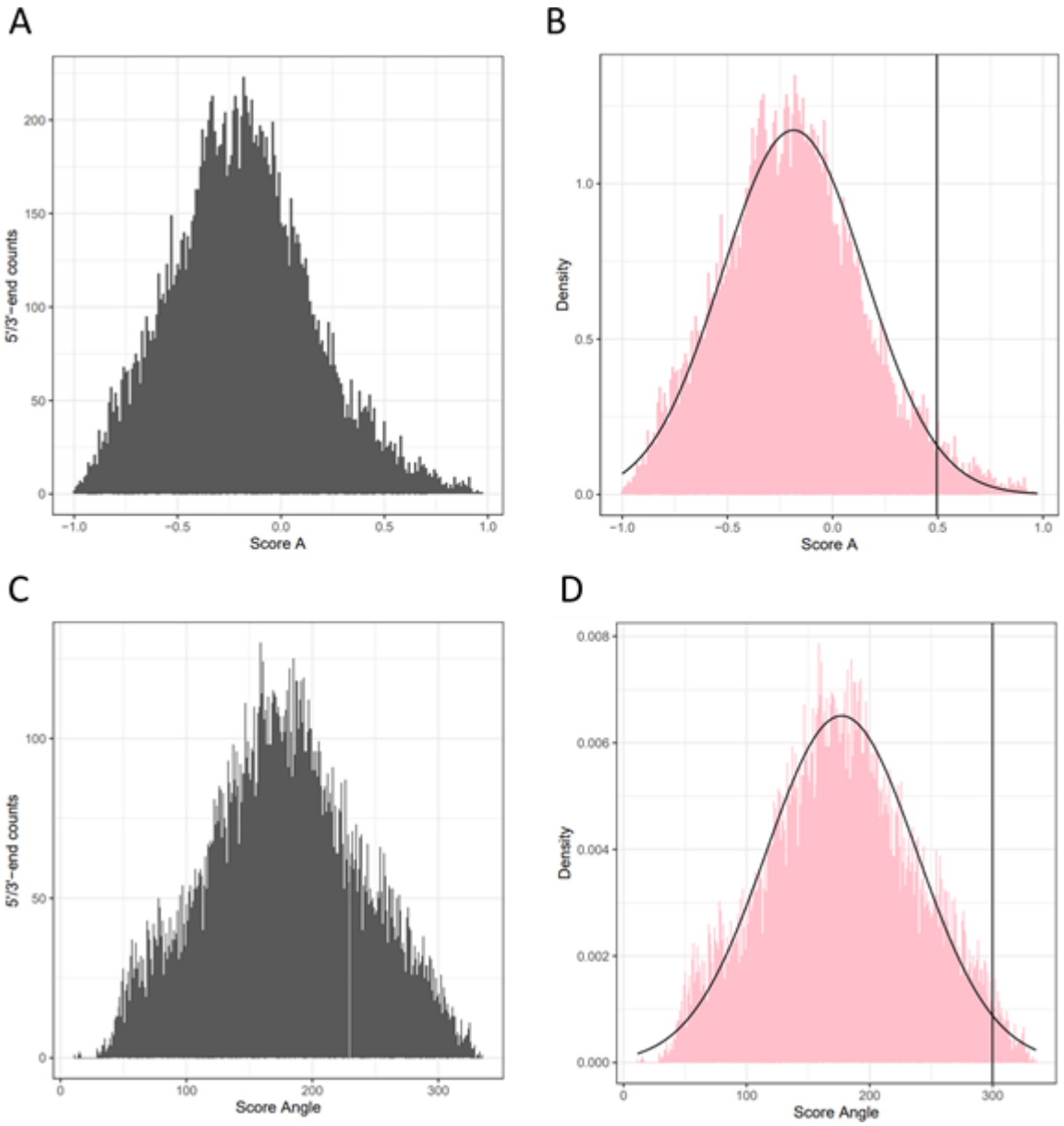


Figure S2 : Noise estimation of the different RiboMethScores from RiboMethSeq. Panels **(A, C)** displays non 2'-O-methylated sites (all tRNA sites not located at position 18, 32, 34, 39, 44 and 54) count distribution according to score A and score Angle, respectively. Panels **(B, D)** display fitting of noise distribution to a Gaussian distribution for score A and score Angle, respectively. The vertical bar indicates the threshold used for both scores which is at the noise average plus 2 times the noise standard deviation, which is 0.495 for score A and 300 for score Angle.

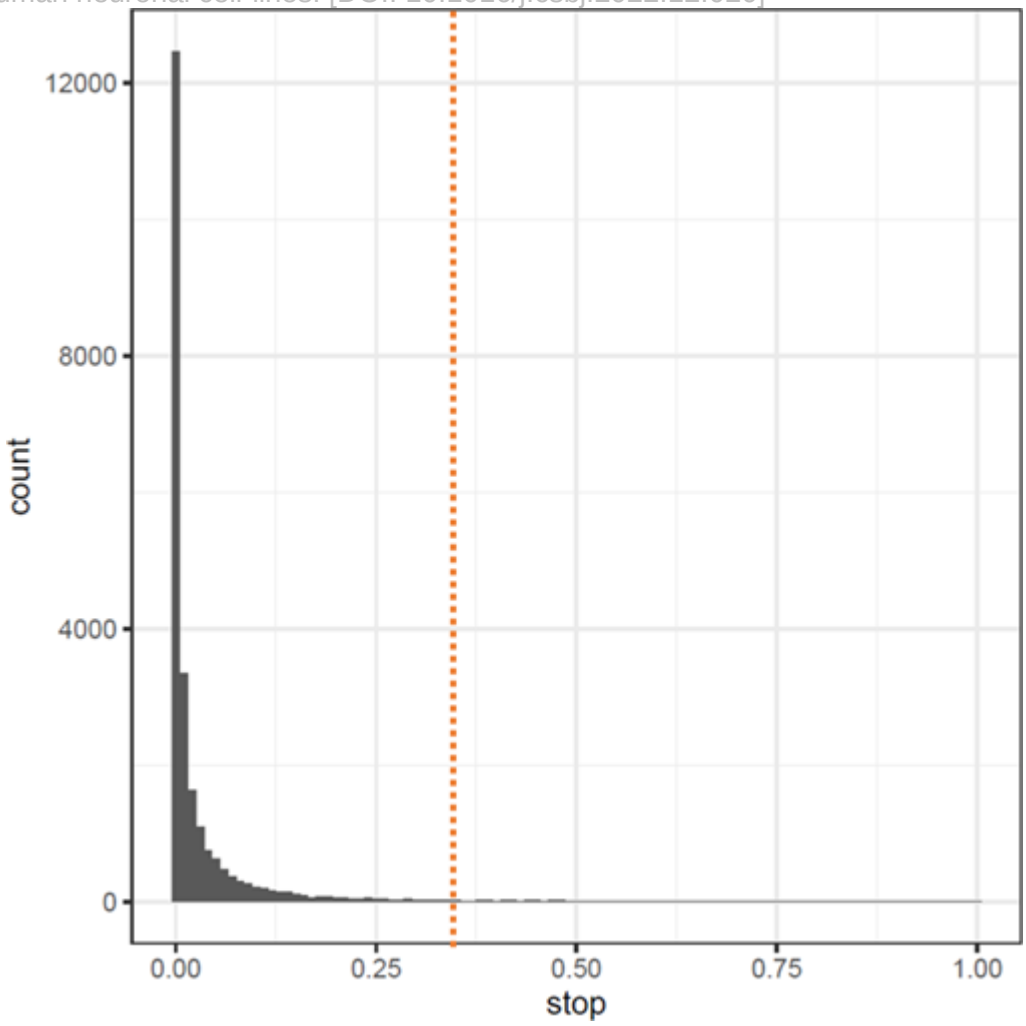


Figure S3 : Noise estimation of stop ratio from AlkAnilineSeq.

Each bar represents the counts of sites unseen by AlkAnilineSeq (no m7G, m3C ou D known enzymatic positions) according to their stop ratio (stop). The distribution of these undetectable sites has been considered as noise for detection threshold characterisation. Threshold determination through noise average + 3* noise standard deviation application occurs within a range from 0.263 (For A sites) to 0.418 (For G sites), with a default threshold of 0.353 while taking in account all nucleotides (represented here by an orange vertical bar).

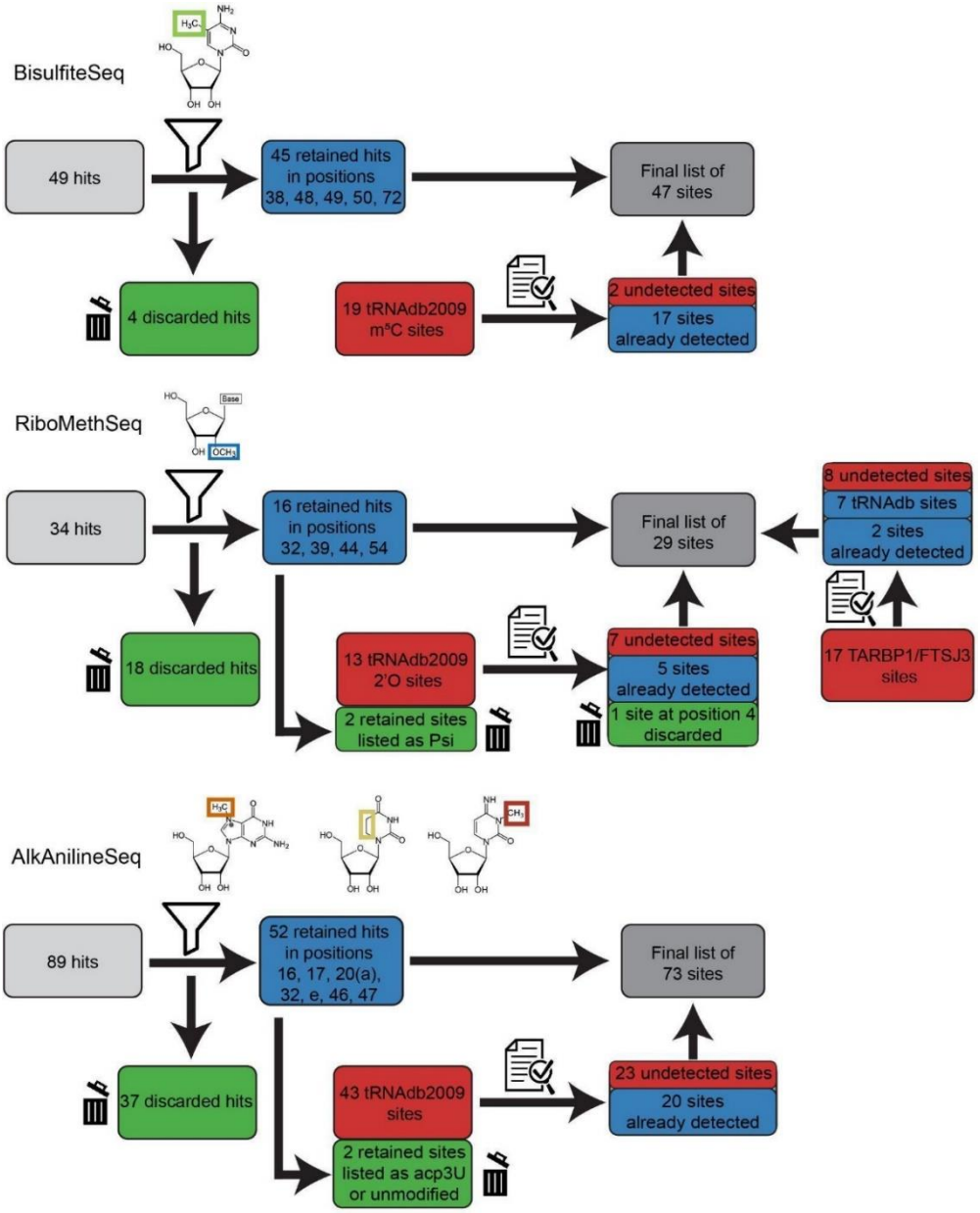


Figure S4 : Flow-chart on sites selection. Counting of retained and discarded hits along the analyses of the three different sequencing methods. On top, the sites selection for Bisulfite sequencing started with 49 total hits where 4 of them didn't belong to known m5C sites. After comparison with tRNAdb2009, 2 sites have been missed and have been added back to the list. For RiboMethSeq in the middle, out of the 34 initial hits, 18 didn't belong to known 2'-O-methylated sites. After comparison with tRNAdb2009 and two papers from the literature (Marchand et al., 2017 and Angelova and al. 2020) with validated sites for Gm18 and Nm32/34, 16 undetectedsites have been added back and 2 hits which are referenced as pseudouridine residues have been discarded. Finally, for AlkAnilineSeq at the bottom, 37 hits have been discarded from the 89 initial ones. Then, by comparing our hits with tRNAdb2009, 23 known sites (mainly dihydrouridine ones) has been missed and added back, while 1 hit which was originally retained has been identified as acp3U and another 1 was known as unmodified. The list of all discarded candidates is available in a separate file called "Supp_Table_all_discarded_candidates.csv"

Tables S2-S6 Wilcoxon's text for assessment of significative differences between datasets

All p-values below have been calculated with the function "wilcox.test" from the *stats* R package with default parameters, which means the alternative hypothesis tested here is "two.sided".

Table S2A: Significant differences estimation for m5C without cell line distinction.

	-Ang_tRNA	-Ang_tRF	+Ang_tRNA	+Ang_tRF
-Ang_tRNA	1			
-Ang_tRF	<0.0001 ****	1		
+Ang_tRNA	0.3966	0.0002 ***	1	
+Ang_tRF	0.0003 ***	0.2058	0.004 **	1

Table S2B: Significant differences estimation for m5C with cell line distinction.

	MZ_-Ang_tRNA	SH_-Ang_tRNA	MZ_-Ang_tRF	SH_-Ang_tRF	MZ_+Ang_tRNA	SH_+Ang_tRNA	MZ_+Ang_tRF	SH_+Ang_tRF
MZ_-Ang_tRNA	1							
SH_-Ang_tRNA	0.9646	1						
MZ_-Ang_tRF	0.0012 **	0.0011 **	1					
SH_-Ang_tRF	0.004 **	0.0065 **	0.7278	1				
MZ_+Ang_tRNA	0.2081	0.2339	0.0154 *	0.0406 *	1			
SH_+Ang_tRNA	0.7534	0.7444	0.0008 ***	0.0023 **	0.1318	1		
MZ_+Ang_tRF	0.0252 *	0.0258 *	0.1431	0.2978	0.2552	0.0175 *	1	
SH_+Ang_tRF	0.0035 **	0.0041 **	0.4420	0.8192	0.0406 *	0.0019 **	0.4091	1

Table S3A: Significant differences estimation for Nm without cell line distinction.

	-Ang_tRNA	-Ang_tRF	+Ang_tRNA	+Ang_tRF
-Ang_tRNA	1			
-Ang_tRF	0.0015 **	1		
+Ang_tRNA	0.6181	0.0039 **	1	
+Ang_tRF	0.003 **	0.6887	0.0129 *	1

Table S3B: Significant differences estimation for Nm with cell line distinction.

	MZ_-Ang_tRNA	SH_-Ang_tRNA	MZ_-Ang_tRF	SH_-Ang_tRF	MZ_+Ang_tRNA	SH_+Ang_tRNA	MZ_+Ang_tRF	SH_+Ang_tRF
MZ_-Ang_tRNA	1							
SH_-Ang_tRNA	0.2043	1						
MZ_-Ang_tRF	0.1823	0.0228 *	1					
SH_-Ang_tRF	0.0294 *	0.0019 **	0.5924	1				
MZ_+Ang_tRNA	0.9814	0.2838	0.1195	0.0177 *	1			
SH_+Ang_tRNA	0.7637	0.5376	0.0954	0.0117 *	0.7274	1		
MZ_+Ang_tRF	0.3631	0.0328 *	0.5028	0.0955	0.3049	0.1937	1	
SH_+Ang_tRF	0.0224 *	0.0023 **	0.3157	0.8034	0.0245 *	0.0135 *	0.0864	1

Table S4A: Significant differences estimation for m7G without cell line distinction.

	-Ang_tRNA	-Ang_tRF	+Ang_tRNA	+Ang_tRF
-Ang_tRNA	1			
-Ang_tRF	0.6476	1		
+Ang_tRNA	0.2239	0.0805	1	
+Ang_tRF	0.8886	0.53	0.2286	1

Table S4B: Significant differences estimation for m7G with cell line distinction.

	MZ_-Ang_tRNA	SH_-Ang_tRNA	MZ_-Ang_tRF	SH_-Ang_tRF	MZ_+Ang_tRNA	SH_+Ang_tRNA	MZ_+Ang_tRF	SH_+Ang_tRF
MZ_-Ang_tRNA	1							
SH_-Ang_tRNA	0.0106 *	1						
MZ_-Ang_tRF	0.0987	0.3057	1					
SH_-Ang_tRF	0.1139	0.3753	0.9729	1				
MZ_+Ang_tRNA	0.8119	0.0131 *	0.0789	0.1309	1			
SH_+Ang_tRNA	0.3571	0.0987	0.4134	0.4536	0.2181	1		
MZ_+Ang_tRF	0.4134	0.0987	0.4958	0.4332	0.3223	1	1	
SH_+Ang_tRF	0.1401	0.2897	0.9729	0.8384	0.0851	0.5177	0.5177	1

Table S5A: Significant differences estimation for m3C without cell line distinction.

	-Ang_tRNA	-Ang_tRF	+Ang_tRNA	+Ang_tRF
-Ang_tRNA	1			
-Ang_tRF	0.0276 *	1		
+Ang_tRNA	0.7901	0.0121 *	1	
+Ang_tRF	0.1011	0.5641	0.11	1

Table S5B : Significant differences estimation for m3C with cell line distinction.

	MZ_-Ang_tRNA	SH_-Ang_tRNA	MZ_-Ang_tRF	SH_-Ang_tRF	MZ_+Ang_tRNA	SH_+Ang_tRNA	MZ_+Ang_tRF	SH_+Ang_tRF
MZ_-Ang_tRNA	1							
SH_-Ang_tRNA	0.0293 *	1						
MZ_-Ang_tRF	0.6052	0.007 **	1					
SH_-Ang_tRF	0.8518	0.0207 *	0.8785	1				
MZ_+Ang_tRNA	0.2824	0.5054	0.065	0.1304	1			
SH_+Ang_tRNA	0.2284	0.6454	0.0499 *	0.1605	0.8785	1		
MZ_+Ang_tRF	0.8518	0.065	0.5737	0.7984	0.3823	0.2786	1	
SH_+Ang_tRF	0.9497	0.0281 *	0.6454	0.8785	0.2786	0.2345	0.9591	1

Table S6A : Significant differences estimation for D without cell line distinction.

	-Ang_tRNA	-Ang_tRF	+Ang_tRNA	+Ang_tRF
-Ang_tRNA	1			
-Ang_tRF	0.4496	1		
+Ang_tRNA	0.918	0.5311	1	
+Ang_tRF	0.5257	0.928	0.6514	1

Table S6B : Significant differences estimation for D with cell line distinction.

	MZ_-Ang_tRNA	SH_-Ang_tRNA	MZ_-Ang_tRF	SH_-Ang_tRF	MZ_+Ang_tRNA	SH_+Ang_tRNA	MZ_+Ang_tRF	SH_+Ang_tRF
MZ_-Ang_tRNA	1							
SH_-Ang_tRNA	0.9109	1						
MZ_-Ang_tRF	0.5193	0.5059	1					
SH_-Ang_tRF	0.6099	0.7624	0.7822	1				
MZ_+Ang_tRNA	0.9002	0.9106	0.6595	0.8089	1			
SH_+Ang_tRNA	0.918	0.9586	0.4787	0.7167	0.829	1		
MZ_+Ang_tRF	0.5847	0.5979	0.8628	0.8357	0.7688	0.6345	1	
SH_+Ang_tRF	0.7037	0.746	0.7889	0.9862	0.8833	0.7297	0.9037	1

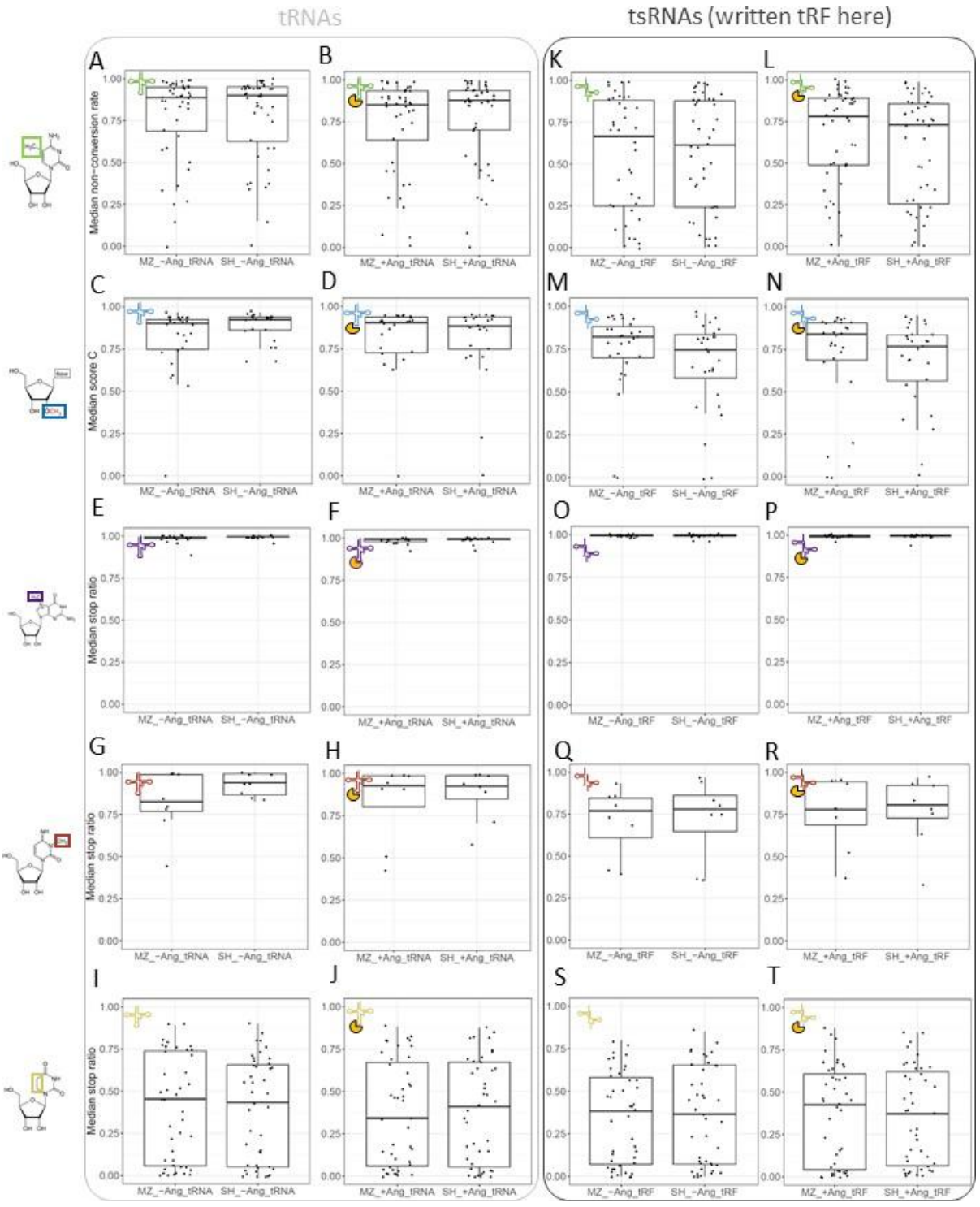


Figure S5: Alternative representation of figure 5

In contrast to Figure 5 separating samples according to their treatment, Panels (A-J) correspond to intact tRNAs samples, while panels (K-T) show the results for tsRNAs samples. Modification levels for m^5C (A, B, K, L), Nm (2'-O-Me) (C, D, M, N), m^7G (E, F, O, P), m^3C (G, H, Q, R) and dihydrouridine D (I, J, S, T) are shown. The values shown here are modification level medians of the respective detection scores obtained for three biological replicates.

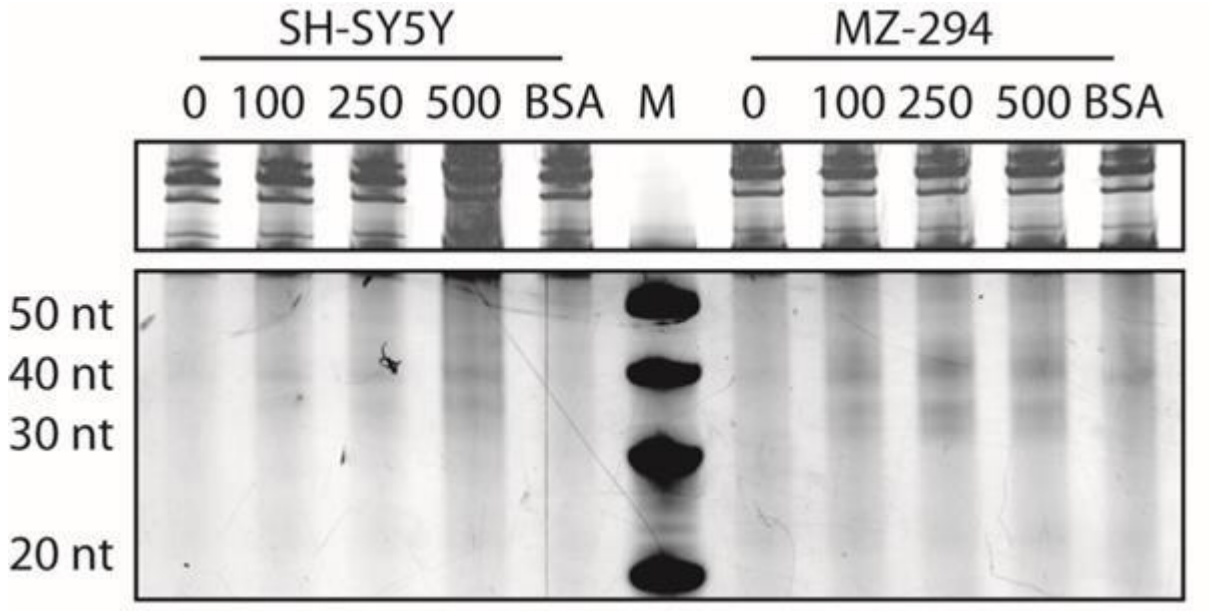


Figure S6: RNA gel showing RNA fragmentation in SH-SY5Y and MZ-294 cells treated with recombinant human Angiogenin. Cells were treated with recombinant human Angiogenin or BSA in Serum free Neurobasal media for 3 hours, and total RNA was loaded on a 15 % polyacrylamide TBE-UREA gel which was stained with SYBR Gold for visualisation. Execution and results of this experiments are in keeping with those previously published in [52] Hogg, M.C., et al., *5'ValCAC tRNA fragment generated as part of a protective angiogenin response provides prognostic value in amyotrophic lateral sclerosis.* Brain Commun, 2020. **2**(2): p. fcaa138.

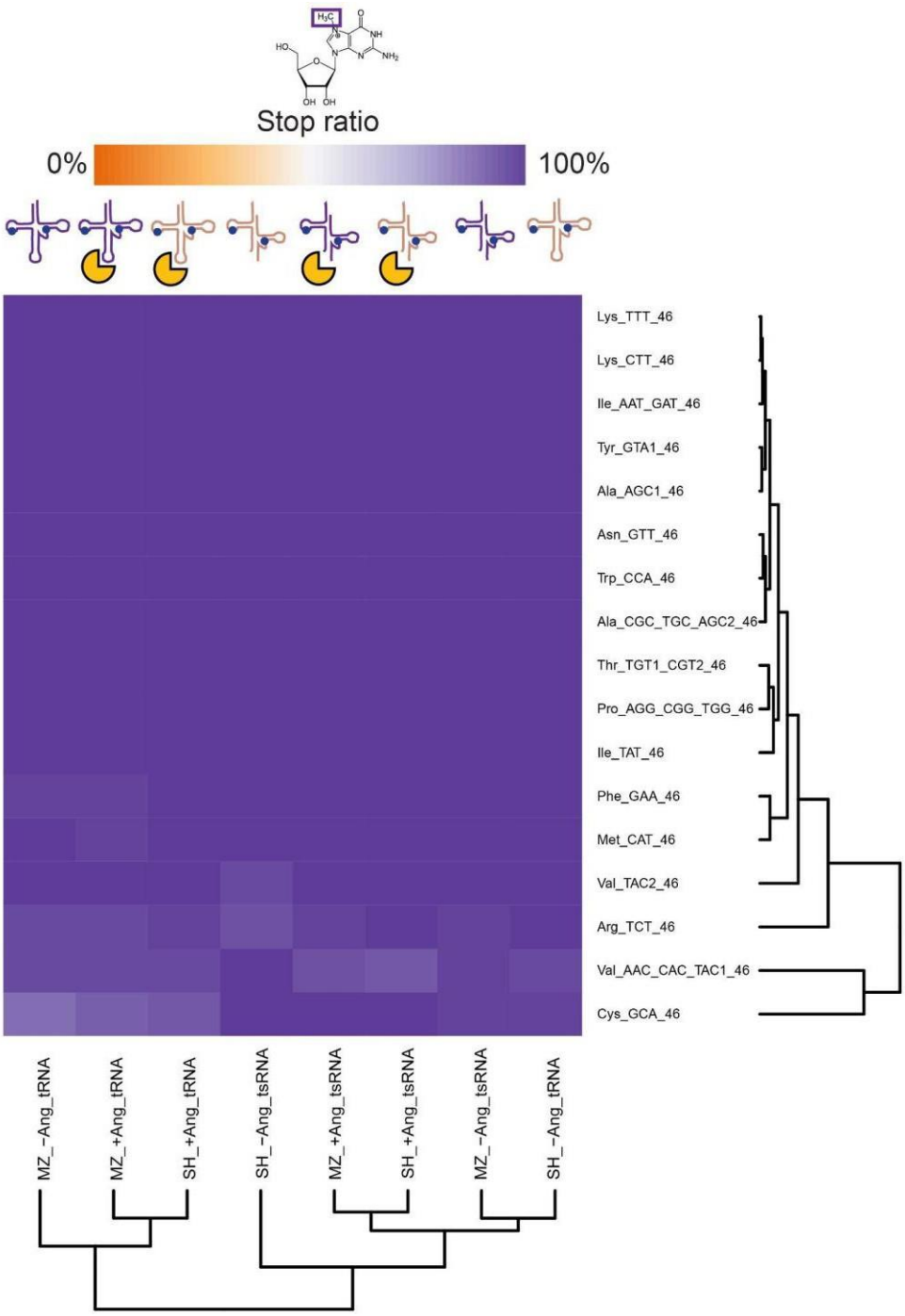


Figure S7 : Heatmap for the assessment of modification changes in individual isoacceptors and isodecoders of m7G sites.

This heatmap displays m7G stoichiometry across the different samples (X-axis) and the different sites retained for analysis (Y-axis) through its AlkAnilineSeq quantitative score stop ratio. This stoichiometry is color-coded from orange for low modification level through white for medium one to finish by purple for high. Each sample and each site are hierarchically clustered based on the Euclidean distances between each subset, represented by dendrograms at the bottom for samples and on the right for sites. No specific sites categorisation have been deduced from this heatmap.

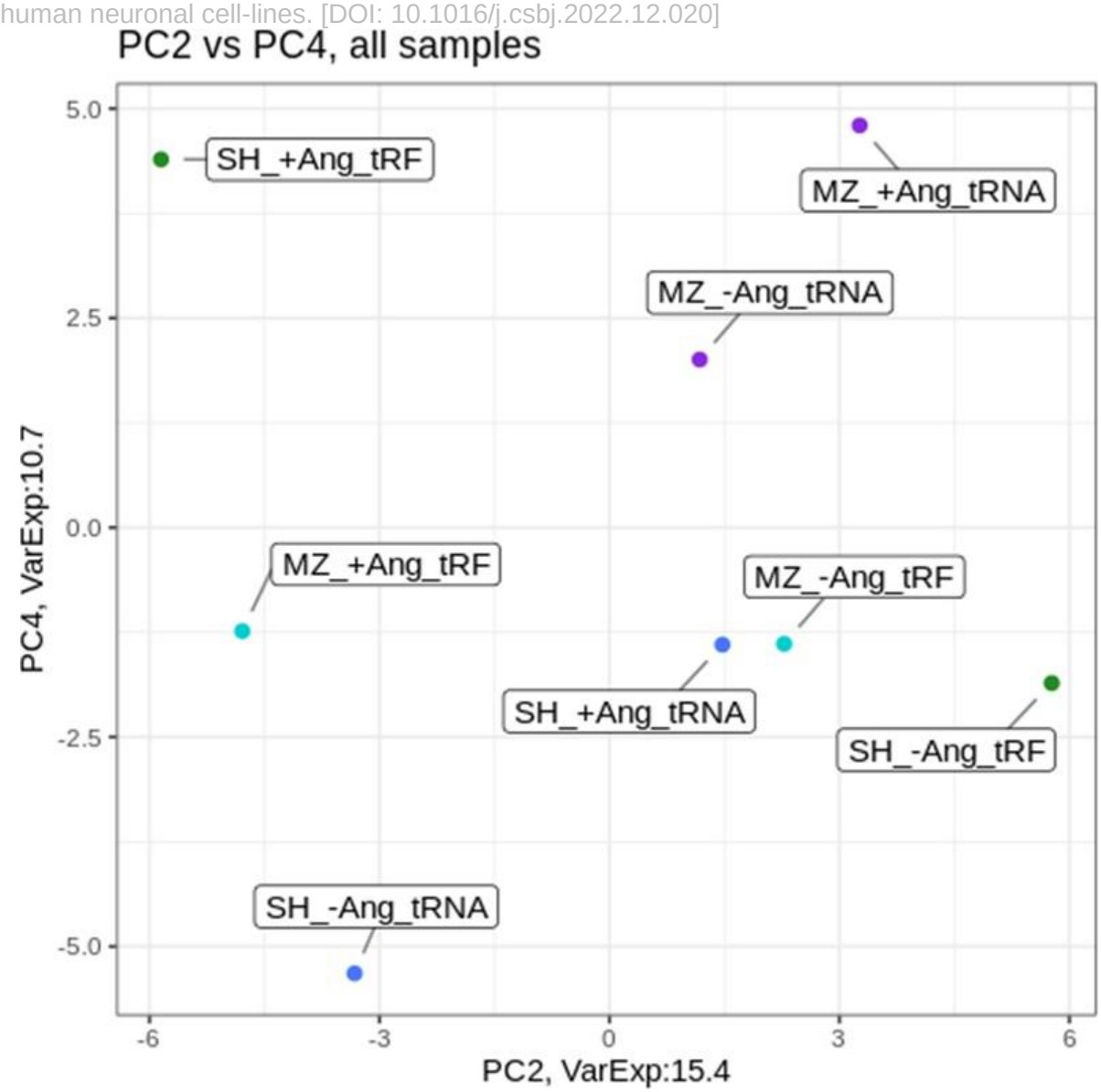


Figure S8 : Plot of PC2 versus PC4 segregates +ANG samples.
The combination of orthogonal vectors PC2 and PC4 from Figure 9 allows to illustrate a clear separation for tRNAs along the diagonal from bottom-left to top-right, while providing a distinct angiogenin clustering effect for tRFs along the X-axis. Correlating these two axes with the different sites provides the sites which best explains the biological differences for tRNAs and angiogenin effect for tsRNAs (here labeled “tRFs”) respectively.

Table S7. Contributors to the 4 first PC listed as loadings vectors. The latter correspond as correlation coefficients since PCA has been done on the data's correlation matrix.

Site	PC1 loadings	PC2 loadings	PC3 loadings	PC4 loadings	Modification
Ala_AGC1_17	0.41	0.17	-0.15	0.69	D
Ala_AGC1_20	0.51	0.20	-0.73	-0.04	D
Ala_AGC1_39	-0.92	-0.05	-0.20	-0.07	Nm
Ala_CGC_TGC_AGC2_49	-0.36	-0.11	0.28	0.74	m5C
Arg_ACG_16	-0.90	0.38	-0.09	0.00	D
Arg_ACG_47	0.01	-0.12	0.37	-0.36	D
Arg_CCG1_TCG_32	-0.42	-0.81	0.39	-0.02	Nm
Arg_CCT_CCG2_16	-0.50	0.08	0.76	0.02	D
Arg_CCT_CCG2_20	0.55	-0.53	0.07	0.17	D
Arg_CCT_CCG2_32	-0.84	0.33	0.29	0.11	m3C
Arg_TCT_16	-0.87	0.07	0.42	0.09	D

Arg_TCT_20a	-0.71	0.30	0.37	-0.51	D
Asn_GTT_16	0.53	0.67	-0.24	-0.12	D
Asn_GTT_20	0.33	0.40	0.81	-0.03	D
Asn_GTT_47	0.70	0.12	0.26	-0.57	D
Asn_GTT_54	-0.95	-0.13	-0.24	-0.09	Nm
Asp_GTC_20	0.76	-0.32	-0.11	-0.11	D
Asp_GTC_38	-0.21	0.66	0.14	-0.03	m5C
Asp_GTC_48	0.42	0.22	0.35	-0.63	m5C
Asp_GTC_49	0.08	0.38	-0.35	-0.80	m5C
Cys_GCA_20	-0.44	0.20	0.47	-0.18	D
Cys_GCA_32	-0.39	0.19	-0.61	-0.08	Nm
Cys_GCA_49	0.10	-0.62	0.52	0.43	m5C
Gln_CTG1_TTG_17	-0.27	0.89	0.09	0.08	D

Gln_CTG1_TTG_18	-0.84	-0.47	-0.12	-0.08	Nm
Gln_CTG1_TTG_20	-0.58	-0.45	0.18	-0.57	D
Gln_CTG1_TTG_20a	-0.76	-0.45	0.25	0.33	D
Gln_CTG1_TTG_32	0.37	-0.31	0.45	-0.66	Nm
Glu_CTC_TTC1_20	-0.72	0.34	0.27	0.38	D
Glu_CTC_TTC1_20a	-0.61	0.09	0.66	-0.39	D
Glu_CTC_TTC1_49	0.48	-0.43	0.33	-0.66	m5C
Glu_CTC_TTC1_50	0.66	-0.34	0.23	-0.60	m5C
Glu_CTC_TTC1_54	-0.80	-0.10	-0.25	-0.23	Nm
Glu_TTC2_20	-0.69	-0.66	0.21	0.08	D
Glu_TTC2_20a	-0.42	-0.83	-0.23	-0.13	D
Glu_TTC2_49	-0.55	0.18	-0.33	-0.67	m5C
Glu_TTC2_50	-0.65	0.24	-0.27	-0.51	m5C

Glu_TTC2_54	-0.96	-0.08	-0.19	-0.17	Nm
Gly_CCC1_GCC_38	0.78	-0.19	0.12	0.23	m5C
Gly_CCC1_GCC_48	0.71	-0.08	-0.32	-0.56	m5C
Gly_CCC1_GCC_49	-0.82	-0.21	-0.21	-0.34	m5C
Gly_CCC1_GCC_50	0.14	-0.60	-0.03	-0.58	m5C
Gly_CCC2_20	-0.89	-0.27	0.14	0.06	D
Gly_CCC2_32	0.56	-0.04	0.81	0.01	Nm
Gly_CCC2_39	-0.55	0.65	-0.06	-0.46	Nm
Gly_CCC2_49	0.57	-0.46	0.39	-0.54	m5C
Gly_CCC2_50	0.72	-0.50	0.15	-0.45	m5C
Gly_TCC_20	-0.74	-0.54	0.00	0.31	D
Gly_TCC_49	0.10	0.69	-0.35	0.47	m5C
Gly_TCC_50	0.56	-0.19	-0.20	-0.69	m5C

His_GTG_17	0.42	-0.23	-0.73	0.09	D
His_GTG_20	0.66	-0.55	-0.05	-0.34	D
His_GTG_20a	0.68	-0.33	0.36	-0.51	D
His_GTG_49	-0.85	-0.47	0.17	-0.06	m5C
His_GTG_50	0.08	-0.47	-0.24	-0.14	m5C
Ile_AAT_GAT_48	-0.96	0.03	0.27	0.05	m5C
Leu_CAG_CAA_32	-0.94	0.02	-0.24	-0.24	Nm
Leu_CAG_CAA_44	-0.81	-0.25	-0.49	0.06	Nm
Leu_CAG_CAA_48	-0.98	-0.02	0.07	-0.09	m5C
Leu_CAG_CAA_e2	0.48	-0.27	0.18	-0.60	m3C
Leu_TAA_48	-0.94	0.17	0.11	-0.03	m5C
Lys_CTT_16	-0.35	0.24	0.32	0.67	D
Lys_CTT_48	-0.97	-0.12	0.16	0.04	m5C

Lys_CTT_54	-0.86	0.15	-0.08	-0.31	Nm
Lys_TTT_16	-0.75	0.57	-0.17	0.12	D
Lys_TTT_49	-0.98	0.14	0.05	0.06	m5C
Lys_TTT_54	-0.80	0.39	-0.14	-0.19	Nm
Met_CAT_20	0.53	0.23	0.13	-0.65	m3C
Met_CAT_34	-0.38	0.56	0.22	-0.10	Nm
Met_CAT_47	-0.24	0.77	-0.10	-0.27	D
Met_CAT_49	-0.84	-0.01	0.17	-0.24	m5C
Phe_GAA_16	-0.80	-0.08	0.27	-0.38	D
Phe_GAA_17	-0.78	-0.03	0.36	-0.13	D
Phe_GAA_34	-0.86	-0.06	-0.18	0.12	Nm
Phe_GAA_47	0.22	0.46	0.80	-0.03	D
Phe_GAA_49	-0.94	-0.08	0.11	-0.28	m5C

Pro_AGG_CGG_TGG_20	-0.33	-0.86	-0.07	0.03	D
Pro_AGG_CGG_TGG_32	-0.89	-0.07	-0.41	-0.16	Nm
Pro_AGG_CGG_TGG_49	-0.74	-0.45	-0.40	-0.12	m5C
Pro_AGG_CGG_TGG_50	-0.89	-0.20	-0.25	-0.24	m5C
Ser_AGA_TGA_17	0.33	-0.65	0.34	0.36	D
Ser_AGA_TGA_18	-0.85	-0.15	-0.49	-0.05	Nm
Ser_AGA_TGA_20	0.22	0.60	0.41	-0.32	D
Ser_AGA_TGA_20a	0.42	-0.57	0.30	0.51	D
Ser_AGA_TGA_32	-0.86	-0.45	-0.14	0.18	m3C
Ser_AGA_TGA_44	-0.84	-0.24	-0.41	0.06	Nm
Ser_AGA_TGA_49	-0.98	0.10	0.09	-0.02	m5C
Ser_GCT_18	-0.79	0.32	-0.14	-0.38	Nm
Ser_GCT_32	-0.77	-0.47	0.14	0.20	m3C

Ser_GCT_39	-0.37	-0.44	-0.76	-0.22	Nm
Ser_GCT_44	-0.81	-0.35	-0.42	0.09	Nm
Ser_GCT_49	-0.94	0.18	0.09	0.02	m5C
Thr_AGT_CGT3_TGT2_32	-0.85	0.33	0.30	-0.25	m3C
Thr_AGT_CGT3_TGT2_47	0.04	-0.90	0.23	-0.10	D
Thr_CGT1_32	-0.76	-0.48	0.34	0.25	m3C
Thr_CGT1_48	-0.99	-0.09	0.07	0.02	m5C
Thr_CGT1_72	-0.81	-0.37	0.45	0.06	m5C
Thr_TGT1_CGT2_20	0.72	0.23	-0.38	-0.14	D
Thr_TGT1_CGT2_32	-0.95	0.16	0.06	-0.06	m3C
Thr_TGT1_CGT2_48	-0.95	-0.23	0.07	0.00	m5C
Thr_TGT1_CGT2_72	-0.71	-0.27	0.34	0.08	m5C
Trp_CCA_39	-0.91	0.29	0.02	-0.13	Nm

Tyr_GTA1_16	0.90	-0.10	-0.13	0.02	D
Tyr_GTA1_17	0.98	0.05	-0.15	0.07	D
Tyr_GTA1_47	-0.76	0.43	0.04	-0.01	D
Tyr_GTA1_48	-0.97	0.09	0.12	0.04	m5C
Val_AAC_CAC_TAC1_17	0.68	-0.37	0.43	0.42	D
Val_AAC_CAC_TAC1_38	0.42	0.42	0.72	-0.14	m5C
Val_TAC2_20	-0.55	0.65	0.38	0.27	D
Val_TAC2_20a	0.67	0.41	0.36	0.17	D
Val_TAC2_48	-0.72	-0.02	0.52	-0.06	m5C
Val_TAC2_49	-0.68	-0.18	0.52	-0.30	m5C

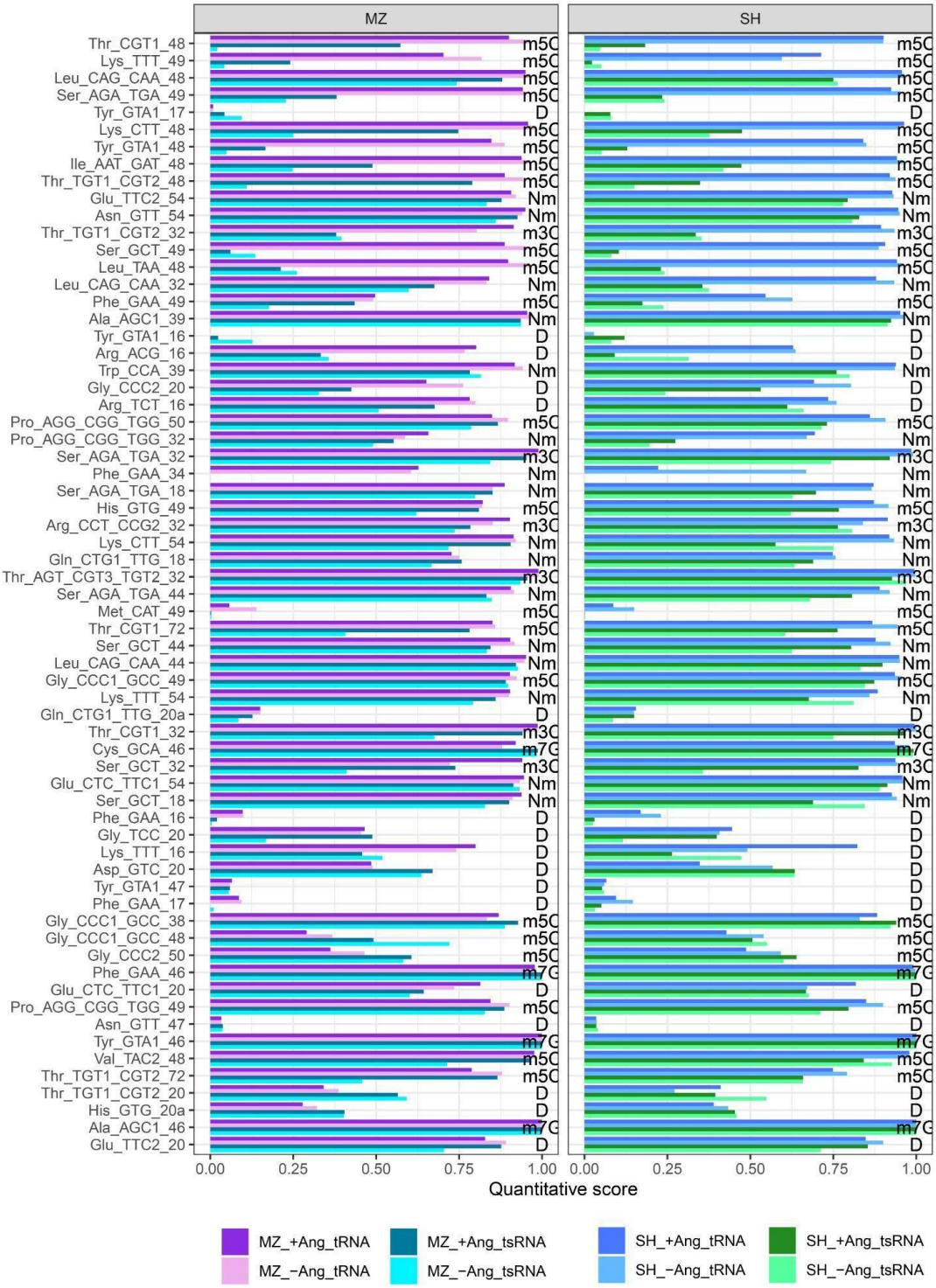


Figure S9: Barplots displaying modification stoichiometry values for main contributors of PC1.

Left panels for MZ cells, right panels for SH cells. The nature of modified nucleotide is shown at the top of each bar, and the identity of tRNA/tsRNA site on the left along the Y-axis. Color code indicates the identity of the sample (tRNA/tsRNA) and treatment with ANG (-Ang/+Ang).

-- End of article --

V. Conclusion and outlook

Optimisation of RiboMethSeq in adequation with biological studies

RiboMethSeq application to biological studies have been shown alongside its pipeline optimisation. The first application of RiboMethSeq on *E. coli* and *S. cerevisiae* tRNA (Section 4.1.1) indicates the potential of RiboMethSeq methodology on tRNAs analysis in an innate immune system project to find new therapeutic targets for viral infections. But this first work could at the time mostly apply to known tRNA sequences and 2'-O-sites (here, Gm18 with known snoRNP) due to high numbers of false positives hits.

Then the subsequent work from the same project (Section 4.1.4) displays new insights on Gm18 impacts on immune response and this is due to the optimisation of the pipeline (Section 4.1.2) decreasing false positive counts and allowing validation of reported Gm18 sites with help to epitranscriptomics-adapted tRNA sequences (Section 4.1.3).

For the latter, multiple articles on tRNA analyses don't explicitly cite their tRNA references for their studies except for their sources based on tRNA genes (gtRNAdb for exemple) or experimentally verified sequences (Modomics database). However, use of these sequences without including any modifications to the references is not adequate for epitranscriptomic studies. This is the main reason why we developed a method to create non-redundant tRNA references (Section 4.1.3) while explicitly describing the different steps along their creations for better reproducibility by the epitranscriptomic community. Moreover, this work has the vocation to give a ground to be criticized by experts in the field in order to improve the creation of tRNA references. This is why, in order to ensure transparency and free access to the different parameters used, the entire set of scripts used for non-redundant tRNA references creation has been published online on GitHub. We expect constructive debate around this work, leading to publication of tRNA references consensually accepted by all. This would ensure easier data comparisons between projects of different groups to streamline ground-breaking discoveries on tRNA.

Importance of statistical analyses applications for error minimization

As proven earlier for the two RiboMethSeq applications, optimisation of the bioinformatics pipeline is essential to deepen our knowledge on modifications. In that regard, use of powerful and adapted statistics is important for in-depth assessment of detection methods efficiency. Knowing the limitations of the different methodologies for modification detection is of the utmost importance for project application, since research groups would be able to correctly interpret the data instead of writing wrong conclusions based on biased results. However, this assessment should rely on spot-on measurement depending on the desired outcome.

As an example of statistical analyses application, one of the most used metrics in detection methods assessments are the sensitivity (which is the rate of modified sites which has been detected), and the specificity (which is the rate of unmodified sites which has been excluded from the predictions). From these metrics calculations, a Receiver Operating Characteristic curve (ROC curve) is usually drawn to represent the evolution of the method's predicted sensitivity in Y-axis along its False Positive rate (1 - specificity) in X-axis. An Area Under Curve (AUC) can be deduced from this plot. The AUC has a value between 0 and 1, and a strong predictive model has an AUC close to 0 or 1, 0.5 being the worst theoretical value. Those terms have been used times and times again by numerous groups[632–634] for their efficiency and simplicity.

However, those terms are sometimes not adapted to correctly assess the power of the detection methods. For example, in the case of epitranscriptomics, the specificity of detection methods is always very high (> 0.90) due to the fact modified residues are rare, meaning correct exclusion of unmodified sites can be easily achieved randomly. As an example, in poly-A transcript RNA, there is one m^6A nucleotide out of 800 other nucleotides[635] while m^6A is known to be enriched in those RNAs. If we randomly select a small fraction of A nucleotides in poly-A transcript and classify them as m^6A, our random selection would have a specificity around 99% since more than 99% of the sites are not m^6A (see Figure 3 below).

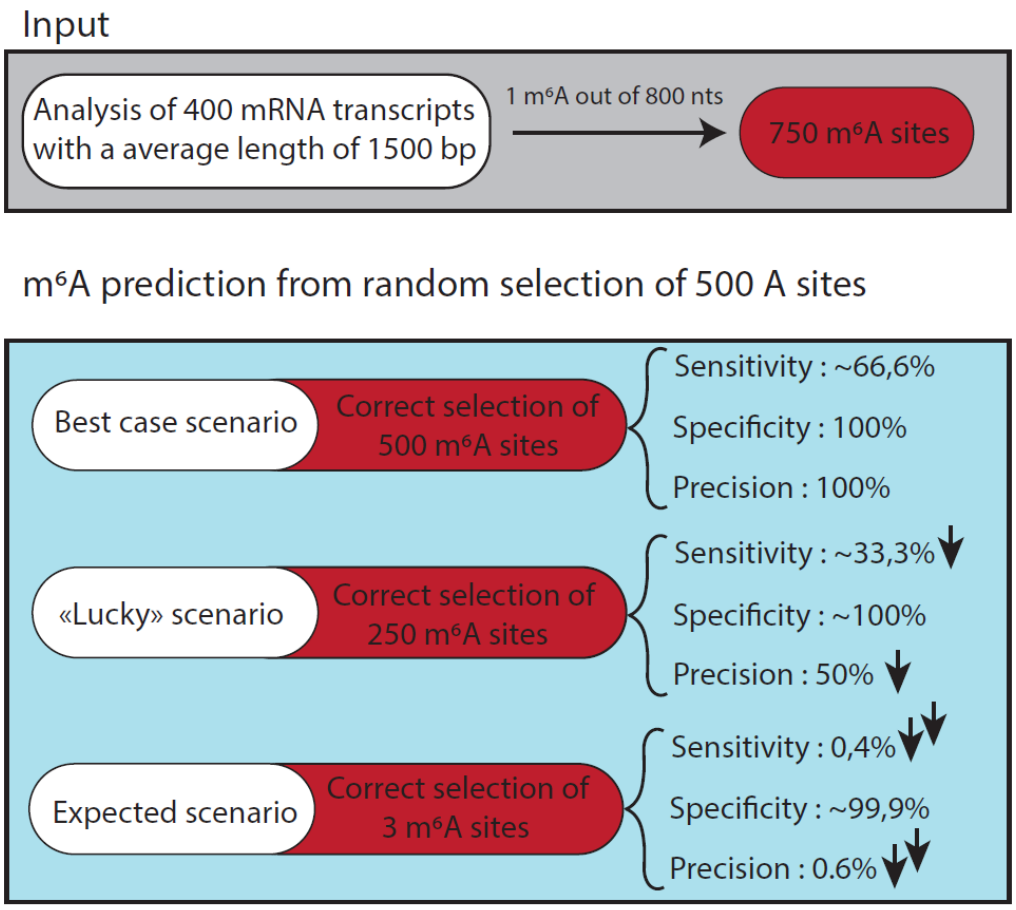


Figure 3: Confusion matrix metrics performance on simulated data.

This figure illustrates the evolution of three commonly used confusion matrix metrics - Sensitivity, Specificity and Precision - on theoretical transcriptomic data. The theoretical data content is specified in the top box in grey, and the theoretical metrics performance as displayed in a blue box below. Three scenarios were chosen from the best outcome (at the top) to the worst outcome (at the bottom) to evaluate metrics' effectiveness.

Specificity should then not be used alone to justify a method's efficiency. Unfortunately, this means also that obtaining a good AUC (so an AUC value far from 0.5) is relatively easy in epitranscriptomics with a high specificity from the start. Moreover, the AUC value itself is purely geometrical so not linked to any possible data interpretation. Sadly, this is a metrics still too routinely used for epitranscriptomics methods assessment despite its major drawbacks.
In line with this rationale, we have tried to find the best metrics to correctly assessed our results, using Matthews Correlation Coefficient (MCC) instead of AUC in our holistic optimisation of RiboMethSeq pipeline (Section 4.1.2) and to the use of precision instead of specificity in our machine learning paper (Section 4.1.5). Other metrics which have almost not been used yet could also be of interest for the epistrancriptomics community such as the Negative Predictive Value (NPV), accuracy (ACC) or the F1 score. Correct use of statistical tools should remain a priority to better gauge the strength and the limitations of current detection techniques in order to improve them for more reliable breakthrough discoveries.

Creation of modifications detection techniques for new epitranscriptomics discoveries

This work showed two different approaches for modifications detection by specific chemical signature entitled HydraPsiSeq (Section 4.2.1) for pseudouridine detection and NOseq (Section 4.2.3) for m^6A detection. Though detection methods already exist for these two modifications, they were also known for their limitations intrinsically linked to the nature of detection (use of chemical adducts or antibodies) and incomplete application of quantification of detected sites. Our new approach now allows quantification at single nucleotide resolution as shown in our work. But most importantly, this is an alternative to the different methods already present for cross-validation of the different newly discovered m^6A and Psi sites since little sites overlap has been found for methods using the same approach[272, 455].
The reasons behind these discrepancies are mainly due to major differences in the bioinformatic pipelines and parameters for analysis. In order to avoid this, we have published the detailed bioinformatic pipeline for HydraPsiSeq (Section 4.2.2) from the trimming and alignment parameters to the complete pipeline of the data treatment after alignment with graphs suggestion. Even though some bioinformatic steps are detailed in the previously cited methods, the entire process and parameters choice are

generally not sufficiently described or justified[453, 634], with mention of “R/python personalized pipeline”, making the reproducibility of the data treatment difficult. While the field needs proper pseudouridine detection methods to study its impact on rRNA regulation[636], the explicit protocol of HydraPsiSeq shows promises to answer the call.

Tool sharpening leading to cutting-edge epitranscriptomics discoveries

Finally, the last work on tsRNA (Section 4.3) points out how optimised pipeline and compilation of quantitative detection methods enhances in-depth analysis of tRNA epitranscriptome, such as trustworthiness of sites detection and modifications quantification of all sites. Even though the treatment and data nature behind the three detection methods are apart from each other, the bioinformatics data treatment is the same: trimming, alignment, scores calculations. This allows the application of the same statistical analysis to each of them displaying the transferability of developed statistical approaches to multiple sequencing techniques. Using the same statistical approaches occurs in unambiguous results comparisons for interpretation. For the latter, we have validated and endorsed previously seen results by other groups, but also find a new putative regulatory role for Dihydrouridine on tRNA fragmentation.
However, even if high-throughput sequencing strength resides in providing results in a relatively short time compared to chromatography methods, the latter are the main reliable methods to confirm presence of putative sites. Putative yet undiscovered modified sites provided by high-throughput sequencing detection methods should be taken with caution.
Nonetheless, epitranscriptomics studies through modified sites detection at nucleotide resolution help the scientific community to gain better understanding of molecular mechanisms central to pathogenicity and drug design for various diseases such as cancers, viral infections and neurodegenerative diseases. Due to the highly beneficial outcome of epitranscriptomics studies applications for society, we have to ensure the methods applied within are up to the task.

Bibliography

[1] CRICK, F. H. On Protein Synthesis. *Symp. Soc. Exp. Biol.*, **1958**, *12*, 138–163.

[2] Buccitelli, C.; Selbach, M. MRNAs, Proteins and the Emerging Principles of Gene Expression Control. *Nat. Rev. Genet.*, **2020**, *21* (10), 630–644. https://doi.org/10.1038/s41576-020-0258-4.

[3] Grabski, D. F.; Broseus, L.; Kumari, B.; Rekosh, D.; Hammarskjold, M.-L.; Ritchie, W. Intron Retention and Its Impact on Gene Expression and Protein Diversity: A Review and a Practical Guide. *Wiley Interdiscip. Rev. RNA*, **2021**, *12* (1), e1631. https://doi.org/10.1002/wrna.1631.

[4] Houseley, J.; Tollervey, D. The Many Pathways of RNA Degradation. *Cell*, **2009**, *136* (4), 763–776. https://doi.org/10.1016/j.cell.2009.01.019.

[5] Lee, Y.; Rio, D. C. Mechanisms and Regulation of Alternative Pre-MRNA Splicing. *Annu. Rev. Biochem.*, **2015**, *84*, 291–323. https://doi.org/10.1146/annurev-biochem-060614-034316.

[6] Curinha, A.; Oliveira Braz, S.; Pereira-Castro, I.; Cruz, A.; Moreira, A. Implications of Polyadenylation in Health and Disease. *Nucl. Austin Tex*, **2014**, *5* (6), 508–519. https://doi.org/10.4161/nucl.36360.

[7] Hodinka, R. L. Respiratory RNA Viruses. *Microbiol. Spectr.*, **2016**, *4* (4). https://doi.org/10.1128/microbiolspec.DMIH2-0028-2016.

[8] Lundstrom, K. Therapeutic Applications for Oncolytic Self-Replicating RNA Viruses. *Int. J. Mol. Sci.*, **2022**, *23* (24), 15622. https://doi.org/10.3390/ijms232415622.

[9] Pardi, N.; Hogan, M. J.; Porter, F. W.; Weissman, D. MRNA Vaccines - a New Era in Vaccinology. *Nat. Rev. Drug Discov.*, **2018**, *17* (4), 261–279. https://doi.org/10.1038/nrd.2017.243.

[10] Alberts, B.; Johnson, A.; Lewis, J.; Raff, M.; Roberts, K.; Walter, P. From DNA to RNA. In *Molecular Biology of the Cell. 4th edition*; Garland Science, 2002.

[11] Weisser, M.; Ban, N. Extensions, Extra Factors, and Extreme Complexity: Ribosomal Structures Provide Insights into Eukaryotic Translation. *Cold Spring Harb. Perspect. Biol.*, **2019**, *11* (9), a032367. https://doi.org/10.1101/cshperspect.a032367.

[12] Dalla Venezia, N.; Vincent, A.; Marcel, V.; Catez, F.; Diaz, J.-J. Emerging Role of Eukaryote Ribosomes in Translational Control. *Int. J. Mol. Sci.*, **2019**, *20* (5), 1226. https://doi.org/10.3390/ijms20051226.

[13] Sloan, K. E.; Warda, A. S.; Sharma, S.; Entian, K.-D.; Lafontaine, D. L. J.; Bohnsack, M. T. Tuning the Ribosome: The Influence of RRNA Modification on Eukaryotic Ribosome Biogenesis and Function. *RNA Biol.*, **2017**, *14* (9), 1138–1152. https://doi.org/10.1080/15476286.2016.1259781.

[14] Lorenz, C.; Lünse, C. E.; Mörl, M. TRNA Modifications: Impact on Structure and Thermal Adaptation. *Biomolecules*, **2017**, *7* (2), 35. https://doi.org/10.3390/biom7020035.

[15] Väre, V. Y. P.; Eruysal, E. R.; Narendran, A.; Sarachan, K. L.; Agris, P. F. Chemical and Conformational Diversity of Modified Nucleosides Affects TRNA Structure and Function. *Biomolecules*, **2017**, *7* (1), 29. https://doi.org/10.3390/biom7010029.

[16] Geslain, R.; Ribas de Pouplana, L. Regulation of RNA Function by Aminoacylation and Editing? *Trends Genet. TIG*, **2004**, *20* (12), 604–610. https://doi.org/10.1016/j.tig.2004.09.012.
[17] Steiner, R. E.; Ibba, M. Regulation of TRNA-Dependent Translational Quality Control. *IUBMB Life*, **2019**, *71* (8), 1150–1157. https://doi.org/10.1002/iub.2080.
[18] Aregger, M.; Cowling, V. H. Regulation of MRNA Capping in the Cell Cycle. *RNA Biol.*, **2017**, *14* (1), 11–14. https://doi.org/10.1080/15476286.2016.1251540.
[19] Mayr, C. What Are 3' UTRs Doing? *Cold Spring Harb. Perspect. Biol.*, **2019**, *11* (10), a034728. https://doi.org/10.1101/cshperspect.a034728.
[20] Hoernes, T. P.; Hüttenhofer, A.; Erlacher, M. D. MRNA Modifications: Dynamic Regulators of Gene Expression? *RNA Biol.*, **2016**, *13* (9), 760–765. https://doi.org/10.1080/15476286.2016.1203504.
[21] Kilchert, C.; Vasiljeva, L. MRNA Quality Control Goes Transcriptional. *Biochem. Soc. Trans.*, **2013**, *41* (6), 1666–1672. https://doi.org/10.1042/BST20130202.
[22] Chen, J.-Q.; Szodoray, P.; Zeher, M. Toll-Like Receptor Pathways in Autoimmune Diseases. *Clin. Rev. Allergy Immunol.*, **2016**, *50* (1), 1–17. https://doi.org/10.1007/s12016-015-8473-z.
[23] Lam, J. K. W.; Chow, M. Y. T.; Zhang, Y.; Leung, S. W. S. SiRNA Versus MiRNA as Therapeutics for Gene Silencing. *Mol. Ther. Nucleic Acids*, **2015**, *4* (9), e252. https://doi.org/10.1038/mtna.2015.23.
[24] Kuscu, C.; Kumar, P.; Kiran, M.; Su, Z.; Malik, A.; Dutta, A. TRNA Fragments (TRFs) Guide Ago to Regulate Gene Expression Post-Transcriptionally in a Dicer-Independent Manner. *RNA N. Y. N*, **2018**, *24* (8), 1093–1105. https://doi.org/10.1261/rna.066126.118.
[25] Park, J.; Ahn, S. H.; Shin, M. G.; Kim, H. K.; Chang, S. TRNA-Derived Small RNAs: Novel Epigenetic Regulators. *Cancers*, **2020**, *12* (10), 2773. https://doi.org/10.3390/cancers12102773.
[26] Cui, H.; Li, H.; Wu, H.; Du, F.; Xie, X.; Zeng, S.; Zhang, Z.; Dong, K.; Shang, L.; Jing, C.; et al. A Novel 3'tRNA-Derived Fragment TRF-Val Promotes Proliferation and Inhibits Apoptosis by Targeting EEF1A1 in Gastric Cancer. *Cell Death Dis.*, **2022**, *13* (5), 471. https://doi.org/10.1038/s41419-022-04930-6.
[27] Hou, J.; Li, Q.; Wang, J.; Lu, W. TRFs and TRNA Halves: Novel Cellular Defenders in Multiple Biological Processes. *Curr. Issues Mol. Biol.*, **2022**, *44* (12), 5949–5962. https://doi.org/10.3390/cimb44120405.
[28] Deryusheva, S.; Talross, G. J. S.; Gall, J. G. SnoRNA Guide Activities: Real and Ambiguous. *RNA N. Y. N*, **2021**, *27* (11), 1363–1373. https://doi.org/10.1261/rna.078916.121.
[29] Ojha, S.; Malla, S.; Lyons, S. M. SnoRNPs: Functions in Ribosome Biogenesis. *Biomolecules*, **2020**, *10* (5), 783. https://doi.org/10.3390/biom10050783.
[30] Boccaletto, P.; Machnicka, M. A.; Purta, E.; Piatkowski, P.; Baginski, B.; Wirecki, T. K.; de Crécy-Lagard, V.; Ross, R.; Limbach, P. A.; Kotter, A.; et al. MODOMICS: A Database of RNA Modification Pathways. 2017 Update. *Nucleic Acids Res.*, **2018**, *46* (D1), D303–D307. https://doi.org/10.1093/nar/gkx1030.
[31] Helm, M.; Motorin, Y. Detecting RNA Modifications in the Epitranscriptome: Predict and Validate. *Nat. Rev. Genet.*, **2017**, *18* (5), 275–291. https://doi.org/10.1038/nrg.2016.169.
[32] Li, X.; Xiong, X.; Yi, C. Epitranscriptome Sequencing Technologies: Decoding RNA Modifications. *Nat. Methods*, **2016**, *14* (1), 23–31. https://doi.org/10.1038/nmeth.4110.

[33] Schwartz, S. Cracking the Epitranscriptome. *RNA N. Y. N*, **2016**, *22* (2), 169–174. https://doi.org/10.1261/rna.054502.115.
[34] Schwartz, S.; Motorin, Y. Next-Generation Sequencing Technologies for Detection of Modified Nucleotides in RNAs. *RNA Biol.*, **2017**, *14* (9), 1124–1137. https://doi.org/10.1080/15476286.2016.1251543.
[35] Baskin, F.; Dekker, C. A. A Rapid and Specific Assay for Sugar Methylation in Ribonucleic Acid. *J. Biol. Chem.*, **1967**, *242* (22), 5447–5449.
[36] Moore, P. B. Methylation of Messenger RNA in Escherichia Coli. *J. Mol. Biol.*, **1966**, *18* (1), 38–47. https://doi.org/10.1016/s0022-2836(66)80074-8.
[37] Nichols, J. L.; Lane, B. G. N-4-Methyl-2'-O-Methyl Cytidine and Other Methyl-Substituted Nucleoside Constituents of Escherichia Coli Ribosomal and Soluble RNA. *Biochim. Biophys. Acta*, **1966**, *119* (3), 649–651. https://doi.org/10.1016/0005-2787(66)90147-x.
[38] Nichols, J. L.; Lane, B. G. In Vitro O2'-Methylation of Sugars in E. Coli RNA. II. Methylation of Ribosomal and Transfer RNA by Homologous Methylases in Crude Cell-Free Extracts and Particulate Suspensions from a Relaxed Mutant of E. Coli. *Can. J. Biochem.*, **1968**, *46* (12), 1487–1495. https://doi.org/10.1139/o68-222.
[39] Gefter, M. L. The in Vitro Synthesis of 2'-Omethylguanosine and 2-Methylthio 6N (Gamma,Gamma, Dimethylallyl) Adenosine in Transfer RNA of Escherichia Coli. *Biochem. Biophys. Res. Commun.*, **1969**, *36* (3), 435–441. https://doi.org/10.1016/0006-291x(69)90583-x.
[40] Lane, B. G.; Tamaoki, T. Methylated Bases and Sugars in 16-S and 28-S RNA from L Cells. *Biochim. Biophys. Acta*, **1969**, *179* (2), 332–340. https://doi.org/10.1016/0005-2787(69)90041-0.
[41] Lau, R. Y.; Lane, B. G. The Sequence Distribution of O-2'-Methylribose in Yeast Ribosomal RNA. *Can. J. Biochem.*, **1971**, *49* (4), 431–436. https://doi.org/10.1139/o71-063.
[42] Nichols, J. L.; Lane, B. G. In Vivo Incorporation of Methyl Groups into the Ribose of Escherichia Coli Ribosomal RNA. *J. Mol. Biol.*, **1967**, *30* (3), 477–489. https://doi.org/10.1016/0022-2836(67)90363-4.
[43] Svensson, I.; Björk, G.; Björk, W.; Johansson, K. E.; Johansson, A. Evidence for Enzymatic Methylation in Vitro of the Ribose Moiety of RNA. *Biochem. Biophys. Res. Commun.*, **1968**, *31* (2), 216–221. https://doi.org/10.1016/0006-291x(68)90733-x.
[44] Björk, G. R.; Svensson, I. Studies on Microbial RNA. Fractionation of TRNA Methylases from Saccharomyces Cerevisiae. *Eur. J. Biochem.*, **1969**, *9* (2), 207–215. https://doi.org/10.1111/j.1432-1033.1969.tb00596.x.
[45] Munns, T. W.; Sims, H. F. Methylation and Processing of Transfer Ribonucleic Acid in Mammalian and Bacterial Cells. *J. Biol. Chem.*, **1975**, *250* (6), 2143–2149.
[46] Grummt, I.; Loening, U.; Slack, J. M. Methylation of Ribosomal-Precursor RNA, Synthesized in Vitro, by Isolated Rat-Liver Nucleoli. *Eur. J. Biochem.*, **1975**, *59* (1), 313–318. https://doi.org/10.1111/j.1432-1033.1975.tb02457.x.
[47] Brand, R. C.; Klootwijk, J.; Van Steenbergen, T. J.; De Kok, A. J.; Planta, R. J. Secondary Methylation of Yeast Ribosomal Precursor RNA. *Eur. J. Biochem.*, **1977**, *75* (1), 311–318. https://doi.org/10.1111/j.1432-1033.1977.tb11531.x.
[48] Hughes, D. G.; Maden, B. E. The Pseudouridine Contents of the Ribosomal Ribonucleic Acids of Three Vertebrate Species. Numerical Correspondence

between Pseudouridine Residues and 2'-O-Methyl Groups Is Not Always Conserved. *Biochem. J.*, **1978**, *171* (3), 781–786. https://doi.org/10.1042/bj1710781.

[49] Hsuchen, C. C.; Dubin, D. T. Methylation Patterns of Mycoplasma Transfer and Ribosomal Ribonucleic Acid. *J. Bacteriol.*, **1980**, *144* (3), 991–998. https://doi.org/10.1128/jb.144.3.991-998.1980.

[50] Vold, B. Modified Nucleosides of Bacillus Subtilis Transfer Ribonucleic Acids. *J. Bacteriol.*, **1976**, *127* (1), 258–267. https://doi.org/10.1128/jb.127.1.258-267.1976.

[51] White, B. N. Studies on Lysine, Glutamine and Glutamic Acid TRNAs from Drosophila. *Biochim. Biophys. Acta*, **1975**, *395* (3), 322–328. https://doi.org/10.1016/0005-2787(75)90203-8.

[52] Viale, G. L. Overmethylated T-RNAs in Human Gliomas. *Riv. Neurobiol. Organo Uff. Della Soc. Dei Neurol. Neuroradiol. E Neurochir. Osp.*, **1969**, *15* (3), 505–516.

[53] Rogg, H.; Müller, P.; Keith, G.; Staehelin, M. Chemical Basis for Brain-Specific Serine Transfer RNAs. *Proc. Natl. Acad. Sci. U. S. A.*, **1977**, *74* (10), 4243–4247. https://doi.org/10.1073/pnas.74.10.4243.

[54] Ro-Choi, T. S.; Choi, Y. C.; Henning, D.; McCloskey, J.; Busch, H. Nucleotide Sequence of U-2 Ribonucleic Acid. The Sequence of the 5'-Terminal Oligonucleotide. *J. Biol. Chem.*, **1975**, *250* (10), 3921–3928.

[55] Dimock, K.; Stoltzfus, C. M. Processing and Function of Undermethylated Chicken Embryo Fibroblast MRNA. *J. Biol. Chem.*, **1979**, *254* (13), 5591–5594.

[56] Heckle, W. L.; Fenton, R. G.; Wood, T. G.; Merkel, C. G.; Lingrel, J. B. Methylated Nucleosides in Globin MRNA from Mouse Nucleated Erythroid Cells. *J. Biol. Chem.*, **1977**, *252* (5), 1764–1770.

[57] Nichols, J. L.; Welder, L. The Modified Nucleotide Constituents of Human Prostatic Cancer Cell (MA-160) Poly(A)-Containing RNA. *Biochim. Biophys. Acta*, **1980**, *608* (1), 1–18. https://doi.org/10.1016/0005-2787(80)90128-8.

[58] Ouellette, A. J.; Frederick, D.; Malt, R. A. Methylated Messenger RNA in Mouse Kidney. *Biochemistry*, **1975**, *14* (20), 4361–4367. https://doi.org/10.1021/bi00691a003.

[59] Khan, M. S.; Maden, B. E. Conformation of Methylated Sequences in HeLa Cell 18-S Ribosomal RNA: Nuclease S1 as a Probe. *Eur. J. Biochem.*, **1978**, *84* (1), 241–250. https://doi.org/10.1111/j.1432-1033.1978.tb12162.x.

[60] Prusiner, P.; Yathindra, N.; Sundaralingam, M. Effect of Ribose O(2')-Methylation on the Conformation of Nucleosides and Nucleotides. *Biochim. Biophys. Acta*, **1974**, *366* (2), 115–123. https://doi.org/10.1016/0005-2787(74)90325-6.

[61] Cundliffe, E. Mechanism of Resistance to Thiostrepton in the Producing-Organism Streptomyces Azureus. *Nature*, **1978**, *272* (5656), 792–795. https://doi.org/10.1038/272792a0.

[62] Cundliffe, E.; Thompson, J. Ribose Methylation and Resistance to Thiostrepton. *Nature*, **1979**, *278* (5707), 859–861. https://doi.org/10.1038/278859a0.

[63] Birkedal, U.; Christensen-Dalsgaard, M.; Krogh, N.; Sabarinathan, R.; Gorodkin, J.; Nielsen, H. Profiling of Ribose Methylations in RNA by High-Throughput Sequencing. *Angew. Chem. Int. Ed Engl.*, **2015**, *54* (2), 451–455. https://doi.org/10.1002/anie.201408362.

[64] Marchand, V.; Blanloeil-Oillo, F.; Helm, M.; Motorin, Y. Illumina-Based RiboMethSeq Approach for Mapping of 2'-O-Me Residues in RNA. *Nucleic Acids Res.*, **2016**, *44* (16), e135. https://doi.org/10.1093/nar/gkw547.

[65] Marchand, V.; Pichot, F.; Thüring, K.; Ayadi, L.; Freund, I.; Dalpke, A.; Helm, M.; Motorin, Y. Next-Generation Sequencing-Based RiboMethSeq Protocol for Analysis of TRNA 2'-O-Methylation. *Biomolecules*, **2017**, *7* (1). https://doi.org/10.3390/biom7010013.

[66] Marchand, V.; Ayadi, L.; El Hajj, A.; Blanloeil-Oillo, F.; Helm, M.; Motorin, Y. High-Throughput Mapping of 2'-O-Me Residues in RNA Using Next-Generation Sequencing (Illumina RiboMethSeq Protocol). *Methods Mol. Biol. Clifton NJ*, **2017**, *1562*, 171–187. https://doi.org/10.1007/978-1-4939-6807-7_12.

[67] Incarnato, D.; Anselmi, F.; Morandi, E.; Neri, F.; Maldotti, M.; Rapelli, S.; Parlato, C.; Basile, G.; Oliviero, S. High-Throughput Single-Base Resolution Mapping of RNA 2´-O-Methylated Residues. *Nucleic Acids Res.*, **2017**, *45* (3), 1433–1441. https://doi.org/10.1093/nar/gkw810.

[68] Zhu, Y.; Pirnie, S. P.; Carmichael, G. G. High-Throughput and Site-Specific Identification of 2'-O-Methylation Sites Using Ribose Oxidation Sequencing (RibOxi-Seq). *RNA N. Y. N*, **2017**, *23* (8), 1303–1314. https://doi.org/10.1261/rna.061549.117.

[69] Dai, Q.; Moshitch-Moshkovitz, S.; Han, D.; Kol, N.; Amariglio, N.; Rechavi, G.; Dominissini, D.; He, C. Nm-Seq Maps 2'-O-Methylation Sites in Human MRNA with Base Precision. *Nat. Methods*, **2017**, *14* (7), 695–698. https://doi.org/10.1038/nmeth.4294.

[70] Brown, R. S.; Rubin, J. R.; Rhodes, D.; Guilley, H.; Simoncsits, A.; Brownlee, G. G. The Nucleoside Sequence of Tyrosine TRNA from Bacillus Stearothermophilus. *Nucleic Acids Res.*, **1978**, *5* (1), 23–36. https://doi.org/10.1093/nar/5.1.23.

[71] Takeda, N.; Nakamura, M.; Yoshizumi, H.; Tatematsu, A. Detection of Ribose-Methylated Nucleotides in Pyrodictium Occultum TRNA by Liquid Chromatography--Frit-Fast Atom Bombardment Mass Spectrometry. *J. Chromatogr. B Biomed. Appl.*, **1994**, *660* (2), 223–233. https://doi.org/10.1016/0378-4347(94)00299-1.

[72] Kuratani, M.; Bessho, Y.; Nishimoto, M.; Grosjean, H.; Yokoyama, S. Crystal Structure and Mutational Study of a Unique SpoU Family Archaeal Methylase That Forms 2'-O-Methylcytidine at Position 56 of TRNA. *J. Mol. Biol.*, **2008**, *375* (4), 1064–1075. https://doi.org/10.1016/j.jmb.2007.11.023.

[73] Renalier, M.-H.; Joseph, N.; Gaspin, C.; Thebault, P.; Mougin, A. The Cm56 TRNA Modification in Archaea Is Catalyzed Either by a Specific 2'-O-Methylase, or a C/D SRNP. *RNA N. Y. N*, **2005**, *11* (7), 1051–1063. https://doi.org/10.1261/rna.2110805.

[74] Droogmans, L.; Grosjean, H. 2'-O-Methylation and Inosine Formation in the Wobble Position of Anticodon-Substituted TRNA-Phe in a Homologous Yeast in Vitro System. *Biochimie*, **1991**, *73* (7–8), 1021–1025. https://doi.org/10.1016/0300-9084(91)90143-o.

[75] Kotelawala, L.; Grayhack, E. J.; Phizicky, E. M. Identification of Yeast TRNA Um(44) 2'-O-Methyltransferase (Trm44) and Demonstration of a Trm44 Role in Sustaining Levels of Specific TRNA(Ser) Species. *RNA N. Y. N*, **2008**, *14* (1), 158–169. https://doi.org/10.1261/rna.811008.

[76] Wilkinson, M. L.; Crary, S. M.; Jackman, J. E.; Grayhack, E. J.; Phizicky, E. M. The 2'-O-Methyltransferase Responsible for Modification of Yeast TRNA at Position 4. *RNA N. Y. N*, **2007**, *13* (3), 404–413. https://doi.org/10.1261/rna.399607.

[77] Guy, M. P.; Phizicky, E. M. Conservation of an Intricate Circuit for Crucial Modifications of the TRNAPhe Anticodon Loop in Eukaryotes. *RNA N. Y. N*, **2015**, *21* (1), 61–74. https://doi.org/10.1261/rna.047639.114.

[78] Guy, M. P.; Podyma, B. M.; Preston, M. A.; Shaheen, H. H.; Krivos, K. L.; Limbach, P. A.; Hopper, A. K.; Phizicky, E. M. Yeast Trm7 Interacts with Distinct Proteins for Critical Modifications of the TRNAPhe Anticodon Loop. *RNA N. Y. N*, **2012**, *18* (10), 1921–1933. https://doi.org/10.1261/rna.035287.112.

[79] Gross, H. J.; Simsek, M.; Raba, M.; Limburg, K.; Heckman, J.; Raj Bhandary, U. L. 2'-O-Methyl Ribothymidine: A Component of Rabbit Liver Lysine Transfer RNA. *Nucleic Acids Res.*, **1974**, *1* (1), 35–43. https://doi.org/10.1093/nar/1.1.35.

[80] Sprinzl, M.; Dank, N.; Nock, S.; Schön, A. Compilation of TRNA Sequences and Sequences of TRNA Genes. *Nucleic Acids Res.*, **1991**, *19 Suppl* (Suppl), 2127–2171. https://doi.org/10.1093/nar/19.suppl.2127.

[81] Païs de Barros, J. P.; Keith, G.; El Adlouni, C.; Glasser, A. L.; Mack, G.; Dirheimer, G.; Desgrès, J. 2'-O-Methyl-5-Formylcytidine (F5Cm), a New Modified Nucleotide at the "wobble" of Two Cytoplasmic TRNAs Leu (NAA) from Bovine Liver. *Nucleic Acids Res.*, **1996**, *24* (8), 1489–1496. https://doi.org/10.1093/nar/24.8.1489.

[82] Glasser, A. L.; el Adlouni, C.; Keith, G.; Sochacka, E.; Malkiewicz, A.; Santos, M.; Tuite, M. F.; Desgrès, J. Presence and Coding Properties of 2'-O-Methyl-5-Carbamoylmethyluridine (Ncm5Um) in the Wobble Position of the Anticodon of TRNA(Leu) (U*AA) from Brewer's Yeast. *FEBS Lett.*, **1992**, *314* (3), 381–385. https://doi.org/10.1016/0014-5793(92)81510-s.

[83] Raviprakash, K. S.; Cherayil, J. D. 2'-O-Methyl-1-Methyl Adenosine: A New Modified Nucleoside in Ragi (Eleusine Coracana) TRNA. *Biochem. Biophys. Res. Commun.*, **1984**, *121* (1), 243–248. https://doi.org/10.1016/0006-291x(84)90713-7.

[84] Bachellerie, J. P.; Cavaillé, J.; Hüttenhofer, A. The Expanding SnoRNA World. *Biochimie*, **2002**, *84* (8), 775–790. https://doi.org/10.1016/s0300-9084(02)01402-5.

[85] Decatur, W. A.; Fournier, M. J. RRNA Modifications and Ribosome Function. *Trends Biochem. Sci.*, **2002**, *27* (7), 344–351. https://doi.org/10.1016/s0968-0004(02)02109-6.

[86] Kimura, S.; Dedon, P. C.; Waldor, M. K. Comparative TRNA Sequencing and RNA Mass Spectrometry for Surveying TRNA Modifications. *Nat. Chem. Biol.*, **2020**, *16* (9), 964–972. https://doi.org/10.1038/s41589-020-0558-1.

[87] Sergeeva, O. V.; Bogdanov, A. A.; Sergiev, P. V. What Do We Know about Ribosomal RNA Methylation in Escherichia Coli? *Biochimie*, **2015**, *117*, 110–118. https://doi.org/10.1016/j.biochi.2014.11.019.

[88] Kimura, S.; Suzuki, T. Fine-Tuning of the Ribosomal Decoding Center by Conserved Methyl-Modifications in the Escherichia Coli 16S RRNA. *Nucleic Acids Res.*, **2010**, *38* (4), 1341–1352. https://doi.org/10.1093/nar/gkp1073.

[89] Purta, E.; O'Connor, M.; Bujnicki, J. M.; Douthwaite, S. YgdE Is the 2'-O-Ribose Methyltransferase RlmM Specific for Nucleotide C2498 in Bacterial 23S RRNA.

Mol. Microbiol., **2009**, *72* (5), 1147–1158. https://doi.org/10.1111/j.1365-2958.2009.06709.x.

[90] Moazed, D.; Noller, H. F. Sites of Interaction of the CCA End of Peptidyl-TRNA with 23S RRNA. *Proc. Natl. Acad. Sci. U. S. A.*, **1991**, *88* (9), 3725–3728. https://doi.org/10.1073/pnas.88.9.3725.

[91] Lövgren, J. M.; Wikström, P. M. The RlmB Gene Is Essential for Formation of Gm2251 in 23S RRNA but Not for Ribosome Maturation in Escherichia Coli. *J. Bacteriol.*, **2001**, *183* (23), 6957–6960. https://doi.org/10.1128/JB.183.23.6957-6960.2001.

[92] Sirum-Connolly, K.; Mason, T. L. Functional Requirement of a Site-Specific Ribose Methylation in Ribosomal RNA. *Science*, **1993**, *262* (5141), 1886–1889. https://doi.org/10.1126/science.8266080.

[93] Kim, D. F.; Green, R. Base-Pairing between 23S RRNA and TRNA in the Ribosomal A Site. *Mol. Cell*, **1999**, *4* (5), 859–864. https://doi.org/10.1016/s1097-2765(00)80395-0.

[94] Widerak, M.; Kern, R.; Malki, A.; Richarme, G. U2552 Methylation at the Ribosomal A-Site Is a Negative Modulator of Translational Accuracy. *Gene*, **2005**, *347* (1), 109–114. https://doi.org/10.1016/j.gene.2004.12.025.

[95] Bügl, H.; Fauman, E. B.; Staker, B. L.; Zheng, F.; Kushner, S. R.; Saper, M. A.; Bardwell, J. C.; Jakob, U. RNA Methylation under Heat Shock Control. *Mol. Cell*, **2000**, *6* (2), 349–360. https://doi.org/10.1016/s1097-2765(00)00035-6.

[96] Caldas, T.; Binet, E.; Bouloc, P.; Costa, A.; Desgres, J.; Richarme, G. The FtsJ/RrmJ Heat Shock Protein of Escherichia Coli Is a 23 S Ribosomal RNA Methyltransferase. *J. Biol. Chem.*, **2000**, *275* (22), 16414–16419. https://doi.org/10.1074/jbc.M001854200.

[97] Caldas, T.; Binet, E.; Bouloc, P.; Richarme, G. Translational Defects of Escherichia Coli Mutants Deficient in the Um(2552) 23S Ribosomal RNA Methyltransferase RrmJ/FTSJ. *Biochem. Biophys. Res. Commun.*, **2000**, *271* (3), 714–718. https://doi.org/10.1006/bbrc.2000.2702.

[98] Guymon, R.; Pomerantz, S. C.; Crain, P. F.; McCloskey, J. A. Influence of Phylogeny on Posttranscriptional Modification of RRNA in Thermophilic Prokaryotes: The Complete Modification Map of 16S RRNA of Thermus Thermophilus. *Biochemistry*, **2006**, *45* (15), 4888–4899. https://doi.org/10.1021/bi052579p.

[99] Srivastava, R.; Gopinathan, K. P. Ribosomal RNA Methylation in Mycobacterium Smegmatis SN2. *Biochem. Int.*, **1987**, *15* (6), 1179–1188.

[100] Dennis, P. P.; Tripp, V.; Lui, L.; Lowe, T.; Randau, L. C/D Box SRNA-Guided 2'-O-Methylation Patterns of Archaeal RRNA Molecules. *BMC Genomics*, **2015**, *16*, 632. https://doi.org/10.1186/s12864-015-1839-z.

[101] Seistrup, K. H.; Rose, S.; Birkedal, U.; Nielsen, H.; Huber, H.; Douthwaite, S. Bypassing RRNA Methylation by RsmA/Dim1during Ribosome Maturation in the Hyperthermophilic Archaeon Nanoarchaeum Equitans. *Nucleic Acids Res.*, **2017**, *45* (4), 2007–2015. https://doi.org/10.1093/nar/gkw839.

[102] Klootwijk, J.; Planta, R. J. Analysis of the Methylation Sites in Yeast Ribosomal RNA. *Eur. J. Biochem.*, **1973**, *39* (2), 325–333. https://doi.org/10.1111/j.1432-1033.1973.tb03130.x.

[103] Yang, J.; Sharma, S.; Kötter, P.; Entian, K.-D. Identification of a New Ribose Methylation in the 18S RRNA of S. Cerevisiae. *Nucleic Acids Res.*, **2015**, *43* (4), 2342–2352. https://doi.org/10.1093/nar/gkv058.

[104] Yang, J.; Sharma, S.; Watzinger, P.; Hartmann, J. D.; Kötter, P.; Entian, K.-D. Mapping of Complete Set of Ribose and Base Modifications of Yeast RRNA by RP-HPLC and Mung Bean Nuclease Assay. *PloS One*, **2016**, *11* (12), e0168873. https://doi.org/10.1371/journal.pone.0168873.

[105] Taoka, M.; Nobe, Y.; Yamaki, Y.; Yamauchi, Y.; Ishikawa, H.; Takahashi, N.; Nakayama, H.; Isobe, T. The Complete Chemical Structure of Saccharomyces Cerevisiae RRNA: Partial Pseudouridylation of U2345 in 25S RRNA by SnoRNA SnR9. *Nucleic Acids Res.*, **2016**, *44* (18), 8951–8961. https://doi.org/10.1093/nar/gkw564.

[106] Sirum-Connolly, K.; Peltier, J. M.; Crain, P. F.; McCloskey, J. A.; Mason, T. L. Implications of a Functional Large Ribosomal RNA with Only Three Modified Nucleotides. *Biochimie*, **1995**, *77* (1–2), 30–39. https://doi.org/10.1016/0300-9084(96)88101-6.

[107] Krogh, N.; Jansson, M. D.; Häfner, S. J.; Tehler, D.; Birkedal, U.; Christensen-Dalsgaard, M.; Lund, A. H.; Nielsen, H. Profiling of 2'-O-Me in Human RRNA Reveals a Subset of Fractionally Modified Positions and Provides Evidence for Ribosome Heterogeneity. *Nucleic Acids Res.*, **2016**, *44* (16), 7884–7895. https://doi.org/10.1093/nar/gkw482.

[108] Sharma, S.; Marchand, V.; Motorin, Y.; Lafontaine, D. L. J. Identification of Sites of 2'-O-Methylation Vulnerability in Human Ribosomal RNAs by Systematic Mapping. *Sci. Rep.*, **2017**, *7* (1), 11490. https://doi.org/10.1038/s41598-017-09734-9.

[109] Erales, J.; Marchand, V.; Panthu, B.; Gillot, S.; Belin, S.; Ghayad, S. E.; Garcia, M.; Laforêts, F.; Marcel, V.; Baudin-Baillieu, A.; et al. Evidence for RRNA 2'-O-Methylation Plasticity: Control of Intrinsic Translational Capabilities of Human Ribosomes. *Proc. Natl. Acad. Sci. U. S. A.*, **2017**, *114* (49), 12934–12939. https://doi.org/10.1073/pnas.1707674114.

[110] Taoka, M.; Nobe, Y.; Yamaki, Y.; Sato, K.; Ishikawa, H.; Izumikawa, K.; Yamauchi, Y.; Hirota, K.; Nakayama, H.; Takahashi, N.; et al. Landscape of the Complete RNA Chemical Modifications in the Human 80S Ribosome. *Nucleic Acids Res.*, **2018**, *46* (18), 9289–9298. https://doi.org/10.1093/nar/gky811.

[111] Incarnato, D.; Anselmi, F.; Morandi, E.; Neri, F.; Maldotti, M.; Rapelli, S.; Parlato, C.; Basile, G.; Oliviero, S. High-Throughput Single-Base Resolution Mapping of RNA 2´-O-Methylated Residues. *Nucleic Acids Res.*, **2017**, *45* (3), 1433–1441. https://doi.org/10.1093/nar/gkw810.

[112] Maden, B. E. Identification of the Locations of the Methyl Groups in 18 S Ribosomal RNA from Xenopus Laevis and Man. *J. Mol. Biol.*, **1986**, *189* (4), 681–699. https://doi.org/10.1016/0022-2836(86)90498-5.

[113] Maden, B. E. Locations of Methyl Groups in 28 S RRNA of Xenopus Laevis and Man. Clustering in the Conserved Core of Molecule. *J. Mol. Biol.*, **1988**, *201* (2), 289–314. https://doi.org/10.1016/0022-2836(88)90139-8.

[114] Rebane, A.; Roomere, H.; Metspalu, A. Locations of Several Novel 2'-O-Methylated Nucleotides in Human 28S RRNA. *BMC Mol. Biol.*, **2002**, *3*, 1. https://doi.org/10.1186/1471-2199-3-1.

[115] Taoka, M.; Nobe, Y.; Yamaki, Y.; Sato, K.; Ishikawa, H.; Izumikawa, K.; Yamauchi, Y.; Hirota, K.; Nakayama, H.; Takahashi, N.; et al. Landscape of the Complete RNA Chemical Modifications in the Human 80S Ribosome. *Nucleic Acids Res.*, **2018**, *46* (18), 9289–9298. https://doi.org/10.1093/nar/gky811.

[116] Kiss-László, Z.; Henry, Y.; Bachellerie, J. P.; Caizergues-Ferrer, M.; Kiss, T. Site-Specific Ribose Methylation of Preribosomal RNA: A Novel Function for Small Nucleolar RNAs. *Cell*, **1996**, *85* (7), 1077–1088. https://doi.org/10.1016/s0092-8674(00)81308-2.

[117] Gumienny, R.; Jedlinski, D. J.; Schmidt, A.; Gypas, F.; Martin, G.; Vina-Vilaseca, A.; Zavolan, M. High-Throughput Identification of C/D Box SnoRNA Targets with CLIP and RiboMeth-Seq. *Nucleic Acids Res.*, **2017**, *45* (5), 2341–2353. https://doi.org/10.1093/nar/gkw1321.

[118] Nazar, R. N.; Sitz, T. O.; Busch, H. Homologies in Eukaryotic 5.8S Ribosomal RNA. *Biochem. Biophys. Res. Commun.*, **1975**, *62* (3), 736–743. https://doi.org/10.1016/0006-291x(75)90461-1.

[119] Nazar, R. N.; Sitz, T. O.; Somers, K. D. Cytoplasmic Methylation of Mature 5.8 S Ribosomal RNA. *J. Mol. Biol.*, **1980**, *142* (1), 117–121. https://doi.org/10.1016/0022-2836(80)90209-0.

[120] Baer, R. J.; Dubin, D. T. Methylated Regions of Hamster Mitochondrial Ribosomal RNA: Structural and Functional Correlates. *Nucleic Acids Res.*, **1981**, *9* (2), 323–337. https://doi.org/10.1093/nar/9.2.323.

[121] Lee, K.-W.; Okot-Kotber, C.; LaComb, J. F.; Bogenhagen, D. F. Mitochondrial Ribosomal RNA (RRNA) Methyltransferase Family Members Are Positioned to Modify Nascent RRNA in Foci near the Mitochondrial DNA Nucleoid. *J. Biol. Chem.*, **2013**, *288* (43), 31386–31399. https://doi.org/10.1074/jbc.M113.515692.

[122] Lee, K.-W.; Bogenhagen, D. F. Assignment of 2'-O-Methyltransferases to Modification Sites on the Mammalian Mitochondrial Large Subunit 16 S Ribosomal RNA (RRNA). *J. Biol. Chem.*, **2014**, *289* (36), 24936–24942. https://doi.org/10.1074/jbc.C114.581868.

[123] Barneche, F.; Gaspin, C.; Guyot, R.; Echeverría, M. Identification of 66 Box C/D SnoRNAs in Arabidopsis Thaliana: Extensive Gene Duplications Generated Multiple Isoforms Predicting New Ribosomal RNA 2'-O-Methylation Sites. *J. Mol. Biol.*, **2001**, *311* (1), 57–73. https://doi.org/10.1006/jmbi.2001.4851.

[124] Higa, S.; Maeda, N.; Kenmochi, N.; Tanaka, T. Location of 2(')-O-Methyl Nucleotides in 26S RRNA and Methylation Guide SnoRNAs in Caenorhabditis Elegans. *Biochem. Biophys. Res. Commun.*, **2002**, *297* (5), 1344–1349. https://doi.org/10.1016/s0006-291x(02)02377-x.

[125] Schnare, M. N.; Gray, M. W. Complete Modification Maps for the Cytosolic Small and Large Subunit RRNAs of Euglena Gracilis: Functional and Evolutionary Implications of Contrasting Patterns between the Two RRNA Components. *J. Mol. Biol.*, **2011**, *413* (1), 66–83. https://doi.org/10.1016/j.jmb.2011.08.037.

[126] Inesta-Vaquera, F.; Cowling, V. H. Regulation and Function of CMTR1-Dependent MRNA Cap Methylation. *Wiley Interdiscip. Rev. RNA*, **2017**, *8* (6), e1450. https://doi.org/10.1002/wrna.1450.

[127] Smietanski, M.; Werner, M.; Purta, E.; Kaminska, K. H.; Stepinski, J.; Darzynkiewicz, E.; Nowotny, M.; Bujnicki, J. M. Structural Analysis of Human 2'-O-Ribose Methyltransferases Involved in MRNA Cap Structure Formation. *Nat. Commun.*, **2014**, *5*, 3004. https://doi.org/10.1038/ncomms4004.

[128] Gu, J.; Patton, J. R.; Shimba, S.; Reddy, R. Localization of Modified Nucleotides in Schizosaccharomyces Pombe Spliceosomal Small Nuclear RNAs: Modified Nucleotides Are Clustered in Functionally Important Regions. *RNA N. Y. N*, **1996**, *2* (9), 909–918.

[129] Meier, U. T. RNA Modification in Cajal Bodies. *RNA Biol.*, **2017**, *14* (6), 693–700. https://doi.org/10.1080/15476286.2016.1249091.
[130] Yu, Y. T.; Shu, M. D.; Steitz, J. A. Modifications of U2 SnRNA Are Required for SnRNP Assembly and Pre-MRNA Splicing. *EMBO J.*, **1998**, *17* (19), 5783–5795. https://doi.org/10.1093/emboj/17.19.5783.
[131] Krogh, N.; Kongsbak-Wismann, M.; Geisler, C.; Nielsen, H. Substoichiometric Ribose Methylations in Spliceosomal SnRNAs. *Org. Biomol. Chem.*, **2017**, *15* (42), 8872–8876. https://doi.org/10.1039/c7ob02317k.
[132] Ji, L.; Chen, X. Regulation of Small RNA Stability: Methylation and Beyond. *Cell Res.*, **2012**, *22* (4), 624–636. https://doi.org/10.1038/cr.2012.36.
[133] Borges, F.; Martienssen, R. A. The Expanding World of Small RNAs in Plants. *Nat. Rev. Mol. Cell Biol.*, **2015**, *16* (12), 727–741. https://doi.org/10.1038/nrm4085.
[134] Gumienny, R.; Jedlinski, D. J.; Schmidt, A.; Gypas, F.; Martin, G.; Vina-Vilaseca, A.; Zavolan, M. High-Throughput Identification of C/D Box SnoRNA Targets with CLIP and RiboMeth-Seq. *Nucleic Acids Res.*, **2017**, *45* (5), 2341–2353. https://doi.org/10.1093/nar/gkw1321.
[135] Sharma, E.; Sterne-Weiler, T.; O'Hanlon, D.; Blencowe, B. J. Global Mapping of Human RNA-RNA Interactions. *Mol. Cell*, **2016**, *62* (4), 618–626. https://doi.org/10.1016/j.molcel.2016.04.030.
[136] Boschi-Muller, S.; Motorin, Y. Chemistry Enters Nucleic Acids Biology: Enzymatic Mechanisms of RNA Modification. *Biochem. Biokhimiia*, **2013**, *78* (13), 1392–1404. https://doi.org/10.1134/S0006297913130026.
[137] Kozbial, P. Z.; Mushegian, A. R. Natural History of S-Adenosylmethionine-Binding Proteins. *BMC Struct. Biol.*, **2005**, *5*, 19. https://doi.org/10.1186/1472-6807-5-19.
[138] Schubert, H. L.; Blumenthal, R. M.; Cheng, X. Many Paths to Methyltransfer: A Chronicle of Convergence. *Trends Biochem. Sci.*, **2003**, *28* (6), 329–335. https://doi.org/10.1016/S0968-0004(03)00090-2.
[139] Hori, H. Transfer RNA Methyltransferases with a SpoU-TrmD (SPOUT) Fold and Their Modified Nucleosides in TRNA. *Biomolecules*, **2017**, *7* (1), 23. https://doi.org/10.3390/biom7010023.
[140] Watanabe, K.; Nureki, O.; Fukai, S.; Ishii, R.; Okamoto, H.; Yokoyama, S.; Endo, Y.; Hori, H. Roles of Conserved Amino Acid Sequence Motifs in the SpoU (TrmH) RNA Methyltransferase Family. *J. Biol. Chem.*, **2005**, *280* (11), 10368–10377. https://doi.org/10.1074/jbc.M411209200.
[141] Watanabe, K.; Nureki, O.; Fukai, S.; Endo, Y.; Hori, H. Functional Categorization of the Conserved Basic Amino Acid Residues in TrmH (TRNA (Gm18) Methyltransferase) Enzymes. *J. Biol. Chem.*, **2006**, *281* (45), 34630–34639. https://doi.org/10.1074/jbc.M606141200.
[142] Persson, B. C.; Jäger, G.; Gustafsson, C. The SpoU Gene of Escherichia Coli, the Fourth Gene of the SpoT Operon, Is Essential for TRNA (Gm18) 2'-O-Methyltransferase Activity. *Nucleic Acids Res.*, **1997**, *25* (20), 4093–4097. https://doi.org/10.1093/nar/25.20.4093.
[143] Kumagai, I.; Watanabe, K.; Oshima, T. A Thermostable TRNA (Guanosine-2')-Methyltransferase from Thermus Thermophilus HB27 and the Effect of Ribose Methylation on the Conformational Stability of TRNA. *J. Biol. Chem.*, **1982**, *257* (13), 7388–7395.

[144] Hori, H.; Suzuki, T.; Sugawara, K.; Inoue, Y.; Shibata, T.; Kuramitsu, S.; Yokoyama, S.; Oshima, T.; Watanabe, K. Identification and Characterization of TRNA (Gm18) Methyltransferase from Thermus Thermophilus HB8: Domain Structure and Conserved Amino Acid Sequence Motifs. *Genes Cells Devoted Mol. Cell. Mech.*, **2002**, *7* (3), 259–272. https://doi.org/10.1046/j.1365-2443.2002.00520.x.

[145] Nureki, O.; Watanabe, K.; Fukai, S.; Ishii, R.; Endo, Y.; Hori, H.; Yokoyama, S. Deep Knot Structure for Construction of Active Site and Cofactor Binding Site of TRNA Modification Enzyme. *Struct. Lond. Engl. 1993*, **2004**, *12* (4), 593–602. https://doi.org/10.1016/j.str.2004.03.003.

[146] Purta, E.; van Vliet, F.; Tkaczuk, K. L.; Dunin-Horkawicz, S.; Mori, H.; Droogmans, L.; Bujnicki, J. M. The YfhQ Gene of Escherichia Coli Encodes a TRNA:Cm32/Um32 Methyltransferase. *BMC Mol. Biol.*, **2006**, *7*, 23. https://doi.org/10.1186/1471-2199-7-23.

[147] Somme, J.; Van Laer, B.; Roovers, M.; Steyaert, J.; Versées, W.; Droogmans, L. Characterization of Two Homologous 2'-O-Methyltransferases Showing Different Specificities for Their TRNA Substrates. *RNA N. Y. N*, **2014**, *20* (8), 1257–1271. https://doi.org/10.1261/rna.044503.114.

[148] Liu, R.-J.; Long, T.; Zhou, M.; Zhou, X.-L.; Wang, E.-D. TRNA Recognition by a Bacterial TRNA Xm32 Modification Enzyme from the SPOUT Methyltransferase Superfamily. *Nucleic Acids Res.*, **2015**, *43* (15), 7489–7503. https://doi.org/10.1093/nar/gkv745.

[149] Jaroensuk, J.; Atichartpongkul, S.; Chionh, Y. H.; Wong, Y. H.; Liew, C. W.; McBee, M. E.; Thongdee, N.; Prestwich, E. G.; DeMott, M. S.; Mongkolsuk, S.; et al. Methylation at Position 32 of TRNA Catalyzed by TrmJ Alters Oxidative Stress Response in Pseudomonas Aeruginosa. *Nucleic Acids Res.*, **2016**, *44* (22), 10834–10848. https://doi.org/10.1093/nar/gkw870.

[150] Benítez-Páez, A.; Villarroya, M.; Douthwaite, S.; Gabaldón, T.; Armengod, M.-E. YibK Is the 2'-O-Methyltransferase TrmL That Modifies the Wobble Nucleotide in Escherichia Coli TRNA(Leu) Isoacceptors. *RNA N. Y. N*, **2010**, *16* (11), 2131–2143. https://doi.org/10.1261/rna.2245910.

[151] Armengod, M.-E.; Moukadiri, I.; Prado, S.; Ruiz-Partida, R.; Benítez-Páez, A.; Villarroya, M.; Lomas, R.; Garzón, M. J.; Martínez-Zamora, A.; Meseguer, S.; et al. Enzymology of TRNA Modification in the Bacterial MnmEG Pathway. *Biochimie*, **2012**, *94* (7), 1510–1520. https://doi.org/10.1016/j.biochi.2012.02.019.

[152] Puri, P.; Wetzel, C.; Saffert, P.; Gaston, K. W.; Russell, S. P.; Cordero Varela, J. A.; van der Vlies, P.; Zhang, G.; Limbach, P. A.; Ignatova, Z.; et al. Systematic Identification of TRNAome and Its Dynamics in Lactococcus Lactis. *Mol. Microbiol.*, **2014**, *93* (5), 944–956. https://doi.org/10.1111/mmi.12710.

[153] Grosjean, H.; Gaspin, C.; Marck, C.; Decatur, W. A.; de Crécy-Lagard, V. RNomics and Modomics in the Halophilic Archaea Haloferax Volcanii: Identification of RNA Modification Genes. *BMC Genomics*, **2008**, *9*, 470. https://doi.org/10.1186/1471-2164-9-470.

[154] Clouet-d'Orval, B.; Gaspin, C.; Mougin, A. Two Different Mechanisms for TRNA Ribose Methylation in Archaea: A Short Survey. *Biochimie*, **2005**, *87* (9–10), 889–895. https://doi.org/10.1016/j.biochi.2005.02.004.

[155] Joardar, A.; Malliahgari, S. R.; Skariah, G.; Gupta, R. 2'-O-Methylation of the Wobble Residue of Elongator Pre-TRNA(Met) in Haloferax Volcanii Is Guided

by a Box C/D RNA Containing Unique Features. *RNA Biol.*, **2011**, *8* (5), 782–791. https://doi.org/10.4161/rna.8.5.16015.

[156] Singh, S. K.; Gurha, P.; Tran, E. J.; Maxwell, E. S.; Gupta, R. Sequential 2'-O-Methylation of Archaeal Pre-TRNATrp Nucleotides Is Guided by the Intron-Encoded but Trans-Acting Box C/D Ribonucleoprotein of Pre-TRNA. *J. Biol. Chem.*, **2004**, *279* (46), 47661–47671. https://doi.org/10.1074/jbc.M408868200.

[157] Cavaillé, J.; Chetouani, F.; Bachellerie, J. P. The Yeast Saccharomyces Cerevisiae YDL112w ORF Encodes the Putative 2'-O-Ribose Methyltransferase Catalyzing the Formation of Gm18 in TRNAs. *RNA N. Y. N*, **1999**, *5* (1), 66–81. https://doi.org/10.1017/s1355838299981475.

[158] Droogmans, L.; Haumont, E.; de Henau, S.; Grosjean, H. Enzymatic 2'-O-Methylation of the Wobble Nucleoside of Eukaryotic TRNAPhe: Specificity Depends on Structural Elements Outside the Anticodon Loop. *EMBO J.*, **1986**, *5* (5), 1105–1109. https://doi.org/10.1002/j.1460-2075.1986.tb04329.x.

[159] Pintard, L.; Lecointe, F.; Bujnicki, J. M.; Bonnerot, C.; Grosjean, H.; Lapeyre, B. Trm7p Catalyses the Formation of Two 2'-O-Methylriboses in Yeast TRNA Anticodon Loop. *EMBO J.*, **2002**, *21* (7), 1811–1820. https://doi.org/10.1093/emboj/21.7.1811.

[160] Guy, M. P.; Shaw, M.; Weiner, C. L.; Hobson, L.; Stark, Z.; Rose, K.; Kalscheuer, V. M.; Gecz, J.; Phizicky, E. M. Defects in TRNA Anticodon Loop 2'-O-Methylation Are Implicated in Nonsyndromic X-Linked Intellectual Disability Due to Mutations in FTSJ1. *Hum. Mutat.*, **2015**, *36* (12), 1176–1187. https://doi.org/10.1002/humu.22897.

[161] Tkaczuk, K. L. Trm13p, the TRNA:Xm4 Modification Enzyme from Saccharomyces Cerevisiae Is a Member of the Rossmann-Fold MTase Superfamily: Prediction of Structure and Active Site. *J. Mol. Model.*, **2010**, *16* (3), 599–606. https://doi.org/10.1007/s00894-009-0570-6.

[162] Towns, W. L.; Begley, T. J. Transfer RNA Methytransferases and Their Corresponding Modifications in Budding Yeast and Humans: Activities, Predications, and Potential Roles in Human Health. *DNA Cell Biol.*, **2012**, *31* (4), 434–454. https://doi.org/10.1089/dna.2011.1437.

[163] Mosquera-Rendón, J.; Cárdenas-Brito, S.; Pineda, J. D.; Corredor, M.; Benítez-Páez, A. Evolutionary and Sequence-Based Relationships in Bacterial AdoMet-Dependent Non-Coding RNA Methyltransferases. *BMC Res. Notes*, **2014**, *7*, 440. https://doi.org/10.1186/1756-0500-7-440.

[164] Punekar, A. S.; Shepherd, T. R.; Liljeruhm, J.; Forster, A. C.; Selmer, M. Crystal Structure of RlmM, the 2'O-Ribose Methyltransferase for C2498 of Escherichia Coli 23S RRNA. *Nucleic Acids Res.*, **2012**, *40* (20), 10507–10520. https://doi.org/10.1093/nar/gks727.

[165] Zhao, M.; Zhang, H.; Liu, G.; Wang, L.; Wang, J.; Gao, Z.; Dong, Y.; Zhang, L.; Gong, Y. Structural Insights into the Methylation of C1402 in 16S RRNA by Methyltransferase RsmI. *PloS One*, **2016**, *11* (10), e0163816. https://doi.org/10.1371/journal.pone.0163816.

[166] Michel, G.; Sauvé, V.; Larocque, R.; Li, Y.; Matte, A.; Cygler, M. The Structure of the RlmB 23S RRNA Methyltransferase Reveals a New Methyltransferase Fold with a Unique Knot. *Struct. Lond. Engl. 1993*, **2002**, *10* (10), 1303–1315. https://doi.org/10.1016/s0969-2126(02)00852-3.

[167] Hager, J.; Staker, B. L.; Jakob, U. Substrate Binding Analysis of the 23S RRNA Methyltransferase RrmJ. *J. Bacteriol.*, **2004**, *186* (19), 6634–6642. https://doi.org/10.1128/JB.186.19.6634-6642.2004.
[168] Tomkuvienė, M.; Ličytė, J.; Olendraitė, I.; Liutkevičiūtė, Z.; Clouet-d'Orval, B.; Klimašauskas, S. Archaeal Fibrillarin-Nop5 Heterodimer 2'-O-Methylates RNA Independently of the C/D Guide RNP Particle. *RNA N. Y. N*, **2017**, *23* (9), 1329–1337. https://doi.org/10.1261/rna.059832.116.
[169] Bonnerot, C.; Pintard, L.; Lutfalla, G. Functional Redundancy of Spb1p and a SnR52-Dependent Mechanism for the 2'-O-Ribose Methylation of a Conserved RRNA Position in Yeast. *Mol. Cell*, **2003**, *12* (5), 1309–1315. https://doi.org/10.1016/s1097-2765(03)00435-0.
[170] Lapeyre, B.; Purushothaman, S. K. Spb1p-Directed Formation of Gm2922 in the Ribosome Catalytic Center Occurs at a Late Processing Stage. *Mol. Cell*, **2004**, *16* (4), 663–669. https://doi.org/10.1016/j.molcel.2004.10.022.
[171] Morello, L. G.; Coltri, P. P.; Quaresma, A. J. C.; Simabuco, F. M.; Silva, T. C. L.; Singh, G.; Nickerson, J. A.; Oliveira, C. C.; Moore, M. J.; Zanchin, N. I. T. The Human Nucleolar Protein FTSJ3 Associates with NIP7 and Functions in Pre-RRNA Processing. *PloS One*, **2011**, *6* (12), e29174. https://doi.org/10.1371/journal.pone.0029174.
[172] Chou, H.-J.; Donnard, E.; Gustafsson, H. T.; Garber, M.; Rando, O. J. Transcriptome-Wide Analysis of Roles for TRNA Modifications in Translational Regulation. *Mol. Cell*, **2017**, *68* (5), 978-992.e4. https://doi.org/10.1016/j.molcel.2017.11.002.
[173] Pintard, L.; Bujnicki, J. M.; Lapeyre, B.; Bonnerot, C. MRM2 Encodes a Novel Yeast Mitochondrial 21S RRNA Methyltransferase. *EMBO J.*, **2002**, *21* (5), 1139–1147. https://doi.org/10.1093/emboj/21.5.1139.
[174] Rorbach, J.; Boesch, P.; Gammage, P. A.; Nicholls, T. J. J.; Pearce, S. F.; Patel, D.; Hauser, A.; Perocchi, F.; Minczuk, M. MRM2 and MRM3 Are Involved in Biogenesis of the Large Subunit of the Mitochondrial Ribosome. *Mol. Biol. Cell*, **2014**, *25* (17), 2542–2555. https://doi.org/10.1091/mbc.E14-01-0014.
[175] Garone, C.; D'Souza, A. R.; Dallabona, C.; Lodi, T.; Rebelo-Guiomar, P.; Rorbach, J.; Donati, M. A.; Procopio, E.; Montomoli, M.; Guerrini, R.; et al. Defective Mitochondrial RRNA Methyltransferase MRM2 Causes MELAS-like Clinical Syndrome. *Hum. Mol. Genet.*, **2017**, *26* (21), 4257–4266. https://doi.org/10.1093/hmg/ddx314.
[176] Yang, Z.; Ebright, Y. W.; Yu, B.; Chen, X. HEN1 Recognizes 21-24 Nt Small RNA Duplexes and Deposits a Methyl Group onto the 2' OH of the 3' Terminal Nucleotide. *Nucleic Acids Res.*, **2006**, *34* (2), 667–675. https://doi.org/10.1093/nar/gkj474.
[177] Yu, B.; Yang, Z.; Li, J.; Minakhina, S.; Yang, M.; Padgett, R. W.; Steward, R.; Chen, X. Methylation as a Crucial Step in Plant MicroRNA Biogenesis. *Science*, **2005**, *307* (5711), 932–935. https://doi.org/10.1126/science.1107130.
[178] Peng, L.; Zhang, F.; Shang, R.; Wang, X.; Chen, J.; Chou, J. J.; Ma, J.; Wu, L.; Huang, Y. Identification of Substrates of the Small RNA Methyltransferase Hen1 in Mouse Spermatogonial Stem Cells and Analysis of Its Methyl-Transfer Domain. *J. Biol. Chem.*, **2018**, *293* (26), 9981–9994. https://doi.org/10.1074/jbc.RA117.000837.
[179] Horwich, M. D.; Li, C.; Matranga, C.; Vagin, V.; Farley, G.; Wang, P.; Zamore, P. D. The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline

PiRNAs and Single-Stranded SiRNAs in RISC. *Curr. Biol. CB*, **2007**, *17* (14), 1265–1272. https://doi.org/10.1016/j.cub.2007.06.030.
[180] Li, J.; Yang, Z.; Yu, B.; Liu, J.; Chen, X. Methylation Protects MiRNAs and SiRNAs from a 3'-End Uridylation Activity in Arabidopsis. *Curr. Biol. CB*, **2005**, *15* (16), 1501–1507. https://doi.org/10.1016/j.cub.2005.07.029.
[181] Chan, C. M.; Zhou, C.; Huang, R. H. Reconstituting Bacterial RNA Repair and Modification in Vitro. *Science*, **2009**, *326* (5950), 247. https://doi.org/10.1126/science.1179480.
[182] Huang, R. H. Unique 2'-O-Methylation by Hen1 in Eukaryotic RNA Interference and Bacterial RNA Repair. *Biochemistry*, **2012**, *51* (20), 4087–4095. https://doi.org/10.1021/bi300497x.
[183] Henras, A. K.; Plisson-Chastang, C.; Humbert, O.; Romeo, Y.; Henry, Y. Synthesis, Function, and Heterogeneity of SnoRNA-Guided Posttranscriptional Nucleoside Modifications in Eukaryotic Ribosomal RNAs. *The Enzymes*, **2017**, *41*, 169–213. https://doi.org/10.1016/bs.enz.2017.03.007.
[184] Lestrade, L.; Weber, M. J. SnoRNA-LBME-Db, a Comprehensive Database of Human H/ACA and C/D Box SnoRNAs. *Nucleic Acids Res.*, **2006**, *34* (Database issue), D158-162. https://doi.org/10.1093/nar/gkj002.
[185] Lowe, T. M.; Eddy, S. R. A Computational Screen for Methylation Guide SnoRNAs in Yeast. *Science*, **1999**, *283* (5405), 1168–1171. https://doi.org/10.1126/science.283.5405.1168.
[186] Lui, L.; Lowe, T. Small Nucleolar RNAs and RNA-Guided Post-Transcriptional Modification. *Essays Biochem.*, **2013**, *54*, 53–77. https://doi.org/10.1042/bse0540053.
[187] Omer, A. D.; Ziesche, S.; Ebhardt, H.; Dennis, P. P. In Vitro Reconstitution and Activity of a C/D Box Methylation Guide Ribonucleoprotein Complex. *Proc. Natl. Acad. Sci. U. S. A.*, **2002**, *99* (8), 5289–5294. https://doi.org/10.1073/pnas.082101999.
[188] Kuhn, J. F.; Tran, E. J.; Maxwell, E. S. Archaeal Ribosomal Protein L7 Is a Functional Homolog of the Eukaryotic 15.5kD/Snu13p SnoRNP Core Protein. *Nucleic Acids Res.*, **2002**, *30* (4), 931–941. https://doi.org/10.1093/nar/30.4.931.
[189] Aittaleb, M.; Rashid, R.; Chen, Q.; Palmer, J. R.; Daniels, C. J.; Li, H. Structure and Function of Archaeal Box C/D SRNP Core Proteins. *Nat. Struct. Biol.*, **2003**, *10* (4), 256–263. https://doi.org/10.1038/nsb905.
[190] Kiss-László, Z.; Henry, Y.; Kiss, T. Sequence and Structural Elements of Methylation Guide SnoRNAs Essential for Site-Specific Ribose Methylation of Pre-RRNA. *EMBO J.*, **1998**, *17* (3), 797–807. https://doi.org/10.1093/emboj/17.3.797.
[191] Reichow, S. L.; Hamma, T.; Ferré-D'Amaré, A. R.; Varani, G. The Structure and Function of Small Nucleolar Ribonucleoproteins. *Nucleic Acids Res.*, **2007**, *35* (5), 1452–1464. https://doi.org/10.1093/nar/gkl1172.
[192] Graziadei, A.; Masiewicz, P.; Lapinaite, A.; Carlomagno, T. Archaea Box C/D Enzymes Methylate Two Distinct Substrate RRNA Sequences with Different Efficiency. *RNA N. Y. N*, **2016**, *22* (5), 764–772. https://doi.org/10.1261/rna.054320.115.
[193] Randau, L. RNA Processing in the Minimal Organism Nanoarchaeum Equitans. *Genome Biol.*, **2012**, *13* (7), R63. https://doi.org/10.1186/gb-2012-13-7-r63.
[194] Su, A. A. H.; Tripp, V.; Randau, L. RNA-Seq Analyses Reveal the Order of TRNA Processing Events and the Maturation of C/D Box and CRISPR RNAs in the

Hyperthermophile Methanopyrus Kandleri. *Nucleic Acids Res.*, **2013**, *41* (12), 6250–6258. https://doi.org/10.1093/nar/gkt317.

[195] Jorjani, H.; Kehr, S.; Jedlinski, D. J.; Gumienny, R.; Hertel, J.; Stadler, P. F.; Zavolan, M.; Gruber, A. R. An Updated Human SnoRNAome. *Nucleic Acids Res.*, **2016**, *44* (11), 5068–5082. https://doi.org/10.1093/nar/gkw386.

[196] Randerath, E.; Gopalakrishnan, A. S.; Gupta, R. C.; Agrawal, H. P.; Randerath, K. Lack of a Specific Ribose Methylation at Guanosine 17 in Morris Hepatoma 5123D TRNASer1IGA. *Cancer Res.*, **1981**, *41* (7), 2863–2867.

[197] Chan, C. T. Y.; Dyavaiah, M.; DeMott, M. S.; Taghizadeh, K.; Dedon, P. C.; Begley, T. J. A Quantitative Systems Approach Reveals Dynamic Control of TRNA Modifications during Cellular Stress. *PLoS Genet.*, **2010**, *6* (12), e1001247. https://doi.org/10.1371/journal.pgen.1001247.

[198] Chionh, Y. H.; McBee, M.; Babu, I. R.; Hia, F.; Lin, W.; Zhao, W.; Cao, J.; Dziergowska, A.; Malkiewicz, A.; Begley, T. J.; et al. TRNA-Mediated Codon-Biased Translation in Mycobacterial Hypoxic Persistence. *Nat. Commun.*, **2016**, *7*, 13302. https://doi.org/10.1038/ncomms13302.

[199] Endres, L.; Dedon, P. C.; Begley, T. J. Codon-Biased Translation Can Be Regulated by Wobble-Base TRNA Modification Systems during Cellular Stress Responses. *RNA Biol.*, **2015**, *12* (6), 603–614. https://doi.org/10.1080/15476286.2015.1031947.

[200] Buchhaupt, M.; Sharma, S.; Kellner, S.; Oswald, S.; Paetzold, M.; Peifer, C.; Watzinger, P.; Schrader, J.; Helm, M.; Entian, K.-D. Partial Methylation at Am100 in 18S RRNA of Baker's Yeast Reveals Ribosome Heterogeneity on the Level of Eukaryotic RRNA Modification. *PloS One*, **2014**, *9* (2), e89640. https://doi.org/10.1371/journal.pone.0089640.

[201] Poole, A. R.; Vicino, I.; Adachi, H.; Yu, Y.-T.; Hebert, M. D. Regulatory RNPs: A Novel Class of Ribonucleoproteins That Potentially Contribute to Ribosome Heterogeneity. *Biol. Open*, **2017**, *6* (9), 1342–1354. https://doi.org/10.1242/bio.028092.

[202] Tycowski, K. T.; Aab, A.; Steitz, J. A. Guide RNAs with 5' Caps and Novel Box C/D SnoRNA-like Domains for Modification of SnRNAs in Metazoa. *Curr. Biol. CB*, **2004**, *14* (22), 1985–1995. https://doi.org/10.1016/j.cub.2004.11.003.

[203] Krogh, N.; Birkedal, U.; Nielsen, H. RiboMeth-Seq: Profiling of 2'-O-Me in RNA. *Methods Mol. Biol. Clifton NJ*, **2017**, *1562*, 189–209. https://doi.org/10.1007/978-1-4939-6807-7_13.

[204] Kawai, G.; Ue, H.; Yasuda, M.; Sakamoto, K.; Hashizume, T.; McCloskey, J. A.; Miyazawa, T.; Yokoyama, S. Relation between Functions and Conformational Characteristics of Modified Nucleosides Found in TRNAs. *Nucleic Acids Symp. Ser.*, **1991**, No. 25, 49–50.

[205] Trim, A. R.; Parker, J. E. Preparation, Purification and Analyses of Thirteen Alkali-Stable Dinucleotides from Yeast Ribonucleic Acid. *Biochem. J.*, **1970**, *116* (4), 589–598. https://doi.org/10.1042/bj1160589.

[206] Bardwell, V. J.; Wickens, M.; Bienroth, S.; Keller, W.; Sproat, B. S.; Lamond, A. I. Site-Directed Ribose Methylation Identifies 2'-OH Groups in Polyadenylation Substrates Critical for AAUAAA Recognition and Poly(A) Addition. *Cell*, **1991**, *65* (1), 125–133. https://doi.org/10.1016/0092-8674(91)90414-t.

[207] Mishler, D. M.; Christ, A. B.; Steitz, J. A. Flexibility in the Site of Exon Junction Complex Deposition Revealed by Functional Group and RNA Secondary

Structure Alterations in the Splicing Substrate. *RNA N. Y. N*, **2008**, *14* (12), 2657–2670. https://doi.org/10.1261/rna.1312808.

[208] Andersen, C. B. F.; Ballut, L.; Johansen, J. S.; Chamieh, H.; Nielsen, K. H.; Oliveira, C. L. P.; Pedersen, J. S.; Séraphin, B.; Le Hir, H.; Andersen, G. R. Structure of the Exon Junction Core Complex with a Trapped DEAD-Box ATPase Bound to RNA. *Science*, **2006**, *313* (5795), 1968–1972. https://doi.org/10.1126/science.1131981.

[209] Iribe, H.; Miyamoto, K.; Takahashi, T.; Kobayashi, Y.; Leo, J.; Aida, M.; Ui-Tei, K. Chemical Modification of the SiRNA Seed Region Suppresses Off-Target Effects by Steric Hindrance to Base-Pairing with Targets. *ACS Omega*, **2017**, *2* (5), 2055–2064. https://doi.org/10.1021/acsomega.7b00291.

[210] Kowalak, J. A.; Dalluge, J. J.; McCloskey, J. A.; Stetter, K. O. The Role of Posttranscriptional Modification in Stabilization of Transfer RNA from Hyperthermophiles. *Biochemistry*, **1994**, *33* (25), 7869–7876. https://doi.org/10.1021/bi00191a014.

[211] Agris, P. F.; Koh, H.; Söll, D. The Effect of Growth Temperatures on the in Vivo Ribose Methylation of Bacillus Stearothermophilus Transfer RNA. *Arch. Biochem. Biophys.*, **1973**, *154* (1), 277–282. https://doi.org/10.1016/0003-9861(73)90058-1.

[212] Kirpekar, F.; Hansen, L. H.; Rasmussen, A.; Poehlsgaard, J.; Vester, B. The Archaeon Haloarcula Marismortui Has Few Modifications in the Central Parts of Its 23S Ribosomal RNA. *J. Mol. Biol.*, **2005**, *348* (3), 563–573. https://doi.org/10.1016/j.jmb.2005.03.009.

[213] Noon, K. R.; Bruenger, E.; McCloskey, J. A. Posttranscriptional Modifications in 16S and 23S RRNAs of the Archaeal Hyperthermophile Sulfolobus Solfataricus. *J. Bacteriol.*, **1998**, *180* (11), 2883–2888. https://doi.org/10.1128/JB.180.11.2883-2888.1998.

[214] Abe, M.; Naqvi, A.; Hendriks, G.-J.; Feltzin, V.; Zhu, Y.; Grigoriev, A.; Bonini, N. M. Impact of Age-Associated Increase in 2'-O-Methylation of MiRNAs on Aging and Neurodegeneration in Drosophila. *Genes Dev.*, **2014**, *28* (1), 44–57. https://doi.org/10.1101/gad.226654.113.

[215] Vázquez, D. Inhibitors of Protein Biosynthesis. *Mol. Biol. Biochem. Biophys.*, **1979**, *30*, i–x, 1–312. https://doi.org/10.1007/978-3-642-81309-2.

[216] La Teana, A.; Gualerzi, C. O.; Dahlberg, A. E. Initiation Factor IF 2 Binds to the Alpha-Sarcin Loop and Helix 89 of Escherichia Coli 23S Ribosomal RNA. *RNA N. Y. N*, **2001**, *7* (8), 1173–1179. https://doi.org/10.1017/s1355838201010366.

[217] Douthwaite, S. Interaction of the Antibiotics Clindamycin and Lincomycin with Escherichia Coli 23S Ribosomal RNA. *Nucleic Acids Res.*, **1992**, *20* (18), 4717–4720. https://doi.org/10.1093/nar/20.18.4717.

[218] Poulsen, S. M.; Karlsson, M.; Johansson, L. B.; Vester, B. The Pleuromutilin Drugs Tiamulin and Valnemulin Bind to the RNA at the Peptidyl Transferase Centre on the Ribosome. *Mol. Microbiol.*, **2001**, *41* (5), 1091–1099. https://doi.org/10.1046/j.1365-2958.2001.02595.x.

[219] Bechthold, A.; Floss, H. G. Overexpression of the Thiostrepton-Resistance Gene from Streptomyces Azureus in Escherichia Coli and Characterization of Recognition Sites of the 23S RRNA A1067 2'-Methyltransferase in the Guanosine Triphosphatase Center of 23S Ribosomal RNA. *Eur. J. Biochem.*, **1994**, *224* (2), 431–437. https://doi.org/10.1111/j.1432-1033.1994.00431.x.

[220] Treede, I.; Jakobsen, L.; Kirpekar, F.; Vester, B.; Weitnauer, G.; Bechthold, A.; Douthwaite, S. The Avilamycin Resistance Determinants AviRa and AviRb Methylate 23S RRNA at the Guanosine 2535 Base and the Uridine 2479 Ribose. *Mol. Microbiol.*, **2003**, *49* (2), 309–318. https://doi.org/10.1046/j.1365-2958.2003.03558.x.

[221] Sofia, H. J.; Chen, G.; Hetzler, B. G.; Reyes-Spindola, J. F.; Miller, N. E. Radical SAM, a Novel Protein Superfamily Linking Unresolved Steps in Familiar Biosynthetic Pathways with Radical Mechanisms: Functional Characterization Using New Analysis and Information Visualization Methods. *Nucleic Acids Res.*, **2001**, *29* (5), 1097–1106. https://doi.org/10.1093/nar/29.5.1097.

[222] Kehrenberg, C.; Schwarz, S.; Jacobsen, L.; Hansen, L. H.; Vester, B. A New Mechanism for Chloramphenicol, Florfenicol and Clindamycin Resistance: Methylation of 23S Ribosomal RNA at A2503. *Mol. Microbiol.*, **2005**, *57* (4), 1064–1073. https://doi.org/10.1111/j.1365-2958.2005.04754.x.

[223] Esguerra, J.; Warringer, J.; Blomberg, A. Functional Importance of Individual RRNA 2'-O-Ribose Methylations Revealed by High-Resolution Phenotyping. *RNA N. Y. N*, **2008**, *14* (4), 649–656. https://doi.org/10.1261/rna.845808.

[224] Dönmez, G.; Hartmuth, K.; Lührmann, R. Modified Nucleotides at the 5' End of Human U2 SnRNA Are Required for Spliceosomal E-Complex Formation. *RNA N. Y. N*, **2004**, *10* (12), 1925–1933. https://doi.org/10.1261/rna.7186504.

[225] Ge, J.; Liu, H.; Yu, Y.-T. Regulation of Pre-MRNA Splicing in Xenopus Oocytes by Targeted 2'-O-Methylation. *RNA N. Y. N*, **2010**, *16* (5), 1078–1085. https://doi.org/10.1261/rna.2060210.

[226] Dong, H.; Fink, K.; Züst, R.; Lim, S. P.; Qin, C.-F.; Shi, P.-Y. Flavivirus RNA Methylation. *J. Gen. Virol.*, **2014**, *95* (Pt 4), 763–778. https://doi.org/10.1099/vir.0.062208-0.

[227] Bradrick, S. S. Causes and Consequences of Flavivirus RNA Methylation. *Front. Microbiol.*, **2017**, *8*, 2374. https://doi.org/10.3389/fmicb.2017.02374.

[228] Chang, D. C.; Hoang, L. T.; Mohamed Naim, A. N.; Dong, H.; Schreiber, M. J.; Hibberd, M. L.; Tan, M. J. A.; Shi, P.-Y. Evasion of Early Innate Immune Response by 2'-O-Methylation of Dengue Genomic RNA. *Virology*, **2016**, *499*, 259–266. https://doi.org/10.1016/j.virol.2016.09.022.

[229] Abbas, Y. M.; Laudenbach, B. T.; Martínez-Montero, S.; Cencic, R.; Habjan, M.; Pichlmair, A.; Damha, M. J.; Pelletier, J.; Nagar, B. Structure of Human IFIT1 with Capped RNA Reveals Adaptable MRNA Binding and Mechanisms for Sensing N1 and N2 Ribose 2'-O Methylations. *Proc. Natl. Acad. Sci. U. S. A.*, **2017**, *114* (11), E2106–E2115. https://doi.org/10.1073/pnas.1612444114.

[230] Züst, R.; Cervantes-Barragan, L.; Habjan, M.; Maier, R.; Neuman, B. W.; Ziebuhr, J.; Szretter, K. J.; Baker, S. C.; Barchet, W.; Diamond, M. S.; et al. Ribose 2'-O-Methylation Provides a Molecular Signature for the Distinction of Self and Non-Self MRNA Dependent on the RNA Sensor Mda5. *Nat. Immunol.*, **2011**, *12* (2), 137–143. https://doi.org/10.1038/ni.1979.

[231] Hoernes, T. P.; Clementi, N.; Faserl, K.; Glasner, H.; Breuker, K.; Lindner, H.; Hüttenhofer, A.; Erlacher, M. D. Nucleotide Modifications within Bacterial Messenger RNAs Regulate Their Translation and Are Able to Rewire the Genetic Code. *Nucleic Acids Res.*, **2016**, *44* (2), 852–862. https://doi.org/10.1093/nar/gkv1182.

[232] Choi, J.; Indrisiunaite, G.; DeMirci, H.; Ieong, K.-W.; Wang, J.; Petrov, A.; Prabhakar, A.; Rechavi, G.; Dominissini, D.; He, C.; et al. 2'-O-Methylation in

MRNA Disrupts TRNA Decoding during Translation Elongation. *Nat. Struct. Mol. Biol.*, **2018**, *25* (3), 208–216. https://doi.org/10.1038/s41594-018-0030-z.

[233] Jung, S.; von Thülen, T.; Laukemper, V.; Pigisch, S.; Hangel, D.; Wagner, H.; Kaufmann, A.; Bauer, S. A Single Naturally Occurring 2'-O-Methylation Converts a TLR7- and TLR8-Activating RNA into a TLR8-Specific Ligand. *PloS One*, **2015**, *10* (3), e0120498. https://doi.org/10.1371/journal.pone.0120498.

[234] Judge, A. D.; Bola, G.; Lee, A. C. H.; MacLachlan, I. Design of Noninflammatory Synthetic SiRNA Mediating Potent Gene Silencing in Vivo. *Mol. Ther.*, **2006**, *13* (3), 494–505. https://doi.org/10.1016/j.ymthe.2005.11.002.

[235] Gehrig, S.; Eberle, M.-E.; Botschen, F.; Rimbach, K.; Eberle, F.; Eigenbrod, T.; Kaiser, S.; Holmes, W. M.; Erdmann, V. A.; Sprinzl, M.; et al. Identification of Modifications in Microbial, Native TRNA That Suppress Immunostimulatory Activity. *J. Exp. Med.*, **2012**, *209* (2), 225–233. https://doi.org/10.1084/jem.20111044.

[236] Rimbach, K.; Kaiser, S.; Helm, M.; Dalpke, A. H.; Eigenbrod, T. 2'-O-Methylation within Bacterial RNA Acts as Suppressor of TLR7/TLR8 Activation in Human Innate Immune Cells. *J. Innate Immun.*, **2015**, *7* (5), 482–493. https://doi.org/10.1159/000375460.

[237] Robbins, M.; Judge, A.; Liang, L.; McClintock, K.; Yaworski, E.; MacLachlan, I. 2'-O-Methyl-Modified RNAs Act as TLR7 Antagonists. *Mol. Ther. J. Am. Soc. Gene Ther.*, **2007**, *15* (9), 1663–1669. https://doi.org/10.1038/sj.mt.6300240.

[238] Kaiser, S.; Rimbach, K.; Eigenbrod, T.; Dalpke, A. H.; Helm, M. A Modified Dinucleotide Motif Specifies TRNA Recognition by TLR7. *RNA N. Y. N*, **2014**, *20* (9), 1351–1355. https://doi.org/10.1261/rna.044024.113.

[239] Zhou, F.; Liu, Y.; Rohde, C.; Pauli, C.; Gerloff, D.; Köhn, M.; Misiak, D.; Bäumer, N.; Cui, C.; Göllner, S.; et al. AML1-ETO Requires Enhanced C/D Box SnoRNA/RNP Formation to Induce Self-Renewal and Leukaemia. *Nat. Cell Biol.*, **2017**, *19* (7), 844–855. https://doi.org/10.1038/ncb3563.

[240] Kierzek, E.; Malgowska, M.; Lisowiec, J.; Turner, D. H.; Gdaniec, Z.; Kierzek, R. The Contribution of Pseudouridine to Stabilities and Structure of RNAs. *Nucleic Acids Res.*, **2014**, *42* (5), 3492–3501. https://doi.org/10.1093/nar/gkt1330.

[241] Li, X.; Ma, S.; Yi, C. Pseudouridine: The Fifth RNA Nucleotide with Renewed Interests. *Curr. Opin. Chem. Biol.*, **2016**, *33*, 108–116. https://doi.org/10.1016/j.cbpa.2016.06.014.

[242] Rintala-Dempsey, A. C.; Kothe, U. Eukaryotic Stand-Alone Pseudouridine Synthases - RNA Modifying Enzymes and Emerging Regulators of Gene Expression? *RNA Biol.*, **2017**, *14* (9), 1185–1196. https://doi.org/10.1080/15476286.2016.1276150.

[243] Cerneckis, J.; Cui, Q.; He, C.; Yi, C.; Shi, Y. Decoding Pseudouridine: An Emerging Target for Therapeutic Development. *Trends Pharmacol. Sci.*, **2022**, *43* (6), 522–535. https://doi.org/10.1016/j.tips.2022.03.008.

[244] Corbett, K. S.; Edwards, D. K.; Leist, S. R.; Abiona, O. M.; Boyoglu-Barnum, S.; Gillespie, R. A.; Himansu, S.; Schäfer, A.; Ziwawo, C. T.; DiPiazza, A. T.; et al. SARS-CoV-2 MRNA Vaccine Design Enabled by Prototype Pathogen Preparedness. *Nature*, **2020**, *586* (7830), 567–571. https://doi.org/10.1038/s41586-020-2622-0.

[245] Nishikura, K. A-to-I Editing of Coding and Non-Coding RNAs by ADARs. *Nat. Rev. Mol. Cell Biol.*, **2016**, *17* (2), 83–96. https://doi.org/10.1038/nrm.2015.4.

[246] Zhang, Y.; Geng, X.; Li, Q.; Xu, J.; Tan, Y.; Xiao, M.; Song, J.; Liu, F.; Fang, C.; Wang, H. M6A Modification in RNA: Biogenesis, Functions and Roles in Gliomas. *J. Exp. Clin. Cancer Res. CR*, **2020**, *39* (1), 192. https://doi.org/10.1186/s13046-020-01706-8.
[247] Sikorski, V.; Vento, A.; Kankuri, E.; IHD-EPITRAN Consortium. Emerging Roles of the RNA Modifications N6-Methyladenosine and Adenosine-to-Inosine in Cardiovascular Diseases. *Mol. Ther. Nucleic Acids*, **2022**, *29*, 426–461. https://doi.org/10.1016/j.omtn.2022.07.018.
[248] Roundtree, I. A.; Evans, M. E.; Pan, T.; He, C. Dynamic RNA Modifications in Gene Expression Regulation. *Cell*, **2017**, *169* (7), 1187–1200. https://doi.org/10.1016/j.cell.2017.05.045.
[249] An, Y.; Duan, H. The Role of M6A RNA Methylation in Cancer Metabolism. *Mol. Cancer*, **2022**, *21* (1), 14. https://doi.org/10.1186/s12943-022-01500-4.
[250] Sun, T.; Wu, R.; Ming, L. The Role of M6A RNA Methylation in Cancer. *Biomed. Pharmacother. Biomedecine Pharmacother.*, **2019**, *112*, 108613. https://doi.org/10.1016/j.biopha.2019.108613.
[251] Ma, S.; Chen, C.; Ji, X.; Liu, J.; Zhou, Q.; Wang, G.; Yuan, W.; Kan, Q.; Sun, Z. The Interplay between M6A RNA Methylation and Noncoding RNA in Cancer. *J. Hematol. Oncol.J Hematol Oncol*, **2019**, *12* (1), 121. https://doi.org/10.1186/s13045-019-0805-7.
[252] Lu, M.; Zhang, Z.; Xue, M.; Zhao, B. S.; Harder, O.; Li, A.; Liang, X.; Gao, T. Z.; Xu, Y.; Zhou, J.; et al. N6-Methyladenosine Modification Enables Viral RNA to Escape Recognition by RNA Sensor RIG-I. *Nat. Microbiol.*, **2020**, *5* (4), 584–598. https://doi.org/10.1038/s41564-019-0653-9.
[253] Hull, C. M.; Bevilacqua, P. C. Discriminating Self and Non-Self by RNA: Roles for RNA Structure, Misfolding, and Modification in Regulating the Innate Immune Sensor PKR. *Acc. Chem. Res.*, **2016**, *49* (6), 1242–1249. https://doi.org/10.1021/acs.accounts.6b00151.
[254] Eigenbrod, T.; Dalpke, A. H. Bacterial RNA: An Underestimated Stimulus for Innate Immune Responses. *J. Immunol. Baltim. Md 1950*, **2015**, *195* (2), 411–418. https://doi.org/10.4049/jimmunol.1500530.
[255] Ringeard, M.; Marchand, V.; Decroly, E.; Motorin, Y.; Bennasser, Y. FTSJ3 Is an RNA 2'-O-Methyltransferase Recruited by HIV to Avoid Innate Immune Sensing. *Nature*, **2019**, *565* (7740), 500–504. https://doi.org/10.1038/s41586-018-0841-4.
[256] El Kazzi, P.; Rabah, N.; Chamontin, C.; Poulain, L.; Ferron, F.; Debart, F.; Canard, B.; Missé, D.; Coutard, B.; Nisole, S.; et al. Internal RNA 2'O-Methylation in the HIV-1 Genome Counteracts ISG20 Nuclease-Mediated Antiviral Effect. *Nucleic Acids Res.*, **2023**, *51* (6), 2501–2515. https://doi.org/10.1093/nar/gkac996.
[257] Teng, P.-C.; Liang, Y.; Yarmishyn, A. A.; Hsiao, Y.-J.; Lin, T.-Y.; Lin, T.-W.; Teng, Y.-C.; Yang, Y.-P.; Wang, M.-L.; Chien, C.-S.; et al. RNA Modifications and Epigenetics in Modulation of Lung Cancer and Pulmonary Diseases. *Int. J. Mol. Sci.*, **2021**, *22* (19), 10592. https://doi.org/10.3390/ijms221910592.
[258] Haruehanroengra, P.; Zheng, Y. Y.; Zhou, Y.; Huang, Y.; Sheng, J. RNA Modifications and Cancer. *RNA Biol.*, **2020**, *17* (11), 1560–1575. https://doi.org/10.1080/15476286.2020.1722449.
[259] Markham, R.; Smith, J. D. Chromatographic Studies of Nucleic Acids; a Technique for the Identification and Estimation of Purine and Pyrimidine Bases,

Nucleosides and Related Substances. *Biochem. J.*, **1949**, *45* (3), 294–298. https://doi.org/10.1042/bj0450294.

[260] Behm-Ansmant, I.; Helm, M.; Motorin, Y. Use of Specific Chemical Reagents for Detection of Modified Nucleotides in RNA. *J. Nucleic Acids*, **2011**, *2011*, 408053. https://doi.org/10.4061/2011/408053.

[261] Ravelet, C.; Grosset, C.; Peyrin, E. Liquid Chromatography, Electrochromatography and Capillary Electrophoresis Applications of DNA and RNA Aptamers. *J. Chromatogr. A*, **2006**, *1117* (1), 1–10. https://doi.org/10.1016/j.chroma.2006.03.101.

[262] Liu, N.; Parisien, M.; Dai, Q.; Zheng, G.; He, C.; Pan, T. Probing N6-Methyladenosine RNA Modification Status at Single Nucleotide Resolution in MRNA and Long Noncoding RNA. *RNA N. Y. N*, **2013**, *19* (12), 1848–1856. https://doi.org/10.1261/rna.041178.113.

[263] Meng, Z.; Limbach, P. A. Mass Spectrometry of RNA: Linking the Genome to the Proteome. *Brief. Funct. Genomic. Proteomic.*, **2006**, *5* (1), 87–95. https://doi.org/10.1093/bfgp/ell012.

[264] Thüring, K.; Schmid, K.; Keller, P.; Helm, M. LC-MS Analysis of Methylated RNA. *Methods Mol. Biol. Clifton NJ*, **2017**, *1562*, 3–18. https://doi.org/10.1007/978-1-4939-6807-7_1.

[265] Dal Magro, C.; Keller, P.; Kotter, A.; Werner, S.; Duarte, V.; Marchand, V.; Ignarski, M.; Freiwald, A.; Müller, R.-U.; Dieterich, C.; et al. A Vastly Increased Chemical Variety of RNA Modifications Containing a Thioacetal Structure. *Angew. Chem. Int. Ed Engl.*, **2018**, *57* (26), 7893–7897. https://doi.org/10.1002/anie.201713188.

[266] Li, P.; Wang, Y.; Sun, Y.; Jiang, S.; Li, J. N 6-Methyladenosine RNA Methylation: From Regulatory Mechanisms to Potential Clinical Applications. *Front. Cell Dev. Biol.*, **2022**, *10*, 1055808. https://doi.org/10.3389/fcell.2022.1055808.

[267] Cock, P. J. A.; Fields, C. J.; Goto, N.; Heuer, M. L.; Rice, P. M. The Sanger FASTQ File Format for Sequences with Quality Scores, and the Solexa/Illumina FASTQ Variants. *Nucleic Acids Res.*, **2010**, *38* (6), 1767–1771. https://doi.org/10.1093/nar/gkp1137.

[268] Werner, S.; Schmidt, L.; Marchand, V.; Kemmer, T.; Falschlunger, C.; Sednev, M. V.; Bec, G.; Ennifar, E.; Höbartner, C.; Micura, R.; et al. Machine Learning of Reverse Transcription Signatures of Variegated Polymerases Allows Mapping and Discrimination of Methylated Purines in Limited Transcriptomes. *Nucleic Acids Res.*, **2020**, *48* (7), 3734–3746. https://doi.org/10.1093/nar/gkaa113.

[269] Carlile, T. M.; Rojas-Duran, M. F.; Gilbert, W. V. Pseudo-Seq: Genome-Wide Detection of Pseudouridine Modifications in RNA. *Methods Enzymol.*, **2015**, *560*, 219–245. https://doi.org/10.1016/bs.mie.2015.03.011.

[270] Dominissini, D.; Moshitch-Moshkovitz, S.; Schwartz, S.; Salmon-Divon, M.; Ungar, L.; Osenberg, S.; Cesarkas, K.; Jacob-Hirsch, J.; Amariglio, N.; Kupiec, M.; et al. Topology of the Human and Mouse M6A RNA Methylomes Revealed by M6A-Seq. *Nature*, **2012**, *485* (7397), 201–206. https://doi.org/10.1038/nature11112.

[271] Linder, B.; Grozhik, A. V.; Olarerin-George, A. O.; Meydan, C.; Mason, C. E.; Jaffrey, S. R. Single-Nucleotide-Resolution Mapping of M6A and M6Am throughout the Transcriptome. *Nat. Methods*, **2015**, *12* (8), 767–772. https://doi.org/10.1038/nmeth.3453.

[272] McIntyre, A. B. R.; Gokhale, N. S.; Cerchietti, L.; Jaffrey, S. R.; Horner, S. M.; Mason, C. E. Limits in the Detection of M6A Changes Using MeRIP/M6A-Seq. *Sci. Rep.*, **2020**, *10* (1), 6590. https://doi.org/10.1038/s41598-020-63355-3.
[273] Schaefer, M. RNA 5-Methylcytosine Analysis by Bisulfite Sequencing. *Methods Enzymol.*, **2015**, *560*, 297–329. https://doi.org/10.1016/bs.mie.2015.03.007.
[274] Marchand, V.; Ayadi, L.; Ernst, F. G. M.; Hertler, J.; Bourguignon-Igel, V.; Galvanin, A.; Kotter, A.; Helm, M.; Lafontaine, D. L. J.; Motorin, Y. AlkAniline-Seq: Profiling of M7 G and M3 C RNA Modifications at Single Nucleotide Resolution. *Angew. Chem. Int. Ed Engl.*, **2018**, *57* (51), 16785–16790. https://doi.org/10.1002/anie.201810946.
[275] Motorin, Y.; Helm, M. Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies. *Genes*, **2019**, *10* (1), 35. https://doi.org/10.3390/genes10010035.
[276] Katanski, C. D.; Watkins, C. P.; Zhang, W.; Reyer, M.; Miller, S.; Pan, T. Analysis of Queuosine and 2-Thio TRNA Modifications by High Throughput Sequencing. *Nucleic Acids Res.*, **2022**, *50* (17), e99. https://doi.org/10.1093/nar/gkac517.
[277] Wang, Y.; Zhao, Y.; Bollas, A.; Wang, Y.; Au, K. F. Nanopore Sequencing Technology, Bioinformatics and Applications. *Nat. Biotechnol.*, **2021**, *39* (11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x.
[278] Liu, H.; Begik, O.; Lucas, M. C.; Ramirez, J. M.; Mason, C. E.; Wiener, D.; Schwartz, S.; Mattick, J. S.; Smith, M. A.; Novoa, E. M. Accurate Detection of M6A RNA Modifications in Native RNA Sequences. *Nat. Commun.*, **2019**, *10* (1), 4079. https://doi.org/10.1038/s41467-019-11713-9.
[279] Lucas, M. C.; Pryszcz, L. P.; Medina, R.; Milenkovic, I.; Camacho, N.; Marchand, V.; Motorin, Y.; Ribas de Pouplana, L.; Novoa, E. M. Quantitative Analysis of TRNA Abundance and Modifications by Nanopore RNA Sequencing. *Nat. Biotechnol.*, **2023**. https://doi.org/10.1038/s41587-023-01743-6.
[280] Fan, W.; Eklund, E.; Sherman, R. M.; Liu, H.; Pitts, S.; Ford, B.; Rajeshkumar, N. V.; Laiho, M. Widespread Genetic Heterogeneity of Human Ribosomal RNA Genes. *RNA N. Y. N*, **2022**, *28* (4), 478–492. https://doi.org/10.1261/rna.078925.121.
[281] Hoffmann, A.; Fallmann, J.; Vilardo, E.; Mörl, M.; Stadler, P. F.; Amman, F. Accurate Mapping of TRNA Reads. *Bioinforma. Oxf. Engl.*, **2018**, *34* (7), 1116–1124. https://doi.org/10.1093/bioinformatics/btx756.
[282] Bakin, A.; Ofengand, J. Four Newly Located Pseudouridylate Residues in Escherichia Coli 23S Ribosomal RNA Are All at the Peptidyltransferase Center: Analysis by the Application of a New Sequencing Technique. *Biochemistry*, **1993**, *32* (37), 9754–9762.
[283] Grozhik, A. V.; Olarerin-George, A. O.; Sindelar, M.; Li, X.; Gross, S. S.; Jaffrey, S. R. Antibody Cross-Reactivity Accounts for Widespread Appearance of M1A in 5'UTRs. *Nat. Commun.*, **2019**, *10* (1), 5126. https://doi.org/10.1038/s41467-019-13146-w.
[284] Cheng, X.; Roberts, R. J. AdoMet-Dependent Methylation, DNA Methyltransferases and Base Flipping. *Nucleic Acids Res.*, **2001**, *29* (18), 3784–3795. https://doi.org/10.1093/nar/29.18.3784.
[285] Motorin, Y.; Seidu-Larry, S.; Helm, M. DNA and RNA Pyrimidine Nucleobase Alkylation at the Carbon-5 Position. *Adv. Exp. Med. Biol.*, **2016**, *945*, 19–33. https://doi.org/10.1007/978-3-319-43624-1_2.

[286] Rana, A. K.; Ankri, S. Reviving the RNA World: An Insight into the Appearance of RNA Methyltransferases. *Front. Genet.*, **2016**, *7*, 99. https://doi.org/10.3389/fgene.2016.00099.
[287] Motorin, Y.; Helm, M. RNA Nucleotide Methylation. *Wiley Interdiscip. Rev. RNA*, **2011**, *2* (5), 611–631. https://doi.org/10.1002/wrna.79.
[288] Gilbert, W. V.; Bell, T. A.; Schaening, C. Messenger RNA Modifications: Form, Distribution, and Function. *Science*, **2016**, *352* (6292), 1408–1412. https://doi.org/10.1126/science.aad8711.
[289] Yue, Y.; Liu, J.; He, C. RNA N6-Methyladenosine Methylation in Post-Transcriptional Gene Expression Regulation. *Genes Dev.*, **2015**, *29* (13), 1343–1355. https://doi.org/10.1101/gad.262766.115.
[290] Sarin, L. P.; Leidel, S. A. Modify or Die?--RNA Modification Defects in Metazoans. *RNA Biol.*, **2014**, *11* (12), 1555–1567. https://doi.org/10.4161/15476286.2014.992279.
[291] Machnicka, M. A.; Milanowska, K.; Osman Oglou, O.; Purta, E.; Kurkowska, M.; Olchowik, A.; Januszewski, W.; Kalinowski, S.; Dunin-Horkawicz, S.; Rother, K. M.; et al. MODOMICS: A Database of RNA Modification Pathways--2013 Update. *Nucleic Acids Res.*, **2013**, *41* (Database issue), D262-267. https://doi.org/10.1093/nar/gks1007.
[292] Hori, H. Methylated Nucleosides in TRNA and TRNA Methyltransferases. *Front. Genet.*, **2014**, *5*, 144. https://doi.org/10.3389/fgene.2014.00144.
[293] Gillingham, D.; Shahid, R. Catalysts for RNA and DNA Modification. *Curr. Opin. Chem. Biol.*, **2015**, *25*, 110–114. https://doi.org/10.1016/j.cbpa.2014.12.025.
[294] El Yacoubi, B.; Bailly, M.; de Crécy-Lagard, V. Biosynthesis and Function of Posttranscriptional Modifications of Transfer RNAs. *Annu. Rev. Genet.*, **2012**, *46*, 69–95. https://doi.org/10.1146/annurev-genet-110711-155641.
[295] Kawamura, T.; Anraku, R.; Hasegawa, T.; Tomikawa, C.; Hori, H. Transfer RNA Methyltransferases from Thermoplasma Acidophilum, a Thermoacidophilic Archaeon. *Int. J. Mol. Sci.*, **2014**, *16* (1), 91–113. https://doi.org/10.3390/ijms16010091.
[296] Phillips, G.; de Crécy-Lagard, V. Biosynthesis and Function of TRNA Modifications in Archaea. *Curr. Opin. Microbiol.*, **2011**, *14* (3), 335–341. https://doi.org/10.1016/j.mib.2011.03.001.
[297] Gustafsson, C.; Reid, R.; Greene, P. J.; Santi, D. V. Identification of New RNA Modifying Enzymes by Iterative Genome Search Using Known Modifying Enzymes as Probes. *Nucleic Acids Res.*, **1996**, *24* (19), 3756–3762. https://doi.org/10.1093/nar/24.19.3756.
[298] Anantharaman, V.; Koonin, E. V.; Aravind, L. SPOUT: A Class of Methyltransferases That Includes SpoU and TrmD RNA Methylase Superfamilies, and Novel Superfamilies of Predicted Prokaryotic RNA Methylases. *J. Mol. Microbiol. Biotechnol.*, **2002**, *4* (1), 71–75.
[299] Koonin, E. V.; Rudd, K. E. SpoU Protein of Escherichia Coli Belongs to a New Family of Putative RRNA Methylases. *Nucleic Acids Res.*, **1993**, *21* (23), 5519. https://doi.org/10.1093/nar/21.23.5519.
[300] Pleshe, E.; Truesdell, J.; Batey, R. T. Structure of a Class II TrmH TRNA-Modifying Enzyme from Aquifex Aeolicus. *Acta Crystallograph. Sect. F Struct. Biol. Cryst. Commun.*, **2005**, *61* (Pt 8), 722–728. https://doi.org/10.1107/S1744309105022980.

[301] Eichler, D. C. Characterization of a Nucleolar 2'-O-Methyltransferase and Its Involvement in the Methylation of Mouse Precursor Ribosomal RNA. *Biochimie*, **1994**, *76* (12), 1115–1122. https://doi.org/10.1016/0300-9084(94)90039-6.

[302] Kawai, G.; Yamamoto, Y.; Kamimura, T.; Masegi, T.; Sekine, M.; Hata, T.; Iimori, T.; Watanabe, T.; Miyazawa, T.; Yokoyama, S. Conformational Rigidity of Specific Pyrimidine Residues in TRNA Arises from Posttranscriptional Modifications That Enhance Steric Interaction between the Base and the 2'-Hydroxyl Group. *Biochemistry*, **1992**, *31* (4), 1040–1046. https://doi.org/10.1021/bi00119a012.

[303] Kita, E.; Kashiba, S. Immunogenicity of Transfer RNA Isolated from a Two-Heptose Rough Mutant of Salmonella Typhimurium LT2 in Mouse Typhoid Infection. *Immunology*, **1983**, *50* (3), 369–376.

[304] Jöckel, S.; Nees, G.; Sommer, R.; Zhao, Y.; Cherkasov, D.; Hori, H.; Ehm, G.; Schnare, M.; Nain, M.; Kaufmann, A.; et al. The 2′-O-Methylation Status of a Single Guanosine Controls Transfer RNA–Mediated Toll-like Receptor 7 Activation or Inhibition. *J. Exp. Med.*, **2012**, *209* (2), 235–241. https://doi.org/10.1084/jem.20111075.

[305] Oldenburg, M.; Krüger, A.; Ferstl, R.; Kaufmann, A.; Nees, G.; Sigmund, A.; Bathke, B.; Lauterbach, H.; Suter, M.; Dreher, S.; et al. TLR13 Recognizes Bacterial 23S RRNA Devoid of Erythromycin Resistance-Forming Modification. *Science*, **2012**, *337* (6098), 1111–1115. https://doi.org/10.1126/science.1220363.

[306] Maden, B. E.; Corbett, M. E.; Heeney, P. A.; Pugh, K.; Ajuh, P. M. Classical and Novel Approaches to the Detection and Localization of the Numerous Modified Nucleotides in Eukaryotic Ribosomal RNA. *Biochimie*, **1995**, *77* (1–2), 22–29. https://doi.org/10.1016/0300-9084(96)88100-4.

[307] Motorin, Y.; Muller, S.; Behm-Ansmant, I.; Branlant, C. Identification of Modified Residues in RNAs by Reverse Transcription-Based Methods. *Methods Enzymol.*, **2007**, *425*, 21–53. https://doi.org/10.1016/S0076-6879(07)25002-5.

[308] Gehrke, C. W.; Kuo, K. C.; McCune, R. A.; Gerhardt, K. O.; Agris, P. F. Quantitative Enzymatic Hydrolysis of TRNAs: Reversed-Phase High-Performance Liquid Chromatography of TRNA Nucleosides. *J. Chromatogr.*, **1982**, *230* (2), 297–308.

[309] Su, D.; Chan, C. T. Y.; Gu, C.; Lim, K. S.; Chionh, Y. H.; McBee, M. E.; Russell, B. S.; Babu, I. R.; Begley, T. J.; Dedon, P. C. Quantitative Analysis of Ribonucleoside Modifications in TRNA by HPLC-Coupled Mass Spectrometry. *Nat. Protoc.*, **2014**, *9* (4), 828–841. https://doi.org/10.1038/nprot.2014.047.

[310] Cao, X.; Limbach, P. A. Enhanced Detection of Post-Transcriptional Modifications Using a Mass-Exclusion List Strategy for RNA Modification Mapping by LC-MS/MS. *Anal. Chem.*, **2015**, *87* (16), 8433–8440. https://doi.org/10.1021/acs.analchem.5b01826.

[311] Abbate, J.; Rottman, F. Gas Chromatographic Method for Determination of 2'-O-Methylation in RNA. *Anal. Biochem.*, **1972**, *47* (2), 378–388. https://doi.org/10.1016/0003-2697(72)90131-5.

[312] Dominissini, D.; Moshitch-Moshkovitz, S.; Salmon-Divon, M.; Amariglio, N.; Rechavi, G. Transcriptome-Wide Mapping of N(6)-Methyladenosine by m(6)A-Seq Based on Immunocapturing and Massively Parallel Sequencing. *Nat. Protoc.*, **2013**, *8* (1), 176–189. https://doi.org/10.1038/nprot.2012.148.

[313] Dominissini, D.; Nachtergaele, S.; Moshitch-Moshkovitz, S.; Peer, E.; Kol, N.; Ben-Haim, M. S.; Dai, Q.; Di Segni, A.; Salmon-Divon, M.; Clark, W. C.; et al. The Dynamic N(1)-Methyladenosine Methylome in Eukaryotic Messenger RNA. *Nature*, **2016**, *530* (7591), 441–446. https://doi.org/10.1038/nature16998.

[314] Schwartz, S.; Bernstein, D. A.; Mumbach, M. R.; Jovanovic, M.; Herbst, R. H.; León-Ricardo, B. X.; Engreitz, J. M.; Guttman, M.; Satija, R.; Lander, E. S.; et al. Transcriptome-Wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of NcRNA and MRNA. *Cell*, **2014**, *159* (1), 148–162. https://doi.org/10.1016/j.cell.2014.08.028.

[315] Squires, J. E.; Patel, H. R.; Nousch, M.; Sibbritt, T.; Humphreys, D. T.; Parker, B. J.; Suter, C. M.; Preiss, T. Widespread Occurrence of 5-Methylcytosine in Human Coding and Non-Coding RNA. *Nucleic Acids Res.*, **2012**, *40* (11), 5023–5033. https://doi.org/10.1093/nar/gks144.

[316] Tserovski, L.; Marchand, V.; Hauenschild, R.; Blanloeil-Oillo, F.; Helm, M.; Motorin, Y. High-Throughput Sequencing for 1-Methyladenosine (m(1)A) Mapping in RNA. *Methods San Diego Calif*, **2016**, *107*, 110–121. https://doi.org/10.1016/j.ymeth.2016.02.012.

[317] Bolger, A. M.; Lohse, M.; Usadel, B. Trimmomatic: A Flexible Trimmer for Illumina Sequence Data. *Bioinforma. Oxf. Engl.*, **2014**, *30* (15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

[318] Langmead, B.; Salzberg, S. L. Fast Gapped-Read Alignment with Bowtie 2. *Nat. Methods*, **2012**, *9* (4), 357–359. https://doi.org/10.1038/nmeth.1923.

[319] Quinlan, A. R.; Hall, I. M. BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features. *Bioinformatics*, **2010**, *26* (6), 841–842. https://doi.org/10.1093/bioinformatics/btq033.

[320] Carlile, T. M.; Rojas-Duran, M. F.; Zinshteyn, B.; Shin, H.; Bartoli, K. M.; Gilbert, W. V. Pseudouridine Profiling Reveals Regulated MRNA Pseudouridylation in Yeast and Human Cells. *Nature*, **2014**, *515* (7525), 143–146. https://doi.org/10.1038/nature13802.

[321] Hauenschild, R.; Tserovski, L.; Schmid, K.; Thüring, K.; Winz, M.-L.; Sharma, S.; Entian, K.-D.; Wacheul, L.; Lafontaine, D. L. J.; Anderson, J.; et al. The Reverse Transcription Signature of N-1-Methyladenosine in RNA-Seq Is Sequence Dependent. *Nucleic Acids Res.*, **2015**, *43* (20), 9950–9964. https://doi.org/10.1093/nar/gkv895.

[322] Schwartz, S. M1A within Cytoplasmic MRNAs at Single Nucleotide Resolution: A Reconciled Transcriptome-Wide Map. *RNA N. Y. N*, **2018**, *24* (11), 1427–1436. https://doi.org/10.1261/rna.067348.118.

[323] Freund, I.; Buhl, D. K.; Boutin, S.; Kotter, A.; Pichot, F.; Marchand, V.; Vierbuchen, T.; Heine, H.; Motorin, Y.; Helm, M.; et al. 2'-O-Methylation within Prokaryotic and Eukaryotic TRNA Inhibits Innate Immune Activation by Endosomal Toll-like Receptors but Does Not Affect Recognition of Whole Organisms. *RNA N. Y. N*, **2019**, *25* (7), 869–880. https://doi.org/10.1261/rna.070243.118.

[324] Piekna-Przybylska, D.; Decatur, W. A.; Fournier, M. J. The 3D RRNA Modification Maps Database: With Interactive Tools for Ribosome Analysis. *Nucleic Acids Res.*, **2008**, *36* (Database issue), D178-183. https://doi.org/10.1093/nar/gkm855.

[325] Gkatza, N. A.; Castro, C.; Harvey, R. F.; Heiß, M.; Popis, M. C.; Blanco, S.; Bornelöv, S.; Sajini, A. A.; Gleeson, J. G.; Griffin, J. L.; et al. Cytosine-5 RNA

Methylation Links Protein Synthesis to Cell Metabolism. *PLoS Biol.*, **2019**, *17* (6), e3000297. https://doi.org/10.1371/journal.pbio.3000297.

[326] Shin, G.; Koo, H. J.; Seo, M.; Lee, S.-J. V.; Nam, H. G.; Jung, G. Y. Transfer RNA-Derived Fragments in Aging Caenorhabditis Elegans Originate from Abundant Homologous Gene Copies. *Sci. Rep.*, **2021**, *11* (1), 12304. https://doi.org/10.1038/s41598-021-91724-z.

[327] Lin, S.; Liu, Q.; Jiang, Y.-Z.; Gregory, R. I. Nucleotide Resolution Profiling of M7G TRNA Modification by TRAC-Seq. *Nat. Protoc.*, **2019**, *14* (11), 3220–3242. https://doi.org/10.1038/s41596-019-0226-7.

[328] Jühling, F.; Mörl, M.; Hartmann, R. K.; Sprinzl, M.; Stadler, P. F.; Pütz, J. TRNAdb 2009: Compilation of TRNA Sequences and TRNA Genes. *Nucleic Acids Res.*, **2009**, *37* (Database issue), D159-162. https://doi.org/10.1093/nar/gkn772.

[329] Boccaletto, P.; Stefaniak, F.; Ray, A.; Cappannini, A.; Mukherjee, S.; Purta, E.; Kurkowska, M.; Shirvanizadeh, N.; Destefanis, E.; Groza, P.; et al. MODOMICS: A Database of RNA Modification Pathways. 2021 Update. *Nucleic Acids Res.*, **2022**, *50* (D1), D231–D235. https://doi.org/10.1093/nar/gkab1083.

[330] Westhof, E.; Thornlow, B.; Chan, P. P.; Lowe, T. M. Eukaryotic TRNA Sequences Present Conserved and Amino Acid-Specific Structural Signatures. *Nucleic Acids Res.*, **2022**, *50* (7), 4100–4112. https://doi.org/10.1093/nar/gkac222.

[331] Chan, P. P.; Lowe, T. M. GtRNAdb 2.0: An Expanded Database of Transfer RNA Genes Identified in Complete and Draft Genomes. *Nucleic Acids Res.*, **2016**, *44* (D1), D184-189. https://doi.org/10.1093/nar/gkv1309.

[332] Clark, W. C.; Evans, M. E.; Dominissini, D.; Zheng, G.; Pan, T. TRNA Base Methylation Identification and Quantification via High-Throughput Sequencing. *RNA N. Y. N*, **2016**, *22* (11), 1771–1784. https://doi.org/10.1261/rna.056531.116.

[333] Yamagami, R.; Sieg, J. P.; Assmann, S. M.; Bevilacqua, P. C. Genome-Wide Analysis of the in Vivo TRNA Structurome Reveals RNA Structural and Modification Dynamics under Heat Stress. *Proc. Natl. Acad. Sci. U. S. A.*, **2022**, *119* (25), e2201237119. https://doi.org/10.1073/pnas.2201237119.

[334] Berg, M. D.; Brandl, C. J. Transfer RNAs: Diversity in Form and Function. *RNA Biol.*, **2020**, 1–24. https://doi.org/10.1080/15476286.2020.1809197.

[335] Lei, L.; Burton, Z. F. Evolution of Life on Earth: TRNA, Aminoacyl-TRNA Synthetases and the Genetic Code. *Life Basel Switz.*, **2020**, *10* (3). https://doi.org/10.3390/life10030021.

[336] Phizicky, E. M.; Hopper, A. K. TRNA Biology Charges to the Front. *Genes Dev.*, **2010**, *24* (17), 1832–1860. https://doi.org/10.1101/gad.1956510.

[337] Wilusz, J. E. Controlling Translation via Modulation of TRNA Levels. *Wiley Interdiscip. Rev. RNA*, **2015**, *6* (4), 453–470. https://doi.org/10.1002/wrna.1287.

[338] Pinkard, O.; McFarland, S.; Sweet, T.; Coller, J. Quantitative TRNA-Sequencing Uncovers Metazoan Tissue-Specific TRNA Regulation. *Nat. Commun.*, **2020**, *11* (1), 4104. https://doi.org/10.1038/s41467-020-17879-x.

[339] Clark, W. C.; Evans, M. E.; Dominissini, D.; Zheng, G.; Pan, T. TRNA Base Methylation Identification and Quantification via High-Throughput Sequencing. *RNA N. Y. N*, **2016**, *22* (11), 1771–1784. https://doi.org/10.1261/rna.056531.116.

[340] Cozen, A. E.; Quartley, E.; Holmes, A. D.; Hrabeta-Robinson, E.; Phizicky, E. M.; Lowe, T. M. ARM-Seq: AlkB-Facilitated RNA Methylation Sequencing Reveals a Complex Landscape of Modified TRNA Fragments. *Nat. Methods*, **2015**, *12* (9), 879–884. https://doi.org/10.1038/nmeth.3508.

[341] Warren, J. M.; Salinas-Giegé, T.; Hummel, G.; Coots, N. L.; Svendsen, J. M.; Brown, K. C.; Drouard, L.; Sloan, D. B. Combining TRNA Sequencing Methods to Characterize Plant TRNA Expression and Post-Transcriptional Modification. *RNA Biol.*, **2020**, 1–15. https://doi.org/10.1080/15476286.2020.1792089.

[342] Lodish, H.; Berk, A.; Zipursky, S. L.; Matsudaira, P.; Baltimore, D.; Darnell, J. Processing of RRNA and TRNA. *Mol. Cell Biol. 4th Ed.*, **2000**.

[343] Drino, A.; Oberbauer, V.; Troger, C.; Janisiw, E.; Anrather, D.; Hartl, M.; Kaiser, S.; Kellner, S.; Schaefer, M. R. Production and Purification of Endogenously Modified TRNA-Derived Small RNAs. *RNA Biol.*, **2020**, 1–12. https://doi.org/10.1080/15476286.2020.1733798.

[344] Kanwal, F.; Lu, C. A Review on Native and Denaturing Purification Methods for Non-Coding RNA (NcRNA). *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.*, **2019**, *1120*, 71–79. https://doi.org/10.1016/j.jchromb.2019.04.034.

[345] Jacob, D.; Thüring, K.; Galliot, A.; Marchand, V.; Galvanin, A.; Ciftci, A.; Scharmann, K.; Stock, M.; Roignant, J.-Y.; Leidel, S. A.; et al. Absolute Quantification of Noncoding RNA by Microscale Thermophoresis. *Angew. Chem. Int. Ed Engl.*, **2019**. https://doi.org/10.1002/anie.201814377.

[346] Coughlin, D. J.; Babak, T.; Nihranz, C.; Hughes, T. R.; Engelke, D. R. Prediction and Verification of Mouse TRNA Gene Families. *RNA Biol.*, **2009**, *6* (2), 195–202. https://doi.org/10.4161/rna.6.2.8050.

[347] Shigematsu, M.; Honda, S.; Loher, P.; Telonis, A. G.; Rigoutsos, I.; Kirino, Y. YAMAT-Seq: An Efficient Method for High-Throughput Sequencing of Mature Transfer RNAs. *Nucleic Acids Res.*, **2017**, *45* (9), e70. https://doi.org/10.1093/nar/gkx005.

[348] Erber, L.; Hoffmann, A.; Fallmann, J.; Betat, H.; Stadler, P. F.; Mörl, M. LOTTE-Seq (Long Hairpin Oligonucleotide Based TRNA High-Throughput Sequencing): Specific Selection of TRNAs with 3'-CCA End for High-Throughput Sequencing. *RNA Biol.*, **2020**, *17* (1), 23–32. https://doi.org/10.1080/15476286.2019.1664250.

[349] Zheng, G.; Qin, Y.; Clark, W. C.; Dai, Q.; Yi, C.; He, C.; Lambowitz, A. M.; Pan, T. Efficient and Quantitative High-Throughput TRNA Sequencing. *Nat. Methods*, **2015**, *12* (9), 835–837. https://doi.org/10.1038/nmeth.3478.

[350] Gogakos, T.; Brown, M.; Garzia, A.; Meyer, C.; Hafner, M.; Tuschl, T. Characterizing Expression and Processing of Precursor and Mature Human TRNAs by Hydro-TRNAseq and PAR-CLIP. *Cell Rep.*, **2017**, *20* (6), 1463–1475. https://doi.org/10.1016/j.celrep.2017.07.029.

[351] Abe, T.; Ikemura, T.; Sugahara, J.; Kanai, A.; Ohara, Y.; Uehara, H.; Kinouchi, M.; Kanaya, S.; Yamada, Y.; Muto, A.; et al. TRNADB-CE 2011: TRNA Gene Database Curated Manually by Experts. *Nucleic Acids Res.*, **2011**, *39* (Database issue), D210-213. https://doi.org/10.1093/nar/gkq1007.

[352] Lowe, T. M.; Eddy, S. R. TRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.*, **1997**, *25* (5), 955–964. https://doi.org/10.1093/nar/25.5.955.

[353] Zou, Q.; Guo, J.; Ju, Y.; Wu, M.; Zeng, X.; Hong, Z. Improving TRNAscan-SE Annotation Results via Ensemble Classifiers. *Mol. Inform.*, **2015**, *34* (11–12), 761–770. https://doi.org/10.1002/minf.201500031.

[354] Sajek, M. P.; Woźniak, T.; Sprinzl, M.; Jaruzelska, J.; Barciszewski, J. T-Psi-C: User Friendly Database of TRNA Sequences and Structures. *Nucleic Acids Res.*, **2020**, *48* (D1), D256–D260. https://doi.org/10.1093/nar/gkz922.

[355] Torres, A. G.; Reina, O.; Stephan-Otto Attolini, C.; Ribas de Pouplana, L. Differential Expression of Human TRNA Genes Drives the Abundance of TRNA-Derived Fragments. *Proc. Natl. Acad. Sci. U. S. A.*, **2019**, *116* (17), 8451–8456. https://doi.org/10.1073/pnas.1821120116.

[356] Galvanin, A.; Ayadi, L.; Helm, M.; Motorin, Y.; Marchand, V. Mapping and Quantification of TRNA 2'-O-Methylation by RiboMethSeq. *Methods Mol. Biol. Clifton NJ*, **2019**, *1870*, 273–295. https://doi.org/10.1007/978-1-4939-8808-2_21.

[357] Fujishima, K.; Kanai, A. TRNA Gene Diversity in the Three Domains of Life. *Front. Genet.*, **2014**, *5*, 142. https://doi.org/10.3389/fgene.2014.00142.

[358] Wald, N.; Margalit, H. Auxiliary TRNAs: Large-Scale Analysis of TRNA Genes Reveals Patterns of TRNA Repertoire Dynamics. *Nucleic Acids Res.*, **2014**, *42* (10), 6552–6566. https://doi.org/10.1093/nar/gku245.

[359] Chen, P.; Qian, Q.; Zhang, S.; Isaksson, L. A.; Björk, G. R. A Cytosolic TRNA with an Unmodified Adenosine in the Wobble Position Reads a Codon Ending with the Non-Complementary Nucleoside Cytidine. *J. Mol. Biol.*, **2002**, *317* (4), 481–492. https://doi.org/10.1006/jmbi.2002.5435.

[360] Marchand, V.; Ayadi, L.; Ernst, F. G. M.; Hertler, J.; Bourguignon-Igel, V.; Galvanin, A.; Kotter, A.; Helm, M.; Lafontaine, D. L. J.; Motorin, Y. AlkAniline-Seq: Profiling of M7 G and M3 C RNA Modifications at Single Nucleotide Resolution. *Angew. Chem. Int. Ed Engl.*, **2018**, *57* (51), 16785–16790. https://doi.org/10.1002/anie.201810946.

[361] Marchand, V.; Pichot, F.; Neybecker, P.; Ayadi, L.; Bourguignon-Igel, V.; Wacheul, L.; Lafontaine, D. L. J.; Pinzano, A.; Helm, M.; Motorin, Y. HydraPsiSeq: A Method for Systematic and Quantitative Mapping of Pseudouridines in RNA. *Nucleic Acids Res.*, **2020**. https://doi.org/10.1093/nar/gkaa769.

[362] Li, H.; Durbin, R. Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform. *Bioinforma. Oxf. Engl.*, **2009**, *25* (14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

[363] Dobin, A.; Davis, C. A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T. R. STAR: Ultrafast Universal RNA-Seq Aligner. *Bioinforma. Oxf. Engl.*, **2013**, *29* (1), 15–21. https://doi.org/10.1093/bioinformatics/bts635.

[364] Lovato, M. A.; Chihade, J. W.; Schimmel, P. Translocation within the Acceptor Helix of a Major TRNA Identity Determinant. *EMBO J.*, **2001**, *20* (17), 4846–4853. https://doi.org/10.1093/emboj/20.17.4846.

[365] Koski, G. K.; Karikó, K.; Xu, S.; Weissman, D.; Cohen, P. A.; Czerniecki, B. J. Cutting Edge: Innate Immune System Discriminates between RNA Containing Bacterial versus Eukaryotic Structural Features That Prime for High-Level, IL-12 Secretion by Dendritic Cells. *J. Immunol.*, **2004**, *172* (7), 3989–3993. https://doi.org/10.4049/jimmunol.172.7.3989.

[366] Eberle, F.; Sirin, M.; Binder, M.; Dalpke, A. H. Bacterial RNA Is Recognized by Different Sets of Immunoreceptors. *Eur. J. Immunol.*, **2009**, *39* (9), 2537–2547. https://doi.org/10.1002/eji.200838978.

[367] Eigenbrod, T.; Pelka, K.; Latz, E.; Kreikemeyer, B.; Dalpke, A. H. TLR8 Senses Bacterial RNA in Human Monocytes and Plays a Nonredundant Role for Recognition of Streptococcus Pyogenes. *J. Immunol. Baltim. Md 1950*, **2015**, *195* (3), 1092–1099. https://doi.org/10.4049/jimmunol.1403173.

[368] Diebold, S. S.; Kaisho, T.; Hemmi, H.; Akira, S.; Reis e Sousa, C. Innate Antiviral Responses by Means of TLR7-Mediated Recognition of Single-Stranded RNA. *Science*, **2004**, *303* (5663), 1529–1531. https://doi.org/10.1126/science.1093616.

[369] Heil, F.; Hemmi, H.; Hochrein, H.; Ampenberger, F.; Kirschning, C.; Akira, S.; Lipford, G.; Wagner, H.; Bauer, S. Species-Specific Recognition of Single-Stranded RNA via Toll-like Receptor 7 and 8. *Science*, **2004**, *303* (5663), 1526–1529. https://doi.org/10.1126/science.1093620.

[370] Ablasser, A.; Poeck, H.; Anz, D.; Berger, M.; Schlee, M.; Kim, S.; Bourquin, C.; Goutagny, N.; Jiang, Z.; Fitzgerald, K. A.; et al. Selection of Molecular Structure and Delivery of RNA Oligonucleotides to Activate TLR7 versus TLR8 and to Induce High Amounts of IL-12p70 in Primary Human Monocytes. *J. Immunol. Baltim. Md 1950*, **2009**, *182* (11), 6824–6833. https://doi.org/10.4049/jimmunol.0803001.

[371] Bergstrøm, B.; Aune, M. H.; Awuh, J. A.; Kojen, J. F.; Blix, K. J.; Ryan, L.; Flo, T. H.; Mollnes, T. E.; Espevik, T.; Stenvik, J. TLR8 Senses Staphylococcus Aureus RNA in Human Primary Monocytes and Macrophages and Induces IFN-β Production via a TAK1-IKKβ-IRF5 Signaling Pathway. *J. Immunol. Baltim. Md 1950*, **2015**, *195* (3), 1100–1111. https://doi.org/10.4049/jimmunol.1403176.

[372] Saitoh, S.-I.; Abe, F.; Kanno, A.; Tanimura, N.; Mori Saitoh, Y.; Fukui, R.; Shibata, T.; Sato, K.; Ichinohe, T.; Hayashi, M.; et al. TLR7 Mediated Viral Recognition Results in Focal Type I Interferon Secretion by Dendritic Cells. *Nat. Commun.*, **2017**, *8* (1), 1592. https://doi.org/10.1038/s41467-017-01687-x.

[373] Barrat, F. J.; Meeker, T.; Gregorio, J.; Chan, J. H.; Uematsu, S.; Akira, S.; Chang, B.; Duramad, O.; Coffman, R. L. Nucleic Acids of Mammalian Origin Can Act as Endogenous Ligands for Toll-like Receptors and May Promote Systemic Lupus Erythematosus. *J. Exp. Med.*, **2005**, *202* (8), 1131–1139. https://doi.org/10.1084/jem.20050914.

[374] Kono, D. H.; Haraldsson, M. K.; Lawson, B. R.; Pollard, K. M.; Koh, Y. T.; Du, X.; Arnold, C. N.; Baccala, R.; Silverman, G. J.; Beutler, B. A.; et al. Endosomal TLR Signaling Is Required for Anti-Nucleic Acid and Rheumatoid Factor Autoantibodies in Lupus. *Proc. Natl. Acad. Sci. U. S. A.*, **2009**, *106* (29), 12061–12066. https://doi.org/10.1073/pnas.0905441106.

[375] Karikó, K.; Buckstein, M.; Ni, H.; Weissman, D. Suppression of RNA Recognition by Toll-like Receptors: The Impact of Nucleoside Modification and the Evolutionary Origin of RNA. *Immunity*, **2005**, *23* (2), 165–175. https://doi.org/10.1016/j.immuni.2005.06.008.

[376] Sioud, M.; Furset, G.; Cekaite, L. Suppression of Immunostimulatory SiRNA-Driven Innate Immune Activation by 2'-Modified RNAs. *Biochem. Biophys. Res. Commun.*, **2007**, *361* (1), 122–126. https://doi.org/10.1016/j.bbrc.2007.06.177.

[377] Schmitt, F. C. F.; Freund, I.; Weigand, M. A.; Helm, M.; Dalpke, A. H.; Eigenbrod, T. Identification of an Optimized 2'-O-Methylated Trinucleotide RNA Motif

Inhibiting Toll-like Receptors 7 and 8. *RNA N. Y. N*, **2017**, *23* (9), 1344–1351. https://doi.org/10.1261/rna.061952.117.

[378] Eberle, F.; Giessler, K.; Deck, C.; Heeg, K.; Peter, M.; Richert, C.; Dalpke, A. H. Modifications in Small Interfering RNA That Separate Immunostimulation from RNA Interference. *J. Immunol. Baltim. Md 1950*, **2008**, *180* (5), 3229–3237. https://doi.org/10.4049/jimmunol.180.5.3229.

[379] Motorin, Y. RNA Modification. In *Encyclopedia of Life Sciences*; John Wiley & Sons, Ltd, 2015; pp 1–18. https://doi.org/10.1002/9780470015902.a0000528.pub3.

[380] Heitzler, J.; Maréchal-Drouard, L.; Dirheimer, G.; Keith, G. Use of a Dot Blot Hybridization Method for Identification of Pure TRNA Species on Different Membranes. *Biochim. Biophys. Acta BBA - Gene Struct. Expr.*, **1992**, *1129* (3), 273–277. https://doi.org/10.1016/0167-4781(92)90503-R.

[381] Beutler, B.; Du, X.; Poltorak, A. Identification of Toll-like Receptor 4 (Tlr4) as the Sole Conduit for LPS Signal Transduction: Genetic and Evolutionary Studies. *J. Endotoxin Res.*, **2001**, *7* (4), 277–280.

[382] Tada, H.; Nemoto, E.; Shimauchi, H.; Watanabe, T.; Mikami, T.; Matsumoto, T.; Ohno, N.; Tamura, H.; Shibata, K.; Akashi, S.; et al. Saccharomyces Cerevisiae- and Candida Albicans-Derived Mannan Induced Production of Tumor Necrosis Factor Alpha by Human Monocytes in a CD14- and Toll-like Receptor 4-Dependent Manner. *Microbiol. Immunol.*, **2002**, *46* (7), 503–512. https://doi.org/10.1111/j.1348-0421.2002.tb02727.x.

[383] Ferwerda, G.; Meyer-Wentrup, F.; Kullberg, B.-J.; Netea, M. G.; Adema, G. J. Dectin-1 Synergizes with TLR2 and TLR4 for Cytokine Production in Human Primary Monocytes and Macrophages. *Cell. Microbiol.*, **2008**, *10* (10), 2058–2066. https://doi.org/10.1111/j.1462-5822.2008.01188.x.

[384] Vierbuchen, T.; Bang, C.; Rosigkeit, H.; Schmitz, R. A.; Heine, H. The Human-Associated Archaeon Methanosphaera Stadtmanae Is Recognized through Its RNA and Induces TLR8-Dependent NLRP3 Inflammasome Activation. *Front. Immunol.*, **2017**, *8*, 1535. https://doi.org/10.3389/fimmu.2017.01535.

[385] Rapino, F.; Robles, E. F.; Richter-Larrea, J. A.; Kallin, E. M.; Martinez-Climent, J. A.; Graf, T. C/EBPα Induces Highly Efficient Macrophage Transdifferentiation of B Lymphoma and Leukemia Cell Lines and Impairs Their Tumorigenicity. *Cell Rep.*, **2013**, *3* (4), 1153–1163. https://doi.org/10.1016/j.celrep.2013.03.003.

[386] Gaidt, M. M.; Rapino, F.; Graf, T.; Hornung, V. Modeling Primary Human Monocytes with the Trans-Differentiation Cell Line BLaER1. *Methods Mol. Biol. Clifton NJ*, **2018**, *1714*, 57–66. https://doi.org/10.1007/978-1-4939-7519-8_4.

[387] Hamm, S.; Latz, E.; Hangel, D.; Müller, T.; Yu, P.; Golenbock, D.; Sparwasser, T.; Wagner, H.; Bauer, S. Alternating 2'-O-Ribose Methylation Is a Universal Approach for Generating Non-Stimulatory SiRNA by Acting as TLR7 Antagonist. *Immunobiology*, **2010**, *215* (7), 559–569. https://doi.org/10.1016/j.imbio.2009.09.003.

[388] Tang, H.; Jin, X.; Li, Y.; Jiang, H.; Tang, X.; Yang, X.; Cheng, H.; Qiu, Y.; Chen, G.; Mei, J.; et al. A Large-Scale Screen for Coding Variants Predisposing to Psoriasis. *Nat. Genet.*, **2014**, *46* (1), 45–50. https://doi.org/10.1038/ng.2827.

[389] Lyons, P. A.; McKinney, E. F.; Rayner, T. F.; Hatton, A.; Woffendin, H. B.; Koukoulaki, M.; Freeman, T. C.; Jayne, D. R. W.; Chaudhry, A. N.; Smith, K. G. C. Novel Expression Signatures Identified by Transcriptional Analysis of Separated Leucocyte Subsets in Systemic Lupus Erythematosus and Vasculitis.

Ann. Rheum. Dis., **2010**, *69* (6), 1208–1213. https://doi.org/10.1136/ard.2009.108043.

[390] Jansen, R. P.; Hurt, E. C.; Kern, H.; Lehtonen, H.; Carmo-Fonseca, M.; Lapeyre, B.; Tollervey, D. Evolutionary Conservation of the Human Nucleolar Protein Fibrillarin and Its Functional Expression in Yeast. *J. Cell Biol.*, **1991**, *113* (4), 715–729. https://doi.org/10.1083/jcb.113.4.715.

[391] Tollervey, D.; Lehtonen, H.; Carmo-Fonseca, M.; Hurt, E. C. The Small Nucleolar RNP Protein NOP1 (Fibrillarin) Is Required for Pre-RRNA Processing in Yeast. *EMBO J.*, **1991**, *10* (3), 573–583. https://doi.org/10.1002/j.1460-2075.1991.tb07984.x.

[392] Van Eenennaam, H.; Vogelzangs, J. H. P.; Bisschops, L.; Te Boome, L. C. J.; Seelig, H. P.; Renz, M.; De Rooij, D.-J.; Brouwer, R.; Pluk, H.; Pruijn, G. J. M.; et al. Autoantibodies against Small Nucleolar Ribonucleoprotein Complexes and Their Clinical Associations. *Clin. Exp. Immunol.*, **2002**, *130* (3), 532–540. https://doi.org/10.1046/j.1365-2249.2002.01991.x.

[393] Hagberg, N.; Rönnblom, L. Systemic Lupus Erythematosus--A Disease with A Dysregulated Type I Interferon System. *Scand. J. Immunol.*, **2015**, *82* (3), 199–207. https://doi.org/10.1111/sji.12330.

[394] David, R. Gm18, a Bacterial “Invisibility Cloak.” *Nat. Rev. Microbiol.*, **2012**, *10* (4), 239–239. https://doi.org/10.1038/nrmicro2767.

[395] Takeuchi, O.; Hoshino, K.; Kawai, T.; Sanjo, H.; Takada, H.; Ogawa, T.; Takeda, K.; Akira, S. Differential Roles of TLR2 and TLR4 in Recognition of Gram-Negative and Gram-Positive Bacterial Cell Wall Components. *Immunity*, **1999**, *11* (4), 443–451. https://doi.org/10.1016/s1074-7613(00)80119-3.

[396] Roger, T.; Froidevaux, C.; Le Roy, D.; Reymond, M. K.; Chanson, A.-L.; Mauri, D.; Burns, K.; Riederer, B. M.; Akira, S.; Calandra, T. Protection from Lethal Gram-Negative Bacterial Sepsis by Targeting Toll-like Receptor 4. *Proc. Natl. Acad. Sci. U. S. A.*, **2009**, *106* (7), 2348–2352. https://doi.org/10.1073/pnas.0808146106.

[397] Sander, L. E.; Davis, M. J.; Boekschoten, M. V.; Amsen, D.; Dascher, C. C.; Ryffel, B.; Swanson, J. A.; Müller, M.; Blander, J. M. Detection of Prokaryotic MRNA Signifies Microbial Viability and Promotes Immunity. *Nature*, **2011**, *474* (7351), 385–389. https://doi.org/10.1038/nature10072.

[398] Barbet, G.; Sander, L. E.; Geswell, M.; Leonardi, I.; Cerutti, A.; Iliev, I.; Blander, J. M. Sensing Microbial Viability through Bacterial RNA Augments T Follicular Helper Cell and Antibody Responses. *Immunity*, **2018**, *48* (3), 584-598.e5. https://doi.org/10.1016/j.immuni.2018.02.015.

[399] Ugolini, M.; Gerhard, J.; Burkert, S.; Jensen, K. J.; Georg, P.; Ebner, F.; Volkers, S. M.; Thada, S.; Dietert, K.; Bauer, L.; et al. Recognition of Microbial Viability via TLR8 Drives TFH Cell Differentiation and Vaccine Responses. *Nat. Immunol.*, **2018**, *19* (4), 386–396. https://doi.org/10.1038/s41590-018-0068-4.

[400] Kellner, S.; Ochel, A.; Thüring, K.; Spenkuch, F.; Neumann, J.; Sharma, S.; Entian, K.-D.; Schneider, D.; Helm, M. Absolute and Relative Quantification of RNA Modifications via Biosynthetic Isotopomers. *Nucleic Acids Res.*, **2014**, *42* (18), e142. https://doi.org/10.1093/nar/gku733.

[401] Chen, B.; Gilbert, L. A.; Cimini, B. A.; Schnitzbauer, J.; Zhang, W.; Li, G.-W.; Park, J.; Blackburn, E. H.; Weissman, J. S.; Qi, L. S.; et al. Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell*, **2013**, *155* (7), 1479–1491. https://doi.org/10.1016/j.cell.2013.12.001.

[402] Dehairs, J.; Talebi, A.; Cherifi, Y.; Swinnen, J. V. CRISP-ID: Decoding CRISPR Mediated Indels by Sanger Sequencing. *Sci. Rep.*, **2016**, *6* (1), 28973. https://doi.org/10.1038/srep28973.
[403] Mahesh, B. *Machine Learning Algorithms -A Review*; 2019. https://doi.org/10.21275/ART20203995.
[404] Liu, Y.; Mazumdar, S.; Bath, P. A.; Alzheimer's Disease Neuroimaging Initiative. An Unsupervised Learning Approach to Diagnosing Alzheimer's Disease Using Brain Magnetic Resonance Imaging Scans. *Int. J. Med. Inf.*, **2023**, *173*, 105027. https://doi.org/10.1016/j.ijmedinf.2023.105027.
[405] Chiorean, A.; Farncombe, K. M.; Delong, S.; Andric, V.; Ansar, S.; Chan, C.; Clark, K.; Danos, A. M.; Gao, Y.; Giles, R. H.; et al. Large Scale Genotype- and Phenotype-Driven Machine Learning in Von Hippel-Lindau Disease. *Hum. Mutat.*, **2022**, *43* (9), 1268–1285. https://doi.org/10.1002/humu.24392.
[406] Song, B.; Chen, K.; Tang, Y.; Ma, J.; Meng, J.; Wei, Z. PSI-MOUSE: Predicting Mouse Pseudouridine Sites From Sequence and Genome-Derived Features. *Evol. Bioinforma. Online*, **2020**, *16*, 1176934320925752. https://doi.org/10.1177/1176934320925752.
[407] Motorin, Y.; Helm, M. TRNA Stabilization by Modified Nucleotides. *Biochemistry*, **2010**, *49* (24), 4934–4944. https://doi.org/10.1021/bi100408z.
[408] Liu, J.; Jia, G. Methylation Modifications in Eukaryotic Messenger RNA. *J. Genet. Genomics Yi Chuan Xue Bao*, **2014**, *41* (1), 21–33. https://doi.org/10.1016/j.jgg.2013.10.002.
[409] Meyer, K. D.; Jaffrey, S. R. The Dynamic Epitranscriptome: N6-Methyladenosine and Gene Expression Control. *Nat. Rev. Mol. Cell Biol.*, **2014**, *15* (5), 313–326. https://doi.org/10.1038/nrm3785.
[410] Zhao, B. S.; Roundtree, I. A.; He, C. Post-Transcriptional Gene Regulation by MRNA Modifications. *Nat. Rev. Mol. Cell Biol.*, **2017**, *18* (1), 31–42. https://doi.org/10.1038/nrm.2016.132.
[411] Dimitrova, D. G.; Teysset, L.; Carré, C. RNA 2'-O-Methylation (Nm) Modification in Human Diseases. *Genes*, **2019**, *10* (2). https://doi.org/10.3390/genes10020117.
[412] Sibbritt, T.; Patel, H. R.; Preiss, T. Mapping and Significance of the MRNA Methylome. *Wiley Interdiscip. Rev. RNA*, **2013**, *4* (4), 397–422. https://doi.org/10.1002/wrna.1166.
[413] O'Connell, M. A.; Mannion, N. M.; Keegan, L. P. The Epitranscriptome and Innate Immunity. *PLoS Genet.*, **2015**, *11* (12), e1005687. https://doi.org/10.1371/journal.pgen.1005687.
[414] Borchardt, E. K.; Martinez, N. M.; Gilbert, W. V. Regulation and Function of RNA Pseudouridylation in Human Cells. *Annu. Rev. Genet.*, **2020**, *54*, 309–336. https://doi.org/10.1146/annurev-genet-112618-043830.
[415] Linder, B.; Jaffrey, S. R. Discovering and Mapping the Modified Nucleotides That Comprise the Epitranscriptome of MRNA. *Cold Spring Harb. Perspect. Biol.*, **2019**, *11* (6). https://doi.org/10.1101/cshperspect.a032201.
[416] Nachtergaele, S.; He, C. Chemical Modifications in the Life of an MRNA Transcript. *Annu. Rev. Genet.*, **2018**, *52*, 349–372. https://doi.org/10.1146/annurev-genet-120417-031522.
[417] Hoernes, T. P.; Erlacher, M. D. Translating the Epitranscriptome. *Wiley Interdiscip. Rev. RNA*, **2017**, *8* (1). https://doi.org/10.1002/wrna.1375.

[418] Suzuki, T.; Ueda, H.; Okada, S.; Sakurai, M. Transcriptome-Wide Identification of Adenosine-to-Inosine Editing Using the ICE-Seq Method. *Nat. Protoc.*, **2015**, *10* (5), 715–732. https://doi.org/10.1038/nprot.2015.037.
[419] Kietrys, A. M.; Velema, W. A.; Kool, E. T. Fingerprints of Modified RNA Bases from Deep Sequencing Profiles. *J. Am. Chem. Soc.*, **2017**, *139* (47), 17074–17081. https://doi.org/10.1021/jacs.7b07914.
[420] Motorin, Y.; Helm, M. Methods for RNA Modification Mapping Using Deep Sequencing: Established and New Emerging Technologies. *Genes*, **2019**, *10* (1). https://doi.org/10.3390/genes10010035.
[421] Motorin, Y.; Marchand, V. Analysis of RNA Modifications by Second- and Third-Generation Deep Sequencing: 2020 Update. *Genes*, **2021**, *12* (2), 278. https://doi.org/10.3390/genes12020278.
[422] Grozhik, A. V.; Jaffrey, S. R. Distinguishing RNA Modifications from Noise in Epitranscriptome Maps. *Nat. Chem. Biol.*, **2018**, *14* (3), 215–225. https://doi.org/10.1038/nchembio.2546.
[423] McIntyre, A. B. R.; Gokhale, N. S.; Cerchietti, L.; Jaffrey, S. R.; Horner, S. M.; Mason, C. E. Limits in the Detection of M6A Changes Using MeRIP/M6A-Seq. *Sci. Rep.*, **2020**, *10* (1), 6590. https://doi.org/10.1038/s41598-020-63355-3.
[424] Grozhik, A. V.; Jaffrey, S. R. Epitranscriptomics: Shrinking Maps of RNA Modifications. *Nature*, **2017**, *551* (7679), 174–176. https://doi.org/10.1038/nature24156.
[425] Wiener, D.; Schwartz, S. The Epitranscriptome beyond M6A. *Nat. Rev. Genet.*, **2020**. https://doi.org/10.1038/s41576-020-00295-8.
[426] Sas-Chen, A.; Schwartz, S. Misincorporation Signatures for Detecting Modifications in MRNA: Not as Simple as It Sounds. *Methods San Diego Calif*, **2019**, *156*, 53–59. https://doi.org/10.1016/j.ymeth.2018.10.011.
[427] Angelova, M. T.; Dimitrova, D. G.; Da Silva, B.; Marchand, V.; Jacquier, C.; Achour, C.; Brazane, M.; Goyenvalle, C.; Bourguignon-Igel, V.; Shehzada, S.; et al. TRNA 2'-O-Methylation by a Duo of TRM7/FTSJ1 Proteins Modulates Small RNA Silencing in Drosophila. *Nucleic Acids Res.*, **2020**, *48* (4), 2050–2072. https://doi.org/10.1093/nar/gkaa002.
[428] Azevedo-Favory, J.; Gaspin, C.; Ayadi, L.; Montacié, C.; Marchand, V.; Jobet, E.; Rompais, M.; Carapito, C.; Motorin, Y.; Sáez-Vásquez, J. Mapping RRNA 2'-O-Methylations and Identification of C/D SnoRNAs in Arabidopsis Thaliana Plants. *RNA Biol.*, **2021**, *18* (11), 1760–1777. https://doi.org/10.1080/15476286.2020.1869892.
[429] Delhermite, J.; Tafforeau, L.; Sharma, S.; Marchand, V.; Wacheul, L.; Lattuca, R.; Desiderio, S.; Motorin, Y.; Bellefroid, E.; Lafontaine, D. L. J. Systematic Mapping of RRNA 2'-O Methylation during Frog Development and Involvement of the Methyltransferase Fibrillarin in Eye and Craniofacial Development in Xenopus Laevis. *PLoS Genet.*, **2022**, *18* (1), e1010012. https://doi.org/10.1371/journal.pgen.1010012.
[430] Pichot, F.; Marchand, V.; Ayadi, L.; Bourguignon-Igel, V.; Helm, M.; Motorin, Y. Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2'-O-Methylations in RNA by RiboMethSeq. *Front. Genet.*, **2020**, *11*, 38. https://doi.org/10.3389/fgene.2020.00038.
[431] Motorin, Y.; Marchand, V. Detection and Analysis of RNA Ribose 2'-O-Methylations: Challenges and Solutions. *Genes*, **2018**, *9* (12). https://doi.org/10.3390/genes9120642.

[432] Marchand, V.; Ayadi, L.; El Hajj, A.; Blanloeil-Oillo, F.; Helm, M.; Motorin, Y. High-Throughput Mapping of 2'-O-Me Residues in RNA Using Next-Generation Sequencing (Illumina RiboMethSeq Protocol). *Methods Mol. Biol. Clifton NJ*, **2017**, *1562*, 171–187. https://doi.org/10.1007/978-1-4939-6807-7_12.

[433] Birkedal, U.; Christensen-Dalsgaard, M.; Krogh, N.; Sabarinathan, R.; Gorodkin, J.; Nielsen, H. Profiling of Ribose Methylations in RNA by High-Throughput Sequencing. *Angew. Chem. Int. Ed Engl.*, **2015**, *54* (2), 451–455. https://doi.org/10.1002/anie.201408362.

[434] Krogh, N.; Birkedal, U.; Nielsen, H. RiboMeth-Seq: Profiling of 2'-O-Me in RNA. *Methods Mol. Biol. Clifton NJ*, **2017**, *1562*, 189–209. https://doi.org/10.1007/978-1-4939-6807-7_13.

[435] Motorin, Y.; Quinternet, M.; Rhalloussi, W.; Marchand, V. Constitutive and Variable 2'-O-Methylation (Nm) in Human Ribosomal RNA. *RNA Biol.*, **2021**, *18* (sup1), 88–97. https://doi.org/10.1080/15476286.2021.1974750.

[436] Nadif, M.; Role, F. Unsupervised and Self-Supervised Deep Learning Approaches for Biomedical Text Mining. *Brief. Bioinform.*, **2021**, *22* (2), 1592–1603. https://doi.org/10.1093/bib/bbab016.

[437] Reel, P. S.; Reel, S.; Pearson, E.; Trucco, E.; Jefferson, E. Using Machine Learning Approaches for Multi-Omics Data Analysis: A Review. *Biotechnol. Adv.*, **2021**, *49*, 107739. https://doi.org/10.1016/j.biotechadv.2021.107739.

[438] Schmidt, B.; Hildebrandt, A. Deep Learning in Next-Generation Sequencing. *Drug Discov. Today*, **2021**, *26* (1), 173–180. https://doi.org/10.1016/j.drudis.2020.10.002.

[439] Zhang, Y.; Hamada, M. DeepM6ASeq: Prediction and Characterization of M6A-Containing Sequences Using Deep Learning. *BMC Bioinformatics*, **2018**, *19* (Suppl 19), 524. https://doi.org/10.1186/s12859-018-2516-4.

[440] Lin, K.-T.; Krainer, A. R. PSI-Sigma: A Comprehensive Splicing-Detection Method for Short-Read and Long-Read RNA-Seq Analysis. *Bioinforma. Oxf. Engl.*, **2019**, *35* (23), 5048–5054. https://doi.org/10.1093/bioinformatics/btz438.

[441] Begik, O.; Lucas, M. C.; Pryszcz, L. P.; Ramirez, J. M.; Medina, R.; Milenkovic, I.; Cruciani, S.; Liu, H.; Vieira, H. G. S.; Sas-Chen, A.; et al. Quantitative Profiling of Pseudouridylation Dynamics in Native RNAs with Nanopore Sequencing. *Nat. Biotechnol.*, **2021**, *39* (10), 1278–1291. https://doi.org/10.1038/s41587-021-00915-6.

[442] Lorenz, D. A.; Sathe, S.; Einstein, J. M.; Yeo, G. W. Direct RNA Sequencing Enables M6A Detection in Endogenous Transcript Isoforms at Base-Specific Resolution. *RNA N. Y. N*, **2020**, *26* (1), 19–28. https://doi.org/10.1261/rna.072785.119.

[443] Silvestre-Ryan, J.; Holmes, I. Pair Consensus Decoding Improves Accuracy of Neural Network Basecallers for Nanopore Sequencing. *Genome Biol.*, **2021**, *22* (1), 38. https://doi.org/10.1186/s13059-020-02255-1.

[444] Wick, R. R.; Judd, L. M.; Holt, K. E. Performance of Neural Network Basecalling Tools for Oxford Nanopore Sequencing. *Genome Biol.*, **2019**, *20* (1), 129. https://doi.org/10.1186/s13059-019-1727-y.

[445] Qin, H.; Ou, L.; Gao, J.; Chen, L.; Wang, J.-W.; Hao, P.; Li, X. DENA: Training an Authentic Neural Network Model Using Nanopore Sequencing Data of Arabidopsis Transcripts for Detection and Quantification of N6-Methyladenosine on RNA. *Genome Biol.*, **2022**, *23* (1), 25. https://doi.org/10.1186/s13059-021-02598-3.

[446] Ayadi, L.; Motorin, Y.; Marchand, V. Quantification of 2'-O-Me Residues in RNA Using next-Generation Sequencing (Illumina RiboMethSeq Protocol). *Methods Mol. Biol. Clifton NJ*, **2018**, *1649*, 29–48. https://doi.org/10.1007/978-1-4939-7213-5_2.
[447] Strobl, C.; Boulesteix, A.-L.; Zeileis, A.; Hothorn, T. Bias in Random Forest Variable Importance Measures: Illustrations, Sources and a Solution. *BMC Bioinformatics*, **2007**, *8*, 25. https://doi.org/10.1186/1471-2105-8-25.
[448] Zhou, Y.; Cui, Q.; Zhou, Y. NmSEER V2.0: A Prediction Tool for 2'-O-Methylation Sites Based on Random Forest and Multi-Encoding Combination. *BMC Bioinformatics*, **2019**, *20* (Suppl 25), 690. https://doi.org/10.1186/s12859-019-3265-8.
[449] Ao, C.; Zou, Q.; Yu, L. NmRF: Identification of Multispecies RNA 2'-O-Methylation Modification Sites from RNA Sequences. *Brief. Bioinform.*, **2022**, *23* (1), bbab480. https://doi.org/10.1093/bib/bbab480.
[450] Li, H.; Chen, L.; Huang, Z.; Luo, X.; Li, H.; Ren, J.; Xie, Y. DeepOMe: A Web Server for the Prediction of 2'-O-Me Sites Based on the Hybrid CNN and BLSTM Architecture. *Front. Cell Dev. Biol.*, **2021**, *9*, 686894. https://doi.org/10.3389/fcell.2021.686894.
[451] Marchand, V.; Bourguignon-Igel, V.; Helm, M.; Motorin, Y. Analysis of Pseudouridines and Other RNA Modifications Using HydraPsiSeq Protocol. *Methods San Diego Calif*, **2021**, S1046-2023(21)00206-1. https://doi.org/10.1016/j.ymeth.2021.08.008.
[452] Khoddami, V.; Yerra, A.; Mosbruger, T. L.; Fleming, A. M.; Burrows, C. J.; Cairns, B. R. Transcriptome-Wide Profiling of Multiple RNA Modifications Simultaneously at Single-Base Resolution. *Proc. Natl. Acad. Sci. U. S. A.*, **2019**, *116* (14), 6784–6789. https://doi.org/10.1073/pnas.1817334116.
[453] Lovejoy, A. F.; Riordan, D. P.; Brown, P. O. Transcriptome-Wide Mapping of Pseudouridines: Pseudouridine Synthases Modify Specific MRNAs in S. Cerevisiae. *PloS One*, **2014**, *9* (10), e110799. https://doi.org/10.1371/journal.pone.0110799.
[454] Li, X.; Zhu, P.; Ma, S.; Song, J.; Bai, J.; Sun, F.; Yi, C. Chemical Pulldown Reveals Dynamic Pseudouridylation of the Mammalian Transcriptome. *Nat. Chem. Biol.*, **2015**, *11* (8), 592–597. https://doi.org/10.1038/nchembio.1836.
[455] Zaringhalam, M.; Papavasiliou, F. N. Pseudouridylation Meets Next-Generation Sequencing. *Methods San Diego Calif*, **2016**, *107*, 63–72. https://doi.org/10.1016/j.ymeth.2016.03.001.
[456] Harcourt, E. M.; Kietrys, A. M.; Kool, E. T. Chemical and Structural Effects of Base Modifications in Messenger RNA. *Nature*, **2017**, *541* (7637), 339–346. https://doi.org/10.1038/nature21351.
[457] Zaccara, S.; Ries, R. J.; Jaffrey, S. R. Reading, Writing and Erasing MRNA Methylation. *Nat. Rev. Mol. Cell Biol.*, **2019**, *20* (10), 608–624. https://doi.org/10.1038/s41580-019-0168-5.
[458] Thüring, K.; Schmid, K.; Keller, P.; Helm, M. Analysis of RNA Modifications by Liquid Chromatography-Tandem Mass Spectrometry. *Methods San Diego Calif*, **2016**, *107*, 48–56. https://doi.org/10.1016/j.ymeth.2016.03.019.
[459] Thüring, K.; Schmid, K.; Keller, P.; Helm, M. LC-MS Analysis of Methylated RNA. *Methods Mol. Biol. Clifton NJ*, **2017**, *1562*, 3–18. https://doi.org/10.1007/978-1-4939-6807-7_1.

[460] Limbach, P. A.; Paulines, M. J. Going Global: The New Era of Mapping Modifications in RNA. *Wiley Interdiscip. Rev. RNA*, **2017**, *8* (1). https://doi.org/10.1002/wrna.1367.
[461] Jora, M.; Lobue, P. A.; Ross, R. L.; Williams, B.; Addepalli, B. Detection of Ribonucleoside Modifications by Liquid Chromatography Coupled with Mass Spectrometry. *Biochim. Biophys. Acta Gene Regul. Mech.*, **2018**. https://doi.org/10.1016/j.bbagrm.2018.10.012.
[462] Novoa, E. M.; Mason, C. E.; Mattick, J. S. Charting the Unknown Epitranscriptome. *Nat. Rev. Mol. Cell Biol.*, **2017**, *18* (6), 339–340. https://doi.org/10.1038/nrm.2017.49.
[463] Zhao, L.; Zhang, H.; Kohnen, M. V.; Prasad, K. V. S. K.; Gu, L.; Reddy, A. S. N. Analysis of Transcriptome and Epitranscriptome in Plants Using PacBio Iso-Seq and Nanopore-Based Direct RNA Sequencing. *Front. Genet.*, **2019**, *10*, 253. https://doi.org/10.3389/fgene.2019.00253.
[464] Dominissini, D.; Moshitch-Moshkovitz, S.; Schwartz, S.; Salmon-Divon, M.; Ungar, L.; Osenberg, S.; Cesarkas, K.; Jacob-Hirsch, J.; Amariglio, N.; Kupiec, M.; et al. Topology of the Human and Mouse M6A RNA Methylomes Revealed by M6A-Seq. *Nature*, **2012**, *485* (7397), 201–206. https://doi.org/10.1038/nature11112.
[465] Meyer, K. D.; Saletore, Y.; Zumbo, P.; Elemento, O.; Mason, C. E.; Jaffrey, S. R. Comprehensive Analysis of MRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell*, **2012**, *149* (7), 1635–1646. https://doi.org/10.1016/j.cell.2012.05.003.
[466] Slama, K.; Galliot, A.; Weichmann, F.; Hertler, J.; Feederle, R.; Meister, G.; Helm, M. Determination of Enrichment Factors for Modified RNA in MeRIP Experiments. *Methods San Diego Calif*, **2019**, *156*, 102–109. https://doi.org/10.1016/j.ymeth.2018.10.020.
[467] Molinie, B.; Wang, J.; Lim, K. S.; Hillebrand, R.; Lu, Z.-X.; Van Wittenberghe, N.; Howard, B. D.; Daneshvar, K.; Mullen, A. C.; Dedon, P.; et al. M(6)A-LAIC-Seq Reveals the Census and Complexity of the m(6)A Epitranscriptome. *Nat. Methods*, **2016**, *13* (8), 692–698. https://doi.org/10.1038/nmeth.3898.
[468] Schaefer, M.; Pollex, T.; Hanna, K.; Lyko, F. RNA Cytosine Methylation Analysis by Bisulfite Sequencing. *Nucleic Acids Res.*, **2009**, *37* (2), e12. https://doi.org/10.1093/nar/gkn954.
[469] Motorin, Y.; Lyko, F.; Helm, M. 5-Methylcytosine in RNA: Detection, Enzymatic Formation and Biological Functions. *Nucleic Acids Res.*, **2010**, *38* (5), 1415–1430. https://doi.org/10.1093/nar/gkp1117.
[470] Legrand, C.; Tuorto, F.; Hartmann, M.; Liebers, R.; Jacob, D.; Helm, M.; Lyko, F. Statistically Robust Methylation Calling for Whole-Transcriptome Bisulfite Sequencing Reveals Distinct Methylation Patterns for Mouse RNAs. *Genome Res.*, **2017**, *27* (9), 1589–1596. https://doi.org/10.1101/gr.210666.116.
[471] Safra, M.; Sas-Chen, A.; Nir, R.; Winkler, R.; Nachshon, A.; Bar-Yaacov, D.; Erlacher, M.; Rossmanith, W.; Stern-Ginossar, N.; Schwartz, S. The M1A Landscape on Cytosolic and Mitochondrial MRNA at Single-Base Resolution. *Nature*, **2017**, *551* (7679), 251–255. https://doi.org/10.1038/nature24456.
[472] Cohn, W. E. Pseudouridine, a Carbon-Carbon Linked Ribonucleoside in Ribonucleic Acids: Isolation, Structure, and Chemical Characteristics. *J. Biol. Chem.*, **1960**, *235* (5), 1488–1498.

[473] Spenkuch, F.; Motorin, Y.; Helm, M. Pseudouridine: Still Mysterious, but Never a Fake (Uridine)! *RNA Biol.*, **2014**, *11* (12), 1540–1554. https://doi.org/10.4161/15476286.2014.992278.
[474] Bakin, A.; Lane, B. G.; Ofengand, J. Clustering of Pseudouridine Residues around the Peptidyltransferase Center of Yeast Cytoplasmic and Mitochondrial Ribosomes. *Biochemistry*, **1994**, *33* (45), 13475–13483.
[475] Massenet, S.; Motorin, Y.; Lafontaine, D. L.; Hurt, E. C.; Grosjean, H.; Branlant, C. Pseudouridine Mapping in the Saccharomyces Cerevisiae Spliceosomal U Small Nuclear RNAs (SnRNAs) Reveals That Pseudouridine Synthase Pus1p Exhibits a Dual Substrate Specificity for U2 SnRNA and TRNA. *Mol. Cell. Biol.*, **1999**, *19* (3), 2142–2154.
[476] Carlile, T. M.; Rojas-Duran, M. F.; Gilbert, W. V. Pseudo-Seq: Genome-Wide Detection of Pseudouridine Modifications in RNA. *Methods Enzymol.*, **2015**, *560*, 219–245. https://doi.org/10.1016/bs.mie.2015.03.011.
[477] Behm-Ansmant, I.; Urban, A.; Ma, X.; Yu, Y.-T.; Motorin, Y.; Branlant, C. The Saccharomyces Cerevisiae U2 SnRNA:Pseudouridine-Synthase Pus7p Is a Novel Multisite-Multisubstrate RNA:Psi-Synthase Also Acting on TRNAs. *RNA N. Y. N*, **2003**, *9* (11), 1371–1382.
[478] Massenet, S.; Ansmant, I.; Motorin, Y.; Branlant, C. The First Determination of Pseudouridine Residues in 23S Ribosomal RNA from Hyperthermophilic Archaea Sulfolobus Acidocaldarius. *FEBS Lett.*, **1999**, *462* (1–2), 94–100.
[479] Ofengand, J. The Function of Pseudouridylic Acid in Transfer Ribonucleic Acid. I. The Specific Cyanoethylation of Pseudouridine, Inosine, and 4-Thiouridine by Acrylonitrile. *J. Biol. Chem.*, **1967**, *242* (21), 5034–5045.
[480] Mengel-Jørgensen, J.; Kirpekar, F. Detection of Pseudouridine and Other Modifications in TRNA by Cyanoethylation and MALDI Mass Spectrometry. *Nucleic Acids Res.*, **2002**, *30* (23), e135. https://doi.org/10.1093/nar/gnf135.
[481] Emmerechts, G.; Herdewijn, P.; Rozenski, J. Pseudouridine Detection Improvement by Derivatization with Methyl Vinyl Sulfone and Capillary HPLC-Mass Spectrometry. *J. Chromatogr. B Analyt. Technol. Biomed. Life. Sci.*, **2005**, *825* (2), 233–238. https://doi.org/10.1016/j.jchromb.2005.06.041.
[482] Lafontaine, D. L.; Bousquet-Antonelli, C.; Henry, Y.; Caizergues-Ferrer, M.; Tollervey, D. The Box H + ACA SnoRNAs Carry Cbf5p, the Putative RRNA Pseudouridine Synthase. *Genes Dev.*, **1998**, *12* (4), 527–537. https://doi.org/10.1101/gad.12.4.527.
[483] Zebarjadian, Y.; King, T.; Fournier, M. J.; Clarke, L.; Carbon, J. Point Mutations in Yeast CBF5 Can Abolish in Vivo Pseudouridylation of RRNA. *Mol. Cell. Biol.*, **1999**, *19* (11), 7461–7472. https://doi.org/10.1128/mcb.19.11.7461.
[484] Jack, K.; Bellodi, C.; Landry, D. M.; Niederer, R. O.; Meskauskas, A.; Musalgaonkar, S.; Kopmar, N.; Krasnykh, O.; Dean, A. M.; Thompson, S. R.; et al. RRNA Pseudouridylation Defects Affect Ribosomal Ligand Binding and Translational Fidelity from Yeast to Human Cells. *Mol. Cell*, **2011**, *44* (4), 660–666. https://doi.org/10.1016/j.molcel.2011.09.017.
[485] Schmitt, M. E.; Brown, T. A.; Trumpower, B. L. A Rapid and Simple Method for Preparation of RNA from Saccharomyces Cerevisiae. *Nucleic Acids Res.*, **1990**, *18* (10), 3091–3092.
[486] Collart, M. A.; Oliviero, S. Preparation of Yeast RNA. *Curr. Protoc. Mol. Biol.*, **2001**, *Chapter 13*, Unit13.12. https://doi.org/10.1002/0471142727.mb1312s23.

[487] Henrionnet, C.; Gillet, P.; Mainard, D.; Vincourt, J.-B.; Pinzano, A. Label-Free Relative Quantification of Secreted Proteins as a Non-Invasive Method for the Quality Control of Chondrogenesis in Bioengineered Substitutes for Cartilage Repair. *J. Tissue Eng. Regen. Med.*, **2018**, *12* (3), e1757–e1766. https://doi.org/10.1002/term.2454.

[488] Peattie, D. A. Direct Chemical Method for Sequencing RNA. *Proc. Natl. Acad. Sci. U. S. A.*, **1979**, *76* (4), 1760–1764. https://doi.org/10.1073/pnas.76.4.1760.

[489] Küpfer, P. A.; Leumann, C. J. The Chemical Stability of Abasic RNA Compared to Abasic DNA. *Nucleic Acids Res.*, **2007**, *35* (1), 58–68. https://doi.org/10.1093/nar/gkl948.

[490] Wang, L. K.; Shuman, S. Mutational Analysis Defines the 5'-Kinase and 3'-Phosphatase Active Sites of T4 Polynucleotide Kinase. *Nucleic Acids Res.*, **2002**, *30* (4), 1073–1080. https://doi.org/10.1093/nar/30.4.1073.

[491] Taoka, M.; Nobe, Y.; Yamaki, Y.; Yamauchi, Y.; Ishikawa, H.; Takahashi, N.; Nakayama, H.; Isobe, T. The Complete Chemical Structure of Saccharomyces Cerevisiae RRNA: Partial Pseudouridylation of U2345 in 25S RRNA by SnoRNA SnR9. *Nucleic Acids Res.*, **2016**, *44* (18), 8951–8961. https://doi.org/10.1093/nar/gkw564.

[492] Decatur, W. A.; Schnare, M. N. Different Mechanisms for Pseudouridine Formation in Yeast 5S and 5.8S RRNAs. *Mol. Cell. Biol.*, **2008**, *28* (10), 3089–3100. https://doi.org/10.1128/MCB.01574-07.

[493] Ma, X.; Yang, C.; Alexandrov, A.; Grayhack, E. J.; Behm-Ansmant, I.; Yu, Y.-T. Pseudouridylation of Yeast U2 SnRNA Is Catalyzed by Either an RNA-Guided or RNA-Independent Mechanism. *EMBO J.*, **2005**, *24* (13), 2403–2413. https://doi.org/10.1038/sj.emboj.7600718.

[494] Rintala-Dempsey, A. C.; Kothe, U. Eukaryotic Stand-Alone Pseudouridine Synthases - RNA Modifying Enzymes and Emerging Regulators of Gene Expression? *RNA Biol.*, **2017**, *14* (9), 1185–1196. https://doi.org/10.1080/15476286.2016.1276150.

[495] Carlile, T. M.; Martinez, N. M.; Schaening, C.; Su, A.; Bell, T. A.; Zinshteyn, B.; Gilbert, W. V. MRNA Structure Determines Modification by Pseudouridine Synthase 1. *Nat. Chem. Biol.*, **2019**, *15* (10), 966–974. https://doi.org/10.1038/s41589-019-0353-z.

[496] Erales, J.; Marchand, V.; Panthu, B.; Gillot, S.; Belin, S.; Ghayad, S. E.; Garcia, M.; Laforêts, F.; Marcel, V.; Baudin-Baillieu, A.; et al. Evidence for RRNA 2'-O-Methylation Plasticity: Control of Intrinsic Translational Capabilities of Human Ribosomes. *Proc. Natl. Acad. Sci. U. S. A.*, **2017**, *114* (49), 12934–12939. https://doi.org/10.1073/pnas.1707674114.

[497] Sharma, S.; Marchand, V.; Motorin, Y.; Lafontaine, D. L. J. Identification of Sites of 2'-O-Methylation Vulnerability in Human Ribosomal RNAs by Systematic Mapping. *Sci. Rep.*, **2017**, *7* (1), 11490. https://doi.org/10.1038/s41598-017-09734-9.

[498] Foster, L. J.; Zeemann, P. A.; Li, C.; Mann, M.; Jensen, O. N.; Kassem, M. Differential Expression Profiling of Membrane Proteins by Quantitative Proteomics in a Human Mesenchymal Stem Cell Line Undergoing Osteoblast Differentiation. *Stem Cells Dayt. Ohio*, **2005**, *23* (9), 1367–1377. https://doi.org/10.1634/stemcells.2004-0372.

[499] Billing, A. M.; Ben Hamidane, H.; Dib, S. S.; Cotton, R. J.; Bhagwat, A. M.; Kumar, P.; Hayat, S.; Yousri, N. A.; Goswami, N.; Suhre, K.; et al.

Comprehensive Transcriptomic and Proteomic Characterization of Human Mesenchymal Stem Cells Reveals Source Specific Cellular Markers. *Sci. Rep.*, **2016**, *6*, 21507. https://doi.org/10.1038/srep21507.

[500] Mueller, M. B.; Fischer, M.; Zellner, J.; Berner, A.; Dienstknecht, T.; Prantl, L.; Kujat, R.; Nerlich, M.; Tuan, R. S.; Angele, P. Hypertrophy in Mesenchymal Stem Cell Chondrogenesis: Effect of TGF-Beta Isoforms and Chondrogenic Conditioning. *Cells Tissues Organs*, **2010**, *192* (3), 158–166. https://doi.org/10.1159/000313399.

[501] Cicione, C.; Muiños-López, E.; Hermida-Gómez, T.; Fuentes-Boquete, I.; Díaz-Prado, S.; Blanco, F. J. Alternative Protocols to Induce Chondrogenic Differentiation: Transforming Growth Factor-β Superfamily. *Cell Tissue Bank.*, **2015**, *16* (2), 195–207. https://doi.org/10.1007/s10561-014-9472-7.

[502] Courtes, F. C.; Gu, C.; Wong, N. S. C.; Dedon, P. C.; Yap, M. G. S.; Lee, D.-Y. 28S RRNA Is Inducibly Pseudouridylated by the MTOR Pathway Translational Control in CHO Cell Cultures. *J. Biotechnol.*, **2014**, *174*, 16–21. https://doi.org/10.1016/j.jbiotec.2014.01.024.

[503] Xue, S.; Barna, M. Specialized Ribosomes: A New Frontier in Gene Regulation and Organismal Biology. *Nat. Rev. Mol. Cell Biol.*, **2012**, *13* (6), 355–369. https://doi.org/10.1038/nrm3359.

[504] Guo, H. Specialized Ribosomes and the Control of Translation. *Biochem. Soc. Trans.*, **2018**, *46* (4), 855–869. https://doi.org/10.1042/BST20160426.

[505] Yu, Y.-T.; Meier, U. T. RNA-Guided Isomerization of Uridine to Pseudouridine--Pseudouridylation. *RNA Biol.*, **2014**, *11* (12), 1483–1494. https://doi.org/10.4161/15476286.2014.972855.

[506] Tortoriello, G.; de Celis, J. F.; Furia, M. Linking Pseudouridine Synthases to Growth, Development and Cell Competition. *FEBS J.*, **2010**, *277* (15), 3249–3263. https://doi.org/10.1111/j.1742-4658.2010.07731.x.

[507] McMahon, M.; Contreras, A.; Ruggero, D. Small RNAs with Big Implications: New Insights into H/ACA SnoRNA Function and Their Role in Human Disease. *Wiley Interdiscip. Rev. RNA*, **2015**, *6* (2), 173–189. https://doi.org/10.1002/wrna.1266.

[508] McMahon, M.; Contreras, A.; Holm, M.; Uechi, T.; Forester, C. M.; Pang, X.; Jackson, C.; Calvert, M. E.; Chen, B.; Quigley, D. A.; et al. A Single H/ACA Small Nucleolar RNA Mediates Tumor Suppression Downstream of Oncogenic RAS. *eLife*, **2019**, *8*. https://doi.org/10.7554/eLife.48847.

[509] Pichot, F.; Marchand, V.; Ayadi, L.; Bourguignon-Igel, V.; Helm, M.; Motorin, Y. Holistic Optimization of Bioinformatic Analysis Pipeline for Detection and Quantification of 2'-O-Methylations in RNA by RiboMethSeq. *Front. Genet.*, **2020**, *11*, 38. https://doi.org/10.3389/fgene.2020.00038.

[510] Adachi, H.; De Zoysa, M. D.; Yu, Y.-T. Post-Transcriptional Pseudouridylation in MRNA as Well as in Some Major Types of Noncoding RNAs. *Biochim. Biophys. Acta Gene Regul. Mech.*, **2019**, *1862* (3), 230–239. https://doi.org/10.1016/j.bbagrm.2018.11.002.

[511] McCown, P. J.; Ruszkowska, A.; Kunkler, C. N.; Breger, K.; Hulewicz, J. P.; Wang, M. C.; Springer, N. A.; Brown, J. A. Naturally Occurring Modified Ribonucleosides. *Wiley Interdiscip. Rev. RNA*, **2020**, *11* (5), e1595. https://doi.org/10.1002/wrna.1595.

[512] Morais, P.; Adachi, H.; Yu, Y.-T. Spliceosomal SnRNA Epitranscriptomics. *Front. Genet.*, **2021**, *12*, 652129. https://doi.org/10.3389/fgene.2021.652129.

[513] Westhof, E. Pseudouridines or How to Draw on Weak Energy Differences. *Biochem. Biophys. Res. Commun.*, **2019**, *520* (4), 702–704. https://doi.org/10.1016/j.bbrc.2019.10.009.
[514] Adachi, H.; DeZoysa, M. D.; Yu, Y.-T. Detection and Quantification of Pseudouridine in RNA. *Methods Mol. Biol. Clifton NJ*, **2019**, *1870*, 219–235. https://doi.org/10.1007/978-1-4939-8808-2_17.
[515] Zhou, K. I.; Clark, W. C.; Pan, D. W.; Eckwahl, M. J.; Dai, Q.; Pan, T. Pseudouridines Have Context-Dependent Mutation and Stop Rates in High-Throughput Sequencing. *RNA Biol.*, **2018**, *15* (7), 892–900. https://doi.org/10.1080/15476286.2018.1462654.
[516] Durairaj, A.; Limbach, P. A. Mass Spectrometry of the Fifth Nucleoside: A Review of the Identification of Pseudouridine in Nucleic Acids. *Anal. Chim. Acta*, **2008**, *623* (2), 117–125. https://doi.org/10.1016/j.aca.2008.06.027.
[517] Addepalli, B.; Limbach, P. A. Mass Spectrometry-Based Quantification of Pseudouridine in RNA. *J. Am. Soc. Mass Spectrom.*, **2011**, *22* (8), 1363–1372. https://doi.org/10.1007/s13361-011-0137-5.
[518] Patteson, K. G.; Rodicio, L. P.; Limbach, P. A. Identification of the Mass-Silent Post-Transcriptionally Modified Nucleoside Pseudouridine in RNA by Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *Nucleic Acids Res.*, **2001**, *29* (10), E49-49. https://doi.org/10.1093/nar/29.10.e49.
[519] Helm, M.; Schmidt-Dengler, M. C.; Weber, M.; Motorin, Y. General Principles for the Detection of Modified Nucleotides in RNA by Specific Reagents. *Adv. Biol.*, **2021**, *5* (10), e2100866. https://doi.org/10.1002/adbi.202100866.
[520] Chang, S. E.; Ish-Horowicz, D. Selective Modification of Cytidine, Uridine, Guanosine and Pseudouridine Residues in Escherichia Coli Leucine Transfer Ribonucleic Acid. *J. Mol. Biol.*, **1974**, *84* (3), 375–388. https://doi.org/10.1016/0022-2836(74)90446-x.
[521] Ofengand, J.; Del Campo, M.; Kaya, Y. Mapping Pseudouridines in RNA Molecules. *Methods San Diego Calif*, **2001**, *25* (3), 365–373. https://doi.org/10.1006/meth.2001.1249.
[522] Lovejoy, A. F.; Riordan, D. P.; Brown, P. O. Transcriptome-Wide Mapping of Pseudouridines: Pseudouridine Synthases Modify Specific MRNAs in S. Cerevisiae. *PloS One*, **2014**, *9* (10), e110799. https://doi.org/10.1371/journal.pone.0110799.
[523] Sas-Chen, A.; Nir, R.; Schwartz, S. Mito-Ψ-Seq: A High-Throughput Method for Systematic Mapping of Pseudouridine Within Mitochondrial RNA. *Methods Mol. Biol. Clifton NJ*, **2021**, *2192*, 103–115. https://doi.org/10.1007/978-1-0716-0834-0_9.
[524] Zhang, W.; Eckwahl, M. J.; Zhou, K. I.; Pan, T. Sensitive and Quantitative Probing of Pseudouridine Modification in MRNA and Long Noncoding RNA. *RNA N. Y. N*, **2019**, *25* (9), 1218–1225. https://doi.org/10.1261/rna.072124.119.
[525] Nachtergaele, S.; He, C. Chemical Modifications in the Life of an MRNA Transcript. *Annu. Rev. Genet.*, **2018**, *52* (1), 349–372. https://doi.org/10.1146/annurev-genet-120417-031522.
[526] Zaccara, S.; Ries, R. J.; Jaffrey, S. R. Reading, Writing and Erasing MRNA Methylation. *Nat. Rev. Mol. Cell Biol.*, **2019**, *20* (10), 608–624. https://doi.org/10.1038/s41580-019-0168-5.

[527] Huang, H.; Weng, H.; Chen, J. M6A Modification in Coding and Non-Coding RNAs: Roles and Therapeutic Implications in Cancer. *Cancer Cell*, **2020**, *37* (3), 270–288. https://doi.org/10.1016/j.ccell.2020.02.004.
[528] Xiao, W.; Adhikari, S.; Dahal, U.; Chen, Y.-S.; Hao, Y.-J.; Sun, B.-F.; Sun, H.-Y.; Li, A.; Ping, X.-L.; Lai, W.-Y.; et al. Nuclear M6A Reader YTHDC1 Regulates MRNA Splicing. *Mol. Cell*, **2016**, *61* (4), 507–519. https://doi.org/10.1016/j.molcel.2016.01.012.
[529] Fustin, J.-M.; Doi, M.; Yamaguchi, Y.; Hida, H.; Nishimura, S.; Yoshida, M.; Isagawa, T.; Morioka, M. S.; Kakeya, H.; Manabe, I.; et al. RNA-Methylation-Dependent RNA Processing Controls the Speed of the Circadian Clock. *Cell*, **2013**, *155* (4), 793–806. https://doi.org/10.1016/j.cell.2013.10.026.
[530] Zheng, G.; Dahl, J. A.; Niu, Y.; Fedorcsak, P.; Huang, C.-M.; Li, C. J.; Vågbø, C. B.; Shi, Y.; Wang, W.-L.; Song, S.-H.; et al. ALKBH5 Is a Mammalian RNA Demethylase That Impacts RNA Metabolism and Mouse Fertility. *Mol. Cell*, **2013**, *49* (1), 18–29. https://doi.org/10.1016/j.molcel.2012.10.015.
[531] Wang, X.; Lu, Z.; Gomez, A.; Hon, G. C.; Yue, Y.; Han, D.; Fu, Y.; Parisien, M.; Dai, Q.; Jia, G.; et al. N 6 -Methyladenosine-Dependent Regulation of Messenger RNA Stability. *Nature*, **2014**, *505* (7481), 117–120. https://doi.org/10.1038/nature12730.
[532] Wang, X.; Zhao, B. S.; Roundtree, I. A.; Lu, Z.; Han, D.; Ma, H.; Weng, X.; Chen, K.; Shi, H.; He, C. N6-Methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell*, **2015**, *161* (6), 1388–1399. https://doi.org/10.1016/j.cell.2015.05.014.
[533] Angelova, M. T.; Dimitrova, D. G.; Dinges, N.; Lence, T.; Worpenberg, L.; Carré, C.; Roignant, J.-Y. The Emerging Field of Epitranscriptomics in Neurodevelopmental and Neuronal Disorders. *Front. Bioeng. Biotechnol.*, **2018**, *6*. https://doi.org/10.3389/fbioe.2018.00046.
[534] Deng, X.; Su, R.; Weng, H.; Huang, H.; Li, Z.; Chen, J. RNA N6-Methyladenosine Modification in Cancers: Current Status and Perspectives. *Cell Res.*, **2018**, *28* (5), 507–517. https://doi.org/10.1038/s41422-018-0034-6.
[535] Klungland, A.; Dahl, J. A. Dynamic RNA Modifications in Disease. *Curr. Opin. Genet. Dev.*, **2014**, *26*, 47–52. https://doi.org/10.1016/j.gde.2014.05.006.
[536] Liu, Z.-X.; Li, L.-M.; Sun, H.-L.; Liu, S.-M. Link Between M6A Modification and Cancers. *Front. Bioeng. Biotechnol.*, **2018**, *6*. https://doi.org/10.3389/fbioe.2018.00089.
[537] Kennedy, E. M.; Cullen, B. R. Gene Editing: A New Tool for Viral Disease. *Annu. Rev. Med.*, **2017**, *68* (1), 401–411. https://doi.org/10.1146/annurev-med-051215-031129.
[538] Tan, B.; Gao, S.-J. The RNA Epitranscriptome of DNA Viruses. *J. Virol.*, **2018**, *92* (22). https://doi.org/10.1128/JVI.00696-18.
[539] Chen, K.; Lu, Z.; Wang, X.; Fu, Y.; Luo, G.-Z.; Liu, N.; Han, D.; Dominissini, D.; Dai, Q.; Pan, T.; et al. High-Resolution N6-Methyladenosine (M6A) Map Using Photo-Crosslinking-Assisted M6A Sequencing. *Angew. Chem. Int. Ed.*, **2015**, *54* (5), 1587–1590. https://doi.org/10.1002/anie.201410647.
[540] Linder, B.; Grozhik, A. V.; Olarerin-George, A. O.; Meydan, C.; Mason, C. E.; Jaffrey, S. R. Single-Nucleotide-Resolution Mapping of M6A and M6Am throughout the Transcriptome. *Nat. Methods*, **2015**, *12* (8), 767–772. https://doi.org/10.1038/nmeth.3453.

[541] Molinie, B.; Wang, J.; Lim, K. S.; Hillebrand, R.; Lu, Z.-X.; Van Wittenberghe, N.; Howard, B. D.; Daneshvar, K.; Mullen, A. C.; Dedon, P.; et al. M(6)A-LAIC-Seq Reveals the Census and Complexity of the m(6)A Epitranscriptome. *Nat. Methods*, **2016**, *13* (8), 692–698. https://doi.org/10.1038/nmeth.3898.

[542] Aviv, H.; Leder, P. Purification of Biologically Active Globin Messenger RNA by Chromatography on Oligothymidylic Acid-Cellulose. *Proc. Natl. Acad. Sci.*, **1972**, *69* (6), 1408–1412. https://doi.org/10.1073/pnas.69.6.1408.

[543] Edmonds, M.; Vaughan, M. H.; Nakazato, H. Polyadenylic Acid Sequences in the Heterogeneous Nuclear RNA and Rapidly-Labeled Polyribosomal RNA of HeLa Cells: Possible Evidence for a Precursor Relationship. *Proc. Natl. Acad. Sci.*, **1971**, *68* (6), 1336–1340. https://doi.org/10.1073/pnas.68.6.1336.

[544] Lindberg, U.; Persson, T. Isolation of MRNA from KB-Cells by Affinity Chromatography on Polyuridylic Acid Covalently Linked to Sepharose. *Eur. J. Biochem.*, **1972**, *31* (2), 246–254. https://doi.org/10.1111/j.1432-1033.1972.tb02527.x.

[545] Horowitz, S.; Horowitz, A.; Nilsen, T. W.; Munns, T. W.; Rottman, F. M. Mapping of N6-Methyladenosine Residues in Bovine Prolactin MRNA. *Proc. Natl. Acad. Sci. U. S. A.*, **1984**, *81* (18), 5667–5671.

[546] Munns, T. W.; Liszewski, M. K.; Sims, H. F. Characterization of Antibodies Specific for N6-Methyladenosine and for 7-Methylguanosine. *Biochemistry*, **1977**, *16* (10), 2163–2168. https://doi.org/10.1021/bi00629a019.

[547] Dominissini, D.; Moshitch-Moshkovitz, S.; Schwartz, S.; Salmon-Divon, M.; Ungar, L.; Osenberg, S.; Cesarkas, K.; Jacob-Hirsch, J.; Amariglio, N.; Kupiec, M.; et al. Topology of the Human and Mouse M6A RNA Methylomes Revealed by M6A-Seq. *Nature*, **2012**, *485* (7397), 201–206. https://doi.org/10.1038/nature11112.

[548] Meyer, K. D.; Saletore, Y.; Zumbo, P.; Elemento, O.; Mason, C. E.; Jaffrey, S. R. Comprehensive Analysis of MRNA Methylation Reveals Enrichment in 3' UTRs and near Stop Codons. *Cell*, **2012**, *149* (7), 1635–1646. https://doi.org/10.1016/j.cell.2012.05.003.

[549] Saletore, Y.; Meyer, K.; Korlach, J.; Vilfan, I. D.; Jaffrey, S.; Mason, C. E. The Birth of the Epitranscriptome: Deciphering the Function of RNA Modifications. *Genome Biol.*, **2012**, *13* (10), 175. https://doi.org/10.1186/gb-2012-13-10-175.

[550] Linder, B.; Jaffrey, S. R. Discovering and Mapping the Modified Nucleotides That Comprise the Epitranscriptome of MRNA. *Cold Spring Harb. Perspect. Biol.*, **2019**, *11* (6). https://doi.org/10.1101/cshperspect.a032201.

[551] Carell, T.; Brandmayr, C.; Hienzsch, A.; Müller, M.; Pearson, D.; Reiter, V.; Thoma, I.; Thumbs, P.; Wagner, M. Structure and Function of Noncanonical Nucleobases. *Angew. Chem. Int. Ed Engl.*, **2012**, *51* (29), 7110–7131. https://doi.org/10.1002/anie.201201193.

[552] Sednev, M. V.; Mykhailiuk, V.; Choudhury, P.; Halang, J.; Sloan, K. E.; Bohnsack, M. T.; Höbartner, C. N6 -Methyladenosine-Sensitive RNA-Cleaving Deoxyribozymes. *Angew. Chem. Int. Ed Engl.*, **2018**, *57* (46), 15117–15121. https://doi.org/10.1002/anie.201808745.

[553] Hong, T.; Yuan, Y.; Chen, Z.; Xi, K.; Wang, T.; Xie, Y.; He, Z.; Su, H.; Zhou, Y.; Tan, Z.-J.; et al. Precise Antibody-Independent M6A Identification via 4SedTTP-Involved and FTO-Assisted Strategy at Single-Nucleotide Resolution. *J. Am. Chem. Soc.*, **2018**, *140* (18), 5886–5889. https://doi.org/10.1021/jacs.7b13633.

[554] Shu, X.; Dai, Q.; Wu, T.; Bothwell, I. R.; Yue, Y.; Zhang, Z.; Cao, J.; Fei, Q.; Luo, M.; He, C.; et al. N6-Allyladenosine: A New Small Molecule for RNA Labeling Identified by Mutation Assay. *J. Am. Chem. Soc.*, **2017**, *139* (48), 17213–17216. https://doi.org/10.1021/jacs.7b06837.

[555] Aschenbrenner, J.; Werner, S.; Marchand, V.; Adam, M.; Motorin, Y.; Helm, M.; Marx, A. Engineering of a DNA Polymerase for Direct M6A Sequencing. *Angew. Chem. Int. Ed.*, **2018**, *57* (2), 417–421. https://doi.org/10.1002/anie.201710209.

[556] Hartstock, K.; Nilges, B. S.; Ovcharenko, A.; Cornelissen, N. V.; Püllen, N.; Lawrence-Dörner, A.-M.; Leidel, S. A.; Rentmeister, A. Enzymatic or In Vivo Installation of Propargyl Groups in Combination with Click Chemistry for the Enrichment and Detection of Methyltransferase Target Sites in RNA. *Angew. Chem. Int. Ed.*, **2018**, *57* (21), 6342–6346. https://doi.org/10.1002/anie.201800188.

[557] Liu, H.; Begik, O.; Lucas, M. C.; Ramirez, J. M.; Mason, C. E.; Wiener, D.; Schwartz, S.; Mattick, J. S.; Smith, M. A.; Novoa, E. M. Accurate Detection of M6A RNA Modifications in Native RNA Sequences. *Nat. Commun.*, **2019**, *10* (1), 4079. https://doi.org/10.1038/s41467-019-11713-9.

[558] Garcia-Campos, M. A.; Edelheit, S.; Toth, U.; Safra, M.; Shachar, R.; Viukov, S.; Winkler, R.; Nir, R.; Lasman, L.; Brandis, A.; et al. Deciphering the “M6A Code” via Antibody-Independent Quantitative Profiling. *Cell*, **2019**, *178* (3), 731-747.e16. https://doi.org/10.1016/j.cell.2019.06.013.

[559] Zhang, Z.; Chen, L.-Q.; Zhao, Y.-L.; Yang, C.-G.; Roundtree, I. A.; Zhang, Z.; Ren, J.; Xie, W.; He, C.; Luo, G.-Z. Single-Base Mapping of M6A by an Antibody-Independent Method. *Sci. Adv.*, **2019**, *5* (7). https://doi.org/10.1126/sciadv.aax0250.

[560] Koh, C. W. Q.; Goh, Y. T.; Goh, W. S. S. Atlas of Quantitative Single-Base-Resolution N 6 -Methyl-Adenine Methylomes. *Nat. Commun.*, **2019**, *10* (1), 1–15. https://doi.org/10.1038/s41467-019-13561-z.

[561] Meyer, K. D. DART-Seq: An Antibody-Free Method for Global m 6 A Detection. *Nat. Methods*, **2019**, *16* (12), 1275–1280. https://doi.org/10.1038/s41592-019-0570-0.

[562] Worpenberg, L.; Jakobi, T.; Dieterich, C.; Roignant, J.-Y. Identification of Methylated Transcripts Using the TRIBE Approach. *Methods Mol. Biol. Clifton NJ*, **2019**, *1870*, 89–106. https://doi.org/10.1007/978-1-4939-8808-2_7.

[563] Legrand, C.; Tuorto, F.; Hartmann, M.; Liebers, R.; Jacob, D.; Helm, M.; Lyko, F. Statistically Robust Methylation Calling for Whole-Transcriptome Bisulfite Sequencing Reveals Distinct Methylation Patterns for Mouse RNAs. *Genome Res.*, **2017**, *27* (9), 1589–1596. https://doi.org/10.1101/gr.210666.116.

[564] Schaefer, M.; Pollex, T.; Hanna, K.; Lyko, F. RNA Cytosine Methylation Analysis by Bisulfite Sequencing. *Nucleic Acids Res.*, **2009**, *37* (2), e12. https://doi.org/10.1093/nar/gkn954.

[565] Tuorto, F.; Liebers, R.; Musch, T.; Schaefer, M.; Hofmann, S.; Kellner, S.; Frye, M.; Helm, M.; Stoecklin, G.; Lyko, F. RNA Cytosine Methylation by Dnmt2 and NSun2 Promotes TRNA Stability and Protein Synthesis. *Nat. Struct. Mol. Biol.*, **2012**, *19* (9), 900–905. https://doi.org/10.1038/nsmb.2357.

[566] Shapiro, Robert.; Pohl, S. H. Reaction of Ribonucleosides with Nitrous Acid. Side Products and Kinetics. *Biochemistry*, **1968**, *7* (1), 448–455. https://doi.org/10.1021/bi00841a057.

[567] Shapiro, R.; Shiuey, S.-J. Reaction of Nitrous Acid with Alkylaminopurines. *Biochim. Biophys. Acta BBA - Nucleic Acids Protein Synth.*, **1969**, *174* (1), 403–405. https://doi.org/10.1016/0005-2787(69)90268-8.
[568] Safra, M.; Sas-Chen, A.; Nir, R.; Winkler, R.; Nachshon, A.; Bar-Yaacov, D.; Erlacher, M.; Rossmanith, W.; Stern-Ginossar, N.; Schwartz, S. The M1A Landscape on Cytosolic and Mitochondrial MRNA at Single-Base Resolution. *Nature*, **2017**, *551* (7679), 251–255. https://doi.org/10.1038/nature24456.
[569] Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.*, **1990**, *215* (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
[570] Ye, H.; Meehan, J.; Tong, W.; Hong, H. Alignment of Short Reads: A Crucial Step for Application of Next-Generation Sequencing Data in Precision Medicine. *Pharmaceutics*, **2015**, *7* (4), 523–541. https://doi.org/10.3390/pharmaceutics7040523.
[571] Li, R.; Li, Y.; Kristiansen, K.; Wang, J. SOAP: Short Oligonucleotide Alignment Program. *Bioinformatics*, **2008**, *24* (5), 713–714. https://doi.org/10.1093/bioinformatics/btn025.
[572] Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S. L. Ultrafast and Memory-Efficient Alignment of Short DNA Sequences to the Human Genome. *Genome Biol.*, **2009**, *10* (3), R25. https://doi.org/10.1186/gb-2009-10-3-r25.
[573] Xi, Y.; Li, W. BSMAP: Whole Genome Bisulfite Sequence MAPping Program. *BMC Bioinformatics*, **2009**, *10* (1), 232. https://doi.org/10.1186/1471-2105-10-232.
[574] R.Bermes, K. Schmeidl. U.S. Patent 3 792 077. February 12, 1974.
[575] Sutandy, F. R.; Hildebrandt, A.; König, J. Profiling the Binding Sites of RNA-Binding Proteins with Nucleotide Resolution Using ICLIP. In *Post-Transcriptional Gene Regulation*; Dassi, E., Ed.; Methods in Molecular Biology; Springer: New York, NY, 2016; pp 175–195. https://doi.org/10.1007/978-1-4939-3067-8_11.
[576] Kulikowska, E.; Kierdaszuk, B.; Shugar, D. Xanthine, Xanthosine and Its Nucleotides: Solution Structures of Neutral and Ionic Forms, and Relevance to Substrate Properties in Various Enzyme Systems and Metabolic Pathways. *Acta Biochim. Pol.*, **2004**, *51* (2), 493–531. https://doi.org/10.18388/abp.2004_3587.
[577] Slama, K.; Galliot, A.; Weichmann, F.; Hertler, J.; Feederle, R.; Meister, G.; Helm, M. Determination of Enrichment Factors for Modified RNA in MeRIP Experiments. *Methods*, **2019**, *156*, 102–109. https://doi.org/10.1016/j.ymeth.2018.10.020.
[578] Liu, N.; Parisien, M.; Dai, Q.; Zheng, G.; He, C.; Pan, T. Probing N6-Methyladenosine RNA Modification Status at Single Nucleotide Resolution in MRNA and Long Noncoding RNA. *RNA N. Y. N*, **2013**, *19* (12), 1848–1856. https://doi.org/10.1261/rna.041178.113.
[579] Maier, D. Hairless: The Ignored Antagonist of the Notch Signalling Pathway. *Hereditas*, **2006**, *143* (2006), 212–221. https://doi.org/10.1111/j.2007.0018-0661.01971.x.
[580] Prehn, J. H. M.; Jirström, E. Angiogenin and TRNA Fragments in Parkinson's Disease and Neurodegeneration. *Acta Pharmacol. Sin.*, **2020**, *41* (4), 442–446. https://doi.org/10.1038/s41401-020-0375-9.
[581] Kishikawa, H.; Wu, D.; Hu, G. Targeting Angiogenin in Therapy of Amyotropic Lateral Sclerosis. *Expert Opin. Ther. Targets*, **2008**, *12* (10), 1229–1242. https://doi.org/10.1517/14728222.12.10.1229.

[582] Li, G.; Manning, A. C.; Bagi, A.; Yang, X.; Gokulnath, P.; Spanos, M.; Howard, J.; Chan, P. P.; Sweeney, T.; Kitchen, R.; et al. Distinct Stress-Dependent Signatures of Cellular and Extracellular TRNA-Derived Small RNAs. *Adv. Sci. Weinh. Baden-Wurtt. Ger.*, **2022**, *9* (17), e2200829. https://doi.org/10.1002/advs.202200829.

[583] Sanadgol, N.; König, L.; Drino, A.; Jovic, M.; Schaefer, M. R. Experimental Paradigms Revisited: Oxidative Stress-Induced TRNA Fragmentation Does Not Correlate with Stress Granule Formation but Is Associated with Delayed Cell Death. *Nucleic Acids Res.*, **2022**, *50* (12), 6919–6937. https://doi.org/10.1093/nar/gkac495.

[584] Rashad, S.; Tominaga, T.; Niizuma, K. The Cell and Stress-Specific Canonical and Noncanonical TRNA Cleavage. *J. Cell. Physiol.*, **2021**, *236* (5), 3710–3724. https://doi.org/10.1002/jcp.30107.

[585] Schaefer, M.; Pollex, T.; Hanna, K.; Tuorto, F.; Meusburger, M.; Helm, M.; Lyko, F. RNA Methylation by Dnmt2 Protects Transfer RNAs against Stress-Induced Cleavage. *Genes Dev.*, **2010**, *24* (15), 1590–1595. https://doi.org/10.1101/gad.586710.

[586] Blanco, S.; Dietmann, S.; Flores, J. V.; Hussain, S.; Kutter, C.; Humphreys, P.; Lukk, M.; Lombard, P.; Treps, L.; Popis, M.; et al. Aberrant Methylation of TRNAs Links Cellular Stress to Neuro-Developmental Disorders. *EMBO J.*, **2014**, *33* (18), 2020–2039. https://doi.org/10.15252/embj.201489282.

[587] Rashad, S.; Han, X.; Sato, K.; Mishima, E.; Abe, T.; Tominaga, T.; Niizuma, K. The Stress Specific Impact of ALKBH1 on TRNA Cleavage and TiRNA Generation. *RNA Biol.*, **2020**, *17* (8), 1092–1103. https://doi.org/10.1080/15476286.2020.1779492.

[588] Fagan, S. G.; Helm, M.; Prehn, J. H. M. TRNA-Derived Fragments: A New Class of Non-Coding RNA with Key Roles in Nervous System Function and Dysfunction. *Prog. Neurobiol.*, **2021**, *205*, 102118. https://doi.org/10.1016/j.pneurobio.2021.102118.

[589] Lyons, S. M.; Fay, M. M.; Ivanov, P. The Role of RNA Modifications in the Regulation of TRNA Cleavage. *FEBS Lett.*, **2018**, *592* (17), 2828–2844. https://doi.org/10.1002/1873-3468.13205.

[590] Studte, P.; Zink, S.; Jablonowski, D.; Bär, C.; von der Haar, T.; Tuite, M. F.; Schaffrath, R. TRNA and Protein Methylase Complexes Mediate Zymocin Toxicity in Yeast. *Mol. Microbiol.*, **2008**, *69* (5), 1266–1277. https://doi.org/10.1111/j.1365-2958.2008.06358.x.

[591] Morita, Y.; Shibutani, T.; Nakanishi, N.; Nishikura, K.; Iwai, S.; Kuraoka, I. Human Endonuclease V Is a Ribonuclease Specific for Inosine-Containing RNA. *Nat. Commun.*, **2013**, *4*, 2273. https://doi.org/10.1038/ncomms3273.

[592] Vik, E. S.; Nawaz, M. S.; Strøm Andersen, P.; Fladeby, C.; Bjørås, M.; Dalhus, B.; Alseth, I. Endonuclease V Cleaves at Inosines in RNA. *Nat. Commun.*, **2013**, *4*, 2271. https://doi.org/10.1038/ncomms3271.

[593] Helm, M. Post-Transcriptional Nucleotide Modification and Alternative Folding of RNA. *Nucleic Acids Res.*, **2006**, *34* (2), 721–733. https://doi.org/10.1093/nar/gkj471.

[594] Fricker, R.; Brogli, R.; Luidalepp, H.; Wyss, L.; Fasnacht, M.; Joss, O.; Zywicki, M.; Helm, M.; Schneider, A.; Cristodero, M.; et al. A TRNA Half Modulates Translation as Stress Response in Trypanosoma Brucei. *Nat. Commun.*, **2019**, *10* (1), 118. https://doi.org/10.1038/s41467-018-07949-6.

[595] Greenway, M. J.; Andersen, P. M.; Russ, C.; Ennis, S.; Cashman, S.; Donaghy, C.; Patterson, V.; Swingler, R.; Kieran, D.; Prehn, J.; et al. ANG Mutations Segregate with Familial and “sporadic” Amyotrophic Lateral Sclerosis. *Nat. Genet.*, **2006**, *38* (4), 411–413. https://doi.org/10.1038/ng1742.

[596] Kieran, D.; Sebastia, J.; Greenway, M. J.; King, M. A.; Connaughton, D.; Concannon, C. G.; Fenner, B.; Hardiman, O.; Prehn, J. H. M. Control of Motoneuron Survival by Angiogenin. *J. Neurosci. Off. J. Soc. Neurosci.*, **2008**, *28* (52), 14056–14061. https://doi.org/10.1523/JNEUROSCI.3399-08.2008.

[597] Skorupa, A.; King, M. A.; Aparicio, I. M.; Dussmann, H.; Coughlan, K.; Breen, B.; Kieran, D.; Concannon, C. G.; Marin, P.; Prehn, J. H. M. Motoneurons Secrete Angiogenin to Induce RNA Cleavage in Astroglia. *J. Neurosci. Off. J. Soc. Neurosci.*, **2012**, *32* (15), 5024–5038. https://doi.org/10.1523/JNEUROSCI.6366-11.2012.

[598] Akiyama, Y.; Tomioka, Y.; Abe, T.; Anderson, P.; Ivanov, P. In Lysate RNA Digestion Provides Insights into the Angiogenin’s Specificity towards Transfer RNAs. *RNA Biol.*, **2021**, *18* (12), 2546–2555. https://doi.org/10.1080/15476286.2021.1930758.

[599] Akiyama, Y.; Kharel, P.; Abe, T.; Anderson, P.; Ivanov, P. Isolation and Initial Structure-Functional Characterization of Endogenous TRNA-Derived Stress-Induced RNAs. *RNA Biol.*, **2020**, *17* (8), 1116–1124. https://doi.org/10.1080/15476286.2020.1732702.

[600] Pereira, M.; Ribeiro, D. R.; Pinheiro, M. M.; Ferreira, M.; Kellner, S.; Soares, A. R. M5U54 TRNA Hypomodification by Lack of TRMT2A Drives the Generation of TRNA-Derived Small RNAs. *Int. J. Mol. Sci.*, **2021**, *22* (6), 2941. https://doi.org/10.3390/ijms22062941.

[601] Richter, F.; Plehn, J. E.; Bessler, L.; Hertler, J.; Jörg, M.; Cirzi, C.; Tuorto, F.; Friedland, K.; Helm, M. RNA Marker Modifications Reveal the Necessity for Rigorous Preparation Protocols to Avoid Artifacts in Epitranscriptomic Analysis. *Nucleic Acids Res.*, **2022**, *50* (8), 4201–4215. https://doi.org/10.1093/nar/gkab1150.

[602] Pichot, F.; Marchand, V.; Helm, M.; Motorin, Y. Machine Learning Algorithm for Precise Prediction of 2’-O-Methylation (Nm) Sites from Experimental RiboMethSeq Datasets. *Methods San Diego Calif*, **2022**, *203*, 311–321. https://doi.org/10.1016/j.ymeth.2022.03.007.

[603] Hetschko, H.; Voss, V.; Horn, S.; Seifert, V.; Prehn, J. H. M.; Kögel, D. Pharmacological Inhibition of Bcl-2 Family Members Reactivates TRAIL-Induced Apoptosis in Malignant Glioma. *J. Neurooncol.*, **2008**, *86* (3), 265–272. https://doi.org/10.1007/s11060-007-9472-6.

[604] Molla-Herman, A.; Angelova, M. T.; Ginestet, M.; Carré, C.; Antoniewski, C.; Huynh, J.-R. TRNA Fragments Populations Analysis in Mutants Affecting TRNAs Processing and TRNA Methylation. *Front. Genet.*, **2020**, *11*, 518949. https://doi.org/10.3389/fgene.2020.518949.

[605] Marchand, V.; Ayadi, L.; Bourguignon-Igel, V.; Helm, M.; Motorin, Y. AlkAniline-Seq: A Highly Sensitive and Specific Method for Simultaneous Mapping of 7-Methyl-Guanosine (M7G) and 3-Methyl-Cytosine (M3C) in RNAs by High-Throughput Sequencing. *Methods Mol. Biol. Clifton NJ*, **2021**, *2298*, 77–95. https://doi.org/10.1007/978-1-0716-1374-0_5.

[606] Pichot, F.; Marchand, V.; Helm, M.; Motorin, Y. Non-Redundant TRNA Reference Sequences for Deep Sequencing Analysis of TRNA Abundance and

Epitranscriptomic RNA Modifications. *Genes*, **2021**, *12* (1), 81. https://doi.org/10.3390/genes12010081.

[607] Amort, T.; Rieder, D.; Wille, A.; Khokhlova-Cubberley, D.; Riml, C.; Trixl, L.; Jia, X.-Y.; Micura, R.; Lusser, A. Distinct 5-Methylcytosine Profiles in Poly(A) RNA from Mouse Embryonic Stem Cells and Brain. *Genome Biol.*, **2017**, *18* (1), 1. https://doi.org/10.1186/s13059-016-1139-1.

[608] Johnson, Z.; Xu, X.; Pacholec, C.; Xie, H. Systematic Evaluation of Parameters in RNA Bisulfite Sequencing Data Generation and Analysis. *NAR Genomics Bioinforma.*, **2022**, *4* (2), lqac045. https://doi.org/10.1093/nargab/lqac045.

[609] Trinklein, T. J.; Warren, C. G.; Synovec, R. E. Determination of the Signal-To-Noise Ratio Enhancement in Comprehensive Three-Dimensional Gas Chromatography. *Anal. Chem.*, **2021**, *93* (24), 8526–8535. https://doi.org/10.1021/acs.analchem.1c01190.

[610] Trixl, L.; Rieder, D.; Amort, T.; Lusser, A. Bisulfite Sequencing of RNA for Transcriptome-Wide Detection of 5-Methylcytosine. *Methods Mol. Biol. Clifton NJ*, **2019**, *1870*, 1–21. https://doi.org/10.1007/978-1-4939-8808-2_1.

[611] Sprinzl, M.; Hartmann, T.; Meissner, F.; Moll, J.; Vorderwülbecke, T. Compilation of TRNA Sequences and Sequences of TRNA Genes. *Nucleic Acids Res.*, **1987**, *15 Suppl* (Suppl), r53-188. https://doi.org/10.1093/nar/15.suppl.r53.

[612] Bohnsack, K. E.; Höbartner, C.; Bohnsack, M. T. Eukaryotic 5-Methylcytosine (M^5C) RNA Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease. *Genes*, **2019**, *10* (2), 102. https://doi.org/10.3390/genes10020102.

[613] Jeltsch, A.; Ehrenhofer-Murray, A.; Jurkowski, T. P.; Lyko, F.; Reuter, G.; Ankri, S.; Nellen, W.; Schaefer, M.; Helm, M. Mechanism and Biological Role of Dnmt2 in Nucleic Acid Methylation. *RNA Biol.*, **2017**, *14* (9), 1108–1123. https://doi.org/10.1080/15476286.2016.1191737.

[614] Cui, J.; Liu, Q.; Sendinc, E.; Shi, Y.; Gregory, R. I. Nucleotide Resolution Profiling of M3C RNA Modification by HAC-Seq. *Nucleic Acids Res.*, **2021**, *49* (5), e27. https://doi.org/10.1093/nar/gkaa1186.

[615] Dalluge, J. J.; Hashizume, T.; McCloskey, J. A. Quantitative Measurement of Dihydrouridine in RNA Using Isotope Dilution Liquid Chromatography-Mass Spectrometry (LC/MS). *Nucleic Acids Res.*, **1996**, *24* (16), 3242–3245. https://doi.org/10.1093/nar/24.16.3242.

[616] Fay, M. P.; Proschan, M. A. Wilcoxon-Mann-Whitney or t-Test? On Assumptions for Hypothesis Tests and Multiple Interpretations of Decision Rules. *Stat. Surv.*, **2010**, *4*, 1–39. https://doi.org/10.1214/09-SS051.

[617] Lyons, S. M.; Fay, M. M.; Akiyama, Y.; Anderson, P. J.; Ivanov, P. RNA Biology of Angiogenin: Current State and Perspectives. *RNA Biol.*, **2017**, *14* (2), 171–178. https://doi.org/10.1080/15476286.2016.1272746.

[618] Hu, G. F.; Riordan, J. F.; Vallee, B. L. A Putative Angiogenin Receptor in Angiogenin-Responsive Human Endothelial Cells. *Proc. Natl. Acad. Sci. U. S. A.*, **1997**, *94* (6), 2204–2209. https://doi.org/10.1073/pnas.94.6.2204.

[619] Hogg, M. C.; Rayner, M.; Susdalzew, S.; Monsefi, N.; Crivello, M.; Woods, I.; Resler, A.; Blackbourn, L.; Fabbrizio, P.; Trolese, M. C.; et al. 5'ValCAC TRNA Fragment Generated as Part of a Protective Angiogenin Response Provides Prognostic Value in Amyotrophic Lateral Sclerosis. *Brain Commun.*, **2020**, *2* (2), fcaa138. https://doi.org/10.1093/braincomms/fcaa138.

[620] Suzuki, T.; Yashiro, Y.; Kikuchi, I.; Ishigami, Y.; Saito, H.; Matsuzawa, I.; Okada, S.; Mito, M.; Iwasaki, S.; Ma, D.; et al. Complete Chemical Structures of Human

Mitochondrial TRNAs. *Nat. Commun.*, **2020**, *11* (1), 4269. https://doi.org/10.1038/s41467-020-18068-6.

[621] Suzuki, T. The Expanding World of TRNA Modifications and Their Disease Relevance. *Nat. Rev. Mol. Cell Biol.*, **2021**, *22* (6), 375–392. https://doi.org/10.1038/s41580-021-00342-0.

[622] Su, Z.; Wilson, B.; Kumar, P.; Dutta, A. Noncanonical Roles of TRNAs: TRNA Fragments and Beyond. *Annu. Rev. Genet.*, **2020**, *54*, 47–69. https://doi.org/10.1146/annurev-genet-022620-101840.

[623] Guzzi, N.; Bellodi, C. Novel Insights into the Emerging Roles of TRNA-Derived Fragments in Mammalian Development. *RNA Biol.*, **2020**, *17* (8), 1214–1222. https://doi.org/10.1080/15476286.2020.1732694.

[624] Di Fazio, A.; Schlackow, M.; Pong, S. K.; Alagia, A.; Gullerova, M. Dicer Dependent TRNA Derived Small RNAs Promote Nascent RNA Silencing. *Nucleic Acids Res.*, **2022**, *50* (3), 1734–1752. https://doi.org/10.1093/nar/gkac022.

[625] Kim, H. K.; Fuchs, G.; Wang, S.; Wei, W.; Zhang, Y.; Park, H.; Roy-Chaudhuri, B.; Li, P.; Xu, J.; Chu, K.; et al. A Transfer-RNA-Derived Small RNA Regulates Ribosome Biogenesis. *Nature*, **2017**, *552* (7683), 57–62. https://doi.org/10.1038/nature25005.

[626] Lu, L.; Li, J.; Wei, R.; Guidi, I.; Cozzuto, L.; Ponomarenko, J.; Prats-Ejarque, G.; Boix, E. Selective Cleavage of NcRNA and Antiviral Activity by RNase2/EDN in THP1-Induced Macrophages. *Cell. Mol. Life Sci. CMLS*, **2022**, *79* (4), 209. https://doi.org/10.1007/s00018-022-04229-x.

[627] Su, Z.; Monshaugen, I.; Wilson, B.; Wang, F.; Klungland, A.; Ougland, R.; Dutta, A. TRMT6/61A-Dependent Base Methylation of TRNA-Derived Fragments Regulates Gene-Silencing Activity and the Unfolded Protein Response in Bladder Cancer. *Nat. Commun.*, **2022**, *13* (1), 2165. https://doi.org/10.1038/s41467-022-29790-8.

[628] Wu, W.; Lee, I.; Spratt, H.; Fang, X.; Bao, X. TRNA-Derived Fragments in Alzheimer's Disease: Implications for New Disease Biomarkers and Neuropathological Mechanisms. *J. Alzheimers Dis. JAD*, **2021**, *79* (2), 793–806. https://doi.org/10.3233/JAD-200917.

[629] Fu, H.; Feng, J.; Liu, Q.; Sun, F.; Tie, Y.; Zhu, J.; Xing, R.; Sun, Z.; Zheng, X. Stress Induces TRNA Cleavage by Angiogenin in Mammalian Cells. *FEBS Lett.*, **2009**, *583* (2), 437–442. https://doi.org/10.1016/j.febslet.2008.12.043.

[630] Finet, O.; Yague-Sanz, C.; Krüger, L. K.; Tran, P.; Migeot, V.; Louski, M.; Nevers, A.; Rougemaille, M.; Sun, J.; Ernst, F. G. M.; et al. Transcription-Wide Mapping of Dihydrouridine Reveals That MRNA Dihydrouridylation Is Required for Meiotic Chromosome Segregation. *Mol. Cell*, **2022**, *82* (2), 404-419.e9. https://doi.org/10.1016/j.molcel.2021.11.003.

[631] Draycott, A. S.; Schaening-Burgos, C.; Rojas-Duran, M. F.; Wilson, L.; Schärfen, L.; Neugebauer, K. M.; Nachtergaele, S.; Gilbert, W. V. Transcriptome-Wide Mapping Reveals a Diverse Dihydrouridine Landscape Including MRNA. *PLoS Biol.*, **2022**, *20* (5), e3001622. https://doi.org/10.1371/journal.pbio.3001622.

[632] Song, B.; Tang, Y.; Wei, Z.; Liu, G.; Su, J.; Meng, J.; Chen, K. PIANO: A Web Server for Pseudouridine-Site (Ψ) Identification and Functional Annotation. *Front. Genet.*, **2020**, *11*, 88. https://doi.org/10.3389/fgene.2020.00088.

[633] Xu, Q.; Chen, K.; Meng, J. WHISTLE: A Functionally Annotated High-Accuracy Map of Human M6A Epitranscriptome. *Methods Mol. Biol. Clifton NJ*, **2021**, *2284*, 519–529. https://doi.org/10.1007/978-1-0716-1307-8_28.
[634] Song, J.; Zhai, J.; Bian, E.; Song, Y.; Yu, J.; Ma, C. Transcriptome-Wide Annotation of M5C RNA Modifications Using Machine Learning. *Front. Plant Sci.*, **2018**, *9*, 519. https://doi.org/10.3389/fpls.2018.00519.
[635] Lavi, U.; Fernandez-Muñoz, R.; Darnell, J. E. Content of N-6 Methyl Adenylic Acid in Heterogeneous Nuclear and Messenger RNA of HeLa Cells. *Nucleic Acids Res.*, **1977**, *4* (1), 63–69. https://doi.org/10.1093/nar/4.1.63.
[636] Barozzi, C.; Zacchini, F.; Asghar, S.; Montanaro, L. Ribosomal RNA Pseudouridylation: Will Newly Available Methods Finally Define the Contribution of This Modification to Human Ribosome Plasticity? *Front. Genet.*, **2022**, *13*, 920987. https://doi.org/10.3389/fgene.2022.920987.

Curriculum Vitae



4 Avenue Théophile Brichon
554530 Pagny-sur-Moselle, FRANCE
+33 6 64 82 07 21
florian.pichot3@gmail.com

FLORIAN PICHOT

DATA ANALYST

FORMATION

September 2013 – July 2018

Faculté des sciences et technologies, Vandoeuvre-les-Nancy, France

Obtained the Licenciate's diploma of Life Sciences with honors in 2016.

Obtained the Master's diploma « BioSciences et Ingénierie de la Santé » (BSIS) with speciality « Ingénierie Moléculaire » (IM) with honors.

Folllow-up of a parellel formation called « Cursus Master en Ingénierie » (CMI) which is a French equivalent of Master of engineering during these 5 years.

Courses based on molecular biology, biochemistry, bioinformatics, biostatistics and Human and Social Sciences.

September 2010 - July 2013

Lycée Louis Bertrand, Briey, France

Obtained the « Baccalauréat S » (French High School Diploma) with honors.

EXPERIENCES

August 2018 – May 2023

RESEARCH ASSOCIATE AT JOHANNES GUTENBERG-UNIVERSITY IN PR. MARK HELM GROUP, MAINZ, GERMANY IN COOPERATION WITH PR. YURI MOTORIN, FRANCE.

Development of bioinformatics tools for Next-Generation Sequencing (NGS) for the detection of chemical modification in RNAs. Development of NO-seq.

June – July 2016 & June – July 2017 & January – July 2018

BIOINFORMATICIAN INTERNEE IN IMOPA (INGENIERIE MOLECULAIRE ET PHYSIOPATHOLOGIE ARTICULAIRE) LABORATORY, VANDOEUVRE-LES-NANCY, FRANCE WITH PR. YURI MOTORIN.

Creation of new bioinformatics tools using R software for RiboMethSeq automatized analysis, allowing 2'O-methylation detection at nucleotide resolution in sequencing data. Use of machines learning algorithms : *Random Forest* and *Support Vectors Machine*.

LANGUAGES

French : Native.

English : B2 level (TOEIC score 855 in June 2017).

German : A2 level.

HOBBIES

Aikido, Football, Biking, E-sports, Manga.