

“Chemical Labeling and Next-Generation Sequencing for Detection of RNA Modifications”

Dissertation

zur Erlangung des Grades

Doktor der Naturwissenschaften

im Promotionsfach Pharmazie
am Fachbereich Chemie, Pharmazie und Geowissenschaften
der

Johannes Gutenberg-Universität
Mainz

Lyudmil Aleksandrov Tserovski

geb. in Sofia, Bulgarien

Mainz, Okt. 2016

Dekan: ...

1. Berichtstatter: ...
2. Berichtstatter: ...

Datum der mündlichen Prüfung: ...

D77 (Dissertation Mainz)

Hiermit versichere ich eidesstattlich:

1. Ich habe die jetzt als Dissertation vorgelegte Arbeit selbst angefertigt und alle benutzten Hilfsmittel (Literatur, Apparaturen, Material) in der Arbeit angegeben.
2. Ich habe oder hatte die jetzt als Dissertation vorgelegte Arbeit nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht.
3. Ich hatte weder die jetzt als Dissertation vorgelegte Arbeit noch Teile davon bei einer anderen Fakultät bzw. einem anderen Fachbereich als Dissertation eingereicht.

Lyudmil Tserovski

I dedicate this work to my parents!

Мамо, тати, благодаря ви за подкрепата!

Abstract

In the field of life sciences RNA modifications start to reveal their important role in the dynamic regulation of gene expression which can be influenced by stress response, immunity reactions or other environmental factors. These modifications are found in archaea, bacteria and eukaryota where they decorate RNA molecules and thus expand the nucleotide repertoire. Although methods were developed in the past ten years that allow the detection of numerous modified RNA nucleosides, these techniques still lack sufficient sensitivity and specificity.

The present PhD thesis addresses several aspects on detection of naturally occurring RNA modifications. First, a library preparation protocol was adapted and optimized to capture reverse transcriptase (RT) events during synthesis of a complementary DNA sequence. Therefore, not only abortive products, but also misincorporations could be traced and later used in a bioinformatic approach for detection and prediction of modified sites.

Second, *N*-1-methyladenosine (m^1A), a well described and highly conserved modification, was used as a model to test this approach. Two major events were observed: (i) the methyl group at *N*-1 of adenosine leads to a substantial number of RT-stops and (ii) a certain amount of read-through is possible, leading to misincorporations at the modified positions. It was demonstrated that RT leaves a specific signature at m^1A sites depending not only on the underlying modification but also on the 3'-neighboring nucleoside. Together, these results led to the discovery of new m^1A positions.

Furthermore, the applicability of osmium tetroxide-bipyridine (os-bipy) as labeling agent for 5-methylcytidine and 5-methyluridine in RNA was evaluated. On the nucleoside level, a five- and ten-fold preference, respectively, over the corresponding unmodified nucleoside was observed. In a short pentanucleotide, however, this preference was strongly reduced. It was demonstrated that the sterical environment in the short oligonucleotide has a strong hindering impact on the reaction rate. Importantly, this effect could be linked to an altered diastereoselectivity which was due to an impediment of the attack of os-bipy toward the preferred *si* side of the diastereotopic 5,6 double bond of the nucleobase. As a result, in the pentanucleotide context preference toward 5-methylcytidine over cytidine was almost lost, whereas for 5-methyluridine it remained about eight times higher than for uridine.

Finally, labeling with osmium tetroxide-bipyridine was used in combination with high-throughput sequencing on a total transfer RNA population of *S. cerevisiae*. Bioinformatic analysis revealed a discrimination between 5-methylcytidine and cytidine. On the contrary, 5-methyluridine containing sites remained undetectable upon reaction with os-bipy. Nevertheless, results obtained from reaction of osmium tetroxide and bipyridine with 5-methylpyrimidines are promising and provide an important basis for the development of appropriate labeling agents for a transcriptome-wide detection of these modifications.

Zusammenfassung

Im Bereich der therapeutischen Lebenswissenschaften nehmen Modifikationen von Ribonukleinsäuren (RNS) eine entscheidende Rolle in der Regulierung der Genexpression ein, beeinflussbar durch eine Stressantwort, Immunreaktion oder andere Umweltfaktoren. RNS-Modifikationen treten in Archaeen, in Bakterien und in Eukaryonten auf und erweitern somit deren Nukleotidrepertoire. Obwohl die Existenz solcher Modifikationen schon seit etwa 70 Jahren bekannt ist, sind ihre vielfältigen Funktionen bis heute nicht vollständig aufgeklärt. Dies kann durch die noch unzureichende Sensitivität und Spezifität der aktuell vorhandenen Detektionsmethoden erklärt werden.

Die vorliegende Dissertation beschäftigt sich mit diversen Aspekten der Detektion und Lokalisierung von natürlich vorkommenden RNS-Modifikationen. Zunächst wurde ein Protokoll angepasst und optimiert, das zur Erfassung besonderer Ereignisse bei der reversen Transkription (RT) von RNS zu komplementärer Deoxyribonukleinsäure (DNS) Anwendung findet. Bei diesen Ereignissen handelt es sich vornehmlich um RT-Abbrüche und den Fehleinbau von Deoxynukleotiden. Mithilfe des Protokolls wurde das charakteristische Verhalten der RT bei natürlich vorkommenden RNS-Modifikationen analysiert und die auftretenden Ereignisse konnten für die Entwicklung einer automatisierten Plattform zur Erkennung von Modifikationen verwendet werden. Dabei wurde 1-Methyladenosin (m^1A) als Modell ausgewählt und sowohl in nativen, als auch in synthetisch hergestellten RNS-Molekülen untersucht. Die Analyse ergab, dass das RT-Verhalten an m^1A -Positionen in RNS sehr spezifisch ist. Neben einem großen Anteil an RT-Abbrüchen konnten Nukleotid-spezifische Fehlinkorporationen beobachtet werden. Darüber hinaus war die Fehlinkorporation stark von dem 3'-benachbarten Nukleotid abhängig. Die Ergebnisse ermöglichten eine Computer-gestützte Erkennung von neuen m^1A -Positionen in verschiedenen Organismen.

Weiterhin beschäftigt sich die vorgelegte Arbeit mit der Möglichkeit, unter Verwendung von Osmiumtetroxid-Bipyridin die modifizierten Pyrimidinnukleoside 5-Methyluridin (m^5U) und 5-Methylcytidin (m^5C) selektiv chemisch zu markieren und durch anschließende Sequenzierung zu detektieren. Als Nukleosid reagierte m^5U etwa zehnfach schneller als sein unmodifiziertes Analogon und m^5C etwa fünffach schneller als Cytidin. In einem Pentanukleotid war diese Reaktivität jedoch deutlich reduziert. Dieser Effekt beruht auf einer veränderten Diastereoselektivität, die auf eine sterische Behinderung der bevorzugten *si*-Seite zurückzuführen ist.

Schließlich wurde die Markierung mit Osmiumtetroxid-Bipyridin an Transfer-RNS von Hefe getestet. Durch Hochdurchsatz-Sequenzierung und nachfolgende bioinformatische Analyse wurde eine Diskriminierung zwischen 5-Methylcytidin und Cytidin beobachtet. Im Gegensatz dazu waren m^5U -Positionen unter den gegebenen Bedingungen nicht nachweisbar. Dennoch sind die Ergebnisse der Reaktion von Osmiumtetroxid und Bipyridin mit 5-Methylpyrimidinen in RNS vielversprechend und bilden die Grundlage für die Entwicklung geeigneter Markierungsreagenzien zum transkriptomweiten Nachweis dieser Modifikationen.

Contents

Abstract	xi
Zusammenfassung	xiii
Abbreviations	xvi
1 Introduction	1
1.1 RNA and RNA modifications	1
1.1.1 Nomenclature of nucleosides	1
1.1.2 Functions of RNA modifications	3
1.2 Osmium tetroxide and its applications in nucleic acids analysis	7
1.2.1 Chemistry of osmium tetroxide	7
1.2.2 Applications in nucleic acids	7
1.2.3 Labeling of 5-methyl deoxycytidine	9
1.3 Determination of RNA modifications	9
1.3.1 HPLC-MS methods	10
1.3.2 Sequencing-based methods	10
1.4 High-throughput sequencing of RNA and DNA	13
1.4.1 Illumina [®] sequencing technology	13
1.4.2 454-Pyrosequencing	15
1.4.3 Ion torrent sequencing	15
1.4.4 SOLiD [™] sequencing technology	16
1.4.5 Single molecule sequencing	16
1.4.6 RNA sequencing using Illumina [®] sequencing technology	17
1.5 Transcriptome-wide detection of RNA modifications	18
1.5.1 Detection of inosine	19
1.5.2 Detection of pseudouridine	20
1.5.3 Detection of 5-methylcytidine	21
1.5.4 Detection of <i>N</i> -6-methyladenosine	23
1.5.5 Detection of <i>N</i> -1-methyladenosine	23
1.5.6 Detection of 2'- <i>O</i> -methylation	24
2 Materials and Methods	26
3 Goal of The Work	27
4 List of Publications	29
5 Results and Discussion	30
5.1 Detection of m ¹ A by NGS	30
5.1.1 Library preparation protocol for RNA sequencing	30
5.1.2 The reverse transcription signature of <i>N</i> -1-methyladenosine in RNA-Seq is sequence dependent	44
5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain	76
5.3 Application of os-bipy reagent and next generation sequencing for detection of 5-methylpyrimidines	98
5.3.1 Stability of Os(VI)-bipy complex during reverse transcription	98
5.3.2 Detection of 5-methylcytidine in yeast tRNA	99

6 Conclusion and Outlook	108
7 References	111
8 Appendix	124
8.1 The plus and minus - sequencing method	124
8.2 Chemistry of the template-directed DNA synthesis	124
8.3 Sequencing of RNA - nucleoside-specific reactions	125
8.4 Inosine base-pairing	126
8.5 Padlock probes	126
8.6 SCARLET method	126
8.7 Os-bipy label of tRNA ^{Ile(TAT)} - differential CSA analysis	127
8.8 Proposed base-pairing properties of osmylated pyrimidines	128
Acknowledgements	129
Lebenslauf	130

Abbreviations

ADAR	adenosine deaminase acting on RNA
CCD	charge coupled device
CMC	N-cyclohexyl-N'-beta(4-methylmorpholinium)ethylcarbodiimide
DNA	deoxyribonucleic acid
cDNA	complementary DNA
dNTP	2'-deoxyribonucleoside 5'-triphosphate
ddNTP	2',3'-dideoxyribonucleoside 5'-triphosphate
dsDNA	double-stranded DNA
ssDNA	single-stranded DNA
ATP	adenosine triphosphate
CTP	cytidine triphosphate
GTP	guanosine triphosphate
TTP	thymidine triphosphate
DTT	dithiothreitol
ICP	inductively coupled plasma
miCLIP	methylation individual nucleotide cross-linking immunoprecipitation
os-bipy	osmium tetroxide-bipyridine
RNA	ribonucleic acid
eRNA	enhancer RNA
lncRNA	long non-coding RNA
mRNA	messenger RNA
miRNA	micro RNA
mt tRNA	mitochondrial tRNA
ncRNA	non-coding RNA
rRNA	ribosomal RNA
snRNA	small nuclear RNA
snoRNA	small nucleolar RNA
siRNA	small interfering RNA
tRNA	transfer RNA
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
RF	random forest
RT	reverse transcription
SAM	<i>S</i> -adenosyl methionine
SNP	single nucleotide polymorphism
TEMED	tetramethylethylenediamine

1 Introduction

The term nucleoside was first introduced by Levene and co-workers about a hundred years ago to define sugar derivatives of purine bases isolated from yeast [1]. Later, this term was extended to denote compounds containing pyrimidine and other heterocyclic bases [2]. Four major nucleosides: adenosine, guanosine, cytidine and uridine are building blocks of a complex family of macromolecules - ribonucleic acid (RNA). Additionally, more than 150 chemically distinct RNA modifications have been discovered to date [3]. Apart from the main RNA species involved in the protein biosynthesis - transfer RNA (tRNA), ribosomal RNA (rRNA) and messenger RNA (mRNA), existence of other seldom RNA species was recently reported [4]. Naturally, questions on their functional role were raised. Additionally, as a result of technological development, numerous RNA modifications were newly discovered in mRNA by high-throughput sequencing methods. This led to the introduction of the term epitranscriptomics that describes the existence, distribution and functional role of RNA modifications on the regulation of gene expression [5].

Still, the methods developed and applied currently show certain discrepancies, a hint that sensitivity and specificity of the latter are not very satisfying. Therefore, improved methods for confident prediction and confirmation of such modified sites are still needed. This was the reason for the investigations initiated by this dissertation.

In the following introductory chapters the term RNA, as well as its building blocks - nucleosides - will be presented. Furthermore, certain chemical properties that provide a basis for detection possibilities will be discussed. Of special interest is the chemical reagent osmium tetroxide that was evaluated as a labeling agent during this work. The introduction concludes with a description of the currently used methods of high-throughput possibilities for RNA sequencing, as well as detection of several important naturally occurring RNA modifications.

1.1 RNA and RNA modifications

The extensive work of Crick and collaborators in the years between 1950 and 1970 has greatly contributed to our understanding of the expression of proteins encoded in the genome. The most important conclusion was drawn in Crick's re-publication of the "Central Dogma of Molecular Biology" [6]. There he stated that RNA is synthesized out of DNA and proteins from RNA. And although certain interconversions are possible, such as synthesis of DNA and RNA from RNA, generating DNA out of proteins was declared prohibited. We now know that the picture drawn at that time was not complete. It was discovered that processes exist that regulate the expression of proteins on a level, that is independent on the genetic code. First such processes were discovered on the level of DNA dynamic alteration, such as DNA methylation and histone acetylation [7]. This regulation, independent on the genetic information, was called epigenetics (greek: $\epsilon\pi\iota$ - outside). In the recent years, development of very sensitive methods allowed detection of numerous RNA modifications in mRNA and processes were postulated that describe the regulation of expression on the RNA level. Thus the term epitranscriptomics was introduced [5] (see figure 1).

1.1.1 Nomenclature of nucleosides

Nucleosides are building blocks of RNA. A nucleoside consists of a nucleobase attached to the ribose typically *via* a *N*-glycosidic bond. There are two nucleosides carrying a purine-derived base: adenosine and guanosine (see figure 2A). The other two nucleosides carry a pyrimidine-derived base: cytidine and uridine (see figure 3A). Of note, in deoxyribonucleic acid (DNA)

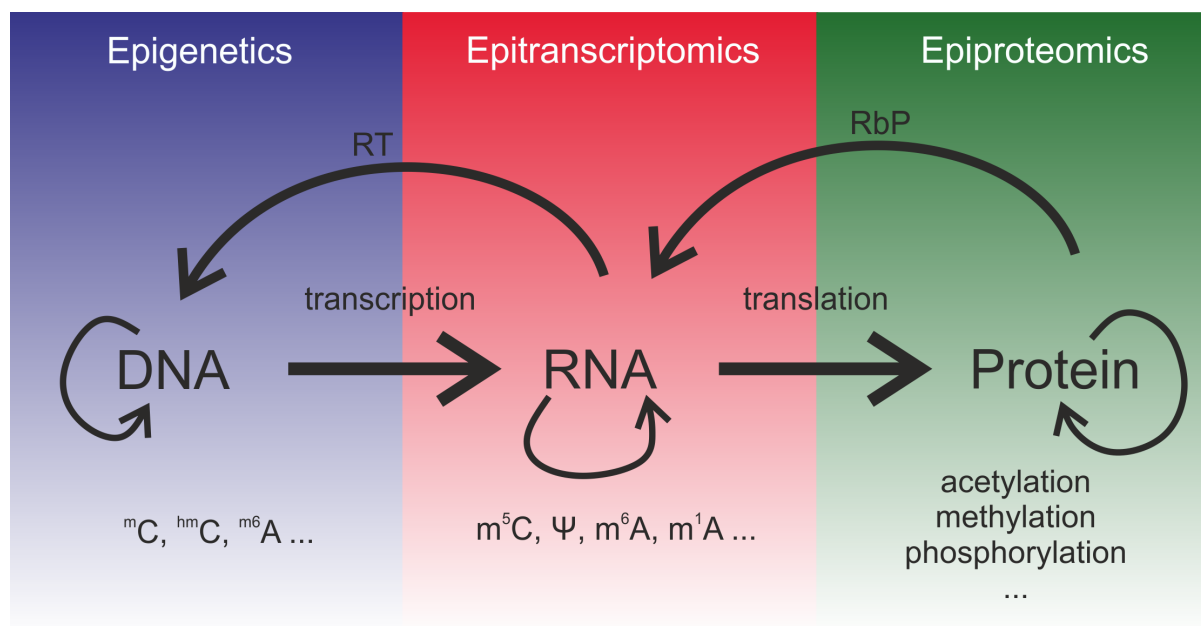


Figure 1: Role of modifications in DNA, RNA and proteins on the expression state of an organism. RT - reverse transcription, RbP - RNA binding protein. Figure adapted from [5].

thymidine is present instead of uridine. If a nucleoside carries a phosphate group it is called nucleotide.

Because nucleosides are involved in a wide range of interactions, a nomenclature has been proposed by the base edges participating in such an interaction [10]. Base-pairing relying on hydrogen bonding at the Watson-Crick Edge determines complementarity by the template-directed DNA or RNA synthesis. Herein, adenosine base-pairs with uridine in RNA or thymidine in the case of DNA *via* two hydrogen bonds, whereas guanosine builds three hydrogen bonds with cytidine. So, unless otherwise stated in text, base-pairing refers to the Watson-Crick base-pairing.

Apart from the four canonical ribonucleosides, numerous naturally occurring modifications have been discovered in the last 70 years. Nomenclature and corresponding abbreviations are based on the RNA modification database [11]. The current version of MODOMICS, an RNA modification pathway database, includes 167 distinct modifications [3] and this list is probably still not complete. According to the established nomenclature, modifications are denoted by the chemical groups they present and the position of occurrence. For example, 5-methylcytidine denotes a methylation at position 5 of the nucleobase of cytidine, whereas 2'-*O*-methyluridine describes a methylation at the 2'-OH of the ribose. Apart from the unabbreviated name that determines a given nucleoside or modification, also short names and single-symbol codes have been introduced. In this case, A corresponds to adenosine, G to guanosine, C stands for cytidine and U for uridine. Then, using the same examples as above, 5-methylcytidine is abbreviated to m^5C or as a single-symbol "?", whereas 2'-*O*-methyluridine is represented by Um or "J", respectively. Thus, in case of base modifications, the modification and corresponding position are denoted in front of the underlying nucleoside. In contrast, ribose methylations are written after the letter denoting the base. For pseudouridine, the first described non-canonical nucleoside, a separate symbol was proposed - Ψ .

Most common modifications presented in RNA species are methylations at differing positions either at the nucleobase, or the 2'-*O* of the ribose. Depending on where such methylation

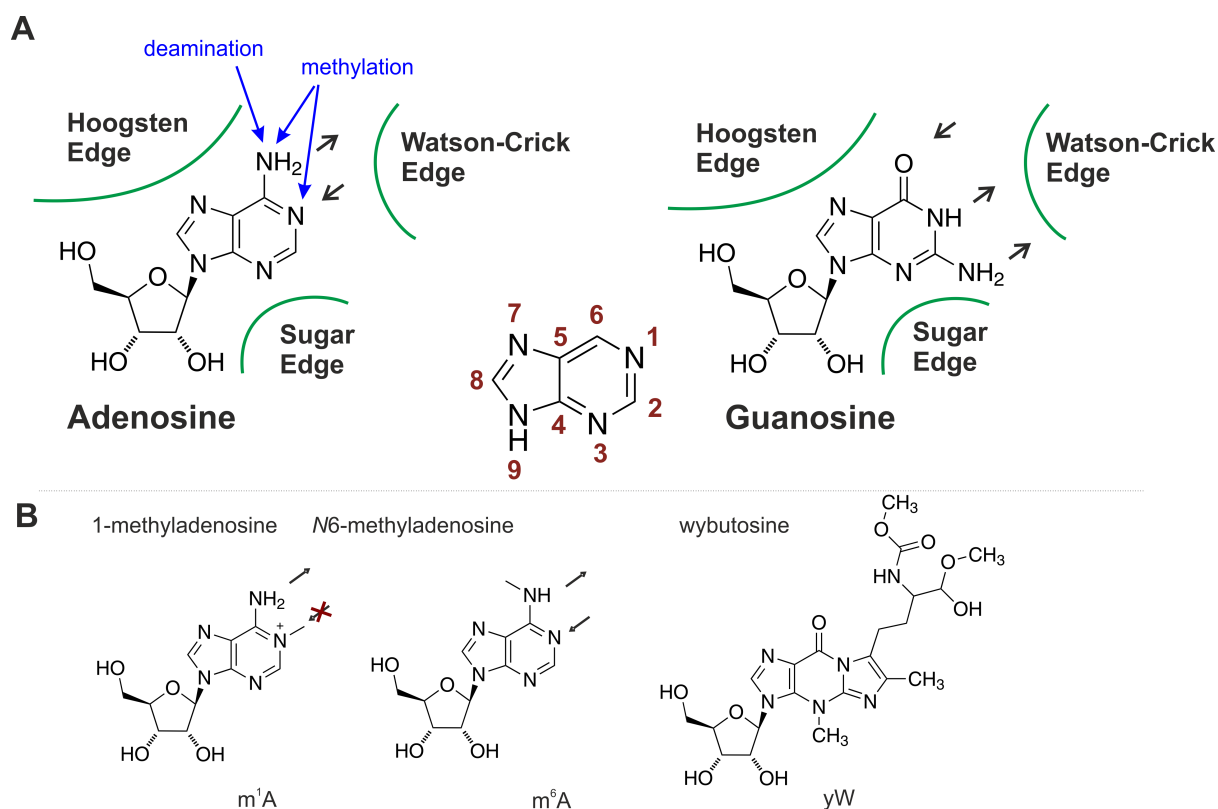


Figure 2: A) Purine nucleosides adenosine and guanosine. Both are connected to the ribose *via* N -9 of the nucleobase. B) Some examples of modified purine-nucleosides. Typical methylations occur either at N -1 or N -6 of adenosine. Wybutosine, a modified guanosine is a typical modification found at position 37 of tRNA^{Phe} of eucaryotic organisms [8]. Black arrows denote the atoms involved in hydrogen bonding at the Watson-Crick Edge. Figure adapted from [9].

occurs, base-pairing properties can be affected (see figures 2 and 3). Numerous modifications appear as a result of atom substitutions. For example, the isomers 2-thiouridine (s^2U) and 4-thiouridine (s^4U) are derived from uridine by substitution of oxygen at position 2 and 4, respectively. Interestingly, both modifications do not exhibit changed base-pairing properties. Pseudouridine is a unique RNA modification also derived from uridine. In contrast to all other RNA building blocks, pseudouridine (Ψ) possesses a rather unusual C–C glycosidic bond. Because of the stability of this bond, Ψ does not yield the nucleobase uracil upon hydrolysis. As a result its identification in the early 1960s was more difficult [12]. The most complex modifications, however, occur relatively seldom and are more conserved at the anti-codon stemloop of tRNA (for example see wybutosine at figure 2). More information on the functions of RNA modifications is given in the next subsection.

1.1.2 Functions of RNA modifications

As already mentioned in the beginning of this section, numerous processes determine the protein expression state of an organism. In the protein biosynthesis, RNA plays an inevitable role. A population of RNA species, messenger RNA (mRNA) is transcribed from DNA and further translated to proteins. Therefore, mRNA is also called coding RNA. In contrast, types of RNA that play rather functional and regulatory role are called non-coding RNA. Non-coding RNA (ncRNA) are for example the highly abundant transfer RNA (tRNA) and ribosomal RNA

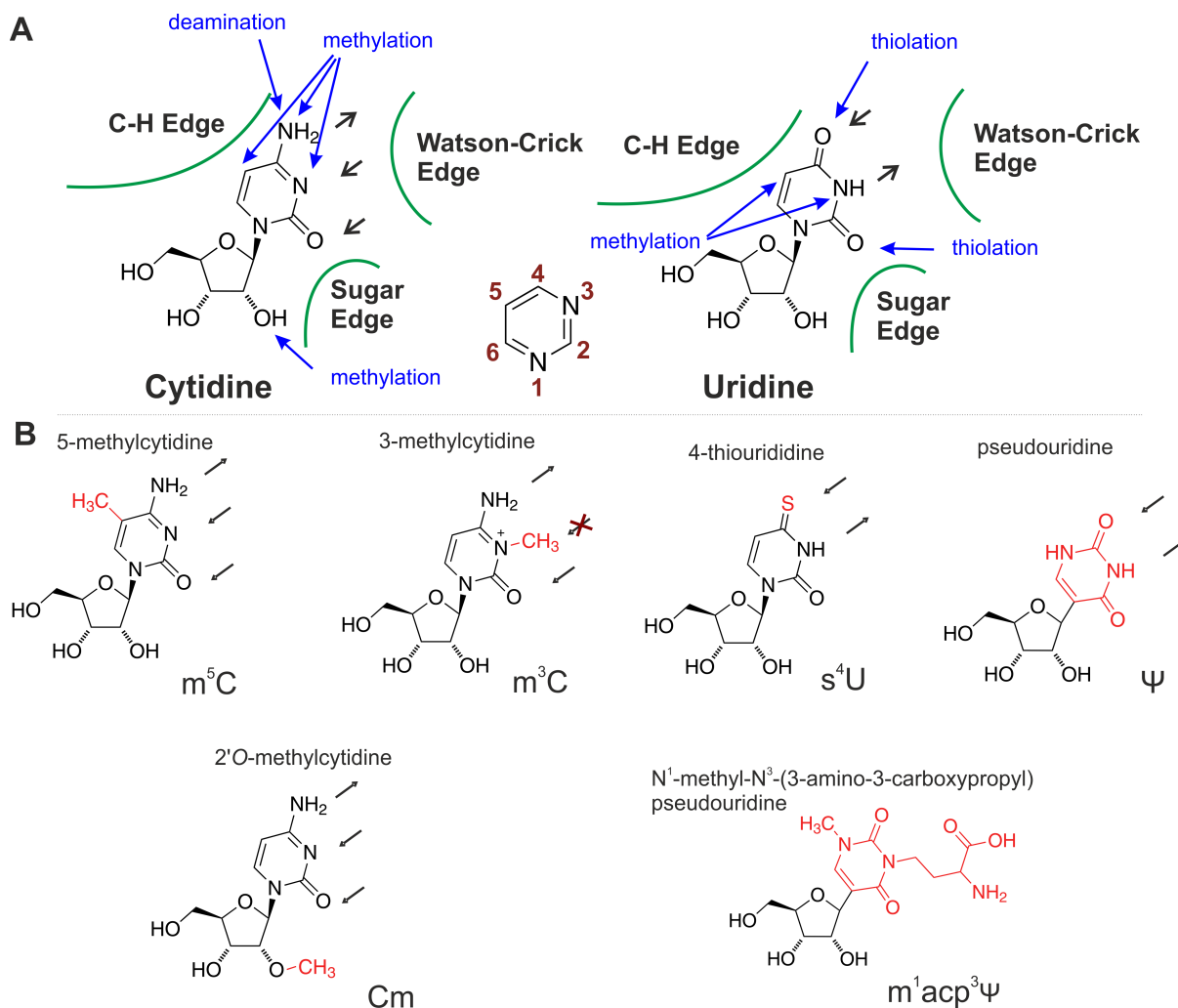


Figure 3: A) Pyrimidine nucleosides cytidine and uridine. Both are connected to the ribose *via* *N*-1 of the nucleobase. B) Some examples of modified pyrimidine-nucleosides. Typical methylations occur either at the nucleobase, for example 5-methylcytidine or 3-methylcytidine, or at the 2'-*O* of the ribose, e.g. 2'-*O*-methylcytidine. More complex modifications include exchange of atoms, such as 4-thiouridine. A special case presents pseudouridine (Ψ) that exhibits an unusual C–C glycosidic bond. Ψ can be further modified to, e.g. $m^1acp^3\Psi$, a modification found in ribosomal RNA. Black arrows denote the atoms involved in hydrogen bonding at the Watson-Crick Edge. Figure adapted from [9].

(rRNA). Both species play a direct role in the protein synthesis. Transfer RNA delivers the amino acids and rRNA together with other proteins build the ribosome where actual protein synthesis takes place. Further examples of ncRNA that are involved in post-transcriptional modification processes include snRNA (small nuclear RNA) and snoRNA (small nucleolar RNA). Others, with mainly regulatory role, are siRNA (small interfering RNA), miRNA (micro RNA), lncRNA (long non-coding RNA) and eRNA (enhancer RNA).

Please refer to subsections 1.3 and 1.5 for methods applied for modification detection.

Modifications in tRNA

The vast majority of all discovered naturally occurring RNA modifications were found in transfer RNA (tRNA) [3]. Not only is this RNA species probably the best described until now, but tRNA also contains numerous highly complex modifications. Certain tRNA base modifica-

tions are important for stabilization of the tertiary, L-shaped structure and thus aminoacylation rate and accuracy are influenced [13]. A conformational change occurs, for instance, upon *N*-1 methylation of adenosine at conserved position 58 in eucaryotic tRNA [14]. Lack of this modification in initiator tRNA^{Met} leads to polyadenylation followed by degradation of this tRNA in the nuclear exosome [15]. Other base modifications play a role in a targeted tRNA cleavage [16]. Schaefer et al. demonstrated that 5-methylcytidine, produced by methylation enzyme DNMT2, reduces the tRNA cleavage that is otherwise induced by stress [17]. Modifications in the anticodon stemloop participate in mRNA and rRNA interactions and therefore the fidelity of translation is influenced [18]. Especially, the wobble position 34 and position 37, that is adjacent to the anticodon loop, are modified to stabilize the specific codon-anticodon interactions. Loss of this stabilization leads to severe diseases [19, 20]. It is a widespread hypothesis that the wobble position 34 increases the capability of tRNA to decode multiple synonymous mRNA codons [21].

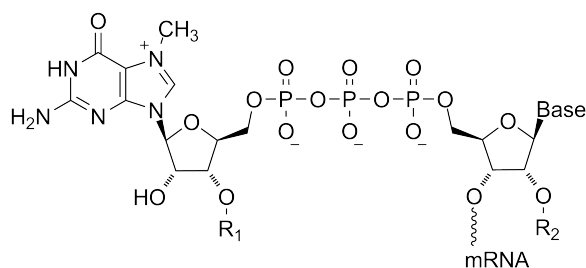
Mitochondria are organelles with a semi-autonomic genome in eukaryotic cells [22]. The mitochondrial genome in mammalian mitochondria encodes for 13 proteins of the oxidative phosphorylation system, two rRNA and all 22 tRNAs required for mitochondrial protein biosynthesis. The tRNAs (mt tRNA) are, in contrast to cytosolic tRNAs, less modified, contain, however, several unique post-transcriptional modifications [23]. The modifications at the anticodon stemloop are crucial for the correct deciphering of all 60-sense codons. For example, 5-formylcytidine was discovered at the wobble position 34 of mt tRNA^{Met} isolated from bovine liver [24]. It was demonstrated that its existence is required for the recognition of the non-universal AUA codon. Others highly-conserved modifications, such as pseudouridine and 5-methyluridine are responsible for the correct formation and stabilization of functional mt tRNA [25]. Interestingly, Helm et al. showed that the existence of *N*-1-methyladenosine (m¹A) at position 9 of human mitochondrial tRNA^{Lys} is necessary and sufficient for the correct folding by hindering the base-pairing between A9 and U64 [26]. An extensive review focusing on tRNA modifications that are important in the regulation of gene expression was recently published by Duechler and collaborators [22].

Modifications in rRNA

During the biogenesis of ribosomal RNA (rRNA) numerous modifications are introduced. The most abundant modifications include isomerisation of uridine to pseudouridine, methylation at the 2'-*O* of the ribose, and various methylations at the nucleobases [27]. Interestingly, it has been discovered that many uridine-to-pseudouridine conversions and ribose methylations in archeen and eukaryots are guided by complementary short RNAs - snoRNA [28]. In contrast, bacterial rRNA contain only a few 2'-*O*-methylations and pseudouridines, which are synthesized by specific enzymes. These modifications occur predominantly at functionally essential regions of rRNA and are therefore expected to contribute to the correct functioning of the ribosome. It has been reported that lack of ribosomal pseudouridines in ribosomal RNA of *Escherichia coli* leads to inhibited cell growth [29]. In contrast, base modifications occur less frequently [30]. Most common base modifications include methylations at positions 1, 6 and 7 of purines and 1, 3 and 5 of pyrimidines. Certain methylations occurring at the Watson-Crick Edge such as *N*-1-methyladenosine and *N*-3-methyluridine interrupt the base-pairing properties of the underlying nucleoside. On the other hand, 5-methylcytidine increases the lipophilic surface of the base, thus promoting base stacking. These properties contribute to the correct folding and structure formation of ribosomal RNA and therefore their proper functions [31].

Modifications in mRNA

It is well-known that eucaryotic cells undergo complex mRNA post-transcriptional processing. Capping of the 5' end of mRNA is essential for protection of mRNA from exonucleases, regulation of nuclear export and promotion of translation [32]. The cap consists of an *N*-7-methylguanosine linked to the mRNA through a 5'-5' triphosphate bond (see figure 4). Furthermore, it was described, that certain organisms possess additionally a 2'-*O*-methylation on the first nucleoside (called cap1) or at the first two nucleosides (cap2) [33]. Very recently it was demonstrated that 2'-*O*-methylation of the cap of RNA coming from Dengue virus is important for the evasion of the virus from host immune response [34].



R₁, R₂: either H or CH₃

Figure 4: Structure of the 5'-cap of eucaryotic organisms and some viruses. Structure adapted from [33].

In the cytoplasm, stability of mRNA is regulated by the length of polyA tail and by RNA interference processes [35]. The first internal modification found in mRNA was *N*-6-methyladenosine (m⁶A) [36]. This modification was abundant enough for detection in bulk messenger RNA. Development in next-generation sequencing, in combination with antibodies, allowed the mapping of m⁶A on a transcriptome-wide basis [37]. Later, improvements of the method led to detection on a single nucleoside resolution [38]. It is known that methylation-specific enzymes called writers (e.g. METTL3, METTL14, WTAP) are responsible for the selective methylation at position *N*-6 of adenosine, and eraser enzymes (FTO) are responsible for corresponding demethylation [39]. In their study, Zhou and co-workers demonstrated, that dynamic methylation in the 5'-untranslated region (5'-UTR) is necessary and sufficient to initiate a cap-independent translation [39]. Another effect attributed to m⁶A is the association with binding protein YTHDF2 that promotes co-localization with decay factors [40].

Pseudouridine (Ψ) is another modification discovered in eucaryotic mRNA by high-throughput sequencing [41, 42, 43]. Although a much less abundant modification, it was demonstrated that numerous Ψ -synthases are responsible for its synthesis that are also active on tRNA and rRNA. Pseudouridine is thought to be an irreversible modification, therefore the observed dynamics are attributed to production or degradation of pseudouridinylated transcripts [44].

Even less is known about the function of 5-methylcytidine (m⁵C) in mRNA [45]. Reported enrichment in the untranslated regions of mRNA in *Homo sapiens* suggests its role in translation regulation [46].

The most recently discovered mRNA modification is *N*-1-methyladenosine (m¹A) [47, 48]. It is predominantly found in structured 5'-UTR and in the proximity of canonical and alternative translation initiation sites. Interestingly, a connection between the existence of m¹A and elevated translation rates was described [47].

1.2 Osmium tetroxide and its applications in nucleic acids analysis

1.2.1 Chemistry of osmium tetroxide

First indications that osmium tetroxide (OsO_4) is reduced by unsaturated species came from Hofmann and co-workers in the early 1912-1913s [49, 50]. In their studies, the authors demonstrated that in presence of sodium- or potassium chlorate osmium tetroxide can be catalytically used for the hydroxylation of alkenes. About 30 years later, Criegee and colleagues showed that in stoichiometric amounts OsO_4 itself can be used as an effective *cis*-hydroxylation agent without secondary oxidants [51, 52]. The corresponding proposed mechanism is shown in figure 5A. It has been assumed that the oxidation by osmium tetroxide is initiated by a direct oxygen attack at the unsaturated center. A 6π -electron transition state would then lead then to the *cis*-addition of OsO_4 to the alkene.

An alternative mechanism, shown in figure 5B, was proposed by Sharpless and collaborators in 1977 [53]. They suggested that although weak nucleophile, the π -electrons of the double bond $\text{C}=\text{C}$ attack the more electropositive osmium center. This hypothesis is supported by the finding that both $\text{C}=\text{N}$ and $\text{C}=\text{O}$ are unreactive toward OsO_4 . The organometallic intermediate rearranges in a rate determining step to the five-membered cyclic ester complex. Important observations were made that electron-withdrawing groups on the alkene reduced reactivity toward osmium tetroxide [54]. In contrast, the same substituents led to increased bis-hydroxylation when MnO_4^- was used instead, indicating that the two reactions proceed *via* different mechanisms.

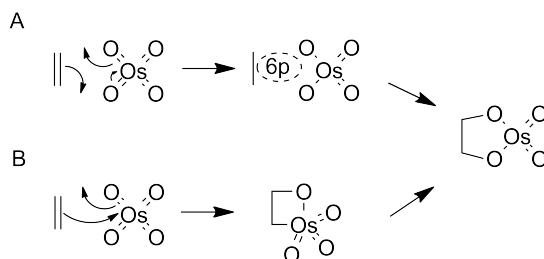


Figure 5: Proposed mechanisms for *cis*-addition of osmium tetroxide to an alkene. A) One-step reaction where a direct attack of oxygen toward unsaturated center leads to a six-electron transition state. B) Proposition from Sharpless and co-workers [53] that an organometallic intermediate is formed upon nucleophilic attack of the alkene toward the osmium center. In a rate-determining second step, rearrangement leads to the *cis* product. Reaction scheme adapted from [55].

The intermediate product, built according to reaction pathway 5B, may explain the drastic increase of formation of osmium (VI) complexes upon addition of aromatic nitrogen donors such as pyridine (see figure 6) [53]. In this case, an electron donation from the nitrogen would induce an osmium-carbon bond cleavage. Several studies on the bis-hydroxylation of thymidine have shown that the reaction is several orders faster in presence of pyridine than with osmium tetroxide alone [56, 57, 58].

1.2.2 Applications in nucleic acids

As already mentioned, several studies concentrated on the applicability of osmium tetroxide in DNA studies. First Beer and co-workers [59, 60], later Behrman et al. [56, 57, 61] showed the preferred reaction of OsO_4 with thymidine over deoxycytidine, and the negligible reaction with purine deoxynucleosides in the presence of tertiary amines. Moreover, it was demonstrated that

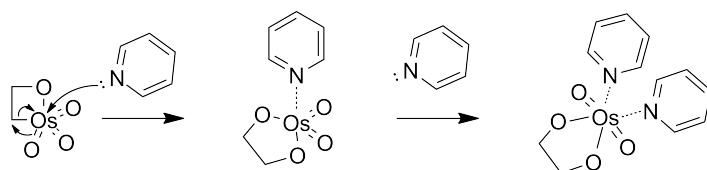


Figure 6: Proposed mechanism for increased reaction rate upon addition of an aromatic nitrogen donor. Electron donation from pyridine to osmium atom would induce osmium-carbon bond cleavage, leading to a formation of intermediate complex. Addition of a second ligand would lead to the final complex. Reaction scheme adapted from [55].

the reaction of osmium tetroxide together with the bidentate chelating agent 2,2'-bipyridine leads to very stable pyrimidine-nucleoside-osmium-bipyridine complexes [56].

Palecek and collaborators used osmium tetroxide alone [62], later in addition of ligands [63, 64], to probe DNA structures. Probing of DNA relies in this case on the fact that OsO_4 selectively reacts with pyrimidine deoxynucleosides not directly involved in base-pairing. Based on their results, Kanavarioti and co-workers investigated the kinetics of complex formation for thymidine and cytidine using capillary electrophoresis [65]. Taking advantage of their analytical device, they also showed the existence of an intermediate complex OsO_4 -bipy (see figure 7). Using optimized conditions that labeled either only thymidines or both, thymidines and cytidines, they developed a Nanopore-based sequencing method for DNA [66, 67, 68].

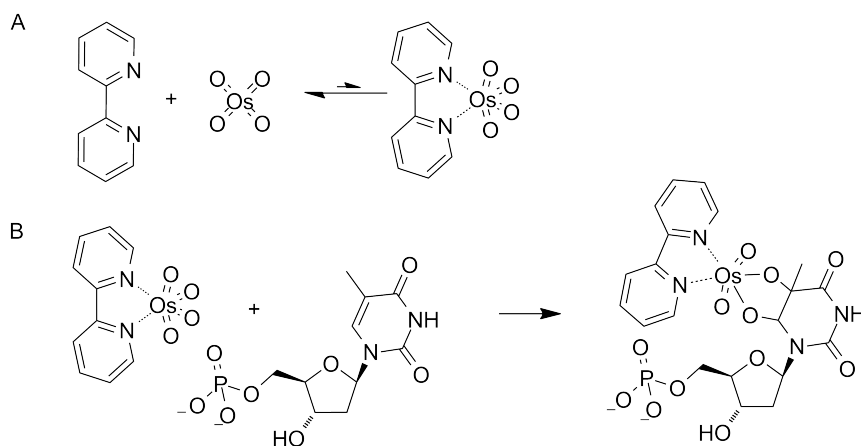


Figure 7: A) Reaction of osmium tetroxide with 2,2'-bipyridine forming an intermediate complex OsO_4 -bipy. B) Reaction of complex OsO_4 -bipy with thymidine-5'-phosphate and formation of corresponding thymidine-os-bipy-complex. Reaction scheme adapted from [65].

While Os(VIII) in OsO_4 adds selectively to unsaturated $\text{C}=\text{C}$ bonds and does not react with vicinal diols, Os(VI) as in K_2OsO_4 adds to vicinal diols. This reactivity was already proposed by Behrman et al. [56] and was successfully exploited by Wrobel and collaborators in combination with another nitrogen donor - tetramethylethylenediamine (TEMED) - to label the vicinal 2',3'-diol of ribose nucleosides (see figure 8). Osmium is a suitable label because of its high ionization efficiency, and low detection limit (in the order of 100 pg L^{-1}). Therefore, osmium (VI) adducts lead to an increase in detection upon Inductively Coupled Plasma (ICP) mass spectrometry [69].

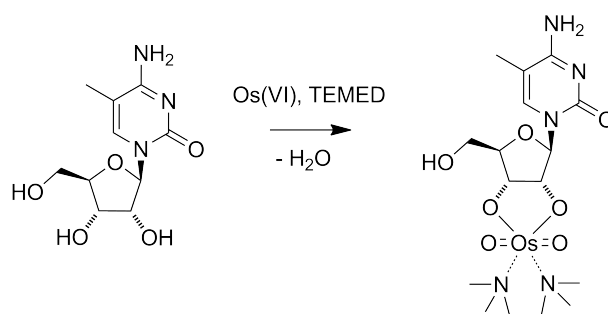


Figure 8: Labeling vicinal 2',3'-diol of 5-methylcytidine with Os(VI) and tetramethylethylenediamine (TEMED). Reaction scheme adapted from [69].

1.2.3 Labeling of 5-methyl deoxycytidine

In contrast to the already established routine of adding OsO_4 and bipy together, Okamoto and collaborators presented an *in situ* generation of OsO_4 by oxidation of potassium osmate(VI) with potassium hexacyanoferrate(III) [70, 71, 72]. Based on the findings by Burton et al. [73], the group of Okamoto demonstrated that 5-methylcytidines can be discriminated against cytidines in DNA [70] upon reaction with OsO_4 –bipy. Their work included a detailed investigation of the formed product 5-methyl-2'-deoxycytidine glycol-dioxoosmium-bipyridine complex. Naturally, the existence of a spacially demanding deoxyribose in direct proximity to the reactive 5,6 double bond of the nucleobase induces a diastereomeric preference. Umemoto and co-workers determined the stereochemical configuration of the preferred diastereomer as 5S,6S [71]. Interestingly, the same preference was found in an independent experiment on bis-hydroxylation of thymidine by osmium tetroxide alone. Among the four products (two *cis* and two *trans*), *cis* 5R¹,6S was by far the most preferred one, and the absolute structure was determined by 2-dimensional ¹H-NMR analysis [74].

1.3 Determination of RNA modifications

Before first reports on the existence of modified nucleosides in both DNA and RNA appeared, appropriate detection methods had to be established. First mention of such possibilities appeared in 1949 by Markham and Smith [75]. Their method was based on 1- and 2-dimensional paper chromatography and was able to separate numerous nucleobases as well as purine- and pyrimidine-ribonucleosides on a microgram scale.

Initial assumptions that a fifth nucleoside in RNA exists, later called pseudouridine, came from Davis and Allen in 1957 [76]. Three years later, Cohn proved the identity of this modification by chromatographic, spectrophotometric and NMR analysis and showed its unusual C–C glycosidic bond [12].

Ribothymidine, also known as 5-methyluridine, was discovered in RNA by Littlefield and Dunn in 1958 [77].

In 1959 Smith and Dunn discovered 2'-*O*-methylation in numerous ribonucleosides [78]. Their finding was based on the stability upon alkaline hydrolysis of dinucleoside phosphates carrying a 2'-*O*-methyl group. This important discovery is the basis for a very recent study on ribosomal 2'-*O*-methylation based on deep sequencing [79, 80] (Please refer to subsection 1.5.6 for application of these findings).

¹5R in thymidine has the same configuration as 5S in 5-methylcytidine because of the higher priority of oxygen at position 4 in thymidine as compared to nitrogen in 5-methylcytidine.

1.3.1 HPLC-MS methods

In the 1980s first publications came out showing that rapidly developing HPLC techniques are suitable for qualification and quantification of nucleosides [81]. In this case, the separation is performed on a reverse-phase C18-column and aqueous-organic eluents containing low amounts of buffer. The detection relied on the UV absorption properties of nucleosides, especially the aromaticity of the nucleobase. Later, improvements in mass spectrometry methods allowed coupling of both analytical devices [82]. Edmonds et al. showed that upon enzymatic digestion of tRNA to nucleosides, 0.1-10 ng were detectable with a signal-to-noise ratio higher than 10 using selected ion monitoring (SIM) [82]. Especially, high resolution mass detection allowed for further development of nucleoside-quantification techniques by addition of stable isotope-labeled internal standards [83].

A major drawback of this nucleoside detection technique is the complete loss of sequence information due to the digestion of RNA macromolecules, first to nucleotides and then to nucleosides. A possibility to circumvent this disadvantage is the direct LC-MS of oligonucleotides developed by Kowalak and co-workers [84]. Usage of RNase T₁ leads to specific hydrolysis of RNA of interest. Upon comparison with unmodified RNA as well as specific mass shifts, corresponding modifications can be uniquely determined. This method was further optimized by the group of Limbach by expanding the pool of applicable endonucleases and thus creating co-called signature digestion products for each RNA of interest [85]. Nevertheless, although it is possible to quantitatively detect pseudouridine in RNA as demonstrated by Addepalli and co-workers [86], a drawback of this technique remains the high amount of analyte needed. This makes the approach applicable predominantly for high abundant RNA species such as rRNA and tRNA.

1.3.2 Sequencing-based methods

Sequencing by synthesis

In 1977 Sanger and coworkers presented the “the dideoxy method”, a technique based on the enzymatic activity of a DNA polymerase [87]². The principle mechanism of the enzymatic elongation of a complementary DNA is shown in Appendix figure A.1. The 3'-OH of the primer attacks the electrophilic α -phosphate of the next complementary deoxynucleoside 5'-triphosphate (dNTP). An internucleotide linkage is formed upon release of a pyrophosphate and the next elongation step can take place. Therefore, if instead of the natural dNTP, a 2',3'-dideoxyribonucleoside 5'-triphosphate (ddNTP) is incorporated, the next elongation step is blocked and the DNA synthesis is interrupted. The method proposed by Sanger relies on the statistical termination of DNA synthesis at one of the four nucleotides. Originally, four reaction mixtures were prepared each containing one of the ddNTPs and a radioactively labeled [α -³²P]dATP. Then, for instance regarding the reaction mixture containing ddCTP, the synthesis is terminated whenever this dideoxynucleotide is incorporated. Loading the reaction mixture on a denaturing PAGE, all oligonucleotides terminating with cytidine can be separated and visualized by autoradiography (see figure 9A). The other three mixtures terminate at adenosine, guanosine and thymidine correspondingly. The four reactions are loaded on a denaturing PAGE and upon autoradiography, the complementary sequence can be directly read (see figure 9).

Further improvements of this technique include the fluorescent labeling of the sequencing primer as demonstrated by Smith et al. in 1985 (see figure 9B) [89]. One year later, the same group extended their initial idea on fluorescent labeling. Instead of using a single fluorophore

²This method is an improvement of the original method called “plus and minus method” developed by Sanger et al. in 1975 [88]. More on this method can be read in Appendix section 8.1.

and separating the four reactions, the authors used a differently labeled primer for each ddNTP reaction mixture. Afterwards, the four reaction mixtures were combined and analyzed in a single PAGE tube (see figure 9C). Moreover, a detector was installed at the end of the gel and eluting DNA fragments could be continuously monitored. Due to the different dyes used, each fragment could then be uniquely determined. And since the detector was connected to a computer, the data could be stored and analyzed [90]. This was the first important step toward automation of the sequencing procedure. Of note, using a reverse transcriptase instead of a DNA polymerase allowed the direct RNA sequencing by the same approach as was demonstrated by Hamlyn et al. [91].

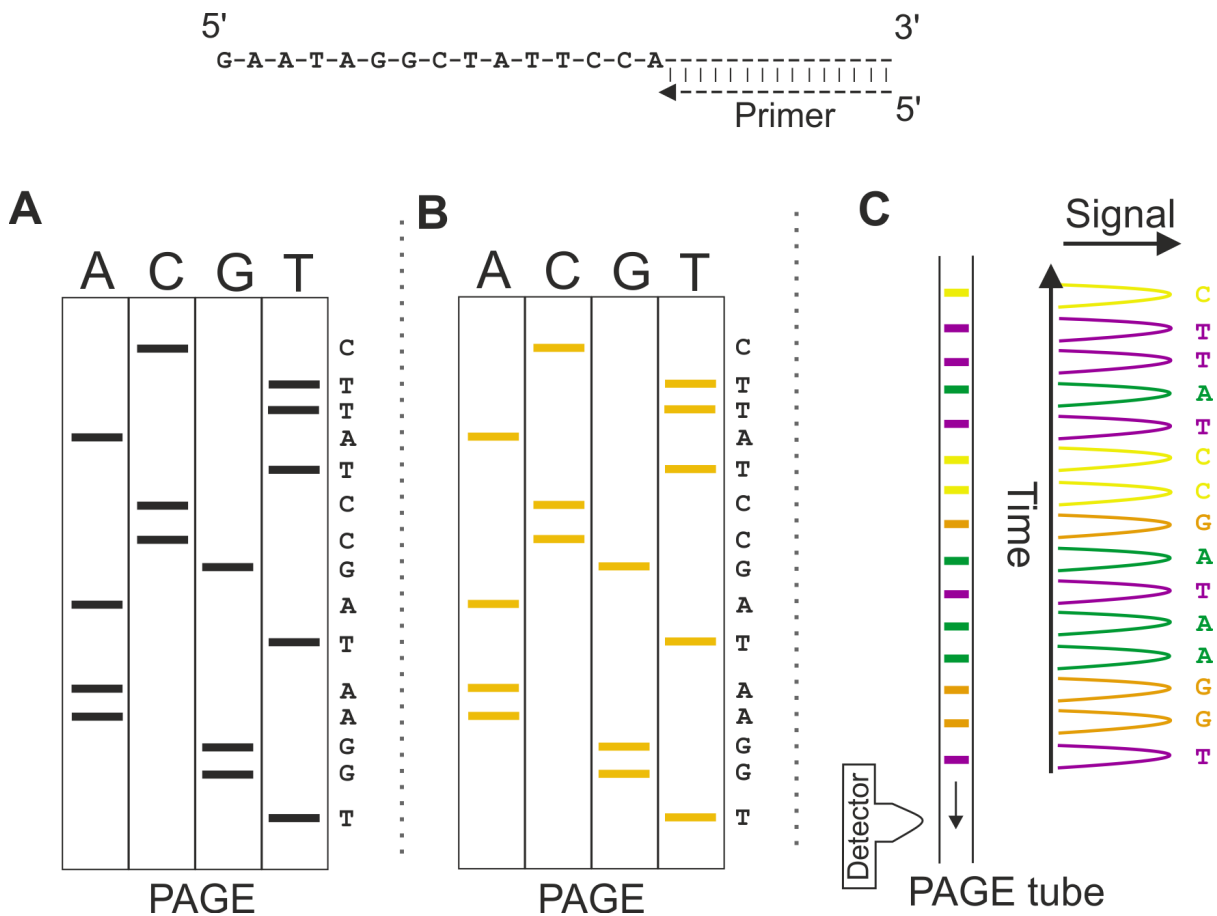


Figure 9: Sanger sequencing using chain terminating inhibitors. DNA synthesis is statistically interrupted by one of the four ddNTPs. Upon electrophoretic analysis the complementary sequence can be directly determined. A) In four reaction sets, one of the four ddNTPs is added together with [α - 32 P]dATP. After completion of the reaction, each set is separated on a PAGE and detected upon autoradiography. B) Instead of radioactivity, a 5'-end fluorescently labeled primer is used. Fluorescent detection reveals the complementary sequence. C) Each of the four reaction sets uses a primer with a different fluorescent dye. This allows mixing the four sets after completion of reaction and loading on a single PAGE tube. A detector coupled to a computer allows the automation of DNA sequencing. Figure adapted from [92, p. 255] and [90].

Sequencing by chain scission

In 1977, Maxam and Gilbert presented yet another method for direct DNA sequencing [93]. It relies on purification of the DNA of interest, followed by a base-specific degradation and detection on a denaturing PAGE. Based on this technique, an analogous chemical method for sequencing of RNA was presented [94]. Four different reaction conditions were adapted for the more labile RNA chain:

- Guanosine reaction: *N*-7 of guanosine is preferentially alkylated by dimethyl sulfate (DMS), followed by borohydride reduction (see figure 10).
- Adenosine > guanosine reaction: under aqueous conditions reaction of diethyl pyrocarbonate leads to carboxyethylation of *N*-7 of both adenosine and guanosine (See Appendix figure A.2).
- Uridine reaction: unprotonated hydrazine reacts *via* a nucleophilic addition to the 5,6 double bond of uridine.
- Cytidine > uridine reaction: presence of 3 M sodium chloride in anhydrous hydrazine leads to a preferred reaction toward the 5,6 double bond of cytidine (see Appendix figure A.3).

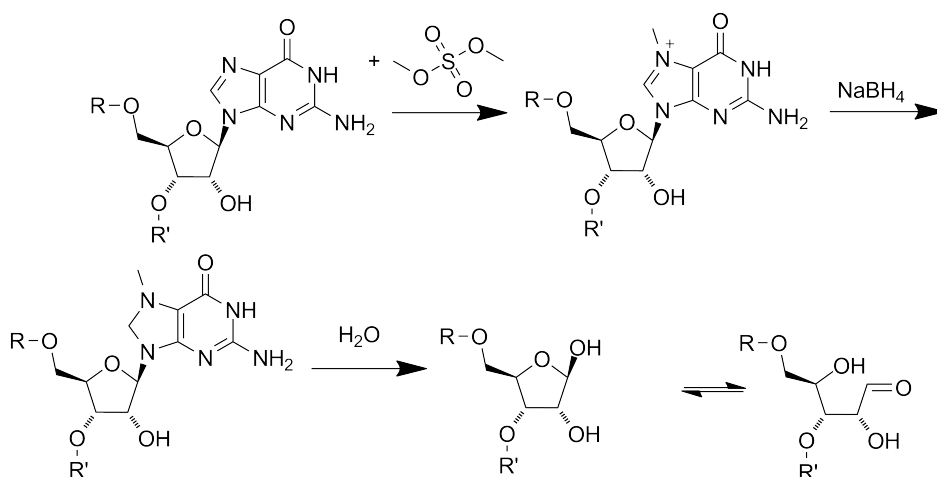


Figure 10: Under mild conditions DMS reacts predominantly with the *N*-7 of guanosine. Since the methylated product is stable at neutral pH, the following reduction with NaBH₄ leads to a site-specific removal of the nucleobase. Reaction adapted from [92, p. 270].

After each of these specific reactions, a chain cleavage at the modified site is performed by reaction with aniline, buffered at its pK_a of 4.6 with acetic acid (see figure 11).

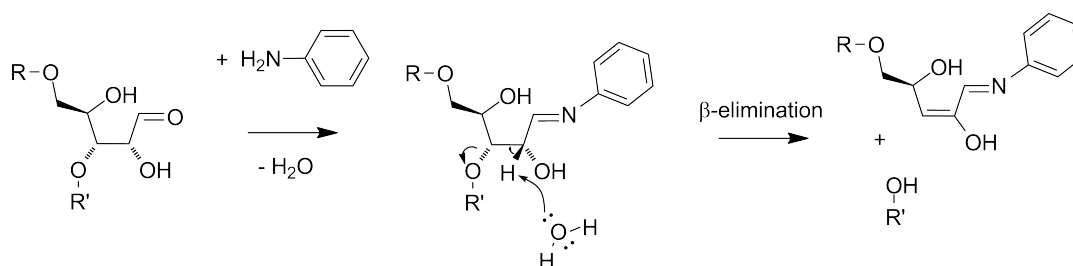


Figure 11: Aniline-assisted chain scission at abasic sites. When a nucleobase is removed, an attack of aniline at the aldehyde group of ribose is eased. Resulting Schiff's base speeds up the following β -elimination, leading to site-specific scission of the RNA chain. Reaction adapted from [92, p. 269].

Chemical labeling of nucleic acids

Especially the techniques developed for sequencing by synthesis, in combination with nucleoside-specific reactions, led to the development of methods for detection of RNA modifications based on a reverse transcription step [95]. It was demonstrated that modifications impairing the base-pairing properties of the underlying nucleoside have a hindering impact on a reverse transcriptase. This effect was observed for *N*-1-methyladenosine in eucaryotic tRNA at position 58 [96, 97], as well as for *N*-2-methylguanosine in 16S rRNA of *E. coli* [98]. Thus, one of the possibilities for detection of modifications, that do not interfere with the base-pairing properties of the corresponding nucleoside is to use a specific chemical, that upon reaction is expected to cause a reverse transcriptase to alter its behavior. For applications of this approach please refer to subsection 1.5, especially the detection of inosine (see 1.5.1) and pseudouridine (see 1.5.2).

1.4 High-throughput sequencing of RNA and DNA

Developments in RNA sequencing technologies in the last 5 to 10 years have greatly contributed to newly discovered RNA-associated processes. Numerous RNA species were characterized and their functions are already beginning to unveil [99]. Concerning RNA sequencing, high-throughput methods still require a reverse transcription step followed by sequencing of produced cDNA. Therefore, unless otherwise stated, the next few subsections describe high-throughput methods for DNA sequencing.

1.4.1 Illumina[®] sequencing technology

First mention of the Illumina[®] sequencing principle appeared in 2008 [100]. In this method DNA containing specific primers are hybridized to a flow cell. By means of polymerase chain reaction (PCR), clonal amplification is performed by so-called cluster-generation. DNA in each cluster is then denatured so that hybridized strands are removed leaving only those covalently bound to the flow cell. After addition of a universal primer, DNA is sequenced by repeated cycles of polymerase-based single nucleotide extension steps (see figure 12).

For this, a set of four reversible terminators, 3'-*O*-azidomethyl 2'-deoxynucleoside triphosphates each labeled with a different removable fluorophore, is used. Because the nucleotides are modified at their 3'-OH, they are added simultaneously, without risking sequential incorporation. After each cycle, the identity of each incorporated nucleotide is determined by laser-induced excitation of the corresponding fluorophore and imaging. Tris(2-carboxyethyl)phosphine is added prior to each following cycle to regenerate the 3'-OH and remove the fluorescent dye

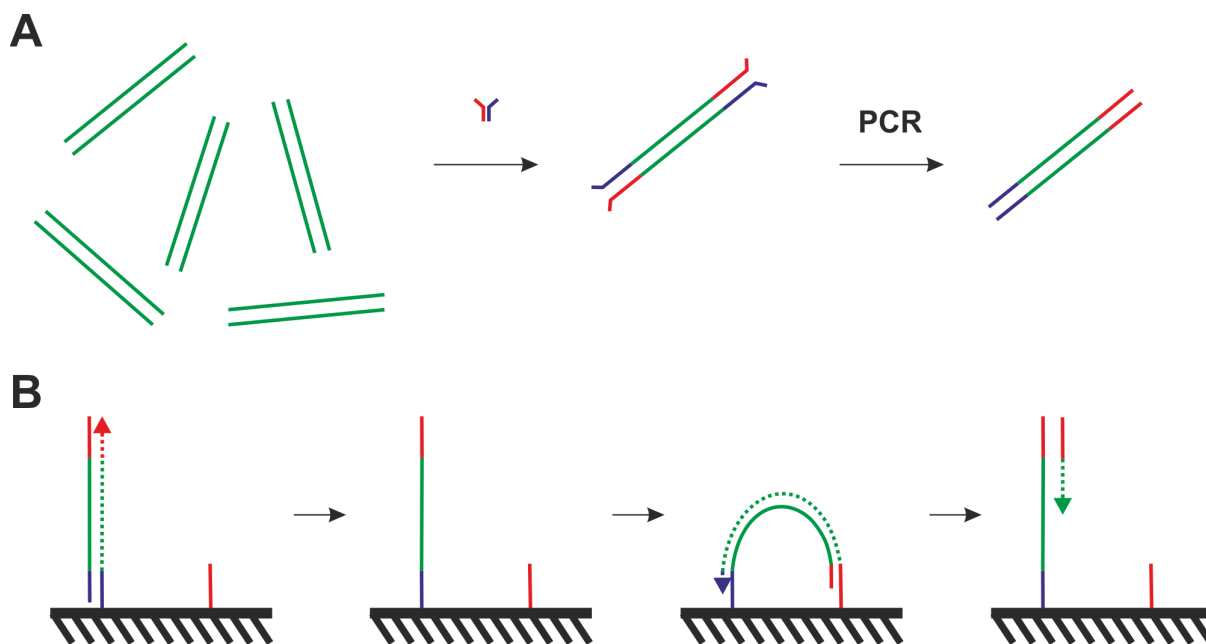


Figure 12: Illumina[®]-based DNA library preparation and sequencing on the flow cell. A) Fragmented DNA is tagged with Y-formed dsDNA adapters compatible with the Illumina[®] sequencing technology followed by a PCR amplification step. B) Denatured ssDNA is hybridized on the flow cell and upon bridge amplification step clusters are formed. Prior to actual sequencing each hybridized DNA is removed, sequencing adapter is added and sequencing is initiated. Figure adapted from [100].

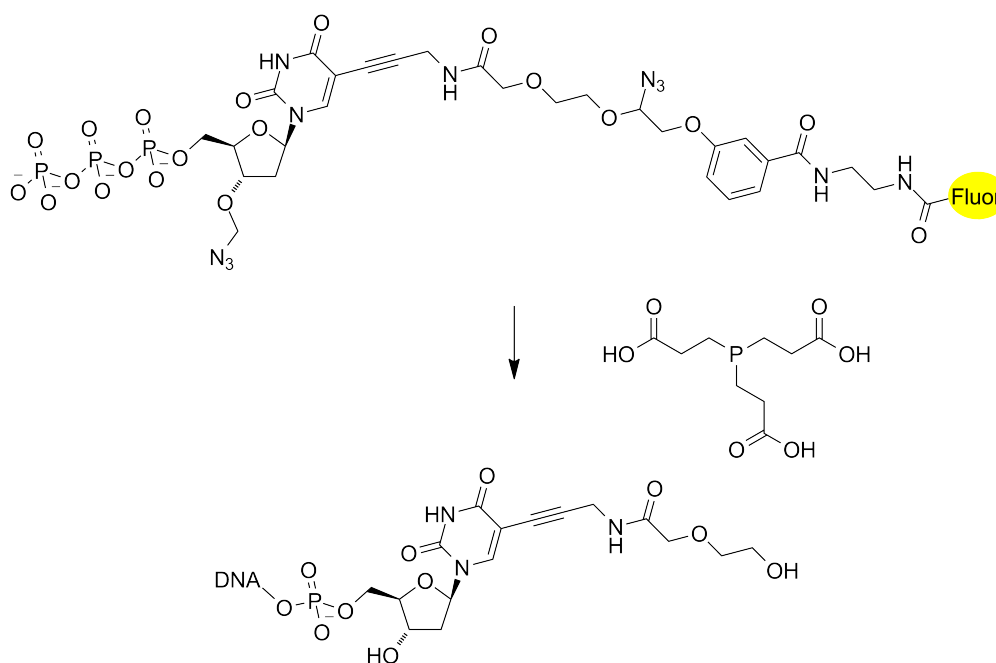


Figure 13: Structure of the reversible terminator 3'-O-azidomethyl 2'-deoxythymidine triphosphate labeled with a removable fluorophore (Fluor). After incorporation into DNA, tris(2-carboxyethyl)phosphine is added. A Staudinger reaction [101] leads to regeneration of the 3'-OH and removal of the fluorophore. Reaction scheme adapted from [100].

(see figure 13). The reaction follows a Staudinger mechanism, where the azide is reduced to an amine, which upon hydrolysis releases an alcohol, amine and an aldehyde [101].

1.4.2 454-Pyrosequencing

454-sequencing, developed in 2005 by 454 Life Sciences (later bought by Roche Diagnostics) [102], is based on a method presented first by Ronaghi et al. in 1996 [103]. The principle of pyrosequencing exploits the release of a pyrophosphate upon incorporation of a deoxynucleoside-phosphate into a growing DNA chain (see figure 14). DNA, hybridized to a sequencing primer, is incubated with a mixture of DNA polymerase, ATP sulfurylase, firefly luciferase and an apyrase (a dNTP-degrading enzyme) [104]. Next, a repetitive cycle of deoxynucleoside-triphosphate additions is performed. A dNTP will only be incorporated into a growing DNA chain, if it is complementary to the template nucleotide. This process is associated with a equimolar release of a pyrophosphate. Released pyrophosphate is converted to ATP by ATP sulfurylase, which in turn is detected by the luciferase. Light produced by this enzyme can be detected by either a luminometer or a CCD (charge coupled device) camera. Unused deoxynucleotide-triphosphates, as well as produced ATP, are degraded and the cycle is repeated by offering the next deoxynucleoside triphosphate.

High-throughput of this method is accomplished by parallel capturing of single DNA fragments bound to beads in emulsion droplets where a initial PCR ensures a clonal amplification. Next, beads are separated and moved into wells of fibre optic slide, where parallel sequencing takes place [102].

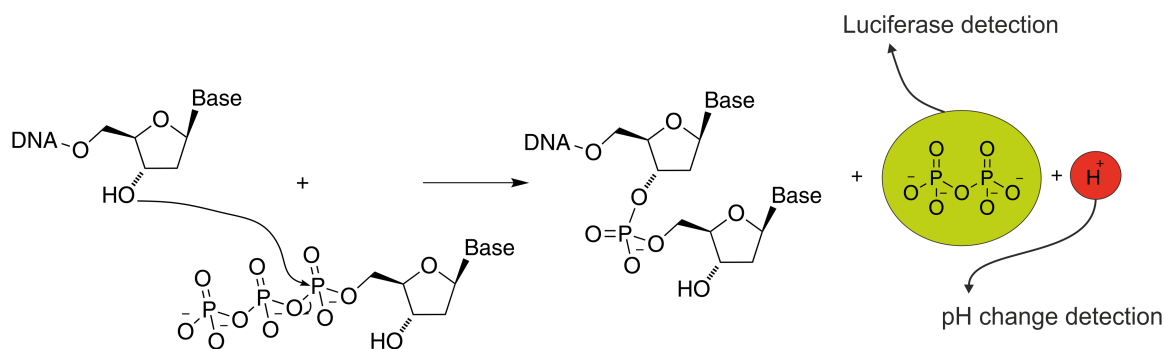


Figure 14: Principle of 454-Pyrosequencing and ion torrent sequencing. Upon incorporation of the correct nucleotide to the DNA, both pyrophosphate and a proton are released. 454-Pyrosequencing detection is based on a coupled luciferase reaction, where pyrophosphate is first used for synthesis of ATP. Ion torrent directly detects pH changes upon release of hydrogen protons.

1.4.3 Ion torrent sequencing

Ion torrent sequencing was developed in 2010 by Ion Torrent Systems Inc. [105] and is based on ion detection. The principle is related to the pyrosequencing method. Upon incorporation of correct dNTP into a growing DNA chain, a proton is released that changes the pH in the solution (see figure 14). This change is then detected by an ion sensor.

1.4.4 SOLiDTM sequencing technology

In contrast to the methods presented above, SOLiDTM sequencing technology uses a ligation approach to determine the sequence of a DNA. In this case, after clonal amplification each DNA template is ligated to a universal adapter sequence. A primer is then hybridized to the adapter and a set of 16 eight-mer fluorescently labeled probes compete for ligation to the primer. Specificity is determined by complementarity of the 1st and 2nd base to the template DNA in each ligation reaction. Multiple cycles of ligation, detection and cleavage of a tri-mer containing the fluorophore take place, after which the extension product is removed and a new primer of length $n - 1$ is hybridized for a second round of ligation cycles. Five rounds of primer reset are completed. Through this process, each base is interrogated in two independent ligation reactions by two different primers. First application of this sequencing technology was used by Valouev and colleagues in creating a high-resolution nucleosome position map of *C. elegans* in [106]. In 2012 Squires and collaborators applied the method in combination with bisulfite treatment for transcriptome-wide detection of 5-methylcytidine in *H. sapiens* [46].

1.4.5 Single molecule sequencing

The methods presented in the above subsections all depend on an amplification step prior to sequencing. This step is necessary due to the nature of detection, which requires multiple substrates in order to generate a sufficiently strong signal that is detectable by the corresponding system. Such amplification steps bear some disadvantages, such as bias during a PCR step. Therefore, alternative methods were and are still under development that would allow a single molecule sequence determination.

PacBio sequencing

One example is PacBio sequencing developed by Pacific BioSciences [107]. The principle, related to Illumina[®], relies on a polymerase immobilized at the bottom of a zero-mode waveguide unit that provides the smallest possible volume for light detection. DNA, tagged with hairpin adapters, is replicated by the polymerase using fluorescently labeled nucleotides. While a nucleotide is being incorporated by the polymerase, a distinct fluorescent signal is recorded. An important advantage of this sequencing technology is the produced read length. The current version of PacBio can generate over 10 kb long reads [107]. Based on the same principle, a direct determination of RNA sequences upon reverse transcription monitoring was published by Vilfan and collaborators [108]. The authors claim that by using PacBio not only the sequence can be determined but also that, based on the different timespan a polymerase needs for incorporation of complementary dNTP, one can distinguish between canonical and modified nucleosides.

Nanopore-based sequencing

Nanopore-based sequencers, also called fourth-generation DNA sequencing technologies, offer another possibility for single molecule sequencing. In contrast to the PacBio sequencer, nanopore technologies allow detection without labeling, and at lower costs [109]. The principle of detection is based on ion channels. Nanometer-sized pores are either embedded in a biological membrane or formed in a solid-state film. In either case, two reservoirs containing conductive electrolytes are connected by the pore. Electrodes are emersed in each compartment. When voltage is applied, electrolytes are moved through the nanopore, thus creating current signal. When an analyte, in this case negatively charged DNA, blocks the pore the current signal is interrupted. The physicochemical properties of the target molecules determine the amplitude

and duration of transient current blockages between translocation events. Apart from the established DNA sequencing applications [110, 111, 112], Oxford Nanopore Technologies reported recently that their Nanopore technology is capable of direct RNA sequencing [113].

1.4.6 RNA sequencing using Illumina[®] sequencing technology

Among the above presented current high-throughput sequencing methods, Illumina[®] technology is the most widely used one. Therefore, in this subsection library preparation options are presented that allow Illumina[®]-RNA sequencing.

A prerequisite for the usage of Illumina[®] sequencing technology is the conversion of RNA to a dsDNA molecule. For this step the ability of a reverse transcriptase to synthesize complementary DNA out of template RNA is used. Additionally, due to the limited length that can be sequenced, a fragmentation of longer RNA (rRNA, mRNA, lncRNA) prior to reverse transcription is performed. Reverse transcriptases, as already mentioned, need a template and a primer on which an elongation at the free 3'-OH end can take place. Therefore, two approaches are currently used:

- An adapter is ligated to the free 3'-OH of RNA (see figure 15AB). Typically, the adapter contains information for subsequent Illumina[®] sequencing.
- Random priming is performed (see figure 15C). In this case, a primer is used that contains the information for a subsequent Illumina[®] sequencing. Additionally the first six nucleotides at the 3' end are random, so that hybridization can take place at random positions of the RNA chain. Of note, in this case G-C rich regions are expected to be preferably reverse transcribed due to the stronger hybridization between guanosines and cytidines.

Furthermore, depending on whether the second adapter is added at RNA level (see figure 15A) or after cDNA synthesis (see figure 15B) either only full-length products are sequenced or also abortive cDNA are captured and sequenced.

At the level of cDNA there are currently two main approaches for introducing the second adapter:

- A ligation step adds the adapter to the free 3'-OH of cDNA.
- After RNA is removed, a new template containing six random nucleotides with the last one blocked at its 3' end is hybridized to the 3' end of cDNA which is then elongated complementary to the template sequence.

A different approach, including a circularization step of synthesized cDNA, was described by Ingolia and co-workers [114]. According to their protocol, upon RNA ligation of an adapter at the free 3'-OH, a reverse transcription step is performed. This includes using a 5'-phosphorylated primer that additionally contains information for the second Illumina[®]-specific sequence. Using an enzyme called CircLigase, newly synthesized cDNA is circularized and finally PCR amplified using Illumina[®]-specific primers. An important optimization of this protocol includes several random nucleotides at the 5' end of RNA adapter as described by Lecanda et al. [115]. The authors show in their study that this random sequence minimizes a bias, that is otherwise introduced by sequence-specific ligation.

Hrdlickova et al. have recently published an extensive review on the currently used library preparation protocols for RNA sequencing [116]. Additionally, in 2014 Illumina[®] issued an extensive overview of the applications achieved by high-throughput sequencing of RNA using their technology [117].

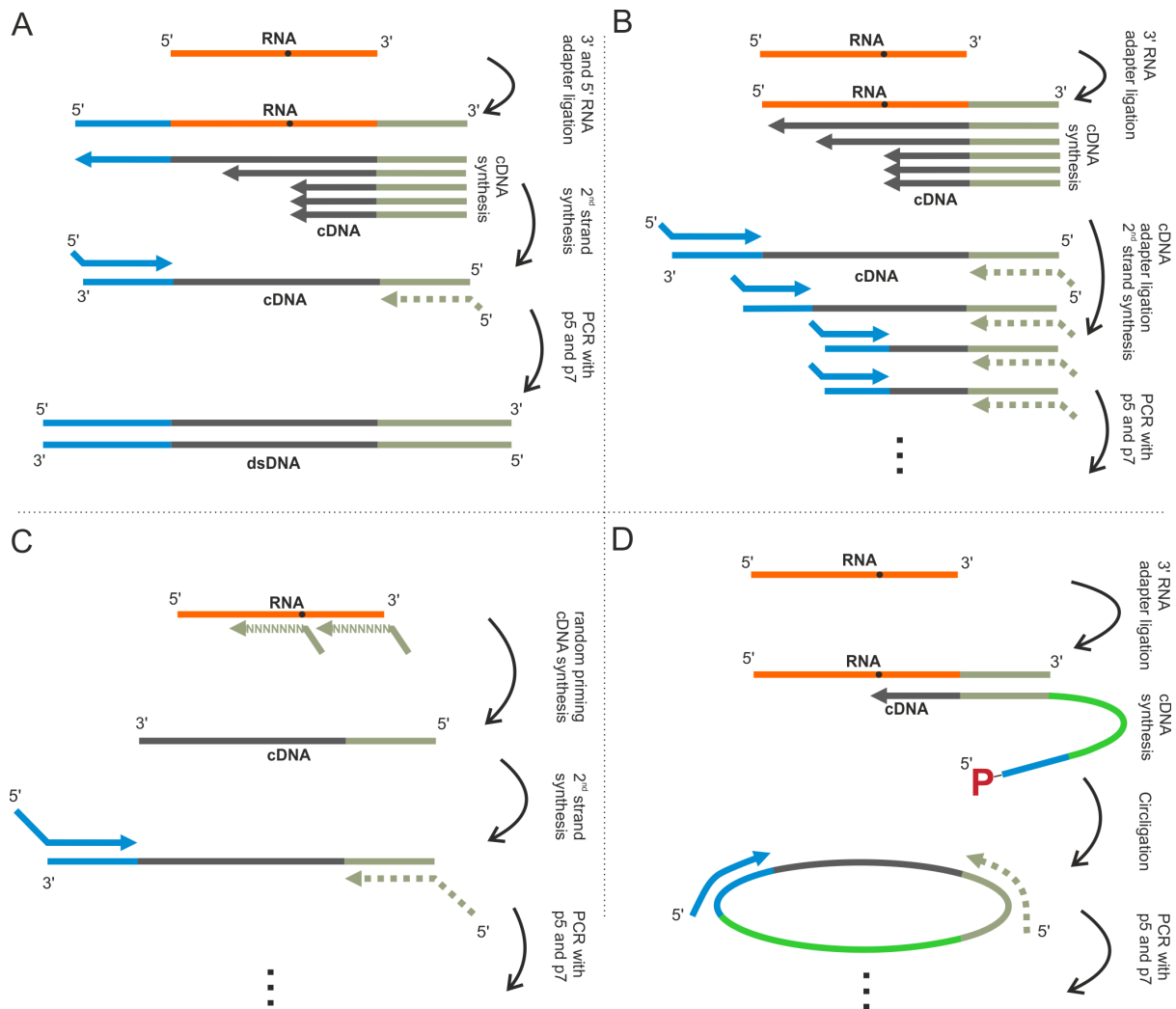


Figure 15: Library preparation strategies for conversion of RNA to dsDNA. A) Both 3' and 5' adapters are added to an RNA molecule. Following an RT step, a PCR amplification step is performed using both full-length Illumina[®] primers p5 and p7. B) One adapter is ligated to the 3' end of RNA. Following a reverse transcription step, a second adapter is added to the 3' end of the newly synthesized cDNA. Finally, a PCR with p5 and p7 primers delivers the dsDNA ready for sequencing. C) Instead of RNA ligation, direct reverse transcription using random primers delivers cDNA. Next, sequence information is added to the 3' end of cDNA either by template elongation or by ligation. D) A different approach offers the CircLigase protocol. Here, a longer RT primer contains the information for both Illumina[®] sequences. After cDNA synthesis, a circularization is performed using the pre-phosphorylated 5' end of RT primer and the free 3'-OH of newly synthesized cDNA. Finally, PCR delivers the dsDNA molecule.

1.5 Transcriptome-wide detection of RNA modifications

First hints that high-throughput sequencing data of RNA contains information for modified nucleosides came from the work of Iida and co-workers [118]. Using the 454-pyrosequencing technology and allowing one mismatch in the produced reads, the researchers found that reads mapping to tRNA of *Arabidopsis thaliana* contained nucleotide-misincorporation sites frequently located at the T arm (U or G instead of correct A), upstream of the D arm (A, U, or C instead of correct G), and downstream of the D arm (U instead of G). It was shown that these misincorporations originate from the underlying modifications: m¹A, m²G and m²₂G, respectively.

Similar results were published by Ebhardt et al. [119] who analyzed data sets for tRNA and miRNA originating from *A. thaliana* and *O. sativa*. A high mismatch composition in tRNA^{Pro} and tRNA^{Val} was related to the conserved modification *N*-1-methylguanosine in eucaryotic organisms. Also, position 58 of tRNA^{Glu} was strongly mismatched which could be linked to the existence of yet another highly conserved modification in eucaryots *N*-1-methyladenosine. Additionally, in the same study numerous editing sites as a result of C-to-U and A-to-I deaminations were discovered in microRNA originating from *A. thaliana*. Findeiss and collaborators [120] investigated a small RNA sequencing set originating from two further organisms: *Homo sapiens* and *Rhesus macaque*, but did not focus much on the mismatch incorporations. Instead, they showed a long known effect from primer extension experiments that RNA modifications impairing a reverse transcriptase lead to truncated cDNA. They observed a nearly 7-fold enrichment of read-starts on tRNA position 59, compared to the modified position 58, that corresponded to reverse transcription products terminating there.

In this case, it is important to know the type of library preparation prior to sequencing. While all preparations for RNA sequencing include the conversion of RNA to cDNA *via* a reverse transcriptase, depending on the order of introduction of needed adapters one may either capture only full-length cDNAs or also abortive products (for more details on this, please refer to subsection 1.4.6).

The first systematic approach to classify and annotate modified nucleotides in RNA based on high-throughput sequencing (HAMR method) was done by Ryvkin et al. [121]. It is important to mention that their analysis is based on misincorporation statistics alone. The library preparation they used included ligation of both 3'- and 5'-Illumina[®] adapters to RNA molecules, followed by a cDNA conversion and PCR. Therefore, only full-length cDNAs could be amplified and sequenced on an Illumina[®] GenomeAnalyzer II. Their method, based entirely on the misincorporation behavior, confirmed results from previous studies. For each modification encountered, the frequency of the three observed non-reference nucleotides were calculated and plotted on a tertiary graph. This way the composition of misincorporation for each modification could be visualized. So, a hypothesis was stated: Depending on the underlying modification a different misincorporation pattern is observed allowing its detection upon comparison analysis. A limitation of their approach is the need for at least two distinct misincorporations.

While next-generation sequencing allowed the transcriptome-wide identification and discovery of new RNA species as well as certain RNA modifications, a precise localization of these modifications requires a more targeted approach. The next subsections present methods currently used for the transcriptome-wide detection of several RNA modifications.

1.5.1 Detection of inosine

Inosine is the product of adenosine deamination, either hydrolytic or enzymatic by enzymes known as ADARs (Adenosine deaminases acting on RNA) [122]. Difficulties in detection of inosine arise from the base-pairing properties of this modification. It was demonstrated that inosine base-pairs better with cytidine than with thymidine (see Appendix figure A.4) [123]. Thus, it appears exclusively as guanosine in sequencing data and is therefore difficult to discriminate it from a single nucleotide polymorphism (SNP). One way to circumvent this was the parallel sequencing of transcriptome and genome presented by Li and coworkers in 2009 [124]. First, a high-throughput RNA sequencing delivered candidate sites. Then, a DNA selection step was performed by usage of specific DNA padlock probes (please refer to Appendix 8.5 for an explanation on a padlock probe). Thus, candidate RNA editing sites were targeted in both genomic DNA and cDNA samples from a single individual. Finding an A in the genome and a G in RNA pointed toward a deamination site, whereas finding G in both RNA and DNA was

an indication for a SNP. The method screened for more than 30,000 computationally predicted sites in the human transcriptome and could find a total of 239 such edited positions [124].

Later, an alternative protocol by Sakurai and co-workers [125] was developed based on differential reactivity of inosine and adenosine toward acrylonitrile³ [126] (see figure 16). Sakurai

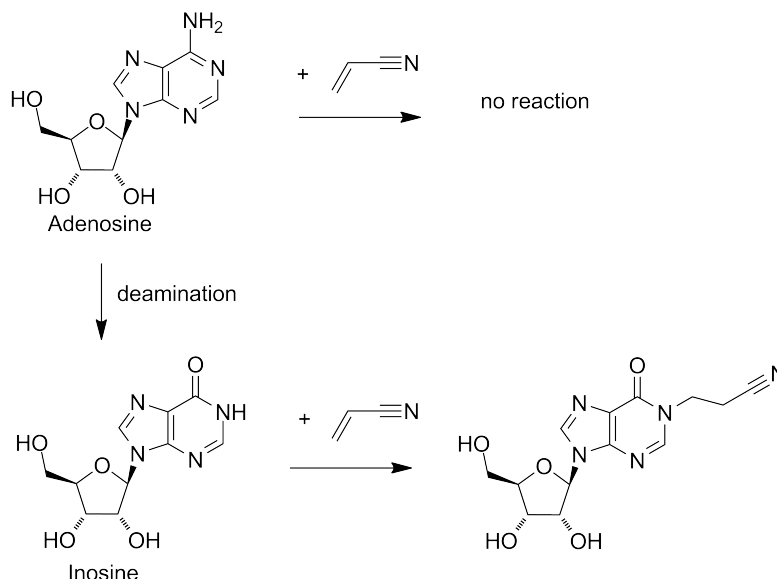


Figure 16: Reaction of inosine with acrylonitrile in a Michael addition manner leads to *N*-1-cyanoethylinosine (ce¹I). Adenosine does not undergo this reaction which allows a discrimination upon reverse transcription of treated RNA. Of note, following the same reaction mechanism, also pseudouridine and uridine react with acrylonitrile to form *N*-1-cyanoethylpseudouridine and *N*-3-cyanoethyluridine, respectively. Nevertheless, under the chosen conditions, Sakurai et al. reported these as minor reactions. Reaction adapted from [125].

et al. [125] performed primer extension experiments after cyanoethylation with tRNA^{Val1} from yeast. In contrast to a control sample, cDNA produced from the cyanoethylated sample showed a specific band ending at A35, proving that ce¹I indeed blocked the reverse transcription. Based on these results, a larger scale verification of the method was performed and over 4,000 new sites in total RNA from *H. sapiens* were determined.

1.5.2 Detection of pseudouridine

Pseudouridine (Ψ) is the product of an isomerization reaction during which *C*-5 and *N*-1 positions of uracil are interconverted. In contrast to all other chemical modifications in RNA, Ψ and its derivatives possess a C-C glycosidic bond. Nonetheless, pseudouridine has the same base-pairing properties as its unmodified equivalent uridine [12] (see also figure 3). As such it is impossible to distinguish both residues directly *via* sequencing. Ryvkin et al. [121] claimed that pseudouridine leaves a characteristic pattern upon reverse transcription. For a more precise identification, however, a chemical treatment must be applied prior to sequencing. This way, after the treatment, a reverse transcriptase is expected to stall at the labeled positions. The most widely spread chemical is *N*-cyclohexyl-*N*'-beta-(4-methylmorpholinium)ethylcarbodiimide p-tosylate (CMC).

³Although the reaction of acrylonitrile with Ψ and U was also reported by Yoshida et al. [126], under the conditions chosen by Sakurai et al., only small amounts of these reacted to *N*-1-cyanoethylpseudouridine and *N*-3-cyanoethyluridine, respectively [125].

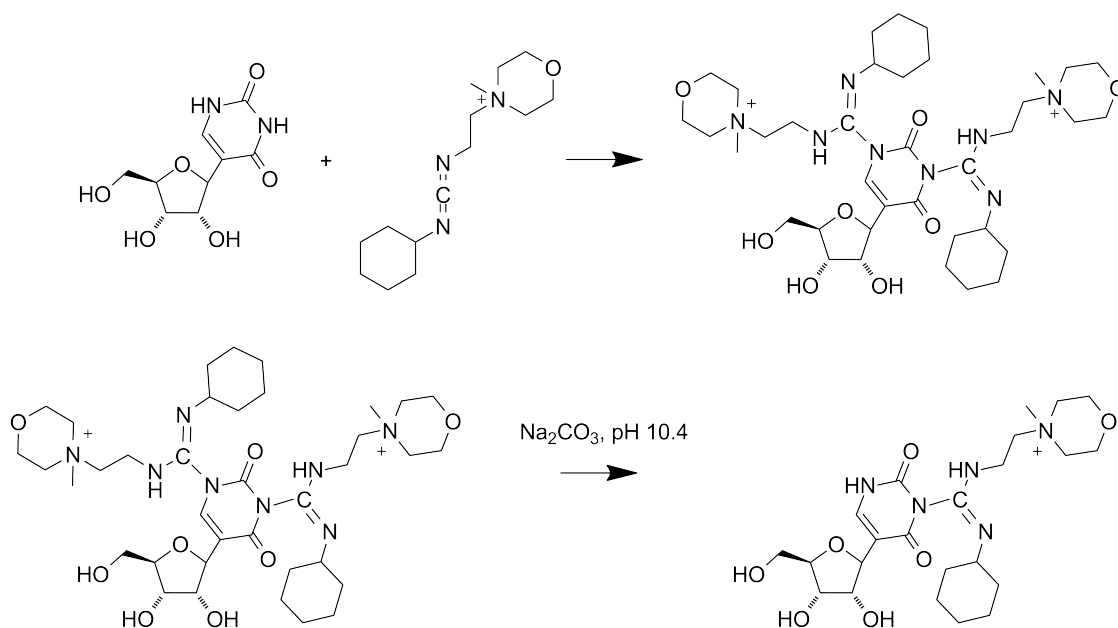


Figure 17: Reaction of pseudouridine with CMC leads to *N,N*-1,3-di-CMC-Ψ. Following a mild alkaline hydrolysis, the *N*-1 substituent is cleaved, leaving only the single substituted *N*-3-CMC-Ψ, that effectively blocks the reverse transcriptase. Reaction adapted from [127].

The method was originally developed by Bakin and Ofengand in 1993 [128]. It exploits the stability of an *N*-3-CMC-Ψ in comparison to uridine and guanosine. *N*-3 of uridine and *N*-1 guanosine, as well as both *N*-1 and *N*-3 of pseudouridine are labeled by CMC, but a mild alkaline hydrolysis at pH 10.4 leads to the only labeled product *N*-3-CMC-Ψ [127]⁴. Since the CMC-conjugate is located in the Watson-Crick Edge of pseudouridine (see figure 17) it can be detected by primer extension experiments [95]. This approach was recently successfully combined with next-generation sequencing for a transcriptome-wide search for pseudouridines in mammalian organisms [42, 41, 43]. Li and co-workers added an azide functionality to CMC [129]. After labeling of Ψ with N₃-CMC, they performed a conjugation reaction with biotin. By using the strong biotin-streptavidine interaction, the authors of the study were able to enrich pseudouridinylated RNA species. They verified their method for a Ψ-site in mRNA by the SCARLET method (please refer to Appendix 8.6 for an explanation of the method).

1.5.3 Detection of 5-methylcytidine

Another example on specific chemical labeling, but with different analytical outcome is the reaction of bisulfite on RNA. Discrimination between cytidine and 5-methylcytidine relies on differential reactivity of these toward sodium bisulfite (NaHSO₃). Thorough investigation on the reactivity of bisulfite with nucleobases was performed by Hayatsu et al. [130]. Under neutral pH, 1 M NaHSO₃ reacts with cytosine and uracil in an addition reaction to the 5,6 double bond of the corresponding nucleobase. Furthermore, it was demonstrated that under mild alkaline conditions cytosine derivative 5,6-dihydrocytosine-6-sulfonate deaminates to the corresponding uracil derivative followed by an elimination reaction that restores it to uracil

⁴Also the CMC-adduct at the *N*-3 of pseudouridine is cleaved under more harsh conditions: 7 M NH₄OH at 100° for 8 min [127].

[131]. In contrast, under these conditions 5-methylcytosine is nearly unreactive⁵. On this basis, a genomic sequencing protocol was developed by Frommer and co-workers [132]. After bisulfite conversion of cytosine residues to uracil specific primers were used for regions of interest. Therefore, all uracil- (formerly cytosine-) and thymine residues were amplified to thymine, whereas 5-methylcytosines were amplified to cytosine. The discrimination was then possible upon classical sequencing. Bisulfite conversion was combined with high-throughput sequencing for a thorough genome methylation investigation. An extensive review on DNA methylation and detection methods was written by Olkhov-Mitsel and Bapat in 2012 [133]. Corresponding applications in RNA analysis were more difficult due to unspecific RNA hydrolysis upon reaction with bisulfite [134]. In 2005, Gu and collaborators showed that after bisulfite reaction with RNA, a primer extension including dideoxyguanosine-triphosphate leads to a stop at the modified positions [135]. This approach was recently successfully applied on a whole transcriptome analysis [136].

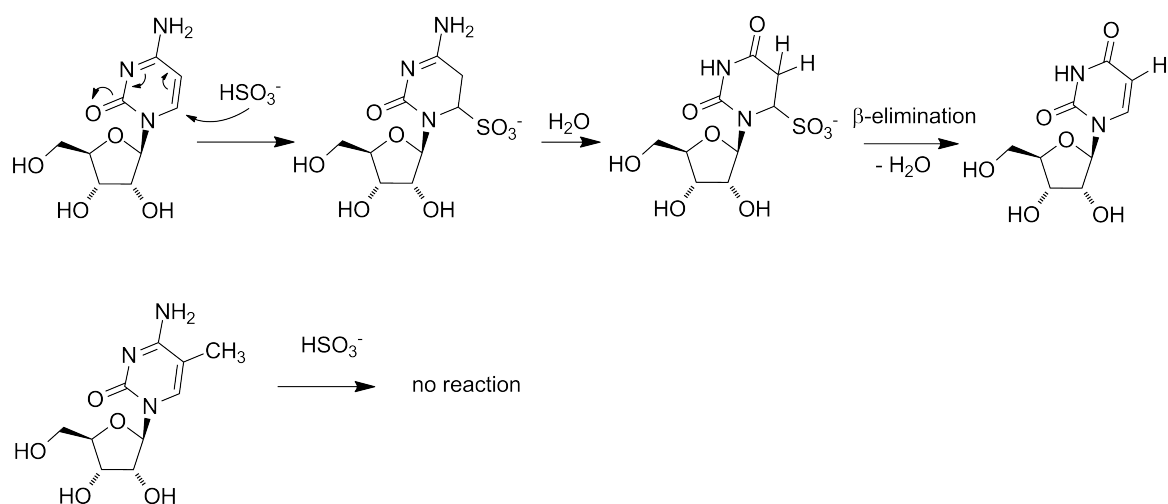


Figure 18: At neutral pH bisulfite (HSO_3^-) adds to the 5,6 double bond of cytidine. Deamination, followed by a β -elimination leads to uridine. In contrast, under same conditions m^5C is nearly unreactive. Reaction adapted from [137].

An alternative method that exploits the methylation activity of methyl-transferases NSun2 and DNMT2 is used in an immunoprecipitation reaction to determine sites of m^5C . For this, 5-azacytidine is added to a cell culture. Cells incorporate this cytidine analog in their RNA. At methylation sites, a covalent bond between methylation enzyme and 5-azacytidine is established. Thus captured RNA fragments are recovered and applied to deep sequencing [138].

Another method, also based on immunoprecipitation, is the miCLIP (methylation individual nucleotide cross-linking immunoprecipitation). Here, the methylation enzyme NSun2 is mutated (C271A) so that upon binding with the target RNA, a covalent bond is built. This allows capturing of RNA and upon next-generation sequencing, target methylation sites are detected [139].

Hussain et al. recently published a comparative review on the three used methods for a transcriptome-wide detection of 5-methylcytidine [140].

⁵Also 5-methylcytosine was converted to thymine under more harsh conditions (3 M NaHSO_3) [130].

1.5.4 Detection of *N*-6-methyladenosine

Because the methyl group at *N*-6 of adenosine does not interact with the base-pairing properties of the underlying nucleoside, it is not detectable by classical extension experiments. Additionally, there is still no chemical reagent specific to *N*-6-methyladenosine (m^6A) over adenosine. Previous detection of (m^6A) was based on the radioactive labeling of *S*-Adenosyl methionine (SAM), a co-substrate used by methylation enzymes (e.g. METTL3), followed by TLC or HPLC detection of labeled product [141].

First successful transcriptome-wide mapping of m^6A was performed in the labs of Jaffrey [142] and Rechavi [37]. The method of Meyer et al. [142] is based on specific antibodies against m^6A developed previously [143]. After testing the antibodies against RNA containing m^6A and determining their sensitivity and specificity, they were combined with next-generation sequencing. The procedure involves random fragmentation of RNA to approximately 100 nucleotides of length. After immunoprecipitation, enriched RNAs are deep sequenced and upon mapping regions containing potential m^6A sites are evaluated.

At the same time, Dominissini and co-workers presented a similar approach toward determination of m^6A positions transcriptome-wide [37]. Their method is an improvement of the previously described one because of the additional sequencing of input RNA. This way, fold enrichment of m^6A containing RNA fragments could be deduced by the results of deep sequencing. A remaining drawback, however, was the lack of single nucleoside resolution.

Very recently, Linder et al. [38] reported further improvement in the detection of m^6A toward single nucleotide resolution. After capturing RNA fragments containing m^6A using antibodies, UV cross-linking leads to a covalent bond between RNA and antibody. Covalently bound antibodies impede reverse transcriptases, leading either to truncated cDNA or specific misincorporation at modified sites. This combined with high-throughput sequencing allows for the first successful analysis of *N*-6-methyladenosine on a single nucleotide resolution [38]. Additionally, using the same approach, the authors were able to detect *N*-6,2'-*O*-dimethyladenosine. This modification is often found at the first nucleotide of certain mRNAs [38]. The authors also applied the SCARLET method (please refer to Appendix 8.6 for an explanation of the method) to validate some of the positions they detected.

1.5.5 Detection of *N*-1-methyladenosine

As was already mentioned in subsection 1.3.2 base-pairing properties of *N*-1-methyl-adenosine (m^1A) are disturbed because of the methyl group located in the Watson-Crick Edge (see figure 2). And although the existence of this modification was long known in tRNA [22] and rRNA [31] little was known about its existence in other RNA species.

Very recently two papers came out describing the transcriptome-wide distribution of m^1A . Dominissini et al. [47] developed an antibody-based approach for enrichment of m^1A sites. As an extra validation, they treated antibody-enriched RNA under alkaline conditions which leads to the conversion of m^1A to m^6A in a Dimroth rearrangement (see figure 19) [144]. This way, converted RNAs now containing m^6A showed lower misincorporation rates when compared to non-treated samples. A bioinformatic approach calculated the fold change of antibody-enriched RNAs to identify methylated regions (m^1A peaks). Drawback of this method is that only regions can be identified, thus single nucleoside resolution is not achieved.

In the same time, based on the ability of m^1A to stall reverse transcriptase, Li et al. [48] developed a m^1A -ID sequencing method. This method is also based on an immunoprecipitation step to enrich regions containing m^1A . Control sample was subjected to ALKB that demethylates m^1A [145] converting it to adenosine. ALKB-treated sample exhibited accumulation of

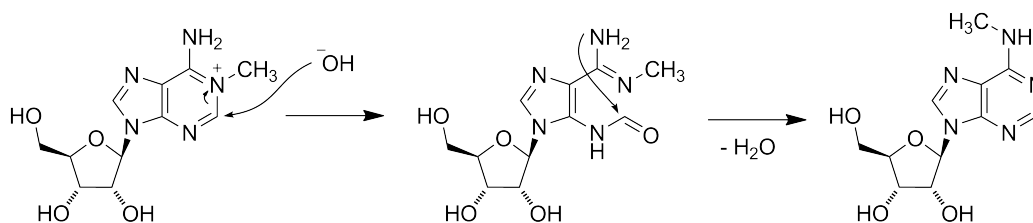


Figure 19: Dimroth rearrangement of m^1A to m^6A under mild alkaline conditions. Rate determining step is the opening of the pyrimidine ring at position 2 upon hydroxide attack. Reaction adapted from [144].

full-length cDNAs in comparison to the untreated one due to the missing methyl group at $N-1$ of adenosine. Thus, by comparing both samples, the authors identified demethylase-sensitive regions that correspond to m^1A containing RNAs.

Both reported studies, although lacking single nucleotide resolution (except for high abundant RNA, such as tRNA and rRNA), report the existence of m^1A in mRNA and other non-coding RNA of *Homo sapiens* [48], *Mus musculus* and *Saccharomyces cerevisiae* [47].

1.5.6 Detection of 2'-*O*-methylation

As already mentioned in subsection 1.3 2'-*O*-methylated ribose nucleosides are stable against mild alkaline hydrolysis. In presence of a base, the 2'-OH of unmethylated riboses are deprotonated (see figure 20). Then the 2'- O^- attacks the electrophilic neighboring 3'-phosphate which leads to a chain scission. One chain leaves with a free 5'-OH, the other with a cyclic 2',3'-phosphate. In contrast, 2'- $O-CH_3$ has reduced nucleophilicity, hence the phosphodiester bond 3' of the methylated ribose is protected from cleavage. Therefore, a following primer extension experiment leads to gaps in a PAGE ladder at modified positions allowing their detection [95].

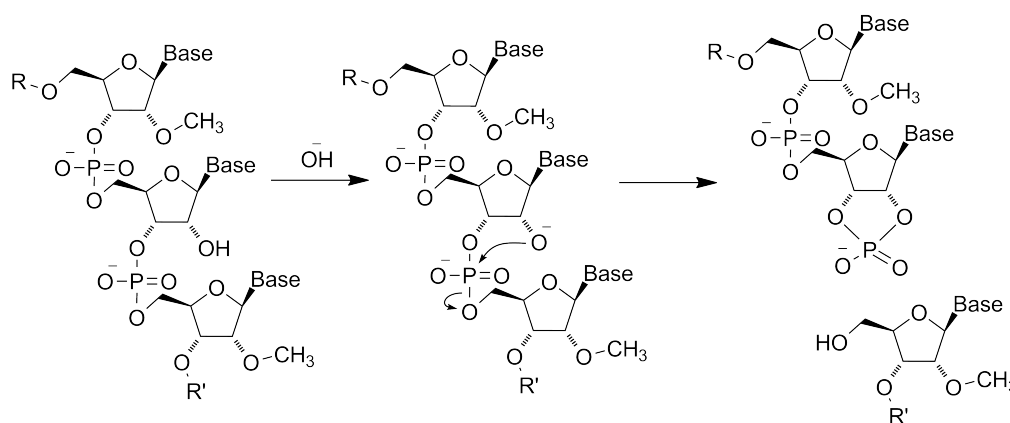


Figure 20: Alkaline hydrolysis of RNA. Under mild alkaline conditions 2'-OH is deprotonated. 2'- O^- attacks the neighboring 3'-phosphate which leads to cleavage of a chain with a free 5'-OH and a chain with cyclic 2',3'-phosphate. In contrast, 2'- $O-CH_3$ is protected from deprotonation, therefore no cleavage is observed. Reaction adapted from [137].

The important observation that 3' extremities of cleaved RNA fragments are non-methylated at their 2'-*O*-position was successfully used in combination with high-throughput sequencing for detection of ribosomal methylation in yeast [79, 80]. Using ion torrent, Birkedal et al. sequenced a library prepared from partially alkaline hydrolyzed RNA. For every mapped read, the

3' extreme base was counted in one set. Therefore, the 2'-*O*-methylnucleotides were underrepresented, due to the reasons already explained. Because the 5'-end of each read results from a free 2'-OH-containing nucleotide 5' next to it, the +1 base upstream of positions corresponding to 5' extremities of each RNA fragment were counted in a second set. Both sets showed a negative image of the 2'-*O*-methylated sites. Finally, a bioinformatic approach combined these sets, converted the data and showed the methylated positions as peaks. Marchand and co-workers adapted the protocol for usage with the Illumina[®] sequencing technology [80]. An advantage of their approach was the usage of commercially available library preparation protocols for the preparation of the RNA fragments for high-throughput sequencing.

An alternative approach on detection of ribose methylations is the usage of limiting concentrations of dNTPs during reverse transcription. This leads to a pause at the modified positions, visible upon PAGE analysis by the appearance of bands at those sites [146]. Also, recently a DNA polymerase was mutated to possess reverse transcriptase activity and to stop at 2'-*O*-methylated positions [147]. This method presents a promising idea for the design of a specific polymerase that reacts to certain modifications [147]. Of note, neither approach has yet been combined with next-generation sequencing.

2 Materials and Methods

Unless otherwise stated, all materials and methods used are described in the included papers and are correspondingly referenced.

3 Goal of The Work

The first goal of this work was to set up a library preparation protocol for high-throughput sequencing of RNA that allows the detection of naturally occurring modifications. Based on a reverse transcription (RT) step, an important feature of it had to be the possibility to capture RT events including stops and misincorporations. It was envisaged that such effects occur predominantly at RNA sites containing modifications that alter the base-pairing properties of the underlying nucleoside. One example is the methyl group at *N*-1 of adenosine (see figures 2 and 21A). The information produced by next generation sequencing could then be used for learning purposes on high abundant tRNAs that allow the search for such modifications transcriptome-wide. Finally, a bioinformatic machine-learning approach in collaboration with Ralf Hauenschild was planned.

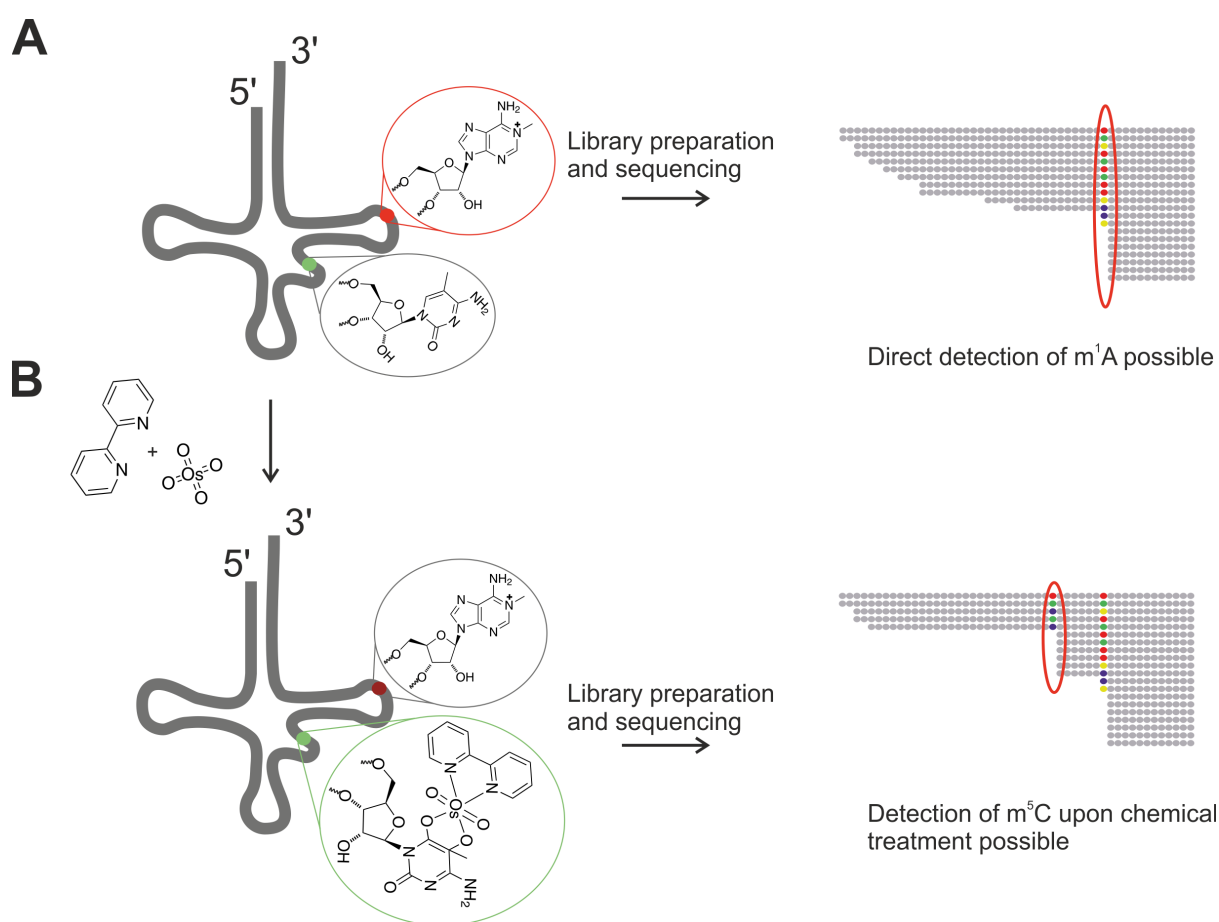


Figure 21: Development of a library preparation protocol for capturing of reverse transcriptase (RT) events (e.g. RT stops and misincorporations) occurring on RNA modification sites. A) Possibility for direct detection of modifications that alter the Watson-Crick base-pairing, such as m¹A. B) Possibility for detection of RT silent modifications, such as m⁵C upon chemical labeling with OsO₄ and 2,2'-bipyridine.

Methyl groups that are located outside the Watson-Crick edge of a nucleoside, as in the case of 5-methyl pyrimidines, remain silent for reverse transcriptases [95]. Thus, it was a second goal of this work to develop a chemical labeling approach that combined with the above-mentioned high-throughput sequencing method would allow the detection of these modifications (see figure 21B). Specifically, since osmium tetroxide-bipyridine was already successfully used to label

deoxy-5-methyluridine and deoxy-5-methylcytidine in DNA, the question was raised whether this chemical is applicable in RNA.

Both tasks were part of a joint project initiated among several PhD students. The goal of it was the establishment of a platform for the detection, localization and quantification of naturally occurring RNA modifications. An overview of this joint project, as well as each individual task is presented in figure 22. Two main approaches were addressed. On one side, a precise LC-MS method for qualitative, as well as quantitative analysis was aimed. This task was performed by Katharina Schmid and Kathrin Thüring. On the other side, a high-throughput method, based on reverse transcription combined with next generation sequencing for precise localization and prediction of modifications was intended. Ralf's objective was then to develop the bioinformatic workflow, that enables the evaluation and learning from well described sites, and optimally predict the occurrence of such at unknown positions. Additionally, Kathrin focused on the development of methods, that allow the isolation of RNAs containing such predicted sites and by means of LC-MS either confirm or reject them.

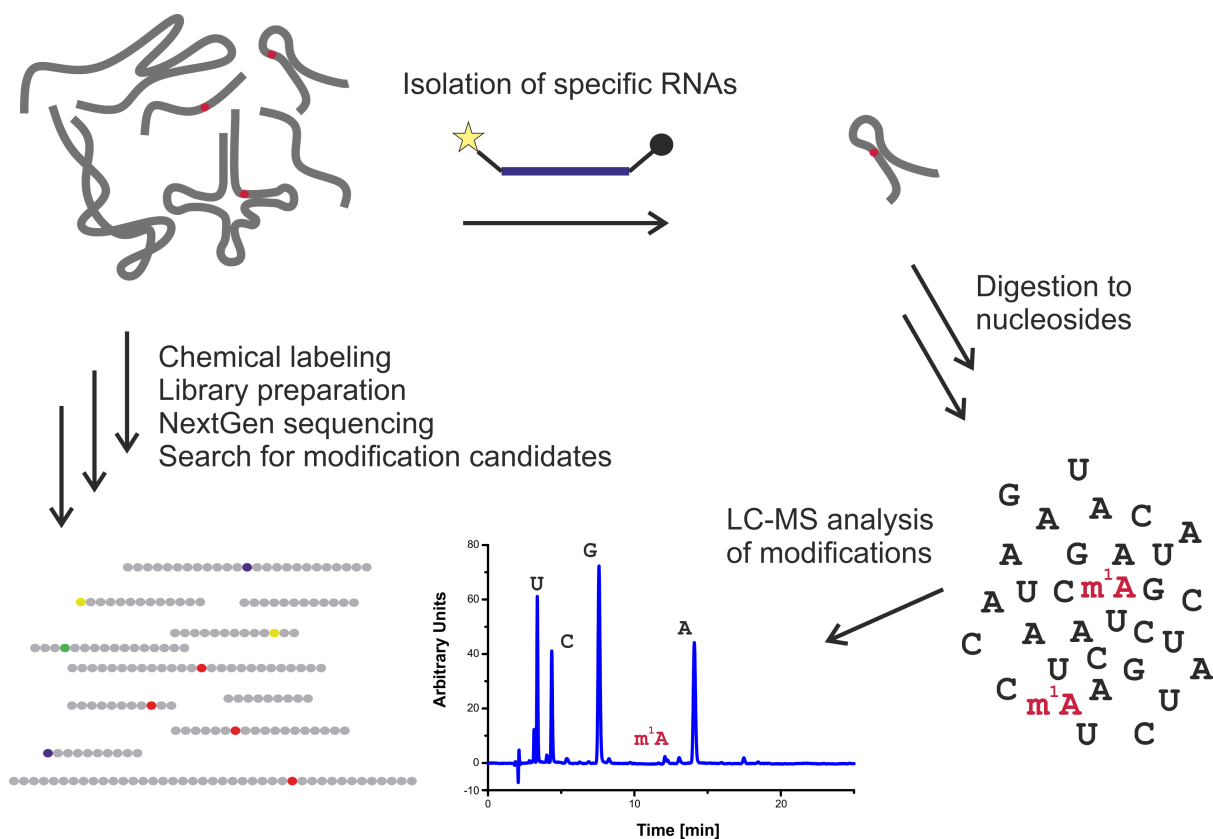


Figure 22: General scheme of the platform for detection of RNA modifications.

4 List of Publications

List as a co-author

1. Tserovski L, Helm M.; Diastereoselectivity of 5-Methyluridine Osmylation Is Inverted inside an RNA Chain.; *Bioconjug Chem.* 2016 Sep 21;27(9):2188-97.
doi: 10.1021/acs.bioconjchem.6b00403.
2. 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.* 2016 Sep 1;107:110-21. doi: 10.1016/j.ymeth.2016.02.012.
*These authors contributed equally.
3. Hauenschild R*, Tserovski L*, Schmid K, Thüring K, Winz ML, Sharma S, Entian KD, Wacheul L, Lafontaine DL, Anderson J, Alfonzo J, Hildebrandt A, Jäschke A, Motorin Y, Helm M.; The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent.; *Nucleic Acids Res.* 2015 Nov 16;43(20):9950-64.
doi: 10.1093/nar/gkv895.
*These authors contributed equally.
4. Hauenschild R, Werner S, Tserovski L, Hildebrandt A, Motorin Y, Helm M.; Coverage-Analyzer (CAn): A Tool for Inspection of Modification Signatures in RNA Sequencing Profiles; *Biomolecules* [Article Accepted]

Contribution as a co-author

1. L.T. designed and conducted the experiments, performed the analysis and wrote the manuscript together with M.H.
Contribution to the publication: 90 %.
2. L.T. was involved in the design of the experiments, performed the library preparations and was involved in the analysis. L.T. also contributed to writing the manuscript.
Contribution to the publication: 40 %.
3. L.T. together with R.H. designed the experiments, conducted the lab experiments and was involved in the analysis. L.T. also contributed to writing the manuscript.
Contribution to the publication: 40 %.
4. L.T. conceived and performed the biomolecular experiments
Contribution to the publication: 10 %.

Declaration of confirmation:

Prof. Mark Helm

Lyudmil Tserovski

5 Results and Discussion

5.1 Detection of *N*-1-methyladenosine (m^1A) in RNA *via* reverse transcription and next generation sequencing

5.1.1 Library preparation protocol for capturing of reverse transcription events at modified sites in RNA *via* next generation sequencing

At the time this project was initiated, there were only a few RNA library preparation protocols suitable for high-throughput sequencing available. And because of the prerequisites that such protocol had to fulfill, namely to capture abortive products, as well as misincorporations, the choice was even more limited. Therefore, a library preparation protocol, developed in the group of Prof. Jäschke, Heidelberg, that was later described in reference [148], was used as a starting point. Their library preparation protocol featured an adapter ligation step at the free 3'-OH group of RNA, that served as a template for the following reverse transcription step. Next, a second adapter was introduced at the free 3'-OH of the newly synthesized cDNA. Both adapters were then used for a final PCR amplification step. This approach had an advantage in comparison to other often used protocols that add both adapters at the RNA molecule. Such protocols capture only full length cDNAs, thus missing all abortive information produced by an RT stop. This was not the case with the protocol of Jäschke's lab, because it introduced the second adapter after cDNA was already synthesized. Yet, there were several issues in their protocol to be considered and changed. During the conversion from RNA to dsDNA their protocol included the ligation of non-standard sequence adapters. This came as an advantage, if different sequencing platforms were planned to be used. In our case, Illumina was the preferred sequencing platform. This meant one additional library preparation step for the ligation of Illumina specific adapters to the dsDNA. Obvious disadvantages were: time consumption for the additional conversion step, increased costs, and lost of sequence information due to the existence of non-standard adapters and their sequencing. Furthermore, Jäschke's protocol used six random nucleotides at the first adapter ligation step to label RNA species. This step is important since it allows the bioinformatical recognition and removal of PCR duplicates. Because their protocol was predominantly used for non-abundant RNA species, these six nucleotides, allowing $4^6 = 4096$ distinct combinations were enough. Our project, however focused on high abundant, but less sequence diverse tRNAs. For this reason, the number of random nucleotides was increased to nine thus allowing $4^9 = 262144$ combinations. This, in combination with the sequence information of the corresponding tRNA allows the efficient recognition and removal of PCR duplicates in a bioinformatical workflow.

Based on these considerations, the following changes in our library preparation protocol were made:

- The RNA adapter sequence was changed to a sequence that allows the hybridization of sequencing primer 1 during Illumina sequencing. Resulting read 1 represents the cDNA, complementary to the original RNA template, starting from its 3' end.
- The cDNA adapter sequence was changed with a sequence that allows the hybridization of an optional sequencing primer 2 during Illumina sequencing. Resulting read 2 represents the template RNA, complementary to the 3' end of reverse transcribed cDNA.
- The random nucleotide sequence in RNA adapter was increased from six to nine, therefore increasing the unique barcoding adapters by a factor of 64.

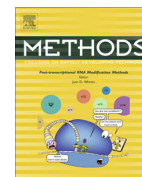
The exact method including a bioinformatic pipeline, as well as an application in the detection and identification of naturally occurring RNA modification *N*-1-methyladenosine was recently published [149]. Of note, the bioinformatic workup was developed by Ralf Hauenschild.



Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth



High-throughput sequencing for 1-methyladenosine (m¹A) mapping in RNA



Lyudmil Tserovski^{a,1}, Virginie Marchand^{b,c,1}, Ralf Hauenschild^a, Florence Blanloeil-Oillo^{b,c}, Mark Helm^a, Yuri Motorin^{b,c,*}

^a Institute of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Staudingerweg 5, 55128 Mainz, Germany

^b IMoPA UMR7365 CNRS-UL, BioPole Lorraine University, 9 avenue de la Foret de Haye, 54505 Vandoeuvre-les-Nancy, France

^c Next-Generation Sequencing Core Facility, FR3209 BMCT, Lorraine University, 9 avenue de la Foret de Haye, 54505 Vandoeuvre-les-Nancy, France

ARTICLE INFO

Article history:

Received 13 January 2016

Received in revised form 22 February 2016

Accepted 22 February 2016

Available online 24 February 2016

Keywords:

RNA modification

High-throughput sequencing

Misincorporation signature

Reverse transcription

ABSTRACT

Detection and mapping of modified nucleotides in RNAs is a difficult and laborious task. Several physico-chemical approaches based on differential properties of modified nucleotides can be used, however, most of these methods do not allow high-throughput analysis. Here we describe in details a method for mapping of rather common 1-methyladenosine (m¹A) residues using high-throughput next generation sequencing (NGS). Since m¹A residues block primer extension during reverse transcription (RT), the accumulation of abortive products as well as the nucleotide misincorporation can be detected in the sequencing data. The described library preparation protocol allows to capture both types of cDNA products essential for further bioinformatic analysis. We demonstrate that m¹A residues produce characteristic arrest and mismatch rates and combination of both can be used for their detection as well as for discrimination of m¹A from other modified A residues present in RNAs.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Modified nucleotide m¹A (1-methyladenosine, Fig. 1) is a rather common RNA modification present in all living domains, but predominantly found in non-coding RNAs in eukaryotic cells [1]. In bacteria, the m¹A modification of rRNA is associated with antibiotic resistance and thus contributes to the virulence of different species [2,3]. m¹A is also present in tRNAs from various archaeobacteria [4,5]. These modified residues are known to contribute both to tRNA thermostability and proper folding [6], however, their exact functions in RNAs remain unclear. Enzymatic formation of m¹A is ensured by dedicated m¹A:RNA-MTases with strict RNA substrate specificity, allowing exclusive modification of only a limited

number of sites in RNAs [7,8]. However, the exact rules for RNA recognition by these m¹A:RNA-MTases are still poorly understood and one cannot exclude that other still poorly characterized RNA species may contain this modified residue.

Methylation of adenosine at position 1 can also be introduced by non-enzymatic chemical methylation, e.g. by dimethylsulfate or other methylating agents. This methylation reaction was extensively used in the past for structural probing of RNA in solution, followed by detection of accessible m¹A sites by the reverse transcription blocks [9,10].

Several methods allow the detection and localization of m¹A residues in RNAs. This modified residue can be detected by physico-chemical approaches, like bi-dimensional thin-layer chromatography (2D-TLC), high-pressure liquid chromatography (HPLC) or HPLC coupled with mass-spectrometry (HPLC-MS(MS)). However, the information on its location is generally lost in these cases. Alternatively, localization of RT-pausing A residues (almost essentially m¹A and m²A) was done by primer extension in RT reaction [11]. Other technologies have also been proposed, notably the usage of DNA chip with specific oligonucleotides, capable to distinguish m¹A from unmodified A by its lower efficiency of hybridization. This method can be eventually used for high-throughput

Abbreviations: DMS, dimethylsulfate; 2D(3D), bi (tri) dimensional; TLC, thin layer chromatography; HPLC, high-pressure liquid chromatography; MS (MS), mass spectrometry (tandem mass spectrometry); RT, reverse transcription; NGS, next generation sequencing; CMCT, N-cyclohexyl-N'-(2-morpholinoethyl)-carbodiimide metho-*p*-toluenesulfonate; RF, Random Forest; PAGE, PolyAcrylamide Gel Electrophoresis.

* Corresponding author at: IMoPA UMR7365 CNRS-UL, BioPole Lorraine University, 9 avenue de la Foret de Haye, 54505 Vandoeuvre-les-Nancy, France.

E-mail address: Yuri.Motorin@univ-lorraine.fr (Y. Motorin).

¹ These authors contributed equally.

<http://dx.doi.org/10.1016/j.ymeth.2016.02.012>

1046-2023/© 2016 Elsevier Inc. All rights reserved.

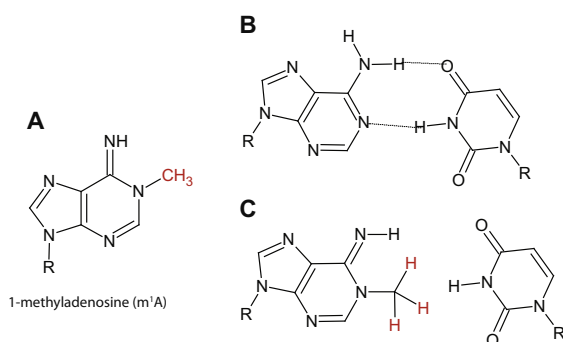


Fig. 1. 1-methyladenosine (m^1A) properties. Chemical structure of m^1A and its disturbed basepairing with U and other bases.

applications, but still requires the pre-existing knowledge on the positions of modified sites to design specific oligos [12].

Recent progress in coupling a specific chemical treatment with NGS allowed mapping of several modified nucleotides in RNA at the transcriptome-wide level (Table 1). First attempts concerned bisulfite mapping of m^5C , and antibody-based enrichment of m^6A residues in RNA [13–19]. Independent efforts of three labs shed light to massive presence of pseudouridines in eukaryotic mRNAs, detected by N-cyclohexyl-N'-(2-morpholinoethyl)-carbodiimide metho-*p*-toluenesulfonate treatment coupled to NGS [20–22]. More recently, detection of 2'-O-Me residues in highly abundant RNAs using NGS was described [23]. Massive and dynamic N¹-methyladenosine modification of eukaryotic mRNAs was just demonstrated using antibody enrichment [24,25]. These efforts clearly demonstrated of anticipated for a long time, but not proved, presence of multiple RNA modifications in all cellular RNAs and their possible important regulatory role in gene expression regulation.

Here we provide a detailed description of experimental and bioinformatic protocols used for high-throughput mapping of m^1A residues in tRNAs of different origins [26].

2. Materials and methods

2.1. RNA extraction

Total human or bacterial RNA was extracted with TRIzol[®] Reagent (Life technologies, Thermo Scientific #15596-026, Dreieich, Germany) according to the manufacturer's protocol. Total RNA extraction from yeast *Saccharomyces cerevisiae* was done using hot acidic phenol [27].

2.2. RNA quality control and quantification

After RNA extraction, RNA samples were quantified on Nanodrop 1000 and loaded at appropriate concentration either on Agilent RNA6000 Pico or Small RNA chips (used with Agilent Bioanalyzer 2100) to check RNA sample quality. Representative traces for total RNA and tRNA fractions are given in Fig. 2.

2.3. Overview of library preparation protocol

Here, we developed a library preparation protocol suitable for the detection of both, abortive cDNA products and nucleotide misincorporation into cDNA (Fig. 3). After an optional RNA fragmentation step and dephosphorylation, a pre-adenylated 3'-adapter was ligated to the de-phosphorylated 3'-extremity of RNA, this adapter

Table 1
High-throughput approaches for mapping of modified residues in RNA.

Modified nucleotide	Detection technique	Reference
m^5C	Bisulfite RNA sequencing	[16,17,19]
m^6A	Anti- m^6A antibody enrichment	[14,15]
m^6A	Anti- m^6A antibody enrichment + PAR-CLIP using s^4U	[13,18]
Ψ	CMCT-treatment induces RT-stop	[20–22]
Ψ	Pull-down of Ψ -containing RNAs using Click-CMCT	[35]
2'-O-Me	Alkaline hydrolysis of RNAs	[23]
m^1A	Specific RT-signature for modified residue	[26]
m^1A	Anti- m^1A antibody enrichment	[24,25]

includes a single C at the 5' end followed by a random 9 nt sequence (N_9), used as individual barcode for every cDNA molecule produced during the RT step. RT step was done with a specific oligonucleotide annealed to 3'-adapter and cDNA 3'-end (corresponds to 5'-end of RNA) was C-tailed using terminal deoxynucleotidyl transferase (TdT) and CTP. A double stranded DNA adapter was then ligated to 3'-tailed cDNA, and the amplicon sequence was amplified by PCR using barcoding p5 and p7 Illumina primers. The resulting amplicon contained p5 and p7 sequences required for flow cell clustering, 8 nt barcodes for multiplexed sequencing, sequences for read1 and read2 sequencing primers and the central part corresponding to the RNA template used for RT. Read1 corresponds to the 3'-end of RNA and read2 to the 5'-end (3'-end of cDNA). Thus, in the majority of cases the generated amplicons were sequenced in paired-end mode (depending on the expected length of the insert, but in general 2×75 nt or 2×80 nt).

2.3.1. RNA fragmentation (optional step)

If required, optional RNA fragmentation was achieved by incubation with 10 mM $ZnCl_2$ in 100 mM Tris-HCl buffer at pH 7.4 at 90 °C for 5 min. Reaction was stopped with EDTA at final concentration 50 mM, and RNA directly fractionated on a 10% denaturing PolyAcrylamide Gel Electrophoresis (PAGE), fraction of about 50–150 nt in size was excised, eluted with 0.5 M NH_4Ac and ethanol precipitated.

2.3.2. Dephosphorylation

RNA (0.1–1 μg) was denatured at 90 °C for 30 s, chilled on ice and de-phosphorylated by 0.5 U of FastAP alkaline phosphatase (Thermo Scientific, #EF0651) for 30 min at 37 °C in final mixture of 10 μL containing 100 mM Tris-HCl, pH 7.4, 20 mM $MgCl_2$, 0.1 mg/mL BSA and 100 mM 2-mercaptoethanol. After 30 min incubation, the denaturation step at 90 °C for 30 s was repeated and another portion of enzyme (0.5 U) was added, followed by another 30 min incubation at 37 °C. Enzyme was then inactivated by a final heating at 75 °C for 5 min.

2.3.3. Chemical preadenylation of RNA adapter

30 μL of 1 mM oligodeoxynucleotide (3'-blocked with a hexylrest, and chemically 5'-phosphorylated, purchased from IBA, Germany) were mixed with equal volume of 200 mM ImpA dissolved in 100 mM $MgCl_2$ (Imidazole of 5'-AMP free acid, a generous gift from Dr. A. Jäschke, Heidelberg University, Germany). Mixture was incubated at 50 °C for 1.5 h, then further 15 μL 200 mM ImpA and 15 μL water were added and the reaction was performed for another 1 h. RNA adapter was separated from unused ImpA by gel filtration (using of Illustra NAP5 column, GE Healthcare #17-0853-02). Adenylation yield was about 50% and pre-adenylated primer was not separated from initial oligodeoxynucleotide before the ligation step. Analysis was performed on a 15% denaturing PAGE.

5 RESULTS AND DISCUSSION

112

L. Tserovski et al./Methods 107 (2016) 110–121

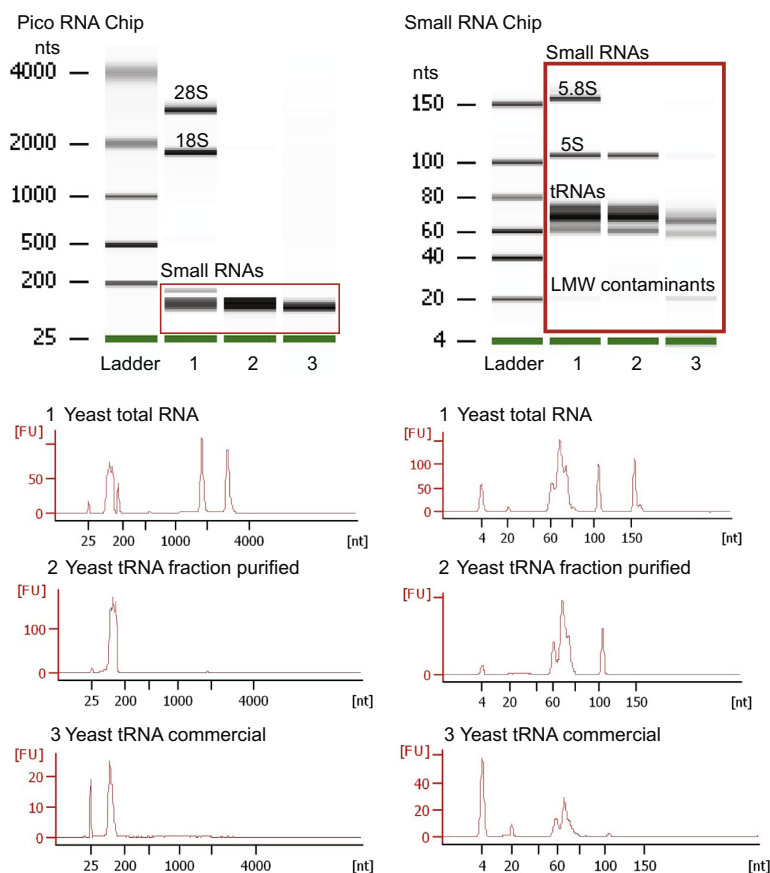


Fig. 2. RNA quality control using Bioanalyzer 2100. Representative traces obtained for total RNA fraction from yeast and two tRNA fractions on Agilent RNA 6000 Pico chip (left) and on Small RNA chip (right). Ladder size is given on the left of each panel.

2.3.4. 3'-adapter ligation

Ligation of pre-adenylated adapter at the RNA 3'-end was performed without additional purification step, using the protocol described in [28]. Ligation was performed in dephosphorylation buffer overnight at 4 °C in a final volume of 20 μ L containing, in addition, 5 μ M of adenylated 3'-RNA adapter, 15% DMSO, 1 U/ μ L T4 RNA ligase 2 truncated (New England Biolabs, #M0242L), 0.5 U/ μ L T4 RNA ligase (Thermo Scientific, #EL0021). RNA ligase was inactivated at 75 °C for 15 min.

2.3.5. Removal of unligated 3'-adapter

Non-ligated pre-adenylated adapter was removed by the combined action of deadenylation enzyme and 5'-P-exonuclease. Ligation mixture from the previous step was complemented by 20 U of 5'-Deadenylase (New England Biolabs #M0331S) and incubated for 30 min at 30 °C. After denaturation for 15 min at 75 °C, the second portion of the enzyme was added and incubation continued for another 30 min at 30 °C. Deadenylated 3'-adapter was cleaved by Lambda exonuclease (Thermo Scientific #EN0561) for 30 min at 37 °C, followed by an additional denaturation step and another portion of the enzyme incubated for 30 min at 37 °C. Final heat inactivation was done at 80 °C for 15 min and RNA was ethanol precipitated with 20 μ g of glycogen as a carrier.

2.3.6. Reverse transcription

RNA pellet from precipitation step was redissolved in 32 μ L of First Strand-Buffer 1x (50 mM Tris-HCl, pH 8.3 at room

temperature, 75 mM KCl, 3 mM MgCl₂, Life Technologies) containing RT primer at 5 μ M final concentration. Annealing was done at 80 °C for 10 min, followed by chilling on ice. Then, the RT mix (8 μ L) containing dNTP mix (0.5 mM final conc. of each), BSA (50 μ g/mL final conc.), DTT (5 mM final conc.) and SuperScript III RT (Life Technologies #18080, 10 U/ μ L final conc.) were added and reaction performed for 1 h at 50 °C.

2.3.7. Purification step

After RT the excess of RT primer was digested by 2 \times 20 U of Lambda exonuclease (Thermo Scientific #EN0561) for two incubations at 37 °C for 30 min, without intermediate denaturation steps to keep DNA/RNA duplexes intact. Digestion was followed by the treatment with 2 \times 80 U of single-strand specific Exonuclease I (Thermo Scientific #EN0582) and two incubations at 37 °C for 30 min. Enzymes were denatured at 80 °C for 15 min and residual dNTPs were de-phosphorylated with 4 U of FastAP alkaline phosphatase (Thermo Scientific #EF0651). This reaction was repeated twice with intermediate heat-denaturation. Next, RNA was degraded by addition of NaOH to a final conc. 0.15 M and heated up to 55 °C for 20 min. Then, an equal amount of acetic acid and sodium acetate to a final conc. 0.15 M was added and cDNA was ethanol precipitated with 20 μ g of glycogen as a carrier.

Combined action of Lambda exonuclease and Exonuclease I is required for efficient elimination of excess primer after RT, and de-phosphorylation of non-incorporated dNTPs is highly recommended to avoid their carry-over to the next step (riboCTP-tailing).

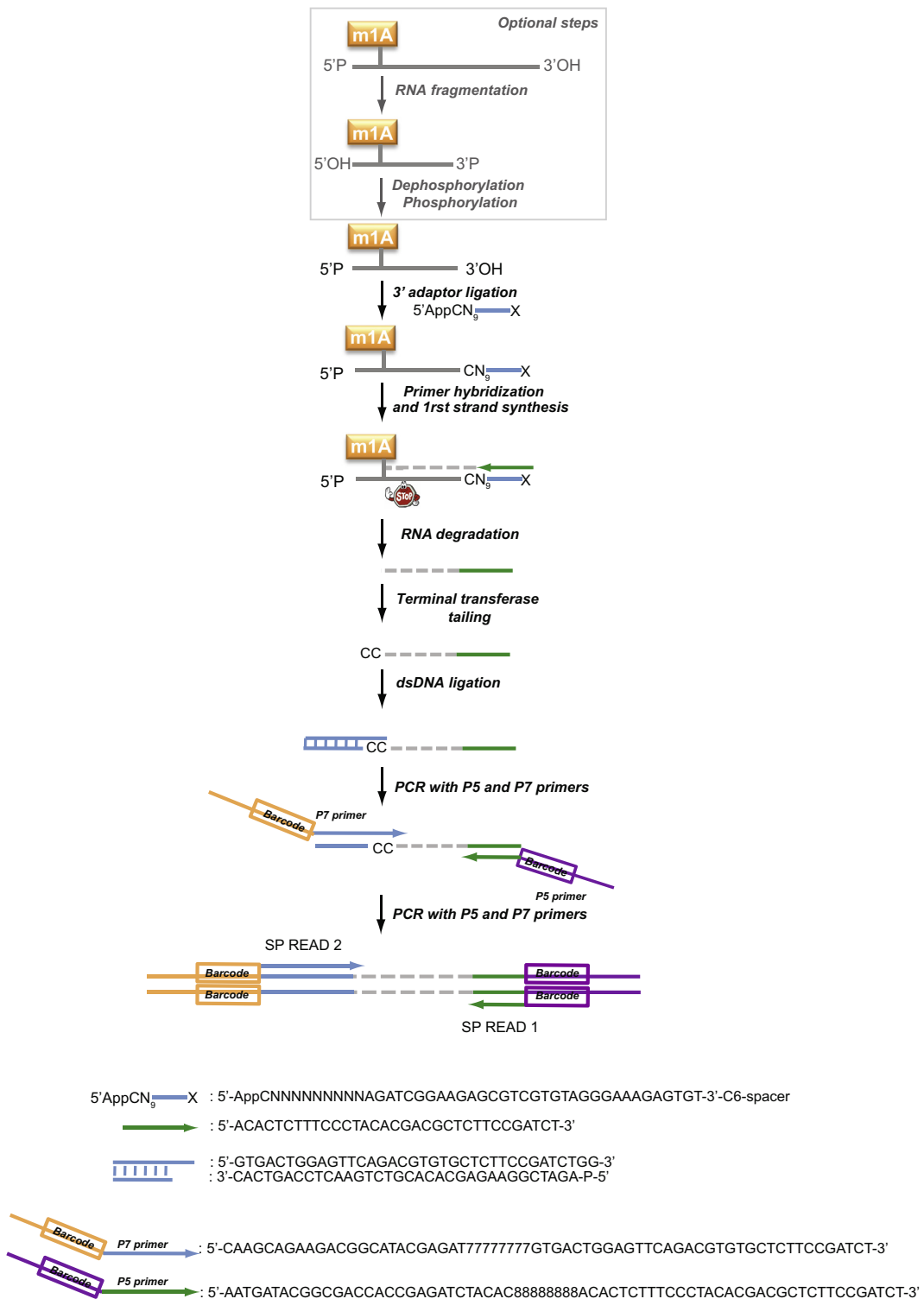


Fig. 3. General overview of the library preparation protocol. Initial fragmentation step is optional and depends on the size of analyzed RNA species. Sequences of oligonucleotides used for library preparation are given at the bottom.

5 RESULTS AND DISCUSSION

114

L. Tserovski et al. / *Methods* 107 (2016) 110–121

2.3.8. 3'-tailing and ligation

3'-cDNA tailing was done using terminal deoxynucleotidyl transferase (TdT, Thermo Scientific #10533-065) in a final mixture of 10 μ L containing TdT Buffer 1X, rCTP 1.25 mM, TdT 1 U/ μ L. RibocTP was used to limit TdT tailing for only a few incorporated nucleotides.

Double stranded DNA adapter was prepared by annealing of equimolar amounts of single-stranded oligonucleotides for 5 min at 90 °C, cooling down to 22 °C, and incubation for 10 min at 22 °C. Ligation with Double stranded DNA adapter (1.25 μ M final concentration of the duplex) was done with T4 DNA ligase (Thermo Scientific 30 U/ μ L, #EL0013, 1.5 Weiss U/ μ L final concentration) and 10 μ M ATP in 50 mM Tris-HCl, pH 7.4, 20 mM MgCl₂. Reaction was performed overnight at 4 °C. Ligated DNA was then ethanol precipitated with the addition of 20 μ g glycogen.

To purify DNA from non-ligated DNA adapters, ligation product was loaded on a 10% denaturing gel and size fraction between 40 and 150 nt was excised, eluted in 0.5 M NH₄Ac and ethanol precipitated by the addition of glycogen.

2.3.9. PCR amplification and barcoding

Final PCR amplification including 8nt barcoding with p5 and p7 primers was done in 7–12 cycles, depending on the amount of cDNA recovered in the previous steps. PCR reaction was done in 20 μ L mixture containing Taq-Polymerase buffer 1 \times , 3 mM MgCl₂, 5 μ M of each primers, 0.5 mM each of dNTP and 0.25 U/ μ L of Taq-Polymerase (Rapidozym #Gen-003-1000). After initial denaturation for 5 min at 95 °C, amplification cycles consisted in annealing for 1 min at 65 °C, elongation for 1 min at 72 °C and denaturation for 1 min at 65 °C. After 7–12 cycles of amplification, final extension was done for 5 min at 72 °C.

2.3.10. Final size fractionation

PCR products were fractionated on 10% denaturing PAGE and size fraction between about 150 and 300 nt was excised, eluted in 0.5 M NH₄Ac and ethanol precipitated (Fig. 4).

2.4. Quality control for libraries and quantification

Libraries were loaded on Agilent High Sensitivity DNA chip (for Agilent Bioanalyzer 2100) after appropriate dilution to concentration in the range of 5–500 pg/ μ L to check for the quality of the library and the eventual presence of adapter dimers. The mean size of each library was calculated using integrated software. Different representative libraries obtained with yeast tRNAs are presented on Fig. 4. Each library was carefully quantified by fluorometry using Qubit dsDNA HS assay kit (Thermo Fischer Scientific, ref #Q32851).

2.5. MiSeq or HiSeq 1000 sequencing

Before sequencing, libraries were pooled using different index reads ensuring diversity of nucleotides for signals in both red and green color channels. Indeed, Illumina HiSeq and MiSeq use a green channel to read G/T nucleotides and a red channel to read A/C nucleotides. With each sequencing cycle at least one nucleotide for each color channel must be present to ensure proper sequencing. In addition, to ensure the sequence diversity, PhiX Control v3 library (Illumina ref # FC-110-3001) was added to libraries to get 3–5% of final concentration.

Libraries were denatured with fresh 2 N NaOH and diluted to 10 pM final concentration to ensure proper clustering on the flow cell. Once clustering was finished, the samples were sequenced on MiSeq using appropriate Illumina reagents. Results of the primary analysis by FastQC (quality score distribution and adapter contamination) for read1 and read2 are given in Fig. 5.

2.6. Bioinformatics analysis pipeline (Fig. 6)

2.6.1. Demultiplexing

Demultiplexing after sequencing consists of identification of barcode sequences attributed to each sample during barcoding step and of separation of sequencing reads into individual FastQ files by sample. Demultiplexing step can be performed either using integrated MiSeq RTA software, or in post-treatment, using BclTo-FastQ utility of CASAVA 1.18.1. Demultiplexing was performed with 0 mismatch allowed in the barcode sequence. The resulting FastQ files were treated and inspected for quality, the presence of adapters and overrepresented sequences using FastQC utility (ver.0.11.2, www.bioinformatics.babraham.ac.uk/projects/).

2.6.2. Adapter trimming

Removal of the adapter sequence was done using Cutadapt v.1.8.1 software taking into account the known sequences of Illumina p5 and p7 adapters, random 10 nt sequence from 3'-adapter oligo at the RNA 3'-end and variable number of 5'-G RNA nucleotides resulting from CTP cDNA tailing.

- a CCCCAGACGGAAGAGCACACGT
- a CCCAGACGGAAGAGCACACGT
- a CCAGACGGAAGAGCACACGT
- a CAGACGGAAGAGCACACGT
- a CCAGATCG
- a AGACGGAAGAGCACACGT
- e 0.2
- After that:
- u 10 -m 10
- For Read 2:
- a AGATCGGAAGAGCGT
- e 0.2
- After that:
- u 3 -u -10 -m 8.

2.6.3. Mapping to the reference sequence

Mapping to the reference sequence (without introns, if any) was performed using bowtie2 aligner with the following custom parameters (end-to-end alignment, without soft clipping, k = 1, seed 6 nt -L 6, one mismatch allowed in the seed -N 1).

2.6.4. Signature extraction

Mapped SAM files were converted to BAM using SAMtools utility, BAM files sorted, indexed and translated into pileup format. An additional conversion step led to a custom tab-separated text file, containing coverage, arrest rate, mismatch rate and nature of all counts for each reference position. Using these data, m¹A signatures were manually compiled upon visual inspection of mapped data.

2.6.5. Supervised analysis

To quantitatively assess how robustly m¹A signatures can distinguish actual modification sites from non-m¹A sites in RNA-Seq data, a supervised prediction of m¹A by machine learning was conducted. The applied setup was a stratified cross-validation. The available m¹A (positive) signature data points from tRNA and rRNA were complemented with an equal number of non-m¹A (negative) adenosine instances randomly drawn from the same sequence pool. In a first step, the data points were shuffled and separated into five folds ensuring the stratification invariant by an alternating order of positive and negative instances. For each of the five folds, we tested a Random Forest model [29] of 500 decision trees to distinguish m¹As from non-m¹As, after training it on the remaining four folds. Repeating the shuffling step and cross-validation ten times, we obtained mean classification performances for

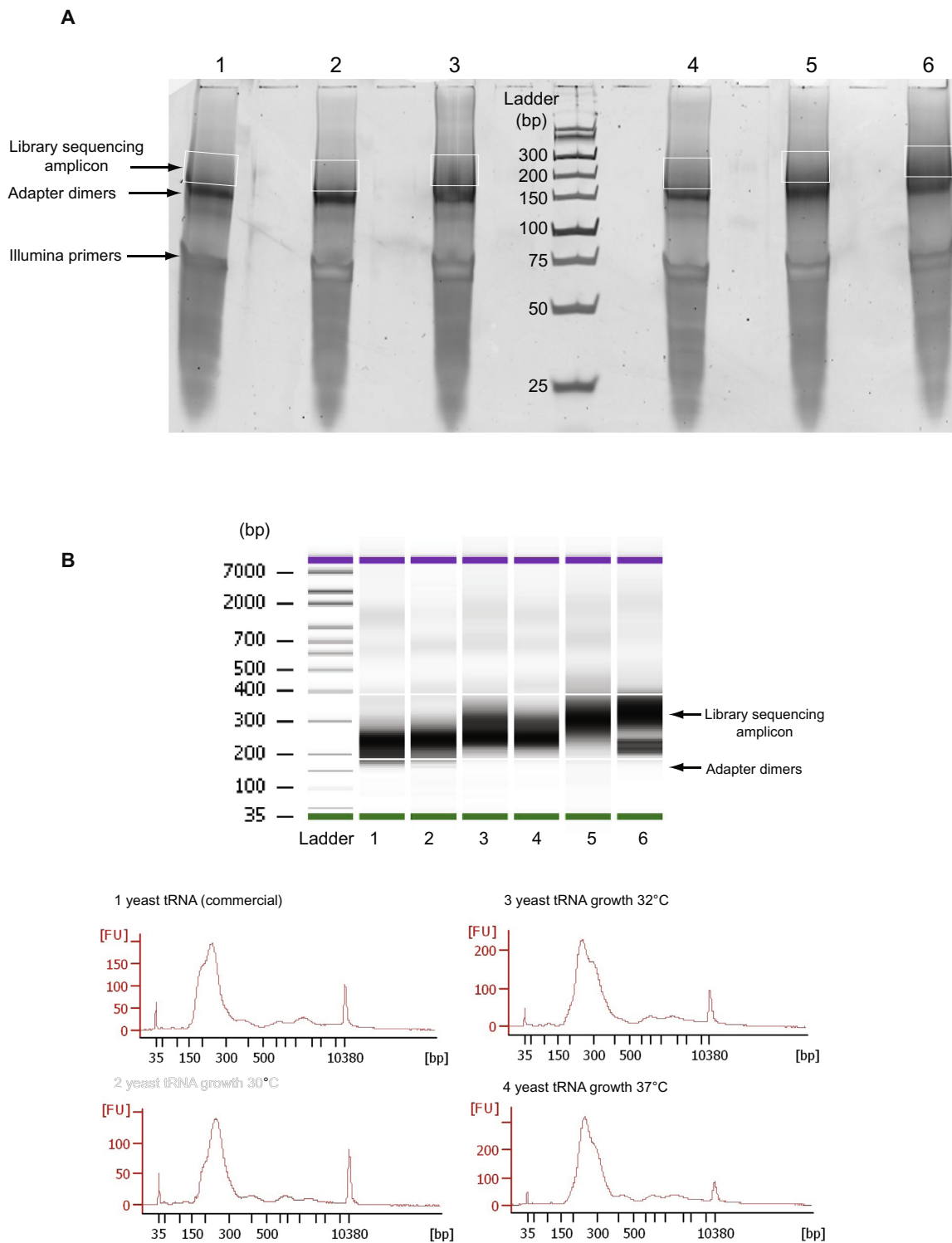


Fig. 4. Final purification and characterization of the libraries. A – Preparative 10% PAGE for separation of amplicons from adapter dimers and unincorporated primers. B – analysis of purified libraries on a High Sensitivity DNA Chip.

5 RESULTS AND DISCUSSION

116

L. Tserovski et al. / *Methods* 107 (2016) 110–121

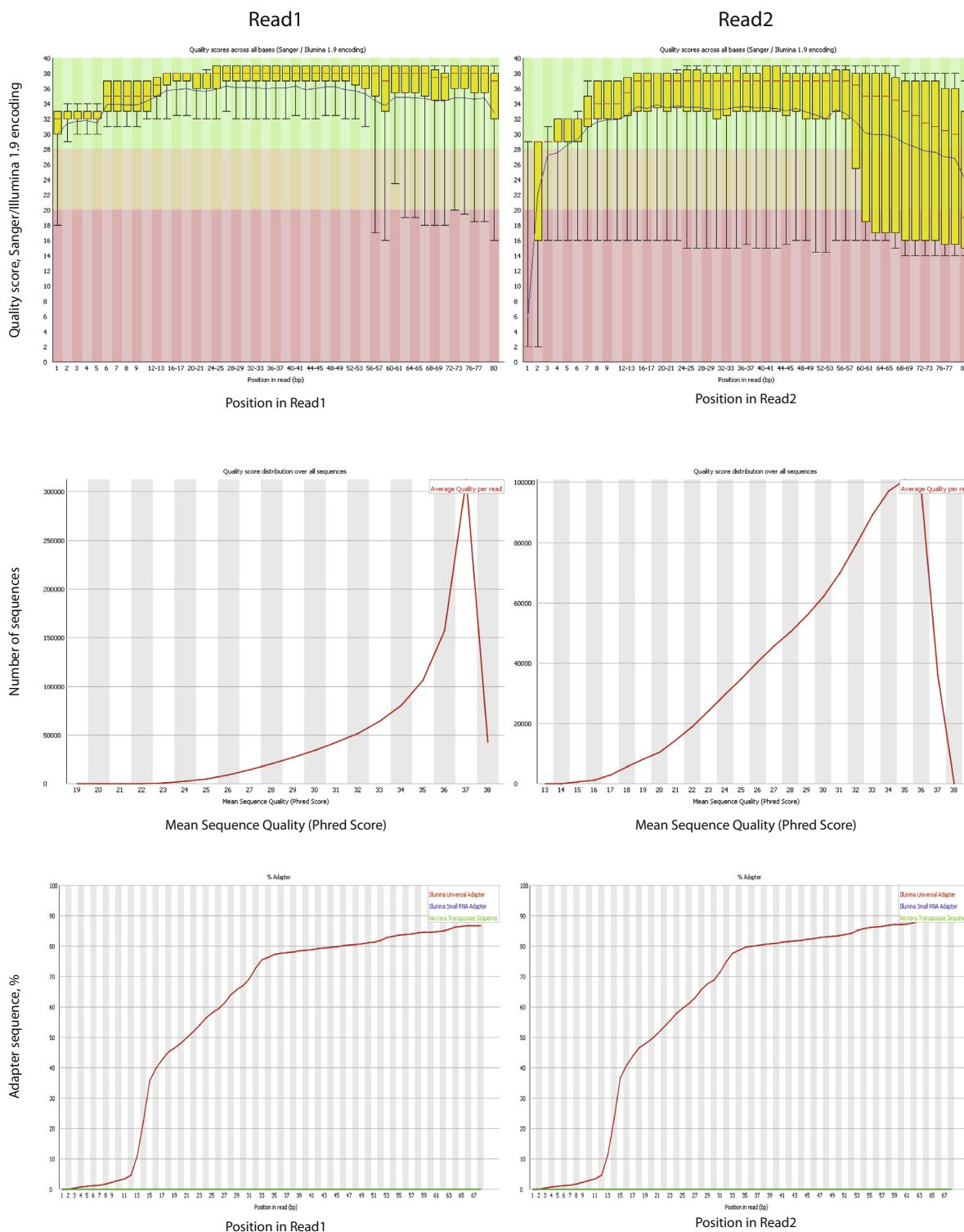


Fig. 5. Quality of sequencing reads obtained on MiSeq. Top panel – Distribution of quality score (QC) during sequencing cycles, Middle panel – distribution of sequencing reads by QC. Bottom panel – contamination by primer adapter sequence in function of the nucleotide position in the read.

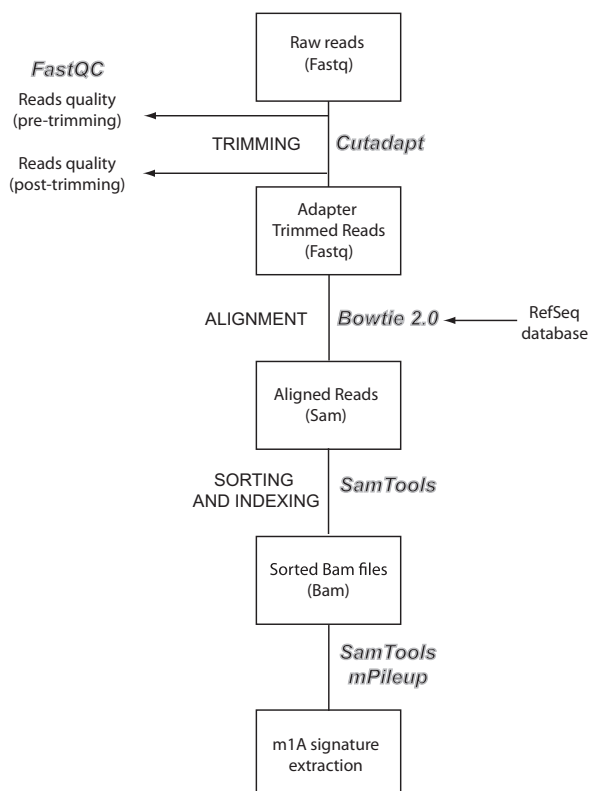


Fig. 6. Schematic illustration of the bioinformatic pipeline used for m^1A signature identification.

sensitivity, specificity, positive and negative predictive values. Finally, we reiterated the procedure using a more stringent configuration of input data, i.e., we introduced mutually exclusive thresholds for features of the non- m^1A instances to guarantee a minimum m^1A similarity.

3. Results

3.1. RT-block and nucleotide misincorporation at m^1A site

The experimental approach described here is designed for mapping of m^1A residues in RNAs using the intrinsic capacity of those modified residues to block (rather efficiently, but not completely) primer extension by RT. Such RT-block can be detected by various approaches and, in its classical low-throughput variant, was extensively used in the past [11]. Here we combined the RT-block with its mapping in RNA using high-throughput NGS technology. Most important parameters of RT-block is its position in the sequence compared to the actual modified residue, the proportion of the abortive products (represented here as arrest rate) and the percentage of nucleotide misincorporation (called mismatch rate). In order to capture all these parameters in the same sequencing experiment, we designed the specific library preparation protocol outlined in Fig. 3. In contrast to many other popular methods used for amplicon generation in NGS (e.g. Small RNA kit or TruSeq Stranded mRNA kit), the designed approach allows to determine both arrest and mismatch rates from the same sequencing library.

Using both naturally modified RNAs and synthetic m^1A -containing oligonucleotides, we demonstrated that the characteristic arrest rate and nucleotide misincorporation correspond

exactly to the position of the modified nucleotide in the sequence (Fig. 7AB).

At optimal Mg^{2+} concentration, the RT Superscript III shows about 60–80% of arrest rate at full m^1A occupancy and, in addition, variable nucleotide misincorporation, the exact percentage of both depending on the nucleotide context and probably other less defined factors.

3.2. Influence of neighboring nucleotides

Visual inspection of different m^1A signals in RNA demonstrated that the percentage of nucleotide misincorporation is far from being identical. Since the same RT was used, these variations are likely related to the RNA nucleotide context, most probably immediate neighboring positions. Previous studies [30] had already observed mismatch composition similar to those shown in the ternary plot (Fig. 7D–F), but could not explain the strong variation. In order to define this parameter more in details, we inspected the influence of positions -1 (5'-to m^1A site) and $+1$ and $+2$ (3'-to m^1A -site). For each position p_i , all data points were assigned to clusters (red for $p_i = U$, green for $p_i = A$, yellow for $p_i = G$ and blue for $p_i = C$).

As shown on Fig. 7DEF, only $+1$ position influences the misincorporation pattern, the other two play only a minor role. Uridines (T) at $+1$ show high T mismatch rates indicating efficient dATP misincorporation into the cDNA at low dCTP and dGTP respectively. Adenosines and especially guanosines at $+1$ balance the G/T mismatch ratio towards equilibrium by increased dCTP misincorporation. Finally, among the few instances of $p+1=C$ available, enhanced C mismatch rates were observed. During RT from the 3' to the 5' end of an RNA template, position $+1$ is processed right before m^1A enters the enzyme's active site. The observed correlation is plausible, since both, the $+1$ residue and its complementary cDNA counterpart are in immediate proximity to the modification such that the m^1A site is directly exposed to their physicochemical properties.

The deterministic predominance of position $+1$ was confirmed by cluster analysis of the data points in Fig. 7DEF, where equally colored data points cohered best and showed clearest separation of cluster centers. This characteristic became even more apparent by the highest Silhouette coefficient [31]. Strongest determinism of misincorporation behavior by the $+1$ base configuration was also inversely demonstrated by the best performance of a Random Forest model trained to predict the $+1$ base from the mismatch pattern (Fig. 7C).

3.3. Deduced parameters for m^1A signature

Several parameters have to be taken into account for analysis of the m^1A signature at the modified site. First of all, these both values depend on RT origin and properties, as well as experimental conditions (buffer, Mg^{2+} , temperature) used for primer extension. In our work these parameters were kept constant since the same RT enzyme (Superscript III) was used. RT behavior can also be modulated by 2D and 3D RNA structure, but these effects are only rarely taken into account. Finally, the RT properties at the modified nucleotide depend on the nucleotide context, e.g. the identity of neighboring nucleotides, especially at $+1$ position.

The dataset of 45 m^1A signatures (coverage ≥ 10 , 3'-adjacent coverage ≥ 15) of tRNA, rRNA and artificial oligonucleotides was fed to a Random Forest (RF) model (500 trees) classifying both kinds of adenosine instances (unmodified A and m^1A). Briefly, an RF [32] is a machine-learning model for object classification by an ensemble of decision trees. Under randomization in training, binary forks are formed in each individual tree, used to differentiate objects according to their features, based on information

5 RESULTS AND DISCUSSION

118

L. Tserovski et al./Methods 107 (2016) 110–121

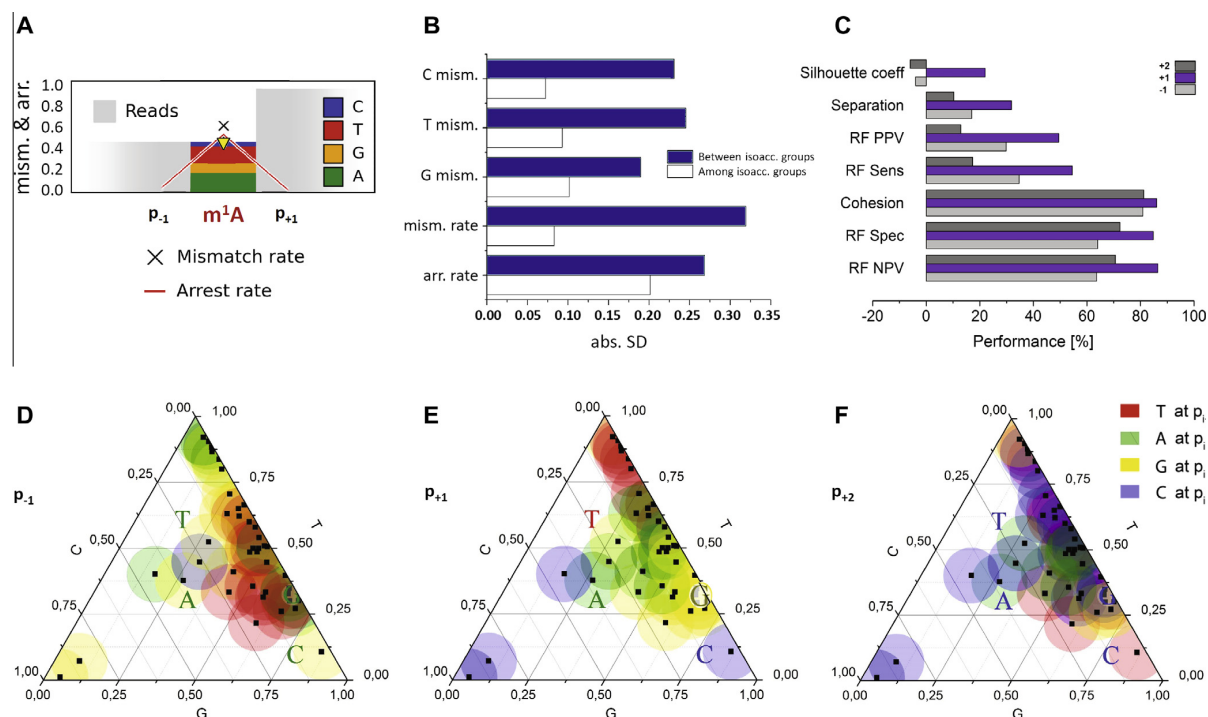


Fig. 7. m^1A signature and its dependence on the neighboring nucleotides. A – Average profile at m^1A position and two neighboring positions -1 and $+1$. B – Standard deviations of average results from A. RNA isotypes were defined taking into account the similarity between tRNA sequences and their cognate amino acid. C – Dependence of mismatch composition on neighboring nucleotide at position $+1$, represented by clustering analysis (cohesion, separation, silhouette coefficient) and Random Forest performance (sensitivity, specificity, positive predictive value – PPV, negative predictive value – NPV) in prediction of the $+1$ nucleotide from m^1A mismatch values G, T and C. D, E and F – Ternary plots of mismatch compositions at m^1A sites colored by nucleotide configuration at positions -1 (D), $+1$ (E) and $+2$ (F). Position $+1$ shows the best color clustering in agreement with the best Random Forest performance in C). Letters A, G, T and C in the ternary plots correspond to mismatch compositions observed in four synthetic oligonucleotides containing m^1A with variegated base configuration (A, G, U and C) at position $+1$.

content. The final object classification is a consensus of all class votes returned by the single trees. Features visible to the RF-classifier included: arrest rate a , mismatch rate m , the m/a ratio, the mismatch composition (fractions of G, T and C), and a parameter that we termed CSA. The latter is defined as the fold change of the site p_i 's a with respect to the median a of its sequence environment of five bases up- and five bases downstream, according to the following formula: $CSA_{p_i} = \frac{a_{p_i}}{\text{median}(a_{p_{i-5}}, \dots, a_{p_{i-1}}, a_{p_{i+1}}, \dots, a_{p_{i+5}})}$

Finally, we have analyzed in depth, which parameters have been retained by the RF model as most informative for the recognition of m^1A sites. From inspection of the trained RF, it became clear, that both arrest and mismatch rates played an important role for performance. A more detailed inspection was conducted by a leave-feature-out analysis, which measures performance in various permutations of incomplete feature combinations. The results, which are shown in Fig. 8, clearly confirm our initial approach to m^1A signature identification, namely that neither arrest rate nor mismatch analysis alone come close to the performance of their combination.

3.4. Candidate assessment in new sequencing data

Given sequencing data from RNAs containing potential m^1A sites, the significance of candidate positions can be assessed by comparison of their profile features with the typical m^1A characteristics. For positions of sufficient coverage, minimum similarity to m^1A signatures should be guaranteed by application of lower thresholds for mismatch and arrest rates. These should be chosen

arbitrarily based on feature distribution at known m^1A sites (regenerate according to example in Appendix A). Calculation of p-values is recommended as decision help based on probability to observe features in certain expression strength at non- m^1A positions. The False Discovery Rate (FDR) [33] should be controlled with respect to sequence length to minimize alpha-error accumulation due to multiple hypothesis testing. Alternatively, significance should be at least down-rated by the number of analyzed sequence positions (in style of Bonferoni correction [34]). In addition, a Random Forest model can be trained to distinguish m^1A from non- m^1A sites on a known RNA landscape. However, the output of this RNA-Seq approach is an indication, not a confirmation. Predictive reliability depends on the nature of the training pool and transferability to a target sequence landscape. Therefore, these auxiliary decision helps cannot replace a proof of m^1A sites by LC-MS/MS.

4. Discussion and perspective

4.1. Quantification of partial m^1A methylation

Experiments with synthetic m^1A -modified and unmodified RNA oligonucleotides [26] demonstrated that the RT-signature is dependent on the proportion of modified oligonucleotide in the mixture, and can be used for relative evaluation of methylation rate. A more detailed inspection of the data shows that the best results in this relative quantification are obtained for the arrest rate (correlation 0.94). Thus, this method can be applied for discrimination between full modification rate and partially modified

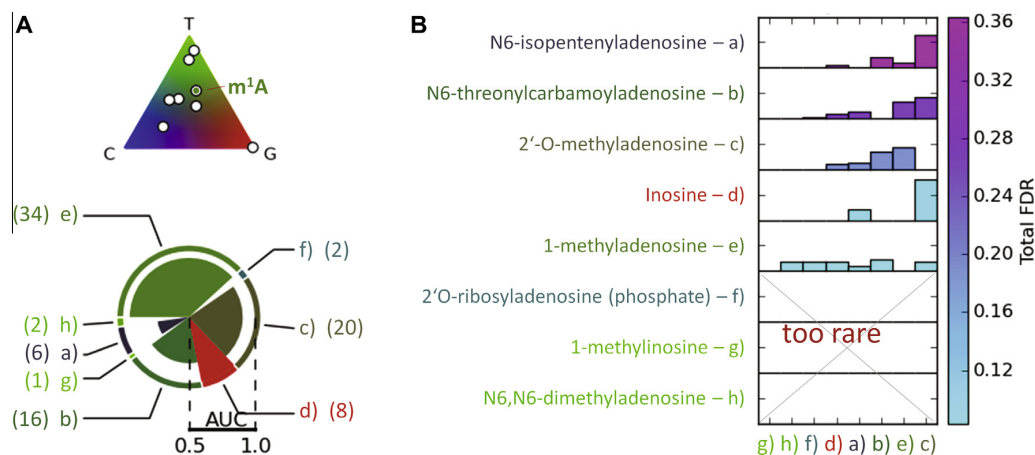


Fig. 8. Discrimination of m^1A residues from other modified adenosines. A – Pie chart showing relative frequencies of analyzed adenosine modification entries from MODOMICS in human mitochondrial tRNAs, yeast tRNAs and rRNAs. Color corresponds to mismatch composition displayed in ternary plot. Pie radii code for area under curve (AUC) from receiver operating characteristic (ROC) curve of a Random Forest model tested for discrimination performance of the modification types. B – Total specific False Discovery Rates (FDR) of modification types (vertical axis and color bar) and relative contributions by other modification types (horizontal axis, bar heights are normalized by relative modification frequencies). Results A and B were determined in 10 repetitions of a 5-fold stratified cross-validation using equal amounts of a specific modification (minimum required frequency = 5) vs. a random composition of “other”-labeled modifications. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

situations. Nevertheless, more precise approaches, like HPLC-MS (MS) should be applied for absolute quantification of the modification rate.

4.2. Reverse transcription stops at other modified nucleotides

RNA contains numerous other modified nucleotides which may affect the efficiency of primer extension by RT. If the additional group is located at the Hoogsteen edge of the base, these modifications are either neutral or still may provoke the RT-pausing, leading to the coverage drop as well as nucleotide misincorporation to some degree. But, as expected, the most prominent RT signals are obtained for nucleotides with modified or hypermodified Watson-Crick edge. In our datasets we also observed such signals corresponding to m^3C , m^3U , m^1G and m^2G . Such signals are readily detectable, but do not correspond to modified A in RNA.

4.3. Discrimination of m^1A and other modified As

At the current state, our machine-learning algorithm can distinguish m^1A from an unmodified adenosine with very good accuracy, if these two possibilities are the only elements in the training data.

However, in addition to m^1A , naturally modified RNAs frequently contain other As modified at the Watson-crick edge, the most important are m^2A , i^6A and t^6A (Fig. 8). Since these modified As may also generate visible RT signals, we analyzed the parameters of the arrest rate and misincorporation in such cases. Our results show that m^1A can easily be distinguished from i^6A , t^6A , Am and Inosine, however the signals of m^2A are too rare for an exhaustive statistical analysis. The signals produced by two m^2A in yeast 18S rRNA show that some m^2A residues may be erroneously classified as m^1A .

4.4. Application of m^1A -mapping to different RNA species

In its current state our method for mapping of m^1A methylation can be applied to short and also long RNA species as demonstrated for tRNAs and rRNAs. With some adaptations of the bioinformatics pipeline it is also suitable for whole transcriptome screening.

However, in the case of long mRNAs one can also use random priming for RT-step, which may be beneficial for a more regular coverage (see also next paragraph).

4.5. Detection of m^1A signatures in other library preparation protocols

In this work we used the original library preparation protocol allowing the simultaneous detection of both arrest rate and misincorporation at a given site. In comparison, the most commonly used methods for RNA conversion to the library (for example adapter ligation to both 3'- and 5' of RNA used in Small RNA kits) are only suitable for detection of misincorporation, all abortive cDNA products are lost during PCR step. In other cases, both arrest rate and misincorporations can be potentially detected (for example for templated 3'-cDNA extension, ScriptSeq v2 protocol, Epicentre-Illumina), while others, like popular Illumina TruSeq stranded mRNA protocol, probably loses such information due to random priming in both directions. In addition, the arrest rate and misincorporation certainly depends on the RT enzyme used in the protocol and other reaction conditions like dNTP and Mg^{2+} concentrations. Hundreds thousands of various RNA libraries have already been sequenced and information is publicly available. Next challenge would be to do comparative analysis of m^1A -signatures generated by different RT-enzymes under standard conditions used in commercial protocols and use those signatures for genome-wide mapping of potential m^1A sites in various transcriptomes obtained for different species under normal conditions as well as for various human pathologies. This will certainly bring new insights into the presence and the role of these modified nucleotides in gene expression regulation.

Acknowledgements

This work was supported by a Joint ANR-DFG project [HTRNA-Mod to Y.M and M.H.] (designations N°ANR-13-ISV8-0001 01 and HE 3397/8-1); THE DFG SPP1784 [HE 3397/13-1, DFG HE3397/14-1] to M.H.

5 RESULTS AND DISCUSSION

120

L. Tserovski et al. / *Methods* 107 (2016) 110–121

Appendix A

Arrest rate	Mism rate	<i>m/a</i>	CSA	G mism	T mism	C mism
0.77	0.46	0.60	7.37	0.66	0.32	0.02
0.65	0.43	0.66	8.08	0.78	0.10	0.12
0.84	0.21	0.25	9.48	0.15	0.61	0.24
0.83	0.15	0.18	8.16	0.32	0.31	0.37
0.01	0.42	51.82	1.52	0.27	0.38	0.35
0.04	0.80	20.36	9.15	0.44	0.54	0.02
0.24	0.38	8.18	12.31	0.32	0.65	0.03
0.96	0.09	0.09	30.99	0.44	0.49	0.08
0.15	0.01	0.09	90.07	0.29	0.63	0.08
0.22	0.56	2.73	2.96	0.51	0.45	0.04
0.01	0.84	68.28	3.36	0.69	0.30	0.01
0.01	0.08	5.90	10.35	0.60	0.40	0.01
0.07	0.74	32.47	9.59	0.48	0.51	0.01
0.02	0.78	177.65	0.64	0.11	0.88	0.00
0.03	0.92	319.06	1.78	0.09	0.91	0.00
0.02	0.91	94.18	5.50	0.16	0.84	0.00
0.06	0.15	15.08	18.09	0.41	0.58	0.01
0.01	0.72	101.16	7.42	0.33	0.66	0.01
0.62	0.51	2.42	47.54	0.12	0.87	0.01
0.41	0.38	0.74	88.45	0.26	0.71	0.03
0.20	0.76	3.93	38.67	0.19	0.80	0.01
0.01	0.03	6.05	0.80	0.86	0.11	0.03
0.16	0.60	23.81	1.73	0.08	0.07	0.84
0.02	0.89	59.61	3.98	0.05	0.01	0.94
0.45	0.10	0.23	5.57	0.17	0.40	0.43
0.75	0.05	0.07	7.92	0.44	0.33	0.22
0.52	0.59	1.14	9.66	0.64	0.33	0.03
0.84	0.09	0.11	11.86	0.46	0.50	0.04
0.88	0.23	0.26	13.98	0.59	0.22	0.19
0.93	0.37	0.40	6.00	0.66	0.30	0.04
0.92	0.25	0.27	6.71	0.56	0.33	0.10
0.42	0.84	2.01	1.77	0.07	0.92	0.01
0.43	0.38	0.89	6.43	0.47	0.51	0.02
0.70	0.20	0.28	4.49	0.57	0.31	0.12
0.25	0.04	0.18	6.23	0.33	0.62	0.04
0.39	0.04	0.12	9.87	0.44	0.50	0.06
0.88	0.13	0.15	7.37	0.65	0.26	0.09
0.56	0.61	1.08	6.52	0.38	0.60	0.02
0.76	0.57	0.75	8.00	0.29	0.45	0.26
0.44	0.39	0.94	6.94	0.63	0.35	0.02
0.61	0.27	0.52	9.61	0.69	0.27	0.04
0.68	0.36	0.62	13.22	0.42	0.41	0.17
0.49	0.46	0.65	3.54	0.46	0.49	0.05
0.16	0.06	0.92	3.48	0.51	0.36	0.13
0.90	0.30	0.34	11.67	0.28	0.52	0.19

References

- [1] M.A. Machnicka, K. Milanowska, O. Osman Oglou, E. Purta, M. Kurkowska, A. Olchowik, et al., MODOMICS: a database of RNA modification pathways–2013 update, *Nucleic Acids Res.* 41 (2013) D262–D267, <http://dx.doi.org/10.1093/nar/gks1007>.
- [2] N. Husain, S. Obranic, L. Kosciński, J. Seetharaman, F. Babic, J.M. Bujnicki, et al., Structural basis for the methylation of A1408 in 16S rRNA by a panaminoglycoside resistance methyltransferase NpmA from a clinical isolate and analysis of the NpmA interactions with the 30S ribosomal subunit, *Nucleic Acids Res.* 39 (2011) 1903–1918, <http://dx.doi.org/10.1093/nar/gkq1033>.
- [3] J. Wachino, K. Shibayama, H. Kurokawa, K. Kimura, K. Yamane, S. Suzuki, et al., Novel plasmid-mediated 16S rRNA m1A1408 methyltransferase, NpmA, found in a clinically isolated *Escherichia coli* strain resistant to structurally diverse aminoglycosides, *Antimicrob. Agents Chemother.* 51 (2007) 4401–4409, <http://dx.doi.org/10.1128/AAC.00926-07>.
- [4] A. Guelorget, M. Roovers, V. Guérineau, C. Barbey, X. Li, B. Golinelli-Pimpaneau, Insights into the hyperthermostability and unusual region-specificity of archaeal *Pyrococcus abyssi* tRNA m1A57/58 methyltransferase, *Nucleic Acids Res.* 38 (2010) 6206–6218, <http://dx.doi.org/10.1093/nar/gkq381>.
- [5] M. Roovers, J. Wouters, J.M. Bujnicki, C. Tricot, V. Stalon, H. Grosjean, et al., A primordial RNA modification enzyme: the case of tRNA (m1A) methyltransferase, *Nucleic Acids Res.* 32 (2004) 465–476, <http://dx.doi.org/10.1093/nar/gkh191>.
- [6] M. Helm, Post-transcriptional nucleotide modification and alternative folding of RNA, *Nucleic Acids Res.* 34 (2006) 721–733, <http://dx.doi.org/10.1093/nar/gkj471>.
- [7] C. Peifer, S. Sharma, P. Watzinger, S. Lamberth, P. Kötter, K.-D. Entian, Yeast Rrp8p, a novel methyltransferase responsible for m1A 645 base modification of 25S rRNA, *Nucleic Acids Res.* 41 (2013) 1151–1163, <http://dx.doi.org/10.1093/nar/gks1102>.
- [8] S. Sharma, P. Watzinger, P. Kötter, K.-D. Entian, Identification of a novel methyltransferase, Bmt2, responsible for the N1-methyl-adenosine base modification of 25S rRNA in *Saccharomyces cerevisiae*, *Nucleic Acids Res.* 41 (2013) 5428–5443, <http://dx.doi.org/10.1093/nar/gkt195>.
- [9] L. Lempereur, M. Nicoloso, N. Riehl, C. Ehresmann, B. Ehresmann, J.P. Bachellerie, Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible, *Nucleic Acids Res.* 13 (1985) 8339–8357.
- [10] P. Tjerina, S. Mohr, R. Russell, DMS footprinting of structured RNAs and RNA-protein complexes, *Nat. Protoc.* 2 (2007) 2608–2623, <http://dx.doi.org/10.1038/nprot.2007.380>.
- [11] Y. Motorin, S. Muller, I. Behm-Ansmant, C. Branlant, Identification of modified residues in RNAs by reverse transcription-based methods, *Methods Enzymol.* 425 (2007) 21–53, [http://dx.doi.org/10.1016/S0076-6879\(07\)25002-5](http://dx.doi.org/10.1016/S0076-6879(07)25002-5).
- [12] S.L. Hiley, J. Jackman, T. Babak, M. Trocheset, Q.D. Morris, E. Phizicky, et al., Detection and discovery of RNA modifications using microarrays, *Nucleic Acids Res.* 33 (2005) e2, <http://dx.doi.org/10.1093/nar/gni002>.
- [13] K. Chen, Z. Lu, X. Wang, Y. Fu, G.-Z. Luo, N. Liu, et al., High-resolution N(6)-methyladenosine (m(6)A) map using photo-crosslinking-assisted m(6)A sequencing, *Angew. Chem. Int. Ed. Engl.* 54 (2015) 1587–1590, <http://dx.doi.org/10.1002/anie.201410647>.
- [14] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, et al., Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq, *Nature* 485 (2012) 201–206, <http://dx.doi.org/10.1038/nature11112>.
- [15] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, G. Rechavi, Transcriptome-wide mapping of N(6)-methyladenosine by m(6)A-seq based on immunocapturing and massively parallel sequencing, *Nat. Protoc.* 8 (2013) 176–189, <http://dx.doi.org/10.1038/nprot.2012.148>.
- [16] S. Edelheit, S. Schwartz, M.R. Mumbach, O. Wurtzel, R. Sorek, Transcriptome-wide mapping of 5-methylcytosine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs, *PLoS Genet.* 9 (2013) e1003602, <http://dx.doi.org/10.1371/journal.pgen.1003602>.
- [17] S. Hussain, J. Aleksic, S. Blanco, S. Dietmann, M. Frye, Characterizing 5-methylcytosine in the mammalian epitranscriptome, *Genome Biol.* 14 (2013) 215, <http://dx.doi.org/10.1186/gb4143>.
- [18] B. Linder, A.V. Grozhik, A.O. Olarerin-George, C. Meydan, C.E. Mason, S.R. Jaffrey, Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome, *Nat. Methods* 12 (2015) 767–772, <http://dx.doi.org/10.1038/nmeth.3453>.
- [19] J.E. Squires, H.R. Patel, M. Nusch, T. Sibbritt, D.T. Humphreys, B.J. Parker, et al., Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA, *Nucleic Acids Res.* 40 (2012) 5023–5033, <http://dx.doi.org/10.1093/nar/gks144>.
- [20] T.M. Carille, M.F. Rojas-Duran, B. Zinshteyn, H. Shin, K.M. Bartoli, W.V. Gilbert, Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells, *Nature* 515 (2014) 143–146, <http://dx.doi.org/10.1038/nature13802>.
- [21] A.F. Lovejoy, D.P. Riordan, P.O. Brown, Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in *S. cerevisiae*, *PLoS ONE* 9 (2014) e110799, <http://dx.doi.org/10.1371/journal.pone.0110799>.
- [22] S. Schwartz, D.A. Bernstein, M.R. Mumbach, M. Jovanovic, R.H. Herbst, B.X. León-Ricardo, et al., Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA, *Cell* 159 (2014) 148–162, <http://dx.doi.org/10.1016/j.cell.2014.08.028>.
- [23] U. Birkedal, M. Christensen-Dalsgaard, N. Krogh, R. Sabarinathan, J. Gorodkin, H. Nielsen, Profiling of ribose methylations in RNA by high-throughput sequencing, *Angew. Chem. Int. Ed. Engl.* 54 (2015) 451–455, <http://dx.doi.org/10.1002/anie.201408362>.
- [24] D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M.S. Ben-Haim, et al., The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA, *Nature* (2016), <http://dx.doi.org/10.1038/nature16998> (advance online publication).
- [25] X. Li, X. Xiong, K. Wang, L. Wang, X. Shu, S. Ma, et al., Transcriptome-wide mapping reveals reversible and dynamic N1-methyladenosine methylome, *Nat. Chem. Biol.* (2016), <http://dx.doi.org/10.1038/nchembio.2040> (advance online publication).

- [26] R. Hauenschild, L. Tserovski, K. Schmid, K. Thüring, M.-L. Winz, S. Sharma, et al., The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent, *Nucleic Acids Res.* 43 (2015) 9950–9964, <http://dx.doi.org/10.1093/nar/gkv895>.
- [27] M.A. Collart, S. Oliviero, Preparation of yeast RNA, in: Frederick M. Ausubel Al. (Ed.), *Curr. Protoc. Mol. Biol.*, 2001, <http://dx.doi.org/10.1002/0471142727.mb1312s23> (Chapter 13, Unit13.12).
- [28] H. Cahová, M.-L. Winz, K. Höfer, G. Nübel, A. Jäschke, NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs, *Nature* 519 (2015) 374–377, <http://dx.doi.org/10.1038/nature14020>.
- [29] A. Liaw, M. Wiener, *Classification and regression by randomForest*, *R. News.* 2 (2002) 18–22.
- [30] P. Rytvkin, Y.Y. Leung, I.M. Silverman, M. Childress, O. Valladares, I. Dragomir, et al., HAMR: high-throughput annotation of modified ribonucleotides, *RNA N. Y. N.* 19 (2013) 1684–1692, <http://dx.doi.org/10.1261/rna.036806.112>.
- [31] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, *J. Comput. Appl. Math.* 20 (1987) 53–65, [http://dx.doi.org/10.1016/0377-0427\(87\)90125-7](http://dx.doi.org/10.1016/0377-0427(87)90125-7).
- [32] L. Breiman, *Random Forests*, *Mach. Learn.* 45 (2001) 5–32.
- [33] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. Ser. B Methodol.* 57 (1995) 289–300.
- [34] C.E. Bonferoni, *Teoria statistica delle classi e calcolo delle probabilità*, *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8 (1936) 3–62.
- [35] X. Li, P. Zhu, S. Ma, J. Song, J. Bai, F. Sun, et al., Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome, *Nat. Chem. Biol.* 11 (2015) 592–597, <http://dx.doi.org/10.1038/nchembio.1836>.

5.1.2 The reverse transcription signature of *N*-1-methyladenosine in RNA-Seq is sequence dependent

After establishing a library preparation protocol, the detection of *N*-1-methyladenosine as a model modification was chosen because of several reasons including:

- The existence of a methyl group at *N*-1 of adenosine is known to impact the natural base-pairing properties of thus modified nucleoside [96, 97]. Therefore, significant reverse transcription stop as well as possible misincorporations at that position were anticipated.
- This modification is highly abundant and conserved within different RNA species, e.g. tRNA [22], rRNA [31] and, as recently discovered, in mRNA [47, 48].
- The importance of this modification lies in its functions such as correct folding [14, 26], or antibiotic resistance [150], but also in structure probing experiments as the product of a DMS-induced methylation of adenosine [151].

Together, these properties provided the basis for testing whether the platform was suitable for the detection of this modification in RNA on a single nucleoside resolution basis.

To evaluate the behavior of a reverse transcriptase toward m^1A , various biological RNAs, featuring this naturally occurring modification, as well as synthetic oligoribonucleotides, were selected. Corresponding knockouts and non-modified synthetic RNAs served as negative controls. Initial analysis involved the evaluation of cytosolic tRNA from *Saccharomyces Cerevisisae*. A typical pattern including strong arrest rate and incorporation of incorrect nucleotides was observed at conserved position 58, a well described m^1A position in eucaryotic tRNA [3]. This pattern completely disappeared in a knockout strain as was shown for tRNA^{Ile}. Similar results were obtained when the 25S ribosomal yeast RNA, possessing two m^1As at positions 645 and 2141, was compared within different strains including wild type, single- and double knockouts.

More importantly, the RT-signatures for m^1A that were captured, although showing some similar characteristics, such as arrest rate, differed considerably, especially in their misincorporation profile. To address this, detailed analysis, including short synthetic oligoribonucleotides was performed. Those RNAs were based on human mitochondrial tRNA^{Lys}, already described to contain m^1A at position 9 [26], as well as varying nucleotides at the +1 position (3') adjacent to it. This way, a correlation could be shown between the nucleotide direct 3' next to m^1A and the corresponding misincorporation profile. For example, 5'- m^1A -U-3' showed an increased misincorporation of dATP into cDNA during reverse transcription. This is visible by a high T proportion, because, the nucleotide information is mapped to the complementary template sequence. Additionally, it was shown, that the neighboring nucleotide had little effect on the overall arrest rate.

It was of special interest, whether the information produced by this sequencing protocol could be used for search of new or homologous sites containing m^1A . Sequencing a sample of total RNA from *Trypanosoma brucei*, that was yet not shown to contain m^1A sites, resulted in 16 tRNAs that visibly featured an m^1A signature. To confirm the actual existence of m^1A in at least one of these tRNA species, a specific biotinylated DNA oligonucleotide, complementary to tRNA^{Arg(UCG)} was designed and by means of hybridization the corresponding tRNA was isolated. The purified tRNA was then submitted to both LC-MS and renewed sequencing, delivering the result that indeed m^1A was present in that tRNA. Analogously, but omitting an LC-MS confirmation, an m^1A signature was present in the mouse rRNA at a position that is homologous to that of human rRNA, where this modification was already found [152].

Structure probing experiments, designed to test the accessibility of, among others, adenosine toward the methylating agent DMS, rely solely on the arrest rate provoked by the resulting

methyl-group at the $N-1$ [151]. To test whether a correlation between modification occupancy and arrest rate exists, an experiment was designed that included the mixing of two synthetic oligonucleotides, one containing m^1A , the other not. A clear dependence of the content of m^1A , determined by LC-MS, and arrest rate, as well as misincorporation degree was observed. The same analysis was then performed on a biological sample. Purified rRNA from yeast, containing two m^1A s with similar occupancy, proven also by LC-MS, yielded different RT profiles. Arrest rate and misincorporation were both visible, yet they no longer correlated with the results produced by the synthetic RNAs. This effect is important, and means that caution must be taken when evaluating results of a structure probing experiment.

Finally, the robustness of the m^1A signature for discrimination between actual modification and non- m^1A sites was tested by supervised prediction assisted by machine learning. Known instances of m^1A were mixed with an equal amount of non- m^1A sites from the adenosine pool and were fed to a random forest (RF) model constructed of 500 trees. The features, visible to the RF classifier were arrest rate, mismatch rate, their ratio, mismatch composition and CSA. The latter is a parameter that describes the fold change of site's arrest rate with respect to its sequence environment of 5 bases up- and downstream. Five-fold stratified cross validation with ten randomizations was performed. Under conditions that tested between m^1A and non-modified adenosine, the model scored better than 97 % for both sensitivity and specificity. Under more strict conditions that tested between m^1A and other modified adenosines, that yielded similar RT signatures, the model scored with about 89 % - sensitivity and about 87 % - specificity. Nevertheless, the chosen RF model was shown superior than a more basic model – the k-Nearest Neighbor. Additionally, it was shown that both the arrest rate and the mismatch rate are most informative for the RF model when compared to either one feature. Statistical tests were performed and evaluated by Ralf Hauenschild.

The results of this research were recently published [153].

The reverse transcription signature of *N*-1-methyladenosine in RNA-Seq is sequence dependent

Ralf Hauenschild^{1,†}, Lyudmil Tserovski^{1,†}, Katharina Schmid¹, Kathrin Thüring¹, Marie-Luise Winz², Sunny Sharma³, Karl-Dieter Entian³, Ludivine Wacheul⁴, Denis L. J. Lafontaine⁴, James Anderson⁵, Juan Alfonzo⁶, Andreas Hildebrandt⁷, Andres Jäschke², Yuri Motorin^{8,*} and Mark Helm^{1,*}

¹Institute of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Staudingerweg 5, 55128 Mainz, Germany, ²Institute of Pharmacy and Molecular Biotechnology (IPMB), Heidelberg University, Im Neuenheimer Feld 364, 69120 Heidelberg, Germany, ³Institute of Molecular Biosciences: Goethe University Frankfurt, Max-von-Laue Street 9, 60438 Frankfurt/M, Germany, ⁴RNA Molecular Biology, Université Libre de Bruxelles, Rue Profs Jeener & Brachet, 12, B-6041 Charleroi-Gosselies, Belgium, ⁵Department of Biological Sciences, Marquette University, 53201-1881, Milwaukee, WI, USA, ⁶Department of Microbiology, The Ohio State University, 43210, Columbus, OH, USA, ⁷Institute for Computer Sciences, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany and ⁸IMoPA UMR7365 CNRS-UL, BioPole de l'Université de Lorraine, 9 avenue de la Forêt de Haye, 54505 Vandoeuvre-les-Nancy, France

Received April 10, 2015; Revised August 26, 2015; Accepted August 27, 2015

ABSTRACT

The combination of Reverse Transcription (RT) and high-throughput sequencing has emerged as a powerful combination to detect modified nucleotides in RNA via analysis of either abortive RT-products or of the incorporation of mismatched dNTPs into cDNA. Here we simultaneously analyze both parameters in detail with respect to the occurrence of *N*-1-methyladenosine (*m*¹A) in the template RNA. This naturally occurring modification is associated with structural effects, but it is also known as a mediator of antibiotic resistance in ribosomal RNA. In structural probing experiments with dimethylsulfate, *m*¹A is routinely detected by RT-arrest. A specifically developed RNA-Seq protocol was tailored to the simultaneous analysis of RT-arrest and misincorporation patterns. By application to a variety of native and synthetic RNA preparations, we found a characteristic signature of *m*¹A, which, in addition to an arrest rate, features misincorporation as a significant component. Detailed analysis suggests that the signature depends on RNA structure and on the nature of the nucleotide 3' of *m*¹A in the template RNA, meaning it

is sequence dependent. The RT-signature of *m*¹A was used for inspection and confirmation of suspected modification sites and resulted in the identification of hitherto unknown *m*¹A residues in trypanosomal tRNA.

INTRODUCTION

1-methyladenosine (*m*¹A) is an RNA modification originating essentially from two different reaction types, one catalyzed by enzymes and the other the result of the reaction of RNA with certain alkylating agents. Correspondingly, the relevance of this modification in RNA-related research is essentially two-fold. On one hand, dimethylsulfate (DMS) is a popular chemical probe of RNA structure in solution; reactivity toward DMS is interpreted as accessibility of the corresponding nitrogen or nucleobase to solvent, and hence a lack of structural involvement. The *N*1 of adenosine is not the only RNA nucleophile to react with DMS (1), as e.g. the *N*3 of cytidines and the *N*7 of guanosines are also probed by this reagent. For the latter two, the resulting chemically modified nucleosides *m*³C and *m*⁷G can be revealed by further chemical treatments leading to chain scission at the modified sites. Since such a treatment has not been developed for *m*¹A, it has traditionally been detected by primer elongation arrest (2,3). The underlying logic is that chemi-

*To whom correspondence should be addressed. Tel: +49 6131 39 25731; Fax: +49 6131 39 20373; Email: mhelm@uni-mainz.de
Correspondence may also be addressed to Yuri Motorin. Tel: +33 3 83 68 55 10; Fax: +33 3 83 68 54 09; Email: motorine5@univ-lorraine.fr
†These authors contributed equally to the paper as first authors.

5.1 Detection of m¹A by NGS

Nucleic Acids Research, 2015, Vol. 43, No. 20 9951

cal alterations blocking the Watson–Crick face should act as an RT-roadblock, as such impair the incorporation of the complementary nucleotide into the cDNA by the reverse transcriptase enzyme, and cause the latter to stall. A typical result of structural probing experiments is thus an arrest signal at the position of the last nucleotide upstream or the 5'-adjacent nucleotide of the cDNA (corresponding to the 3'-nucleotide of m¹A on the RNA template). While the traditional method for detecting such RT-arrest signals involves the resolution of labeled primer extension products by polyacrylamide gelelectrophoresis (PAGE) or capillary electrophoresis, recent developments in structural probing make use of the power of deep sequencing methods (4). Of note, in the entire field, the tacit assumption for decades has been that m¹A is quantitative in its RT-arrest capacity, i.e. structural probing experiments were interpreted as if every encounter of an m¹A by an RT enzyme led to abortion of primer elongation. On the other hand, m¹A is also a prominent and frequently occurring member of the growing family of 150 or so chemically different naturally occurring RNA modifications. It is typically found at position 58 of many eukaryotic and archaeal tRNAs (5), as well as in eukaryotic (6) and bacterial (7–9) rRNA. Further occurrences are known from position 9 of metazoan mitochondrial tRNAs (9–11), and as a mediator of antibiotic resistance in rRNA of *Streptomyces pactum* (12).

Interestingly, several recent papers, including applications of different RNA-Seq protocols have created data containing mismatched nucleoside signals at sites known or postulated to contain m¹A (13,14). This strongly suggests that reverse transcriptase is capable of reading through this altered Watson–Crick face, thereby incorporating non-matching nucleotides in the process and leaving unobtrusive traces of the m¹A modification in cDNA data. Such behavior is known from DNA polymerase bypassing sites of DNA lesion (15). In a comprehensive investigation of misincorporation caused by various RNA modifications, Ryvkin *et al.* have recently reported a common misincorporation pattern for adenosine modifications (16). However, the protocols for library preparation in the reported RNA-Seq approach were unsuited to detect the abortive cDNA products described above in the structural probing context. Failure to detect RT-arrest signals may originate from details of the library preparation protocols, for example when both primer binding sites are introduced via ligation on the RNA level. For the detection of RT-arrest signals e.g. in tRNA (17), the second primer binding site must be introduced at the level of cDNA.

Here, we use a library preparation protocol suitable for the detection of both, abortive cDNA and misincorporation (Figure 1). Application to RNA preparations containing known or suspected m¹A sites revealed an RT-signature left by m¹A residues which includes characteristic misincorporation patterns as well as typical RT-arrest rates. Most interestingly, we find a dependence on the type of the m¹A-preceding nucleotide in the RNA template (i.e. to the 3' of m¹A), whose nature correlates with misincorporation patterns. These findings have important bearings for both areas: in structural probing, proper interpretation of RT-arrest assays of DMS treated RNA should include the no-

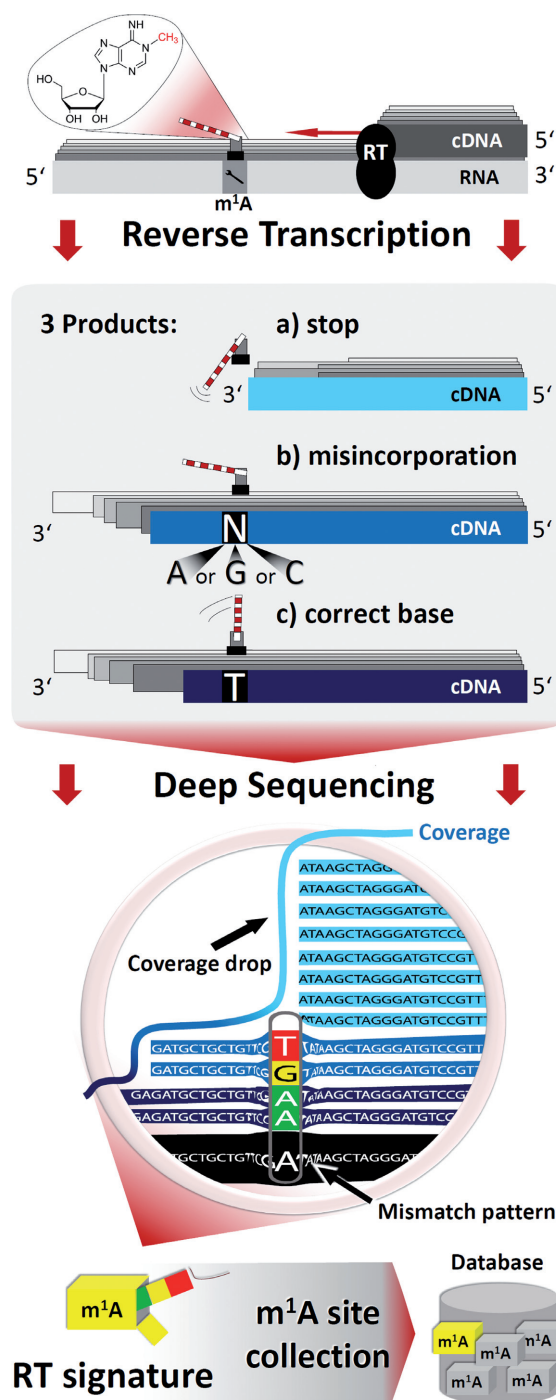


Figure 1. Principle of generation and analysis of RNA-Seq data for the detection of m¹A residues.

5 RESULTS AND DISCUSSION

9952 *Nucleic Acids Research*, 2015, Vol. 43, No. 20

tion of incomplete detection as well as a sequence context around the adenosine residue under investigation.

MATERIALS AND METHODS

Unless otherwise specified, all synthetic nucleic acids were from IBA (Göttingen, Germany). Details including sequence information are given in Supplementary Table S1. Yeast rRNA was prepared as described in (6), yeast tRNA as described in (18). *S. pactum* DSM40530 (DSMZ, Braunschweig, Germany) was cultivated as recommended for liquid media growth (19) with slight modifications. Total bacterial RNA was extracted with TRIzol[®] Reagent (Life Technologies, Thermo Scientific, Dreieich, Germany) according to the manufacturer's protocol.

Library preparation protocol

This protocol was slightly varied from a previously published version (18, 20) as follows: the RNA adapter of 5 random (N) nucleotides + 1 constant cytidine (C) at the 3' end was replaced by a 9 N + 1 C Illumina p5 template sequence. The custom sequence of the cDNA adapter was substituted with an Illumina p7 template sequence. Instead of 6 nt at each 5' end of primers as barcode, full length Illumina compatible (derived from Nextera 2 platform) primers with dual barcodes N501-N508 and N701-N7012 were used. True-Seq DNA amplicon library preparation for introducing Illumina compatible sequences before sequencing was not required.

Fragmentation. Total or ribosomal RNA was fragmented in a volume of 10 μ l containing 10 mM ZnCl₂, and 100 mM Tris-HCl, pH 7.4, at 90°C for 5 min. The reaction was stopped by addition of ethylenediaminetetraacetic acid (EDTA) to a final concentration of 50 mM. Next, the RNA was size separated by 10% denaturing PAGE and bands of size 50–150 nt were excised, eluted in 0.3 M ammonium acetate and ethanol precipitated.

Dephosphorylation. RNA (about 0.5 μ g) was then dephosphorylated on both extremities in dephosphorylation mixture (10 μ l total) consisting of 100 mM Tris-HCl, pH 7.4, 20 mM MgCl₂, 0.1 mg/ml BSA, 100 mM 2-mercaptoethanol, and 0.5 U FastAP (Thermo Scientific) at 37°C for 30 min. Prior to addition of enzyme, RNA was denatured at 90°C for 30 s, then chilled on ice (in the following, this will be referred to as 'heat denaturation'). After 30 min reaction time, RNA was again heat denatured and an equal amount of enzyme was added to perform a second cycle of dephosphorylation.

3'-adapter ligation. Next, an adapter was ligated to the 3'-end of dephosphorylated RNA. To this end, dephosphorylated RNA was complemented to yield a ligation mixture (final volume 20 μ l) and to perform ligation as described in (18). Here, RAdapter (IBA, Goettingen, Germany; see Supplementary Table S1 for Sequence) was used in a concentration of 5 μ M. After the reaction the enzymes were inactivated at 75°C for 15 min.

Removal of excess adapters. Before the reverse transcription step, the excess of RAdapter was removed. To this end, the mixture was heat denatured. Then 20 U of 5'-Deadenylase (New England Biolabs, Frankfurt, Germany) were added to the ligation mix, followed by incubation at 30°C for 30 min. After a second heat-denaturation an equal amount of enzyme was added and the reaction was repeated. Next, the single stranded RAdapter (now completely monophosphorylated) was digested by adding 10 U of Lambda exonuclease (Thermo Scientific, Dreieich, Germany) to the reaction mixture and incubating at 37°C for 30 min. After heat denaturation an equal amount of enzyme was added to repeat the reaction. Finally, both enzymes were heat-inactivated at 80°C for 15 min, after which RNA was ethanol precipitated with the addition of 1 μ l glycogen (Thermo Scientific, Dreieich, Germany) per sample.

Reverse transcription. Composition of the reverse transcription mixture (here, 40 μ l) was previously described (18). Here, the pellet was first redissolved in 32 μ l RT-Primer (IBA, Goettingen, Germany; see Supplementary Table S1 for sequence) in a final concentration of 5 μ M in FS Buffer (Life Technologies) and heat denatured at 80°C for 10 min, then chilled on ice. Then, dNTP mix BSA, dithiothreitol and SuperScript III reverse transcriptase (10 U/ μ l, Life Technologies) were added. Reactions were performed at 50°C for 1 h and no heat inactivation was performed.

Removal of excess primers and dNTPs. Next, the RT-Primer was digested. To this end, 20 U of Lambda exonuclease were added to the reverse transcription mix, and incubated at 37°C for 30 min. The reaction was repeated by addition of an equal amount of enzyme, without prior heat denaturation, to avoid denaturation of RNA:DNA hybrids. Following this, 80 U of single-strand specific Exonuclease I (Thermo Scientific) were added and incubated at 37°C for 30 min. Again, the reaction was repeated by addition of an equal amount of enzyme, without prior heat-denaturation. Finally, all enzymes were heat-inactivated at 80°C for 15 min. After that, dNTPs were dephosphorylated. For this, 4 U of FastAP thermosensitive alkaline phosphatase were added to the mixture and incubated at 37°C for 30 min. The reaction was repeated upon heat-denaturation. Finally, RNA was hydrolyzed as described (18). The reaction was stopped by neutralizing with an equal amount of acetic acid and precipitating with ethanol.

3'-tailing and ligation of cDNA. The obtained cDNA was reacted with TdT (Thermo Scientific, Dreieich, Germany) as published in a volume of 10 μ l. Oligocytidine overhangs were generated using cytidine triphosphate (CTP) under optimized conditions affording >90% addition of three cytidines (18). For the ligation of the second adapter the TdT mixture was complemented to yield a final ligation mixture consisting of 50 mM Tris-HCl, pH 7.4, 20 mM MgCl₂, 1.25 μ M DAnchor (DAnchorA annealed to DAnchorB, IBA, Goettingen, Germany; see Supplementary Table S1 for sequence), 10 μ M ATP and 1.5 Weiss U/ μ l T4 DNA ligase (Thermo Scientific) in a total volume of 40 μ l. Reaction was performed and ligation products purified as described (18).

5.1 Detection of m¹A by NGS

Nucleic Acids Research, 2015, Vol. 43, No. 20 9953

PCR amplification and barcoding. Each sample was finally polymerase chain reaction (PCR) amplified using the respective barcoded P7 and P5 primers (IBA, Goettingen, Germany, see Supplementary Table S1). PCR products were size-separated by 10% denaturing PAGE and regions of interest (above 150, which is the size of adapter dimers and below 300, which is the maximum size of PCR amplicons) were excised, DNA was eluted in 0.3 M ammonium acetate and ethanol precipitated. The resuspended DNA was then sequenced on the MiSeq platform (see Supplementary Table S2 for details).

Deep-Seq data processing and mapping

The sequence libraries specified in Supplementary Table S2 (end-types, lengths, platform) were processed in a custom bioinformatic pipeline. Corresponding to the library preparation settings, a Python based workflow accommodated demultiplexing, removal of primers, adapters, barcodes and ligation-assistance overhangs. Mapping was performed using Bowtie2 with solely tRNA or rRNA references obtained from MODOMICS (9) for the corresponding organism. Alignment mode was set to global (end-to-end, no soft-clipping) with one mismatch tolerated in the seed of 6 nt. Splicing was not part of the mapping strategy. Mapping to all references simultaneously, only one ($k = 1$) alignment declared as valid by Bowtie2 was reported for each read.

Signature extraction

Mapping was followed by format conversion using SAM-tools. From SAM files, sorted and indexed BAM files were generated, which were translated to Pileup format. An additional conversion lead to a custom tab-separated text file format, termed *Profile* (details in Supplementary Table S3), providing all parameters of relevance for inspection of modification candidates. Herein, for each reference position the listed properties include coverage c , arrest rate a , mismatch content m as well as the counts for each base type. All presented RT signatures were compiled manually during visual inspection of the mapping results. Database entries of m¹A sites listed in MODOMICS were retrieved and confirmed by evaluation of arrest rate characteristics and mismatch patterns. The extracted signatures were complemented by those of m¹As from homologous identification performed via ClustalW2 sequence alignments of related organisms. Identification was performed by isolated visual inspection. By manual selection, positional shifts of m¹A₅₈ to e.g. positions 57 or 59 due to variable loops were correctly recognized and from all sites listed in Modomics, those could be determined that obtained a signature projected by our approach.

Supervised prediction

The uniqueness of m¹A's RT signature was evaluated by supervised prediction, i.e. machine learning mediated detection of known m¹A instances within a pool of non-methylated adenosine sites with m¹A-resembling or differing sequencing profiles. The general workflow is shown in Supplementary Figure S6. Mean prediction performances

(sensitivity, specificity) were calculated from 10 repetitions of a five-fold stratified cross-validation, training and testing a Random Forest (RF) *R* package implementation (21). The training sets contained equal amounts of instances of both classes. Attributes used for classification input were its arrest rate a , relative mismatch content m , relative mismatch composition values (G, T and C content), m/a and the fold change of a w.r.t. the mean a within the site's -5 and $+5$ bp neighborhood, termed context sensitive arrest rate (CSA). The input format of training material is detailed in Supplementary Table S4. In the first input setting, (i), all 45 m¹A signatures from tRNA (already averaged for isotypes), rRNA and synthetic oligoribonucleotides were merged with 45 random non-m¹A instances. The isotype averaging ensures that for any distribution of the data into training and testing sets, the classifier is facing unseen data in a test run. From (i), setting (ii) was derived, which allowed only non-m¹A of a minimum m¹A signature resemblance w.r.t. at least one of the thresholds $a \geq 0.2$, $m \geq 0.2$ or at least two mismatch type with ≥ 0.1 share of an $m \geq 0.1$. Setting (iii) corresponded to (i) except that the training set was generated from tRNA instances (*Saccharomyces cerevisiae* cytosolic and *Homo sapiens* mitochondrial) only, while the test set contained rRNA sites (*S. cerevisiae*, *S. pactum*) exclusively. To demonstrate the advantage of our prediction model, we compared the supervised prediction power of the RF with that of a basic k-nearest neighbor (kNN) classifier (Supplementary Figure S7).

LC-MS/MS analysis

HPLC-DAD-MS/MS analysis.

Isolation of single tRNA species from Trypanosoma brucei. Single tRNA species were isolated from *Trypanosoma brucei* total RNA (22) by hybridization with complementary, biotinylated DNA-oligonucleotides followed by immobilization on streptavidin-coated magnetic beads (Dynabeads® MyOne™ Streptavidin T1, Life Technologies, Darmstadt, Germany). Target tRNA was tRNA^{Arg(U_{CG})} (oligonucleotide 4309, sequence: biotin-CGGCAGGACTCGAACCTGCAACCCTCA). The hybridization step was performed in 5× SSC buffer (20×: 3 M NaCl, 300 mM trisodium citrate, pH 7.0) using 100 pmol biotinylated oligonucleotide and 150 μg total RNA per 25 μl beads. Samples were denatured at 90°C for 3 min and subsequently hybridized at 65°C for 10 min and cooled to room temperature. Dynabeads® were washed three times using Binding and Washing buffer (5 mM Tris-HCl (pH 7.5), 0.5 mM EDTA, 1 M NaCl) according to the manual and then equilibrated once in 5× SSC buffer before adding the hybridized samples. 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 1× SSC buffer and three times in 0.1× SSC buffer. Finally, the beads were resuspended in MilliQ water and heated to 75°C for 3 min to elute the target tRNA. To exclude the presence of remaining DNA-oligonucleotide and non-target RNAs, the eluted RNA was further purified

5 RESULTS AND DISCUSSION

9954 *Nucleic Acids Research*, 2015, Vol. 43, No. 20

by 10% denaturing polyacrylamide gel electrophoresis and ethanol precipitation.

Sample preparation. Prior to LC-MS/MS analysis, RNA samples were digested into nucleosides according to the following protocol: samples were incubated in presence of 1/10 volume of 10× nuclease P1 buffer (0.2 M ammonium acetate pH 5.0, ZnCl₂ 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 10× fast alkaline phosphatase buffer (Fermentas, St Leon-Roth, Germany) and 1 U fast alkaline phosphatase (Fermentas, St Leon-Roth, Germany) were added, and samples were incubated for additional 60 min at 37°C. After digestion, 1/10 volume of ¹³C-labeled total RNA (*S. cerevisiae*, 10 ng/μl), prepared as described in (23), was added as internal standard for m¹A quantification.

HPLC parameters. The calibration solutions and digested RNA samples 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 parameters. Prior to entering the mass spectrometer, the effluent from the column was measured photometrically at 254 nm by the DAD. The triple quadrupole 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. For the detection and quantification of m¹A, time-segmented multiple reaction monitoring (MRM mode) was applied in order to ensure the separation of m¹A from other methylated adenosine derivatives. The elution of m¹A took place in the time segment from 5 to 8.5 min, while e.g. m⁶A could be detected in the last segment starting at 14 min, thus the segmentation allowed the exclusive detection of m¹A. Mass transitions and QQQ parameters used can be found in Table 1. Peak areas were determined employing Agilent MassHunter Qualitative Analysis Software. In the case of adenosine, peak areas were extracted from the recorded UV chromatograms in order to avoid saturation of the mass signals.

m¹A and A quantification. In order to quantify the m¹A content of the RNA samples, ¹³C-labeled total RNA from *S. cerevisiae* was used as a stable isotope-labeled internal

standard (SIL-IS) as described for total RNA from *Escherichia coli* previously (24). Briefly, 10 calibration solutions containing 0.01–500 fmol/μl m¹A (Sigma Aldrich, Munich, Germany) and 10 ng/μl SIL-IS were prepared and analyzed by LC-MS/MS (injection volume 10 μl/sample). For determination of a nucleoside–isotope response factor for m¹A, the ratio of the extracted areas of the ¹²C-m¹A and ¹³C-m¹A peaks was calculated for each calibration solution. The resulting response factor was then used for m¹A quantification in the RNA samples.

Quantification of A was performed by running an external calibration series (5–1000 pmol) and extracting the peak areas from the recorded UV chromatogram. For inter-sample comparability, the detected m¹A amount was normalized to the A content for each sample (% m¹A/A). For the synthetic RNA samples with defined sequence, the quantification of A enables the calculation of the analyzed amount of RNA as well as the percentage of RNA molecules carrying an m¹A modification. Results are displayed in Supplementary Table S7.

RESULTS

Our approach to capture a comprehensive profile of reverse transcription encounters with m¹A containing templates is depicted in Figure 1. Like in conventional RNA-Seq procedures, RNA preparations were reverse transcribed into cDNA libraries and submitted to Illumina sequencing. However, in contrast to typical library preparations, which include numerous biochemical steps prone to result in biased amplification of certain RNA species, we applied a specifically optimized protocol (18,20). This was designed to minimize such biases, as well as to capture abortive reverse transcription products originating in particular from encounters of the enzyme with nucleotide modifications in the template. The choice of RNA preparations as starting material was guided by the necessity to assess or eliminate, by proper control samples, other factors known to influence the RT-signature. Thus, we compared known m¹A sites in tRNA and rRNA with that of null mutants to assess the influence of strongly structured RNA domains on the RT-profile. Short synthetic m¹A containing oligonucleotides were included to assess the influence of the nucleotide directly 3'-to the m¹A site, as this is the last one to be conventionally reverse transcribed before the direct encounter of the RT-enzyme active site with the m¹A modification. For all known m¹A sites, the resulting reads were inspected for their arrest rate at the m¹A site, and in reads bypassing the modification site, the ratio of all four nucleotides was determined and extensively analyzed.

Library preparation

An overview over the library preparation is given in Supplementary Figure S1. It was slightly adapted from a previously published protocol (18). The first step included in an optional fragmentation, applied to preparations containing RNAs significantly longer than tRNAs, such as e.g. rRNA. It consisted in incubation with ZnCl₂, followed by excision from preparative PAGE of a size range denoted by the 50 and 150 nt bands of a size standard. Treatment with alkaline phosphatase was performed to remove phosphates

and cyclic 2'-3'-phosphates that might block the 3'-end for the subsequent ligation with a DNA adapter. The adapter contained the following sequence elements (details in 'Materials and Methods' section): (i) pre-adenylated 5'-cytidine (18) (ii) nine randomized nucleotides ('N') for indexing of individual molecules (iii) P5 Illumina sequencing template (iv) 3'-blocking non-nucleosidic building block. In a third step, adapter ligation was followed by specific hybridization and elongation of a primer complementary to the non-random part of the adapter, resulting in a cDNA library. This library contained the pertinent information, namely the length and sequence of RT-events resulting from read-through events at the modification site, as well as of abortive products. To quantitatively convert this information into ds DNA libraries ready for Illumina sequencing, the RNA was degraded by alkaline hydrolysis, and the cDNA was submitted to CTP tailing by terminal transferase. The oligocytidine overhang was used as an anchor to hybridize a secondary adapter of double-stranded DNA, containing one helper strand in addition to the principle primer (18). The latter contained the following elements: (v) 5'-phosphate for ligation and (vi) the P7 Illumina sequencing element (sequences in Supplementary Table S1). The complementary helper strand contained an additional two guanines as an overhang on its 3'-end to improve ligation efficiency by hybridization to the oligocytidine tail of the cDNA. This library was amplified in two PCR steps, the first one using only the P5 and P7 sequencing primers. After gel purification and excision, the second PCR was conducted using the full length P5 and P7 primers containing indices *i5* and *i7* for dual barcoding of multiple samples in a single sequencing run, as well as flow cell anchoring sequences. This latter step allows direct sequencing on the Illumina platform, circumventing an additional step, normally contained in the TruSeq kit protocol. A total of 20 libraries were prepared for this paper, annotated with various relevant characteristics as listed in Supplementary Table S2.

Characterization of RT-signature at known m^1A sites

To find particular m^1A -related signatures in thus prepared RNA-seq libraries, the reads were mapped onto a minimal target genome consisting only of rRNA and tRNA sequences. The respective sequences have been obtained from the Modomics database (9) and do thus not contain any unspliced or unprocessed sequences, but, importantly, sequences of known modifications status. This is in some contrast to the previously published HAMR method (16), which relied on the generation of tRNA families from the ensemble of genetic copies of tRNA genes. The precise mapping parameters are of some concern, because among the various tRNAs sequences present in e.g. yeast, there are many strong similarities, especially among isoacceptors. Isoacceptors are tRNA species related by the amino acid they decode and are charged with on their 3'-end (25). As a result of such similarity, a significant fraction of reads may be assigned to targets that they do not biochemically originate from also outside the isoacceptor context. To evaluate the degree to which such mismatching might influence a potential m^1A -signature, we used a parameter called the Levenshtein distance, essentially the number of mutation steps

necessary to interconvert both species (26). As this is a measure of the relative similarity of two given RNA sequences, it inversely correlates with the probability for mismatching between the two species. The comparison of yeast tRNA sequences based on Levenshtein distance (details shown in Supplementary Figure S2) impressively shows, that most concerns for mismatching must be directed toward isoacceptors, while the sequence similarities outside these groups are minor in comparison. Therefore, for reads with multiple potential mapping sites, a regime termed 'k1' was applied, which reports one valid mapping site only. This and the treatment of other details on the mapping strategy must be relegated to the discussion part of this manuscript, because most of the relevant aspects are yet to be developed below. Thus, for example we later on comment on the clear advantages of the k1 regime when compared to the k3 regime (results displayed in Supplementary Figure S3), which reports up to three valid mapping sites.

To initially circumvent the above problems, a first assessment of RT-signatures was conducted with yeast 25S rRNA, which has known m^1A sites at positions 645 and 2142 of the large ribosomal subunit (entry 5 in Supplementary Table S2). Pure 25S rRNA, isolated from whole ribosomes as described in (6) showed a distinct occurrence of both, abortive RT-products and misincorporation of non-adenosine signals at positions suggestive of a causal connection to the presence of m^1A . Importantly, both aspects were absent in negative controls obtained from either single or double knockout strains (6) of the methyltransferases responsible for the respective methylation (Figure 2A). Similarly, comparable RT-signatures were detected at position 58 of various yeast tRNAs, of which one example is shown in Figure 2B, whereas the remainder is detailed in Supplementary Figure S4 and an average signature is compiled in Figure 2C, which also lists the corresponding deviations. These signatures were absent in tRNA preparations from a knockout strain of the respective tRNA m^1A methyltransferase (Figure 2B) (27,28). This clearly demonstrates that m^1A residues leave a distinct signature even in RNA species whose stable structures are known to affect RT-arrest rates.

These RT-signatures displayed common characteristics in the mismatch incorporation of nucleotides into the cDNA at the positions corresponding to m^1A in the RNA template. However, significant variation is evident, which also applies to the RT-arrest rate between m^1A and the position to its 3', as indicated by a red line in Figure 2. Clearly, a significantly larger number of instances must be investigated for a comprehensive picture. Therefore, we analyzed the RT-signatures of known m^1A residues in further RNA preparations with known m^1A sites, including yeast tRNA, human mitochondrial tRNA, human rRNA and rRNA from *S. pactum* (samples listed in Supplementary Table S2). The latter is of particular interest, because its m^1A residue, which mediates an antibiotic resistance, is the only one situated in small subunit rRNA.

m^1A 's RT signature is dependent on the sequence context of the RNA template

For all instances from Table 2, the m^1A -dependent mismatch composition was analyzed as a function of the

5 RESULTS AND DISCUSSION

9956 Nucleic Acids Research, 2015, Vol. 43, No. 20

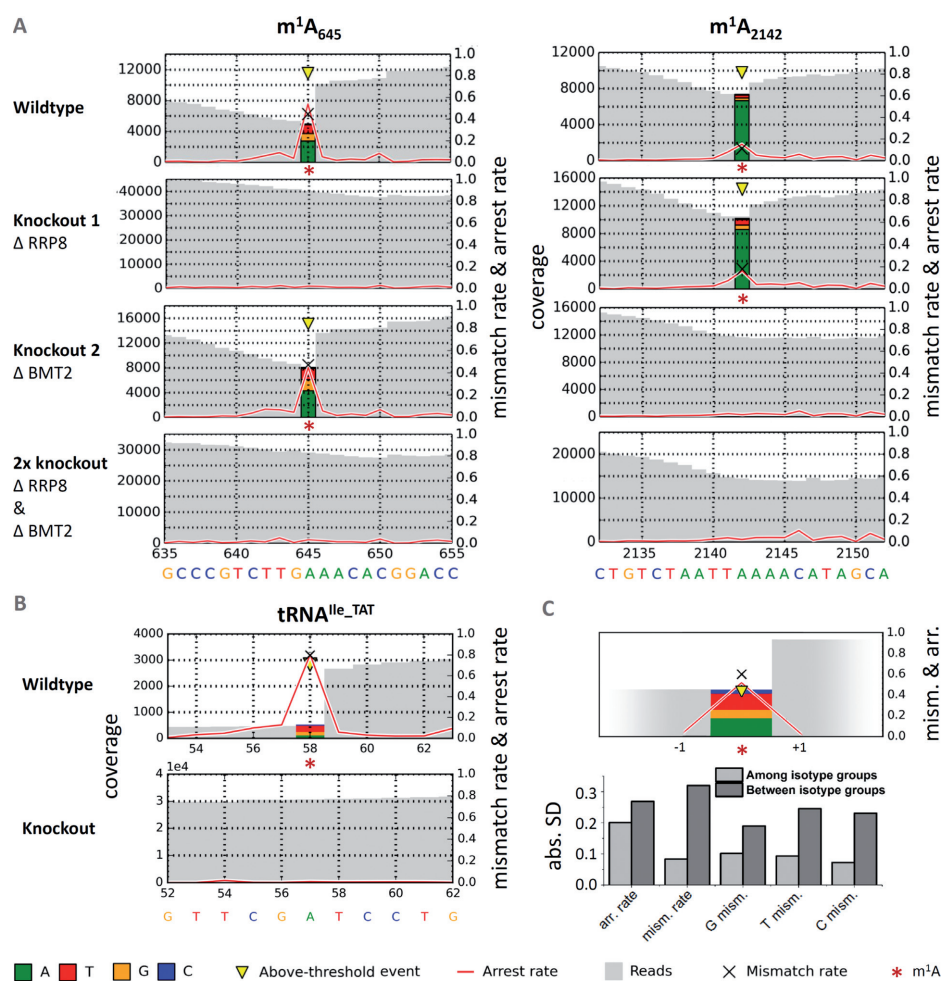


Figure 2. Detection of m¹A signatures in deep sequencing data. The representations illustrate the coverage of a given site in gray, the arrest rate is plotted as a red line, and the mismatch composition is visualized by colored stacks at the m¹A sites. For a position p, the arrest rate reflects the relative amount of mapped reads ending at p + 1, i.e. not covering p. (A) Sequencing profiles from single and double methyltransferase knockouts of *Saccharomyces cerevisiae*'s LSU rRNA with m¹A sites 645 and 2142. Signatures of m¹A residues are clearly apparent in the wild-type, and disappear in the corresponding knockout constructs. (B) Sequencing profiles of tRNA^{Ile}-TAT from wild-type and Trm6-knockout strains. The signature clearly disappears in RNA from a knockout strain of the enzyme, which is responsible for synthesis of m¹A₅₈ in tRNAs (28). tRNA^{Ile}-TAT was chosen as an example out of 37 signatures, which are detailed in Supplementary Figure S4. Positions are labeled according to absolute length of reference sequences, including variable regions. (C) Average signature of said 37 yeast cytosolic tRNAs at m¹A₅₈ complemented with absolute standard deviations of signature features among and between groups of isotypes. For the displayed profile, signatures were averaged among isotypes first, before calculating the final means.

neighboring sequence context, including one upstream nucleotide and two downstream nucleotides of the RNA template, denoted -1, +1 and +2, respectively. This implies that nucleotides +1 and +2 are reverse transcribed, before the m¹A residues acts as a template in the RT active site, with +1 denoting the characteristic position after which RT-arrest occurs. Consequentially, the -1 position only enters the RT active site after the enzyme has bypassed the m¹A residue. In a first instance, the influence of each position was analyzed independently from the others. A distinct influence of a given nucleotide would result in a clustering of signals in a ternary plot (16) of the mismatch composition. While no significant impact of nucleotide identity on positions -1

and +2 was observable in such plots (Supplementary Figure S5 A and C), the ternary plot of position +1, visualized in Figure 3A and B, stands out. For example, the 5'-m¹A-U-3' motif leads to very efficient misincorporation of dATP into cDNA.

Since nucleotide information is mapped to the template sequence, this corresponds to high T signal, as well as to low G and low C signals. This characteristic misincorporation pattern is visually recognizable by clustering of 5'-m¹A-U-3' derived red data points in the upper end of the ternary diagram in Figure 3A. Similarly, m¹A-G (yellow) and m¹A-A (green) give rise to distinct clusters with overall low C signal, while data points for m¹A-C are more spread

5.1 Detection of m^1A by NGS

Nucleic Acids Research, 2015, Vol. 43, No. 20 9957

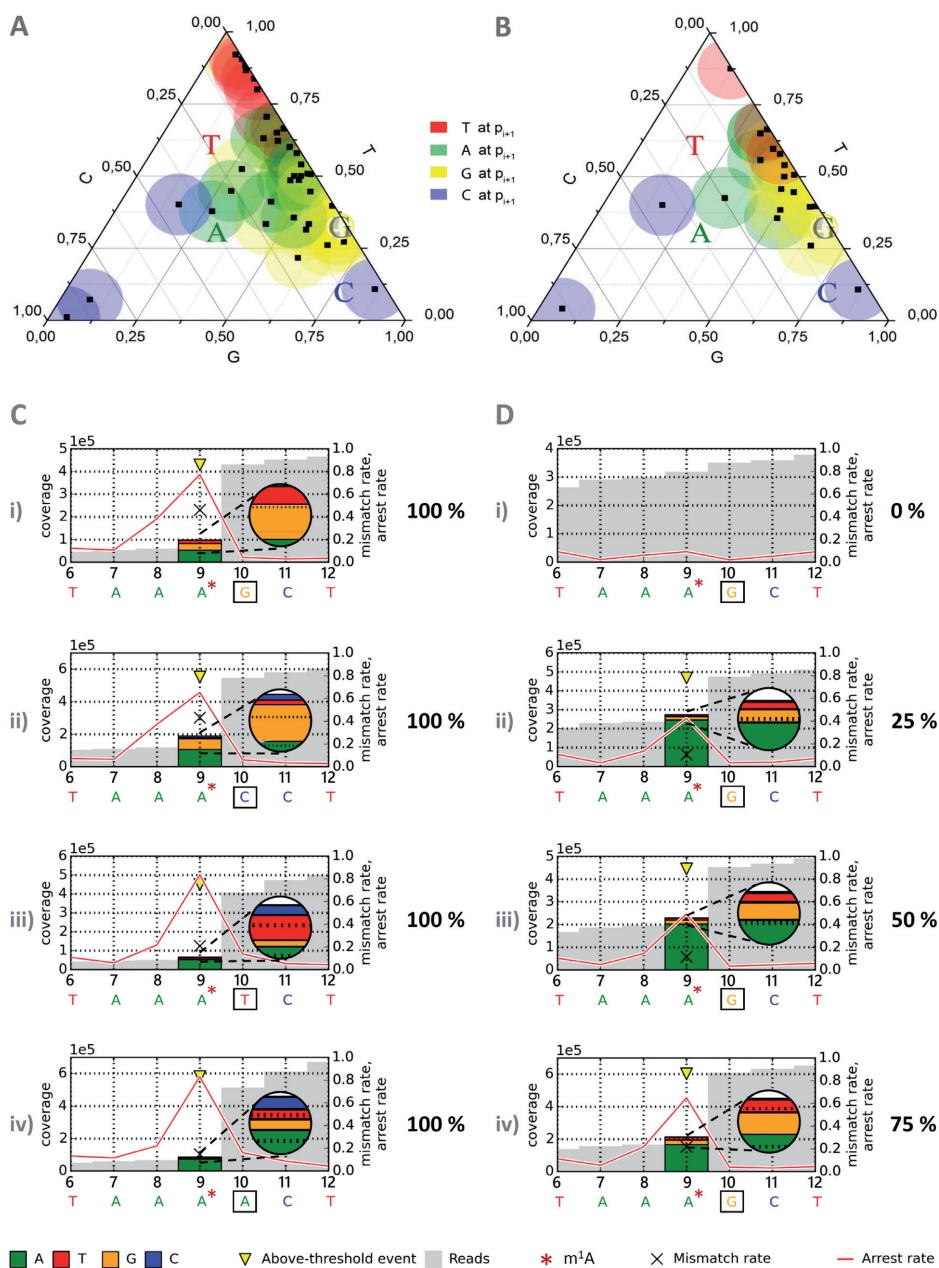


Figure 3. Revolver assay. Revolver oligonucleotides feature permutation of the four major nucleotides at a position of interest, here the +1 position ($3'$ to m^1A). For a position p , the arrest rate reflects the relative amount of reads ending at $p + 1$ (i.e. not covering p) out of all reads covering $p + 1$. (A) Ternary plot of mismatch composition of 41 natural m^1A sites (black dots) for base configurations guanosine (yellow), cytidine (blue), uridine (red, T in mapping profile) and adenosine (green) at position +1 w.r.t. m^1A . Data points from revolver oligonucleotides are represented as colored letters corresponding to the color code also used in (C) and (D). (B) Twenty-two hierarchically clustered data points derived from initial 41 measurements in (A). (C) Mismatch composition at m^1A site for base configurations guanosine (i), cytidine (ii), uridine (iii, T in mapping profile) and adenosine (iv) at position +1 w.r.t. m^1A in sequencing profiles of synthetic oligonucleotides. (D) RT signature by modification level. Arrest rates and mismatch contents at different ratios of modified and unmodified equivalents of revolver oligonucleotide are shown: 0% m^1A in (D-i), 25% in (D-ii), 50% in (D-iii), 75% in (D-iv) and 100% m^1A in (C-i).

5 RESULTS AND DISCUSSION

9958 Nucleic Acids Research, 2015, Vol. 43, No. 20

Table 1. QQQ parameters of the dynamic MRM method

Mod. nucleoside	Precursor ion [m/z]	Product ion [m/z]	Fragm. voltage [V]	Coll. energy [eV]	Cell accel. voltage [V]	Time segment [min]
¹² C-m ¹ A	282	150	92	17	2	5–8.5
¹³ C-m ¹ A	293	156	92	17	2	5–8.5

Table 2. m¹A sites

RNA spec.	Position	Organism	Distinct RNAs	Replicates
Confirmed				
tRNA cyt.	58	Yeast	20	2
tRNA mit.	9	Human	13	1
rRNA	645	Yeast	1	2
rRNA	2142	Yeast	1	2
rRNA	1309	Human	1	1
rRNA	964	<i>Streptomyces pactum</i>	1	1
Artif. oligo.	9*		2	2
Revolver oligo.	9*		4	1
tRNA ^{Arg-UCG} cyt.	58	<i>Trypanosoma brucei</i>	1	2
Unconfirmed				
rRNA	1136	Mouse	1	2
tRNA mit.	9	Human	1	1
tRNA cyt.	58	<i>Trypanosoma brucei</i>	15	1

Confirmed instances include published and self-designed m¹A sites, whereas unconfirmed sites rely on homologous identification. Distinct RNAs refers to the number of non-redundant RNAs, in which m¹A signatures were found. * labeled synthetic oligoribonucleotides contain m¹A₉ in a sequence derived from tRNA^{Lys} of *Homo sapiens*.

out, and share high cytidine and low thymidine content as a common characteristic.

Of the originally 54 m¹A instances present in our data, experimental replicates were averaged, leading to the 41 data points plotted in Figure 3A. Because certain sequence contexts were over-represented in that dataset, we further reduced the dataset by averaging data points from RNAs of over 95% sequence identity (e.g. tRNA sequences containing SNPs) and a final averaging step left only sequences differing at positions –1, +1 and +2, relative to the m¹A site. This dataset, which is plotted in Figure 3B, only incompletely covers the permutation space of said three positions (Supplementary Figure S5 D). Therefore, to cover more of the remaining sequence space, we investigated the RT profiles of synthetic m¹A-containing oligoribonucleotides. These oligoribonucleotides were derived from the naturally occurring m¹A containing sequence of human mitochondrial tRNA^{Lys} (11). In what we termed ‘revolver’ concept, position +1 was systematically variegated, such that the influence of the respective nucleotide could be assessed in direct comparison. The resulting RT-profiles, which are visualized in Figure 3C, point out a pronounced effect of position +1. As is apparent by visual inspection of Figure 3A and B, in which the revolver data are highlighted by colored letters, they reflect well the clustering of the respective natural instances. This visually apparent clustering was statistically verified (computational details in Supplementary Figure S5 E, F and Method S1). Further computational inspection of the revolver-extended experimental dataset (as detailed in the supplement) did not reveal any significant influence of positions +2 and –1. Note that all revolver oligonucleotides in Figure 3B show similar arrest rates, suggesting that the sequence context at the +1 position does not significantly influence the reverse transcription arrest rate.

Quantification of m¹A occupancy

Synthetic oligoribonucleotides were also used to gauge the effect of incomplete occupancy of the modification site. Figure 3D shows profiles obtained from the unmodified oligoribonucleotide of wild-type sequence mixed with increasing amounts of the corresponding m¹A containing oligoribonucleotides. Clearly visible, both, the arrest rate and the misincorporation increase linearly with the fraction of m¹A. This suggests, that some of the biological samples might be incompletely modified and that RT-profiles may eventually be used to gauge modification efficiency after thorough calibration. Therefore, in addition to verifying the presence or absence of m¹A, we have quantified the m¹A content by LC-MS, using a recently developed biosynthetic stable isotope labeled standard (24).

Figure 4 shows chromatograms of the four revolver-oligoribonucleotides, from which an m¹A content of about 80 ± 10% at position 9 was calculated. Only traces of m⁶A, a known rearrangement product of m¹A, were found, therefore a possibility of incomplete occupancy at m¹A₉ even in synthetic samples remains. Not surprisingly, a plot of mismatch rate and arrest rate as a function of m¹A content in Figure 4C suggests a linear dependence of both parameters, but neither correlation is precise enough to confirm or discard the possibility of incomplete m¹A modification in the revolver oligoribonucleotides. LC-MS quantification of the m¹A sites (Figure 4B) in yeast rRNA by analysis of the single knockouts and double knockout yielded 0.7 mol m¹A per mol rRNA for each of both sites, which is consistent with a total of 1.4 mol m¹A per mol rRNA in the wild-type. Interestingly, the profiles vary strongly, although both sites have similar occupancy. Thus, while arrest and mismatch rate at position m¹A₆₄₅ (Figure 2) correlate at least roughly with the m¹A content, the profile at

5.1 Detection of m¹A by NGS

Nucleic Acids Research, 2015, Vol. 43, No. 20 9959

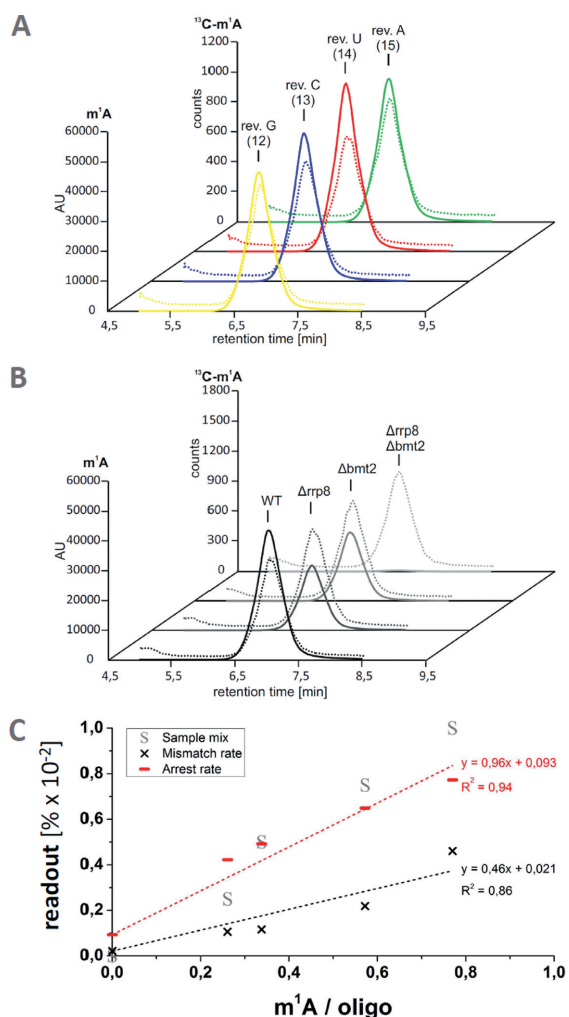


Figure 4. Quantification of m¹A by LC-MS using a biosynthetic internal standard. LC-MS/MS chromatograms showing the m¹A and ¹³C-labeled m¹A peaks in the revolver oligonucleotides (A) and in 25S rRNA from wild-type and *rrp8/bmt2* knockout yeast (B). Continuous lines represent the peaks of unlabeled m¹A, dotted lines those of ¹³C-labeled m¹A added as an internal standard (24). To ensure inter-sample comparability of the m¹A peaks, the peak heights were adjusted to the respective ¹³C-m¹A peaks and normalized to the injected amount of oligonucleotide or 25S rRNA. The amount of analyzed oligonucleotide or 25S rRNA was determined by calculating the amount of adenosine in the respective samples using the UV peak of adenosine and dividing the amount by the number of adenosines per molecule. AU—arbitrary units. (C) Plot of RT signature occupancy by m¹A content.

m¹A₂₁₄₂ incorrectly suggests a much lower modification occupancy. Of note, LC-MS analysis (Figure 4B) and RNA-Seq analysis were conducted with aliquots from the same rRNA preparation, and while the quantification of fractional occupancy is apparently fraught with a ~10% error in precision, the relative comparison of rRNA from both knockout mutants is largely more precise, because numerous error sources average out (24). Comparison of the pro-

files of m¹A₆₄₅ and m¹A₂₁₄₂ clearly show differences despite identical occupancy. In the case of m¹A₂₁₄₂, a large fraction of correctly incorporated dTTP, visualized as green bar, erroneously suggests a significant fraction of unmodified A residues in the RNA template. This leads to the conclusion that RT-profiles have limited use in the quantification of fractional occupancy, in that they tend to underestimate the degree of fractional occupation because the RT does occasionally incorporate the correct dTTP even when challenged by m¹A. Inversely however, the sum of RT-arrest and misincorporation provides a plausible lower limit, since these events clearly derive from events occurring only on an m¹A containing RNA template. This finding has deeper implications and consequences, in particular for structural probing experiments with DMS.

Confirmation of m¹A sites predicted by homology

Despite the variations described above, the signature of m¹A appears characteristic already by visual inspection of RNA-Seq representations in Figures 2 and 3. Of obvious interest is the use of such data for the detection of m¹A residues where they have not been detected by other methods. Arguably the easiest application is the qualitative confirmation of m¹A at putative sites that show plausible homology to known sites. For example, human 28S rRNA was reported to contain an m¹A residue at position 1309 (29), while the corresponding rRNA from mouse has not yet been analyzed for this modification. Figure 5 shows a strong m¹A signature at the corresponding position in human rRNA, as well as at position 1136 of mouse rRNA, which is homologous to the human site. Another example is the m¹A signature at position 9 of human mitochondrial tRNA^{Asn} (Figure 5B), of which the bovine homolog has recently been sequenced (30). These profiles plausibly show that an RT-signature can qualitatively confirm the presence of m¹A.

To apply this identification by computer-aided visual inspection in a more challenging biological question, we have applied it to an organism in which the occurrence of m¹A in tRNA was little investigated, namely *T. brucei*. From a dataset obtained by application of the library preparation protocol outlined above to total RNA (22), we isolated the profiles of tRNAs. By visual inspection, 16 species showed a clear m¹A signature, as shown in Supplementary Figure S10A. Importantly, the sequence dependence of their mismatch distribution, which is plotted in Supplementary Figure S10C, agrees very well with the authentic one in Figure 3B. To further verify the actual existence of m¹A in at least one of these species, we isolated tRNA^{ARG.UGG} by hybridization with a biotinylated cDNA, and subsequent sequestration on streptavidin-beads, as detailed in the ‘Materials and Methods’ section. The purified tRNA was submitted to both-LC-MS analysis and renewed RNA Seq. Both confirmed the presence of m¹A. LC-MS analysis suggested near complete occupancy, i.e. one m¹A residue per tRNA molecule (Supplementary Table S7), and the m¹A signature obtained from the isolated tRNA (Supplementary Figure S10B) is in excellent agreement with that of the bulk tRNA, experimentally confirming that mismatching effects are indeed minor.

5 RESULTS AND DISCUSSION

9960 Nucleic Acids Research, 2015, Vol. 43, No. 20

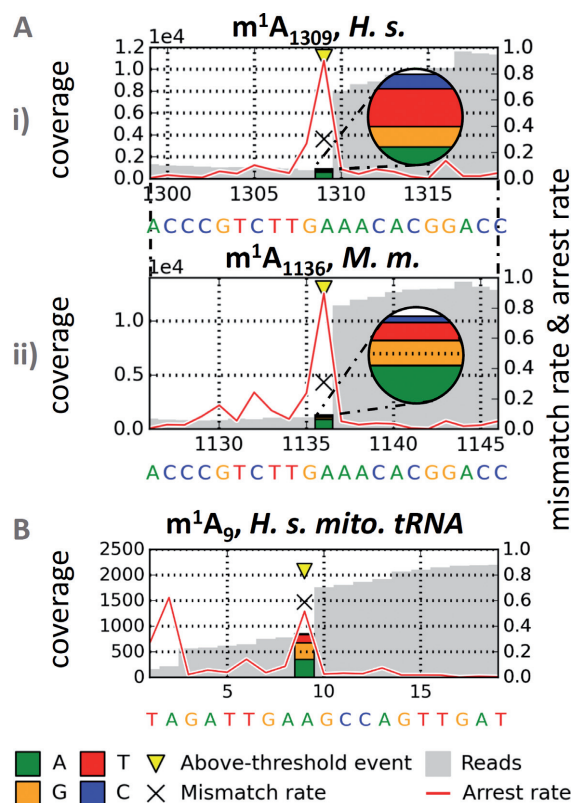


Figure 5. Homology based confirmation of m¹A. For a position p, the arrest rate reflects the relative amount of mapped reads ending at p + 1, i.e. not covering p. (A) Homologous identification of m¹A₁₁₃₆ in murine 28S rRNA (i) by alignment to human sequence containing m¹A₁₃₀₉ (ii). (B) m¹A₉ in human mitochondrial tRNA, identified by alignment to identical bovine sequence with published m¹A₉.

Supervised prediction of m¹A by machine learning

To quantitatively assess how robustly m¹A signatures can distinguish actual modification sites from non-m¹A sites in RNA-Seq data, a supervised prediction of m¹A by machine learning was conducted. Known instances of m¹A and non-m¹A sites in equal numbers were fed to a machine learning algorithm (overall workflow depicted in Supplementary Figure S6). Thus, 45 m¹A signatures (coverage ≥ 10 , 3'-adjacent coverage ≥ 15 , taking isoacceptors into account as separate entities) of tRNA, rRNA and artificial oligonucleotides were merged with an equal amount of data points from non-m¹A sites randomly drawn from the adenosine pool of *bona fide* m¹A-containing datasets: mitochondrial (human) and cytosolic (yeast) tRNA and rRNA (yeast and mouse), according to setting (i) in the 'Materials and Methods' section. This dataset was fed to a RF model (500 trees) classifying both kinds of adenosine instances. Briefly, an RF (31) is a machine learning model for object classification by an ensemble of decision trees. Under randomization in training, binary forks are formed in each individual tree, used to differentiate objects according to their features,

10 rep. 5-fold cross-validation

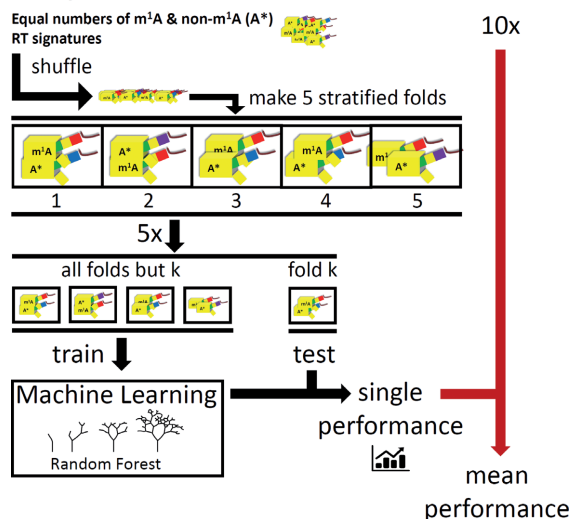


Figure 6. Validation outline for supervised prediction. RT signatures (yellow) of m¹A and non-m¹A (A*) sites and are distributed into subsamples, termed folds, with uniform ratios (stratification) m¹A / A*. The system was tuned toward both, sensitivity and specificity by equal abundance of each class, minimizing learning biases due to *a priori* class probabilities. In each of 10 repetitions (10 \times), the Random Forest was trained on another four of five possible fold combinations (5 \times) and tested on the respective left-out fold.

based on information content. The final object classification is a consensus of all class votes returned by the single trees. Features visible to the RF-classifier included: arrest rate *a*, mismatch rate *m*, the *m/a* ratio, the mismatch composition (fractions of G, T and C), and a parameter that we termed CSA. The latter is defined as the fold change of the site's *a* with respect to its sequence environment of five bases up- and five bases downstream (details in the 'Materials and Methods' section).

Despite its documented impact we did not include the identity of the +1 neighboring base, in order to avoid biases and overfitting side effects in the training run. We applied a five-fold stratified cross-validation (Figure 6), i.e. the data were divided into parts parts, of which four-fifth were used for a training run and one-fifth as a test dataset, for which the model was tasked to classify all adenines. In a total of 10 runs, during which the five parts were permuted between training and testing sets, the model scored better than 97% for both, sensitivity and specificity (setting (i), detailed in 'Materials and Methods' section). In setting (ii), a more stringent variation of this validation, the non-m¹A sites fed to the RF were deliberately chosen among those that showed the closest resemblance to m¹A-signatures among the non-m¹A sites. Under these circumstances, the values dropped to 89% (SD = $\pm 2.4\%$) for sensitivity, and 87% (SD = $\pm 2.8\%$) for specificity (averaged from ten repetitions, all statistics are given in Supplementary Table S5). Of interest is the deliberate inclusion of other modified adenosine residues in the training set for 'non-m¹A', in particular of two ubiquitous consecutive m^{6,6}A rRNA residues at posi-

tions 1781 and 1782 of yeast 18S rRNA. This modification type features a methyl group on the Watson–Crick face, as does m¹A and was reported to show a similar misincorporation pattern (16). Indeed, m^{6,6}A₁₇₈₁ shows a signature (shown in Supplementary Figure S9) that, by visual inspection, is indistinguishable from that of m¹A (Figure 2C) and is classified as such by the algorithm as well. On the other hand, the adjacent m^{6,6}A₁₇₈₂ also shows a clear signature, which is however, different from the typical m¹A signature. Accordingly, it is correctly classified as ‘non-m¹A’ by the algorithm. Of note, when the model was trained on the entire amount of available tRNA instances of m¹A with as many random non-m¹A adenosine signatures (setting (iii)), all presented (five) rRNA sites of m¹A (2× *S. cerevisiae*, 1× *S. pactum*, 1× *M. musculus*, 1× *H. sapiens*) were correctly identified with a specificity >99.9%.

In addition, a leave-one-out cross-validation was performed. This roughly corresponds to the above stratification concept with the number of folds maximally increased, such that the test-fold contains only a single positive and a single negative m¹A instance. As expected, performance increased with availability of additional training instances (Supplementary Table S6). This thus provides additional confirmation of the overall feasibility of the concept and underlines the need to maximize training instances for efficient machine learning.

Out of concern that our dataset of 45 positive m¹A instances might be too limited for complex classifiers such as RF, we compared the performance of the latter with a more basic method, namely k-Nearest Neighbor (kNN). In a so called receiver operating characteristic (ROC) analysis, the area under the curve corresponds to the probability with which a model scores a random positive m¹A instance higher than a negative one. As can be deduced from the various curves plotted in Supplementary Figure S7, the RF model consistently outperforms various kNN setups by a large margin. The shape of the RF curve also illustrates that the RF achieved considerable sensitivity while maintaining high specificity.

Finally, we have analyzed in depth, which parameters have been retained by the RF model as most informative for the recognition of m¹A sites. From inspection of the trained RF, it became clear, that both, arrest rate and at mismatch rate played an important role for performance. A more detailed analysis was conducted by a leave-feature-out analysis, which measures performance in various permutations of incomplete feature combinations. The results, which are shown in Supplementary Figure S8, clearly confirm our initial approach to m¹A signature identification, namely that neither arrest rate nor mismatch analysis alone come close to the performance of their combination.

DISCUSSION

Encounters of reverse transcriptase enzymes with non-canonical nucleotides are about to become a focus of intense research. Early investigations into the effect of m¹A were mostly concerned with the application in structural probing *in vitro* (1), and the replication of the HIV genome *in vivo*, which actually strongly relies on RT-arrest induced by m¹A₅₈ of the HIV-primer tRNA^{Lys3} (32–34). The topic is

subject to renewed impetus, as RNA-Seq based approaches are being developed to detect RNA modifications on a transcriptome wide scale. Such analyses for m⁵C (35), m⁶A (36) and pseudouridine (37–39) have recently revolutionized the RNA modification field, and the common belief is that more modifications types are to be found in transcriptomes by similar methods. So-called PSI-Seq (37–39) relies completely on RT-arrest upon encounter of the enzyme with a CMC modified nucleoside and an understanding of the efficiency of RT-arrest clearly will improve the accuracy of such approaches.

Here, we present an in-depth investigation of the effect of m¹A in an RNA template on the composition of cDNA fragments generated during reverse transcription. Where previous studies have provided a more general picture of various modifications in parallel (16,17), we focused on a single modification species and characterized the resulting arrest rate as well as the misincorporation pattern for over 50 RNA sequences. We taught the common characteristics to a computer learning program for supervised prediction and identification. The method is well capable of qualitatively confirming the presence of m¹A at a defined candidate position, such as e.g. A₉₆₄ in rRNA from *S. pactum*. Since this methylation mediates resistance to the pactamycin (12), our approach can conceivably be applied to the detection of antibiotic resistance. Because m^{6,6}A blocks the Watson–Crick face of an adenosine like m¹A and leaves signatures as well, we expect that moderate adaptation of parameters will allow the monitoring of m^{6,6}A at position 1519 in bacterial rRNA, which mediates resistance to kasugamycin (40). From our results, we can project that the limiting step in this endeavor is likely to be a larger training set of *bona fide* m^{6,6}A sites.

Parameters that shape the RT-signature

As an important message, the presented data suggest, that the amount of misincorporation by the RT enzyme is very substantial, resulting in a non-negligible read-through efficiency. It is known from the literature, that read-through by RT-enzymes *in vitro* may depend on a variety of parameters, including e.g. the dNTP concentration (41) in the case of 2'-OMe modifications (42). Certainly, the nature of the enzyme itself is important in the encounter with an RNA modification (43), and we can expect key parameters of *in vitro* conditions such as pH, ion strength and divalent cations to be important as well. The present study has kept these parameters constant and focused on the identification of features residing in the RNA template itself. Our investigations into the influence of neighboring nucleotides –1, +1 and +2 revealed a clear influence of the nature of the +1 nucleotide, situated 3' to the m¹A residue. Beyond the scope of detecting m¹A residues at new positions in transcriptomes, this insight has significant implications for the interpretations of structural probing data obtained by primer extension. With respect to the interpretation of structural probing data of m¹A residues generated using DMS, the classification of RT-arrest signals as weak, intermediate, or strong (44), may now be refined by taking into account the penultimate nucleotide.

5 RESULTS AND DISCUSSION

9962 *Nucleic Acids Research*, 2015, Vol. 43, No. 20

Higher order structure of RNA has long been known to negatively affect the efficiency of primer extension, a fact frequently encountered in structural probing of e.g. rRNA, where a strong noise from RT-arrest signals made data interpretation difficult. Our data, however, suggest that RNA structure may, in certain cases, even facilitate read-through. This is exemplified by comparison of the signatures of rRNA (Figure 2) with revolver oligonucleotides (Figure 3). The latter can reasonably be assumed to be weakly structured (11). In each case, the m¹A residue causes >80% arrest rate, while the two rRNA sites show strongly diverging arrest albeit being equally modified to ~70%. Said discrepancy must mostly stem from outside the immediate neighboring sequence context, which is quite similar between both sites.

Limitations

From the above discussion flow a number of limitations of the presented method at its current state. Clearly, estimates of fractional occupancy of an m¹A site can be semi-quantitative at best, and only after calibration as shown in Figure 3D. The differential strength of equally modified sites in rRNA (Figure 2A) points to further factors that influence the strength of the m¹A signature, whose identification must await further work. In contrast, analyses of different mapping strategies as detailed in Supplementary Figure S3 show, that the influence of the mapping strategy was efficiently minimized in the k1 regime we applied. This analysis revealed that for tRNA-related reads, mismapping in a k3 regime, which allows up to three mappings per read, strongly depends on the tRNA species, and may potentially outnumber the reads derived from conservative k1 regime mapping by more than an order of magnitude (Supplementary Figure S3A), although the k3-related mismapping was mostly inside isoacceptor groups. Interestingly, differences in the signature-relevant parameters arrest rate and mismatch content were relatively minor (Supplementary Figure S3C), when k1, k3 and 'best' (default) settings were compared, and this was exemplarily visualized for a selected tRNA species in Supplementary Figure S3D. For scenarios with higher cross-mapping rates, reporting the best alignment on cost of computation time may be considered, although this can lead to undesired suppression of mismatch information. In this study, variation of signature parameters due to mapping artefacts is clearly smaller than what we expect from experimental parameters. Given that salt conditions, temperature and the type of enzyme are known to affect polymerization characteristics e.g. in PCR reactions (45), we will turn our attention to these parameters in the near future.

Potential applications, prediction performance and scope

In the rapidly developing field of RNA modifications there is an urgent need for new methods, which are sought for applications to a variety of biological questions. These include transcriptome-wide searches to detect new modifications sites as well as quantification e.g. in the context of a response to outside stress. With respect to the latter, elevated temperatures were recently shown to ablate a thiol-modification

in yeast (46,47). In analogy, we have investigated potential changes in the m¹A-signatures of tRNAs from yeast raised at normal temperature versus 39°C, but failed to identify any differences (data not shown). Although this might be due to the limited quantification accuracy (compare Figure 3D and related material), total ablation would have been detectable.

At the current state, our machine learning algorithm can distinguish m¹A from an unmodified adenosine with very good accuracy, if these two possibilities are the only elements in the training data. Not unexpectedly, when other modifications are forcibly included as non-m¹A training data, the performance drops. The erroneous classification of an m^{6,6}A as m¹A is readily rationalized: m^{6,6}A carries a methyl group on its Watson-Crick face, and therefore leads to RT-arrest as well. Furthermore, a previous study suggests that all the adenosine modifications have similar misincorporation patterns (16). The latter argument must be attenuated somewhat, since our current analysis shows strong variability even within m¹A samples (Supplementary Figure S4). Still, this instance once more illustrates, that once a candidate site is identified, further evidence, such as sequence homology to known sites, RNA-Seq data from relevant knockout organisms, or biochemical analysis is needed for confirmation.

The performance of a prospective large-scale prediction depends on the quality and quantity of both, positive and negative training instances. Our m¹A pool covers a large number of sequence contexts, but is clearly biased in that some portions of the sequence space are missing in the training pool. Obviously, the sequence context of m¹A occurrence in nature is not random, but biased by biological evolution, e.g. of the m¹A methyltransferases (48,49). Since the algorithm is based on learning, its current version will be more successful at predicting m¹A sites situated in a similar sequence context, and it is prone to perform poorly in the prediction of sites in a radically new sequence context, including in particular such situated in clusters containing multiple different modifications. The training pool of non-m¹A instances determines the success along similar lines.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We are grateful to F. Blanloeil-Oillo (NGS core facility of FR3209, BioPole, Nancy) for excellent technical assistance in MiSeq sequencing of libraries.

FUNDING

Joint ANR-DFG project [HTRNAMod to Y.M and M.H.] (designations N° ANR-13-ISV8-0001 01 and HE 3397/8-1); THE DFG SPP1784 [HE 3397/13-1, DFG HE3397/14-1]; Fonds National de la Recherche (FRS/FNRS) (to D.L.J.L.); Walloon Region (DGO6, 'CWALity') (to D.L.J.L.); National Institutes of Health GM084065 (to J.D.A.). Funding for open access charge: University of Mainz.

5.1 Detection of m¹A by NGS

Nucleic Acids Research, 2015, Vol. 43, No. 20 9963

Conflict of interest statement. None declared.

REFERENCES

- Lempereur, L., Nicoloso, M., Riehl, N., Ehresmann, C., Ehresmann, B. and Bachellerie, J.P. (1985) Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible. *Nucleic Acids Res.*, **13**, 8339–8357.
- Motorin, Y., Muller, S., Behm-Ansmant, I. and Branlant, C. (2007) Identification of modified residues in RNAs by reverse transcription-based methods. *Methods Enzymol.*, **425**, 21–53.
- Behm-Ansmant, I., Helm, M. and Motorin, Y. (2011) Use of specific chemical reagents for detection of modified nucleotides in RNA. *J. Nucleic Acids*, **2011**, 408053.
- Mortimer, S.A., Trapnell, C., Aviran, S., Pachter, L. and Lucks, J.B. (2012) SHAPE-Seq: high-throughput RNA structure analysis. *Curr. Protoc. Chem. Biol.*, **4**, 275–297.
- Roovers, M., Wouters, J., Bujnicki, J.M., Tricot, C., Stalon, V., Grosjean, H. and Droogmans, L. (2004) A primordial RNA modification enzyme: the case of tRNA (m¹A) methyltransferase. *Nucleic Acids Res.*, **32**, 465–476.
- Sharma, S., Watzinger, P., Kotter, P. and Entian, K.D. (2013) Identification of a novel methyltransferase, Bmt2, responsible for the N-1-methyl-adenosine base modification of 25S rRNA in *Saccharomyces cerevisiae*. *Nucleic Acids Res.*, **41**, 5428–5443.
- Schmidt, W., Arnold, H.H. and Kersten, H. (1975) Biosynthetic pathway of ribothymidine in *B. subtilis* and *M. lysodeikticus* involving different coenzymes for transfer RNA and ribosomal RNA. *Nucleic Acids Res.*, **2**, 1043–1051.
- Srivastava, R. and Gopinathan, K.P. (1987) Ribosomal RNA methylation in *Mycobacterium smegmatis* SN2. *Biochem. Int.*, **15**, 1179–1188.
- Machnicka, M.A., Milanowska, K., Osman Oglou, O., Purta, E., Kurkowska, M., Olchowiak, A., Januszewski, W., Kalinowski, S., Dunin-Horkawicz, S., Rother, K.M. et al. (2013) MODOMICS: a database of RNA modification pathways—2013 update. *Nucleic Acids Res.*, **41**, D262–D267.
- Juhling, F., Morl, M., Hartmann, R.K., Sprinzl, M., Stadler, P.F. and Putz, J. (2009) tRNAdb 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.
- Helm, M., Brule, H., Degoul, F., Cepanec, C., Leroux, J.P., Giege, R. and Florentz, C. (1998) The presence of modified nucleotides is required for cloverleaf folding of a human mitochondrial tRNA. *Nucleic Acids Res.*, **26**, 1636–1643.
- Ballesta, J.P. and Cundliffe, E. (1991) Site-specific methylation of 16S rRNA caused by pct, a pactamycin resistance determinant from the producing organism, *Streptomyces pactum*. *J. Bacteriol.*, **173**, 7213–7218.
- Blanco, S., Dietmann, S., Flores, J.V., Hussain, S., Kutter, C., Humphreys, P., Lukk, M., Lombard, P., Treps, L., Popis, M. et al. (2014) Aberrant methylation of tRNAs links cellular stress to neuro-developmental disorders. *EMBO J.*, **33**, 2020–2039.
- Hodgkinson, A., Idaghdour, Y., Gbeha, E., Grenier, J.C., Hip-Ki, E., Bruat, V., Goulet, J.P., de Malliard, T. and Awadalla, P. (2014) High-resolution genomic analysis of human mitochondrial RNA sequence variation. *Science*, **344**, 413–415.
- Zhang, Y., Yuan, F., Wu, X., Rechkoblit, O., Taylor, J.S., Geacintov, N.E. and Wang, Z. (2000) Error-prone lesion bypass by human DNA polymerase ϵ . *Nucleic Acids Res.*, **28**, 4717–4724.
- Ryvkin, P., Leung, Y.Y., Silverman, I.M., Childress, M., Valladares, O., Dragomir, I., Gregory, B.D. and Wang, L.S. (2013) HAMR: high-throughput annotation of modified ribonucleotides. *RNA*, **19**, 1684–1692.
- Findeiss, S., Langenberger, D., Stadler, P.F. and Hoffmann, S. (2011) Traces of post-transcriptional RNA modifications in deep sequencing data. *Biol. Chem.*, **392**, 305–313.
- Cahová, H., Winz, M.L., Höfer, K., Nübel, G. and Jäschke, A. (2015) NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs. *Nature*, **519**, 374–377.
- Shepherd, M.D., Kharel, M.K., Bosserman, M.A. and Rohr, J. (2010) Laboratory maintenance of *Streptomyces* species. *Curr. Protoc. Microbiol.*, Chapter 10, Unit 10E 11.
- Winz, M.L. (2014) *Biological, chemical and computational investigations on RNA function and modification*. Ph.D. Thesis, Heidelberg University.
- Liaw, A. and Wiener, M. (2002) Classification and regression by randomForest. *R NEWS*, **2**, 18–22.
- Rubio, M.A., Paris, Z., Gaston, K.W., Fleming, I.M., Sample, P., Trotta, C.R. and Alfonzo, J.D. (2013) Unusual noncanonical intron editing is important for tRNA splicing in *Trypanosoma brucei*. *Mol. Cell*, **52**, 184–192.
- Kellner, S., Neumann, J., Rosenkranz, D., Lebedeva, S., Ketting, R.F., Zischler, H., Schneider, D. and Helm, M. (2014) Profiling of RNA modifications by multiplexed stable isotope labelling. *Chem. Commun.*, **50**, 3516–3518.
- Kellner, S., Ochel, A., Thüring, K., Spenkuch, F., Neumann, J., Sharma, S., Entian, K.D., Schneider, D. and Helm, M. (2014) Absolute and relative quantification of RNA modifications via biosynthetic isotopomers. *Nucleic Acids Res.*, **42**, e142.
- Giege, R., Sissler, M. and Florentz, C. (1998) Universal rules and idiosyncratic features in tRNA identity. *Nucleic Acids Res.*, **26**, 5017–5035.
- Levenshtein, V. (1965) Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, **10**, 845–848.
- Anderson, J., Phan, L. and Hinnebusch, A.G. (2000) The Gcd10p/Gcd14p complex is the essential two-subunit tRNA(1-methyladenosine) methyltransferase of *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. U.S.A.*, **97**, 5173–5178.
- Kadaba, S., Krueger, A., Trice, T., Krecic, A.M., Hinnebusch, A.G. and Anderson, J. (2004) Nuclear surveillance and degradation of hypomodified initiator tRNA^{Met} in *S. cerevisiae*. *Genes Dev.*, **18**, 1227–1240.
- Maden, B.E. (1990) The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucleic Acid Res. Mol. Biol.*, **39**, 241–303.
- Suzuki, T. and Suzuki, T. (2014) A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs. *Nucleic Acids Res.*, **42**, 7346–7357.
- Breiman, L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.
- Auxilien, S., Keith, G., Le Grice, S.F. and Darlix, J.L. (1999) Role of post-transcriptional modifications of primer tRNA^{Lys3} in the fidelity and efficacy of plus strand DNA transfer during HIV-1 reverse transcription. *J. Biol. Chem.*, **274**, 4412–4420.
- Wildauer, M., Zemora, G., Liebeg, A., Heisig, V. and Walschid, C. (2014) Chemical probing of RNA in living cells. *Methods Mol. Biol.*, **1086**, 159–176.
- Renda, M.J., Rosenblatt, J.D., Klimatcheva, E., Demeter, L.M., Bambara, R.A. and Planelles, V. (2001) Mutation of the methylated tRNA(Lys)(3) residue A58 disrupts reverse transcription and inhibits replication of human immunodeficiency virus type 1. *J. Virol.*, **75**, 9671–9678.
- Schaefer, M., Pollex, T., Hanna, K. and Lyko, F. (2009) RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.*, **37**, e12.
- Domimissini, D., Moshitch-Moshkovitz, S., Schwartz, S., Salmon-Divon, M., Ungar, L., Osenberg, S., Cesarkas, K., Jacob-Hirsch, J., Amariglio, N., Kupiec, M. et al. (2012) Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq. *Nature*, **485**, 201–206.
- Carlile, T.M., Rojas-Duran, M.F., Zinshteyn, B., Shin, H., Bartoli, K.M. and Gilbert, W.V. (2014) Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, **515**, 143–146.
- Lovejoy, A.F., Riordan, D.P. and Brown, P.O. (2014) Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in *S. cerevisiae*. *PLoS One*, **9**, e110799.
- Schwartz, S., Bernstein, D.A., Mumbach, M.R., Jovanovic, M., Herbst, R.H., Leon-Ricardo, B.X., Engreitz, J.M., Guttman, M., Satija, R., Lander, E.S. et al. (2014) Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA. *Cell*, **159**, 148–162.
- Helser, T.L., Davies, J.E. and Dahlberg, J.E. (1971) Change in methylation of 16S ribosomal RNA associated with mutation to kasugamycin resistance in *Escherichia coli*. *Nat. New Biol.*, **233**, 12–14.

5 RESULTS AND DISCUSSION

9964 *Nucleic Acids Research*, 2015, Vol. 43, No. 20

41. Goldschmidt, V., Didierjean, J., Ehresmann, B., Ehresmann, C., Isel, C. and Marquet, R. (2006) Mg²⁺ dependency of HIV-1 reverse transcription, inhibition by nucleoside analogues and resistance. *Nucleic Acids Res.*, **34**, 42–52.
42. Maden, B.E. (2001) Mapping 2'-O-methyl groups in ribosomal RNA. *Methods*, **25**, 374–382.
43. Vilfan, I.D., Tsai, Y.C., Clark, T.A., Wegener, J., Dai, Q., Yi, C., Pan, T., Turner, S.W. and Korlach, J. (2013) Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription. *J. Nanobiotechnol.*, **11**, 1477–3155.
44. Talkish, J., May, G., Lin, Y., Woolford, J.L. Jr and McManus, C.J. (2014) Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA*, **20**, 713–720.
45. Tee, K.L. and Wong, T.S. (2013) Polishing the craft of genetic diversity creation in directed evolution. *Biotechnol. Adv.*, **31**, 1707–1721.
46. Alings, F., Sarin, L.P., Fufezan, C., Drexler, H.C. and Leidel, S.A. (2015) An evolutionary approach uncovers a diverse response of tRNA 2-thiolation to elevated temperatures in yeast. *RNA*, **21**, 202–212.
47. Han, L., Kon, Y. and Phizicky, E.M. (2015) Functional importance of Psi38 and Psi39 in distinct tRNAs, amplified for tRNAGln(UUG) by unexpected temperature sensitivity of the s2U modification in yeast. *RNA*, **21**, 188–201.
48. Hamdane, D., Guelorget, A., Guerineau, V. and Golinelli-Pimpaneau, B. (2014) Dynamics of RNA modification by a multi-site-specific tRNA methyltransferase. *Nucleic Acids Res.*, **42**, 11697–11706.
49. Takuma, H., Ushio, N., Minoji, M., Kazayama, A., Shigi, N., Hirata, A., Tomikawa, C., Ochi, A. and Hori, H. (2015) Substrate tRNA Recognition Mechanism of Eubacterial tRNA (m1A58) Methyltransferase (TrmI). *J. Biol. Chem.*, **290**, 5912–5925.

Supplementary material to

The reverse transcription signature of N¹-methyladenosine in RNA-Seq is sequence dependent

Ralf Hauenschild¹, Lyudmil Tserovski¹, Katharina Schmid¹, Kathrin Thuring¹, Marie-Luise Winz², Sunny Sharma³, Karl-Dieter Entian³, Ludivine Wacheul⁴, Denis L.J. Lafontaine⁴, James Anderson⁵, Juan Alfonzo⁶, Andreas Hildebrandt⁷, Andres Jäschke⁸, Yuri Motorin^{9,*} and Mark Helm^{1,*}

Table of Contents

	Page
Library preparation - Sequence elements	Tab. S1 2
Library preparation – Outline	Fig. S1 3
Sequence libraries	Tab. S2 4
Pairwise Levenshtein edit distances of cytosolic tRNA sequences of <i>S. cerevisiae</i>	Fig. S2 5
RT signature vs. multiple mappings	Fig. S3 6
m¹A₅₈ signature compilation	Fig. S4 7
RT sequence context	Fig. S5 8
Computational analysis of sequence context mediated mismatch composition	Met. S1 9
Workflow - Data processing for prediction	Fig. S6 10
Profile format	Tab. S3 10
Classification input format	Tab. S4 10
Random Forest performance - 10 rep. 5-fold cross-validation	Tab. S5 10
Random Forest performance - 10 rep. leave-one-m ¹ A-out cross-validation	Tab. S6 10
Receiver operating characteristic (ROC)	Fig. S7 11
Prediction quality vs. number and category of predictors	Fig. S8 12
RT signatures of m^{6,6}As 1781 and 1782	Fig. S9 13
Trypanosomal m¹A₅₈	Fig. S10 14
LC-MS/MS quantification of m¹A	Tab. S7 15

5 RESULTS AND DISCUSSION

Table S1: Library preparation - Sequence elements.

Element	Sequence		
RAdapter	5'-P-CNNNNNNNAGATCGGAAGAGCGTCGTAGGGAAAGAGTGT-3'-C6-spacer		
RTPrimer	5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'		
DAnchorA	5'-GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGG-3'		
DAnchorB	5'-P- AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3'-C6-spacer		
PCR P7 primer	5'-CAAGCAGAAGACGGCATACGAGAT77777777GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT-3'		
PCR P5 primer	5'-AATGATACGGCGACCACCGAGATCTACAC55555555ACACTCTTCCCTACACGACGCTCTTCCGATCT-3'		
Barcodes used with P7	Sequence	Barcodes used with P7	Sequence
N701	TCGCCTTA	N501	TAGATCGC
N702	CTAGTACG	N502	CTCTCTAT
N703	TTCTGCCT	N503	TATCCTCT
N704	GCTCAGGA	N504	AGAGTAGA
N705	AGGAGTCC	N505	GTAAGGAG
N706	CATGCCTA	N506	ACTGCATA
N707	GTAGAGAG	N507	AAGGAGTA
N708	CCTCTCTG	N508	CTAAGCCT
N709	AGCGTAGC		
N710	CAGCCTCG		
N711	TGCCTCTT		
N712	TCCTCTAC		
Oligoribonucleotide type	Sample ID	Sequence	
A-G	17	5'-CACUGUAAAGCUAACUUAGC-3'	
revolver m ¹ A-G	12	5'-CACUGUAAm ¹ AGCUAACUUAGC-3'	
revolver m ¹ A-C	13	5'-CACUGUAAm ¹ ACCUAACUUAGC-3'	
revolver m ¹ A-U	14	5'-CACUGUAAm ¹ AUCUAACUUAGC-3'	
revolver m ¹ A-A	15	5'-CACUGUAAm ¹ AACUAACUUAGC-3'	
hybridization oligo for tRNA ^{Arg} _{UCG}	S24	biotin-CGGCAGGACTCGAACCTGCAACCCTCA	

5.1 Detection of m^1A by NGS

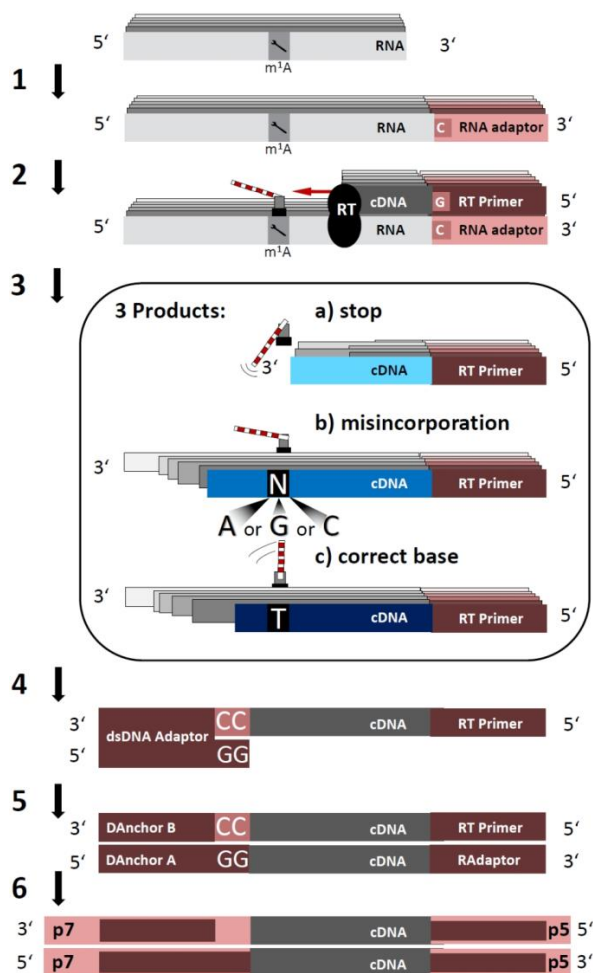


Figure S1: Library preparation - Outline. Note that the presented procedure captures cDNA resulting from abortive reverse transcription events. Primer information is given in Table S1.

5 RESULTS AND DISCUSSION

Table S2: Sequence libraries. Raw reads denote the number of FASTQ sequences obtained from Illumina prior to bioinformatic processing. Mapped reads reflect the relative number of processed reads mappable to references provided to Bowtie2. Value pairs refer to reads from paired end libraries or replicates (sample 1). Bp is the length of these reads in base pairs, where 150 bp were used in the single end (s) mode of sequencing and 151 or 35 + 88 bp for the reads in paired end (p) mode of libraries. For comparability, the average number ($\bar{\theta}$ reads / kb ref) of reads mapped on a kilobase (kb) of an m¹A-annotated target reference sequence is listed, e.g. tRNA^{lⁿⁱ} in sample 3 and rRNA in sample 5. The mean coverage ($\bar{\theta}$ cov. 3' of m¹A) at +1 positions 3' of m¹A provides a reference for the arrest rate, ($\bar{\theta}$ arr. (m¹A)), at the Sm¹A position and $\bar{\theta}$ mism. (m¹A) provides the mismatch contents. Sample 1 values originate from replicates (N=2). Numbers annotated with ^{lni} refer to tRNA^{lⁿⁱ} only. Entries in brackets refer to pairs of m¹A sites in the corresponding sample, such as m¹A₆₄₅ and m¹A₂₁₄₂ in rRNA. The mean is given in front of each bracket.

ID	Interest	Material	Organism	Raw reads	Mapped [%]	Bp	End	$\bar{\theta}$ Reads / kb ref	$\bar{\theta}$ cov. 3' of m ¹ A	$\bar{\theta}$ arr. (m ¹ A) [%]	$\bar{\theta}$ mism. (m ¹ A) [%]
1	Signature pool	total tRNA	<i>S. cerevisiae</i>	2.95 M / 2.23 M	40.0 / 40.7	151	p	370 k / 260 k	10.0 k / 6.6 k	21 / 22	54 / 54
2	Signature pool	total mito. RNA	<i>H. sapiens</i>	2.1 M	14.3	151	p	189 k	1.6 k	62	32
3	m ¹ A ₆₄₅ knockout	total tRNA	<i>S. cerevisiae</i> Δ trm6	5.58 M	89.9 / 89.1	35, 88	p	16.4 M ^{lni}	776 k ^{lni}	0.1 ^{lni}	0.4 ^{lni}
4	Positive control	total tRNA	<i>S. cerevisiae</i>	6.37 M	77.3 / 82.1	35, 88	p	579 k ^{lni}	1.25 k ^{lni}	18.1 ^{lni}	7.0 ^{lni}
5	Positive control	rRNA	<i>S. cerevisiae</i>	4.95 M	47.9	150	s	213 k	9.3 k (10.0 / 8.6 k)	35 (54.0 / 15.0)	28.0 (44.7 / 11.2)
6	single knockout m ¹ A ₆₄₅	rRNA	<i>S. cerevisiae</i> Δ trp8	5.40 M	63.1	150	s	407 k	25.6 k (39.1 k / 12.0 k)	9.15 (2.0 / 16.3)	9.6 (1.2 / 18.0)
7	single knockout m ¹ A ₂₁₄₂	rRNA	<i>S. cerevisiae</i> Δ bmt2	3.82 M	76.4	150	s	308 k	12.6 (13.7 k / 11.6 k)	23.35 (44.7 / 2.0)	24.3 (47.0 / 1.6)
8	double knockout m ¹ A ₆₄₅ and m ¹ A ₂₁₄₂	rRNA	<i>S. cerevisiae</i> Δ trp8 + Δ bmt2	7.37 M	68.33	150	s	402 k	21.6 k (28.3 / 14.8 k)	2.75 (3.2 / 2.3)	0.8 (0.8 / 0.8)
9	m ¹ A on SSU of rRNA	total RNA	<i>S. pectum</i>	6.27 M	1.2 / 0.6	35, 88	p	250 k	6.4 k	76.0	56.9
10	Homologous identification	rRNA	<i>H. sapiens</i>	5.14 M	77.7 / 72.0	35, 88	p	12 k	8.0 k	90.0	30.3
11	Homologous identification	rRNA	<i>M. musculus</i>	7.18 M	71.0 / 68.7	35, 88	p	150 k	11.4 k	89.7	30.9
12	RT sequence context dependency	oligo.	synthetic	1.42 M	92.6 / 86.8	35, 88	p	18.6 M	430 k	76.0	48.9
13	RT sequence context dependency	oligo.	synthetic	1.66 M	87.9 / 82.3	35, 88	p	23.9 M	550 k	54.4	56.7
14	RT sequence context dependency	oligo.	synthetic	1.56 M	84.1 / 80.1	35, 88	p	20.1 M	410 k	82.7	24.5
15	RT sequence context dependency	oligo.	synthetic	1.89 M	88.4 / 68.0	35, 88	p	27.1 M	510 k	81.4	22.2
16	RT sequence context dependency	2 oligo.	ligate	0.37 M	42.7 / 15.1	151	p	(500 k / 400 k)	2.9 k	(44.1 / 60.5)	(39.1 / 27.2)
17	Signature vs. occupancy	oligo.	in vitro transcr.	1.77 M	89.1 / 81.1	35, 88	p	16.0 M	350 k	8.2	3.1
18	Signature vs. occupancy	oligo.	synthetic	2.00 M	90.1 / 83.1	35, 88	p	21.2 M	480 k	41.5	11.7
19	Signature vs. occupancy	oligo.	synthetic	1.72 M	91.0 / 84.4	35, 88	p	20.0 M	450 k	48.5	12.8
20	Signature vs. occupancy	oligo.	synthetic	2.17 M	91.8 / 86.0	35, 88	p	26.4 M	610 k	64.0	25.0
S21	Positive control	total tRNA	<i>S. cerevisiae</i>	1.95 M	44.4 / 49.7	80, 80	p	145 k	5.6 k	51.9	60.2
S22	m ¹ A ₆₄₅ knockout	total tRNA	<i>S. cerevisiae</i>	3.21 M	58.8 / 51.9	80, 80	p	571 k	11.2 k	0.2	0.3
S23	Novel sites	total tRNA	<i>T. brucei</i>	1.36 M	4.8 / 3.9	80, 80	p	7.5 k	0.4 k	36.6	83.0
S24	Signature vs. occupancy	tRNA ^{lⁿⁱ}	<i>T. brucei</i>	1.79 M	13.1 / 12.4	80, 80	p	2.6 M	85 k	18.3	82.8

5.1 Detection of m^1A by NGS

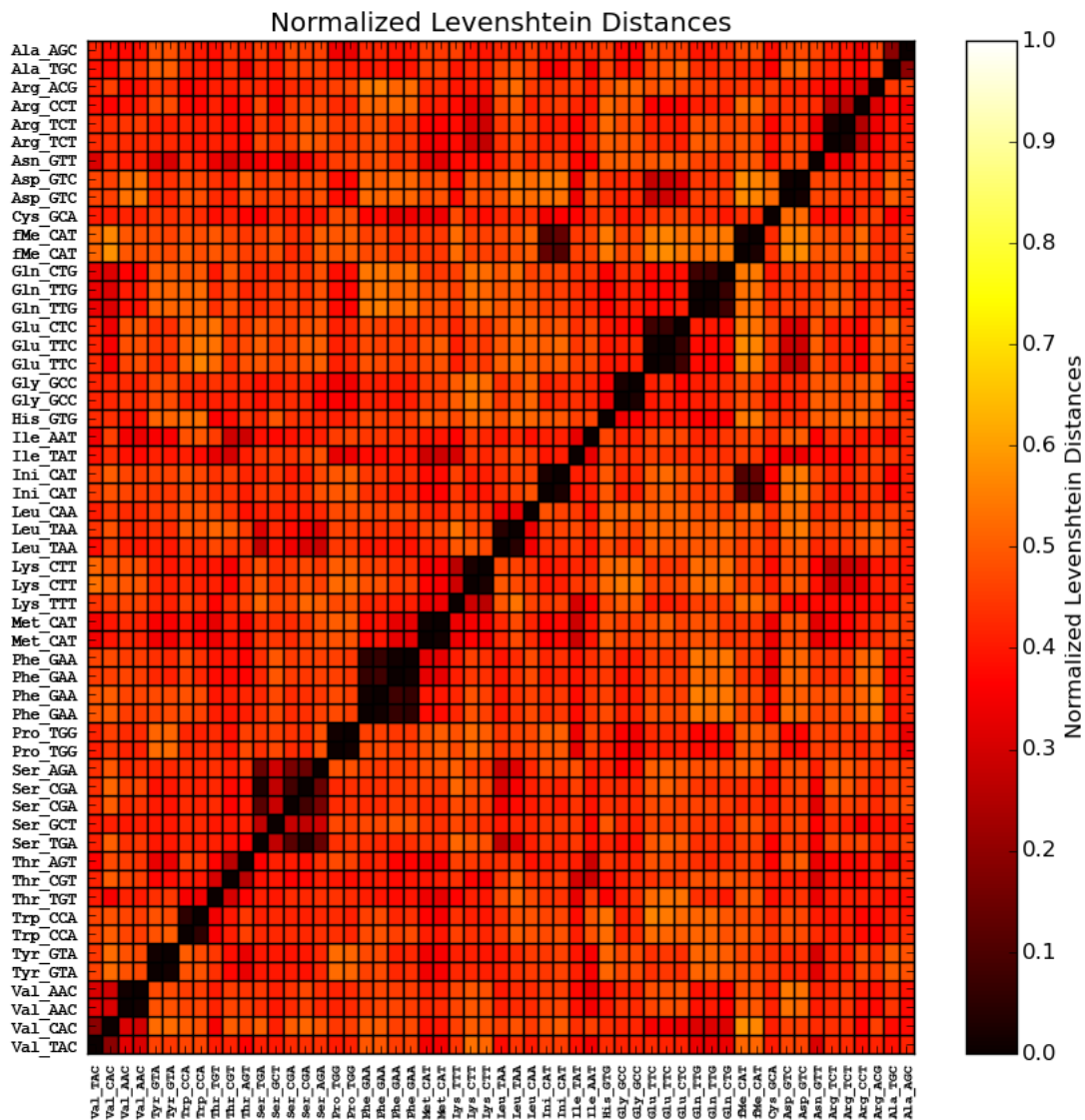


Fig. S2: Pairwise Levenshtein edit distances of cytosolic tRNA sequences of *S. cerevisiae*. Normalization was done by division of each distance value by the length of the longer of two compared sequences.

5 RESULTS AND DISCUSSION

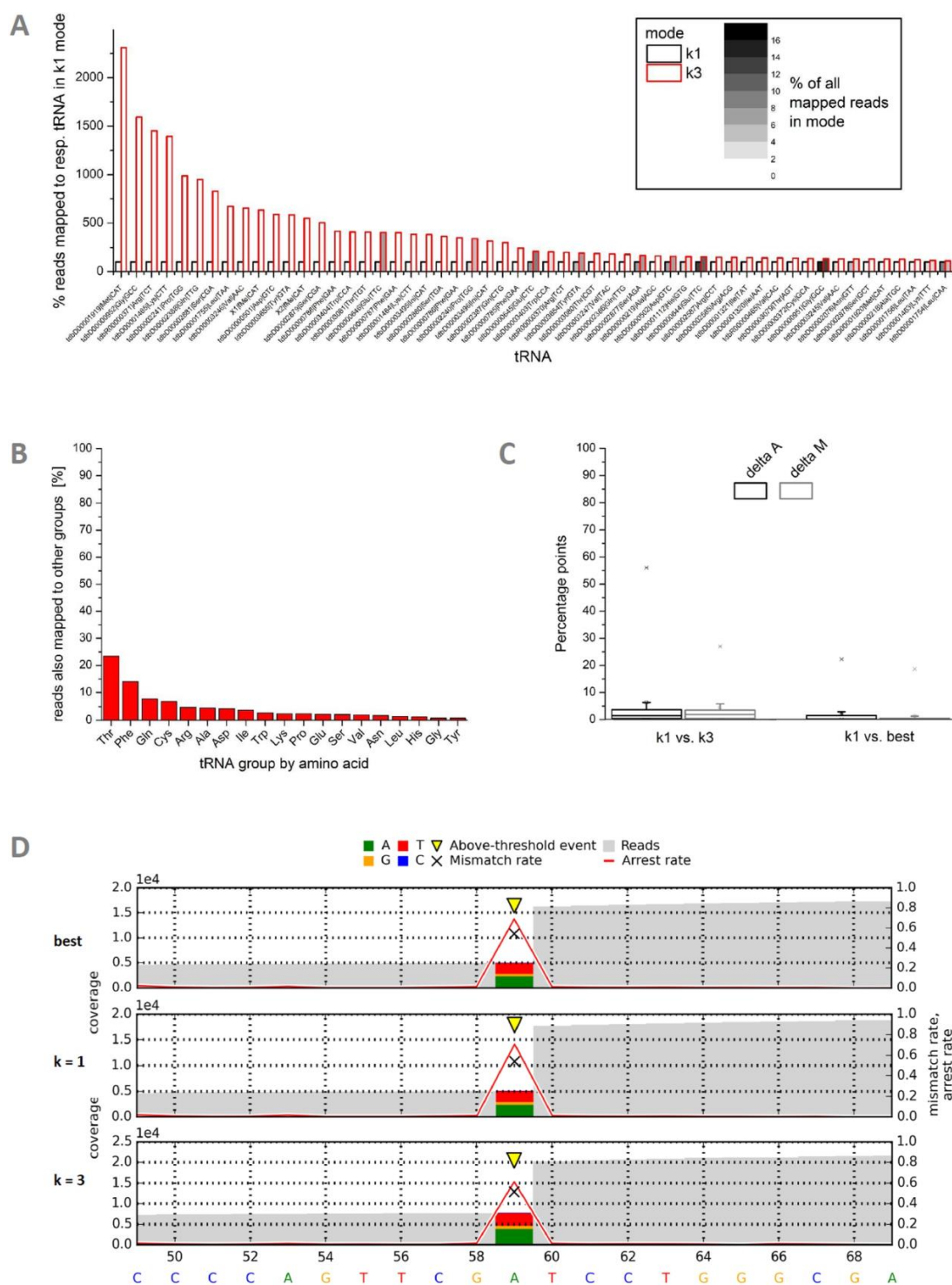


Fig. S3: RT signature vs multiple mappings. A) Relative read count comparison of $k = 1$ (“report best only”, set as 100 % for each tRNA) and $k = 3$ (“report best three”) modes for valid alignments by Bowtie2. B) Isotype confusion behavior for $k = 3$. The red bars indicate the relative amount of mapped reads per tRNA also mapped to tRNA(s) of a different acceptor group. C) Distribution of absolute difference in arrest and mismatch rates: $k = 1$ vs. $k = 3$ and $k = 1$ vs. “best”. D) Exemplary comparison of m^1A_{58} (sequence position 59) RT signatures for reporting modes $k = 1$, $k = 3$ and “best” (default) in yeast’s cytosolic $tRNA^{Val}_{AAC}$.

5.1 Detection of m^1A by NGS

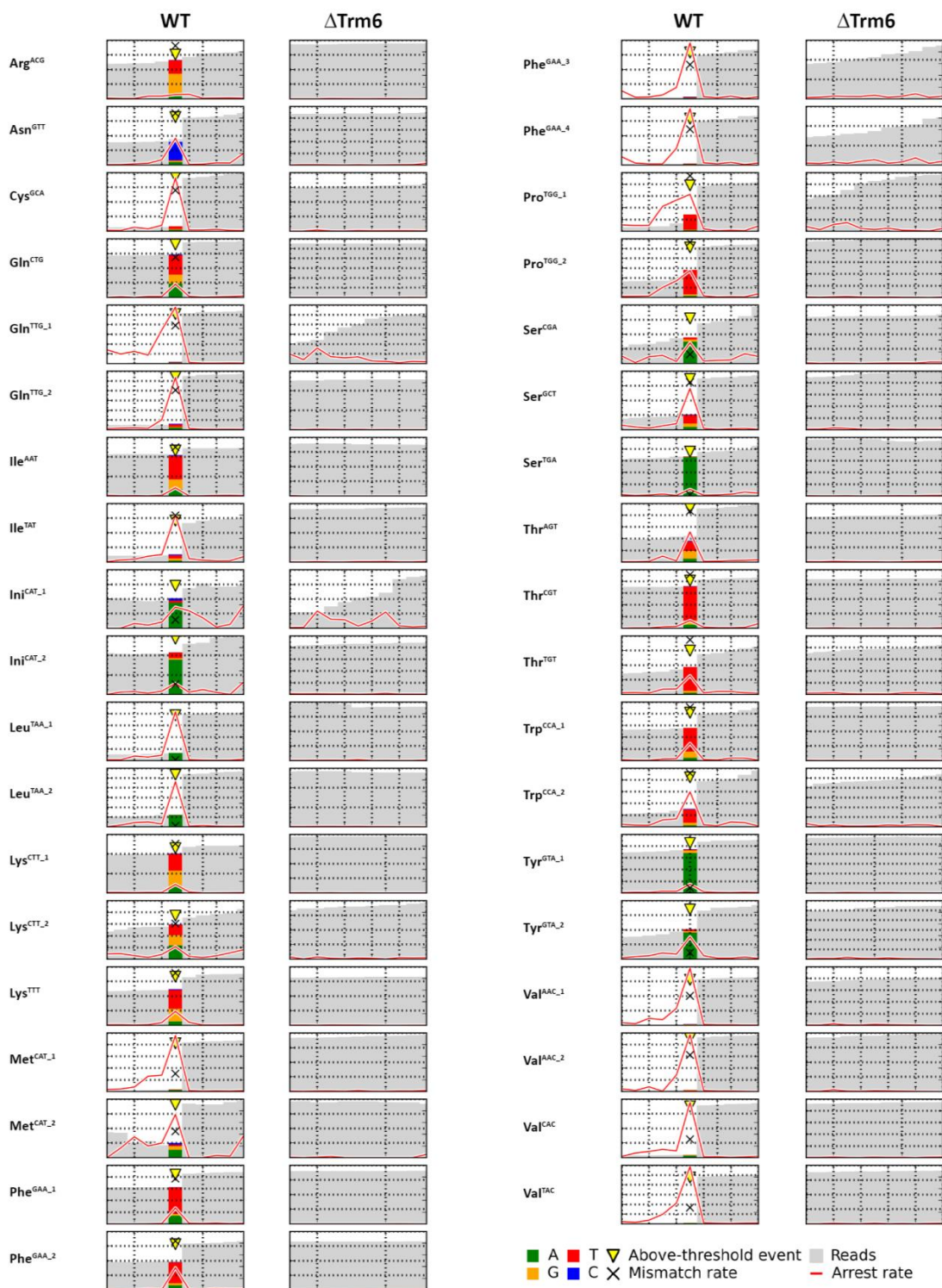


Fig. S4: m^1A_{58} signature compilation from 37 cytosolic tRNAs of wildtype and m^1A -negative ($\Delta Trm6$) *S. cerevisiae*. Plot scope: 5 bp upstream and 5 bp downstream of m^1A . Arrest and mismatch rates range on a [0, 1] scale as in the main article.

5 RESULTS AND DISCUSSION

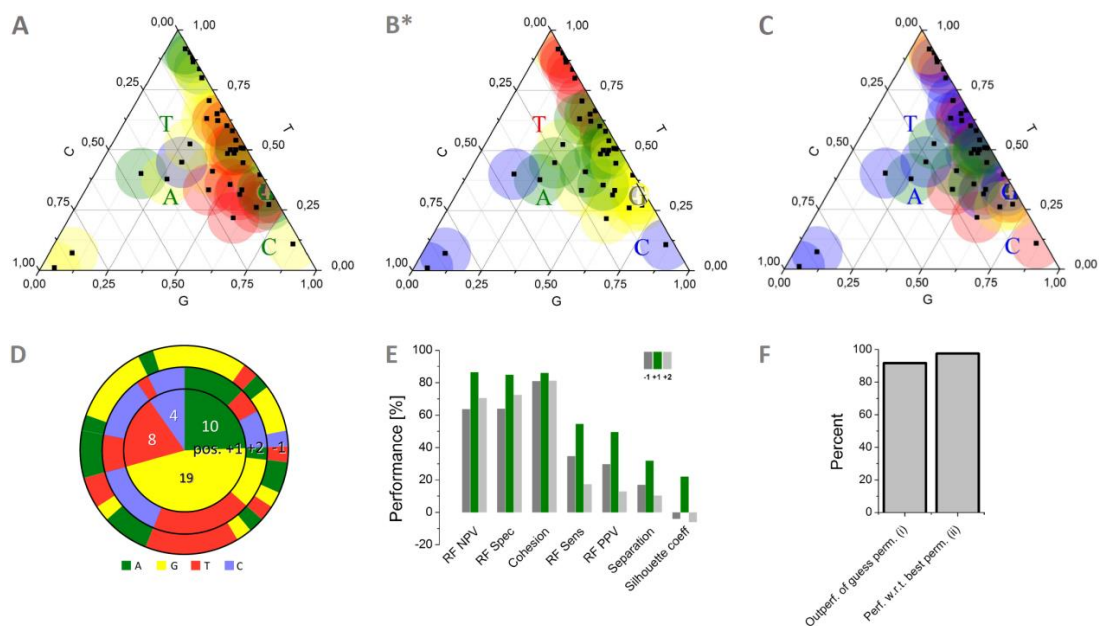


Figure S5: RT sequence context. Mismatch composition at m¹A site by nucleotide configuration at -1 (A), +1 (B*, identical with Figure 4 A from main document) and +2 (C). Data points from revolver oligonucleotides are specified by base at +1. D) Observed combinations of base configurations at positions +1, +2 and -1 relative to m¹A. E) Positional comparison by clustering measures cohesion, separation & silhouette coefficient (1) as well as Random Forest prediction performance in ten repetitions of seven-fold stratified cross-validation. Means for negative and positive predictive values (NPV & PPV), sensitivity and specificity indicate the model's performance predicting the base configuration at position -1, +1 or +2 from the m¹A site's mismatch composition. F) Accordance of revolver assay with m¹A pool. Knowing that the mismatch compositions of the synthetic instances correspond to four distinct populations of the global pool, 22 of 23 alternative permutations are outperformed (i) by the actual assignment, based on the mean of the four corresponding distances to cluster centers (MDC). The correct assignment ranks at 97.6 % of the mean MDC of the best-performing permutation (ii).

Method S1: Computational analysis of sequence context mediated mismatch composition

Evaluation of determinism regarding mismatch compositions driven by neighboring bases was done using both, descriptive and inferential statistics. Fig. S1, E shows seven measures, comparing positions -1, +1 and +2 in what is generalized as *performance* in RT context-mediated misincorporation pattern determinism. Separation describes the average inter-cluster center distance, while cohesion is the inverse of average intra-cluster deviation from the corresponding center. Edit distances were calculated for clusters with more than one data points only. They are defined as the alteration of the two least-diverging mismatch types $mism_k$ of an arbitrary pair of misincorporation patterns (i,j) required to transform one data point into the other, equivalent to:

$$d_{i,j} := 1 - \max_k |m_{i,k} - m_{j,k}|, \text{ where } m_{i,k} \in \{mism_G, mism_T, mism_C\}$$

In case of cohesion, performance scale normalization was performed constituting the weakest performance as the maximum possible average deviation of data points from cluster means, amounting to 2/3 in the hypothetic worst case of uniform distribution of a cluster's data points to the three corners of the ternary plot. Mean calculation from the four cohesion values was done using weights corresponding to the number of data points in each cluster. The maximum theoretical average inter-cluster center distance, used for normalization of the separation parameter, was set to 100 percentage points, corresponding to three perfectly condensed clusters of contrary misincorporation characteristics. Determinism analysis was supplemented by performance assessment of a Random Forest classifier ((2), 500 trees) inferring the base type of a corresponding neighbor position from the mismatch composition. In a four-fold stratified cross validation, we obtained sensitivities of (34.7, 54.5 and 17.3), specificities of (64, 84.8, 72.4) as well as positive and negative predictive values of (29.8, 49.5, 12.9) resp. (63.6, 86.5, 70.6) for the model in average for positions -1, +1 and +2. Fig S1, F shows how the measurements from the revolver oligonucleotides fit the clusters of the remaining data points. Distances from revolver oligonucleotide data points to cluster centers were normalized by average *intra* cluster distances, accounting for variance. Additional normalization with cluster member counts corrected for *a priori* class likelihoods.

5 RESULTS AND DISCUSSION

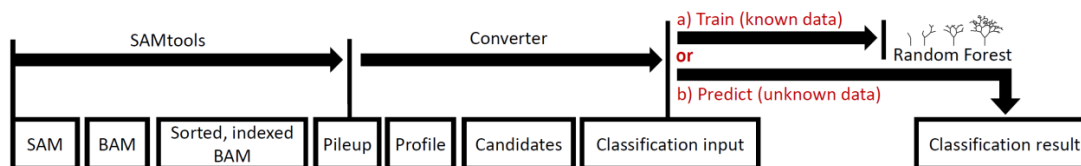


Fig. S6: Workflow - Data processing for prediction.

Table S3: Profile format.

ref_seg	pos	refbase	cov	matches	A	G	T	C	arrest	mismatch
tdbD00003245 Sacch...cer... 4932 Val AAC	59	A	7793	3822	3822	858	3057	56	0.612	0.515

Table S4: Classification input format. CSA denotes context sensitive arrest rate as defined in methods section.

arrest	mismatch	mismatch per arrest	CSA	G mism	T mism	C mism	mod. type
0.76	0.569	0.749	8	0.29	0.448	0.262	'm1A'

Table S5: Random Forest performance - 10 rep. 5-fold cross-validation. SD^R is the absolute standard deviation of a mean value calculated from 10 runs. SD^F is the mean SD of 10x5 = 50 foldwise outcomes for the corresponding performance measure.

performance for class m ¹ A (avg. of 10 runs)	low resemblance	high resemblance
sensitivity [%] (+/- SD) ^R (+/- SD) ^F	96.2 (+/- 1.0) ^R (+/- 6.2) ^F	88.9 (+/- 1.4) ^R (+/- 9.1) ^F
specificity [%] (+/- SD) ^R (+/- SD) ^F	96.9 (+/- 2.0) ^R (+/- 4.1) ^F	87.0 (+/- 2.8) ^R (+/- 10.5) ^F
positive predictive value (PPV) [%] (+/- SD) ^R (+/- SD) ^F	97.1 (+/- 1.8) ^R (+/- 3.8) ^F	87.4 (+/- 2.4) ^R (+/- 8.9) ^F
negative predictive value (NPV) [%] (+/- SD) ^R (+/- SD) ^F	96.6 (+/- 0.9) ^R (+/- 5.5) ^F	89.4 (+/- 1.1) ^R (+/- 8.1) ^F

Table S6: Random Forest performance - 10 rep. leave-one-m¹A-out cross-validation

performance for class m ¹ A (avg. of 10 runs)	low resemblance	high resemblance
sensitivity [%] +/- SD	96.0 +/- 0.9	89.3 +/- 1.9
specificity [%] +/- SD	98.0 +/- 1.8	86.7 +/- 3.8
positive predictive value (PPV) [%] +/- SD	95.1 +/- 1.4	82.8 +/- 3.2
negative predictive value (NPV) [%] +/- SD	96.1 +/- 2.0	81.4 +/- 4.2

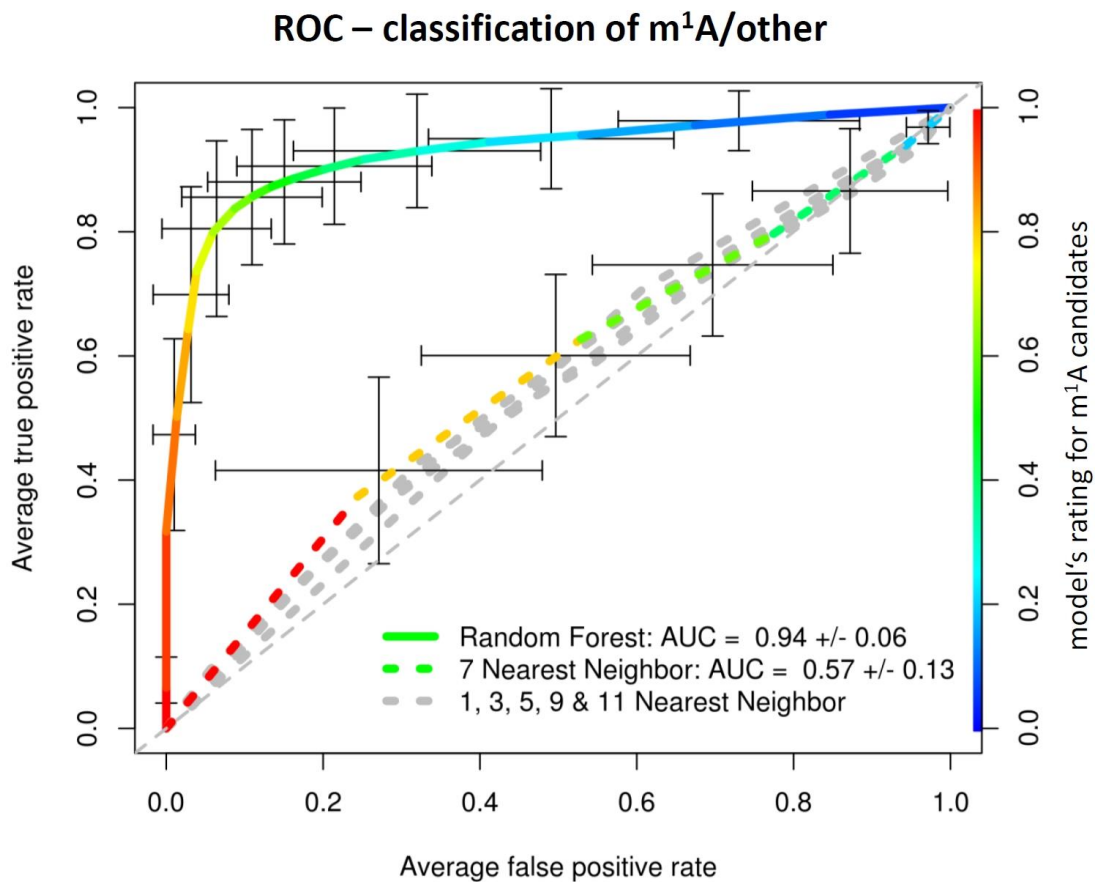


Fig. S7: Receiver operating characteristic (ROC) plot (3) showing the areas under curve (AUC) for Random Forest and k Nearest Neighbor (NN) supervised prediction of m¹A vs. other sites. The curves are averaged from 10 repetitions of a 5-fold cross validation. Error bars show the standard deviations of the ROC curve at the models' rating scores attributed to the m¹A candidates. The grey diagonal corresponds to the performance of a guessing classifier.

5 RESULTS AND DISCUSSION

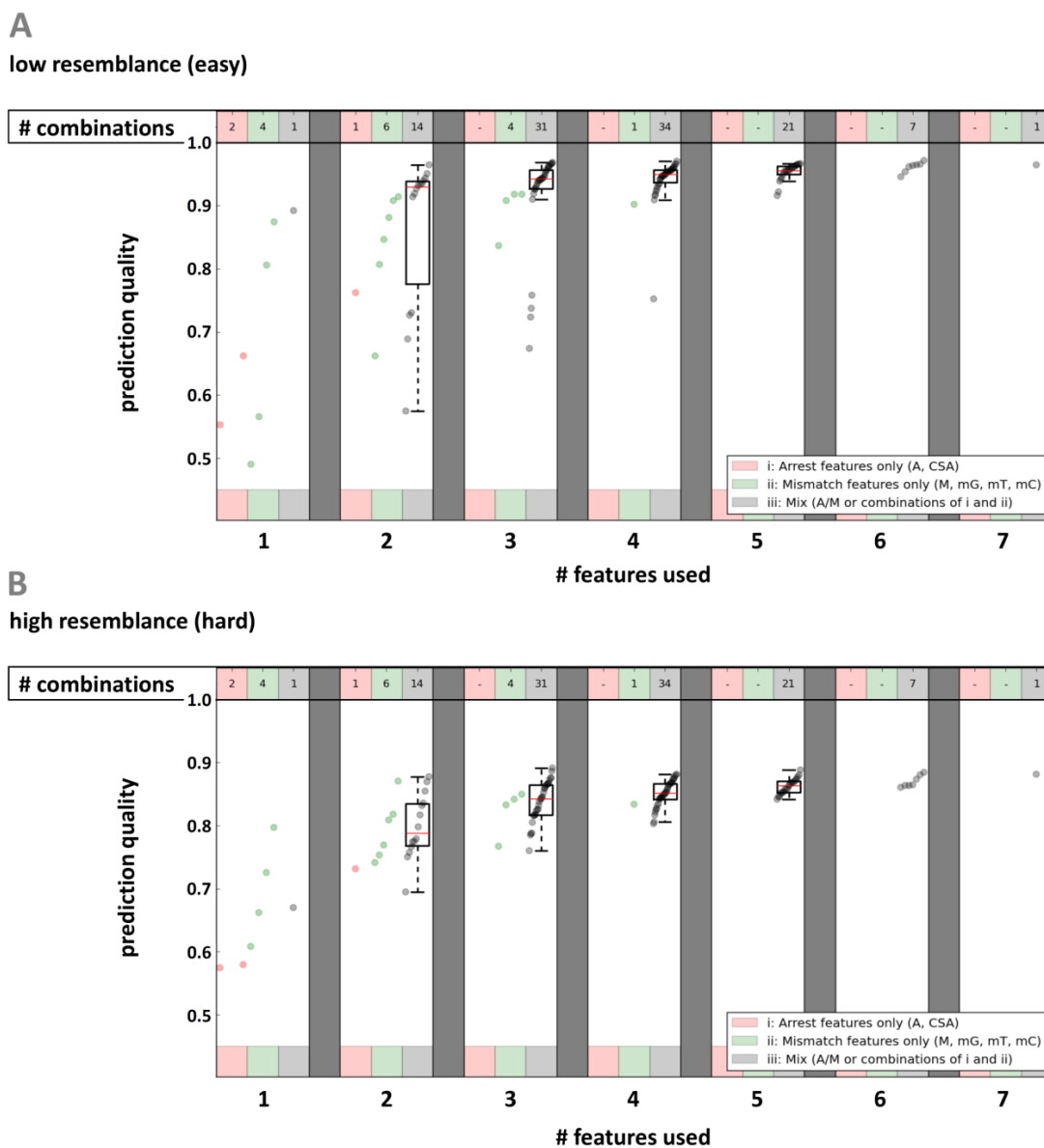


Fig. S8: Prediction quality vs. number and category of predictors. Quality is given as mean of sensitivity, specificity, positive and negative predictive value. A) Low and B) high mean resemblance of non-m¹As to m¹As as specified in subsection ‘Supervised prediction of m¹A by machine learning’ results. 2⁷-1=127 combinations of 7 used or omitted features were tested for classification by Random Forest. The number of data points in each column corresponds to the possible combinations for the categories i)-iii) using the respective feature count 1-7.

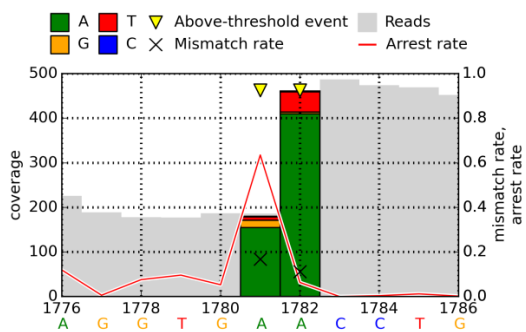


Fig. S9: RT signatures of m^6As 1781 and 1782 in yeast 18S rRNA. For a position p , the arrest rate reflects the relative amount of mapped reads ending at $p+1$, i.e. not covering p .

5 RESULTS AND DISCUSSION

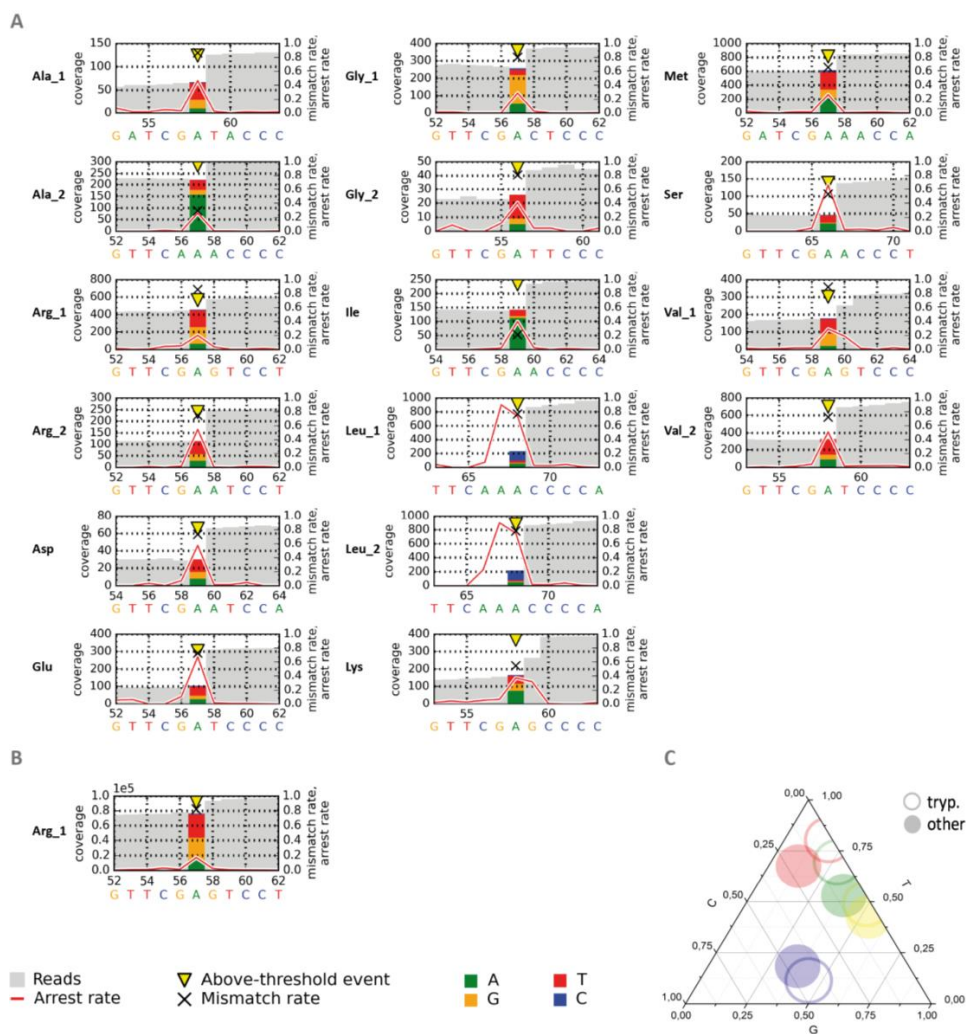


Fig. S10: Trypanosomal m¹A₅₈. A) 16 single tRNA profiles from total tRNA preparation from *Trypanosoma*. B) Sequencing profile of a purified sample of tRNA^{Arg-UCG}. C) Mismatch composition by base configuration at pos +1. The data points taken from Figure S10A were treated in the same way as described for yeast and averaged, then visualized as open circles. For comparison, yeast data averaged from Figure 3B are plotted in closed circles.

Table S7: LC-MS/MS quantification of m¹A.

ID	Sample	Material	Interest	m ¹ A/molecule	% m ¹ A/A
3	<i>S. cer.</i> Δ trm6	total tRNA	m ¹ A ₅₈ knockout		0.04
4	<i>S. cer.</i> wt	total tRNA	positive control		3.89
5	<i>S. cer.</i> 25S wt	rRNA	wildtype	1.50	
6	<i>S. cer.</i> 25S Δrrp8	rRNA	single knockout m ¹ A ₆₄₅	0.73	
7	<i>S. cer.</i> 25S Δbmt2	rRNA	single knockout m ¹ A ₂₁₄₂	0.77	
8	<i>S. cer.</i> 25S Δrrp8 + Δbmt2	rRNA	double knockout m ¹ A ₆₄₅ and m ¹ A ₂₁₄₂	0.01	
9	<i>S. pactum</i>	total RNA	m ¹ A on SSU of rRNA		0.15
10	<i>H. sapiens</i>	rRNA	Homologous identification		0.15
11	<i>M. musculus</i>	rRNA	Homologous identification		0.09
12	revolver m ¹ A-G	synthetic oligo.	RT sequence context dependency	0.77	
13	revolver m ¹ A-C	synthetic oligo.	RT sequence context dependency	0.82	
14	revolver m ¹ A-U	synthetic oligo.	RT sequence context dependency	0.92	
15	revolver m ¹ A-A	synthetic oligo.	RT sequence context dependency	0.79	
17	A-G	<i>in vitro</i> transcr.	negative control	0.00	
S24	Signature vs. occupancy	tRNA ^{Arg} _{UCG}	Novel site	1.01	

References

1. Rousseeuw, P.J. (1987) Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Computational and Applied Mathematics*, **20**, 53-65.
2. Liaw, A., & Wiener, M. (2002) Classification and regression by randomForest. *R news*, **2**, 18-22.
3. Sing, T., Sander, O., Beerenwinkel, N. and Lengauer, T. (2005) ROCr: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940-3941.

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

Methylation on the Hoogsteen edge of pyrimidine nucleosides is known to have little or no effect on reverse transcriptases [95]. Therefore, detection of 5-methyluridine (m^5U) and 5-methylcytidine (m^5C), based on a next generation sequencing method, requires a preceding chemical labeling step. That step would ideally alter the modifications, and thus produce a specific reverse transcription signature. Since there were already published successful experiments on labeling those nucleotides in DNA with osmium tetroxide - bipyridine (os-bipy), [71, 154, 66] a focus was set to evaluate this agent with RNA.

Experiments performed on a nucleoside level revealed that Os-bipy was about ten time more selective for m^5U than its non-methylated equivalent uridine and about 5 times more selective over m^5C . 5-methylcytidine reacted about 5 times faster than cytidine.

Additionally, upon reaction of osmium tetroxide-bipyridine with the double bond of 5-methyluridine, two formed diastereomers were chromatographically separated and with the help of 1- and 2-dimensional 1H -NMR characterized and determined. It was shown that in analogy to DNA, the ribose attached to the C-1 of the nucleobase induces the same favored *si* attack of the Os-bipy label toward the 5,6 double bond of the pyrimidine leading to the formation of a preferred product with *5R,6S* configuration. Of note, all products of Os-bipy adducts were confirmed by high resolution mass spectrometry attached to an HPLC device.


Because of the high preference of Os-bipy toward 5-methyl pyrimidines on a nucleoside level, further experiments were performed to investigate whether this preference was preserved in short oligonucleotide chains. For this, two 5-mer RNAs were designed that featured either 5-methylcytidine or 5-methyluridine at the central position, flanked by two unreactive purines and terminated with pyrimidines at the extremities. Surprisingly, when comparing m^5C incorporated in the oligonucleotide and as a nucleoside, its reactivity was reduced by a factor of 4, which made it indistinguishable from the unmodified pyrimidine nucleosides. To some contrast, the reactivity toward m^5U was only reduced by half, and since the reactivity for uridine dropped by roughly the same factor, selectivity remained high (roughly 8 fold). Further investigations on the formation of corresponding diastereomers offered an explanation for the observed reduced reactivity. On nucleoside level, the preferred osmylation product for 5-methylpyrimidines resulted from a *si* attack of the Os-bipy label. This site was shielded by the RNA chain and was significantly slowed down. In the case of m^5U the reduction was about 3.5 fold, for m^5C it was about 12 fold. This led to a change in the formation ratio of osmylation from roughly 80:20 (*si/re*) on a nucleoside level, to about 40:60 (*si/re*) in the oligonucleotides.

Furthermore, accessibility of uridine in a 5-mer oligonucleotide chain containing four unreactive adenosines was investigated. The effects of sterical shielding and following change in the formation ratio of osmylation reaction were not as pronounced as was the case with 5-methylpyrimidines. Nevertheless, a tendency was observed, that uridines located at either 3'- or 5' extremity lead to a formation ratio similar to that of a nucleoside, whereas approaching the middle of the chain, the ratio was nearly equalized. These results imply that osmium tetroxide-bipyridine is extremely sensitive toward the chemical environment of pyrimidines within an RNA chain, and can be further exploited in e.g. structure probing experiments. The results of this study were recently published [155].

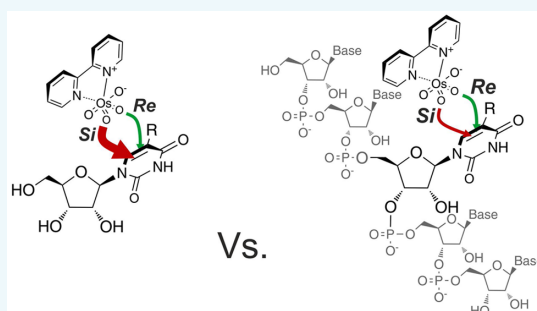
Diastereoselectivity of 5-Methyluridine Osmylation Is Inverted inside an RNA Chain

Lyudmil Tserovski and Mark Helm*

Institute of Pharmacy and Biochemistry, University of Mainz, D-55128 Mainz, Germany

 Supporting Information

ABSTRACT: In this study, we investigated the reaction of the osmium tetroxide–bipyridine complex with pyrimidines in RNA. This reagent, which reacts with the diastereotopic 5–6 double bond, thus leading to the formation of two diastereomers, was used in the past to label thymidine and 5-methylcytosine in DNA. In light of the growing interest in post-transcriptional RNA modifications, we addressed the question of whether this reagent could be used for labeling of the naturally occurring RNA modifications 5-methylcytosine and 5-methyluridine. On nucleoside level, 5-methylcytosine and 5-methyluridine revealed a 5- and 12-fold preference, respectively, over their nonmethylated equivalents. Performing the reaction on an RNA level, we could show that the steric environment of a pentanucleotide has a major detrimental impact on the reaction rate of osmylation. Interestingly, this drop in reactivity was due to a dramatic change in diastereoselectivity, which in turn resulted from impediment of the preferred attack via the *si* side. Thus, while on the nucleoside level, the absolute configuration of the major product of osmylation of 5-methyluridine was (*SR,6S*)-5-methyluridine glycol-dioxoosmium-bipyridine, reaction with an RNA pentanucleotide afforded the corresponding (*SS,6R*)-diastereomer as the major product. The change in diastereoselectivity lead to an almost complete loss of selectivity toward 5-methylcytosine in a pentanucleotide context, while 5-methyluridine remained about 8 times more reactive than the canonical pyrimidines. On the basis of these findings, we evaluate the usefulness of osmium tetroxide–bipyridine as a potential label for the 5-methyluridine modification in transcriptome-wide studies.



INTRODUCTION

Interest in the field of RNA modifications has recently surged as a consequence of technical breakthroughs in detection and analytics, which led to the discovery of a large number of new modification sites on a transcriptome-wide scale.^{1–4} This development concerns numerous chemically distinct RNA modifications, and their newly discovered distribution and functional impact is of considerable importance in the regulation of gene expression, to an extent that has recently coined the term “epitranscriptome”,^{5–13} which is meant to include all post-transcriptional RNA alterations that affect the epigenetic state of a cell. This development is largely driven by the appearance of antibodies against various modifications such as *N*⁶-methyladenosine,⁴ *N*¹-methyladenosine,¹ and 5-methylcytosine.¹⁴ The detection of the latter modification transcriptome-wide was spearheaded by the adaptation of bisulfite sequencing from DNA to RNA.⁵ This method results in the conversion of unmodified cytosines to uridines, whereas leaving 5-methylcytosines intact allows their distinction upon sequencing. Bisulfite sequencing has maintained the status of gold standard in the field of DNA methylation (more specifically, the detection of 5-methyldeoxycytosine (5dmC)). However, despite the applicability of bisulfite sequencing and antibodies against 5dmC, as well as other detection methods such as

AzaIP,¹⁵ detection of the corresponding modification in RNA, abbreviated m⁵C, is still controversial,^{11,15,16} and new detection methods are needed. Of note is the fact that no sequencing-based method for the detection of the related pyrimidine modification 5-methyluridine (m⁵U) has been described to date.

While quantification of m⁵C or m⁵U by liquid chromatography–mass spectrometry (LC–MS) techniques is routinely conducted with high sensitivity, this method incurs hydrolysis of the RNA sample to mononucleosides and thus a complete loss of sequence information.¹⁷ Alternatively, the modifications can be detected by mass spectrometry in oligonucleotides obtained by diverse RNases,¹⁸ although a drawback of this method is the requirement for large amounts of material and limited sequence information near the modified site. Methods that supply sequence information are typically based on an enzymatic amplification step using a polymerase. Recently, it was shown that products of next-generation sequencing contain information about RNA modifications that were shown to be useful in determination of the underlying modification.¹⁹

Received: July 20, 2016

Revised: August 19, 2016

Published: August 19, 2016

5 RESULTS AND DISCUSSION

Bioconjugate Chemistry

Article

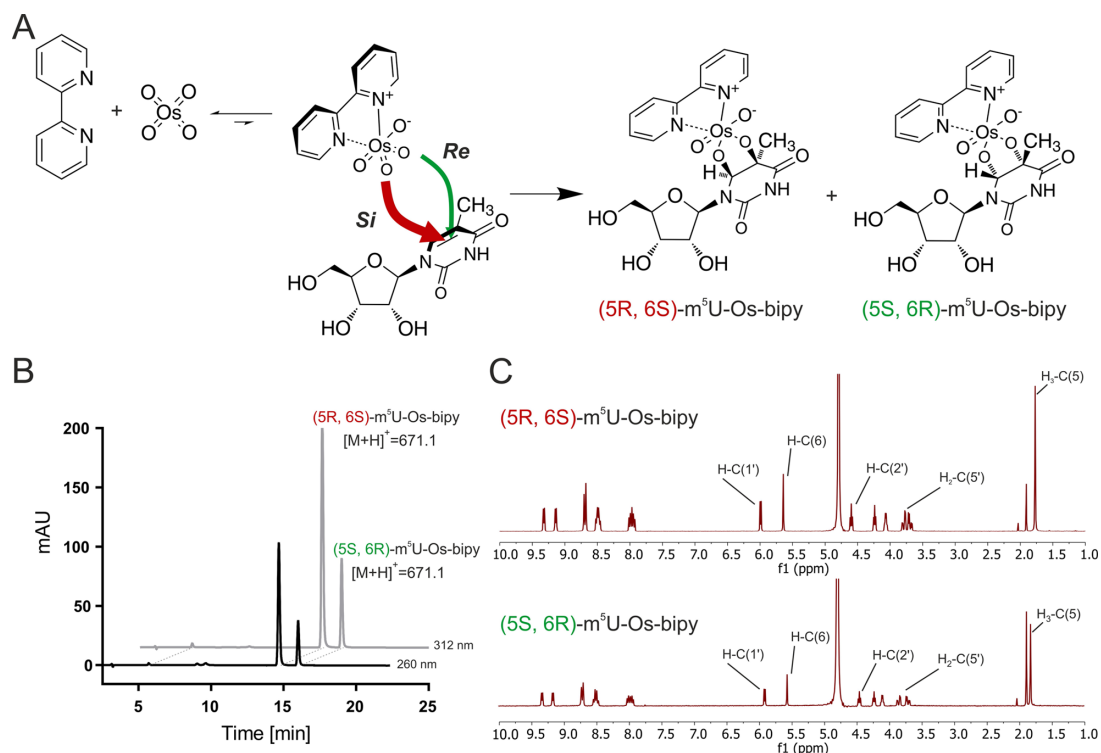


Figure 1. General reaction scheme of osmylation reaction toward m^5U and identification of products by LC–MS and 1H NMR. (A) Formation of a OsO_4 -bipy complex by reaction of osmium tetroxide with 2,2'-bipyridine followed by the reaction with 5-methyluridine and the formation of two stereoselective products. (B) Detection and separation of two products by HPLC. Detection of consumed substrate m^5U at 260 nm (black lane) and detection of products at 312 nm (gray lane). (C) 1H NMR of chromatographically separated products. Signals at 1.89 and 2.01 correspond to mobile-phase acetonitrile and acetate.

However, the 5-methylgroup in m^5C and its sibling species, m^5U , are situated on the Hoogsteen edge, and substituents at this position are known to have very little effect on polymerase activity.¹⁶ Indeed, 5-methylpyrimidines are recognized, like their canonical unmodified equivalents, by typical polymerases, the most obvious example being the occurrence and replication of the m^5U analog thymidine in DNA. Thus, polymerase erases all information on 5-methylpyrimidine in nucleic acid templates unless, as in the case of bisulfite sequencing, information is added in the form of a selective chemical derivatization.³ This information may manifest itself by different means in the cDNA of a reverse-transcribed RNA template. However, applications of bisulfite sequencing to RNA suffer limitations from RNA degradation and background signals, especially for less-abundant RNAs. Therefore, and because of the complete lack of reagents for the detection of m^5U -containing sequences in RNA, there is a need for selective reagents for 5-methylpyrimidines. We here investigate the possibility of selective derivatization of 5-methylpyrimidines in RNA by agents known to yield the product of a *cis*-addition reaction.²⁰

A first indication for selectivity of such reagents is their use in DNA footprinting studies, conducted with permanganate that selectively reacted with thymidine and 5-methyl deoxycytosine but not with cytosine.^{21–23} Osmium tetroxide (OsO_4) has been used in similar applications in a pure form²⁴ as well as in the presence of nitrogen-containing ligands such as TEMED, pyridine, and 2, 2'-bipyridine (bipy).^{25–32} The latter forms a

well-characterized complex with OsO_4 , which allows experiments to be conducted under very mild conditions.³³

Because of its stability, this complex was already applied as a covalent label for the detection of single- and double stranded DNA and RNA in the lower nanomolar range by means of electrochemical hybridization.^{34–38} Furthermore, Palecek et al. have worked on the osmium tetroxide–bipyridine label for footprint applications in DNA.²⁷ More recently, the concept has been applied to selective labeling of thymidine for nanopore-based DNA Base Discrimination by Kanavarioti et al.^{31,39} This work included the development of derivatization conditions for labeling of either all pyrimidines or for selective labeling of thymidines only. Clear demonstration that osmium reagents can discriminate 5-methylcytidines against cytidines in DNA came from Okamoto, who generated OsO_4 in situ from potassium osmate(VI) and potassium hexacyanoferrate(III).^{40,41} His work included the thorough stereochemical characterization of the 5-methyl-2'-deoxycytidine glycol-dioxosmium–bipyridine ternary complex obtained from the reaction of oligodeoxynucleotides, which revealed by X-ray crystallography the preferred reaction product (SR,6S) resulting from a *si* attack of the osmylation reagent.³³

So far, Okamoto's reagent has not been investigated in the context of 5-methylpyrimidine osmylation in RNA, possibly because the starting reagents osmate (VI) and 2, 2'-bipyridine were found to react with the vicinal hydroxylgroups of the ribose moiety.^{26,32} The latter reaction is not undergone by the

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

Bioconjugate Chemistry

Article

OsO₄-bipy combination.^{26,42} Detailed analysis of the kinetics of the labeling reaction of thymidine, cytosine, and uracil was performed by Reske et al.²⁵ Nevertheless, a comparison of the kinetics between the canonical pyrimidines cytosine and uridine and their 5-methyl equivalents still lacks. We therefore decided to investigate the selectivity of the OsO₄-bipy complex in osmylation of 5-methylpyrimidines in comparison to unmethylated pyrimidines in RNA.

RESULTS

Diastereoselectivity of *cis*-Osmylation of 5-Methyluridine. To determine if 5-methyl-ribopyrimidine nucleosides are preferentially attacked by a *si* attack in analogy to the reports by Okamoto,³³ 5-methyluridine (m⁵U) was reacted with osmium tetroxide (OsO₄) and 2,2'-bipyridine (bipy). Preliminary data showed that a reaction with OsO₄ alone, i.e., in the absence of bipy, was very slow (Figure S1). This confirmed literature data for the slow reactivity of the deoxy equivalent thymidine with OsO₄ alone.⁴³ Hence, a 10 mM solution of m⁵U was incubated in the presence of 10 mM bipy and 20 mM OsO₄. HPLC analysis indicated the reversible formation of an OsO₄-bipy complex in approximately 10%. The reaction was monitored by HPLC, and concomitant with the disappearance of the starting material, the formation of two product peaks at *t* = 14.7 min and *t* = 16.1 min upon separation on an RP-18 column was observed (Figure 1B). The masses of these two products, as determined by LC-MS (electrospray ionization (ESI)-ion trap) corresponded to that of the expected (5*R*,6*S*)- and (5*S*,6*R*)-5-methyluridine glycol-dioxoosmium-bipyridine compounds ([M + H]⁺, calc: 671.095; experimental: 671.1) (Figure 1B).

After 1 h, the disappearance of starting material indicated completion of the reaction, and at this point, the reaction mixture was treated with water-saturated corn oil to remove OsO₄ and a major fraction of the bipy. The remaining aqueous phase of the reaction mixture was lyophilized, redissolved in water, and subjected to separation on a semipreparative RP-18 column. Analysis by ¹H NMR of both products (Figures 1C, S2, and S4) revealed high similarities with the thymidine glycol-dioxoosmium-bipyridine complexes described for DNA.³³ To determine their absolute configuration, an in-depth NMR characterization was conducted including COSY and NOESY experiments, as shown in Figures S3 and S5. In agreement with the results found by Okamoto et al.³³ and Vaishnav et al.,⁴³ including similar lipophilicity and ratio of corresponding diastereomers, it was determined that the major product was of 5*R*,6*S* configuration and, thus, the result of a *si* attack by the OsO₄-bipy-complex. Consequently, the less-efficient *re* attack lead to the minor product of 5*S*,6*R* configuration. Some characteristic ¹H NMR signals that support this conclusion are compiled in Table S1. From these data, it was concluded that the major product of the reaction with the nucleoside m⁵U had a 5*R*,6*S* configuration and the minor had 5*S*,6*R*, respectively.

In a separate experiment, the reaction was conducted in an NMR tube, and the ratio of the two products was determined by using integrals of the well-resolved signals (Figure 1C) that had been attributed to H₃-C(5), H-C(6), H-C(1'), and H-C(2'), respectively. Quantification of these signals, as detailed in Table 1, was in excellent agreement with the product ratio measured by the 312 nm absorption of an HPLC run of the same reaction mixture. The relative quantification by NMR therefore allowed us to state that the extinction coefficients of

Table 1. Relevant ¹H NMR Signals Used for the Relative Quantification of Diastereomeric Products 5-Methyluridine Glycol-Dioxoosmium-Bipyridine^a

¹ H NMR(300 MHz, D ₂ O)	m ⁵ U complex (integral of signal)		ratio of P1 to P2
	product 1	product 2	
H ₃ -C(5)	1167.63	393.94	3.0
H-C(6)	390.43	131.49	3.0
H-C(1')	396.42	117.06	3.4
H-C(2')	381.09	133.75	2.8
mean:			3.0
HPLC [312 nm]	1938.4	654.24	3.0

^aComparison with the peak areas from HPLC were detected at 312 nm. Similar experiments were performed with further pyrimidine nucleosides (m³C, U, and C; see Tables S2-S4).

both diastereomers at 312 nm are comparable within the accuracy of these two detection methods and could thus be used for quantification in the determination of kinetic parameters by HPLC in what follows.

From the above flow, two conclusions relevant to the further course of this study were drawn. First, it was shown that the ribose moiety of 5-methyluridine induces the same preferred attack of osmylation reagent OsO₄-bipy toward the double-bond 5-6 of the nucleobase, thus delivering the same preferred diastereomer that has been described for both deoxynucleosides and oligodeoxynucleotides in DNA,³³ namely the (5*R*,6*S*)-5-methyluridine glycol-dioxoosmium-bipyridine complex. Second, with the combined analysis of NMR and HPLC, it could be shown that a relative quantification of two osmylated products for each reacted pyrimidine nucleoside was possible and could be used for an in-depth analysis of formation kinetics.

Selective Osmylation of 5-Methylpyrimidines. To evaluate selectivity of the OsO₄-bipy reagent in a potential RNA labeling reaction, various nucleosides were reacted under conditions that afforded pseudo-first-order kinetics, including 8 mM OsO₄, 4 mM bipy, 400 mM phosphate buffer pH 7, 7 M urea, and 50 μM nucleoside. Under similar conditions, purines did not undergo oxidation for several hours at room temperature, as evidenced by NMR (Figures S6 and S7). In contrast, m⁵U, 5-methyl ribo-cytidine (m³C), ribo-cytidine (C), and uridine (U) all underwent oxidation, as observed by both NMR and HPLC. To determine kinetic parameters, aliquots were drawn at various time points from reaction mixtures containing one of the pyrimidines and 50 μM adenosine. The latter was used as internal standard in the subsequent HPLC analysis, from which the remaining concentrations of pyrimidines were determined. Figure 2A shows decaying concentrations of the four pyrimidines, whose oxidation was essentially completed within 2 h. In a logarithmic plot, a linear dependence of concentration against time was evident, confirming pseudo-first-order kinetics (Figure S8) and allowing the extraction of an overall reaction rate constant *k* for each individual nucleoside. When the reaction was conducted with all four pyrimidines in the same reaction mixture, the kinetics as shown in Figure 2B were essentially unchanged compared to individual kinetics (Figure 2A), suggesting the absence of any interaction among the nucleoside species. Consequently, the resulting values for *k* were found to be similarly unchanged, as shown in Figure 2C. The latter figure also illustrates strikingly that the reaction of m⁵U was by far the fastest, about 5 times

5 RESULTS AND DISCUSSION

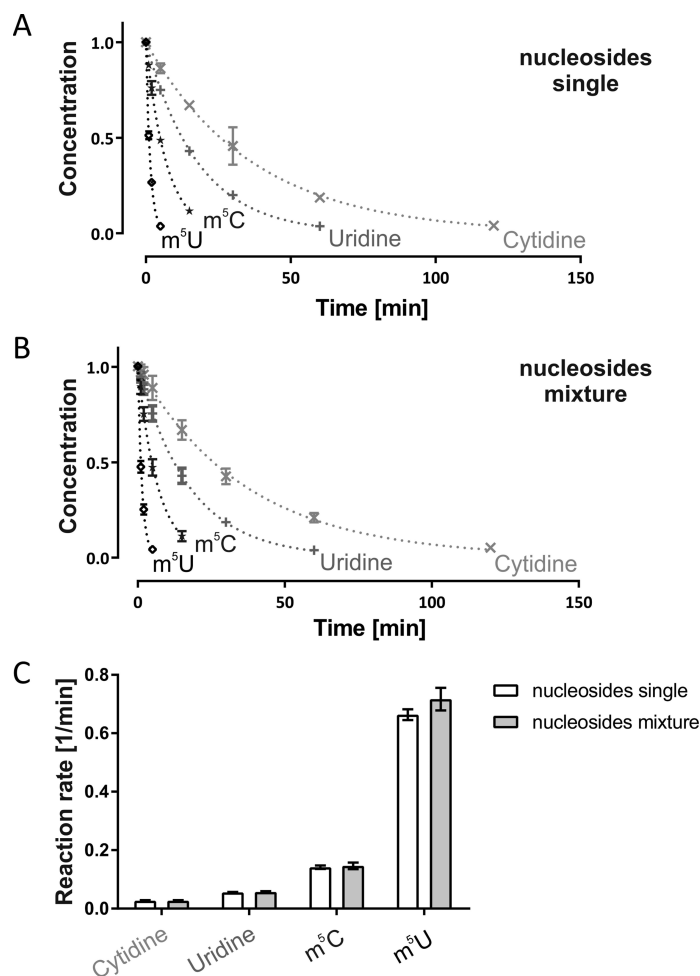


Figure 2. Reaction rates of pyrimidine nucleosides toward the OsO₄–bipy complex. Error bars indicate the standard deviation from three independent experiments. The dotted lines in (A) and (B) show the fit of a pseudo-first-order reaction type. (A) Normalized substrate decay in a time-dependent manner for single reacted pyrimidines. (B) Normalized substrate decay in a time-dependent manner for pyrimidines reacted in a mixture. (C) Comparison of overall reaction rates for pyrimidines reacted either as single nucleosides or in a mixture.

higher than that of the second fastest, namely m⁵C, and 12 times higher than the osmylation of uridine.

From these data that show significant selectivity in terms of different *k* values, we hypothesized that it should be possible to selectively label both m⁵C and m⁵U within oligoribonucleotides and thus distinguish them from the nonmethylated pyrimidines inside RNA sequences.

Osmylation of Uridines in Oligonucleotides. Because the presence of phosphate groups is one of the major differences when moving from nucleosides to oligonucleotides, osmylation kinetics of a pyrimidine mononucleotide under the conditions described above were investigated next. We chose uridine 5′-monophosphate as the more reactive of the two canonical pyrimidines occurring in RNA. Workup of the reaction prior to HPLC analysis now additionally included a treatment with alkaline phosphatase FastAP to remove the phosphate. To also characterize the difference between monomeric and oligomeric nucleotides, a series of 5-mers was investigated that were composed of four inert adenosine

residues and a single uridine at varying positions within the chain. After osmylation reaction under standard conditions and prior to HPLC analysis, RNA was enzymatically digested to nucleotides with a mixture of nuclease P1 and snake venom phosphodiesterase and then further dephosphorylated with FastAP, as described in ref 17. Again, normalized substrate concentrations were plotted against time, and pseudo-first-order reaction rate constants *k* were calculated for each uridine in RNA and UMP, respectively (Figure S9)

Comparison of the reaction rate of uridine nucleoside with that of uridine monophosphate (Figure 3) shows that the addition of a phosphate group to the 5′ of uridine alone slows the reaction slightly. Furthermore, inclusion in a short RNA chain of four adenines led to even more, now-considerable reduction of the reaction rate, which strongly depended on the relative position of the uridine within the 5-mer chain. The reaction was slowest in a 5-mer with a 3′-terminal uridine. As uridine was moved along the chain toward the 5′ end, the reaction rate increased correspondingly, achieving similar or

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

Bioconjugate Chemistry

Article

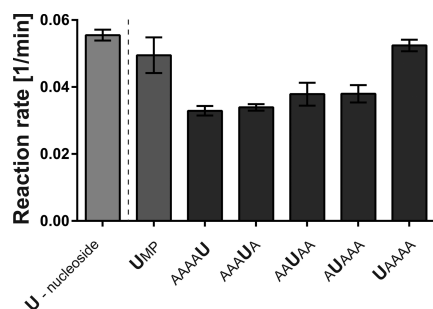


Figure 3. Comparison of the reaction rate of OsO_4 -bipy toward uridine in the context of a nucleoside or nucleotide or incorporated in a short RNA chain. Error bars indicate the standard deviation of the k value calculated by the pseudo-first-order fit of three independent experiments.

slightly higher reactivity for 5'-UAAAA-3' than for UMP. From this data, it was evident that the position of the pyrimidine in the chain played a role in its reactivity toward OsO_4 -bipy.

Given the structural similarity of the different 5-mers, the most likely explanation for differential reactivity was accessibility to the reagent as a consequence of differential steric encumbering. Because this, in turn, could be expected to affect the diastereoselectivity previously observed (Figure 1), we reinvestigated the experiments presented in Figure 2. Whereas previous reaction rates had taken account of the decay of pyrimidine starting material without consideration of the diastereomeric ratio of products, the latter was now quantified from HPLC traces, as shown in Figure 4A,B. From the overall reaction rate constant k , deduced from the decay of starting material and the ratio of products, the corresponding k_1 and k_2 values were calculated.

Figure 4B shows a comparison of characteristic HPLC traces monitoring the nucleoside glycol-dioxoosmium-bipy complex at 312 nm, obtained from reactions of uridine nucleoside, uridine monophosphate, and one of the 5-mer oligonucleotides featuring the uridine residue in a central position. Clearly visible, the diastereomeric ratio was strongly affected by the presence of a single phosphate and the oligonucleotide context, respectively. A systematic analysis of the respective reaction rates, as shown in Figure 4C, reveals pronounced effects, especially for uridine residues situated inside the RNA chain. Although uridines at the 3' and 5' extremities show a strong bias toward high k_1 values, in-chain uridines approach a near equalized ratio of k_1 and k_2 . This suggests that the overall reaction rate drops within the oligomer because the attack from the preferred side of the free nucleoside, reflected in the k_1 values, is more susceptible to steric hindrance caused by an extended RNA chain than is the attack from the opposing side. Of note is the fact that all reactions were performed in high urea concentrations to denature as much as possible the portions of RNA structure that rely on hydrogen bonds.

Osmylation of 5-Methylpyrimidines in Oligonucleotides. Taking into account, as noted above, (i) that the pyrimidines at the 5'-end of oligonucleotides appear to be especially exposed because they react faster than internal position and that (ii) 5-methylpyrimidine nucleosides show a 5–10-fold increased reaction rate compared to that of canonical nucleosides, we designed oligomers to assess the selectivity of osmylation of 5-methylpyrimidines in oligonucleotides. In two 5-mers (5'-CG[m⁵C]AU-3 and 5'-CG[m⁵U]AU-3'), the m⁵C

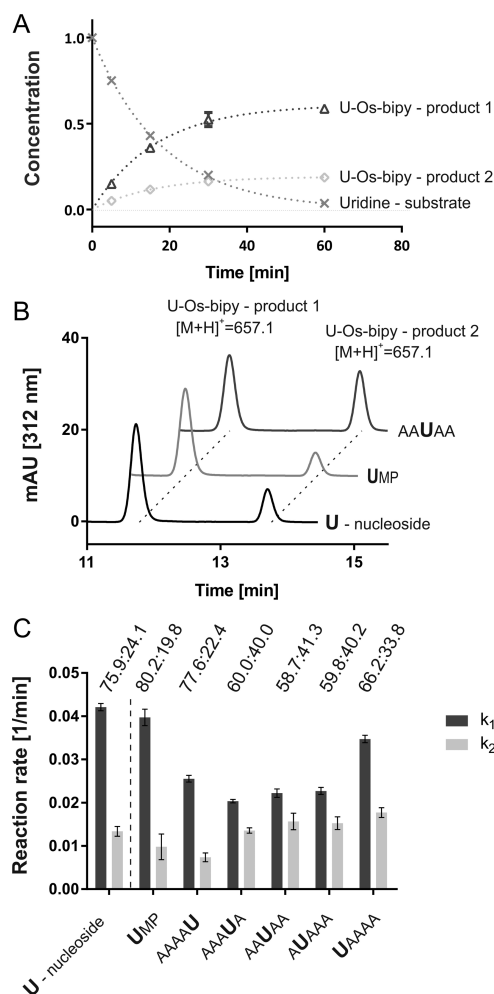


Figure 4. Reaction of OsO_4 -bipy complex with uridine and corresponding diastereomeric product formation. (A) Substrate decay in a time-dependent manner and simultaneous formation of products 1 and 2. The dotted lines show a fit of a parallel first-order type. The error bars indicate the standard deviation of three independent experiments. (B) LC-MS traces of selected uridine glycol-dioxoosmium-bipy complexes detected at 312 nm. Those chromatograms were taken after 30 min of reaction with OsO_4 -bipy. (C) Comparison of reaction rates k_1 and k_2 responsible for the formation of product 1 and product 2, respectively, for the reaction of OsO_4 -bipy with uridine in the context of a nucleoside, a nucleotide, and incorporated in a short RNA chain. The error bars indicate the standard deviation of calculated k values. The ratios over the bars indicate the ratio of formation kinetics, k_1/k_2 .

and m⁵U residues, respectively, were placed at the central position, flanked by two unreactive purines at the penultimate position of the chain, and terminated again with pyrimidines at the extremities. After these oligonucleotides were reacted under standard conditions, overall reaction rates for all pyrimidine nucleosides were determined from the decay of the respective peaks in HPLC chromatograms as described above (compare this with Figure 2). Figure 5B shows a comparison of the decay rates in the 5'-CG[m⁵C]AU-3 oligomer, which reveals little

5 RESULTS AND DISCUSSION

Bioconjugate Chemistry

Article

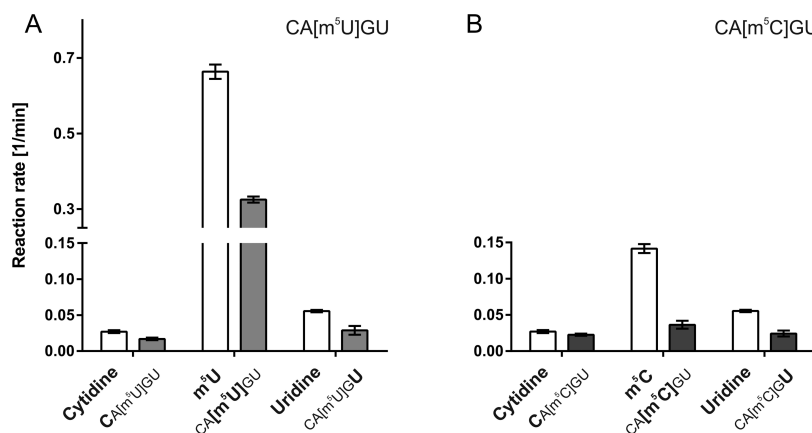


Figure 5. Comparison of reaction rates of pyrimidines in the context of nucleosides and incorporated in short RNA chains containing either m⁵U (A) or m⁵C (B). Error bars indicate the standard deviation of the calculated k value of three independent experiments.

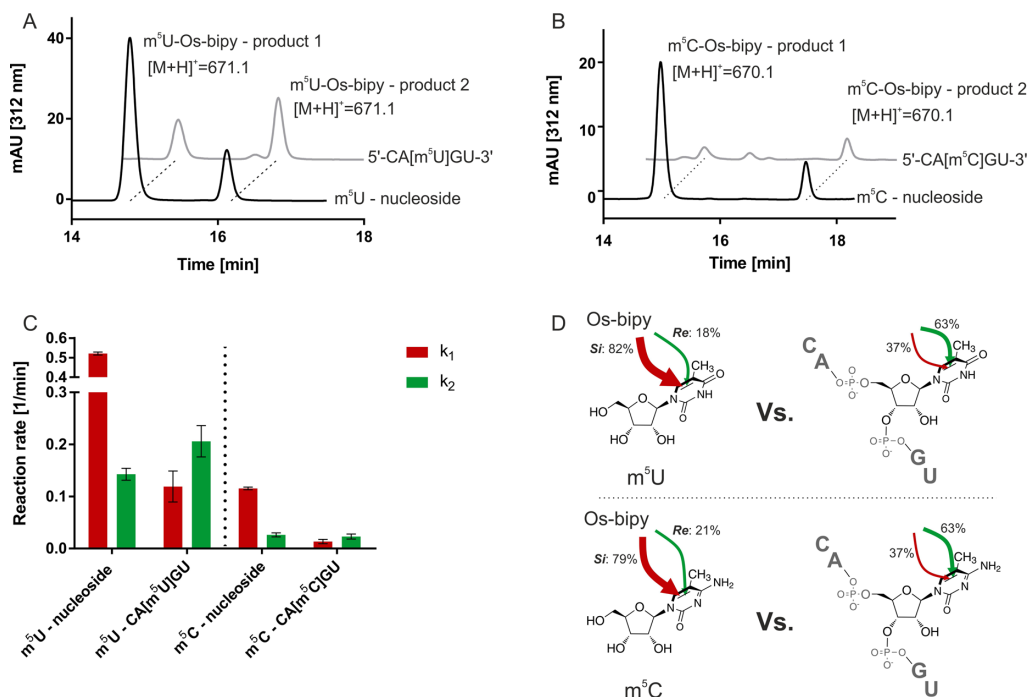


Figure 6. Reaction of OsO₄-bipy complex with methylated pyrimidines incorporated in a short RNA and detection of diastereomeric product formation. (A) LC-MS traces of 5-methyluridine glycol-dioxoosmium-bipy complex formation in the case of free nucleoside and incorporated in a short RNA chain detected at 312 nm. Both chromatograms were taken after 5 min of reaction with OsO₄-bipy. (B) LC-MS traces of 5-methylcytidine glycol-dioxoosmium-bipy complex formation in the case of free nucleoside and incorporated in a short RNA chain detected at 312 nm. Both chromatograms were taken after 5 min of reaction with OsO₄-bipy. (C) Comparison of formation rates k_1 and k_2 for the reaction of OsO₄-bipy with either 5-methyl uridine or 5-methyl cytidine in the context of nucleosides and incorporated in a short RNA chain. Error bars indicate the standard deviation of calculated k_1 and k_2 values of three independent experiments. (D) Change of stereoselective attack toward m⁵U and m⁵C, depending on whether methylated pyrimidine was used as a free nucleoside or incorporated in an RNA chain.

change for the cytidine reaction and a clearly visible decrease in the uridine reaction that is comparable to what was previously observed (Figure 3). Finally, a dramatic decrease of m⁵C reactivity inside the 5-mer was observed. In comparison to the free nucleoside, it dropped by a factor of 4 to a level that barely distinguishes it from the unmodified pyrimidines. In some

contrast, m⁵U in the analogous 5-mer suffered only a 2-fold reduction. Because the reaction rate of uridine dropped by roughly the same factor, the selectivity for m⁵U was still ~8 fold.

As before (Figure 4C), the relative formation rate of diastereomers in both oligonucleotides provided some details

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

Bioconjugate Chemistry

Article

on the intriguingly different behavior of m^5C versus m^5U in the pentanucleotide context. Panels A and B of Figure 6 show characteristic traces of 5-methyl-pyrimidine glycol–dioxosmium–bipy complexes monitored by HPLC. Clearly visible, the effect of steric shielding in the central position of the pentanucleotide was even more pronounced for methylated pyrimidines than for uridine. Most strikingly, formation of the product of a *si* attack on m^5U (designated as product 1 in Figure 6A), which was clearly predominant in reactions with free nucleoside, is slowed by about ~ 3.5 fold and is now the slower for both reactions, inverting the selectivity from 82:18 (*si/re*) to about 37:63 (Figure 6C,D). Even more striking is the slowdown of product 1 formation in the case of m^5C , as is clearly visible in the HPLC traces in Figure 6B. This ~ 12 -fold slowdown is the principal reason for the loss of osmylation selectivity of m^5C when moving from nucleosides to oligonucleotides. Figure 6C shows a direct comparison of rate constants k_1 and k_2 for each m^5C and m^5U , and an overview over all determined rate constants is given in the Supporting Information (Table S5).

DISCUSSION

The accessibility of functional groups in nucleobases in general is of interest for several reasons, including the aspects of structural probing in solution⁴⁴ and the tagging of nonstandard nucleotide modifications for either detection^{2,45,46} or isolation.⁴⁷ Here, we present a detailed study on the reactivity of the 5–6 double bond in pyrimidines toward an osmium tetroxide–bipyridine complex. A comparison of reaction rates in uridine, uridine-5'-monophosphate, and the 5'-UAAAA-3' pentamer strongly suggests that the presence of phosphates does not affect reactivity via electronic effects, leaving indeed steric accessibility as the governing factor. The present work contributes interesting insights into factors affecting accessibility and, as will be detailed toward the end of this discussion, provides a perspective for the selective tagging of naturally occurring post-transcriptional methylations of pyrimidines at the 5-position, as has been pioneered by Okamoto in the DNA field.⁴⁸

Configuration of Preferred Osmylation Product of 5-Methyluridine. The present work picks up an interesting observation of Okamoto's group in DNA, namely that the chiral deoxyribose moiety induces substantial diastereoselectivity in the osmylation of the planar nucleobase.³³ After HPLC separation of the diastereomers, the major adduct of SdmC had been deaminated to the corresponding deoxythymidine adduct and then characterized by NMR and crystallography. The latter allowed the determination of its absolute configuration to be 5*R*,6*S*,³³ which had resulted from a preferential *si* attack. The same configuration was also inferred by Vaishnav et al. after oxidation of deoxythymidine with OsO_4 and subsequent hydrolysis to the vicinal diol.⁴³ Our reenactment of osmylation with ribothymidine (m^5U) was carried out with slightly different reagents; while Okamoto's protocol generated an Os-(VIII) species by oxidation of osmate-(VI) with Fe(III)-hexacyanoferrate, we directly generated the osmium complex from OsO_4 and bipy. After the resulting diastereomers were separated by HPLC, investigation by two-dimensional nuclear magnetic resonance spectroscopy (2D-NMR) similarly lead us to conclude that a predominant *si* attack, which is superficially seen, an anticipated result.

RNA Structure Impedance of the *si* Attack of Osmium Tetroxide–Bipyridine toward Methylated Pyrimidines.

Surprisingly, however, we observed a change of the predominant attack from *si* to *re* when moving from free nucleoside to pentanucleotide oligomers (Figure 6). This is in stark contrast to the observation on SdmC in DNA, where this preference was conserved between the free nucleoside and an oligodeoxynucleotide.³³ A change of accessibility from 82:18 to 37:63 (*si/re* product) was observed for ribothymidine (m^5U), but a similar effect was also seen for 5-methyl ribocytidine, with which the ratio of 79:21 of the nucleoside changed to 37:63 in the pentanucleotide (Figure 6D). For uridine, the ratio changed from 75:25 to a range of 77:23 to 59:41, depending on the position in the pentanucleotide. The latter observation suggests an exquisite sensitivity of the osmylation reagent to the steric environment of the targeted double bond, making it potentially very useful for mapping accessibilities, e.g., as in structural probing experiments.⁴⁹ The current bottleneck here is the lack of position specific detection of osmylation in RNA oligonucleotides. Our efforts to induce strand scission at osmylated sites by aniline treatment alone, or a combination of borohydride reduction and aniline treatment failed. Also, the osmate–bipy adducts could not be mapped in reverse transcription experiments (not shown).

Potential for Enrichment of 5-Methyl Uridine Containing RNA. The higher reactivity of OsO_4 –bipy reagent toward m^5U compared to the other pyrimidine nucleosides in RNA offers the possibility for selective tagging and isolation of fragments containing this modification. A prerequisite for this would be, e.g., a modified 2,2'-bipyridine carrying an azide moiety that allows its conjugation to biotin via click chemistry and subsequent streptavidin–biotin interaction. Another important factor is that the reactivity toward thus-modified OsO_4 –bipy derivative remains similarly unchanged.

As a case in point, we evaluated the theoretically possible enrichment of m^5U containing fragments in an example RNA molecule. This hypothetical scenario is based on a number of model assumptions, including a reaction rate constants of $k = 0.33 \text{ min}^{-1}$ for m^5U and $k = 0.03 \text{ min}^{-1}$ for U. The model RNA molecule was assumed to be 1000 nucleotides in length and to contain equally distributed canonical nucleosides, plus a single m^5U residue. Prior to the osmylation reaction, the RNA was hydrolyzed into fragments of approximately 15 nucleotides, producing roughly 66 fragments per RNA molecule. The interest was now to determine if a hypothetical reagent with the properties described would be able to isolate the one fragment containing the m^5U residue and, in particular, to what purity. Side reactions would obviously occur when nonmethylated uridines were tagged as a consequence of the selectivity described above. Assuming that the osmylation reaction was stopped after one half-life (i.e., 2.1 min), 0.50 mol/mol m^5U would have reacted. In the same time, only 15 of 250 uridines (or 0.06 mol/mol U) would have been tagged. In the worst-case scenario in which each uridine comes from a different fragment, 15 out of 66 fragments would have reacted. This corresponds to 0.23 mol of labeled fragment per mol of total fragment. Assuming a near-perfect separation of labeled fragments from the rest, the best enrichment factor would be $0.5/0.23 = 2.2$ for the fragment containing m^5U . Although fragment size and reaction time clearly do modify this enrichment factor, the current state of selectivity obviously is too low for a whole-transcriptome enrichment approach like that recently employed with a pseudouridine-selective reagent.⁴⁷ The osmium tetroxide–bipy complex may, however, become highly useful if a method for the sequence-specific

5 RESULTS AND DISCUSSION

Bioconjugate Chemistry

Article

detection of the resulting osmylation products involving, e.g., reverse-transcription signatures⁵⁰ can be developed.

EXPERIMENTAL PROCEDURES

Materials. Osmium tetroxide (OsO₄), ammonium acetate, uridine, uridine-5′ monophosphate, cytidine, 5-methyl uridine, and LC–MS-grade acetonitrile were purchased from Sigma-Aldrich, and 5-methyl cytidine was purchased from Berry and associates. Bipyridine (bipy), disodium–hydrogen phosphate, and sodium–dihydrogen phosphate were purchased from Carl Roth. Urea was purchased from Acros. RNA oligonucleotides were purchased from IBA.

All reagents were dissolved in either Milli-Q water or deuterium oxide for NMR experiments.

Caution: osmium tetroxide in the solid state is very volatile and toxic. Perform experiments only under the hood!

Methods. Reaction for Pyrimidine Nucleosides with Osmium Tetroxide and Bipyridine. Pyrimidine nucleoside (50 μM) was mixed with 8 mM OsO₄, 4 mM bipy, and 7 M urea in a phosphate buffer (0.4 M, pH 7) and 50 μM adenosine in a glass vial covered with a PTFE cap to prevent the evaporation of osmium tetroxide. The reaction was performed at 25 °C, and depending on the pyrimidine used, an aliquot was drawn at different time points and the reaction stopped by dilution and the addition of corn oil. The oily phase was then removed, and the aqueous phase was directly used for HPLC analysis.

For statistical reasons, each reaction was performed in independent triplicate.

Reaction for Uridine 5′-Monophosphate with Osmium Tetroxide and Bipyridine. The same reaction conditions were used as above except that, prior to HPLC analysis, 1 U of FastAP (Thermo Scientific) including corresponding buffer was added to the aqueous phase, and the reaction was performed for 1 h at 37 °C.

For statistical reasons, each reaction was performed in independent triplicate.

Reaction for 5-mer RNA with Osmium Tetroxide and Bipyridine. The same reaction conditions were used as above. Additionally, prior to HPLC analysis, a digestion step (involving SPV and NP1; duration of 2 h) followed by a dephosphorylation step (fastAP, 1U) were performed as described in ref 17.

For statistical reasons, each reaction was performed in independent triplicate.

HPLC Analysis. An Agilent 1100 was equipped with a quat pump, autosampler, DAD detector, optional mass selective detector (LC–MSD–Trap-SP_10180).

HPLC conditions:

1. Analytical:

Column: YMC-Triart C18/S-3 μm/12 nm, column size: 150 × 3.0 mm I.D.

Buffer A: 5 mM ammonium acetate, pH 5.3

Buffer B: acetonitrile

Flow rate: 0.4 mL/min

Gradient I: 0–4 min, %B to 2.7% constant; 4–23 min, %B to 10%; 23–25 min, %B to 70%; 25–30 min, %B to constant 70%; 30–32 min, %B to 2.7%; 32–40 min, equilibrate

Detection: 260 nm for nucleosides and 312 nm for nucleoside–osmium–bipyridine complexes

2. Semipreparative:

Column: Phenomenex-Clarity Oligo-RP, 5 μm, column size: 250 × 10 mm I.D.

Buffer A: 5 mM ammonium acetate, pH 5.3

Buffer B: acetonitrile

Flow rate: 5 mL/min

Gradient II: 0–30 min, %B to constant 4%

Detection: 260 nm for nucleosides and 312 nm for nucleoside–osmium–bipyridine complexes.

Mass Spectrometry. In addition to the UV detection, a mass spectrometer was used for the identification of the osmium tetroxide–bipyridine complexes. Settings used included ion polarity: positive; ion Source: ESI; dry temp: 350 °C; nebulizer: 50 psi; dry gas: 12 l/min; trap drive: 43.2; Octapole RF amplitude: 142.5 Vpp; capillary exit: 125.1 V; skimmer: 40.0 V; Oct 1 DC: 12.0 V; Oct 2 DC: 1.9 V; scan range: 105–1000 m/z; averages: 2 spectra; maximum accumulation time: 500 ms; and ICC target: 100 000.

NMR. A Bruker 300 MHz NMR instrument was used.

Experiments performed: ¹H (differing numbers of scans), 2D- COSY, and NOESY

Determination of Reaction Rate Constants *k*, *k*₁, and *k*₂. For a competitive pseudo-first-order reaction, it holds that $A \xrightarrow{k} B + C$, with $A \xrightarrow{k_1} B$ and $A \xrightarrow{k_2} C$, where *k* is the overall reaction rate constant, and *k*₁ and *k*₂ are the formation rate constants for the two competing products. From the reaction law, it is known that

$$\frac{d[A]}{dt} = -k[A] = -(k_1 + k_2)[A] \quad (r1)$$

After integration,

$$[A] = A_0 e^{-(k_1+k_2)t} \quad (r2)$$

Additionally,

$$\frac{d[B]}{dt} = k_1[A] \quad (r3)$$

and

$$\frac{d[C]}{dt} = k_2[A] \quad (r4)$$

Substituting *A* from r2 in r3 and r4 and after integration, one gets:

$$[B] = \frac{k_1 A_0}{k_1 + k_2} (1 - e^{-(k_1+k_2)t}) \quad (r5)$$

and

$$[C] = \frac{k_2 A_0}{k_1 + k_2} (1 - e^{-(k_1+k_2)t}) \quad (r6)$$

This means that

$$[B]/[C] = k_1/k_2 \quad (r7)$$

From the decay of the substrate, one therefore can calculate the overall reaction rate as $k = k_1 + k_2$. Furthermore, from the ratio of the two products, one can calculate the ratio k_1/k_2 . From both reactions, one can thus calculate each individual *k*, *k*₁, and *k*₂ value.

The determination of *k* was calculated by fitting the normalized concentrations for each time point to a first-order reaction by using GraphPad Prism version 7.0 for Windows, GraphPad Software, La Jolla, CA. For fitting the parameters of a

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

Bioconjugate Chemistry

Article

first-order reaction type, the starting and the end concentrations were constrained to 1 and 0, respectively. The normalized starting concentrations was experimentally determined by performing a negative control for each reaction by excluding osmium tetroxide from the reaction mixture.

k_1 and k_2 for each substrate used were calculated using the mean of the ratio of both products for each measured time point. The standard deviation of the ratio was then used to calculate the standard deviation of k_1 and k_2 , respectively.

■ ASSOCIATED CONTENT

S Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.bioconjchem.6b00403.

Figures showing NMR spectra and other analysis results, the linear dependency of the natural logarithm of normalized concentration over time, and normalized uridine decay in a time-dependent manner. Tables containing NMR comparison analysis and quantification analysis and a summary of reaction rate constants. Additional details on NMR analysis. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mhelm@uni-mainz.de; tel: +49 6131 39 25731; fax: +49 6131 39 20373.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank all members of the Helm lab for technical assistance. Funding was provided by DFG FOR 1082 and HE 3397/6-2 to M.H.

■ REFERENCES

- 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. (2016) The Dynamic N(1)-Methyladenosine Methylome in Eukaryotic Messenger RNA. *Nature* 530, 441–446.
- 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. (2014) Transcriptome-Wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA. *Cell* 159, 148–162.
- Edelheit, S., Schwartz, S., Mumbach, M. R., Wurtzel, O., and Sorek, R. (2013) Transcriptome-Wide Mapping of 5-Methylcytidine RNA Modifications in Bacteria, Archaea, and Yeast Reveals m5C within Archaeal mRNAs. *PLoS Genet.* 9, e1003602.
- Dominissini, D., Moshitch-Moshkovitz, S., Salmon-Divon, M., Amariglio, N., and Rechavi, G. (2013) Transcriptome-Wide Mapping of N(6)-Methyladenosine by m(6)A-Seq Based on Immunocapturing and Massively Parallel Sequencing. *Nat. Protoc.* 8, 176–189.
- Shafik, A., Schumann, U., Evers, M., Sibbritt, T., and Preiss, T. (2016) The Emerging Epitranscriptomics of Long Noncoding RNAs. *Biochim. Biophys. Acta, Gene Regul. Mech.* 1859, 59–70.
- du Toit, A. (2016) RNA: Expanding the mRNA Epitranscriptome. *Nat. Rev. Mol. Cell Biol.* 17, 201.
- Schwartz, S. (2016) Cracking the Epitranscriptome. *RNA* 22, 169–174.
- Burgess, A., David, R., and Searle, I. R. (2016) Deciphering the Epitranscriptome: A Green Perspective. *J. Integr. Plant Biol.* DOI: 10.1111/jipb.12483.
- Dominissini, D. (2014) Genomics and Proteomics. Roadmap to the Epitranscriptome. *Science* 346, 1192.
- O'Connell, M. A., Mannion, N. M., and Keegan, L. P. (2015) The Epitranscriptome and Innate Immunity. *PLoS Genet.* 11, e1005687.
- Hussain, S., Aleksic, J., Blanco, S., Dietmann, S., and Frye, M. (2013) Characterizing 5-Methylcytosine in the Mammalian Epitranscriptome. *Genome Biol.* 14, 215.
- Saletore, Y., Meyer, K., Korch, J., Vilfan, I. D., Jaffrey, S., and Mason, C. E. (2012) The Birth of the Epitranscriptome: Deciphering the Function of RNA Modifications. *Genome Biol.* 13, 175.
- Witkin, K. L., Hanlon, S. E., Strasburger, J. a, Coffin, J. M., Jaffrey, S. R., Howcroft, T. K., Dedon, P. C., Steitz, J. a, Daschner, P. J., and Read-Connole, E. (2015) RNA Editing, Epitranscriptomics, and Processing in Cancer Progression. *Cancer Biol. Ther.* 16, 21–27.
- Mishima, E., Jinno, D., Akiyama, Y., Itoh, K., Nankumo, S., Shima, H., Kikuchi, K., Takeuchi, Y., Elkordy, A., Suzuki, T., et al. (2015) Immuno-Northern Blotting: Detection of RNA Modifications by Using Antibodies against Modified Nucleosides. *PLoS One* 10, e0143756.
- Khoddami, V., and Cairns, B. R. (2013) Identification of Direct Targets and Modified Bases of RNA Cytosine Methyltransferases. *Nat. Biotechnol.* 31, 458–464.
- Schaefer, M., Pollex, T., Hanna, K., and Lyko, F. (2008) RNA Cytosine Methylation Analysis by Bisulfite Sequencing. *Nucleic Acids Res.* 37, e12.
- Thüring, K., Schmid, K., Keller, P., and Helm, M. (2016) Analysis of RNA Modifications by Liquid Chromatography–tandem Mass Spectrometry. *Methods*, 1–9.
- Wagner, T. M., Nair, V., Guymon, R., Pomerantz, S. C., Crain, P. F., Davis, D. R., and McCloskey, J. A. (2004) A Novel Method for Sequence Placement of Modified Nucleotides in Mixtures of Transfer RNA. *Nucleic Acids Symp. Ser. (1979-2000)* 48 (48), 263–264.
- Ryvkin, P., Leung, Y. Y., Silverman, I. M., Childress, M., Valladares, O., Dragomir, I., Gregory, B. D., and Wang, L.-S. (2013) HAMR: High-Throughput Annotation of Modified Ribonucleotides. *RNA* 19, 1684–1692.
- Sharpless, K. B., Teranishi, A. Y., and Backvall, J. E. (1977) Chromyl Chloride Oxidations of Olefins. Possible Role of Organometallic Intermediates in the Oxidations of Olefins by Oxo Transition Metal Species. *J. Am. Chem. Soc.* 99, 3120–3128.
- Hayatsu, H., and Iida, S. (1969) Studies on the Chemical Modifications of Nucleic Acids. The Permanganate Oxidation of Thymine. *Tetrahedron Lett.* 10, 1031–1034.
- Fritzsche, E., Hayatsu, H., Igloi, G. L., Iida, S., and Kössel, H. (1987) The Use of Permanganate as a Sequencing Reagent for Identification of 5-Methylcytosine Residues in DNA. *Nucleic Acids Res.* 15, 5517–5528.
- Bui, C. T., Rees, K., and Cotton, R. G. H. (2003) Permanganate Oxidation Reactions of DNA: Perspective in Biological Studies. *Nucleosides, Nucleotides Nucleic Acids* 22, 1835–1855.
- Beer, M., Stern, S., Carmalt, D., and Mohlhenrich, K. H. (1966) Determination of Base Sequence in Nucleic Acids with the Electron Microscope. V. The Thymine-Specific Reactions of Osmium Tetroxide with Deoxyribonucleic Acid and Its Components *. *Biochemistry* 5, 2283–2288.
- Reske, T., Surkus, A.-E., Duwensee, H., and Flechsig, G.-U. (2009) Kinetics of the Labeling Reactions of Thymine, Cytosine and Uracil with Osmium Tetroxide Bipyridine. *Microchim. Acta* 166, 197–201.
- Daniel, F. B., and Behrman, E. J. (1976) Osmium (VI) Complexes of the 3', 5'-dinucleoside Monophosphates, ApU and UpA. *Biochemistry* 15, 565–568.
- Jelen, F., Karlovský, P., Makaturová, E., Pecinka, P., and Paleček, E. (1991) Osmium Tetroxide Reactivity of DNA Bases in Nucleotide Sequencing and Probing of DNA Structure. *Gen. Physiol. Biophys.* 10, 461–473.
- Chang, C. H., Beer, M., and Marzilli, L. G. (1977) Osmium-Labeled Polynucleotides. The Reaction of Osmium Tetroxide with

5 RESULTS AND DISCUSSION

Bioconjugate Chemistry

Article

Deoxyribonucleic Acid and Synthetic Polynucleotides in the Presence of Tertiary Nitrogen Donor Ligands. *Biochemistry* 16, 33–38.

(29) Subbaraman, L. R., Subbaraman, J., and Behrman, E. J. (1971) The Reaction of Osmium Tetroxide-Pyridine Complexes with Nucleic Acid Components. *Bioinorg. Chem.* 1, 35–55.

(30) Kanavarioti, A., Greenman, K. L., Hamalainen, M., Jain, A., Johns, A. M., Melville, C. R., Kemmish, K., and Andregg, W. (2012) Capillary Electrophoretic Separation-Based Approach to Determine the Labeling Kinetics of Oligodeoxynucleotides. *Electrophoresis* 33, 3529–3543.

(31) Kanavarioti, A. (2015) Osmylated DNA, a Novel Concept for Sequencing DNA Using Nanopores. *Nanotechnology* 26, 134003.

(32) Wrobel, K., Rodriguez Flores, C., Chan, Q., and Wrobel, K. (2010) Ribonucleoside Labeling with Os(VI): A Methodological Approach to Evaluation of RNA Methylation by HPLC-ICP-MS. *Metallomics* 2, 140–146.

(33) Umemoto, T., and Okamoto, A. (2008) Synthesis and Characterization of the 5-Methyl-2'-deoxycytidine Glycol-Dioxosmium-Bipyridine Ternary Complex in DNA. *Org. Biomol. Chem.* 6, 269–271.

(34) Bartosik, M., Hrstka, R., Palecek, E., and Vojtesek, B. (2014) Magnetic Bead-Based Hybridization Assay for Electrochemical Detection of microRNA. *Anal. Chim. Acta* 813, 35–40.

(35) Duwensee, H., Jacobsen, M., and Flechsig, G.-U. (2009) Electrochemical Competitive Hybridization Assay for DNA Detection Using Osmium Tetroxide-Labeled Signalling Strands. *Analyst* 134, 899–903.

(36) Reske, T., Mix, M., Bahl, H., and Flechsig, G.-U. (2007) Electrochemical Detection of Osmium Tetroxide-Labeled PCR-Products by Means of Protective Strands. *Talanta* 74, 393–397.

(37) Palecek, E., Fojta, M., and Jelen, F. (2002) New Approaches in the Development of DNA Sensors: Hybridization and Electrochemical Detection of DNA and RNA at Two Different Surfaces. *Bioelectrochemistry* 56, 85–90.

(38) Sopha, H., Wachholz, F., and Flechsig, G.-U. (2008) Cathodic Adsorptive Stripping Voltammetric Detection of tRNA by Labelling with Osmium Tetroxide. *Electrochem. Commun.* 10, 1614–1616.

(39) Henley, R. Y., Vazquez-Pagan, A. G., Johnson, M., Kanavarioti, A., and Wanunu, M. (2015) Osmium-Based Pyrimidine Contrast Tags for Enhanced Nanopore-Based DNA Base Discrimination. *PLoS One* 10, 1–12.

(40) Tanaka, K., Tainaka, K., Umemoto, T., Nomura, A., and Okamoto, A. (2007) An Osmium-DNA Interstrand Complex: Application to Facile DNA Methylation Analysis. *J. Am. Chem. Soc.* 129, 14511–14517.

(41) Okamoto, A., Tainaka, K., and Kamei, T. (2006) Sequence-Selective Osmium Oxidation of DNA: Efficient Distinction between 5-Methylcytosine and Cytosine. *Org. Biomol. Chem.* 4, 1638–1640.

(42) Daniel, F. B., and Behrman, E. J. (1975) Reactions of Osmium Ligand Complexes with Nucleosides. *J. Am. Chem. Soc.* 97, 7352–7358.

(43) Vaishnav, Y., Holwitt, E., Swenberg, C., Lee, H. C., and Kan, L. S. (1991) Synthesis and Characterization of Stereoisomers of 5,6-Dihydro-5,6-Dihydroxy-Thymidine. *J. Biomol. Struct. Dyn.* 8, 935–951.

(44) Mortimer, S. A., Trapnell, C., Aviran, S., Pachter, L., and Lucks, J. B. (2012) SHAPE-Seq: High-Throughput RNA Structure Analysis. *Curr. Protoc. Chem. Biol.* 4, 275–297.

(45) Lovejoy, A. F., Riordan, D. P., and Brown, P. O. (2014) Transcriptome-Wide Mapping of Pseudouridines: Pseudouridine Synthases Modify Specific mRNAs in *S. Cerevisiae*. *PLoS One* 9, e110799.

(46) Carlile, T. M., Rojas-Duran, M. F., Zinshteyn, B., Shin, H., Bartoli, K. M., and Gilbert, W. V. (2014) Pseudouridine Profiling Reveals Regulated mRNA Pseudouridylation in Yeast and Human Cells. *Nature* 515, 143–146.

(47) Li, X., Zhu, P., Ma, S., Song, J., Bai, J., Sun, F., and Yi, C. (2015) Chemical Pulldown Reveals Dynamic Pseudouridylation of the Mammalian Transcriptome. *Nat. Chem. Biol.* 11, 592–597.

(48) Okamoto, A. (2014) DNA-Osmium Complexes: Recent Developments in the Operative Chemical Analysis of DNA Epigenetic Modifications. *ChemMedChem* 9, 1958–1965.

(49) Kwok, C. K., Tang, Y., Assmann, S. M., and Bevilacqua, P. C. (2015) The RNA Structure: Transcriptome-Wide Structure Probing with next-Generation Sequencing. *Trends Biochem. Sci.* 40, 221–232.

(50) 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. (2015) The Reverse Transcription Signature of N-1-Methyladenosine in RNA-Seq Is Sequence Dependent. *Nucleic Acids Res.* 43, 9950–9964.

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

Supplementary information

Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain.

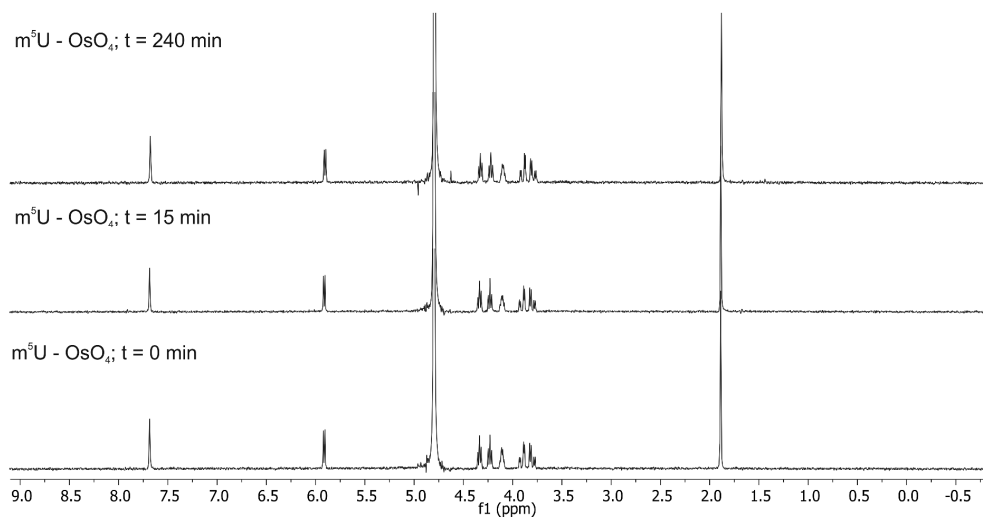
Lyudmil Tserovski¹ and Mark Helm^{1,*}

¹Institute of Pharmacy and Biochemistry, University of Mainz, D-55128 Mainz, Germany

*To whom correspondence should be addressed. Tel: +49 6131 39 25731; Fax: +49 6131 39 20373;
Email: mhelm@uni-mainz.de

Supplementary figure S1	2
Supplementary figure S2	3
Supplementary figure S3	4
Supplementary figure S4	5
Supplementary figure S5	6
Supplementary figure S6	7
Supplementary figure S7	7
Supplementary figure S8	8
Supplementary figure S9	8
Supplementary table S1	9
Supplementary table S2	9
Supplementary table S3	9
Supplementary table S4	10
Supplementary table S5	10

5 RESULTS AND DISCUSSION



Supplementary Figure S1: Reaction of 3 mM m⁵U with 8 mM OsO₄. Reaction was monitored by ¹H-NMR.

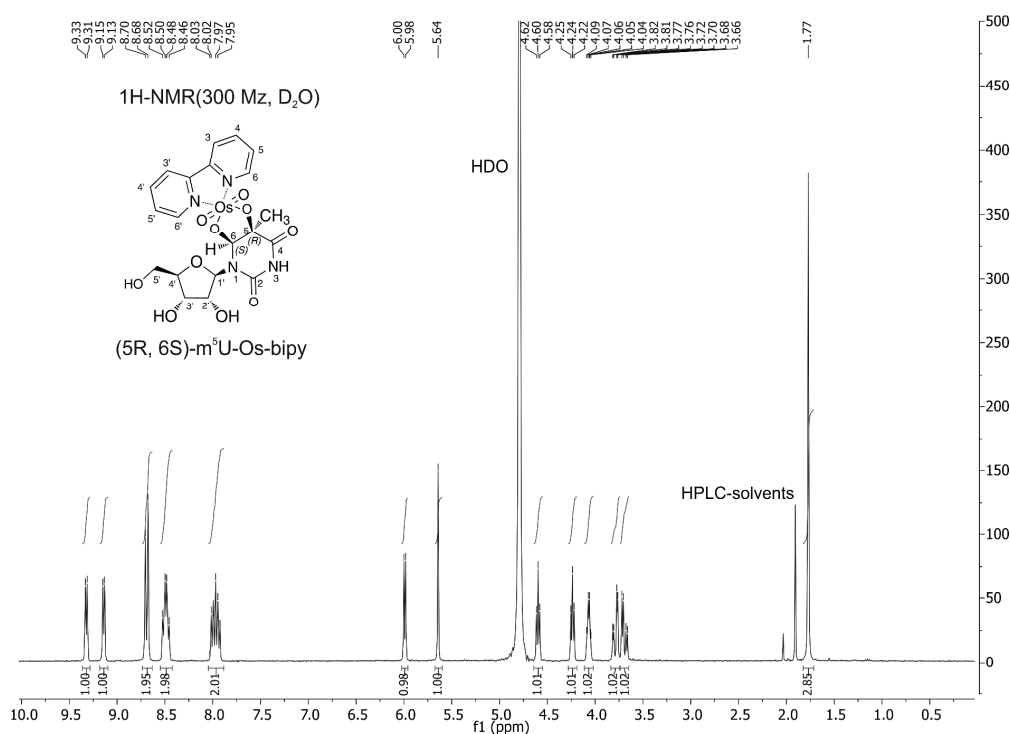
Supplementary Figure S1 shows, that after 4 hours of reaction, minimum oxidation of m⁵U is observed.

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

NMR analysis of HPLC separated diastereomers of 5-methyluridine glycol dioxosmium bipyridine

Settings used for the COSY scans: Temperature: ambient, Number of scans: 32, Receiver gain: 32, Relaxation delay: 1.7397, Pulse width 12.5, Acquisition time: 0.3355

Settings used for the NOESY scans: Temperature: ambient, Number of scans 64, Receiver gain: 32, Relaxation delay: 1.7397, Pulse width 12.5, Acquisition time: 0.3355



Supplementary Figure S2: ¹H-NMR of major diastereomer: (5R, 6S) 5-methyluridine glycol-dioxosmium bipyridine complex

(5R, 6S)-m⁵U-Os-Bipy (major isomer). ¹H NMR (D₂O, 300 MHz), δ_H 9.32 (1H, d, *J* = 5.5 Hz, H-C(6-bipy)*), 9.14 (1H, d, *J* = 5.4 Hz, H-C(6'-bipy)*), 8.69 (2H, d, *J* = 8.1 Hz, H-C(3, 3'-bipy)), 8.49 (2H, q, *J* = 7.7 Hz, H-C(4, 4'-bipy)), 8.04 – 7.91 (2H, m, H-C(5, 5'-bipy)), 5.99 (1H, d, *J* = 5.2 Hz, H-C(1')), 5.64 (1H, s, H-C(6)), 4.60 (1H, t, *J* = 5.4 Hz, H-C(2')**), 4.24 (1H, t, *J* = 5.0 Hz, H-C(3')**), 4.07 (1H, q, *J* = 4.1 Hz, H-C(4')), 3.79 (1H, dd, *J* = 12.6, 3.2 Hz, H'-C(5')***), 3.69 (1H, dd, *J* = 12.6, 4.7 Hz, H''-C(5')***), 1.77 (3H, s, H₃C(5)).

Notes:

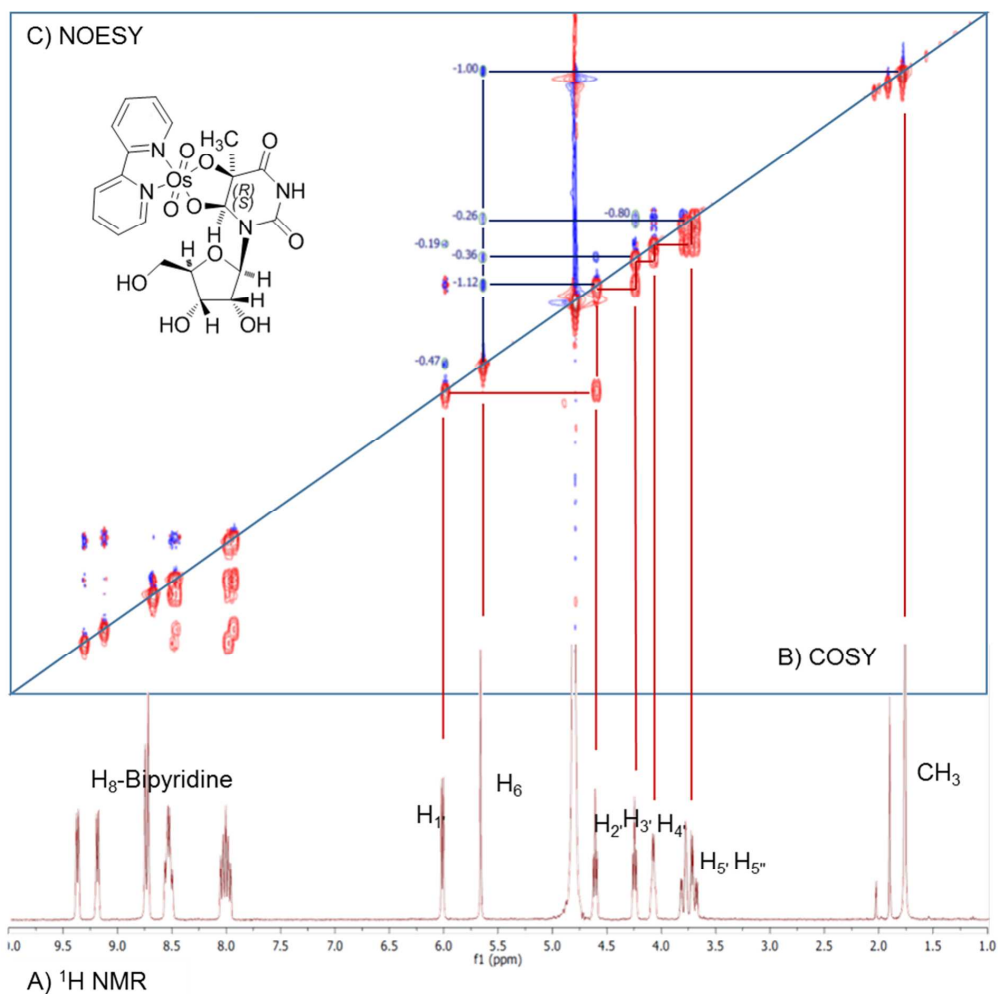
Signals at 1.89 and 2.08 belong to the HPLC mobile phase acetonitrile and acetate

* Two protons at 6 and 6' of 2, 2'-bipyridine have different chemical shifts due to different surrounding chemical environment.

5 RESULTS AND DISCUSSION

** These signals are apparent triplets. Since each proton couples to two neighboring protons, a doublet of doublets was expected.

*** Both protons at C(5') are diastereotopic, therefore also magnetic inequivalent.

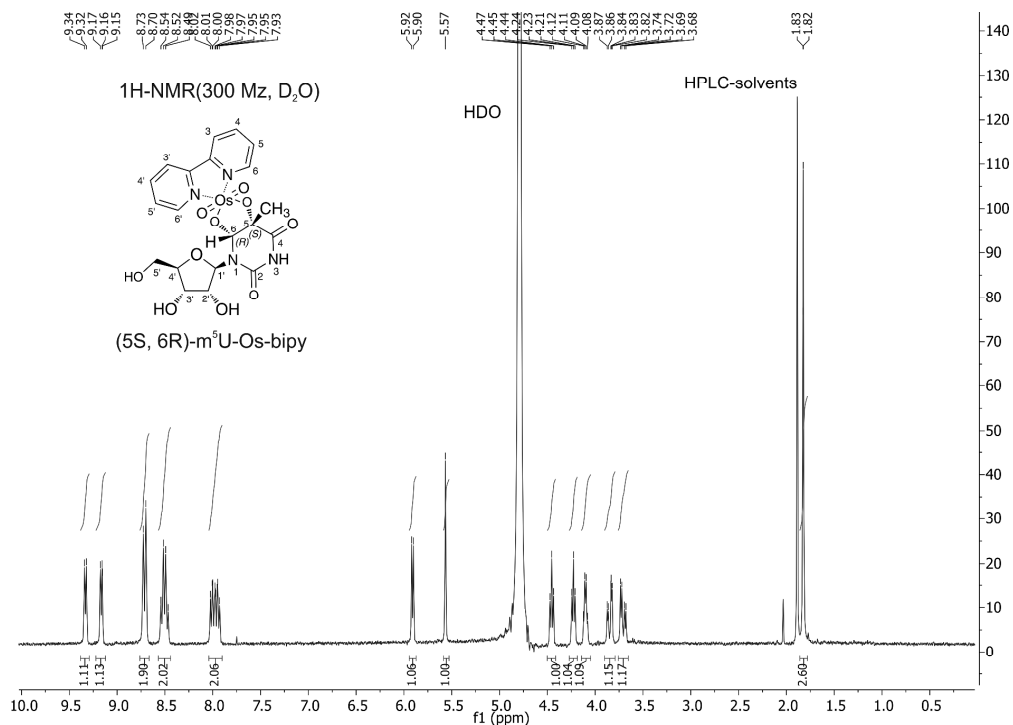


Supplementary Figure S3: 1- and 2-dimensional NMR analysis of major product (5R, 6S) 5-methyl uridine glycol dioxosmium bipyridine A) ^1H NMR spectrum B) COSY to determine the neighboring/coupling protons; C) NOESY experiment to determine spatially neighboring protons

Using COSY analysis, the neighboring protons of the ribose were determined. H-(1') couples to H-(2') which in turn couples to H-(3'). H-(3') is coupled to H-(4'), which is coupled to both H₂-(5').

In the NOESY analysis, the signal of cis configured H-(6) and H₃-(C5) was normalized to -1 and used to determine the remaining spatially neighboring signals. Of special interest was the signal coming from H-(6) and H-(2') measured to be -1.12.

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain



Supplementary Figure S4: ¹H-NMR of minor diastereomer: (5S, 6R) 5-methyluridine glycol-dioxosmium bipyridine complex

(5S, 6R)-m⁵U-Os-Bipy (minor isomer). ¹H NMR (D₂O, 300 MHz), δ_H 9.33 (1H, d, *J* = 5.6 Hz, H-C(6-bipy)*), 9.17 (1H, d, *J* = 5.2 Hz, H-C(6'-bipy)*), 8.71 (2H, d, *J* = 8.2 Hz, H-C(3, 3'-bipy)), 8.50 (2H, q, *J* = 7.6 Hz, H-C(4, 4'-bipy)), 8.05 – 7.91 (2H, m, H-C(5, 5'-bipy)), 5.91 (1H, d, *J* = 5.4 Hz, H-C(1')), 5.57 (1H, s, H-C(6)), 4.45 (1H, t, *J* = 5.5 Hz, H-C(2')**), 4.23 (1H, t, *J* = 5.1 Hz, H-C(3')**), 4.10 (1H, q, *J* = 4.4 Hz, H-C(4')), 3.85 (1H, dd, *J* = 12.6, 3.3 Hz, H'-C(5')***), 3.71 (1H, dd, *J* = 12.6, 4.6 Hz, H''-C(5')***), 1.82 (3H, s, H₃C(5)).

Notes:

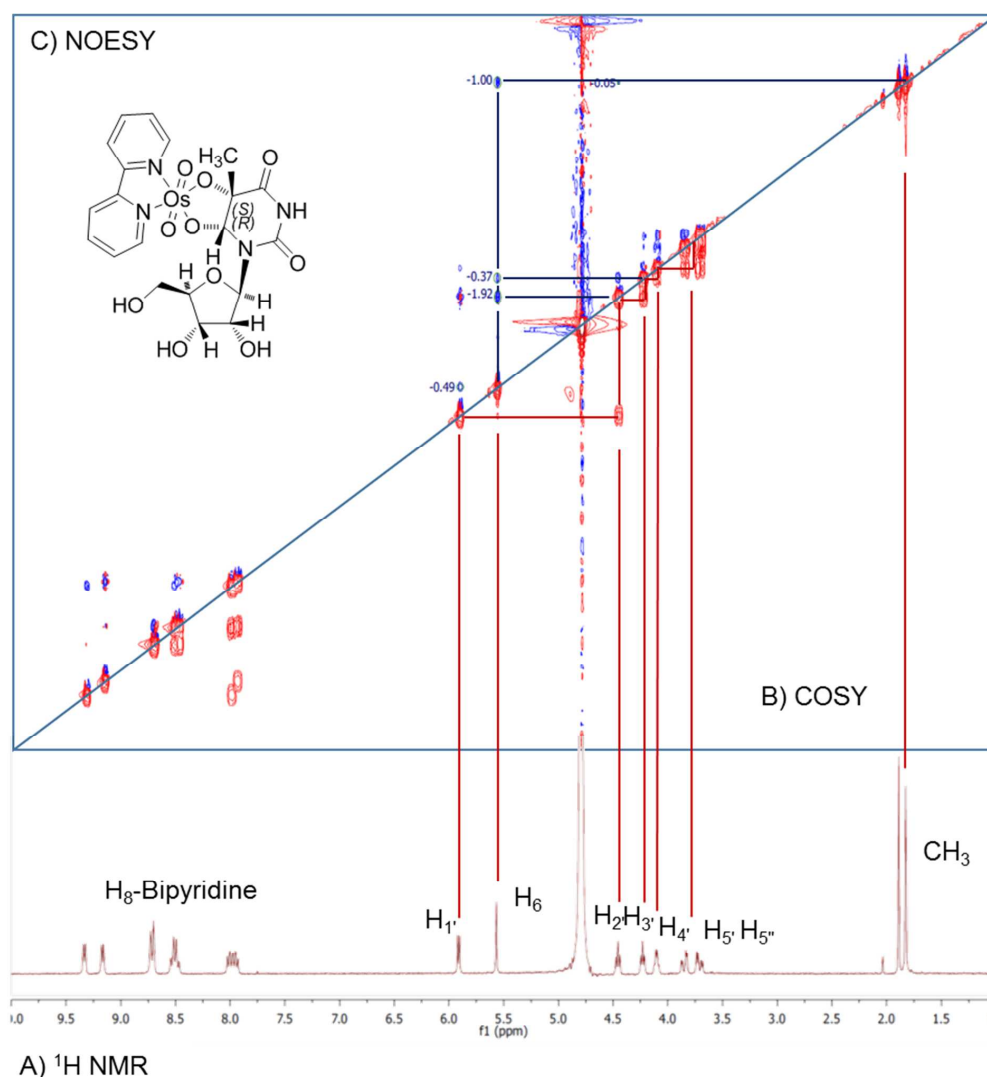
Signals at 1.89 and 2.08 belong to the HPLC mobile phase acetonitrile and acetate

* Two protons at 6 and 6' of 2, 2'-bipyridine have different chemical shifts due to different surrounding chemical environment.

** These signals are apparent triplets. Since each proton couples to two neighboring protons, a doublet of doublets was expected.

*** Both protons at C(5') are diastereotopic, therefore also magnetic inequivalent.

5 RESULTS AND DISCUSSION

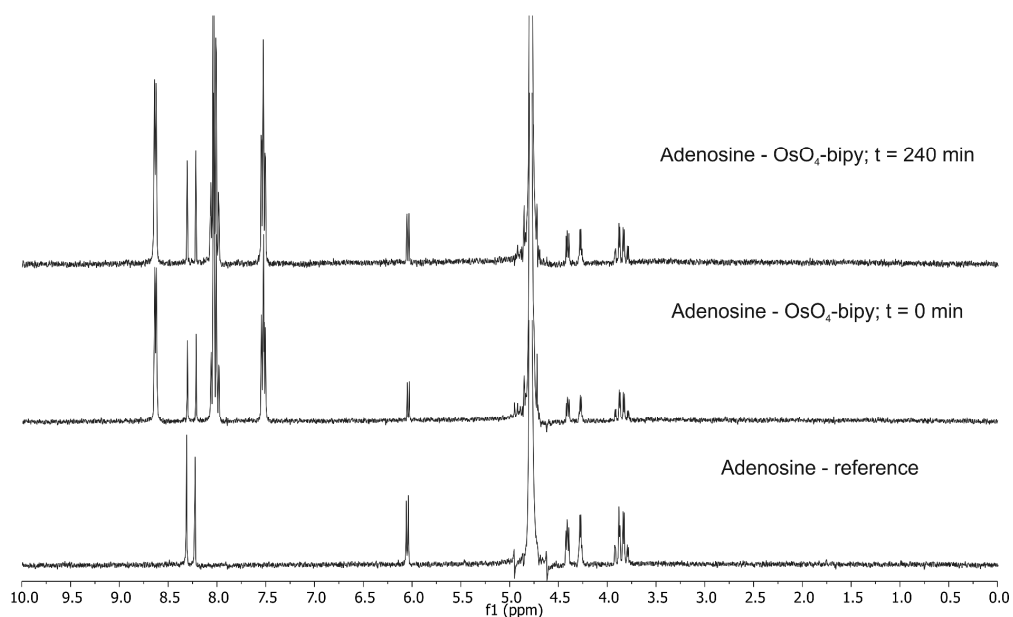


Supplementary Figure S5: 1- and 2-dimensional NMR analysis of minor product (5S, 6R) 5-methyluridine glycol dioxoosmium bipyridine A) ¹H NMR spectrum B) COSY to determine the neighboring/coupling protons; C) NOESY experiment to determine spatially neighboring protons.

Using COSY analysis, the neighboring protons of the ribose were determined. H-(1') couples to H-(2') which in turn couples to H-(3'). H-(3') is coupled to H-(4'), which is coupled to both H₂-(5').

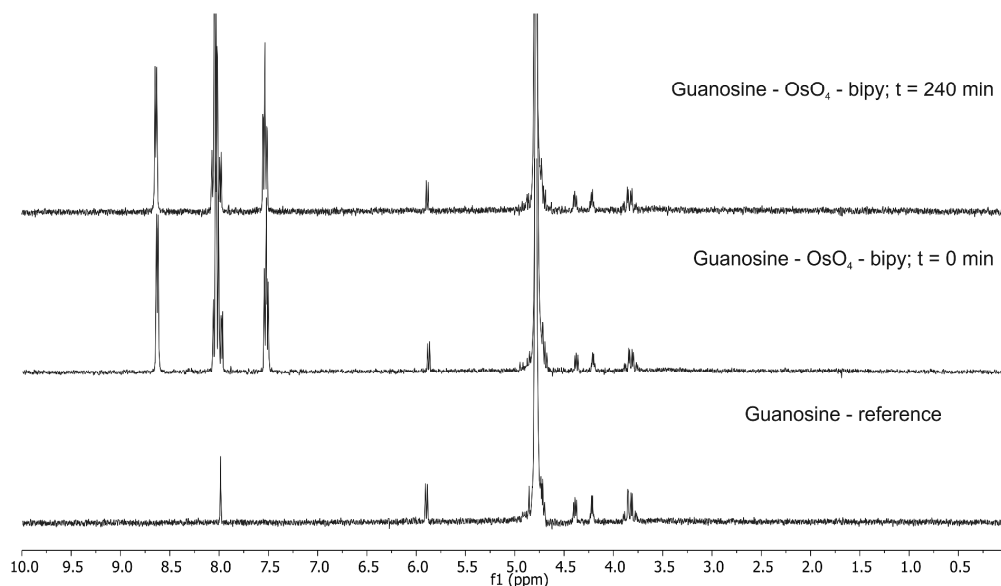
In the NOESY analysis, the signal of cis configured H-(6) and H₃-(C5) was normalized to -1 and used to determine the remaining spatially neighboring signals. In contrast to the major product (5R, 6S), in this case a much stronger signal coming from H-(6) and H-(2') was measured (-1.92). This indicated that those protons are spatially nearer to each other. This signal was used already analyzed by Vaishnav et al.¹ in characterizing the stereoisomers of 5,6-dihydro-5,6-dihydroxy-thymidine.

5.2 *Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain*



Supplementary figure S6: Reaction of 1 mM adenosine with 8 mM OsO₄ and 4 mM bipy. Reaction was monitored by ¹H-NMR.

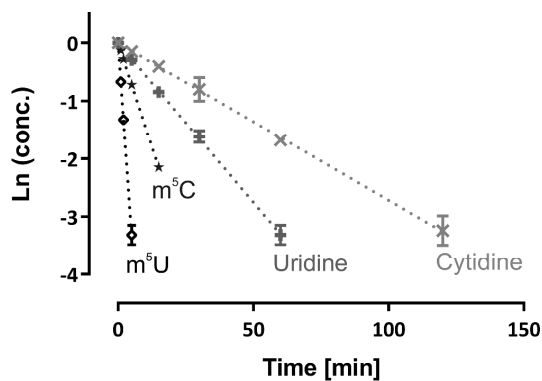
Supplementary Figure S6 shows that under these conditions, no reaction between adenosine and OsO₄-bipy after 4 hours was observed.



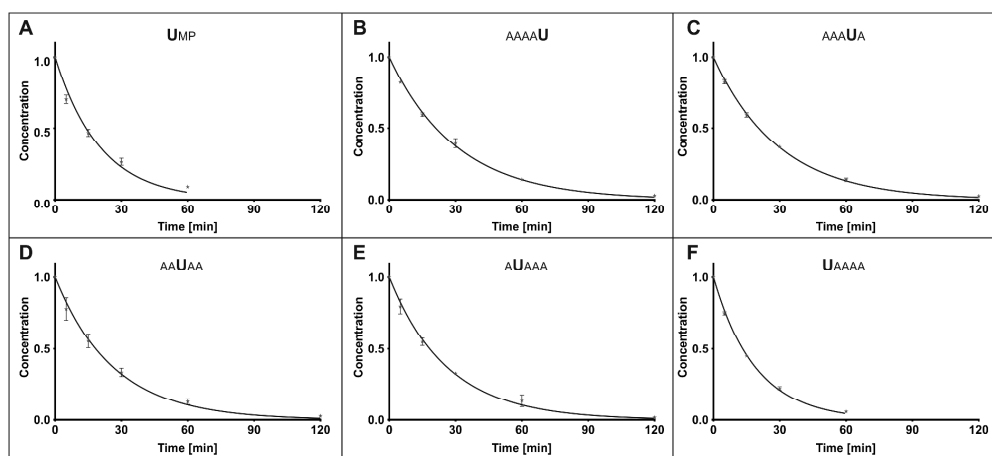
Supplementary Figure S7: Reaction of 1 mM guanosine with 8 mM OsO₄ and 4 mM bipy. Reaction was monitored by ¹H-NMR.

Supplementary Figure S7 shows that under these conditions, no reaction between guanosine and OsO₄-bipy after 4 hours was observed.

5 RESULTS AND DISCUSSION



Supplementary Figure S8: Linear dependency of natural logarithm of normalized concentration over time that confirmed pseudo first order reaction for osmylation of pyrimidines.



Supplementary Figure S9: Normalized uridine decay in a time dependent manner in the context of uridine monophosphate (A) and incorporated in short RNA chains (B-F). The error bars indicate the standard deviation of three independent experiments. The solid lines correspond to a fit of a first order type.

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

¹ H NMR(300 Mhz, D ₂ O)	thymidine glycol-dioxoosmium-bipyridine ²		5-methyl-uridine glycol-dioxoosmium-bipyridine	
	5R, 6S	5S, 6R	5R, 6S	5S, 6R
H ₃ -C(5)	1.53	1.56	1.77	1.82
H-C(6)	5.25	5.22	5.64	5.56
H-C(1')	6.23	6.09	5.99	5.91

Supplementary Table S1: Comparison of relevant ¹H-NMR signals for osmylated products of 5-methyl uridine and its DNA equivalent thymidine.

¹ H NMR(300 Mhz, D ₂ O)	m5C-Complex (Integral of signal)		
	Product 1	Product 2	Ratio P1/P2
H ₃ -C(5)	5193.06	1387.06	3.7
H-C(6)	1724.83	448.84	3.8
H-C(1')	1745.78	492.65	3.5
H-C(2')	1657.86	423.92	3.9
Mean:			3.8
HPLC [312 nm]	1222.19	329.64	3.7

Supplementary Table S2: Relevant ¹H NMR signals used for the relative quantification of diastereomeric products 5-methyl-cytidine glycol-dioxoosmium-bipy. Comparison with the peak areas from HPLC detected at 312 nm.

¹ H NMR(300 Mhz, D ₂ O)	U-complex (Integral of signal)		
	Product 1	Product 2	Ratio P1/P2
H-C(6)+H-C(1')	4653.68	1950.29	2.4
H-C(2')	2295.98	946.62	2.4
Mean:			2.4
HPLC [312 nm]	942.26	394.29	2.4

Supplementary Table S3: Relevant ¹H NMR signals used for the relative quantification of diastereomeric products uridine glycol-dioxoosmium-bipy. Comparison with the peak areas from HPLC detected at 312 nm.

5 RESULTS AND DISCUSSION

¹ H NMR(300 Mhz, D ₂ O)	C-complex (Integral of signal)		Ratio P1/P2
	Product 1	Product 2	
H-C(6)+H-C(1')	2640.78	875.43	3.0
H-C(2')	908.27	277.88	3.3
Mean:			3.1
HPLC [312 nm]	378.33	123.25	3.1

Supplementary Table S4: Relevant ¹H NMR signals used for the relative quantification of diastereomeric products cytidine glycol-dioxoosmium-bipy. Comparison with the peak areas from HPLC detected at 312 nm.

	Overall reactivity (<i>k</i>)	Product 1(<i>k</i> ₁)	Product 2 (<i>k</i> ₂)
Nucleoside			
C	0.027	-	-
U	0.055	0.042	0.013
m5C	0.142	0.115	0.026
m5U	0.664	0.521	0.143
Nucleotide			
UMP	0.049	0.040	0.010
5-mer RNA			
AAAAU	0.033	0.026	0.007
AAUA	0.034	0.020	0.014
AAUAA	0.038	0.022	0.016
AUAAA	0.038	0.023	0.015
UAAAA	0.052	0.035	0.018
CG[m5C]AU			
C	0.023	-	-
U	0.024	-	-
m5C	0.036	0.013	0.023
CG[m5U]AU			
C	0.017	-	-
U	0.029	-	-
m5U	0.325	0.119	0.206

Supplementary Table S5: Compilation of all determined reaction rate values *k*, *k*₁ and *k*₂.

5.2 Diastereoselectivity of 5-methyluridine osmylation is inverted inside an RNA chain

References:

- (1) Vaishnav, Y., Holwitt, E., Swenberg, C., Lee, H. C., Kan, L. S. (1991) Synthesis and Characterization of Stereoisomers of 5,6-Dihydro-5,6-Dihydroxy-Thymidine. *J. Biomol. Struct. Dyn.* 8, 935–951.
- (2) Umemoto, T., Okamoto, A. (2008) Synthesis and Characterization of the 5-Methyl-2'-deoxycytidine Glycol-Dioxoosmium-Bipyridine Ternary Complex in DNA. *Org. Biomol. Chem.* 6, 269–271.

5.3 Application of os-bipy reagent and next generation sequencing for detection of 5-methylpyrimidines

5.3.1 Stability of Os(VI)-bipy complex during reverse transcription

Oxo-amine-osmium(VI) complexes are very stable. Nevertheless, a cleavage with a reductive agent such as bisulfite or lithium aluminium hydride is described as the most effective method for *cis*-hydroxylation of alkenes [55]. Because an RNA library preparation protocol requires a reverse transcription in a buffer solution (RT-buffer), the question was raised whether the pyrimidine-os-bipy complex is stable toward this step. Reductive agent that is present in the reverse transcription reaction is for example dithiothreitol (DTT, 5 mM final concentration). In order to evaluate its effect, a short 5-mer RNA, featuring four adenosines and a single uridine in the central position, was reacted with osmium tetroxide and bipyridine under the conditions already described [155]. Upon completion of the reaction and purification, the os-bipy-complexed oligonucleotide was separated in three vials. One was prepared for reverse transcription according to the method described in [149] omitting the RT-primer. Another was prepared analogously, but omitting additionally the enzyme reverse transcriptase. The last was a control, without the buffer and without the reverse transcriptase. All three vials were kept at 50 °C for one hour in order to simulate the conditions of an actual reverse transcription. Finally, the RNA was digested according to [155] and products were analyzed using HPLC.

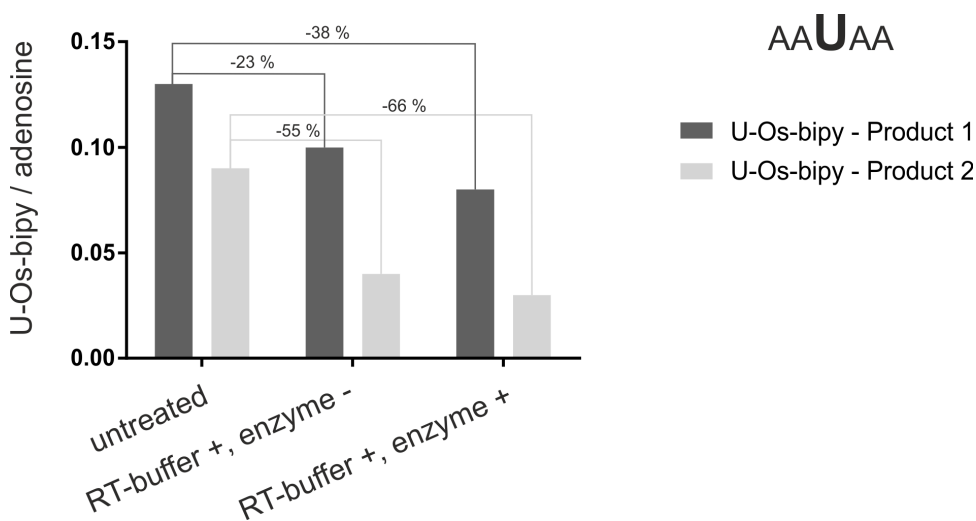


Figure 23: Effect of RT-buffer, with and without reverse transcriptase, on the stability of two uridine-os-bipy complexes. Products in each reaction were related to the adenosine content.

Figure 23 shows clearly that the overall product, normalized to the adenosine content, was decreased in both reactions where RT-buffer was added. The effect was more pronounced when the enzyme was present. Most striking is the observation that both diastereomers were unevenly affected. It was already demonstrated in section 5.2 that the stereochemical environment of the RNA chain has a substantial effect on the accessibility of os-bipy toward the 5,6 double bond of uridine. There the accessibility for formation of product 1 in a short RNA chain was hindered in comparison to the single nucleoside. Comparably, here the chain had a more protective effect for product 1, which led to its slower hydrolysis. Of note, expected hydrolyzed products 5,6-dihydroxy-5,6-dihydrouridine were not observed by the UV-detector due to loss of aromaticity. Also, these are the results of a single experiment, so validation is still missing.

5.3.2 Detection of 5-methylcytidine in yeast tRNA

Although the results from previous subsection 5.3.1 showed some instability of os-bipy complex during the reverse transcription step, an attempt was made to label 5-methylpyrimidines in yeast total tRNA. Cytosolic tRNA from *Saccharomyces cerevisiae* was chosen because of the numerous well-annotated 5-methylcytidine- and 5-methyluridine sites. Total tRNA⁶ was treated with Os-bipy under the conditions described in reference [155]. To determine possible time-dependent effects, two reaction times were chosen: 30 minutes and 1 hour. Both reactions were stopped by precipitation⁷ and samples were prepared for deep sequencing according to the protocol described in [149]. A corresponding control contained untreated tRNAs.

Figure 24A shows the sequencing profile of tRNA^{Ile(TAT)}. Clearly visible is the signature of m¹A at conserved position 58. The coverage is high starting at the 3' end of tRNA because of the nature of the library preparation protocol, that includes a cDNA synthesis starting from that end. Due to the complex 3D structure of tRNAs, as well as the high modification rate at the anticodon loop, the coverage profile is continuously dropping along the tRNA toward the 5' end. Recently developed software in the Helm group by Ralf Hauenschild [publication accepted] [156] allows a differential analysis and visualization of two samples, either by comparing the mismatch pattern-, or the context sensitive arrest rate change between them. Visual comparison between untreated and Os-bipy-treated samples shows no significant difference in the arrest rate (Figures 24AB and A.5).

In contrast, differential mismatch pattern analysis (Figure 24) shows significant changes comparing untreated- with Os-bipy-treated samples. The observation that mainly pyrimidines are effected by the chemical treatment, (denoted by X in figure 24C) confirms results shown in reference [155]. Not only does Os-bipy react with pyrimidine nucleosides but this also leads to a changed RT signature. Important is the fact, that among those pyrimidines, cytidine at position 48, annotated as 5-methylcytidine in MODOMICS [3], shows pronounced mismatch change. While the control sample showed no misincorporation, in the 30 min. treated sample a very high thymidine content was present (about 50 %). The misincorporation rate was even higher in the 1 hour treated sample (about 56 %). This means that reverse transcriptase incorporated also adenosines instead of correct guanosines. Two explanations for this effect are possible. Either Os-bipy as a labeling agent changes the base-pairing properties of 5-methylcytidine, or a deamination occurs leading to false base-pairing (see figure A.6).

Visual inspection of position 54, annotated as 5-methyluridine in MODOMICS [3], shows no significant increase in misincorporation. This, and the lack of arrest rate on osmylated positions, shows that under the given conditions and chemicals it is not possible to detect this modification.

The results presented in this chapter are to some extent contradictory with the previous one (Chapter 5.2). In reference [155] it was shown, that the reactivity of Os-bipy toward m⁵C, incorporated in a short RNA, and cytidine at the 5' extremity of the same oligonucleotide, was comparable. Here, a discrimination between both methylated- and non-methylated cytidines was visible. One possible explanation is, that in this case all pyrimidines are inside an RNA chain, and therefore the reactivity is expected to be similarly reduced. This effect was already observed for uridine incorporated at differing positions in a short RNA [155]. Also, it was expected that, upon reaction, Os-bipy-labeled pyrimidines would offer a road-block for a reverse transcriptase. Thus, a reaction with 5-methyluridine was expected to lead to a strong arrest

⁶Total tRNA from yeast was purchased from Roche, REF: 10109525001

⁷Precipitation of RNA was performed as follows: NH₄OAc to 0.5 M final conc. and three volumes of ice-cold ethanol were added. Samples were kept at -80° for 30 min, centrifugated at 13,000 rpm for 1 hour, washed with 70 % ethanol and dried. Finally, the pellets were redissolved in pure water.

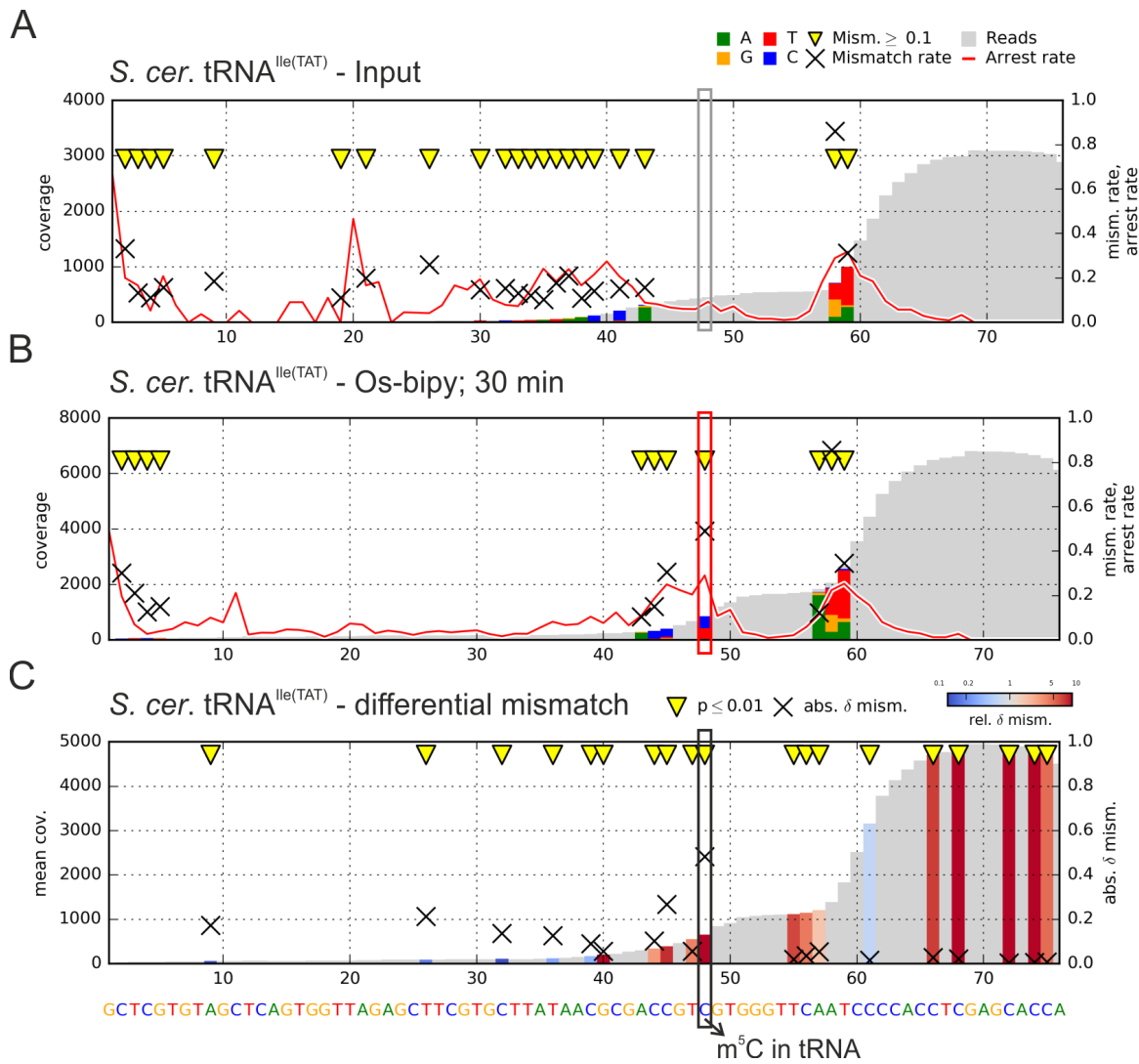


Figure 24: Comparison of sequencing profiles for tRNA^{Ile(TAT)}. Profiles were exported from CoverageAnalyzer, software developed in the Helm group by Ralf Hauenschild [156]. Red boxes denote the annotated position for m⁵C from MODOMICS. A) Sequencing profile of untreated sample. B) Sequencing profile of sample treated with Os-bipy for 30 min. C) Differential mismatch analysis of untreated- and Os-bipy treated samples. Mismatch change upon Os-bipy treatment is denoted by X.

rate at that position. This effect was not observed, which is probably due to, either the small size of the labeling agent, or strong capability of reverse transcriptases to read-through over modified sites. Of note, thirteen different reverse transcriptases were tested and upon visual inspection, no one lead to a significant arrest rate on osmylated positions (data not shown). One way to circumvent this, would be the exchange of the nitrogen donor 2,2'-bipyridine with a more sterically demanding substituent.

The results of this study were partially used in a recently published article describing the software CoverageAnalyzer developed by Ralf Hauenschild [156].



Article

CoverageAnalyzer (CAn): A Tool for Inspection of Modification Signatures in RNA Sequencing Profiles

Ralf Hauenschild ^{1,*}, Stephan Werner ¹, Lyudmil Tserovski ¹, Andreas Hildebrandt ², Yuri Motorin ³ and Mark Helm ^{1,*}

¹ Institute of Pharmacy and Biochemistry, Johannes Gutenberg University Mainz, Staudingerweg 5, 55128 Mainz, Germany; stwerner@uni-mainz.de (S.W.); ltserovs@uni-mainz.de (L.T.)

² Institute for Computer Sciences, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany; Andreas.Hildebrandt@uni-mainz.de

³ IMoPA UMR7365 CNRS-UL, BioPole de l'Université de Lorraine, 9 avenue de la Foret de Haye, 54505 Vandoeuvre-les-Nancy, France; motorine5@univ-lorraine.fr

* Correspondence: ralf.hauenschild@uni-mainz.de (R.H.); mhelm@uni-mainz.de (M.H.); Tel.: +49-6131-39-25731 (M.H.)

Academic Editor: Valérie de Crécy-Lagard

Received: 29 August 2016; Accepted: 21 October 2016; Published: 10 November 2016

Abstract: Combination of reverse transcription (RT) and deep sequencing has emerged as a powerful instrument for the detection of RNA modifications, a field that has seen a recent surge in activity because of its importance in gene regulation. Recent studies yielded high-resolution RT signatures of modified ribonucleotides relying on both sequence-dependent mismatch patterns and reverse transcription arrests. Common alignment viewers lack specialized functionality, such as filtering, tailored visualization, image export and differential analysis. Consequently, the community will profit from a platform seamlessly connecting detailed visual inspection of RT signatures and automated screening for modification candidates. CoverageAnalyzer (CAn) was developed in response to the demand for a powerful inspection tool. It is freely available for all three main operating systems. With SAM file format as standard input, CAn is an intuitive and user-friendly tool that is generally applicable to the large community of biomedical users, starting from simple visualization of RNA sequencing (RNA-Seq) data, up to sophisticated modification analysis with significance-based modification candidate calling.

Keywords: RNA modifications; reverse transcription; reverse transcription (RT) signature; RNA sequencing (RNA-Seq); Next-Generation Sequencing (NGS); candidate screening; alignment viewer

1. Introduction

The detection of RNA modifications has recently re-emerged as a very timely topic of current research. Coupled to new detection methods came new insights into the function of RNA modifications in the regulation of RNA stability [1], regulation of gene expression [2–5], and immunity [6]. RNA modifications are structurally highly diverse, and among the approximately 150 chemically different structures in the Modomics database [7], all major classes of natural product compounds can be found [8,9]. Furthermore, there is evidence that the diversity may yet increase with the discovery of more modifications [10]. Despite this high diversity, some common denominators apply to both function and detection. Here, two important features for detection are reverse transcription (RT) arrest and misincorporation during complementary DNA (cDNA) synthesis. Before the advent of methods that are nowadays subsumed as deep sequencing, RT reverse transcription arrest was traditionally analyzed by gel or capillary electrophoresis [11]. A model modification for misincorporation, inosine, the product of an A-to-I deamination, is reliably reverse transcribed into

5 RESULTS AND DISCUSSION

Biomolecules 2016, 6, 42

2 of 7

a cytidine rather than a thymidine residue in the resulting cDNA. This misincorporation has led to the first transcriptome-wide mapping of an RNA modification [12]. The combined appearance of both RT arrest and misincorporation at modification sites was analyzed in early work [13,14]. Detailed analysis showed correlation between modification type and the relative composition of misincorporated nucleotides [15]. Also, chemical treatments that selectively alter the properties of a given modification [16,17] may therefore be exploited as an additional layer of information in single RNA species or in transcriptome-wide mapping [18–20]. Collection [7] and curation [21] of RNA sequences containing modifications underline a central problem in the field, arising from the vast number of candidate sites in large datasets. Because of these vast numbers, experimental verification of candidate sites by independent methods must typically be restricted to a small subset. Before engaging in such an endeavor, the experimentalist, and potential user of the software presented here, may want to assess the significance of an identification event, and visually inspect parameters at a given site. In principle, a variety of so-called alignment viewers like IGV, Tablet, Savant, UGENE and Persephone provide more or less detailed graphical representations of mapping results, typically resolving the base composition and orientation of reads covering a reference sequence. However, our recent application of machine learning approaches to the identification of modification sites has uncovered an unmet need for particular features in said tools. Specifically, the combination of mismatch patterns and a newly defined RT arrest rate has emerged as the central feature allowing efficient identification of 1-methyladenosine residues [22]. In response, CoverageAnalyzer (CAn) was specifically created for analysis of modification signatures in deep sequencing data. Distinct from variant caller and single nucleotide polymorphisms (SNP) identification tools, it allows the definition of a highly detailed query, based on combinations of arrest rates and mismatch composition, as well as a Context Sensitive Arrest rate (CSA). A differential visualization tool is particularly useful to compare signatures upon differential chemical treatment, or between wild-type and knockout mutants e.g., of a methyltransferase [22]. CAn combines a data processing pipeline with flexible controls for independent or differential visualization and automated screening for modification candidates based on complex RT signatures.

2. Results

CAn was optimized to allow rapid pre-selection and convenient visualization of such sites in transcriptome data, which display conspicuous RT signatures and are therefore potential candidates for further scrutiny, e.g., by visual inspection. The RT signatures in question may comprise nucleotide misincorporation or transcription arrest, and frequently originate from nucleotide modification at the position of interest. Several library preparation protocols have been published that capture cDNA from abortive RT [16,20,22,23] and can therefore be fully exploited by CAn. However, even preparation methods that do not capture abortive cDNA may provide useful information by providing misincorporation signals that may be analyzed by CAn. It is hence recommended that the user familiarizes himself with details of the various preparations beforehand ([24]). A typical CAn-session, conceived to identify, highlight, and visually inspect modification candidates, is depicted in Figure 1. The user is required to input a dataset in SAM format, containing RNA sequencing (RNA Seq) reads mapped to a genome or transcriptome. This is converted to the internally used *Profile* format by an automated pipeline (Figure 1a). To optimally detect stalled RT events, a parameter called CSA was introduced, which queries a local background arrest rate near the inspection site and takes it into account. CSA was defined as the fold change of a site i 's arrest rate A [22] with respect to the median A of its sequence environment of r bases up- and r bases downstream (here $r = 5$):

$$CSA^r(i) = \frac{A_i}{\text{median}(A_{i-5}, A_{i-1}, A_{i+1}, \dots, A_{i+5})} \quad (1)$$

5.3 Application of os-bipy reagent and next generation sequencing for detection of 5-methylpyrimidines

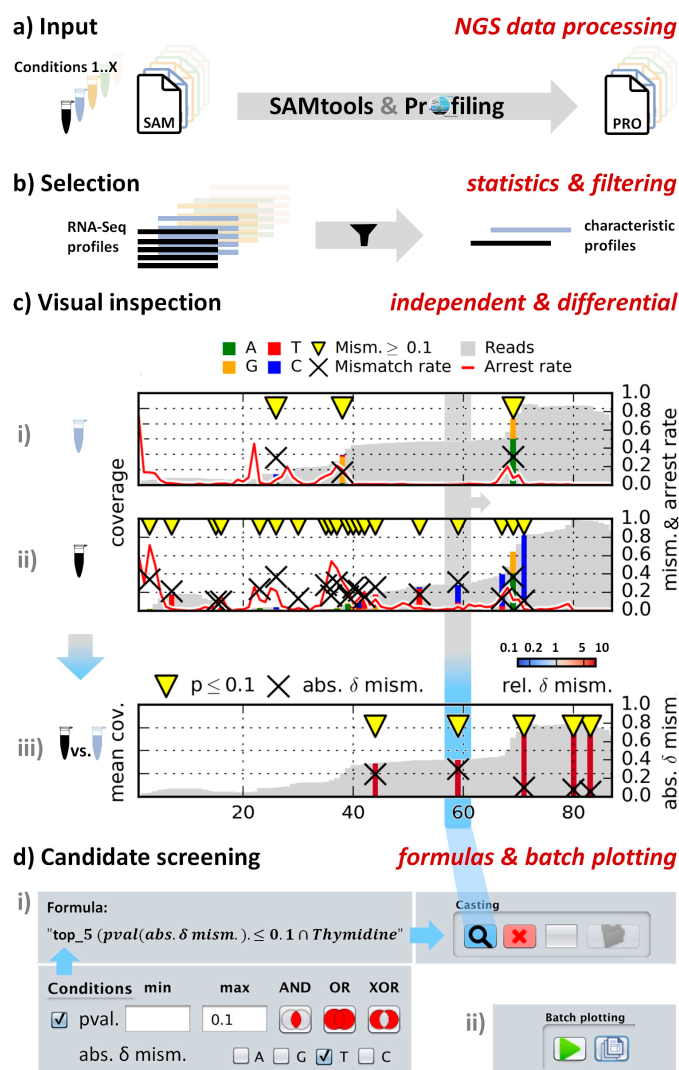


Figure 1. Workflow for a typical CA session (a) Input SAM files are processed to a positional profile; (b) Sorting and filtering of data by various statistical criteria. From the depicted result table, users select sequences for visualization; (c) Visualization tab. Independent plots and differential comparison for mismatch and/or arrest parameters with marked above-threshold sites (yellow triangle). Display of base sequences is enabled automatically depending on the horizontal plot dimension; (d) Candidate casting tab; (i) Formula editor: Specification of screening thresholds. Conditions are combined with Boolean operators AND, OR and XOR and can be parenthesized; (ii) Control panel for serial plotting of a resulting candidate batch.

The CSA feature, since it maps cDNA from abortive RT events, can only be meaningfully applied to data from library preparation protocols that specifically include such reads. Whether or not a protocol does so, typically hinges upon the incorporation step that introduces the second primer binding site. The installation package provides test datasets which were obtained by a library preparation protocol that captures abortive RT reads by ligation of a second adapter to the cDNA, as described in [22]. After selection of an RNA sequence of interest (Figure 1b), the software displays the sequence for visual inspection in a window (Figure 1c(i)). Events are labeled (yellow triangles), where values for misincorporation or peaks of CSA exceed adjustable thresholds. Additional profiles of the same

5 RESULTS AND DISCUSSION

Biomolecules 2016, 6, 42

4 of 7

sequence can be loaded and displayed in parallel plots (Figure 1c(ii)) for comparison of samples of variegated modification status, for example a wild-type RNA preparation *versus* one from a knockout organism lacking a certain modification activity [22]. Another application of interest is exemplified by the included test dataset, namely a chemical treatment suspected to alter the profile of certain modifications. With *test data 1* as the naive sample in window (i) and *test data 2* treated with an agent causing partial deamination of 5-methylcytidines in window (ii); a differential plot was generated in window (iii), where differences are displayed according to self-defined threshold criteria, and combinations thereof. The candidate casting tab (snippet shown in Figure 1d) offers a formula editor to generate filter rules of arbitrary complexity using thresholds combined by Boolean expressions and brackets. The resulting *candidates* files can be submitted to batch plotting for fast visual inspection of many candidate positions. With high flexibility in image dimensions, parameters, and legend details, these data can be exported as publication-ready images.

3. Discussion

CAn is a tool that allows the visualization and assisted inspection of deep sequencing data in the search for RNA modifications. Perusal of vast amounts of data is facilitated by a toolbox that allows to automatically highlight sites, where noticeably unusual combinations of RT arrest and misincorporation hint at the potential presence of modifications. Of note, there are no predefined thresholds that the program uses to flag unusual instances. Rather, it is up to the user to define threshold values for different parameters, and to combine them by Boolean operators. CAn is not meant to predict a modification event, or even to decipher the chemical structure of a potential modification. The program is rather designed to point attention to special candidate sites for its visual inspection. Inspection of large datasets automatically increases the statistical likelihood of the occurrence of conspicuous signals without a biochemical cause. Therefore, it is prudent to increase the stringency in such a case. While it is left to the user to decide how *p*-values are used to gauge the significance of findings, we recommend to use techniques like the Bonferroni correction [25] in order to account for the number of tested positions. In addition, the False Discovery Rate (FDR) can be controlled in the manner of Benjamini and Hochberg [26]. Outside these two approaches, which are rooted in statistics, a number of experimental approaches are open to the user to improve confidence by experimentally validating the candidate sites proposed by CAn. We urgently propose to call and treat these sites as “candidates” until validated by further experiments, e.g., by biochemical interrogation of a suggested site. In this context, we again emphasize the comparison feature, through which CAn specifically provides the possibility to inspect profiles before and after treatment with specific chemicals known to alter the RT-profile of a given modification. These may include, e.g., a Dimroth rearrangement of 1-methyladenosine (m¹A) by alkaline treatment [27], acrylonitrile treatment for the detection of inosine [28], and others [16].

4. Materials and Methods

4.1. Implementation

The graphical user interface (GUI) and the core of CAn are written in Java. The Miniconda Python based plotting component uses Matplotlib [29], Numpy [30] and Scipy. The software is distributed as self-extracting archive (~100 MB) for Windows (64-bit) and as zip files installed via included script setup routines for Linux and Mac OS X. Dependencies are downloaded automatically. On Mac OS X, latest Homebrew is installed to setup SAMtools [31]. Java 1.7+ is expected to be installed by the user, whereas Linux version installs dependencies via *apt-get*. Test data, a getting-started screencast and a user manual are included.

5.3 Application of os-bipy reagent and next generation sequencing for detection of 5-methylpyrimidines

Biomolecules 2016, 6, 42

5 of 7

4.2. Workflow

From unseen SAM input data files from N user-specified samples and the original FASTA mapping reference, CAN creates sorted and indexed Binary Alignment/Map (BAM) and finally the *Pileup* format. Users may replace the generated results on the hard drive with own files if they prefer different SAMtools parameters. In *Pileup*, periods and commas indicate matches, As, Gs, Ts and Cs mismatches and the arrest rate A of position i can be calculated as quotient of circumflexes at $i + 1$ and coverage at $i + 1$. Thus, a tabular *Profile* format is created, listing sequence positions line-wise with columns providing information on: position, reference base, coverage, mismatch rate M , number of (#) As, #Gs, #Ts, #Cs, and arrest rate A . *Profile* is divided into subfiles named by an $x_y.txt$ tag, where x represents the reference number and y the y^{th} file of a 1 kb block of subsequence of reference x . For example, a file named 3_7.txt contains data for positions 6001–6430, if the third reference has 6430 nt. Hence, hashing allows fast access to a query region without reading or memorizing leading positions, when accessing ends of long reference sequences. Thus, although the scope is on short sequences of RNA, chromosomes can be handled, too. In parallel, statistics are gathered for reference sequences (Figure 1b): ID, file path, length, sequence (first 100 nucleotides (nt)), coverage peak, number of high-arrest sites (S_A), high mismatch sites (S_M), heterogeneous mismatch sites (S_H) and mapped reads. This facilitates manual sorting and filtering by the user for visualization. Let c be the coverage at position i of reference f of length n . Let R be the reference base at i . Let $F_b(f, i) \stackrel{\text{obs.}(b,i)}{c(f_i)} :=$ where $b \in \{A, G, T, C\}$ be the observed frequency of base type b covering i in f . Thus, $mF := \{F_b(f, i), \text{ with } b \neq R\}$ is the set of mismatching $F_b(f, i)$. All i with $c \geq 20$ contribute to S_{H_f} , if two or more mismatch types exhibit a minimum mismatch rate of 0.1:

$$S_{H_f} := \sum_{i=1}^n x, \text{ where } x = 1 \text{ if } c(f_i) \geq 20 \text{ and } \text{median}_k mF(f, i)_k \geq 0.1, 0 \text{ else.} \quad (2)$$

S_A and S_M are calculated similarly, for arrest or mismatch rates exceeding a threshold normalized with coverage c , such that low arrest rates are considered insignificant at low c , but captured if c is high.

5. Conclusions

CAN was developed as a cross-platform open-source software running on most current computers. It allows efficient inspection of RNA Seq profiles for RT signatures of modifications, such as m¹A [22]. The user is provided with assistance to identify unusual patterns, to compare different datasets containing the same sequence, and to perform significance-based candidate calling. Important to the field is the implementation of both misincorporation patterns and RT arrest, including also the CSA format as defined during our recent extraction of RT signatures by machine learning [22]. CAN is highly conducive to the extraction of complete RT signatures, by providing full control of all thresholds for visualization, identification and discrimination to the user.

Supplementary Materials: CoverageAnalyzer setup files (incl. source code), instructions and documentation can be downloaded at: <https://zenodo.org/record/164811> (doi:10.5281/zenodo.164811) or <https://sourceforge.net/projects/coverageanalyzer/>.

Acknowledgments: This work was supported by the DFG SPP1784, MH3397/12-1, MH3397/14-1 to M.H.; by an ANR HTRNAMod ANR-13-ISV8-0001 grant to Y.M. and by a fellowship to R.H. from the International PhD Programme (IPP) at the Institute of Molecular Biology (IMB) Mainz, funded by the Boehringer Ingelheim Foundation.

Author Contributions: R.H. and M.H. conceived the software and wrote the paper; R.H. developed the software; S.W. contributed to realization of software distribution formats; L.T. conceived and performed the biomolecular experiments; Y.M. provided the sequencing service and tested the software together with A.H. and S.W.

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

5 RESULTS AND DISCUSSION

Biomolecules 2016, 6, 42

6 of 7

References

- Motorin, Y.; Helm, M. tRNA stabilization by modified nucleotides. *Biochemistry* **2010**, *49*, 4934–4944. [[CrossRef](#)] [[PubMed](#)]
- Chen, K.; Zhao, B.S.; He, C. Nucleic acid modifications in regulation of gene expression. *Cell Chem. Biol.* **2016**, *23*, 74–85. [[CrossRef](#)] [[PubMed](#)]
- Frye, M.; Jaffrey, S.R.; Pan, T.; Rechavi, G.; Suzuki, T. RNA modifications: What have we learned and where are we headed? *Nat. Rev. Genet.* **2016**, *17*, 365–372. [[CrossRef](#)] [[PubMed](#)]
- Spenkuch, F.; Motorin, Y.; Helm, M. Pseudouridine: Still mysterious, but never a fake (uridine)! *RNA Biol.* **2014**, *11*, 1540–1554. [[CrossRef](#)] [[PubMed](#)]
- 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.* **2016**, 1–16. [[CrossRef](#)] [[PubMed](#)]
- Dalpke, A.; Helm, M. RNA mediated Toll-like receptor stimulation in health and disease. *RNA Biol.* **2012**, *9*, 828–842. [[CrossRef](#)] [[PubMed](#)]
- 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*, D262–D267. [[CrossRef](#)] [[PubMed](#)]
- Motorin, Y.; Helm, M. RNA nucleotide methylation. *Wiley Interdiscip. Rev. RNA* **2011**, *2*, 611–631. [[CrossRef](#)] [[PubMed](#)]
- Helm, M.; Alfonzo, J.D. Posttranscriptional RNA Modifications: Playing metabolic games in a cell's chemical legoland. *Chem. Biol.* **2014**, *21*, 174–185. [[CrossRef](#)] [[PubMed](#)]
- Kellner, S.; Neumann, J.; Rosenkranz, D.; Lebedeva, S.; Ketting, R.F.; Zischler, H.; Schneider, D.; Helm, M. Profiling of RNA modifications by multiplexed stable isotope labelling. *Chem. Commun.* **2014**, *50*, 3516–3518. [[CrossRef](#)] [[PubMed](#)]
- Lempereur, L.; Nicoloso, M.; Riehl, N.; Ehresmann, C.; Ehresmann, B.; Bachellerie, J.P. Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible. *Nucleic Acids Res.* **1985**, *13*, 8339–8357. [[CrossRef](#)] [[PubMed](#)]
- Levanon, E.Y.; Eisenberg, E.; Yelin, R.; Nemzer, S.; Hallegger, M.; Shemesh, R.; Fligelman, Z.Y.; Shoshan, A.; Pollock, S.R.; Sztybel, D.; et al. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat. Biotechnol.* **2004**, *22*, 1001–1005. [[CrossRef](#)] [[PubMed](#)]
- Ebhardt, H.A.; Tsang, H.H.; Dai, D.C.; Liu, Y.; Bostan, B.; Fahlman, R.P. Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.* **2009**, *37*, 2461–2470. [[CrossRef](#)] [[PubMed](#)]
- Findeiss, S.; Langenberger, D.; Stadler, P.F.; Hoffmann, S. Traces of post-transcriptional RNA modifications in deep sequencing data. *Biol. Chem.* **2011**, *392*, 305–313. [[CrossRef](#)] [[PubMed](#)]
- Ryvkin, P.; Leung, Y.Y.; Silverman, I.M.; Childress, M.; Valladares, O.; Dragomir, I.; Gregory, B.D.; Wang, L.S. HAMR: High-throughput annotation of modified ribonucleotides. *RNA* **2013**, *19*, 1684–1692. [[CrossRef](#)] [[PubMed](#)]
- Behm-Ansmant, I.; Helm, M.; Motorin, Y. Use of specific chemical reagents for detection of modified nucleotides in RNA. *J. Nucleic Acids* **2011**, 2011. [[CrossRef](#)] [[PubMed](#)]
- Schaefer, M.; Pollex, T.; Hanna, K.; Lyko, F. RNA cytosine methylation analysis by bisulfite sequencing. *Nucleic Acids Res.* **2009**, *37*. [[CrossRef](#)] [[PubMed](#)]
- 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*, 143–146. [[CrossRef](#)] [[PubMed](#)]
- 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*, e110799. [[CrossRef](#)] [[PubMed](#)]
- Schwartz, S.; Bernstein, D.A.; Mumbach, M.R.; Jovanovic, M.; Herbst, R.H.; Leon-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*, 148–162. [[CrossRef](#)] [[PubMed](#)]

5.3 Application of os-bipy reagent and next generation sequencing for detection of 5-methylpyrimidines

21. Sun, W.J.; Li, J.H.; Liu, S.; Wu, J.; Zhou, H.; Qu, L.H.; Yang, J.H. RMBase: A resource for decoding the landscape of RNA modifications from high-throughput sequencing data. *Nucleic Acids Res.* **2015**. [[CrossRef](#)] [[PubMed](#)]
22. Hauenschild, R.; Tserovski, L.; Schmid, K.; Thuring, K.; Winz, M.L.; Sharma, S.; Entian, K.D.; Wacheul, L.; Lafontaine, D.L.; Anderson, J.; et al. The reverse transcription signature of N-1-methyladenosine in RNA-seq is sequence dependent. *Nucleic Acids Res.* **2015**, *43*, 9950–9964. [[CrossRef](#)] [[PubMed](#)]
23. 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. [[PubMed](#)]
24. Head, S.R.; Komori, H.K.; LaMere, S.A.; Whisenant, T.; Van Nieuwerburgh, F.; Salomon, D.R.; Ordoukhanian, P. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques* **2014**, *56*, 61–64. [[CrossRef](#)] [[PubMed](#)]
25. Bonferroni, C. Sulle medie multiple di potenze. *Bollettino dell'Unione Matematica Italiana* **1950**, *5*, 267–270. (In Italian)
26. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300.
27. 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*, 441–446. [[CrossRef](#)] [[PubMed](#)]
28. 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*, 715–732. [[CrossRef](#)] [[PubMed](#)]
29. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [[CrossRef](#)]
30. Van der Walt, S.; Colbert, S.C.; Varoquaux, G. The numpy array: A structure for efficient numerical computation. *Comput. Sci. Eng.* **2011**, *13*, 22–30. [[CrossRef](#)]
31. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics* **2009**, *25*, 2078–2079. [[CrossRef](#)] [[PubMed](#)]



© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (<http://creativecommons.org/licenses/by/4.0/>).

6 Conclusion and Outlook

The topic of this thesis lies within the dynamic field of rapidly developing tools for detection of RNA and its naturally occurring modifications. In general, several achievements upon the successful detection of RNA modifications have to be mentioned:

- HPLC-MS methods were developed for a sensitive discrimination of modified nucleosides. The methods that include stable isotope-labeled internal standards are especially useful in quantitative analyses [83, 157]. Drawback of these methods, however, is the loss of sequence information.
- Reverse transcription (RT) methods are used for sequence-specific detection of modified nucleosides, when the modification directly interrupts the Watson-Crick base-pairing properties of the underlying nucleoside [95]. Alternatively, for modifications that do not impede the reverse transcriptase, specific chemicals were developed that upon reaction lead to an RT stop at the modified position [95].
- Chemical labeling, or usage of antibodies against certain modifications, in combination with high-throughput sequencing led to the detection of RNA modifications on a transcriptome-wide level. These include, for example, the detection of 5-methylcytidine [136], *N*-6-methyladenosine [142, 37, 38] and pseudouridine [42, 41, 43].

The project presented here focused on two subjects.

Detection of *N*-1-methyladenosine

First, the development of a reverse transcriptase-based high-throughput sequencing method for detection of naturally occurring RNA modifications was aimed at. For this, a library preparation protocol was adapted that was able to capture reverse transcriptase events occurring during the synthesis of complementary DNA from RNA. For modifications that interrupt the base-pairing properties of the underlying nucleoside, such RT events include, among others, abortive cDNA products and misincorporations.

N-1-methyladenosine (m^1A) was used as a model modification for reasons already described in the section 5.1.2. It was demonstrated, that this modification, leaves a very specific RT-signature in sequencing data. This signature, highly dependent on the immediate 3'-neighbor, allowed for the first time to show that a computer-assisted prediction of the existence of m^1A at previously unknown positions is possible. One such position was validated by a complementary method. It included the targeted isolation of RNA of interest, followed by a digestion step and HPLC-MS analysis.

Using the developed approach, detection of m^1A modifications in high abundant RNAs such as tRNAs and rRNAs was achieved. However, detection of this modification in rare RNA species such as mRNA was not done yet. At the time this work was written, two approaches on the transcriptome-wide detection of m^1A were published [48, 47]. Both described methods are based on an immunoprecipitation step and thus output only regions of m^1A presence. Additionally, Chen et al. [158] implemented a support vector machine for the prediction of existent m^1A sites using the data sets produced by the work of Dominissini et al. [47]. This computer-based method calculates a feature space out of the chemical nature of the nucleosides that build the RNA of interest. On the contrary, the method developed in our group implements random forest as a machine-learning approach and uses the mismatch composition and the arrest rate as the most informative features.

In further experiments, the method presented here, has to be applied to a total RNA, in order to assess its applicability on a transcriptome-wide basis. Design of corresponding complementary experiments is important as well. One possibility is the mild alkaline treatment of RNA, upon which m¹A sites are converted to m⁶A in a Dimroth rearrangement reaction. This approach was successfully applied by Dominissini et al. [47]. The other possibility is the usage of a demethylase enzyme ALKB, that among other methylated nucleosides converts *N*-1-methyladenosine to adenosine, as was demonstrated by Li et al. [48]. In both cases, a change in the behavior of the reverse transcriptase is aimed at. This change can then be evaluated by differential analysis e.g. by CoverageAnalyzer, a software that was recently developed in our group by Ralf Hauenschild [publication accepted].

Detection of 5-methylpyrimidines

As was already mentioned in the introduction section 1.3.2, methylations outside the Watson-Crick Edge typically do not impede a reverse transcriptase. Therefore, 5-methylcytidine as well as 5-methyluridine cannot be directly detected by sequencing approaches. Several methods have been developed and already applied to the transcriptome-wide detection of 5-methylcytidine (see section 1.5.3). On the contrary, while 5-methyluridine was first discovered about 60 years ago [77], knowledge on the existence of this modification is still limited to tRNA and rRNA. Also, transcriptome-wide detection for this modification is still missing.

Therefore, a second focus in this work was the evaluation of osmium tetroxide and bipyridine (os-bipy) as a labeling agent for both 5-methylpyrimidines. The results obtained after reaction of os-bipy with tRNA, presented in section 5.3.2, show high potential for this labeling agent toward 5-methylcytidine. 5-methyluridine showed high reactivity in single nucleosides and short oligonucleotides. Nonetheless, in the context of total tRNA, in combination with high-throughput sequencing, it could not be discriminated against uridine upon differential analysis performed with CoverageAnalyzer. Some possible explanations for these effects were already discussed in the corresponding section 5.3.2.

General considerations for RNA labeling

An important observation is that the stereochemical environment of an RNA molecule plays an essential role in the chemical labeling of nucleosides. This is relevant not only in labeling experiments for detection of modified nucleosides, e.g. pseudouridine, inosine, 5-methylcytidine, but also in structure-probing experiments. We have already encountered two *N*-1-methyladenosine positions in the same 25S yeast rRNA molecule with similar modification occupancy, but substantial difference in reverse transcriptase behavior. This shows that extreme caution should be taken upon interpretation of results coming from a chemical treatment in combination with RT-dependent sequencing approaches.

Single nucleoside sequencers offer a possibility to circumvent such problems. Of special interest are the Nanopore methods that do not rely on an enzymatic step. As already mentioned in the introductory section 1.4.5, the read-out of a Nanopore sequencer is typically an amplitude and duration of transient current blockage between translocation events. A drawback is that this two-dimensional read-out may be insufficient to allow the discrimination of all modifications known to date. Nevertheless, it was recently reported that the Oxford Technologies Nanopore sequencer is capable of discriminating between *N*-6-methyladenosine and adenosine [113].

Recently developed sequencing-based methods emerge as effective tools in the detection of modified nucleosides. Especially the high-throughput methods that allow detection and localization of modifications in low abundant RNA gain special interest.

Nevertheless, several important aspects need to be considered:

- For a given RNA modification employing different, possibly complementary methods to increase the probability of correct detection.
- Performing experiments in different laboratories in order to prove their robustness.
- If possible, validation of the existence of the target modification by applying appropriate biochemical and biophysical methods, e.g. LC-MS methods of isolated and purified RNA species, SCARLET methods, or primer extension methods.

In conclusion, the results presented in this thesis are important in the field of RNA detection. Reverse transcriptase events have proven effective in determination of given modified nucleosides, as was demonstrated for *N*-1-methyladenosine. Additionally, the chemical reagent osmium tetroxide-bipyridine was used with RNA for the first time in the context of selective labeling of 5-methylpyrimidines and its possible application in the detection of 5-methylcytidine was demonstrated.

7 References

References

- [1] P. A. Levene and J. A. Mandel, “Über die Konstitution der Thymo-nucleinsäure,” *Berichte der deutschen chemischen Gesellschaft*, vol. 41, pp. 1905–1909, may 1908.
- [2] P. A. Levene and W. A. Jacobs, “Über die Hefe-Nucleinsäure,” *Berichte der deutschen chemischen Gesellschaft*, vol. 42, pp. 2474–2478, apr 1909.
- [3] M. A. Machnicka, K. Milanowska, O. Osman Oglou, E. Purta, M. Kurkowska, A. Olchowik, W. Januszewski, S. Kalinowski, S. Dunin-Horkawicz, K. M. Rother, M. Helm, J. M. Bujnicki, and H. Grosjean, “MODOMICS: a database of RNA modification pathways–2013 update.,” *Nucleic acids research*, vol. 41, pp. D262–7, jan 2013.
- [4] D. Veneziano, S. Di Bella, G. Nigita, A. Laganà, A. Ferro, and C. M. Croce, “Non-coding RNA: Current Deep Sequencing Data Analysis Approaches and Challenges.,” *Human mutation*, pp. 1–57, 2016.
- [5] Y. Saletore, K. Meyer, J. Korfach, I. D. Vilfan, S. Jaffrey, and C. E. Mason, “The birth of the Epitranscriptome: deciphering the function of RNA modifications.,” *Genome biology*, vol. 13, no. 10, p. 175, 2012.
- [6] F. Crick, “Central dogma of molecular biology.,” *Nature*, vol. 227, pp. 561–3, aug 1970.
- [7] B. Lewin, “The mystique of epigenetics.,” *Cell*, vol. 93, pp. 301–3, may 1998.
- [8] Y. Suzuki, A. Noma, T. Suzuki, M. Senda, T. Senda, R. Ishitani, and O. Nureki, “Crystal structure of the radical SAM enzyme catalyzing tricyclic modified base formation in tRNA.,” *Journal of molecular biology*, vol. 372, pp. 1204–14, oct 2007.
- [9] Y. Motorin and M. Helm, “tRNA stabilization by modified nucleotides.,” *Biochemistry*, vol. 49, pp. 4934–44, jun 2010.
- [10] N. B. Leontis, J. Stombaugh, and E. Westhof, “The non-Watson-Crick base pairs and their associated isostericity matrices.,” *Nucleic acids research*, vol. 30, pp. 3497–531, aug 2002.
- [11] W. A. Cantara, P. F. Crain, J. Rozenski, J. A. McCloskey, K. A. Harris, X. Zhang, F. A. P. Vendeix, D. Fabris, and P. F. Agris, “The RNA Modification Database, RNAMDB: 2011 update.,” *Nucleic acids research*, vol. 39, pp. D195–201, jan 2011.
- [12] W. E. COHN, “Pseudouridine, a carbon-carbon linked ribonucleoside in ribonucleic acids: isolation, structure, and chemical characteristics.,” *The Journal of biological chemistry*, vol. 235, pp. 1488–98, may 1960.
- [13] E. Madore, C. Florentz, R. Giegé, S.-i. Sekine, S. Yokoyama, and J. Lapointe, “Effect of modified nucleotides on Escherichia coli tRNA Glu structure and on its aminoacylation by glutamyl-tRNA synthetase,” *European Journal of Biochemistry*, vol. 266, pp. 1128–1135, dec 1999.
- [14] P. B. Sigler, “An analysis of the structure of tRNA.,” *Annual review of biophysics and bioengineering*, vol. 4, no. 00, pp. 477–527, 1975.

- [15] S. Kadaba, A. Krueger, T. Trice, A. M. Krecic, A. G. Hinnebusch, and J. Anderson, “Nuclear surveillance and degradation of hypomodified initiator tRNA^{Met} in *S. cerevisiae*,” *Genes & development*, vol. 18, pp. 1227–40, jun 2004.
- [16] Z. Durdevic and M. Schaefer, “tRNA modifications: necessary for correct tRNA-derived fragments during the recovery from stress?,” *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 35, pp. 323–7, apr 2013.
- [17] M. Schaefer, T. Pollex, K. Hanna, F. Tuorto, M. Meusburger, M. Helm, and F. Lyko, “RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage,” *Genes & development*, vol. 24, pp. 1590–5, aug 2010.
- [18] N. Manickam, K. Joshi, M. J. Bhatt, and P. J. Farabaugh, “Effects of tRNA modification on translational accuracy depend on intrinsic codon-anticodon strength,” *Nucleic acids research*, vol. 44, pp. 1871–81, feb 2016.
- [19] T. Karlsborn, H. Tükenmez, C. Chen, and A. S. Byström, “Familial dysautonomia (FD) patients have reduced levels of the modified wobble nucleoside mcm(5)s(2)U in tRNA,” *Biochemical and biophysical research communications*, vol. 454, pp. 441–5, nov 2014.
- [20] A. G. Torres, E. Batlle, and L. Ribas de Pouplana, “Role of tRNA modifications in human diseases,” *Trends in molecular medicine*, vol. 20, pp. 306–14, jun 2014.
- [21] P. F. Agris, F. A. P. Vendeix, and W. D. Graham, “tRNA’s wobble decoding of the genome: 40 years of modification,” *Journal of molecular biology*, vol. 366, pp. 1–13, feb 2007.
- [22] M. Duechler, G. Leszczyńska, E. Sochacka, and B. Nawrot, “Nucleoside modifications in the regulation of gene expression: focus on tRNA,” *Cellular and Molecular Life Sciences*, vol. 73, pp. 3075–3095, aug 2016.
- [23] T. Suzuki and T. Suzuki, “A complete landscape of post-transcriptional modifications in mammalian mitochondrial tRNAs,” *Nucleic acids research*, vol. 42, pp. 7346–57, jun 2014.
- [24] J. Moriya, T. Yokogawa, K. Wakita, T. Ueda, K. Nishikawa, P. F. Crain, T. Hashizume, S. C. Pomerantz, and J. A. McCloskey, “A Novel Modified Nucleoside Found at the First Position of the Anticodon of Methionine tRNA from Bovine Liver Mitochondria,” *Biochemistry*, vol. 33, pp. 2234–2239, mar 1994.
- [25] C. I. Jones, A. C. Spencer, J. L. Hsu, L. L. Spremulli, S. A. Martinis, M. DeRider, and P. F. Agris, “A counterintuitive Mg²⁺-dependent and modification-assisted functional folding of mitochondrial tRNAs,” *Journal of molecular biology*, vol. 362, pp. 771–86, sep 2006.
- [26] M. Helm, R. Giegé, and C. Florentz, “A Watson-Crick base-pair-disrupting methyl group (m1A9) is sufficient for cloverleaf folding of human mitochondrial tRNA^{Lys},” *Biochemistry*, vol. 38, pp. 13338–46, oct 1999.
- [27] W. A. Decatur and M. J. Fournier, “rRNA modifications and ribosome function,” *Trends in biochemical sciences*, vol. 27, pp. 344–51, jul 2002.

- [28] T. Kiss, “NEW EMBO MEMBER’S REVIEW: Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs,” *The EMBO Journal*, vol. 20, pp. 3617–3622, jul 2001.
- [29] S. Raychaudhuri, L. Niu, J. Conrad, B. G. Lane, and J. Ofengand, “Functional effect of deletion and mutation of the Escherichia coli ribosomal RNA and tRNA pseudouridine synthase RluA.,” *The Journal of biological chemistry*, vol. 274, pp. 18880–6, jul 1999.
- [30] D. L. J. Lafontaine, “Noncoding RNAs in eukaryotic ribosome biogenesis and function.,” *Nature structural & molecular biology*, vol. 22, pp. 11–9, jan 2015.
- [31] S. Sharma and D. L. J. Lafontaine, “‘View From A Bridge’: A New Perspective on Eukaryotic rRNA Base Modification.,” *Trends in biochemical sciences*, vol. 40, pp. 560–75, oct 2015.
- [32] A. J. Shatkin, “Capping of eucaryotic mRNAs.,” *Cell*, vol. 9, pp. 645–53, dec 1976.
- [33] P. Fechter and G. G. Brownlee, “Recognition of mRNA cap structures by viral and cellular proteins.,” *The Journal of general virology*, vol. 86, pp. 1239–49, may 2005.
- [34] D. C. Chang, L. T. Hoang, A. N. Mohamed Naim, H. Dong, M. J. Schreiber, M. L. Hibberd, M. J. A. Tan, and P.-Y. Shi, “Evasion of early innate immune response by 2’-O-methylation of dengue genomic RNA,” *Virology*, vol. 499, pp. 259–266, dec 2016.
- [35] J. Jia, P. Yao, A. Arif, and P. L. Fox, “Regulation and dysregulation of 3’UTR-mediated translational control.,” *Current opinion in genetics & development*, vol. 23, pp. 29–34, feb 2013.
- [36] R. Perry and D. Kelley, “Existence of methylated messenger RNA in mouse L cells,” *Cell*, vol. 1, pp. 37–42, jan 1974.
- [37] D. Dominissini, S. Moshitch-Moshkovitz, S. Schwartz, M. Salmon-Divon, L. Ungar, S. Osenberg, K. Cesarkas, J. Jacob-Hirsch, N. Amariglio, M. Kupiec, R. Sorek, and G. Rechavi, “Topology of the human and mouse m6A RNA methylomes revealed by m6A-seq.,” *Nature*, vol. 485, pp. 201–6, may 2012.
- [38] B. Linder, A. V. Grozhik, A. O. Olarerin-George, C. Meydan, C. E. Mason, and S. R. Jaffrey, “Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome.,” *Nature methods*, vol. 12, pp. 767–72, aug 2015.
- [39] J. Zhou, J. Wan, X. Gao, X. Zhang, S. R. Jaffrey, and S.-B. Qian, “Dynamic m(6)A mRNA methylation directs translational control of heat shock response.,” *Nature*, vol. 526, pp. 591–4, oct 2015.
- [40] X. Wang, Z. Lu, A. Gomez, G. C. Hon, Y. Yue, D. Han, Y. Fu, M. Parisien, Q. Dai, G. Jia, B. Ren, T. Pan, and C. He, “N6-methyladenosine-dependent regulation of messenger RNA stability.,” *Nature*, vol. 505, pp. 117–20, jan 2014.
- [41] A. F. Lovejoy, D. P. Riordan, and P. O. Brown, “Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in *S. cerevisiae*.,” *PloS one*, vol. 9, no. 10, p. e110799, 2014.

- [42] T. M. Carlile, M. F. Rojas-Duran, B. Zinshteyn, H. Shin, K. M. Bartoli, and W. V. Gilbert, "Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells," *Nature*, vol. 515, no. 7525, pp. 143–6, 2014.
- [43] S. Schwartz, D. A. Bernstein, M. R. Mumbach, M. Jovanovic, R. H. Herbst, B. X. León-Ricardo, J. M. Engreitz, M. Guttman, R. Satija, E. S. Lander, G. Fink, and A. Regev, "Transcriptome-wide mapping reveals widespread dynamic-regulated pseudouridylation of ncRNA and mRNA.," *Cell*, vol. 159, pp. 148–62, sep 2014.
- [44] W. V. Gilbert, T. A. Bell, and C. Schaening, "Messenger RNA modifications: Form, distribution, and function," *Science*, vol. 352, pp. 1408–1412, jun 2016.
- [45] T. P. Hoernes, A. Hüttenhofer, and M. D. Erlacher, "mRNA modifications: Dynamic regulators of gene expression?," *RNA biology*, vol. 6286, no. August, p. 0, 2016.
- [46] J. E. Squires, H. R. Patel, M. Nusch, T. Sibbritt, D. T. Humphreys, B. J. Parker, C. M. Suter, and T. Preiss, "Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA.," *Nucleic acids research*, vol. 40, pp. 5023–33, jun 2012.
- [47] D. Dominissini, S. Nachtergaele, S. Moshitch-Moshkovitz, E. Peer, N. Kol, M. S. Ben-Haim, Q. Dai, A. Di Segni, M. Salmon-Divon, W. C. Clark, G. Zheng, T. Pan, O. Solomon, E. Eyal, V. Hershkovitz, D. Han, L. C. Doré, N. Amariglio, G. Rechavi, and C. He, "The dynamic N(1)-methyladenosine methylome in eukaryotic messenger RNA.," *Nature*, vol. 530, pp. 441–6, feb 2016.
- [48] X. Li, X. Xiong, K. Wang, L. Wang, X. Shu, S. Ma, and C. Yi, "Transcriptome-wide mapping reveals reversible and dynamic N(1)-methyladenosine methylome.," *Nature chemical biology*, vol. 12, pp. 311–6, may 2016.
- [49] K. A. Hofmann, "Sauerstoff-übertragung durch Osmiumtetroxyd und Aktivierung von Chlorat-Lösungen," *Berichte der deutschen chemischen Gesellschaft*, pp. 3329–3336, 1912.
- [50] K. A. Hofmann, O. Ehrhart, and O. Schneider, "Aktivierung von Chloratlösungen durch Osmium. II. Mitteilung," *Berichte der deutschen chemischen Gesellschaft*, vol. 46, no. 2, pp. 1657–1668, 1913.
- [51] R. Criegee, "Osmiumsäure-ester als Zwischenprodukte bei Oxydationen," *Justus Liebigs Annalen der Chemie*, vol. 522, no. 1, 1936.
- [52] C. R., M. B., and W. H., "Zur Kenntnis der organischen Osmium-Verbindungen. II. Mitteilung," *Justus Liebigs Annalen der Chemie*, vol. 550, p. 99, 1942.
- [53] K. B. Sharpless, A. Y. Teranishi, and J. E. Backvall, "Chromyl chloride oxidations of olefins. Possible role of organometallic intermediates in the oxidations of olefins by oxo transition metal species," *Journal of the American Chemical Society*, vol. 99, pp. 3120–3128, apr 1977.
- [54] H. B. Henbest, W. R. Jackson, and B. C. G. Robb, "Electronic effects in the reactions of olefins with permanganate ion and with osmium tetroxide," *Journal of the Chemical Society B: Physical Organic*, no. 0, pp. 803–807, 1966.
- [55] M. Schroeder, "Osmium tetroxide cis hydroxylation of unsaturated substrates," *Chemical Reviews*, vol. 80, pp. 187–213, apr 1980.

- [56] F. B. Daniel and E. J. Behrman, "Reactions of osmium ligand complexes with nucleosides.," *Journal of the American Chemical Society*, vol. 97, pp. 7352–8, dec 1975.
- [57] F. B. Daniel and E. J. Behrman, "Osmium (VI) complexes of the 3', 5'-dinucleoside monophosphates, ApU and UpA.," *Biochemistry*, vol. 15, pp. 565–8, feb 1976.
- [58] L. Subbaraman, J. Subbaraman, and E. Behrman, "The reaction of osmium tetroxide-pyridine complexes with nucleic acid components," *Bioinorganic Chemistry*, vol. 1, pp. 35–55, jan 1971.
- [59] M. Beer, S. Stern, D. Carmalt, and K. H. Mohlhenrich, "Determination of Base Sequence in Nucleic Acids with the Electron Microscope. V. The Thymine-Specific Reactions of Osmium Tetroxide with Deoxyribonucleic Acid and Its Components *," *Biochemistry*, vol. 5, pp. 2283–2288, jul 1966.
- [60] C. H. Chang, M. Beer, and L. G. Marzilli, "Osmium-labeled polynucleotides. The reaction of osmium tetroxide with deoxyribonucleic acid and synthetic polynucleotides in the presence of tertiary nitrogen donor ligands.," *Biochemistry*, vol. 16, pp. 33–8, jan 1977.
- [61] C.-H. Chang, H. Ford, and E. Behrman, "Reactions of cytosine and 5-methylcytosine with osmium(VIII) reagents: Synthesis and deamination to uracil and thymine derivatives," *Inorganica Chimica Acta*, vol. 55, pp. 77–80, jan 1981.
- [62] G. C. Glikin, M. Vojtískova, L. Rena-Descalzi, E. Palecek, M. Vojtískova, L. Rena-Descalzi, and E. Paleček, "Osmium tetroxide: a new probe for site-specific distortions in supercoiled DNAs.," *Nucleic Acids Research*, vol. 12, pp. 1725–1736, feb 1984.
- [63] E. Palecek, E. Rasovská, and P. Boublíková, "Probing of DNA polymorphic structure in the cell with osmium tetroxide.," *Biochemical and biophysical research communications*, vol. 150, pp. 731–8, jan 1988.
- [64] E. Palecek, P. Boublíková, and K. Nejedlý, "Probing of DNA structure with osmium tetroxide. Effect of ligands.," *Biophysical chemistry*, vol. 34, pp. 63–8, sep 1989.
- [65] A. Kanavarioti, K. L. Greenman, M. Hamalainen, A. Jain, A. M. Johns, C. R. Melville, K. Kemmish, and W. Andregg, "Capillary electrophoretic separation-based approach to determine the labeling kinetics of oligodeoxynucleotides," *Electrophoresis*, vol. 33, no. 23, pp. 3529–3543, 2012.
- [66] A. Kanavarioti, "Osmylated DNA, a novel concept for sequencing DNA using nanopores.," *Nanotechnology*, vol. 26, p. 134003, mar 2015.
- [67] R. Y. Henley, A. G. Vazquez-Pagan, M. Johnson, A. Kanavarioti, and M. Wanunu, "Osmium-Based Pyrimidine Contrast Tags for Enhanced Nanopore-Based DNA Base Discrimination," *PLoS ONE*, vol. 10, no. 12, pp. 1–12, 2015.
- [68] Y. Ding and A. Kanavarioti, "Single pyrimidine discrimination during voltage-driven translocation of osmylated oligodeoxynucleotides via the α -hemolysin nanopore," *Beilstein Journal of Nanotechnology*, vol. 7, pp. 91–101, jan 2016.
- [69] K. Wrobel, C. Rodríguez Flores, Q. Chan, and K. Wrobel, "Ribonucleoside labeling with Os(VI): a methodological approach to evaluation of RNA methylation by HPLC-ICP-MS.," *Metallomics : integrated biometal science*, vol. 2, no. 2, pp. 140–146, 2010.

- [70] A. Okamoto, K. Tainaka, and T. Kamei, "Sequence-selective osmium oxidation of DNA: efficient distinction between 5-methylcytosine and cytosine.," *Organic & biomolecular chemistry*, vol. 4, pp. 1638–40, may 2006.
- [71] T. Umemoto and A. Okamoto, "Synthesis and characterization of the 5-methyl-2'-deoxycytidine glycol-dioxosmium-bipyridine ternary complex in DNA.," *Organic & biomolecular chemistry*, vol. 6, pp. 269–71, jan 2008.
- [72] K. Tanaka, K. Tainaka, T. Umemoto, A. Nomura, and A. Okamoto, "An osmium-DNA interstrand complex: application to facile DNA methylation analysis.," *Journal of the American Chemical Society*, vol. 129, pp. 14511–7, nov 2007.
- [73] K. Burton and W. Riley, "Selective degradation of thymidine and thymine deoxynucleotides," *Biochemical Journal*, vol. 98, pp. 70–77, jan 1966.
- [74] Y. Vaishnav, E. Holwitt, C. Swenberg, H. C. Lee, and L. S. Kan, "Synthesis and characterization of stereoisomers of 5,6-dihydro-5,6-dihydroxy-thymidine.," *Journal of biomolecular structure & dynamics*, vol. 8, pp. 935–51, apr 1991.
- [75] R. MARKHAM and J. D. SMITH, "Chromatographic studies of nucleic acids; a technique for the identification and estimation of purine and pyrimidine bases, nucleosides and related substances.," *The Biochemical journal*, vol. 45, no. 3, pp. 294–8, 1949.
- [76] F. F. DAVIS and F. W. ALLEN, "Ribonucleic acids from yeast which contain a fifth nucleotide.," *The Journal of biological chemistry*, vol. 227, pp. 907–15, aug 1957.
- [77] J. W. LITTLEFIELD and D. B. DUNN, "Natural Occurrence of Thymine and Three Methylated Adenine Bases in Several Ribonucleic Acids," *Nature*, vol. 181, pp. 254–255, jan 1958.
- [78] J. Smith and D. Dunn, "An additional sugar component of ribonucleic acids," *Biochimica et Biophysica Acta*, vol. 31, pp. 573–575, feb 1959.
- [79] U. Birkedal, M. Christensen-Dalsgaard, N. Krogh, R. Sabarinathan, J. Gorodkin, and H. Nielsen, "Profiling of ribose methylations in RNA by high-throughput sequencing.," *Angewandte Chemie (International ed. in English)*, vol. 54, pp. 451–5, jan 2015.
- [80] V. Marchand, F. Blanloeil-Oillo, M. Helm, and Y. Motorin, "Illumina-based RiboMethSeq approach for mapping of 2'-O-Me residues in RNA," *Nucleic Acids Research*, p. gkw547, jun 2016.
- [81] C. W. Gehrke, K. C. Kuo, and R. W. Zumwalt, "Chromatography of nucleosides," *Journal of Chromatography A*, vol. 188, pp. 129–147, jan 1980.
- [82] C. G. Edmonds, M. L. Vestal, and J. A. McCloskey, "Thermospray liquid chromatography-mass spectrometry of nucleosides and of enzymatic hydrolysates of nucleic acids," *Nucleic Acids Research*, vol. 13, no. 22, pp. 8197–8206, 1985.
- [83] S. Kellner, A. Ochel, K. Thüring, F. Spenkuch, J. Neumann, S. Sharma, K.-D. Entian, D. Schneider, and M. Helm, "Absolute and relative quantification of RNA modifications via biosynthetic isotopomers.," *Nucleic acids research*, vol. 42, p. e142, oct 2014.

- [84] J. A. Kowalak, S. C. Pomerantz, P. F. Crain, and J. A. McCloskey, "A novel method for the determination of posttranscriptional modification in RNA by mass spectrometry," *Nucleic Acids Research*, vol. 21, no. 19, pp. 4577–4585, 1993.
- [85] M. Hossain and P. A. Limbach, "Multiple endonucleases improve MALDI-MS signature digestion product detection of bacterial transfer RNAs," *Analytical and bioanalytical chemistry*, vol. 394, pp. 1125–35, jun 2009.
- [86] B. Addepalli and P. A. Limbach, "Mass Spectrometry-Based Quantification of Pseudouridine in RNA," *Journal of The American Society for Mass Spectrometry*, vol. 22, pp. 1363–1372, aug 2011.
- [87] F. Sanger, S. Nicklen, and a. R. Coulson, "DNA sequencing with chain-terminating inhibitors.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 5463–7, dec 1977.
- [88] F. Sanger and A. R. Coulson, "A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase.," *Journal of molecular biology*, vol. 94, pp. 441–8, may 1975.
- [89] L. M. Smith, S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, and L. E. Hood, "The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis.," *Nucleic acids research*, vol. 13, pp. 2399–412, apr 1985.
- [90] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, C. R. Connell, C. Heiner, S. B. Kent, and L. E. Hood, "Fluorescence detection in automated DNA sequence analysis.," *Nature*, vol. 321, no. 6071, pp. 674–9, 1986.
- [91] P. H. Hamlyn, G. G. Brownie, C. C. Cheng, M. J. Gait, and C. Milstein, "Complete sequence of constant and 3' noncoding regions of an immunoglobulin mRNA using the dideoxynucleotide method of RNA sequencing.," *Cell*, vol. 15, pp. 1067–75, nov 1978.
- [92] Z. Shabarova and A. Bogdanov, eds., *Advanced Organic Chemistry of Nucleic Acids*. Weinheim, Germany: Wiley-VCH Verlag GmbH, jun 1994.
- [93] a. M. Maxam and W. Gilbert, "A new method for sequencing DNA.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 74, pp. 560–4, feb 1977.
- [94] D. a. Peattie, "Direct chemical method for sequencing RNA.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, pp. 1760–4, apr 1979.
- [95] Y. Motorin, S. Muller, I. Behm-Ansmant, and C. Branlant, "Identification of modified residues in RNAs by reverse transcription-based methods.," *Methods in enzymology*, vol. 425, pp. 21–53, 2007.
- [96] M. J. Renda, J. D. Rosenblatt, E. Klimatcheva, L. M. Demeter, R. A. Bambara, and V. Planelles, "Mutation of the Methylated tRNA³Lys Residue A58 Disrupts Reverse Transcription and Inhibits Replication of Human Immunodeficiency Virus Type 1," *Journal of Virology*, vol. 75, pp. 9671–9678, oct 2001.

- [97] B. Wittig and S. Wittig, "Reverse transcription of tRNA," *Nucleic Acids Research*, vol. 5, pp. 1165–1178, apr 1978.
- [98] D. C. Youvan and J. E. Hearst, "Reverse transcriptase pauses at N²-methylguanine during in vitro transcription of Escherichia coli 16S ribosomal RNA.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 76, pp. 3751–4, aug 1979.
- [99] B. J. Reon and A. Dutta, "Biological Processes Discovered by High-Throughput Sequencing.," *The American journal of pathology*, vol. 186, pp. 722–32, apr 2016.
- [100] D. R. Bentley, N. A. Gormley, S. Balasubramanian, P. A. Baybayan, V. A. Benoit, H. P. Swerdlow, G. P. Smith, J. A. Bridgham, A. A. Brown, J. Milton, A. A. Bundu, C. G. Brown, P. A. Granieri, K. P. Hall, T. A. Huw Jones, D. J. Evers, C. L. Barnes, M. A. Laurent, J. A. Loch, H. R. Bignell, J. M. Boutell, M. A. Osborne, J. Bryant, M. A. M. E. Smith, R. J. Carter, R. Keira Cheetham, A. J. Cox, D. J. Ellis, M. R. Flatbush, S. J. Humphray, L. J. Irving, M. S. Karbelashvili, S. M. Kirk, H. Li, X. Liu, K. S. Maisinger, L. J. Murray, B. Obradovic, T. Ost, M. L. Parkinson, M. R. Pratt, I. M. J. Rasolonjatovo, M. T. Reed, R. Rigatti, C. Rodighiero, M. T. Ross, A. Sabot, S. V. Sankar, A. Scally, G. P. Schroth, M. A. M. E. Smith, V. P. Smith, A. Spiridou, P. E. Torrance, S. S. Tzonev, E. H. Vermaas, K. Walter, X. Wu, L. Zhang, M. D. Alam, C. Anastasi, I. C. Aniebo, D. M. D. Bailey, I. R. Bancarz, S. Banerjee, S. G. Barbour, K. F. Benson, C. Bevis, P. J. Black, A. Boodhun, J. S. Brennan, R. C. Brown, D. H. Buermann, J. C. Burrows, N. P. Carter, N. Castillo, M. Chiara E Catenazzi, S. Chang, R. Neil Cooley, N. R. Crake, O. O. Dada, K. D. Diakoumakos, B. Dominguez-Fernandez, D. J. Earnshaw, U. C. Egbujor, D. W. Elmore, S. S. Etchin, M. R. Ewan, M. Fedurco, L. J. Fraser, K. V. Fuentes Fajardo, W. Scott Furey, D. George, K. J. Gietzen, C. P. Goddard, G. S. Golda, D. E. Green, D. L. Gustafson, N. F. Hansen, K. Harnish, C. D. Haudenschild, N. I. Heyer, M. M. Hims, J. T. Ho, A. M. Horgan, K. Hoschler, S. Hurwitz, D. V. Ivanov, M. Q. Johnson, T. James, G.-D. Kang, T. H. Kerelska, A. D. Kersey, I. Khrebtukova, A. P. Kindwall, Z. Kingsbury, P. I. Kokko-Gonzales, A. Kumar, C. T. Lawley, S. E. Lee, X. Lee, A. K. Liao, M. Lok, S. Luo, R. M. Mammen, J. W. Martin, P. G. McCauley, P. McNitt, P. Mehta, K. W. Moon, J. W. Mullens, T. Newington, Z. Ning, B. Ling Ng, S. M. Novo, M. J. O'Neill, A. Osnowski, O. Ostadan, L. L. Paraschos, L. Pickering, A. C. A. C. Pike, A. C. A. C. Pike, D. Chris Pinkard, D. P. Pliskin, J. Podhasky, V. J. Quijano, C. Racz, V. H. Rae, S. R. Rawlings, A. Chiva Rodriguez, P. M. Roe, J. J. Rogers, M. C. Rogert Bacigalupo, N. Romanov, A. Romieu, R. K. Roth, N. J. Rourke, S. T. Ruediger, E. Rusman, R. M. Sanches-Kuiper, M. R. Schenker, J. M. Seoane, R. J. Shaw, M. K. Shiver, S. W. Short, N. L. Sizto, J. P. Sluis, J. Ernest Sohna Sohna, E. J. Spence, K. Stevens, N. Sutton, L. Szajkowski, C. L. Tregidgo, G. Turcatti, S. Vandevondele, Y. Verhovsky, S. M. Virk, S. Wakelin, G. C. Walcott, J. Wang, G. J. Worsley, J. Yan, L. Yau, M. Zuerlein, J. J. Rogers, J. C. Mullikin, M. E. Hurles, N. J. McCooke, J. S. West, F. L. Oaks, P. L. Lundberg, D. Klenerman, R. Durbin, A. J. Smith, M. Sun, J. E. Paciga, R. I. Feldman, Z. Q. Yuan, D. Coppola, You Yong Lu, S. A. Shelley, S. V. Nicosia, and J. Q. Cheng, "Accurate whole human genome sequencing using reversible terminator chemistry.," *Nature*, vol. 456, no. 7218, pp. 53–9, 2008.
- [101] W. Q. Tian and Y. A. Wang, "Mechanisms of Staudinger reactions within density functional theory.," *The Journal of organic chemistry*, vol. 69, pp. 4299–308, jun 2004.
- [102] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. a. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes,

- B. C. Godwin, W. He, S. Helgesen, C. H. Ho, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. a. Vogt, G. a. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg, "Genome sequencing in microfabricated high-density picolitre reactors.," *Nature*, vol. 437, pp. 376–80, sep 2005.
- [103] M. Ronaghi, S. Karamohamed, B. Pettersson, M. Uhlén, and P. Nyrén, "Real-Time DNA Sequencing Using Detection of Pyrophosphate Release," *Analytical Biochemistry*, vol. 242, pp. 84–89, nov 1996.
- [104] M. Ronaghi, "DNA SEQUENCING:A Sequencing Method Based on Real-Time Pyrophosphate," *Science*, vol. 281, pp. 363–365, jul 1998.
- [105] N. Rusk, "Torrents of sequence," *Nature Methods*, vol. 8, pp. 44–44, jan 2011.
- [106] A. Valouev, J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire, and S. M. Johnson, "A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning," *Genome Research*, vol. 18, pp. 1051–1063, jul 2008.
- [107] A. Rhoads and K. F. Au, "PacBio Sequencing and Its Applications," *Genomics, Proteomics & Bioinformatics*, vol. 13, pp. 278–289, oct 2015.
- [108] I. D. Vilfan, Y.-c. Tsai, T. A. Clark, J. Wegener, Q. Dai, C. Yi, T. Pan, S. W. Turner, and J. Korlach, "Analysis of RNA base modification and structural rearrangement by single-molecule real-time detection of reverse transcription," *Journal of Nanobiotechnology*, vol. 11, no. 1, p. 8, 2013.
- [109] Y. Feng, Y. Zhang, C. Ying, D. Wang, and C. Du, "Nanopore-based fourth-generation DNA sequencing technology," *Genomics, proteomics & bioinformatics*, vol. 13, pp. 4–16, feb 2015.
- [110] M. Ayub, D. Stoddart, and H. Bayley, "Nucleobase Recognition by Truncated α -Hemolysin Pores.," *ACS nano*, vol. 9, pp. 7895–903, aug 2015.
- [111] D. B. Wells, M. Belkin, J. Comer, and A. Aksimentiev, "Assessing graphene nanopores for sequencing DNA.," *Nano letters*, vol. 12, pp. 4117–23, aug 2012.
- [112] A. B. Farimani, K. Min, and N. R. Aluru, "DNA base detection using a single-layer MoS₂," *ACS nano*, vol. 8, pp. 7914–22, aug 2014.
- [113] D. R. Garalde, E. A. Snell, D. Jachimowicz, A. J. Heron, M. Bruce, J. Lloyd, A. Warland, N. Pantic, T. Admassu, J. Ciccone, S. Serra, J. Keenan, S. Martin, L. McNeill, J. Wallace, L. Jayasinghe, C. Wright, J. Blasco, B. Sipos, S. Young, S. Juul, J. Clarke, and D. J. Turner, "Highly parallel direct RNA sequencing on an array of nanopores," tech. rep., aug 2016.
- [114] N. T. Ingolia, G. A. Brar, S. Rouskin, A. M. McGeachy, and J. S. Weissman, "The ribosome profiling strategy for monitoring translation in vivo by deep sequencing of ribosome-protected mRNA fragments.," *Nature protocols*, vol. 7, pp. 1534–50, aug 2012.

- [115] A. Lecanda, B. S. Nilges, P. Sharma, D. D. Nedialkova, J. Schwarz, J. M. Vaquerizas, and S. A. Leidel, “Dual randomization of oligonucleotides to reduce the bias in ribosome-profiling libraries,” *Methods (San Diego, Calif.)*, vol. 107, pp. 89–97, sep 2016.
- [116] R. Hrdlickova, M. Toloue, and B. Tian, “RNA-Seq methods for transcriptome analysis,” *Wiley interdisciplinary reviews. RNA*, may 2016.
- [117] Illumina, “Sequencing Methods Review (A review of publications featuring Illumina Technology),” vol. 199, no. 27-28, p. 27, 2014.
- [118] K. Iida, H. Jin, and J.-K. Zhu, “Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in *Arabidopsis thaliana*,” *BMC Genomics*, vol. 10, no. C, p. 155, 2009.
- [119] H. A. Ebhardt, H. H. Tsang, D. C. Dai, Y. Liu, B. Bostan, and R. P. Fahlman, “Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications,” *Nucleic Acids Research*, vol. 37, no. 8, pp. 2461–2470, 2009.
- [120] S. Findeiss, D. Langenberger, P. F. Stadler, and S. Hoffmann, “Traces of post-transcriptional RNA modifications in deep sequencing data,” *Biological chemistry*, vol. 392, pp. 305–13, apr 2011.
- [121] P. Ryvkin, Y. Y. Leung, I. M. Silverman, M. Childress, O. Valladares, I. Dragomir, B. D. Gregory, and L.-S. Wang, “HAMR: high-throughput annotation of modified ribonucleotides,” *RNA (New York, N.Y.)*, vol. 19, pp. 1684–92, dec 2013.
- [122] R. W. Wagner, J. E. Smith, B. S. Cooperman, and K. Nishikura, “A double-stranded RNA unwinding activity introduces structural alterations by means of adenosine to inosine conversions in mammalian cells and *Xenopus* eggs,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, pp. 2647–51, apr 1989.
- [123] C. BASILIO, A. J. WAHBA, P. LENGYEL, J. F. SPEYER, and S. OCHOA, “Synthetic polynucleotides and the amino acid code. V,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 48, pp. 613–6, apr 1962.
- [124] J. B. Li, E. Y. Levanon, J.-K. Yoon, J. Aach, B. Xie, E. LeProust, K. Zhang, Y. Gao, and G. M. Church, “Genome-Wide Identification of Human RNA Editing Sites by Parallel DNA Capturing and Sequencing,” *Science*, vol. 324, pp. 1210–1213, may 2009.
- [125] M. Sakurai, T. Yano, H. Kawabata, H. Ueda, and T. Suzuki, “Inosine cyanoethylation identifies A-to-I RNA editing sites in the human transcriptome,” *Nature chemical biology*, vol. 6, pp. 733–40, oct 2010.
- [126] M. Yoshida and T. Ukita, “Modification of nucleosides and nucleotides. VII. Selective cyanoethylation of inosine and pseudouridine in yeast transfer ribonucleic acid,” *Biochimica et biophysica acta*, vol. 157, pp. 455–65, may 1968.
- [127] N. W. Ho and P. T. Gilham, “Reaction of pseudouridine and inosine with N-cyclohexyl-N'-beta-(4-methylmorpholinium)ethylcarbodiimide,” *Biochemistry*, vol. 10, pp. 3651–7, sep 1971.
- [128] A. Bakin and J. Ofengand, “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*, vol. 32, pp. 9754–62, sep 1993.

- [129] X. Li, P. Zhu, S. Ma, J. Song, J. Bai, F. Sun, and C. Yi, “Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome.,” *Nature chemical biology*, vol. 11, pp. 592–7, aug 2015.
- [130] H. Hayatsu, Y. Wataya, K. Kai, and S. Iida, “Reaction of sodium bisulfite with uracil, cytosine, and their derivatives.,” *Biochemistry*, vol. 9, pp. 2858–65, jul 1970.
- [131] R. Shapiro, R. E. Servis, and M. Welcher, “Reactions of Uracil and Cytosine Derivatives with Sodium Bisulfite,” *Journal of the American Chemical Society*, vol. 92, pp. 422–424, jan 1970.
- [132] M. Frommer, L. E. McDonald, D. S. Millar, C. M. Collis, F. Watt, G. W. Grigg, P. L. Molloy, and C. L. Paul, “A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, pp. 1827–31, mar 1992.
- [133] E. Olkhov-Mitsel and B. Bapat, “Strategies for discovery and validation of methylated and hydroxymethylated DNA biomarkers.,” *Cancer medicine*, vol. 1, pp. 237–60, oct 2012.
- [134] M. Schaefer, T. Pollex, K. Hanna, and F. Lyko, “RNA cytosine methylation analysis by bisulfite sequencing.,” *Nucleic acids research*, vol. 37, p. e12, feb 2009.
- [135] W. Gu, R. L. Hurto, A. K. Hopper, E. J. Grayhack, and E. M. Phizicky, “Depletion of *Saccharomyces cerevisiae* tRNA(His) guanylyltransferase Thg1p leads to uncharged tRNA^{His} with additional m(5)C.,” *Molecular and cellular biology*, vol. 25, pp. 8191–201, sep 2005.
- [136] S. Edelheit, S. Schwartz, M. R. Mumbach, O. Wurtzel, and R. Sorek, “Transcriptome-wide mapping of 5-methylcytidine RNA modifications in bacteria, archaea, and yeast reveals m5C within archaeal mRNAs.,” *PLoS genetics*, vol. 9, p. e1003602, jun 2013.
- [137] I. Behm-Ansmant, M. Helm, and Y. Motorin, “Use of Specific Chemical Reagents for Detection of Modified Nucleotides in RNA,” *Journal of Nucleic Acids*, vol. 2011, pp. 1–17, 2011.
- [138] V. Khoddami and B. R. Cairns, “Identification of direct targets and modified bases of RNA cytosine methyltransferases.,” *Nature biotechnology*, vol. 31, no. 5, pp. 458–464, 2013.
- [139] S. Hussain, A. A. Sajini, S. Blanco, S. Dietmann, P. Lombard, Y. Sugimoto, M. Paramor, J. G. Gleeson, D. T. Odom, J. Ule, and M. Frye, “NSun2-mediated cytosine-5 methylation of vault noncoding RNA determines its processing into regulatory small RNAs.,” *Cell reports*, vol. 4, pp. 255–61, jul 2013.
- [140] S. Hussain, J. Aleksic, S. Blanco, S. Dietmann, and M. Frye, “Characterizing 5-methylcytosine in the mammalian epitranscriptome.,” *Genome biology*, vol. 14, no. 11, p. 215, 2013.
- [141] Y. Niu, X. Zhao, Y.-S. Wu, M.-M. Li, X.-J. Wang, and Y.-G. Yang, “N6-methyl-adenosine (m6A) in RNA: an old modification with a novel epigenetic function.,” *Genomics, proteomics & bioinformatics*, vol. 11, pp. 8–17, feb 2013.

REFERENCES

- [142] K. D. Meyer, Y. Saletore, P. Zumbo, O. Elemento, C. E. Mason, and S. R. Jaffrey, "Comprehensive analysis of mRNA methylation reveals enrichment in 3' UTRs and near stop codons.," *Cell*, vol. 149, pp. 1635–46, jun 2012.
- [143] T. W. Munns, M. K. Liszewski, and H. F. Sims, "Characterization of antibodies specific for N 6 -methyladenosine and for 7-methylguanosine," *Biochemistry*, vol. 16, pp. 2163–2168, may 1977.
- [144] J. B. Macon and R. Wolfenden, "1-Methyladenosine. Dimroth rearrangement and reversible reduction," *Biochemistry*, vol. 7, pp. 3453–3458, oct 1968.
- [145] A. E. Cozen, E. Quartley, A. D. Holmes, E. Hrabeta-Robinson, E. M. Phizicky, and T. M. Lowe, "ARM-seq: AlkB-facilitated RNA methylation sequencing reveals a complex landscape of modified tRNA fragments.," *Nature methods*, vol. 12, pp. 879–84, sep 2015.
- [146] B. E. Maden, M. E. Corbett, P. A. Heeney, K. Pugh, and P. M. Ajuh, "Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA.," *Biochimie*, vol. 77, no. 1-2, pp. 22–9, 1995.
- [147] J. Aschenbrenner and A. Marx, "Direct and site-specific quantification of RNA 2'-O-methylation by PCR with an engineered DNA polymerase.," *Nucleic acids research*, vol. 44, pp. 3495–502, may 2016.
- [148] H. Cahova, M. L. Winz, K. Hofer, G. Nubel, and A. Jaschke, "NAD captureSeq indicates NAD as a bacterial cap for a subset of regulatory RNAs," *Nature*, vol. 519, no. 7543, pp. 374–377, 2015.
- [149] L. Tserovski, V. Marchand, R. Hauenschild, F. Blanloeil-Oillo, M. Helm, and Y. Motorin, "High-throughput sequencing for 1-methyladenosine (m(1)A) mapping in RNA.," *Methods (San Diego, Calif.)*, pp. 1–12, feb 2016.
- [150] J. P. Ballesta and E. Cundliffe, "Site-specific methylation of 16S rRNA caused by pct, a pactamycin resistance determinant from the producing organism, *Streptomyces pactum*." *Journal of bacteriology*, vol. 173, pp. 7213–8, nov 1991.
- [151] L. Lempereur, M. Nicoloso, N. Riehl, C. Ehresmann, B. Ehresmann, and J. P. Bachellerie, "Conformation of yeast 18S rRNA. Direct chemical probing of the 5' domain in ribosomal subunits and in deproteinized RNA by reverse transcriptase mapping of dimethyl sulfate-accessible.," *Nucleic acids research*, vol. 13, pp. 8339–57, dec 1985.
- [152] T. Waku, Y. Nakajima, W. Yokoyama, N. Nomura, K. Kako, A. Kobayashi, T. Shimizu, and A. Fukamizu, "NML-mediated rRNA base methylation links ribosomal subunit formation to cell proliferation in a p53-dependent manner.," *Journal of cell science*, vol. 129, pp. 2382–93, jun 2016.
- [153] R. Hauenschild, L. Tserovski, K. Schmid, K. Thüring, M.-L. Winz, S. Sharma, K.-D. Entian, L. Wacheul, D. L. J. Lafontaine, J. Anderson, J. Alfonzo, A. Hildebrandt, A. Jäschke, Y. Motorin, and M. Helm, "The reverse transcription signature of N-1-methyladenosine in RNA-Seq is sequence dependent.," *Nucleic acids research*, vol. 43, pp. 9950–64, nov 2015.
- [154] A. Okamoto, "DNA-osmium complexes: recent developments in the operative chemical analysis of DNA epigenetic modifications.," *ChemMedChem*, vol. 9, pp. 1958–65, sep 2014.

-
- [155] L. Tserovski and M. Helm, “Diastereoselectivity of 5-Methyluridine Osmylation Is Inverted inside an RNA Chain.,” *Bioconjugate chemistry*, p. 1, sep 2016.
- [156] R. Hauenschild, S. Werner, L. Tserovski, A. Hildebrandt, Y. Motorin, and M. Helm, “CoverageAnalyzer (CAn): A Tool for Inspection of Modification Signatures in RNA Sequencing Profiles,” *Biomolecules*, vol. 6, no. 4, p. 7, 2016.
- [157] K. Thüring, K. Schmid, P. Keller, and M. Helm, “Analysis of RNA modifications by liquid chromatography–tandem mass spectrometry,” *Methods*, pp. 1–9, mar 2016.
- [158] W. Chen, P. Feng, H. Tang, H. Ding, and H. Lin, “RAMPred: identifying the N1-methyladenosine sites in eukaryotic transcriptomes,” *Scientific Reports*, vol. 6, p. 31080, aug 2016.
- [159] G. Brownlee and E. Cartwright, “Rapid gel sequencing of RNA by primed synthesis with reverse transcriptase,” *Journal of Molecular Biology*, vol. 114, pp. 93–117, jul 1977.
- [160] K. Nishikura, “A-to-I editing of coding and non-coding RNAs by ADARs.,” *Nature reviews. Molecular cell biology*, vol. 17, pp. 83–96, feb 2016.
- [161] M. Nilsson, H. Malmgren, M. Samiotaki, M. Kwiatkowski, B. P. Chowdhary, and U. Landegren, “Padlock probes: circularizing oligonucleotides for localized DNA detection.,” *Science (New York, N.Y.)*, vol. 265, pp. 2085–8, sep 1994.
- [162] N. Liu, M. Parisien, Q. Dai, G. Zheng, C. He, and T. Pan, “Probing N6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA.,” *RNA (New York, N.Y.)*, vol. 19, pp. 1848–56, dec 2013.
- [163] Y. T. Yu, M. D. Shu, and J. A. Steitz, “A new method for detecting sites of 2'-O-methylation in RNA molecules.,” *RNA (New York, N.Y.)*, vol. 3, pp. 324–31, mar 1997.

8 Appendix

8.1 The plus and minus - sequencing method

This method is based on the original work of Sanger and Coulson on DNA sequencing [88], later adapted for RNA [159]. It involves the *in vitro* synthesis of complementary ^{32}P -labeled products that are then separated into two sets of systems. The first system, called the minus system, is based on the fact that polymerases stop synthesis at positions where a single deoxynucleoside triphosphate (dNTP) is absent. The second system, the plus system, exploits that Klenow enzymes in presence of a single dNTP degrade a DNA strand from its 3' end up to a residue corresponding to the one present in the mixture. This way, loading all eight samples on a denaturing PAGE (four plus systems for each dNTP missing in the mixture, and four minus systems with each single dNTP present in the mixture) allows the determination of a sequence of up to 150 nucleotides of RNA or DNA of interest.

8.2 Chemistry of the template-directed DNA synthesis

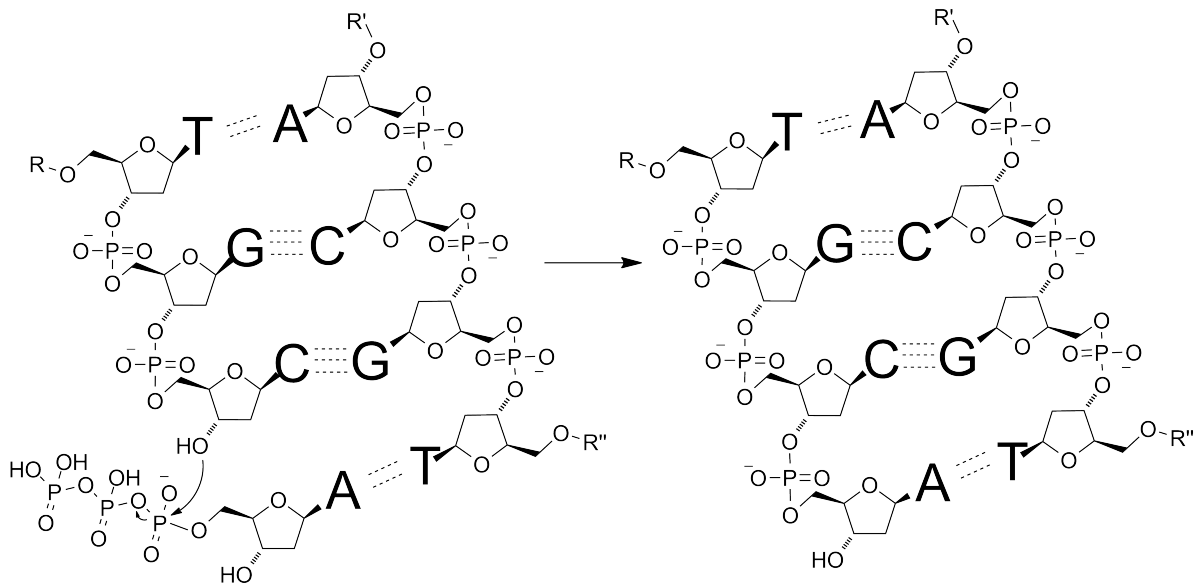


Figure A.1: Template-directed elongation of the DNA chain at its 3'-end. Attack of the 3'-OH of the DNA chain toward α -phosphate of dNTP, in this case deoxyadenosine triphosphate, leads to internucleotide linkage. This mechanism allows the radioactive labeling when $[\alpha\text{-}^{32}\text{P}]\text{dNTP}$ is used. Reaction is adapted from [92, p. 253].

8.3 Sequencing of RNA - nucleoside-specific reactions

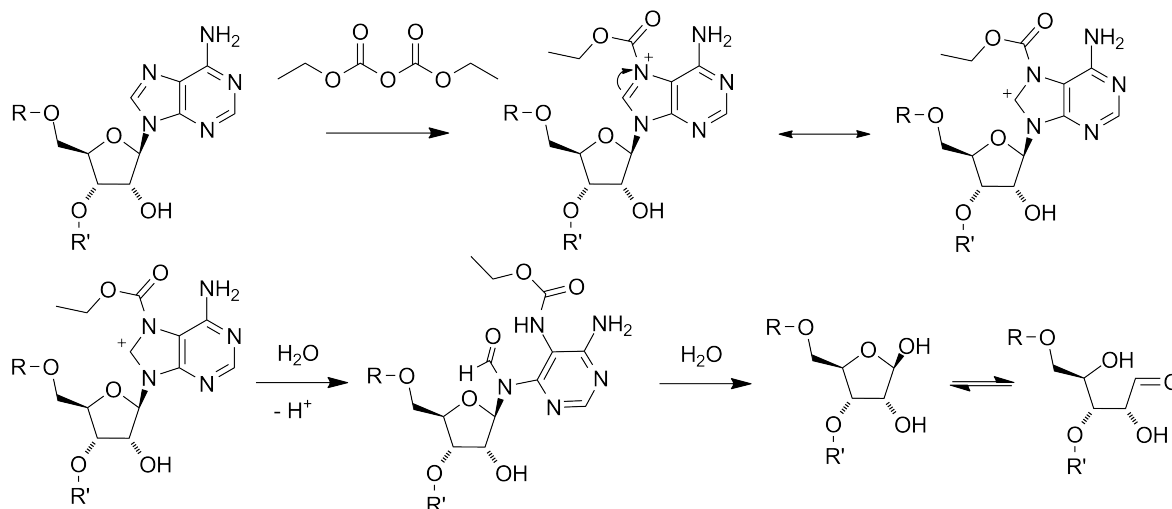


Figure A.2: Diethylpyrocarbonate (DEPC) reacts selectively with *N*-7 of purine nucleotides, which leads to a loss of aromaticity followed by a base cleavage. Reaction is shown for adenosine, adapted from [92, p. 270].

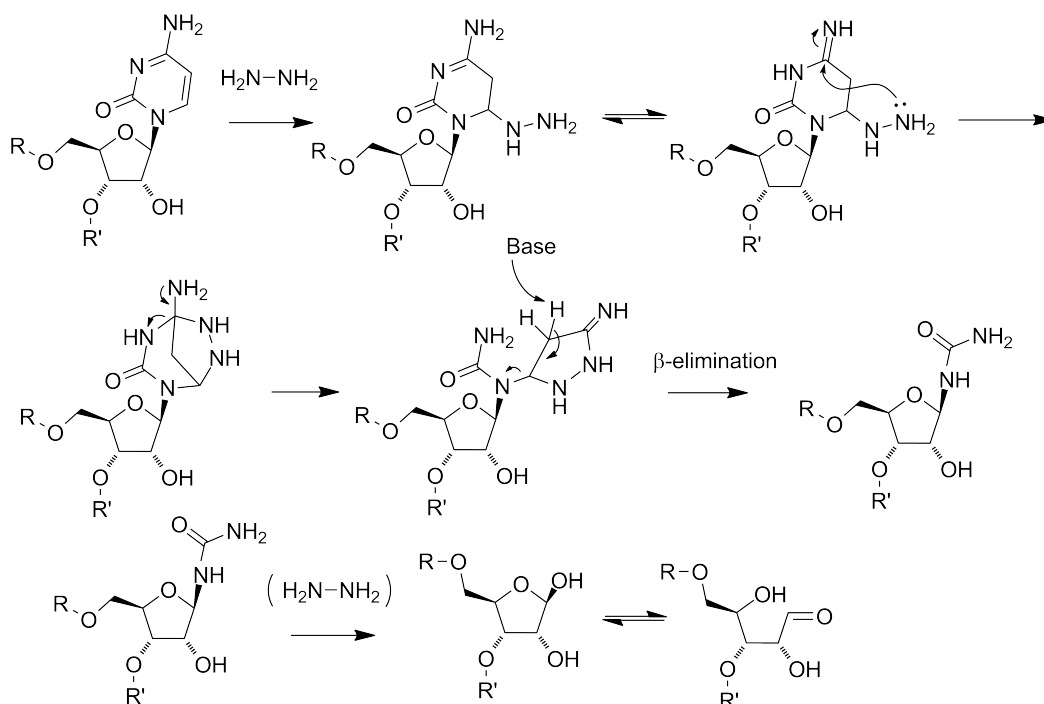


Figure A.3: Hydrazine attacks the 5,6 double bond of pyrimidine nucleotides. The reaction leads to a selective degradation of the nucleobase, allowing the site-specific hydrolysis of the RNA chain at those positions. Under appropriate choice of reaction conditions, either uridines or cytidines are reactive toward hydrazine. Reaction is shown for cytosine, adapted from [92, p. 237].

8.4 Inosine base-pairing

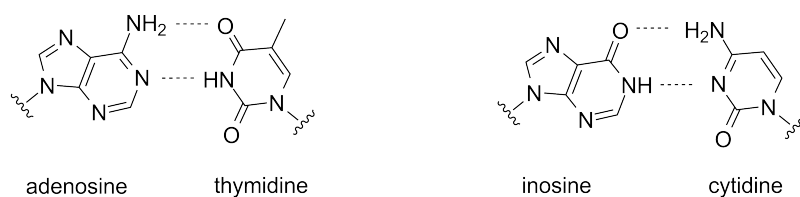


Figure A.4: Base-pairing properties of adenosine and inosine. While adenosine base-pairs with thymidine, its deaminated product inosine base-pairs better with cytidine. Figure adapted from [160].

8.5 Padlock probes

Padlock probes were first described by Nilsson and co-workers for the localized detection of DNA [161]. For this, an oligonucleotide was designed whose 3' end hybridizes to the 3' side of the questioned sequence and the 5' end to the 5' side. Then, a polymerase, dNTPs and a ligase are added which leads to a circularization of the padlock probe. Finally, a PCR step is performed and upon sequencing, the unknown region of the DNA can be determined.

8.6 SCARLET method

Site-specific Cleavage and radioactive-labeling followed by ligation-assisted extraction and TLC (SCARLET) was originally developed by Liu and colleagues for the specific detection of *N*-6-methyladenosine in low abundant RNA species such as mRNA and lncRNA [162]. The method is based on results obtained previously by Yu et al. [163]. An RNA molecule, hybridized to a chimeric oligonucleotide containing 2'-deoxynucleotides flanked by 2'-*O*-methyl ribonucleotides is cleaved at the position complementary to the first 2'-deoxynucleotide of the chimeric sequence. Therefore, for a cite of interest specific chimeras are designed and upon hybridization with the target RNA, RNase H treatment is performed. Cleaved RNA is then ³²P-radioactively labeled and with the assistance of a splint oligonucleotide ligated to a DNA of approximately 100 nucleotides length. Ligation product is then purified on a denaturing PAGE, excised and eluted. Finally, purified product is digested to nucleotides and modification status is evaluated by thin-layer chromatography. The method was also demonstrated effective for detection of pseudouridine by Li et al. [129] and is probably also applicable for other RNA modifications [162].

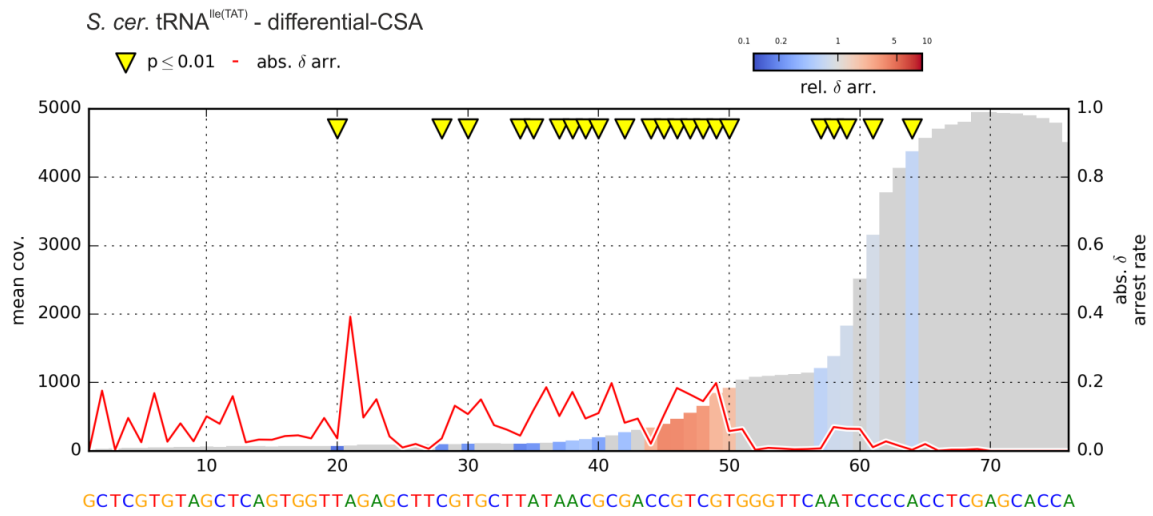
8.7 Os-bipy label of tRNA^{Ile(TAT)} - differential CSA analysis

Figure A.5: Differential Context Sensitive Arrest rate analysis of tRNA^{Ile(TAT)} before- and after 30 min Os-bipy treatment.

8.8 Proposed base-pairing properties of osmylated pyrimidines

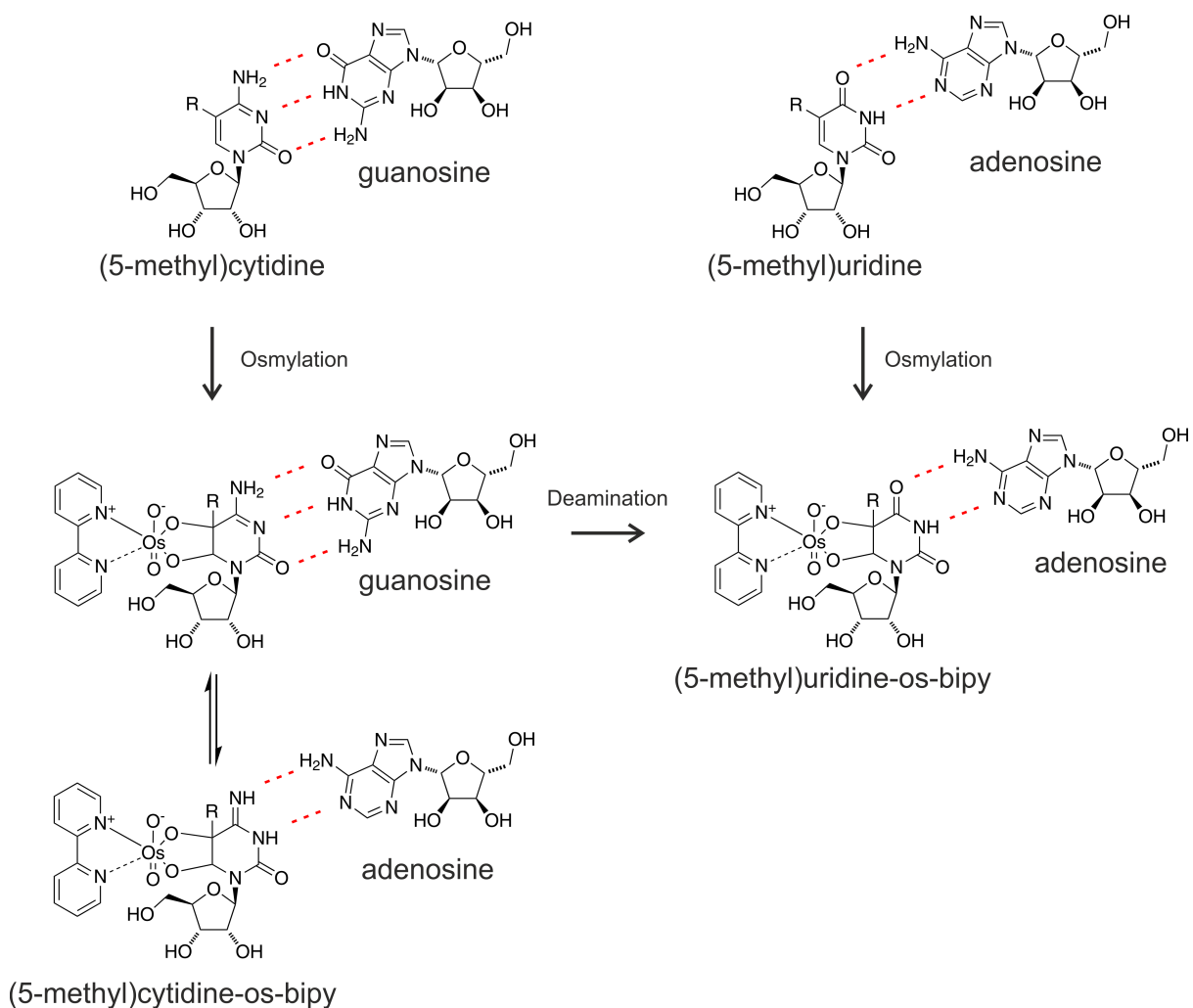


Figure A.6: Proposed base-pairing properties for pyrimidine nucleosides and their osmylated equivalents

Acknowledgments

Lebenslauf

Lyudmil Aleksandrov Tserovski

Ausbildung

Praktika/Nebentätigkeiten

Ort, Datum

Lyudmil Tserovski