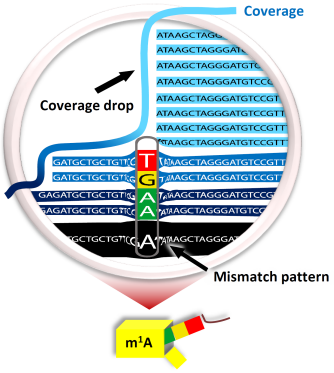
RNA-Seq and *CoverageAnalyzer* reveal sequence dependent reverse transcription signature of *N*-1-methyladenosine



Dissertation

zur Erlangung des Grades

Doktor der Naturwissenschaften

im Promotionsfach Pharmazie
am Fachbereich Chemie, Pharmazie und Geowissenschaften
der

Johannes Gutenberg-Universität Mainz

Ralf Hauenschild

geb. in Erlenbach am Main

Mainz, 11. Mai 2016

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

INTERNATIONAL PhD PROGRAMME MAINZ

Dekan: Prof. Dr. XXXX XXXXXXXXX

1. Berichterstatter: Prof. Dr. XXXX XXXX
2. Berichterstatter: Junior Prof. Dr. XXXXX XXXX
3. Berichterstatter: Prof. Dr. XXXXXX XXXXXXX

Datum der mündlichen Prüfung: 22. Juni 2016
Gesamtnote: *summa cum laude*

D77 (Dissertation Mainz)

Institutions:

Supervision:

Prof. Dr. XXXX XXXX
JGU, Institute of Pharmacy and Biochemistry

Thesis Advisory Committee:

Prof. Dr. XXXX XXXX
JGU, Institute of Pharmacy and Biochemistry
Prof. Dr. XXXXXXX XXXXXXXXXXX
JGU, Institute of Computer Science
Prof. Dr. XXXXXX XXXXXXX
JGU, Institute of Molecular Genetics

Examination Board:

Prof. Dr. XXXX XXXX
JGU, Institute of Pharmacy and Biochemistry
Prof. Dr. XXXXXXX XXXXXXXXXXX
JGU, Institute of Computer Science
Prof. Dr. XXXXXX XXXXXXX
JGU, Institute of Molecular Genetics
Junior Prof. Dr. XXXXX XXXX
JGU, Institute of Pharmacy and Biochemistry
PD Dr. XXXX XXXXXXX
JGU, Institute of Pharmacy and Biochemistry

Main collaborators:

XXXXXXX XXXXXXXXX, XXXXXXX XXXXXX
XXXXXXXXX XXXXXX, XXXXXXX XXXXXXX, Prof. Dr. XXXX XXXXXXX

Zusammenfassung

Die Entdeckung von Pseudouridin (Ψ) als fünftem Sequenzbaustein der RNA vor 60 Jahren gab den Auftakt zu einer fortlaufenden Erweiterung des bekannten Alphabets von Ribonukleinsäuren auf derzeit rund 150 verschiedene Nukleotid-Derivate. Kartierung und funktionelle Assoziation dieser Modifikationen sind die wesentlichen Schwerpunkte eines der aktuellsten und dynamischsten Gebiete der modernen Lebenswissenschaften, der Erforschung des Epitranskriptoms. Über den fortgeschrittenen Kenntnisstand im Bereich der dicht und systematisch modifizierten tRNAs und rRNAs hinaus gelangen während der letzten Jahre entscheidende Durchbrüche in der Kategorie kodierender Transkripte. Detektionsgrundlage ist ein modifikationsspezifisches Übersetzungsverhalten der Reversen Transkriptase (RT) bei der Abschrift von RNA zu cDNA, eine *RT Signatur*. Die Kombination von Next Generation Sequencing (NGS) mit spezifischem Labeling oder auch Immunopräzipitation offenbarte individuelle Modifikationslandschaften in mRNAs für z.B. Ψ, m^5C und m^6A, zum Teil mit Anhaltspunkten für regulatorische Bedeutung.
Diese Doktorarbeit befasste sich mit der Entwicklung bioinformatischer Methoden zur Beschreibung und Identifikation von Nukleotidmodifikationen anhand von Deep Sequencing-Daten. Das Konzept wurde durch die Charakterisierung der RT-Signatur von *N*-1-Methyladenosin (m^1A) demonstriert. Dieses an der Watson-Crick-Edge methylierte Adenosin kommt in tRNAs von Bakterien, Archaen und Eukaryoten vor und erregte mit seiner kürzlichen Entdeckung in zahlreichen Säuger-mRNAs Aufsehen. Während die in der Arbeit entwickelte Software auch den Vergleich von RT-Effekten nach differenzieller chemischer Behandlung erlaubt, erfolgte die Analyse von m^1A ausschließlich anhand nativer Signaturen, d.h. ohne spezifisches Labeling oder antikörperbasierte Anreicherung. Künstlich erzeugte m^1A-Instanzen sind in der Strukturaufklärung von RNAs von Interesse, bei der man den lokalen Methylierungserfolg als Lösemittelzugänglichkeit von Nukleotiden, d.h. als Strukturierungsgrad von RNA-Strängen interpretiert. Die Detektion basiert auf der Tendenz der Modifikation zur RT-Blockade, welche sich in der Gelelektrophorese oder in Sequenzierprofilen von Primer Extension-Assays als Häufung von Abbruchprodukten an der betreffenden Position äußert. Read-Through-Produkte wiederum weisen laut Studien ein bevorzugtes Verhältnis an missinkorporierten cDNA-Bausteinen an m^1A-Stellen auf.
Die somit duale RT-Signatur von m^1A, bestehend aus Abbruch- und Missinkorporationsraten, wurde durch die vorliegende Arbeit anhand natürlicher Instanzen in tRNA und rRNA charakterisiert und differenziert, zwecks verbesserter Auflösung und erweiterten Erkennungspotentials. Abbruch- und Read-Through-Produkte wurden durch ein spezialisiertes Protokoll zur Präparation sequenzierbereiter cDNA-Bibliotheken erfasst. Die digitale Analyse erfolgte durch Abgleich der Sequenzierdaten mit Referenzsequenzen. Kern des Workflows ist die eigenständige Software *CoverageAnalyzer*, entwickelt im Rahmen dieser Arbeit als universelle Plattform zur Prozessierung, Visualisierung und Filterung von Sequenzierprofilen nach Signaturmerkmalen. Damit wurden m^1A-Signaturen extrahiert und sodann durch deskriptive und inferentielle Statistik analysiert, auch auf Unterscheidbarkeit von un- oder anderweitig modifizierten Adenosinen auffälliger RT-Merkmale. Überwachtes Machine Learning mit Random Forest-Modellen zur Erkennung von m^1A in Adenosin-Pools, abgestuft nach Unterscheidungsschwierigkeit, gab Aufschluss über das Nutzungspotential acht formulierter Features, darunter ein kontext-sensitiver Deskriptor für RT-Stops. Es zeigte weiterhin den Vorteil simultaner Verwendung mismatch- und arrestbezogener Information und hob die Sonderstellung von m^1A unter nativen RT-Signaturen von Adenosinderivaten hervor, welche die sensitive und spezifische Detektion von m^1A erlaubt.
Erfolge in der Entdeckung unbekannter m^1A-Stellen in Mensch, Maus und *T. brucei* gelangen per Signaturabgleich und Sequenzhomologie. Mithilfe synthetischer Oligoribonukleotide wurde das Bild um Effekte unvollständiger Modifikationslevels verfeinert. Künstliche Instanzen bestätigten zudem ein Hauptergebnis der Studie: Die Mismatch-Zusammensetzung in m^1A's RT-Signatur ist abhängig vom Sequenzkontext, nämlich der Identität des 3'-gelegenen Nachbarnukleotids.
Die entwickelte Analysemethodik, spezialisierte Software sowie Erkenntnisse zur RT-Signatur von m^1A mit Implikationen für andere Modifikationen sind wegbereitend für Prüfungen bestehender Vorhersagen und den Ausbau der Kartierungsstrategie für das Epitranskriptom.

Summary

The discovery of pseudouridine (Ψ) as the fifth sequence residue of RNA 60 years ago marked the beginning of a successive extension of the known alphabet of ribonucleic acids up to currently around 150 different nucleotide derivatives. Mapping and functional association of these modifications are the essential emphases of one of the most topical and dynamic areas of modern life sciences, the exploration of the epitranscriptome. Beyond the advanced state of knowledge concerning the densely and systematically modified tRNAs and rRNAs, major breakthroughs were achieved in the class of coding transcripts during the last years. Basis for detection is a modification-specific behavior of Reverse Transcriptase (RT) in the transcription of RNA to cDNA, an *RT signature*. The combination of Next Generation Sequencing (NGS) with specific labeling or immunoprecipitation revealed individual modification landscapes in mRNA for e.g. Ψ, m^5C and m^6A, partially with evidence for regulatory relevance.

This PhD thesis addressed the development of bioinformatic methods for description and identification of nucleotide modifications based on Deep Sequencing data. The concept was demonstrated by the characterization of the RT signature of *N*-1-methyladenosine (m^1A). This adenosine residue, methylated at the Watson-Crick edge, occurs in tRNAs of bacteria, archea and eukarya, and called attention by its recent discovery in numerous mammalian mRNAs. Whereas the software developed in this project also allows comparison of RT effects after differential chemical treatment, analysis of m^1A relied on native signatures only, i.e. without specific labeling or antibody-mediated enrichment. Artificially induced m^1A instances are of interest in structural probing of RNA, wherein the local methylation efficiency is interpreted as the accessibility of nucleotides to the solvent, i.e. as the degree of structuring of RNA strands. The detection is based on the tendency of the modification to block RT, which is reflected by accumulation of abortive products at the respective position in gel electrophoresis or in sequencing profiles of primer extension assays. In turn, according to previous studies, read-through products exhibit a preferred composition of misincorporated cDNA residues at m^1A sites.

The hence dual RT signature of m^1A, consisting of arrest and misincorporation rates, was characterized and differentiated by the present work based on natural instances in tRNA and rRNA, for the purpose of improved resolution and enhanced recognition potential. Arrest and read-through products were captured by a specialized protocol for preparation of cDNA libraries ready for sequencing. The digital analysis was carried out by comparison of sequencing data to reference sequences. Core of the workflow is the standalone software *CoverageAnalyzer*, which was engineered in the scope of this work as a universal platform for processing, visualization and screening of sequencing profiles for signature features. In this way, m^1A signatures were extracted and then analyzed by descriptive and inferential statistics, also in terms of their capability of discrimination from non- or otherwise modified adenosines with noticeable RT features. Supervised machine learning with Random Forest models for recognition of m^1A in adenosine pools staggered by distinction difficulty shed light on usage potential of eight formulated features, including a context-sensitive descriptor of RT stops. Furthermore, it showed the benefit of simultaneous utilization of mismatch- and arrest related information and highlighted the special nature of m^1A among native RT signatures of adenosine derivatives, which allows the sensitive and specific detection of m^1A.

Achievements in discovery of unreported m^1A sites in human, mouse and *T. brucei* were made by signature comparison and sequence homology. With the help of synthetic oligoribonucleotides, the picture was refined by effects of incomplete levels of modification. Artificial instances moreover confirmed a central result of this study: the composition of mismatches in m^1A's RT signature depends on the sequence context, namely the identity of the 3'-adjacent nucleotide.

The developed analytical methodology, the specialized software as well as findings regarding m^1A's RT signature with implications for other modifications prepare the ground for revisal of existing predictions and for advancement of mapping strategies for the epitranscriptome.

Contents

List of Figures

List of Tables

Abbreviations

A, rA, dA	adenosine, riboadenosine, deoxyriboadenosine (monophosphate)
AUC	area under curve
BAM	Binary Sequence Alignment/Map format
bp	base pairs (sequence length)
C, rC, dC	cytidine, ribocytidine, deoxyribocytidine (monophosphate)
cDNA	copy DNA (formerly termed complementary DNA)
CSA	context sensitive arrest rate
DMS	dimethyl sulfate
dNTP	deoxyribonucleoside triphosphate
dTTP	deoxyribothymidine triphosphate
DNA	deoxyribonucleic acid
FDR	false discovery rate (1-PPV)
FNR	false negative rate (1-TPR)
FOR	false omission rate (1-NPV)
G, rG, dG	guanosine, riboguanosine, deoxyriboguanosine (monophosphate)
GUI	graphical user interface
I	inosine
kNN	k-Nearest Neighbor
LC-MS/MS	liquid chromatography - tandem mass spectrometry
m^1A	N-1-methyladenosine
m^1G	N-1-methylguanosine
$m^{2,2}G$	$N2$,$N2$-dimethylguanosine
m^3C	N-3-methylcytidine
m^5C	5-methylcytidine
m^6A	N-6-methyladenosine
$m^{6,6}A$	$N6$,$N6$-dimethyladenosine
mRNA	messenger RNA
MRM	multiple reaction monitoring
NPV	negative predictive value (1-FOR)
NGS	Next Generation Sequencing
nt	nucleotides (sequence length)
PAGE	polyacrylamide gel electrophoresis
PPV	positive predictive value (1-FDR)
Ψ, Psi	pseudouridine
RF	Random Forest
RNA	ribonucleic acid
ROC	receiver operating characteristic
rRNA	ribosomal ribonucleic acid
RT	reverse transcription, reverse transcriptase
s^4U	4-thiouridine
SAM	Sequence Alignment/Map format; S-adenosyl-L-methionine
Seq	sequencing
SNP/SNV	single nucleotide polymorphism/variant
ssRNA	single stranded RNA
T, dT	thymidine, deoxyribothymidine (monophosphate)
TPR	true positive rate (sensitivity, 1-FNR)
TNR	true negative rate (specificity, 1-FPR)
tRNA	transfer ribonucleic acid
QQQ	triple quadrupole
U, rU	uridine, ribouridine (monophosphate)
UTR	untranslated region

"Was macht die Mathe?"
—XXXXX XXXX

1 Introduction

1.1 Motivation and Background

Discovery and exploration of epigenetic mechanisms have drastically changed our understanding of the relationship between genotypes and phenotypes of cells beyond mere translation of invariant code. Besides DNA associated phenomena such as methylation, chromatin remodeling and histone modification, substantial contribution to gene expression and cell differentiation was revealed also in the world of RNA. A post-transcriptional regulatory machinery, including alternative splicing [1], alternative polyadenylation (APA) [2], RNA interference (RNAi) [3] and other mechanisms orchestrates the fate of molecular messages by qualitative and quantitative modulation on their way to ribosomal translation. Another crucial impact resides in RNA modifications, which occur in the form of more than 150 chemical derivatives [4] of the four canonical ribonucleotides rA, rG, rU and rC, being found in coding as well as in non-coding RNAs of all three phylogenetic domains of life. During RNA maturation, sequence residues are modified by numerous specific enzymes, such as methyltransferases [5]. Decades after the first discovery of a modified nucleotide, pseudouridine (Ψ), in 1957 [6], the interest in identification and especially localization of modification sites underwent a renaissance, driven by progressive evolution of analytical methods and increasing awareness of the functional role of such events. The *epitranscriptome* was born [7].

1.1.1 Role of RNA Modifications

Highly modified: tRNAs. Despite the ubiquitous occurrence of non-canonical nucleotides, the progress in understanding of biological functions is still in its infancy for many RNA modifications, and most success is achieved in positional mapping. Therein, tRNAs represent a special category, already being well-annotated [4, 8] with identities and positions of nucleotide modifications, in parts due to good experimental accessibility of this RNA species (∼15% of total RNA in rapidly growing mammalian cells [9]), but also due to tRNAs' conserved structural domains, decorated by modifications in exceptional density and diversity [10]. Closer inspection shows that mitochondrial tRNAs exhibit lower modification densities (9.5% and 7.5% of residues in yeast [11] and bovine [12] mtRNA positions) compared to cytosolic ones (16.4% of residues in yeast), suggesting an evolutionary trend towards intensely decorated tRNAs, analogous to observations made between archea, eubacteria and eukaryotes [13]. With a median of 8 modified residues per molecule in sequence lengths from 70 to 100 nt (cross-species compilation [11, 14]), cytosolic tRNAs are a popular system to study the roles of RNA modifications.

Structure and function. Similarly to rRNAs, where Ψ and 2'-O-methylations accumulate in functional and structural domains and are therefore involved in aspects of ribosome assembly and translation [15], also tRNA modifications can fulfill structural tasks. Correct formation of cloverleaf secondary structure and folding into L-like tertiary conformation is dependent on various modification types, e.g. Ψ at positions 32 and 39 having a critical role in shaping the anticodon stem loop [16]. While organisms living in hot environments tend to stabilize RNA structures by increased modification levels, higher abundance of dihydrouridine can be used to maintain nucleotide flexibility at cold temperatures [17]. Interestingly, s^4U_8 between the D-loop and the acceptor stem is highly conserved in prokaryotes and archea, and does not only stabilize [18, 19] tRNA fold, but is also described as a sensor for near-UV radiation [20, 21]. Another example of a functional effect is Gm_{18}, by which bacteria suppress activation of immune response [22, 23]. A more common fact is that modifications and editing (A→I) are involved in functional fine tuning of tRNAs, especially in the wobble position 34 at the 5' end of the anticodon (34-36) [10]. Thus, they are overall essential for accurate and efficient translation [24]. Strikingly, various modifications at position 37 can prevent frameshifts [25], such as m^1G_{37} [26]. In studies of a more medical angle, mutations in certain tRNA modification enzymes and corresponding lack of

modifications (e.g. m^5C and s^4U) were shown to be associated with several human diseases, such as intellectual disability, cancer and diverse mitochondria linked disorders (reviewed in *Torres et al. 2014* [27]). Regarding all the above structural and functional aspects, it is not surprising that hypomodified tRNAs are even targeted for degradation [28].

Non-coding and coding context. Also outside of tRNAs and rRNAs, the role of RNA modifications gained increased attention. A long-known example for non-coding context are U2 snRNAs, which are decorated by a 5'-trimethylguanosine cap, ten 2'-O-methylated residues and 13 pseudouridines, acquired (e.g. guided by C/D box) in the nucleus and essential for splicing function of U2 [29]. But also coding RNAs undergo modification, even though to a lower degree. Under a dozen of known mRNA modifications, the 5' cap with its variants is the most famous example, since it e.g. protects transcripts from exonuclease degradation [30]. Recently, an enormous interest in m^6A mapping has emerged, for example because of the finding that a fat mass- and obesity-associated gene encodes a specific demethylase (FTO) converting m^6A to adenosine [31]. If the enzyme is dysfunctional, significant alterations in metabolism are the consequence. FTO mutations have also been linked with higher risk for Alzheimer's disease and decrease in brain mass, too [32, 33]. These observations suggest potential physiological roles of m^6A in signaling as well as in neurodegeneration. Later studies described a dynamic m^6A epitranscriptome regulated by 'writer' and 'eraser' enzymes [34]. On the quest for effects of mRNA modifications in codon translation, recent analyses confirmed 'rewiring' of genetic code by occurrence of 2'-O-methylation, m^5C, m^6A and Ψ [35]. Another functional facet can be found in the context of therapeutic mRNAs designed to change cell fate, where m^5C and m^6A are of potential use to evade innate immune responses in transfected cells [36].

Recent momentum. Before one further prominent representative of RNA modifications, namely *N*-1-methyladenosine (m^1A), is introduced as the main subject of this work (see section 1.1.3), we present relevant techniques for identification of modified nucleotides in sequence context. In doing so, the current distributional and functional picture of some at present highly attended mRNA modifications such as Ψ, m^5C and m^6A is reviewed along with underlying cutting-edge techniques developed in a strong recent momentum of the field. Thereupon, we motivate the focus on m^1A by its wide-spread occurrence, diverse function and importantly eligibility for sequencing based detection.

1.1.2 Deep Sequencing based Detection

From confirmation to localization. Identification of modified nucleotides is feasible by several analytical techniques, including 2D Thin-Layer Chromatography (TLC), High Pressure Liquid Chromatography (HPLC) and Liquid Chromatography combined with Mass Spectrometry (LC-MS). These approaches exploit the physicochemical properties of a target analyte, the latter method with enormous sensitivity: tandem mass spectrometry (LC-MS/MS) has limits of quantifications (LOQ) in single-digit femtomolar and limits of detection (LOD) even in attomolar range [37]. However, information on location of modifications in sequence context is not accessible, with some exceptions where mass measurements from RNA fragments of a limited molecular pool can be unambiguously linked to reference sequences, allowing to reconstruct some sequential modification profiles [38]. Alternative methods use DNA chips prepared with specific oligonucleotides that distinguish modified from unmodified residues by differential hybridization efficiency [39], but they require prior knowledge of the modified sites and are thus not suitable for *de novo* detection.
Usage of reverse transcription arrest as a modification indicator is one of the key approaches developed in the past. The underlying idea is that bulky modifications prevent RT read-through, observable by an accumulation of truncated primer extension products, which is traditionally analyzed by polyacrylamide gel electrophoresis (PAGE) or capillary electrophoresis [40]. A dif-

ferent principle is applied in the case of inosine (I) resulting from A→I deamination. The modified residue is read as guanosine by reverse transcriptases (RT), and therefore reliably transcribed into cytidine in cDNA. This 'misreading' property allowed the first transcriptome-wide mapping of an RNA modification [41] even before the advent of methods that are nowadays subsumed as deep sequencing for RNA [42].

Combining treatments and sequencing. Modern approaches for localization of RNA modifications are still mostly based on the principle of primer extension by RT [43]. Whereas efficient quantitative sequencing of highly modified RNAs can require diligent methods (e.g. DM-TGIRT-seq [44]) for enzymatic demethylation of m^1G, m^1A and m^3C, the impeding effects on RT are a fundamental prerequisite and readout for modification detection. Therefore, a prevalent technique is the combination of specific chemical treatments and Next Generation Sequencing (NGS), in order to induce or enhance RT impeding effects at modified sites.
For instance, RT-blocking properties of CMC-labeled Ψ were exploited in studies that shed light on numerous occurrences of this modification in eukaryotic mRNAs, recently [45, 46, 47]. The key reagent N-**c**yclohexyl-N-(2-**m**orpholinoethyl)-**c**arbodiimide metho-*p*-**t**oluenesulfonate (CMCT) is applied prior to RT, such that NGS profiles indicate accumulated cDNA ends next to Ψ sites labeled by CMCT's bulky carbodiimide moiety (CMC). Additional click chemistry allows targeted pulldown of Ψ-featuring RNA molecules in an approach named N_3-CMC-enriched pseudouridine sequencing (CeU-Seq) [48]. Ψ is formed in a complex mechanism involving base detachment of uracil from the ribose, flipping and reattachment [49], catalyzed by either dyskerin Ψ synthase or a family of Ψ synthase (Pus) enzymes [50]. Occurring in tRNA and rRNA, as well as in small nuclear RNAs (snRNAs) and further noncoding RNAs (ncRNAs), it is the most abundant base modification in cells [50]. However, according to PSI-Seq, Pseudo-Seq and Ψ-Seq [45, 46, 47] this modification has much less non-redundant sites in mRNAs (only 100-400 in human cell lines) than e.g. has m^6A, and apparently features no regional preferences such as affinity to UTRs or coding sequences (CDS). Hence, the picture of functional roles of the fifth nucleotide in coding transcripts is still mostly limited to exemplary observations of half-life prolongation [51] and some vague speculations. To the latter belong possible amino acid substitution by Ψ occurring in open reading frames [45] or nonsense suppression by read-through at transcription termination sites [52]. In the near future, identification of transcripts with maximum Ψ stoichiometry, and mutagenesis of those residues, could reveal functional details [53].
An RT stop based example from the *RNA editing* field is cyanoethylation of inosines and subsequent differential comparison to sequencing profiles from untreated samples [54]. Conversely, alkaline hydrolysis forces RT stops at all but 2'-O-methylated sites by selective strand cleavage, detectable as depletion of sequence read ends at 2'-O-methylation sites [55]. A repertoire of further chemical reagents with exploitable modification specificity was reviewed in *Behm-Ansmant et al. 2011* [56].
Instead of RT stop, examples for transcriptome wide mappings of m^5C [57, 58, 59, 60] rely on information from successful read-through, and exploit the selective conversion of (5'-)unmethylated cytidines to uridines by bisulfite (HSO_3^-) ions. Replacement of Cs by Ts (Us) in reference sequences allows highlighting of the unaffected m^5C residues as C mismatches in mapping profiles. In this way, more than 10,000 m^5C sites could be identified in eukaryal mRNAs [60].

Real-time alternative. Recently established photonic nanostructures, so called zero-mode waveguides (ZMWs), are used for single-molecule resolved real-time sequencing (SMRT® technology) and allow identification of RNA base modifications by analysis of the kinetics of RTs, as e.g. demonstrated for m^6A [61]. Yet, this young technique is still under development and relatively cost intensive, while distinction power across modification species is unclear. Thus, most RNA-Seq projects still rely on conventional techniques, which separate the RT step from sequencing.

Enrichment. During the last five years, the field has experienced a contesting development of novel NGS based approaches for modification prediction on transcriptomal scale. The mentioned clickable CMC used for Ψ detection is only one example of targeted enrichment techniques for modified RNA molecules, a rising trend, which has completely revolutionized the research area in terms of throughput and knowledge about distribution and functional context of modifications. The major advantage is to bundle the available sequencing depth only at those molecules bearing the modification of interest with high confidence. To this class of approach belong m^6A-Seq [62, 63] and methyl-RNA-immunoprecipitation-sequencing (MeRIP-Seq) [64], which were independently developed for m^6A site detection in eukaryotic mRNA by antibody based enrichment of the respective RNA molecules. The earlier established photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) is a method to identify target sites of RNA-binding proteins in high resolution [65]. It incorporates photoactivatable ribonucleosides (s^4U or s^6G) into mRNA, which can covalently crosslink with nearby aromatic amino acids in RNA-binding proteins (or e.g. m^6A-bound antibodies) upon UV irradiation. Recently, PAR-CLIP was utilized to improve accuracy of MeRIP-Seq [66] and m^6A-Seq [67], yielding highly reliable m^6A annotations on transcriptomal scale. Although taking advantage of base transitions (occurring in RT-PCR) at crosslinked residues and sequence motifs in order to increase accuracy, these methods still generate unsharp probability peaks, since m^6A itself neither causes RT stop nor mismatch information. Finally, so called miCLIP advanced the idea by antibody-induced mismatch and truncation patterns, unique enough to call m^6A sites at single nucleotide resolution [68]. Improved reliability will contribute significantly to deeper insights into biological aspects of this most abundant mRNA modification [69]. Examples are dynamic enzymatic regulation [31, 34], tissue specific enrichment (as found for brain [64]) and regional preference in transcripts, namely near stop codons and in 3'UTRs, potentially influencing miRNA binding and mRNA half-life [64]. The principle to combine immunoprecipitation with sequencing was successfully adopted in mapping studies of other modifications, too. hMeRIP-seq, an analog of MeRIP-Seq was recently developed to reveal 5-hydroxymethylcytosine (hm^5C) distribution in the transcriptome of *D. melanogaster* [70]. In doing so, evidence on hydroxylation of m^5C to hm^5C, which is catalyzed by a Tet methyldioxygenase, was found prevalent in brain, where a knockout of the responsible enzyme caused impaired tissue development accompanied by low hm^5C levels. Another two recently published enrichment-based studies, undertaking transcriptome-wide detection of m^1A, are introduced in section 1.2.1, whereas section 1.2.2 sheds light on implied bioinformatic challenges in downstream analysis.

1.1.3 Prominence of m^1A

Subject of this work is *N*-1-methyladenosine (m^1A), an RNA modification eligible by threefold interest: i) structural and functional importance, ii) abundant and diverse natural occurrence and iii) amenability to sequencing based detection.

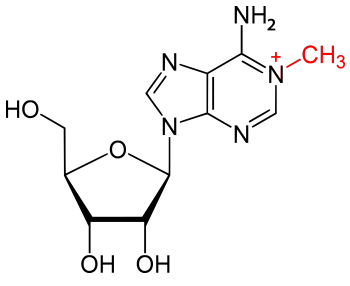


Figure 1: m^1A's chemical structure [4, 71].

Occurrence and Function. Shared by organisms from all three domains of life (eukaryotes, bacteria, archea) [5], m^1A is most famous for its conserved occurrence at position 58 in the TΨC-loop of many tRNA species. At this position it forms a reverse Hoogsteen base pair with s^2T_{54}, effecting a stabilization of the L-shaped tertiary tRNA structure [72], with e.g. increased heat tolerance of extremophilic bacteria as a documented consequence [73]. m^1A_{58} deficiency in yeast mutants lacking the responsible methyltransferase (*Trm*6, formerly Gcd10/Gcd14 complex) even has a lethal effect due to the modification's role in processing and stability of initiator tRNA [74]. In order to restore viability, degradation of the hypomodified $tRNA^{Ini}$ can be compensated by overexpression [75]. Another

conserved m^1A site in eukaryotic mtRNA, as well as prokaryotic and archeal tRNA is position 9 [4, 76]. Its importance was demonstrated for nematodes, which showed poor aminoacylation of tRNAs lacking the T-arm and m^1A_9 [77]. This is not overly surprising with respect to earlier findings of misfolded human mitochondrial $tRNA^{Lys}$ cloverleafs in absence of m^1A_9 [78]. Ensuing FRET based studies led to the conclusion that the methyl group controls a conformational equilibrium [79]. Furthermore, m^1A is common in eukaryotic [80, 81] and bacterial [82, 83, 4] rRNA, where in *Streptomyces pactum* it mediates antibiotic resistance [84]. Recent news are widespread m^1A occurrences in mRNA, which are dynamically regulated by stress stimuli like heat, starvation and oxidizing treatment, and show notable conservation across higher eukaryotic organisms, [85, 71]. The underlying approaches for enrichment and detection of modified transcripts were published by *Dominissini et al. 2016* and *Li et al. 2016*, and will be introduced in section 1.2 along with resultant findings.

Natural origin. In nature, formation of m^1A is catalyzed by site-specific methyltransferases (MTases), wherein *S*-adenosyl-L-methionine (AdoMet/SAM) acts as a methyl group donor being hydrolyzed to adenosine and homocysteine *via* S-adenosylhomocysteine [86]. Mutation experiments demonstrated that methyltransferases, such as *Trm*I responsible for m^1A formation in eubacterial tRNA, recognize their substrate regions by base identities, e.g. in the combination of aminoacyl stem, the variable region and the T-loop [87]. For archea, *Trm*I was shown to perform even bi-specific methylation of two adjacent adenosines 57 and 58 [88]. Another special case in human is a subcomplex of mitochondrial RNase P, reported to have a bifunctional methyltransferase activity responsible for m^1A_9 and m^1G_9 formation [89]. In the context of dynamic regulation of m^1A occurrence, also reversibility of the modification by ALKBH3 DNA/RNA demethylase was described [71].

Structural probing. Besides natural occurrences, also chemically induced m^1A sites are relevant. Analysis of 2D and 3D RNA structures is of utmost importance for understanding of functional aspects. In a popular method for structural probing, the reactivity towards the alkylating reagent dimethylsulfate (DMS) is interpreted as accessibility of nucleobase atoms to solvents, as observable e.g. in single-stranded loops [40]. While other DMS-induced methylations, such as m^3C and m^7G can be revealed by further treatments, artificial m^1A sites are detected by RT stops in primer extension [90, 56]. The readout are cDNA 3'-ends accumulating opposite to the RNA position 3'-adjacent to m^1A, since the modification causes the enzyme to stall. In contrast to later findings, the readout of DMS structural probing was, for decades, tacitly assumed quantitative, as if every encounter of the RT with m^1A led to arrest. Meanwhile, less base-specific probing methods have emerged: FragSeq and **p**arallel **a**nalysis of **R**NA **s**tructure (PARS) use structure-specific nucleases cleaving either single- or double-stranded RNAs [91, 92]. SHAPE-Seq instead, relies on 1-methyl-7-nitroisatoic anhydride (1M7) formation *via* **s**elective 2'-**h**ydroxyl **a**cylation **a**nalyzed by **p**rimer **e**xtension (SHAPE) combined with Deep-Seq for high-throughput *in vitro* probing [93, 94]. However, recent powerful DMS methods renew the actuality of m^1A using its pronounced RT blocking effect as readout for *in vivo* structural probing [95].

Misincorporation. While m^1A-mediated RT stop even has a biological role in HIV replication [96, 97], also misincorporations caused by this modification are of interest. Misincorporation has bearings in bypass of DNA lesions by Polymerases, which is essential during replication but can be error-prone, such that DNA modifications have mutagenic consequences [98]. Studies on mutagenic potential of m^1A in DNA, where T(*anti*)·A(*anti*) Watson-Crick base pairing is changed to a T(*anti*)·m^1A(*syn*) Hoogsteen base pair [99], allowed insights into m^1A flipping for AlkB mediated demethylation [100]. They found unexpected base pairing preferences of m^1A, an essential characteristic addressed in our work. A more detailed review of the state of knowledge about specific polymerase behavior towards m^1A and the challenge to utilize these properties to identify the modification is presented in the next section.

1.2 Status Quo and Challenges

Understanding of m^1A's RT effect has gradually improved, and so have the possibilities to utilize this knowledge. A major portion of challenges resides on the computational side, beginning with detection and extending to validation and downstream analysis towards biological aspects.

1.2.1 Picture of m^1A's RT Effects

Dual evidence. Early work on deep sequencing data [101, 102] has pointed out the fact that m^1A and other RNA modifications leave cryptic traces by simultaneously causing RT stops and misincorporation. Using different RNA-Seq protocols, several recent studies generated data featuring mismatch contents at positions known or postulated as m^1A sites [103, 104]. This clearly suggests a read-through capability of the RT encountering m^1A's altered Watson-Crick face, which results in unobtrusive traces in cDNA. More comprehensive studies of mismatch patterns revealed correlation between a given modification and the relative composition of misincorporated nucleotides [105]. However, their protocol for preparation of cDNA libraries for NGS was unsuited for simultaneous capture of mismatch and arrest information, which are the major constituents of what we would term m^1A's *RT signature*, a characteristic fingerprint allowing to detect the modification by computational analysis of NGS profiles. The fact that even recent work fails to fully resolve the dual error pattern, signalizes a lasting challenge in sequencing-based m^1A identification, only to be solved by specialized approaches, which can ideally be directly applied without tedious specific chemical labeling or enrichment.

Latest developments. While the lack of a fully resolved native RT signature of m^1A motivated realization of this PhD thesis, (see section 2), valuable insights into a transcriptome-wide m^1A landscape were published during the last months, wherein incomplete access to evidence on RT errors was compensated by antibody-mediated enrichment and differential treatment [85, 71]. Using their m^1A-ID-seq approach with random priming, *Li et al. 2016* [71] managed to identify $\sim$900 m^1A peaks in coding and noncoding RNA from 600 human genes. They abandoned to exploit mismatch information, pointing out the issue of sequencing depth, which is especially diminished at RT blocking m^1A sites. However, they claim massive enrichment of m^1A-bearing sequence regions by highly specific immunoprecipitation ($>$ 500-fold in case of m^1A_{1322} in 28S rRNA), recognizable as coverage peaks of mapped sequence reads obtained *via* RT and NGS. Confidence is further increased by comparison of peaks from native samples to those from replicates treated with *E. coli*'s AlkB demethylase, which converts m^1A to A. m^1A peaks generated *via* random priming typically show a central coverage trough, separating the reads originating from transcripts primed 3' (downstream) of the RT-blocking m^1A, from those reads of cDNAs primed 5' (upstream) of m^1A, while the latter fraction overlaps with read-through cases of the first fraction. Thus, superimposition of native m^1A peaks with those from demethylated (AlkB-treated) replicate RNA samples, allows calculation of a demethylation sensitivity (DS) score, indicating high-confidence m^1A regions, where m^1A-troughs disappear in favor of augmented read-through sequences. Finally, *Li et al. 2016* used Dimroth rearrangement as orthogonal method for confirmation of their findings.
This reaction converts m^1A to m^6A under alkaline conditions and was recently employed in the m^1A-seq (adapted from MeRIP-seq) approach by *Dominissini et al. 2016* [85] as primary confidence indicator of m^1A peaks from antibody-enriched RNA. Mismatch contents within sequence peaks indicate read-trough cDNAs at m^1A sites, if they disappear upon Dimroth rearrangement. While the protocol allowed only blurred resolution of the stop-component of m^1A's RT effect, the differential mismatch information provided sufficient evidence to call $\sim$7000 peaks, in $\sim$200 cases at single-nucleotide resolution. According to *Li et al. 2016*, m^1A/A ratio is 0.02% in mRNA, meaning an m^1A level of about 5-10% of the global m^6A level [71].

1.2.2 Computational Methods and Tools

Detection and evidence. Access to misincorporation or RT stop information as native or induced RT effects of modifications requires mapping of NGS reads to reference transcript sequences by alignment programs like Bowtie2 [106], tolerating mismatches to retain evidence on modified sites. Whether for prediction [45, 46, 47, 85, 71] or for characterization studies [101, 105], indicative parameters of annotated instances need to be extracted from the alignment profiles, in order to compile a collective representative recognition measure, e.g. referred to as MetaPsi by *Carlile et al. 2014* [45] in case of Ψ.
Few guidelines exist on read counts to be produced by sequencing platform in RNA-Seq, since this parameter highly depends on the investigated biological question, addressing small nucleotide polymorphisms (SNPs), differential expression (DE), splice variants or other phenomena (reviewed in *Sims et al. 2014* [107]). Also in RNA-Seq based modification detection, the decisive element of required sequencing depth is the expression level of the least abundant RNA species of interest, which can theoretically range from one to millions of copies per cell. By means of depletion or enrichment of certain over- or underrepresented RNA fractions, the total necessary read count can be lowered. However, definition of a local coverage cutoff for confidence, as needed for instance in modification site calling, is still up to the analyst and depends on the desired tradeoff between quality and quantity of predictions. Thanks to highly specific immunoprecipitation, *Dominissini et al. 2016* could afford *de novo* prediction of peaks based on a required minimum of only 20 reads and a mismatch content greater than 0.1 at 'reportable' m^1A sites. In contrast, analyses of m^1A's detailed RT characteristics based on confirmed sites need to ensure at least an order of magnitude more in coverage, if description of individual mismatch components in single-digit percentage precision is desired, even though the problem of site authenticity does not apply in such studies. The RT-stop effect of m^1A, entailing a coverage drop at the site of mismatch measurement aggravates the provision of sufficient sequencing depth. Nevertheless, the HAMR method, in which m^1A was analyzed in parallel to many other modifications, applied a minimum threshold of only ten reads per reference base, although actual coverages were not published [105].

Downstream analysis. Computational analysis of predicted sites by region-normalized superimposition of transcripts revealed an enrichment of m^1A in 5'-UTRs, particularly in structured domains in the vicinity of translation start codons (AUG) upstream of first splice sites [85, 71]. Conversely, m^1A is under-represented in 3'-UTRs in contrast to m^6A, which is associated with stop codons [62, 64], allowing for speculations on a potential complementing interplay of these marks in mRNA metabolism and translation [85].
While *Li et al. 2016* had to make use of combination of replicate samples only to narrow down peaks to ∼130 nt diameter [71], studies on m^6A achieved single-nucleotide resolution by bioinformatic prediction in peaks with a single clear summit based on the 'DRACH' sequence motif. Whereas success of this method is limited if motifs occur outside of peak centers or if multiple m^6A residues are clustered [64], *Linder et al. 2015* proceeded from MeRIP-Seq peaks of 100-200 nt in width to single-nucleotide resolution *via* mutational RT effects induced by the previously mentioned miCLIP method, instead of relying on motif search [68].
Further computational downstream analysis performed by *Li et al. 2016* based on Gene Ontology (GO) terms, revealed association of m^1A-bearing mRNAs with transcription-factor binding and RNA binding [71]. While widespread occurrence in mRNA could alter interactions with RNA-binding proteins by m^1A's positive charge under physiological conditions, findings in coding sequence (CDS) have even more striking bearings with regard to impact on translation. Nevertheless, expression was found positively correlated with overall methylation level [85].

Models. One can distinguish simple threshold-based methods from complex machine learning models, both applied to maximize discriminatory power in modification site calling. An example of the first kind is a logistic classifier function used by *Schwartz et al. 2014* [47] combining two positional readouts in Ψ-Seq, namely the relative read stop frequency (Ψ-ratio) in treated sam-

ples as independent component and the fold-change of Ψ-ratios upon CMCT treatment (Ψ-fc) as differential component. Another example of threshold-based technique is a peak score calculated from normalized differential (treated vs. untreated) counts of read stops within a sliding-window, as applied by *Carlile et al. 2014* [45]. These thresholds are then calibrated towards optimal discriminatory power on an annotated training landscape like rRNA in a supervised prediction scenario. Examples from the machine learning side are highly efficient black-box models such like Random Forests (RFs), as used for RNA/DNA Difference (RDD) detection [108], or Support Vector Machines (SVMs), as applied for uridine modification prediction in tRNA based on sequence context and structural data [109]. Such models are preferably used when descriptors become numerous or difficult to rate.

Evaluation. As a guideline, maximum comparability to various existing or future studies should be esteemed in prediction or characterization of RNA modifications. Hence, it was desirable, if RT effects like the one of m^1A were evaluated by both, descriptive and inferential statistics, also by means of machine learning models. Previous publications (discussed later) unfortunately provide only very limited information on distinction power of modification signatures and underlying reasons. By reporting sensitivities (= True Positive Rate, TPR) and specificities (= True Negative Rate, TNR), behavior of prediction models can be reflected from the perspective of defined modification landscapes based on the relative amount of achieved recognitions among existing target instances and successful rejections among the remainder respectively. Conversely, by Positive and Negative Predictive Values (PPV, NPV), reliability of predictions can be rendered transparent under specific settings for input data textures. Ideally, these measures are complemented by further discrete parameters, e.g. scenario-specific False Discovery Rate (FDR), which is the fraction of false positives among called modification sites and serves as reference point for expectable reliability. A continuous representation of model performance, preferably published along with *de novo* predicted modifications, is the Receiver Operating Characteristic (ROC), describing the selectivity of model-specific scores attributed to candidate sites by the corresponding tradeoff between TPR and False Positive Rate (FPR, = 1-TNR). ROC curves can be found e.g. in *Schwartz et al. 2014* [47] and *Carlile et al. 2014* [45] for evaluation of Ψ-prediction. However, the performance measures must be interpreted with caution, regarding the common problem of α-error accumulation by multiple hypothesis testing [110] observed in large-scale sequencing data analysis towards highly underrepresented events like RNA modifications. In fact, the requirements for prediction performance highly depend on the application scope. Using tRNAs and rRNAs as examples *Schaefer et al. 2009* [57] demonstrated that cytosine methylation can be reproducibly and quantitatively detected by bisulfite sequencing. In turn, for Ψ studies, reviews exposed poor cut-sets of predicted sites [111], in parts explainable by differences in read depth and stringency criteria [53].
Whereas the recent approaches to m^1A mapping can be considered milestones in terms of volume and claimed accuracy, they require a complex workup, including diligent immunoprecipitation with highly specific antibodies and ensuing application of differential treatment. Peaks detected by e.g. the MACS2 algorithm [85] need to be reinspected for differential information, namely demethylation sensitivity near troughs or high-mismatch sites, while only the latter allow calls in single-nucleotide resolution. Clearly, a remaining challenge is the validation of a vast amount of predictions. One promising way to achieve this task would be additional consultation of a characteristic, highly detailed description of m^1A's native RT effects, which should be recognizable at candidate positions. However, a hurdle is the introduced backward state of progress in characterization of such a fingerprint.

Big Data: quantity vs. quality. The availability of NGS has evoked a wide range of computer programs developed for analysis of sequencing profiles in manifold aspects, such as differential expression, regulation and variant calling. Like the SNP detection area, also the RNA modification field is challenged by a vast number of predicted candidates. While collection [4] and

curation [112] of the modification sites in databases is steadily promoted, experimental verification of predictions by independent methods must typically be restricted to a small subset of sites. Before engaging in effortful verification or subsequent in-depth investigations of biological aspects of given sites, the experimentalist needs to assess the significance of an identification event, and visually inspect typical features of RT effects.

Unmet software demands. In principal, a huge variety of so called alignment viewers like IGV, Tablet, Savant, UGENE and Persephone provides more or less detailed graphical representations of mapping results, typically resolving the base composition and orientation of reads covering a reference sequence. While powerful in navigation and sometimes rich in database-driven annotation tracks, these visualizers are missing essential functions for efficient and comprehensive characterization of RNA modifications' RT signatures. Options to reduce or organize sequence positions by signature-relevant feature thresholds are very limited. As becomes clear from the introduced state of the art, in-depth characterization of modification effects requires parallel access to both, graphical and numerical representation of a tailored set of parameters beyond the scope of standard viewers and variant detectors. Besides read end counts, a key aspect of primer extension based modification detection, which can typically neither be plotted nor exported by existing viewers, there is also a lack of options for context-sensitive rating of RT stop frequencies, and thus no possibility to account for potential regional tendencies in RNA sequences. Such lack of specialized functions in existing tools was decisive for the tasks and goals of this work, formulated in section 2.

1.3 Interdisciplinary Task Force

This work is part of a general research effort towards Deep-Seq based detection of RNA modifications, which is undertaken by a task force of PhD students working on interdigitating topics. Since modern Life Science projects (like the one for m^1A) require interdisciplinary skills, techniques and resources, close cooperation and diligent coordination within the group was essential. The major responsibilities resided in library preparation (Lyudmil Tserovski), RNA isolation & LC-MS/MS based quantification (Kathrin Thüring, Katharina Schmid) and bioinformatics (Ralf Hauenschild). The joint research progress benefited from mutual support across projects in accordance with individual priorities.

2 Goal of the Work

This PhD project aimed at the development of bioinformatic analysis methods for identification of RNA modifications based on RT signatures in Deep Sequencing data. As a proof of principle, the suitability of the corresponding approach was supposed to be demonstrated by application to m^1A, a modification species uniting important criteria: biological relevance, well annotated prevalence in experimentally accessible RNA species, and prior record of pronounced impact on RT behavior.

Based on the introduced state of knowledge about m^1A's typical RT stop and mismatch tendencies, one important goal was an advanced characterization of m^1A's native RT signature, i.e. a detailed description of the fingerprint this modification leaves in sequencing profiles without a need for specific chemical labeling. This task included the identification and formal definition of suitable, characteristic parameters and features, in order to assess the variability of the signature and to improve its resolution beyond the hitherto picture. Besides qualitative confirmation of m^1A sites and potential evidence based identification of previously unreported occurrences, it was also planned to estimate the extent, to which m^1A levels can be narrowed down by readouts from NGS profiles. While large-scale prediction of m^1A sites was beyond the scope of this work, a major objective was to evaluate the distinction power of m^1A's RT signature by means of machine learning. In this process, attention should be drawn to thorough statistical evaluation under different conditions and scenarios of m^1A instances intermixed with data points from other adenosines. Ideally, suitable models would also be used to shed light on importance of various signature features, and to identify other eligible modification types amenable to the developed method.

Major tasks of this project resided in both, elaboration of an experimental concept for the characterization of m^1A's signature as well as in the design of a tailored computational workflow. The requirements for an extensive analysis pipeline from raw NGS data to highly resolved RT signatures comprised versatile aspects. Major premises included efficiency in processing of sequencing profiles and compatibility with common sequence and alignment formats. Under compliance with flexibility and adaptability, the workflow should be designed not only towards requirable adjustment to and reflection of potential parametrical changes *in vivo* (modification levels and landscapes in RNA), *in vitro* (treatments, RT enzymes, library preparation settings) and *in silico* (sequencing properties, trimming, mapping), but also allow transfer to modification types other than m^1A. Aiming at a modular conception of the pipeline, optionality between different application scenarios was envisaged: (i) in-depth characterization of RT signatures by ensuing statistical analysis based on a defined feature set and facultatively supported by machine learning, (ii) utilization of established signatures for detection of unreported modification sites, and (iii) comparative quick-runs of analysis schemes under varied input conditions.

A particular goal was the engineering of a graphical user interface (GUI) software, for an ideally seamless connection of specialized visualization and directed numerical extraction of site-specific features relevant for modification detection based on RT signatures. Aiming at maximum synergy between scientific outcome of RT signature studies and according functional repertoire of the GUI software, the plan provided a development of the application in parallel to m^1A characterization. Thus, features identified as most characteristic for RT signatures should be taken into account in software conception. Provided such a program would prove substantial assistance in characterization of m^1A's RT signature and show potential beyond, re-engineering of the software in a distributable, platform independent format was envisaged, allowing for a continued standalone application in various endeavors of Deep-Seq based modification detection.

3 Results & Discussion

Overview. This chapter is divided into three sections according to a modular project conception in response to the defined goals of this work. Briefly, a computational workflow for detection of RNA modifications based on Deep-Seq profiles was developed and applied for characterization of m^1A's RT signature as a proof of principle. The core component of our analysis pipeline was conceived as a standalone graphical user interface (GUI) software, named *CoverageAnalyzer* (*CAn*). The project parts are presented in an order most conducive for understanding:

- Section 3.1 characterizes m^1A's RT signature and thereby demonstrates an application scenario of the workflow.
- Section 3.2 addresses the corresponding functional repertoire of the analysis interface *CoverageAnalyzer*, discussing its use for signature analysis.
- Section 3.3 finally provides a technical outline of the entire workflow for NGS-based modification detection, here used for examination of m^1A.

3.1 RT Signature of m^1A

The term *RT signature* of a particular ribonucleotide species universally entitles the behavior of reverse transcriptases towards this residue during cDNA synthesis. From a theoretical point of view, this would comprise enzymatic responses of all RT types, described by a set of differential equations based on a high-dimensional space of experimental settings (e.g. temperatures, dNTP concentrations, pH milieu etc.) as variables. For a practical application however, it is sufficient, to reduce this complex system to a small set of descriptive and robust features observed under one reproducible experimental setup for a conventional RT enzyme (see Methods section 5.2).

The following sections progressively present such a robust RT signature allowing for a sensitive and specific identification of m^1A in deep sequencing data.

3.1.1 Technical Approach

As common in RNA-Seq procedures, RNA preparations were reverse transcribed to cDNA libraries and submitted to sequencing on an NGS platform, here Illumina. According to the introduced literature, three types of cDNA products are expected in reverse transcription of m^1A sites in RNA (Fig. 2). Among the stop products (a), a substantial fraction originates from m^1A induced termination of RT after complementing its 3'-adjacent neighbor residue in the template RNA, as mentioned in the structural probing context. On the other hand, read-through products (b) and (c) contain either misincorporated (dA, dG, dC) or correct (dT) residues at the cDNA position corresponding to m^1A. Most RNA-Seq protocols can afford certain loss of sequence information from cDNA products not reaching the second primer binding site due to RT drop-off. Our study addresses m^1A sites mainly in tRNAs, where spontaneous RT arrests are accompanied by massive stop frequencies at roadblock modifications. Missing abortive products not only eliminates type (a), but also substantially decimates type (b) and (c) products, leading to low coverage and bad signature resolution.

Therefore, in this work, we employed a library preparation protocol suitable for simultaneous capture of both, abortive and full-length cDNA products. The steps were conducted by Lyudmil Tserovski. A key feature of the protocol is a primer ligation for second-strand cDNA synthesis rather at the cDNA level (at 3' end of first-strand cDNA products a, b, c) instead of at the RNA level, where the RT primer for first-strand cDNA synthesis is ligated (3' end of RNA) [114, 115].

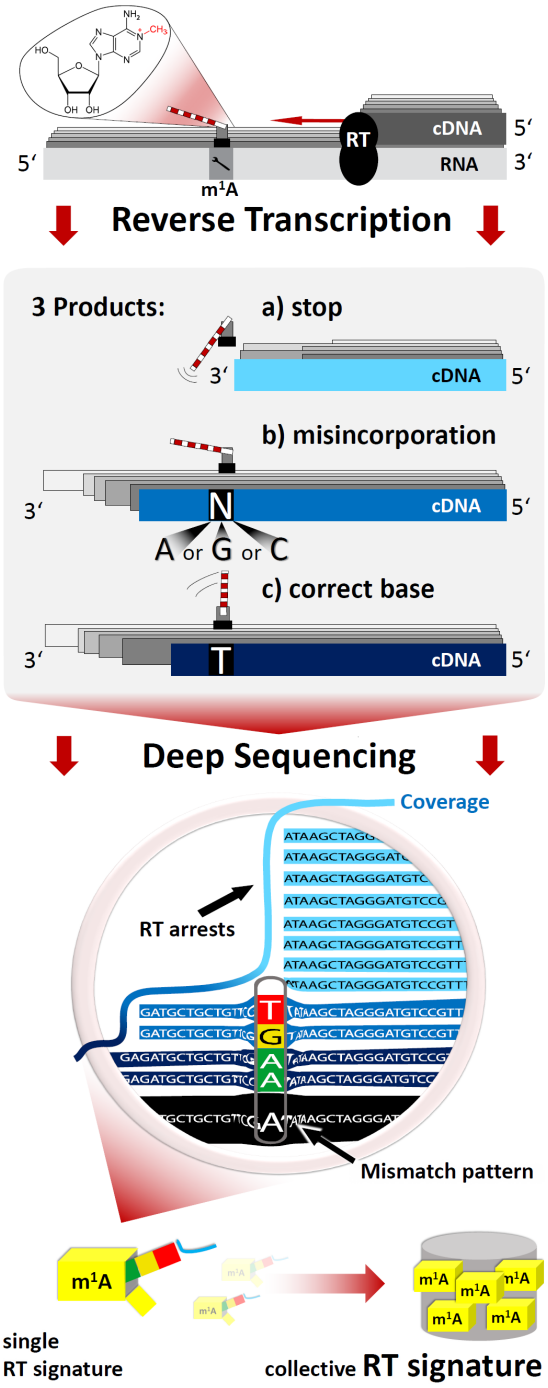


Figure 2: Principle of RNA-Seq data generation for the detection of m^1A residues. Adopted from *Hauenschild et al. 2015* [113].

Thus, in contrast to conventional methods, which are prone to biased amplification of RT products, the combined readout of modification-caused RT arrests and detailed mismatch information in NGS data could be utilized for extraction of single m^1A fingerprints to establish a collective RT signature. To obtain the best possible resolution of m^1A signatures, profiles of particularly high coverage were generated by increased sequencing depths (see Table 1) for targeted RNA pools. Further improvements of the protocol are the result of Lyudmil Tserovski's work within a greater research effort addressing Deep-Seq based RNA modification detection. The experimental details are specified in section 5.2, whereas the resulting sequence libraries are presented in Table 1. From the raw reads, several key parameters were scrutinized in post-processing

steps, and will be referred to throughout the Results section. Details on sequencing, trimming, mapping and signature extraction are provided in Materials and Methods sections 5.2-5.6. As illustrated in Fig. 2, the specified goal of a well-described RT signature of m^1A is pursued *via* collection of single-site observations. The latter followed a strategic concept, presented in the next section.

3.1.2 Characterization Strategy

For systematic characterization of m^1A's RT signature throughout this project, a powerful bioinformatic analysis platform, *CoverageAnalyzer*, was developed gradually, synchronized with its application to m^1A, allowing for insights into the corresponding features of the modification. This key component for targeted visual inspection of sequencing profiles and efficient access to statistical information by automated screening is presented in section 3.2. We characterized m^1A's RT signature based on RNA preparations chosen guided by a holistic concept, including examination of annotated m^1A sites, knockout examples as negative controls and other considerable presets to capture a comprehensive picture, such as fractional occupancy and variation of sequence context. By comparison of known m^1A positions in native tRNA and rRNA samples to those from null mutants, our analysis covered also signatures potentially influenced by strongly structured RNA domains. Quantitative analysis was complemented by studies of qualitative signal dependence on the local nucleotide environment. In this context, synthetic oligoribonucleotides were designed to assess the influence of the 3'-neighboring nucleotide of an m^1A site, which is the last conventionally reverse transcribed residue before the direct encounter of the RT's active site with the modification.

Once characterized, the signature was tested for identification of *bona-fide* m^1A sites, based on known modified positions. Indeed, co-occurrence of the described patterns in homologous sequence contexts of related organisms provided sufficient evidence for calling of unreported m^1A sites. Subsequent exemplary LC-MS/MS-based confirmation of m^1A's presence in an isolated tRNA demonstrated the viability of such predictions.

Major concerns are the choice of reference sequences and the corresponding mapping strategy. The previously published HAMR [105] method relied on definition of tRNA families from the ensemble of all genetic loci predicted by tRNAscan-SE [116]. In contrast, we rather preferred a well-annotated modification landscape of spliced and maturated sequences, such as available at MODOMICS [4]. Our tRNA reference pools from this database (see Methods section 5.4) cover the acceptor-space to a large extent, representing 20 amino acid types and including sequences both, with and without m^1A annotation. The latter were intentionally added in order to allow for evidence based identification of hitherto unreported m^1A sites and to include negative control instances. RNA sequences that were experimentally available and provided with database annotations of m^1A sites include yeast cytosolic tRNAs, human mitochondrial tRNAs, as well as yeast, murine and human rRNA (see Table 2). Furthermore, rRNA from *Streptomyces pactum* was of particular interest, featuring the only m^1A site in the pool situated in the small ribosomal subunit. Notably, the modification mediates antibiotic resistance [84] of this bacterium.

Operating on this sequence pool, we providently assessed the possibility of multiple mapping targets for reads originating from a tRNA with several isoforms. Isoacceptors, supposed to be charged with the same amino acid [117], frequently show strong sequence similarities. As a consequence, cross-mapping rates are usually higher between isoacceptors than between other tRNAs. The conclusive results of a closer analysis are shown in section 3.1.11.2. Therein, report settings addressing cases of multiple mappings were compared in their effects. From those, we used a regime termed 'k1', reporting one valid mapping site per read only. Initial characterization of m^1A signature was conducted using yeast rRNA, allowing to circumvent the above tRNA-related challenges. The latter are discussed in more detail in section 3.1.11.2.

Table 1: Sequence libraries. Adopted from *Hauenschild et al. 2015* [113].

ID	Interest	Material	Organism	Raw reads	Mapped [%]	Bp	End	∅ Reads / kb ref	∅ cov. 3' of m^1A	∅ arr. (m^1A) [%]	∅ mism. (m^1A) [%]
1	Signature pool	total tRNA	*S. cerevisiae*	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	*H. sapiens*	2.1 M	14.3	151	p	189 k	1.6 k	62	32
3	m^1A_{58} knockout	total tRNA	*S. cerevisiae* Δ *trm6*	5.58 M	89.9 / 89.1	35, 88	p	16.4 M^{Ini}	776 k^{Ini}	0.1^{Ini}	0.4^{Ini}
4	Positive control	total tRNA	*S. cerevisiae*	6.37 M	77.3 / 82.1	35, 88	p	579 k^{Ini}	1.25 k^{Ini}	18.1^{Ini}	7.0^{Ini}
5	Positive control	rRNA	*S. cerevisiae*	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^1A_{645}	rRNA	*S. cerevisiae* Δrrp8	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^1A_{2142}	rRNA	*S. cerevisiae* Δbmt2	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^1A_{645} and m^1A_{2142}	rRNA	*S. cerevisiae* Δrrp8 + Δbmt2	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^1A on SSU of rRNA	total RNA	*S. pactum*	6.27 M	1.2 / 0.6	35, 88	p	250 k	6.4 k	76.0	56.9
10	Homologous identification	rRNA	*H. sapiens*	5.14 M	77.7 / 72.0	35, 88	p	12 k	8.0 k	90.0	30.3
11	Homologous identification	rRNA	*M. musculus*	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	*S. cerevisiae*	1.95 M	44.4 / 49.7	80, 80	p	145 k	5.6 k	51.9	60.2
S22	m^1A_{58} knockout	total RNA	*S. cerevisiae*	3.21 M	58.8 / 51.9	80, 80	p	571 k	11.2 k	0.2	0.3
S23	Novel sites	total tRNA	*T. brucei*	1.36 M	4.8 / 3.9	80, 80	p	7.5 k	0.4 k	36.6	83.0
S24	Signature vs. occupancy	$tRNA^{Arg}\_1$	*T. brucei*	1.79 M	13.1 / 12.4	80, 80	p	2.6 M	85 k	18.3	82.8

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 (∅ reads / kb ref) of reads mapped on a kilobase (kb) of an m^1A-annotated target reference sequence is listed, e.g. $tRNA^{Ini}$ in sample 3 and rRNA in sample 5. The mean coverage (∅ cov. 3' of m^1A) at +1 positions 3' of m^1A provides a reference for the arrest rate, ∅ arr. (m^1A), at the m^1A position and ∅ mism. (m^1A) provides the mismatch contents. Sample 1 values originate from replicates (N=2). Numbers annotated with *Ini* refer to $tRNA^{Ini}$ only. Entries in brackets refer to pairs of m^1A sites in the corresponding sample, such as m^1A_{645} and m^1A_{2142} in rRNA. The mean is given in front of each bracket.

Table 2: m^1A sites

RNA spec.	Position	Organism	Distinct RNAs	Replicates
Confirmed				
tRNA cyt.	58	*S. cerevisiae*	20	2
tRNA mit.	9	*H. sapiens*	13	1
rRNA LSU 25S	645	*S. cerevisiae*	1	2
rRNA LSU 25S	2142	*S. cerevisiae*	1	2
rRNA LSU 28S	1309	*H. sapiens*	1	1
rRNA SSU 16S	964	*S. pactum*	1	1
Artif. oligo	9*		2	2
Revolver oligo	9*		4	1
$tRNA^{Arg\_UCG}$ cyt.	58	*T. brucei*	1	2
Unconfirmed				
rRNA LSU 28S	1136	*M. musculus*	1	2
tRNA mit.	9	*H. sapiens*	1	1
tRNA cyt.	58	*T. brucei*	15	1

Confirmed instances include published and self-designed m^1A sites, whereas **unconfirmed** sites rely on homologous identification. 'Distinct RNAs' refers to the number of non-redundant RNAs, in which m^1A signatures were found. LSU - large subunit. SSU - small subunit. * asterisk-labeled synthetic oligoribonucleotides contain m^1A_9 in a sequence derived from mitochondrial $tRNA^{Lys}$ of *Homo sapiens*.

3.1.3 Arrest Rate and Mismatch Rate

Various studies already described m^1A's effect as road-block [90] or indicated a composed mismatch rate in NGS profiles caused by misincorporations of dA, dG and dC [101] at the cDNA position opposite to m^1A. Nevertheless, most approaches lack simultaneous capture of abortive products [105] and mismatch information [95].

A first demonstration of the combined resolution potential of both kinds of RT errors is given in the sequencing profiles from pure yeast 25S ribosomal RNA shown in Fig. 3A. This large ribosomal subunit features known m^1A sites at positions 645 and 2142. Samples isolated from whole ribosomes of a wildtype strain as described in [81] showed clear coverage drops and increased amounts of mismatches at the modified positions compared to the surrounding sequence profile. Since ribosomal RNAs of several kilobases exceed the read length by far, fragmentation of rRNA was necessary in order to cover the whole template sequence by a sufficient density of RT start sites. As a consequence of such start sites overlapping with RT stops, coverage drops may provide only an underestimate readout of RT arrest frequency.

Thus, we defined arrest rate a_i of a position i as the relative portion of read alignments starting at $i+1$ (covering $i+1$ but not i) among all reads covering $i+1$, referred to as coverage c_{i+1}. If $s_i + 1$ is the number of reads starting at $i+1$, the arrest rate of position i is defined as: $a_i = \frac{s_{i+1}}{c_{i+1}}$. Let d_i be the number of read alignments that cover i with a base different from reference base at i. Then, mismatch rate is defined as $m_i = \frac{d_i}{c_i}$. These simple definitions of a and m are abstracted from the corresponding computational steps presented in Material and Methods section 5.6.

Importantly, the features are absent at one or both sites in profiles from single or double knockout mutant samples, where the corresponding responsible methyltransferases [81] are missing. As expected, *Rrp*8-deficient yeast lacks the signature at position 645, which respectively holds at position 2142 for the *Bmt*2 null mutants. Similarly, signatures at m^1A_{58} sites in tRNA disappear, if the responsible enzyme *Trm*6 is missing [74, 75], as demonstrated for yeast $tRNA^{Ile\_TAT}$ and $tRNA^{Cys\_GCA}$ in Fig. 3B exemplarily.

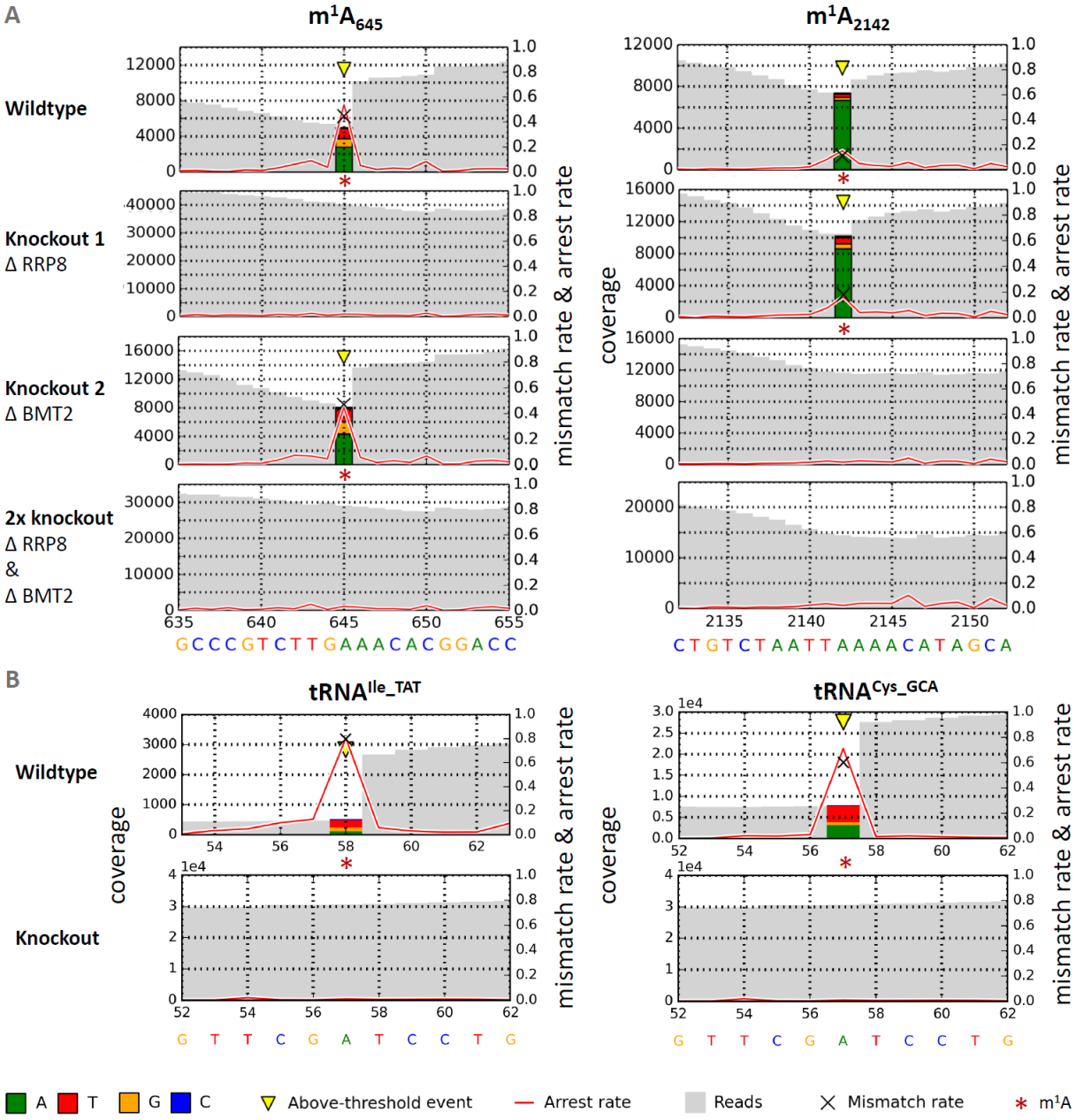


Figure 3: Detection of m^{1}A signatures in deep sequencing data. The representations illustrate the coverage of a given site in gray, the arrest rate is plotted as a red line, and the mismatch composition is visualized by colored stacks at the m^{1}A sites. For a position p, the arrest rate reflects 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 m^{1}A sites 645 and 2142. Signatures of m^{1}A residues are clearly apparent in the wild-type, and disappear in the corresponding knockout constructs. (**B**) Sequencing profiles of tRNA^{Ile_TAT} and tRNA^{Cys_GCA} from wild-type and *Trm*6-knockout yeast strains. The signatures clearly disappear in RNA from a knockout strain of the enzyme, which is responsible for synthesis of m^{1}A$_{58}$ in tRNAs [75]. tRNA^{Ile_TAT} and tRNA^{Cys_GCA} are depicted as examples out of 37 signatures, which are detailed in Appendix subsection 6.2, figure 27. Positions are labeled according to absolute length of reference sequences, including variable regions. Adopted from *Hauenschild et al. 2015* [113].

On the one hand, the results clearly demonstrate a distinct m^{1}A-induced signature even in RNA species whose stable structures are known to affect RT arrest rates [94]. However, a significant variation is evident already when comparing the two m^{1}A sites of the wildtype rRNA sample. Interestingly, while m and a are close in value, both parameters are higher at m^{1}A site 645 ($m = 45.9\%$, $a = 49.3\%$) than at m^{1}A$_{2142}$ ($m = 11.4\%$, $a = 15.6\%$).

Contrary to the impression that m^1A site 2142 might possess a lower modification occupancy than position 645, a control analysis of the wildtype, single and double knockout samples by LC-MS/MS (Fig. 4) verified equal levels of 0.7 mol m^1A per mol rRNA at both sites, consistent with 1.4 mol m^1A per mol rRNA in the wildtype. While the readout at position 645 correlates roughly with the determined m^1A content, the decreased signal at 2142 suggests a much lower modification level. Of note, quantification of fractional occupancy can be fraught with a ~10% error in overall precision. However, since the analyses were conducted with aliquots from the same rRNA preparation, numerous error sources are equalized in case of relative comparison of rRNA from both knockout mutants [37]. The fact that signatures can vary strongly between sites of similar occupancy leads to the conclusion that arrest and mismatch rates are of limited reliability for quantification of modification level. Nonetheless, the sum of both rates serves as lower limit estimate of m^1A content, since the signals disappear at unmodified adenosine residues. Deeper implications and consequences of this finding apply particularly to DMS experiments for structural probing, where RT stops serve as a semi-quantitative readout of RNA's accessibility to a methylating reagent [95].

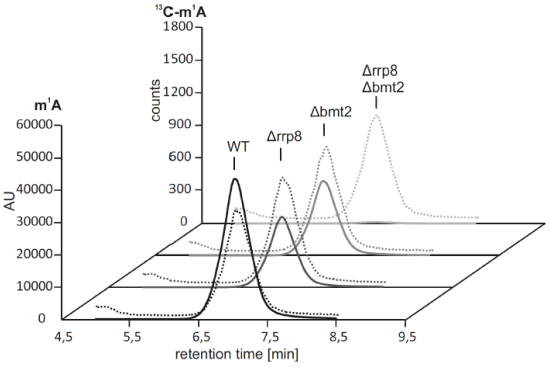


Figure 4: Quantification of ribosomal m^1As by LC-MS using a biosynthetic internal standard. LC-MS/MS chromatograms showing the m^1A and ^{13}C-labeled peaks in 25S rRNA from wild-type and *Rrp*8/*Bmt*2 knockout yeast. Continuous lines represent the peaks of unlabeled m^1A, dotted lines those of ^{13}C-labeled m^1A added as an internal standard [37]. To ensure inter-sample comparability of the m^1A peaks, the peak heights were adjusted to the respective ^{13}C-m^1A peaks and normalized to the injected amount of 25S rRNA. The amount of analyzed 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. Adopted from *Hauenschild et al. 2015* [113]. Measurements done by Katharina Schmid and Kathrin Thüring.

Motivated by the conclusion that signature intensities may be modulated not only by the modification level itself but also by RT sensitivity to the modification's structural context, we introduced a new parameter. What we termed Context Sensitive Arrest rate (CSA), normalizes a positional arrest rate a with respect to the median of surrounding arrest rates:

$$CSA^r(i) = \frac{a_i}{median(a_{i-r},...,a_{i-1},a_{i+1},...a_{i+r})}$$

Here, r denotes the number of contributing positions (visual range) upstream or downstream of position i. If not specified otherwise, we used CSA^5 ($r = 5$), providing a visual range long enough to compensate up to four positions of increased arrest due to potentially higher modification density or aggregated RNA fragment ends: 15.5% of positions in our yeast cytosolic tRNA reference pool are annotated with modifications. By using the median instead of mean surrounding arrest rate, high coverage drops due to such modifications within the $2 \cdot r$ neighboring positions are less likely to bias CSA. On the other hand, $r = 5$ is still short enough to reflect the local profile of RT arrests, e.g. in loop structures.

3.1.4 Signature by m^{1}A Occupancy

Once the modification level is verified for certain arrest and mismatch rates, lower signature intensities at the same position should reflect fractional occupancy reliably. In order to gauge the effect of incomplete modification in a real scenario, we used a synthetic analog of yeast mitochondrial tRNA^{Lys}'s first 20 bases bearing an m^{1}A at position 9 [78]. This wild-type sequence was mixed with equivalents of an unmodified twin (*in vitro* transcript) in equidistant ratios as shown in figure 5A (i-v). Obviously, the rates of arrest and misincorporation increase linearly with the content of m^{1}A. Additional verification of the modification levels in the mixtures in figure 5A was performed by quantitative LC-MS/MS using the recently developed biosynthetic stable isotope labeled standard [37].

Although neither of the readouts is precise enough for definite quantification of incomplete modification, Figure 5B demonstrates the clear linear correlation of both, with these results. The mixture ratios (i-v), indicated by 'S' markers are located at equidistant positions on the ordinate, whereas their positions on the abscissa reflect the *de facto* m^{1}A levels. Importantly, this discrepancy does not affect the quality of dependency, since the linear equations were fitted based on the abscissa values determined by LC-MS/MS. In parts, the incomplete m^{1}A level even in the 100% mixture can be attributed to traces of m^{6}A, a rearrangement product of m^{1}A [118], we identified even in synthetic samples. In total, the results of the mixture assay suggest that many natural m^{1}A sites are probably incompletely modified. Thoroughly calibrated RT profiles may be used to estimate the underlying modification efficiency to a remarkable degree, though.

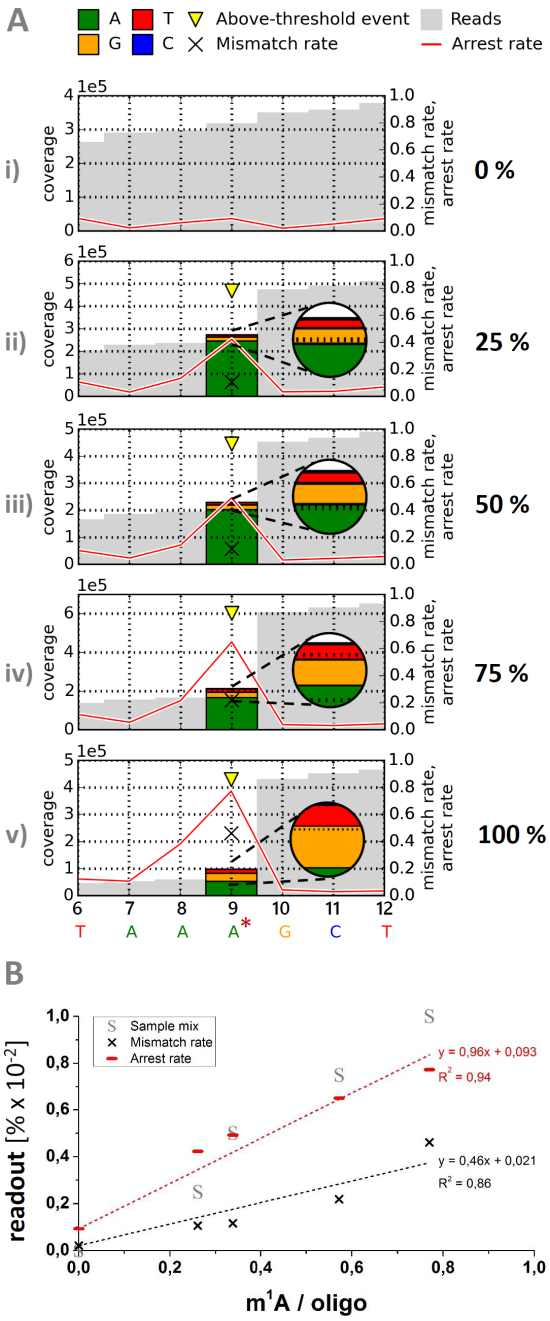


Figure 5: Signature intensity by m^{1}A content. (**A**) Arrest and mismatch rates at different ratios of modified (synthetic) and unmodified (*in vitro* transcript) equivalents of a human tRNA^{Lys}-derived oligonucleotide are shown: 0% synth. in (i), 25% in (ii), 50% in (iii), 75% in (iv) and 100% synth. in (v). (**B**) Correlation of m^{1}A level (quantified by LC-MS/MS) in sample mixes from (**A**) and signature intensities. Adopted from *Hauenschild et al. 2015* [113].

While causality between common signature characteristics and the presence of m^{1}A residues was demonstrated, the signatures of different natural m^{1}A sites vary not only in quantity, but also in quality. This applies in particular to mismatch composition. Clearly, a larger number of m^{1}A instances has to be analyzed in order to obtain a comprehensive picture.

3.1.5 Mismatch Composition

As a central goal of this work, we refine m^1A's RT signature beyond its established description to a level of maximum resolution. Therein, we optimize its value when approaching prediction of m^1A sites by computational screening of NGS data.

Closer assessment of the signatures at multiple m^1A sites revealed that arrest and mismatch rates are not the only parameters to fluctuate. This is demonstrated in Fig. 6, which shows an average signature of 37 m^1A sites from yeast cytosolic tRNAs (Appendix Fig. 27). Differences in both, read-through efficiency and misincorporation frequency can be mainly ascribed to heterogeneous fractional modification occupancies. In contrast, the law of large numbers (Bernoulli/Poisson) would suggest a G/T/C mismatch composition converging towards a narrow range as a consequence of the huge amount of molecular copies reflected by the Deep-Seq results of each m^1A site. However, variation of mismatch contributions among the analyzed sites is evident.

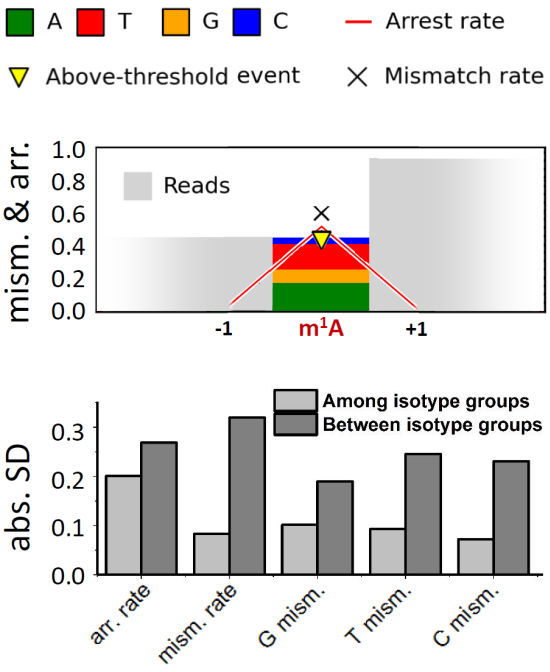


Figure 6: Average m^1A signature. Arithmetic mean of profiles of 37 m^1A sites in yeast cytosolic tRNAs (supplementary Fig. 27) and standard deviations (SD) of arrest and mismatch parameters among and between tRNA isoacceptor groups. The threshold for displayed events was set to a minimum mismatch rate of 5%. -1 and +1 refer to the sequence positions 5'- and 3'-adjacent to the m^1A site. Adopted from *Hauenschild et al. 2015* [113].

Previous studies [101] had described increased misincorporation rates of mainly dA and dC into the cDNA, accompanied by a small fraction of dG opposite to an m^1A residue. The resulting G/T mismatch ratio underlies significant variation, as found by further analysis more recently [105]. However, the reason for this variation had remained unclear, leaving a blurred picture of the mismatch composition as a signature component.

The described G/T-driven distribution of mismatch compositions was verified by the average profile. In order to obtain a more detailed picture, we extracted signatures from the entirety of annotated m^1A sites in Table 2. Next, the mismatch compositions were gathered in a ternary plot [105] (Fig. 7A), resulting in a pattern consistent with the described preference of data points scattering along the T-axis, indicating the G/T-driven mismatch range. As per the goal of signature refinement, this unexplained scattering was investigated from various angles.

Reconsideration of the average profile (Fig. 6) yielded the decisive hint, when it became clear that the mismatch components differ much more between distinct groups of tRNA isoacceptors than among tRNAs of the same isoacceptor group. Thus, we correlate sequence context with mismatch patterns in the subsequent section.

3.1.6 Sequence Dependence

Corresponding to their occurrence in different tRNAs, the 41 m^1A instances from Fig. 7 were found to be located within different sequence contexts. The fact that the reverse transcriptase proceeds from 3' to 5' of the RNA template during cDNA synthesis, implies that position +1 (see Fig. 6) is reverse transcribed before the m^1A site, defined as position 0, enters the active site of the enzyme. Consequently, the -1 position can only act as a template after the RT has bypassed the m^1A site.

Reasoning that the immediate molecular environment is most likely to affect RT behavior in its direct encounter with m^1A, the positions -1, +1 and +2 were chosen for assessment of their influence on the signature. Next, the signature parameters of all m^1A instances in Tab. 2 were analyzed as a function of the base configurations of neighboring nucleotides. Whereas arrest rate and mismatch rate did not show notable dependence on the sequence context, the composition of mismatches was sensitive all the more. This could be demonstrated visually: distinct influences of the given base identity at either of the three positions would be reflected by a clustering of mismatch compositions, when plotted in a ternary diagram.

Strikingly, the ternary plot showed clear clustering of colors, when the mismatch composition values were colored according to the +1 base configuration (Fig. 7B). The result demonstrates that a 5'-m^1A-U-3' motif, present in red data points clustering in the upper corner of the ternary plot, leads to highly efficient dATP misincorporation into cDNA. These misincorporations are reflected by a high contribution of T mismatches in the mapping profiles using the RNA sequence as reference, which obviously involves low relative amounts of G and T mismatches. Similarly, the motifs 5'-m^1A-A-3' (green) and 5'-m^1A-G-3' (yellow) lead to distinct clusters of generally low misincorporation frequency of dGTP corresponding to C mismatches. The 5'-m^1A-G-3' motif has the stronger G mismatch tendency here. Moreover, four spread-out 5'-m^1A-C-3' motifs were ob-

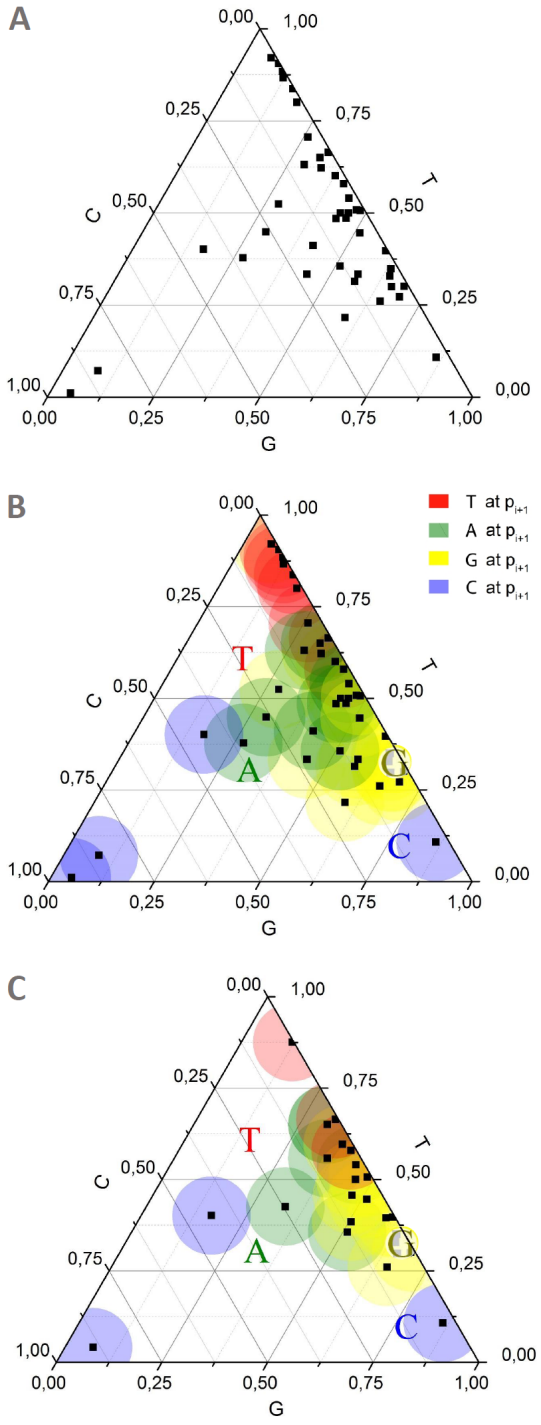


Figure 7: Mismatch composition: +1 dependence & binning. G/T/C balance at 41 natural m^1A sites (**A**) in a ternary plot are colored by base configurations guanosine (yellow), cytidine (blue), uridine (red, T in mapping profile) and adenosine (green) at position +1 relative to m^1A. (**B**) Data points from revolver oligonucleotides are represented as colored letters (same color code). (**C**) 22 hierarchically binned data points derived from initial 41 measurements in (**A**) and (**B**). Adopted from *Hauenschild et al. 2015* [113].

served. They share low T mismatch signals, but a larger set of instances would be required to characterize a more distinct trend.

In the light of this sequence dependence, it is important to reconsider the underlying assumptions and applicative perspectives of m^1A's RT signature definition. Obviously, the sequence context pool analyzed here, is biased by evolution, by database annotation and even by the modification level which is decisive of a detectable mismatch signal. On the one hand, it is arguable that the entirety of accessible m^1A instances at least roughly approximates the natural distribution of the modification's sequence context at positions -1, +1 and +2 relative to the m^1A site. This would suggest acquisition of an overall signature from the given distribution in order to yield the most transferable, robust and therefore valuable description, in particular for the analyzed tRNA and rRNA landscape. On the other hand though, when describing the impartial influence of a fixed +1 nucleotide configuration X minimally biased by nonuniform frequencies of -1 and +2 configurations, 3'-m^1A-X-5' instances of identical (over-represented) -1 and +2 sequence contexts need to be averaged before being pooled with such 3'-m^1A-X-5' from other -1 and +2 sequence contexts. This procedure already makes analogous averaging for fixed -1 and +2 positions obsolete, since it yields only unique combinations of -1, +1 and +2 configurations.

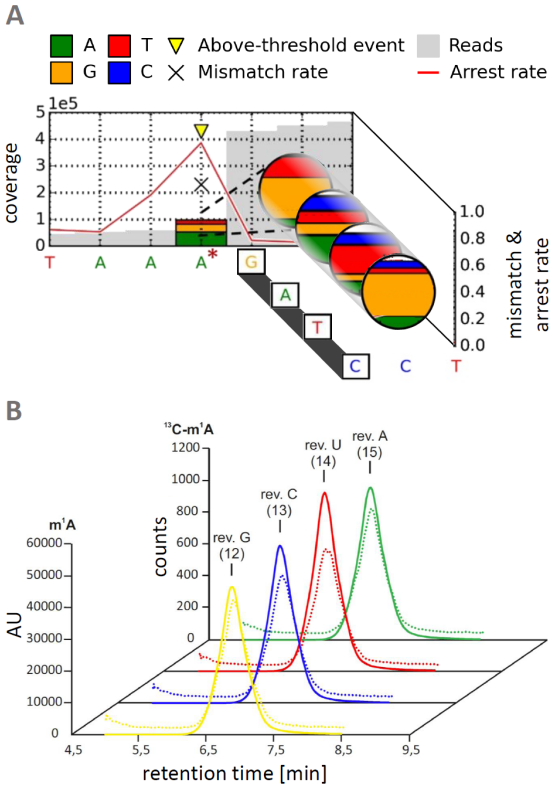


Figure 8: Revolver Assay. (**A**) RT profiles of synthetic oligonucleotides with variegated +1 base configuration, termed *revolvers*. The m^1A_9 site is marked by an asterisk. (**B**) Quantification of m^1A in revolver oligonucleotides by LC-MS/MS using a biosynthetic internal standard. Chromatograms show the m^1A and ^{13}C-labeled peaks. Continuous lines represent the peaks of unlabeled m^1A, dotted lines those of ^{13}C-labeled m^1A added as an internal standard [37]. To ensure inter-sample comparability of the m^1A peaks, the peak heights were adjusted to the respective ^{13}C-m^1A peaks and normalized to the injected amount of oligonucleotide. Amounts of oligonucleotides were 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. Adopted from *Hauenschild et al. 2015* [113]. LC-MS/MS measurements done by K. Schmid and K. Thüring.

First, from originally 54 m^1A instances, experimental replicates were averaged, which led to the 41 data points in Fig. 7A. Next, overrepresentation of certain sequences was taken account of by averaging instances from mapping references exceeding 95% similarity (see Methods section 5.7), e.g. tRNAs featuring SNPs. Now, as described in the previous paragraph, the final averaging step reduced the data points to those mismatch compositions observed under unique sequence contexts. Crucially, the described sequence dependent mismatch composition persists also in the resulting 22 hierarchically averaged (=binned) data points presented in Fig. 7C. Fig. 9E presents the observed combinations in the 41 data points, which cover the theoretical combinatorial space of $4^3 = 64$ configurations by roughly one third (22 unique wedges in pie chart).

Pursuing two goals, verification of the pronounced effect of the +1 nucleotide and a first expansion of the incomplete pool of sequence contexts, an additional four data points were generated. In what we termed 'revolver' concept, four versions of synthetic oligonucleotides were used, which contained m^1A in a uniform sequence context, but featured variegated +1 base configuration (Tab. 6). Exactly as in section 3.1.4, the sequence was derived from natural human mitochondrial $tRNA^{Lys}$ [78] bearing an m^1A_9. If the assumed +1 dependence was true, the m^1A mediated mismatch compositions of the four revolver oligoribonucleotides differing only at +1 position would turn out in an allocation ideally congruent to the corresponding colored clusters in Figures 7B and 7C. Indeed, the revolver sequencing profiles presented in Fig. 8 and their mismatch compositions, plotted as colored letters A, G, T (U) and C in the ternary plots respectively, reflect well the trends detected in the natural instances. Statistical analysis verified this in a permutation test (Methods section 5.7), the results of which are shown in Fig. 9F. The overall mismatch and arrest rates were similar between the oligoribonucleotides, which is in agreement with previously observed absence of sequence context impact on these parameters when analyzed for the natural m^1As.

As a consequence of the described impact of the +1 sequence context on m^1A induced mismatch composition, knowledge of the +1 base configuration improves the resolution of this signature component significantly. Theoretically, knowledge of all three investigated base configurations could refine this resolution even slightly further, but acceptable reliability would require

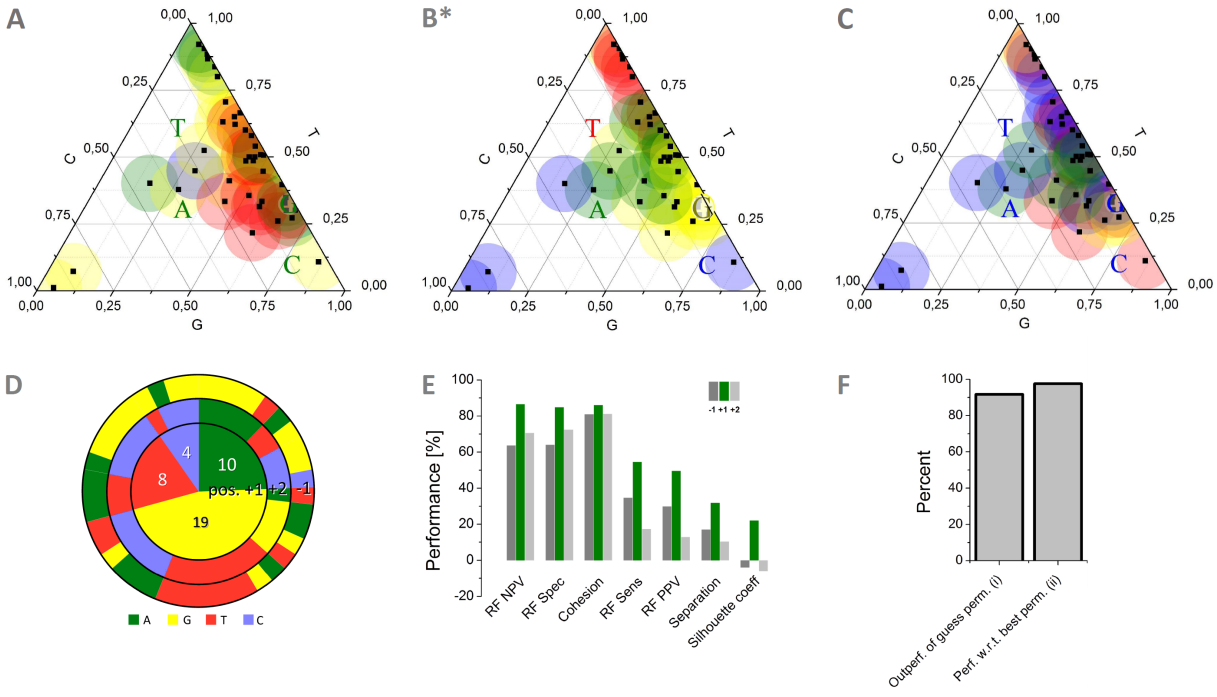


Figure 9: Signature dependence on sequence context. Mismatch composition at m^1A site by nucleotide configuration at -1 (**A**), +1 (**B***, identical with Figure 7B), and +2 (**C**). Data points from revolver oligonucleotides are specified by base at +1. (**D**) Observed combinations of base configurations at positions +1, +2 and -1 relative to m^1A. (**E**) Positional comparison by clustering measures cohesion, separation & silhouette coefficient [119] 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 m^1A site's mismatch composition. (**F**) Accordance of revolver assay with m^1A 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), detailed in section 5.7. The correct assignment ranks at 97.6% of the mean MDC of the best-performing permutation (ii). *Hauenschild et al. 2015* [113].

a much larger volume of data points, preferably covering a maximum variety of sequence context space. An ideal data set should include a sufficient number of technical and biological replicates in order to obtain a proper estimate of the variance in redundant, homogenous data. The next level of variation would then be implemented by multiple of such replicate sets sharing the -1, +1 and +2 context but differing in RNA species origin. Provided such data sets for each of the 64 combinations of analyzed sequence context, a potential cooperative impact of -1, +1 and +2 configurations may be assessed.

Nevertheless, the available m^1A signature pool allows evaluation of the independent impact of individual base configurations. Fig. 9A, B and C illustrate a direct comparison of mismatch compositions with respect to base configurations at positions -1, +1 and +2. Clearly, the +2 configuration does not have a systematic individual impact in the 41 misincorporation patterns, as is reflected by a promiscuous arrangement of colors. Position -1, although the ternary plot appears slightly better organized in color distribution, had only one single ribocytidine (blue) configuration and resolution is clearly inferior to that of the +1 diagram B*. These visually apparent differences were verified by both, descriptive and inferential statistics, also including the four revolver data points. Cohesion, separation and the Silhouette Coefficient [119] were chosen as indicators of clustering quality. For each of the three context positions, these parameters were calculated for the individual color clusters and then reported as a mean weighted by cluster sizes. The indicators were calculated based on an edit distance measure $d_{i,j}$ between two mismatch compositions i and j, defined as the minimum required alteration of mismatch components $mism_k \in \{mism_G, mism_T, mism_C\}$ to transform i into j:

$$d_{i,j} := (\sum_{mism_k} |mism_{k,i} - mism_{k,j}|) - \max_k |mism_{k,i} - mism_{k,j}|$$

More details on weighting and normalization can be found in Methods section 5.7.

As illustrated in Fig. 9E, the clusters colored by the +1 base configuration exhibit the best cohesion and are more separated from each other leading to less overlap. This is exposed even more evidently by comparing the average Silhouette Coefficients at -1, +1 and +2.

Practical benefit of the improvement of mismatch composition resolution by the impact of a given +1 nucleotide was initially evaluated by using a Random Forest [120, 121] machine learning model. If such a model could learn to predict the +1 adjacent base configurations of the m^1A sites based on their mismatch composition (i.e. the data point's location in the G/T/C triangle), then one could reciprocally rate new m^1A candidates by the plausibility of their G/T/C balance occurring under their corresponding +1 nucleotide. As expected, when trained and tested in separated cross-validation setups for positions -1, +1 and +2, the model showed best prediction results in each performance category, when predicting the +1 base configuration. The results are depicted in Fig. 9E. Utilization of the +1 dependence of mismatch composition for computational prediction of m^1A candidates is discussed in section 3.1.9.

From a biochemical point of view, the predominant impact of the +1 base configuration is plausible. As illustrated based on m^1A's sequence context in the revolver oligoribonucleotides in Fig. 10, at the very moment of the enzyme's encounter with m^1A, both, the -1 and +1 residue are closer to the RT's active site than is the +2 ribonucleotide. Crucially, the +1 neighbor is already base-paired, which extends the spatial sphere of influence by the properties of the complementary deoxynucleotide. Since canonical Watson-Crick base pairing shown in (viii) is apparently impeded by a clash due to m^1A's methyl group (i-vii), alternatives, such as Hoogsteen / C-H edge pairing [100, 122] must be considered. For what physicochemical reasons pyrimidines or purines, or even specific deoxynucleotides are preferred on the cDNA side under a given +1 configuration, remains to be clarified in extensive studies and is beyond the scope of this work.

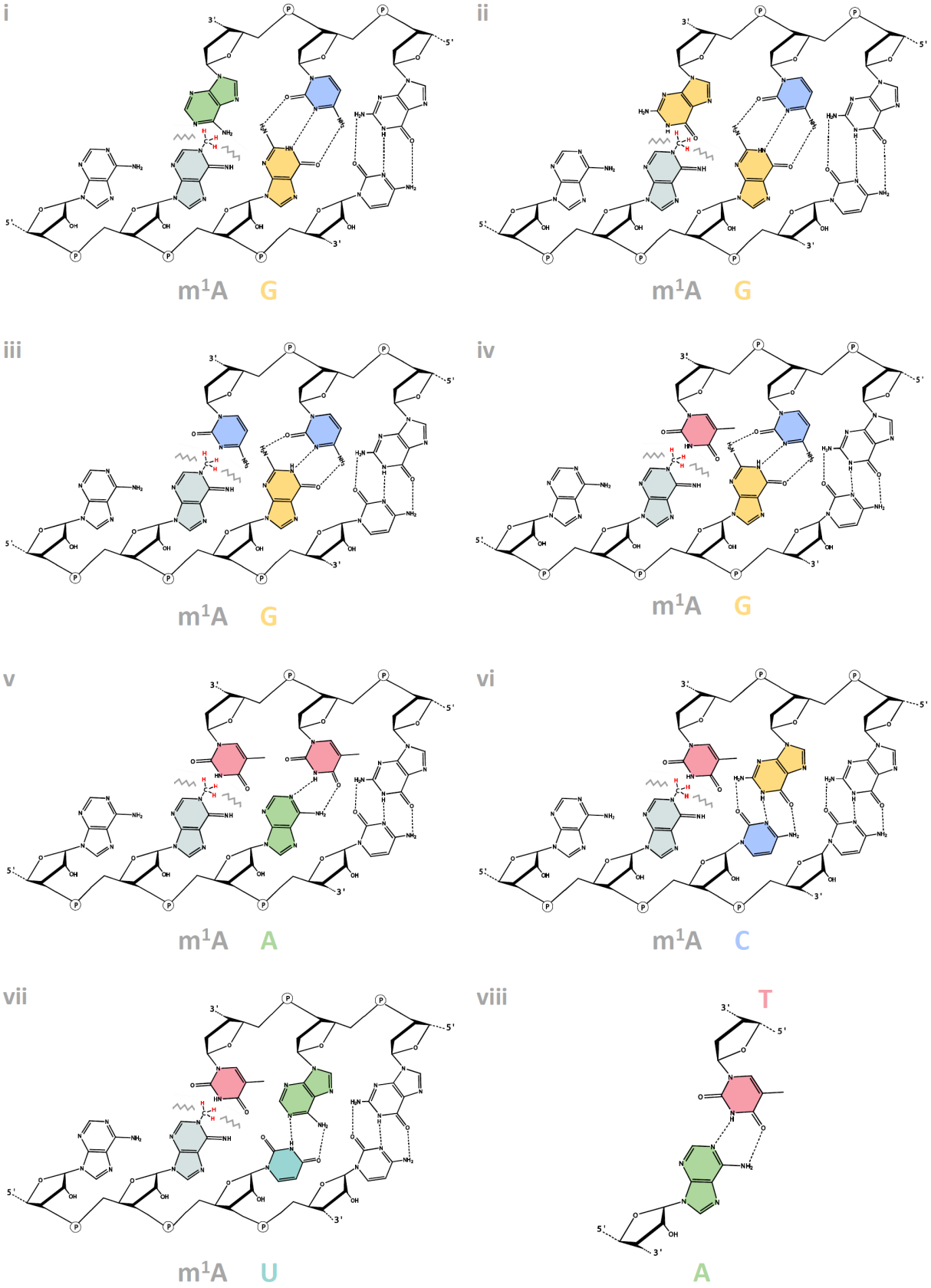


Figure 10: RT sequence context and m^{1}A base pairing. Scheme of positions 8-11 of the tRNA^{Lys} revolver sequence sketches how methyl group of m^{1}A with constant +1 neighbor rG, impairs base pairing with mismatching residues (i-iii) dA (green), dG (yellow), and dC (blue) or with matching dT (red) in cDNA (iv). Matching dT (red) in cDNA is furthermore shown in different +1 base configurations (v-vii) rA (green), rC (blue) and rU (turquoise). Steric clashes are indicated by jagged lines in contrast to standard rA-dT base pairing with conventional H-bond formation (viii).

3.1.7 Homologous Identification

Adopting the above described improvements of signature resolution, the identification of previously unreported m^1A sites comes within reach, already by visual inspection of RNA-Seq profiles. Arguably the easiest deployment of the characteristic signature is the qualitative confirmation of putative m^1As, where they have not yet been detected by other methods, but show plausible sequence homology to annotated sites. For example, the obvious homology between human and murine 28S rRNA allowed the identification of m^1A_{1136} in mouse by analysis of the murine profile in the region corresponding to the known m^1A_{1309} [123] in the human sequence. Fig. 11A shows intense arrest and mismatch rates at both sites. Similarly, m^1A's signature was found in human mitochondrial $tRNA^{Asn}$ identical to the sequence of the bovine homolog, which was recently sequenced and annotated with an m^1A_9 site [8].

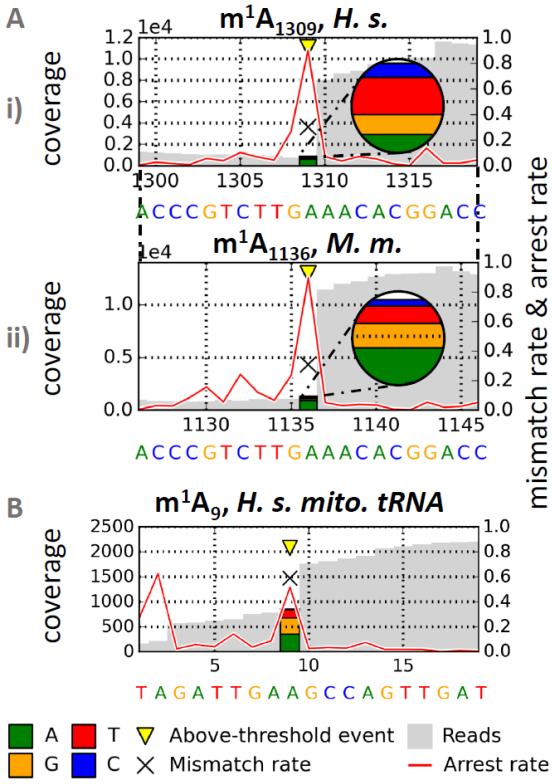


Figure 11: Homology based confirmation of m^1A. For a position p, the arrest rate reflects the relative amount of mapped reads ending at $p+1$, i.e. not covering p. (**A**) Homologous identification of m^1A_{1136} in murine 28S rRNA (i) by alignment to human sequence containing m^1A_{1309} (ii). (**B**) m^1A_9 in human mitochondrial tRNA, identified by alignment to identical bovine sequence with published m^1A_9. Adopted from *Hauenschild et al. 2015* [113].

During the final phase of this project, the great potential of such homologous identifications by software-aided visual inspection was thereupon made use of in a more challenging endeavor. *Trypanosoma brucei*'s tRNA had so far been poorly annotated in terms of m^1A sites. Total RNA [124] of this eukaryote was submitted to the same library preparation protocol as the previous samples. Sequencing yielded typical m^1A signatures at positions 58 (deviations due to variable loops) in 16 tRNA species when inspected visually, as demonstrated in Appendix section 6.5, Fig. 29. Importantly, this applies also to the mismatch composition and dependence on the +1 base configuration, which shows impressive agreement with the findings in Fig. 7B. As to verify the presence of the modification in at least one of the species, $tRNA^{Arg_UCG}$ was isolated by hybridization with a biotinylated cDNA and subsequently sequestered on streptavidin-beads (see details in Methods section 5.1). Now, the purified tRNA was submitted to LC-MS/MS analysis. The results suggested near complete modification, namely one m^1A residue per molecule tRNA as presented in Appendix Tab. 10. Subsequently, an additional sequencing run was performed for the isolated tRNA, yielding a coverage profile in excellent agreement with that of the bulk-sequenced version. This experiment confirmed that the tRNA-specific mismapping effects discussed in section 3.1.2 are indeed minor. However, we intensify this topic in section 3.1.11.2.

3.1.8 Supervised Prediction

Distinctiveness is the crucial benchmark of m^1A's RT signature in predictive application. To evaluate how reliably the signature can distinguish the modification from non-m^1A sites, a machine learning based supervised prediction was conducted.

Model choice. A Random Forest [120] model (RF) was deployed using an R package [121]. Though reported as inferior to support vector machines (SVMs) [125], this versatile data mining method is widely used in Life Sciences [126, 127, 128] gaining popularity by high prediction accuracy and unique advantages, e.g. variable importance readout [129]. Importantly, with respect to our sparse m^1A data, RFs master scenarios with many variables but few samples [130], a.k.a. 'course of dimensionality' problems [126] being robust to overfitting and outliers [129].

Random Forest principle. RFs take advantage of the **b**ootstrap **agg**regation (='bagging') principle related to boosting, which combines a set of potentially weak learners, here decision trees, to obtain a strong meta learner. Each tree is grown based on a new random 2/3 'bootstrap' sample of the available training data drawn with replacement. Recursively at every branching (node) in the tree, a random subset of features describing these data is drawn, a.k.a random subspace method [131]. Next, the bootstrap sample is split at that node by a binary fork using the variable providing the highest increase in purity of the child nodes, i.e. best separation of items by class labels based on a decisive threshold value of that variable. Growing ends, when the terminal nodes of all trees have reached 'leaf' status, i.e. they contain items of one class type only. A new item is classified by the majority consensus of all votes returned by the single trees after processing that item along their decision forks and assigning class labels in the leafs.

Validation scheme. Throughout growing, RFs already perform self-evaluation based on the average classification error on out-of-bag (OOB) data sets presented to the respective 2/3 fractions of trees grown without that data. However, regarding the small m^1A pool, we preferred a validation scheme, which allows to use a larger fraction ($> 2/3$) of the available data as training background in the moment of RF assessment on unseen data. By a post-training k-fold cross-validation with $k = 5 > 3$ (Fig. 12), we slightly reduced the expected overall model performance, since only $k - 1 = 4$ folds are used for training. However, this setup allows to estimate the error on the respective i^{th}'s of k folds data, truly unseen of the entire RF, which now knows 4/5 instead of 2/3 of maximum training background in the moment of OOB testing. This way, also other reported biases [132] in built-in error estimation are circumvented. Instead of a plain OOB error, an ensemble of performance measures was calculated based on the average results from 10 repetitions of a 5-fold cross-validation, including sensitivity, specificity as well as positive and negative predictive value (PPV & NPV) (details in Methods section 5.8). Briefly, the 45 collected m^1A signatures from

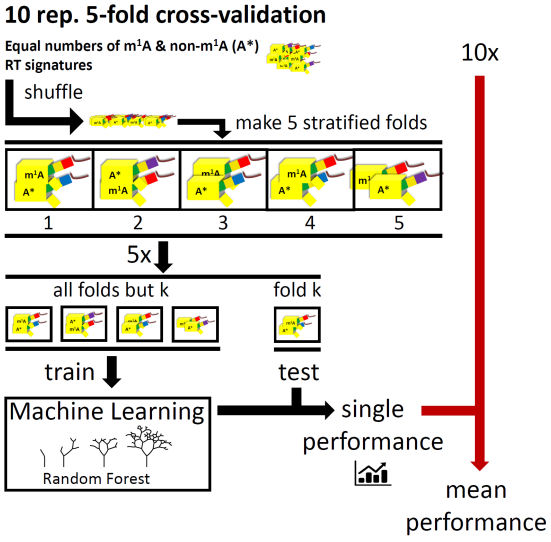


Figure 12: Validation scheme for supervised prediction. RT signatures (yellow) of m^1A and non-m^1A (A*) sites and are distributed into sub-samples ('folds'), with uniform ratios (stratification) m^1A/A^*. The system was tuned toward both, sensitivity and specificity by equal class abundances, minimizing learning biases due to *a priori* class probabilities. In each of 10 repetitions, a Random Forest was trained all 5 possible combinations of 4 folds and tested on the respective 5^{th} fold. Adopted from *Hauenschild et al. 2015* [113].

tRNA, rRNA and synthetic oligoribonucleotides were merged with equal amounts (1:1 ratio) of data points (coverage ≥ 10, +1 pos. coverage ≥ 15) randomly drawn from all non-m^1A sites of the *bona fide* m^1A containing landscape: mitochondrial (human) and cytosolic (yeast) tRNA and rRNA (yeast and mouse), named setting (i) in Methods section 5.8. As shown in Fig. 12, each cross-validation run included shuffling of m^1A (*positive*) and non-m^1A (*negative*) instances. Then, the data were split into five 'stratified' folds, preserving a 1:1 ratio by alternation of pos. and neg. examples. The RF was trained once on each of five possible combinations of four folds, while the respective fifth fold served as test data.

Feature selection. Modern machine learning approaches successively move from hand-crafted features to automatically engineered features [133] learned from raw input. Instead, preferring transparency in this early-stage study of very limited data, we relied on intuitive features that either provide a direct quantitative or qualitative visual readout, such as arrest rate a, mismatch rate m and mismatch components m_G, m_T and m_C, or at least have interpretable semantics like the m/a ratio and the Context Sensitive Arrest rate (CSA, see section 3.1.3), when used by black-box models. Thus, features visible to the RF were a, m, m/a, m_G, m_T, m_C and CSA.

Sequence dependence. Despite its documented impact, the +1 nucleotide identity was not used in this setup, in order to avoid overfitting to the biased sequence context of collected m^1A sites. Otherwise, special precautions in data preparation and model settings would be necessary, to prevent the RF from 'cheating', i.e. learning m^1A signatures by abusing the over-representation of certain neighboring bases (section 3.1.6, Fig. 9D) first off, rather than taking advantage of the correlation with mismatch composition. Compared to sequence motifs of later reported m^1A sites in mRNA [85, 71] differing in +1 base frequencies from our tRNA/rRNA m^1A pool, our mismatch composition is biased in a similar way as their determinant, the +1 configuration. However, approaching this bias by averaging (section 3.1.6), is not suitable, training towards prediction of real m^1A instances: the latter may carry their maximum information content and detective potential in the original signatures. Averaging would further reduce the limited training instances down to an intolerable amount and diversity. Alternative omission of the entire mismatch composition information, for a training bias reduction when aiming at transcriptome-wide mapping [71], should also be refrained from, facing the bad trade-off between the valuable G/T/C balance information lost and the expectable correction of a bias neither fully verified nor quantified.

Inducing m^1A resemblance. The RF scored solid 97% in specificity and PPV with over 96% in sensitivity and NPV, when predicting m^1A in the standard setting (i) (details in Appendix section 6.6, Tab. 7). In a more stringent variation, setting (ii), random non-m^1As were selected, only if they showed a minimum 'resemblance' to m^1A signatures. This way, the impact of a more difficult scenario could be assessed. In principle, distance measures for resemblance can be extracted from an RF itself (e.g. similarity scores or regression outputs) or established by importance-based (see section 3.1.9) weighting. However, this implies undesired circular reasoning, i.e. a pointless conclusion, when the RF is challenged by issues defined by itself. Alternative calculation of neutral distances (e.g. Euklidian) in the manner of clustering methods (e.g. k-Means) or a principal component analysis (PCA) based strategy either lack in sufficient importance-weighted consideration or else in intuition, qualitative interpretation and control. Such downsides were circumvented by a custom filter rule fulfilled by eligible non-m^1As, suffic-ing any of the conditions $a \geq 0.2$, $m \geq 0.2$ or at least two mismatch types with ≥ 0.1 share of an $m \geq 0.1$ (setting (ii), Methods section 5.8). From these readouts, which showed being instrumental in visual inspection, the third is most informative and characteristic for m^1A: it neither occurs in SNPs nor plays a greater role in many other inspected signatures of adenosine modifications. If still shared by several modification types, the actual combination of mismatch components as variable sub-features can provide further potential for distinction. According to

said priority order established in feature inspection, the demand for m^1A-specialized, transparent weighting of these parameters, gave rise to *diversity* score, a primitive weight-based qualitative indicator of signature characteristics. Since this index proved proficient in preliminary investigations on prediction dynamics for different data textures beyond (i) and (ii) (Appendix section 6.6, Fig. 31), it was integrated in the software, *CoverageAnalyzer* (*CAn*) (section 3.2). The impact of various combinations of training and testing textures on predictive performances underlines the importance of appropriate training background for models deployed on unseen RNA sequences.

Performance. Setting (ii) was a harder task for the RF than setting (i), but since both, training and test sets featured non-m^1A profile patterns of a minimum m^1A resemblance, performance dropped to still remarkable ∼89% sensitivity and ∼89% NPV, along with ∼87% specificity and ∼87% PPV. Appropriate usage of scenario-specifically trained models is discussed in section 3.1.11.2. The accuracy drop under setting (ii) can be ascribed to deliberate inclusion of other adenosine modifications passing the selection filter, e.g. two consecutive $m^{6,6}A$ sites in yeast's 18S rRNA. Like m^1A, this species has a methyl group on the Watson-Crick face, causing a mismatch pattern [105]. Indeed, the signature of $m^{6,6}A_{1781}$ is indistinguishable from m^1A's, even by visual inspection (Appendix section 6.6, Fig. 30). Consequently, the algorithm misclassified this position as m^1A site. Its neighbor however, $m^{6,6}A_{1782}$, albeit also featuring a pronounced pattern, deviated from the typical m^1A signature, and was correctly reported as non-m^1A, therefore.

Comparison: Out of concern whether limited data sets, such as the 2x 45 instances from setting (ii), are suitable for complex methods like RFs, the model's Receiver Operating Characteristic (ROC) was compared to that of a more basic classifier, namely a *k*-Nearest Neighbor (*k*NN). In a ROC analysis, the area under curve (AUC) in Fig. 13 corresponds to the probability, the corresponding classifier scores a random positive (m^1A) instance higher than a random negative

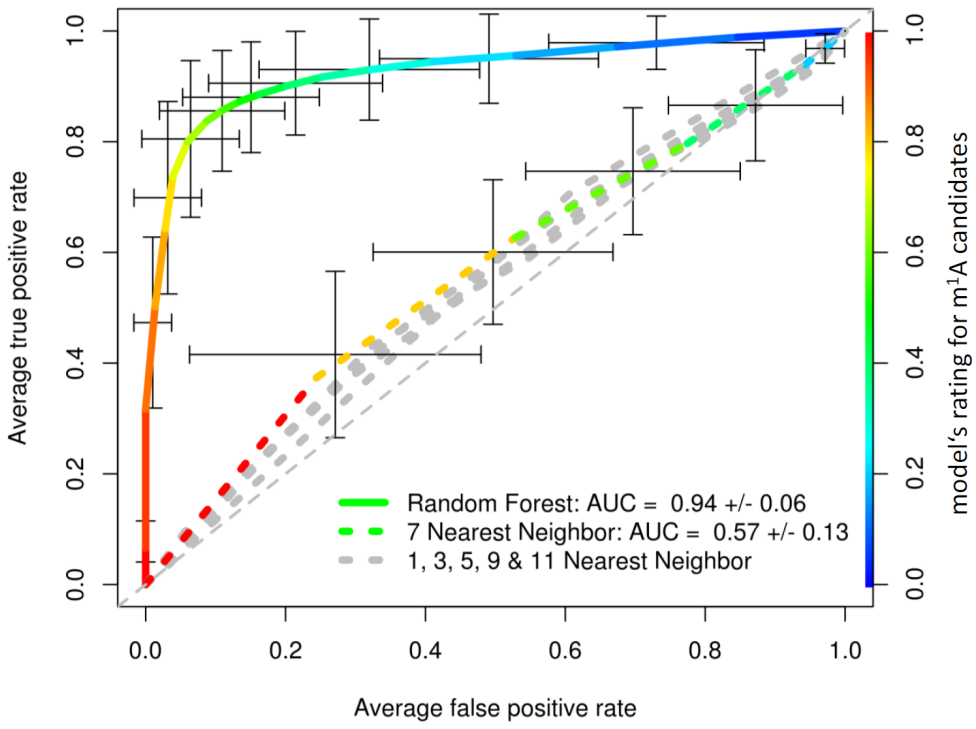


Figure 13: Receiver Operating Characteristic (ROC) plot [134] showing the areas under curve (AUC) for Random Forest and *k*-Nearest Neighbor (*k*NN) in supervised prediction of m^1A vs. other sites (setting (ii), described in the text). Curves are averaged from 10 repetitions of a 5-fold cross validation. Error bars show standard deviations of the ROC curve at the models' rating scores attributed to the m^1A candidates. The gray diagonal corresponds to the performance of a guessing classifier. Adopted from *Hauenschild et al. 2015* [113].

(non-m^1A) one. With an average AUC of 94%, the RF consistently outperformed various kNN settings (max. 57%) by a huge margin. The shape of the RF curve indicates its excellent scoring-tradeoff, which achieves high sensitivities at low cost in specificity.

Maximum training set: In order to obtain a closer estimate of the upper performance limit of a model trained on the entire available data, we developed the idea of a large-k cross-validation further, leading to a 'leave-one-out' cross-validation. More precisely, with a maximum of 45 folds, only a *single* pair of positive and negative instances was left in each test folder. As expected, the average performance increased with the availability of training data (Appendix section 6.6, Tab. 8). This not only reconfirms the feasibility of our concept, but also underlines the need for a maximum number of training instances, essentially including pronounced signatures of non-m^1As. As a rule of thumb, statisticians advise an at least 5-10-fold number of training instances in each class for a given number of learning features [135]. This should be kept in mind, when we go significantly below the recommended ratio in a scenarios presented later on in section 3.1.10. In contrast, the $\frac{4}{5} \cdot 45 = 36$ training instances for both classes (m^1A, non-m^1A), used in each constellation within the 5-fold cross-validation schemes so far, can be regarded as just sufficient for the 7 used features. Some dimensions are not fully orthogonal, which, while lowering the information content on the one hand, slightly relativizes the problem of increasing sparsity in the curse of dimensionality on the other hand.

Transferability: As a final challenge, the RF was trained on the entire amount of available tRNA instances of m^1A with as many random non-m^1As of the same sequence pool (setting (iii)). When tested (10-repetitions 5-fold stratified cross-validation) on an analogously composed unseen rRNA data set, the five ribosomal m^1A sites (2x *S. cerevisiae*, 1x *S. pactum*, 1x *M. musculus*, 1x *H. sapiens* were consistently identified with 100% accuracy. Handling the scenario without a single false-positive classification, the method provided further evidence of its robustness.

3.1.9 Feature Importance

In order to assess, which of the characteristic features identified in visual inspection are predominantly exploited by the RF model, we repeated the cross-validation scheme from section 3.1.8 (setting (ii)). This time, the data set was extended by *bona fide* m^1A and non-m^1A instances from trypanosomal tRNAs (section 3.1.7), leading to a slightly improved results of ∼81% sens., ∼92% spec., ∼88% PPV and ∼88% NPV, due to the additional training material. Next, as an experiment, also the critically discussed (section 3.1.8) +1 base information was provided to the RF, which again raised the performance to ∼86% sens., ∼97% spec., ∼95% PPV and ∼91% NPV. When we extracted the feature importances from the RF, a clear hierarchy was observed (Fig. 14). The underlying rational is that important features, typically promote class splitting efficiency when used at the RF's decision forks, which is measurable as high Gini-Index [136]. The latter indicates the resulting inequality of item class frequencies in the corresponding child nodes (a.k.a. *node purity*) of decision trees. Important variables not only tend to increase the Gini-Index, but permutation of their values easily leads to deteriorated prediction accuracy. The results indicate considerable importance of both types, arrest and mismatch related features. They also confirm the proficiency of a Context Sensitive Arrest rate (CSA) for m^1A detection in contexts rich in RT stops.

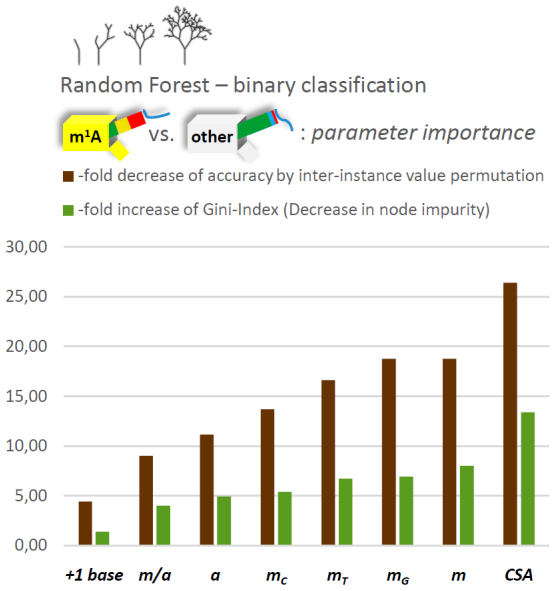


Figure 14: Feature importance. 10 repetitions of a 5-fold stratified cross-validation training a Random Forest with equal amounts of *bona fide* 45 m^1A signatures and non-m^1A counter-examples from cyt. (yeast, tryp.) and mitoch. (human) tRNA as well as from rRNA (human, mouse, yeast) and synthetic oligoribonucleotides. (■) A high Gini-Index [136] $\in [0, 1]$ indicates a high node purity, i.e. inequality of item class frequencies, corresponding to a progressed class separation. Important variables tend to increase the Gini-Index. (■) If prediction accuracy is sensitive to value permutation of a certain variable between instances of different classes, that variable tends to be important. Variables insensitive to permutation usually have low impact on prediction accuracy.

The fact that the RF attributed an only minor importance to the +1 neighboring base, having access to seven other descriptors, can be explained by the data texture combined with the model's working principle. While signature assessment by human visual inspection is not bound to a strict order by information content, the computational model takes a greedy approach using the available variable subspace for item splitting. Particularly in scenarios, in which a high percentage of the non-m^1A signatures are plain adenosines without notable RT patterns (i) or even for more m^1A-resembling instances (ii), the RF can simply apply cutoffs for mismatch or arrest. A fallback on the more sophisticated features ($m_G/m_T/m_C$ balance, +1 base) is seldom necessary. Increased importance of the +1 base information would be expected for higher m^1A-similarity of non-m^1As, which remains an interesting aspect to be revealed in future projects.

Aware that interpretability of above measures can suffer from biases [137] and importances do not necessarily correlate with feature selection frequencies [138] inside the RF, we performed an independent, more straightforward analysis to identify combinations promotive to classification power, when visible to the RF. For the original two setups (i) and (ii) from section 3.1.8, we generated all possible feature combinations, then cross-validated the RF under each. For presentation reasons, performances were generalized as arithmetic mean of sensitivity, specificity, PPV and NPV (showing high mutual correlation). The results were plotted according to the number

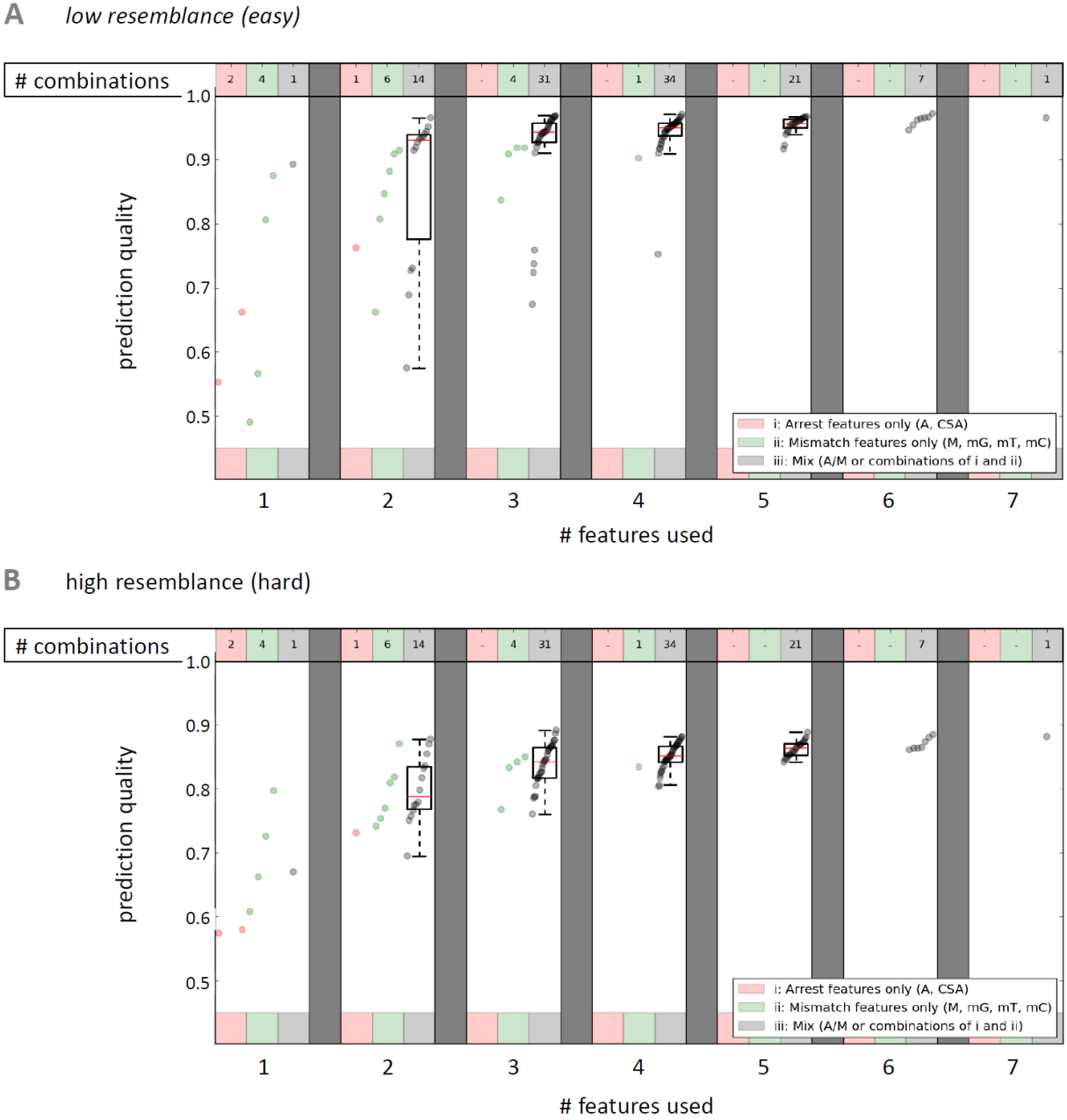


Figure 15: Prediction quality vs. number and category of predictors. Quality is generalized as arithmetic mean of sensitivity, specificity, positive and negative predictive value. (**A**) Low and (**B**) high avg. resemblance of non-m^1As to m^1As as specified in section 3.1.8. 27-1=127 combinations 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. Adopted from *Hauenschild et al. 2015* [113].

of involved features (1-7) and colored by categories 'mismatch info. only', 'arrest info. only' and 'mixed info.' leading to the representations in Fig. 15. Where readability suffered from large numbers of combinations, box plots were chosen to facilitate evaluations. As an example, in the first column of panel A, the only hybrid feature m/a outperforms any other single feature. In the second column, most mixed feature pairs lead to higher performance than homogeneous pairs. Column 7 of Fig. 15A and B corresponds to the full feature set provided to the RF, resulting in the performances reported in section 3.1.8. Not only, these findings demonstrate improved learning with increasing number of features. Underlining the key aspect of our library preparation, this analysis clearly signalizes the need for combined capture and usage of mismatch and arrest information, to yield prediction performances neither of them alone can come close to.

3.1.10 Discrimination from other A-Modifications

Above, we demonstrated the distinction potential of m^1A's signature, accessed by the combined use of characteristic features in machine learning based detection. Considering the differential prediction outcome for scenarios of modulated minimum signature intensities, an essential practical requirement is the discrimination of m^1A from other adenosine derivatives. This aspect was addressed in what we termed an *eligibility chart*, a study conceived to obtain a comprehensive, bundled readout of the modifications' distinction potential, based on their typical RT profiles and frequencies in the modification landscape of the primary RNA pool examined in this work. By means of such vis-à-vis representations for modifications of all four main nucleotides, one can not only recognize m^1A's unrivaled role among adenosine modification signatures (Fig. 16), but also identify other promising modification species of guanosines (section 3.1.11.3, Fig. 18), uridines and cytidines (Appendix 6.7 Fig. 32 and 33) of comparable pivotal rank, eligible for future deployment of the concept.

For the generation of *eligibility charts*, we used reference sequences from yeast's cytosolic tRNA and rRNA as well as human mitochondrial tRNA, non-redundantly compiled from the entire available modification-annotated pools at MODOMICS [4] and Sprinzl tRNAdb [139]. After mapping the corresponding NGS samples 1, 2 and 5 from Tab. 1, a 5 bp margin (regions prone to bias/artifacts and CSA^5 is undefined) at both ends of each resulting profile was discarded, yielding 9425 sequence positions left for analysis. Under those, 621 were annotated modification sites, which, by applying a minimum coverage threshold of 10 reads, were further decimated to a total of 604, defined as setting (iv), a scenario based on modification sites exclusively: 88 (A, 8 types), 120 (G, 7 types), 327 (U, 12 types) and 69 (C, 6 types). Due to the strongly uneven distribution of modification type frequencies, a Random Forest binary (type X vs. non-X) classification scheme was used. It was carried out in 5+3+3+5=16 feasible (X has $\geq$5 occurrences) from 8+7+12+6=33 theoretical setups of a 10-repetitions 5-fold stratified cross-validations, each opposing all available instances of a respective modification type X with as many random non-X instances. As per usual, features visible to the RF were a, m, m/a, m_G, m_T, m_C and CSA.

From the pie chart in Fig. 16A, the suitability of m^1A becomes apparent by its relative numeric prevalence among the annotated adenosines, providing enough training material for a proper learning performance. Obviously, its mismatch composition can be easily distinguished from that of inosine, a species also subject to RNA-Seq based approaches [54] and even Random Forest based prediction [108]. To a limited extent, the scattering misincorporation values overlap with those of other modification types (the color legend in form of a ternary plot shows means only), such as the rare $m^{6,6}A$ (rRNA) discussed in section 3.1.8 or the more frequent t^6A (tRNA). Nevertheless, the previously yielded average ~94% AUC from the ROC analysis could be reproduced even in this modifications-only scenario, also entailing a total False Discovery Rate (FDR) as low as 8%, thereby outvying even inosine among the modifications of $\geq$5 occurrences (Fig. 16B). Interestingly, no predominant contributor can be observed for m^1A, in contrast to t^6A, i^6A and inosine, which all were preferably miscalled on actual 2'-O-methyladenosine sites (contributions normalized by frequencies). Fig. 16C introduces two of the remaining key parameters, arrest and mismatch rate, which together shift m^1A's signature into an isolated position within the feature space, resolving uncertainties on the mismatch composition level to a large extent. Refined by m/a and CSA, m^1A can be discriminated highly accurately from other modifications, a crucial prerequisite when envisaging large-scale *de novo* prediction. In contrast, premature conclusions based on sizeable AUC values of species such as 2'-O-methyladenosine that lack any arrest and mismatch rates (here referred to as 'freeloaders') should be refrained from: although theoretically even low CSA values could be indicative for certain modification species, good classification results of 'freeloaders' (if seen as *positive* class) can often be ascribed to the indirect recognition of characteristic inosine and m^1A signatures, which are overrepresented among counterexamples (*negative*) in the binary classification setup (iv). For relative contributions to

FDRs (B), addressing the aspect of *a priori* confusion potential, this bias was accounted for by normalization of *negatives* by class counts. However, compilation of counterexamples in the cross-validation scheme was done without forced uniformity of corresponding representatives of *negative* sub-classes, in favor of a more realistic constellation of modification frequencies.

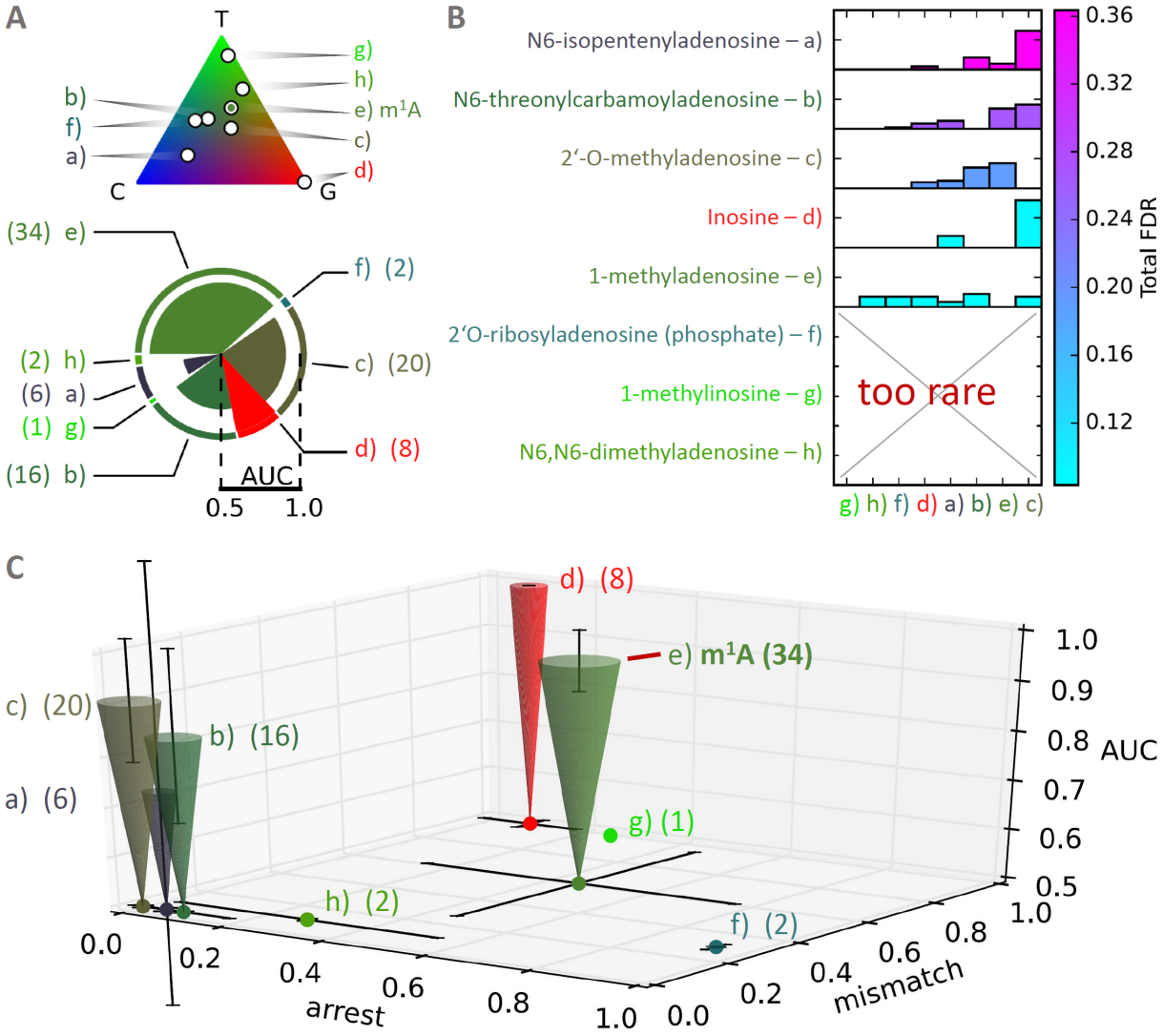


Figure 16: Eligibility chart - Random Forest performance by RT signatures: adenosines. All annotated (MODOMICS) modification sites in yeast cyt. tRNA & rRNA and human mitoch. tRNA sequences were grouped by modification type. Results were determined in 10 repetitions of a 5-fold stratified cross-validation using equal amounts of a specific modification (minimum required frequency = 5) vs. a random composition of 'other'-labeled modifications. **(A)** Colors of pie chart code for mismatch composition displayed in ternary plot. Pie radii reflect Area Under Curves (AUC) from Receiver Operating Characteristic (ROC) curves of a Random Forest model tested for discrimination performance of the modification types. Circular fractions of pies represent the relative frequencies (abs. frequencies displayed in round brackets) of 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 medication frequencies). **(C)** Random Forest performance (AUC) represented as cone height vs. arrest rates, mismatch rates and mismatch compositions (colors). Radii were squared (=normalized) for presentation reasons. Black whiskers indicate standard deviations.

3.1.11 Positioning in Current RNA Modification Field

Encounters of RT enzymes with non-canonical RNA residues become more and more a focus of intense research. Early investigations into m^1A's RT effect were mostly concerned with the application in structural probing *in vitro* [40], and the replication of the HIV genome, which remarkably was found to rely on RT arrest at m^1A_{58} of the HIV primer $tRNA^{Lys3}$ [96, 97]. *In vivo* methods followed [140, 141]. The topic experienced renewed impetus, since RNA-Seq based approaches are being developed to detect RNA modifications on a transcriptome-wide scale. Sequencing methods for m^5C [57], m^6A [62] and pseudouridine [45, 46, 47] have recently revolutionized the RNA modification field, and lately published proceedings in large scale m^1A mapping [85, 71] affirm the common belief that transcriptomes harbor a range of modification types left to be traced by similar approaches. A better understanding of RT arrest behaviors will clearly improve accuracy, in particular of approaches like Psi-Seq, which completely relies on RT stop upon the enzyme's encounter with a CMC modified nucleotide.

Here, we present an in-depth analysis of m^1A's effect on the composition of cDNA products generated during reverse transcription of RNA templates. In contrast to later published transcriptome-wide *de novo* prediction studies [85, 71], this work concentrated on known m^1A instances with well-curated database annotations and prioritizes a comprehensive capture of those feature combinations that make m^1A's RT signature maximally unique. Whereas previous studies aimed at a general picture of various modification types in parallel [105, 101], we focused on a single modification species and characterized its heterogeneous arrest and mismatch patterns in dependence on the sequence environment in over 50 RNA sequences. The herein detected variability of m^1A instances was presented to a computational model, taught to identify the modification in various supervised scenarios.

While large scale prediction is beyond the scope of this project, the various strong machine learning performances indicate feasibility of future m^1A candidate calling, essentially without chemical treatment or enrichment strategy. A direct comparison of accuracy between the mentioned related methods is hardly possible, in parts because they were validated to very different degrees. More importantly, each of them evaluated their performance on highly individual scenarios, distinct modification types of unequally complicated signatures and therefore under dissimilar quality requirements. Bearing that in mind, however, our approach can be ranked at least on a par with prevailing published variants, as demonstrated in Tab. 3 based on supplied information. Essentially, our reference settings (i-iii) (section 3.1.8) represented highly altering, but never trivial challenges. From the outcomes, one can derive that corresponding margins to perfect prediction performance can be ascribed almost exclusively to occasional overlap of m^1A's signature (e.g. with $m^{6,6}A$), rather than to selection and extraction of features, the machine learning model itself or the library preparation approach. Hypothesized that 100% accuracy is a justified common demand for the exemplary cases listed in Tab. 3, our approach outperforms e.g. the HAMR [105] method, which claimed its performance for 'RT-affecting' modifications, thus of course handicapped by modification species of more blurred signature, but also missing out e.g. the arrest based criteria we could exploit maximally in full focus on m^1A. Concurrent Random Forest based methods [142, 108] are challenged by SNP/SNV miscalling in RNA/DNA difference (RDD) detection, and cannot keep up with our m^1A approach under the typical scenario (i). *Dominissini et al.* [85] profit from the additional differential readout for mismatch information upon Dimroth rearrangement of m^1A to m^6A, but don't provide any performance estimate, except an applied 5% FDR limit ensuring that 95% of detected peaks indeed reflect enriched RNA regions, which does not refer to actual m^1A site confidence. Our specified FDR of still acceptable 8% results from the worst case scenario (iv), in which m^1A was identified in a modification-only data set. We can also compete with FDRs claimed in Ψ mappings [45]. Their ROC curve dominates our TPR/FPR tradeoff under very conservative acceptance thresholds, but exhibits early massive breakdowns of specificity, when attempting sensitivities above $\sim$87%.

Table 3: Prediction performance of current methods. Numbers marked with asterisks (*) are valid for specific application scenarios. Values under Hauenschild et al. 2015 marked with (i-iv) result from the corresponding settings in section 3.1.8 (low and high minimum m^1A similarity of non-m^1As) and 3.1.10 (m^1A vs. other modifications). The ribosomal m^1As in setting (iii) were predicted 100 % accurately. Numbers marked with '∼' were estimated/approximated in all conscience based on the authors' claims for certain validation schemes. '-' indicates performance values from *Hauenschild et al. 2015* not specified by the other authors: *Carlile et al. 2014* [45], *Ryvkin et al. 2013* [105], *St. Laurent et al. 2013* [142], *Kim et al. 2016* [108] and *Dominissini et al. 2016* [85] (modification species is specified under author/year). FPR = False Positive Rate (fraction among negative instances, which is erroneously reported as positive) = 1-specificity. FDR = False Discovery Rate (fraction of actually negative instances among all instances reported as positive).

	Carlile et al. 2014 Ψ	**HAMR, Ryvkin et al. 2013** various	**St. Laurent et al., 2013** Inosine	**RDDpred, Kim et al. 2016** RDDs - RNA-editing: A→I, C→U	**Dominissini et al. 2016** m^1A	**Hauenschild et al. 2015** m^1A (+ other eligible)
Approach	CMC + RNA-Seq (Illumina)	RNA-Seq	RNA-Seq (SMS)	RNA-Seq	Antibody capture + RNA-Seq	RNA-Seq (Illumina)
RNA species	whole transcriptome	small RNAs	ribo$^-$ total RNA	mRNA / whole transcriptome	mRNA	tRNA, rRNA, synth. RNA
Readouts	Arrest - -	- Mismatch -	- Mismatch Alignment qualities	- - Alignment qualities	Arrest Mismatch + Dimroth rearrangement -	Arrest Mismatch (-)
Prediction	hardcoded threshold custom procedure	Basic classifier: *k*-Nearest Neighbor (*k*NN)	Machine Learning: Random Forest (RF)	Machine Learning: Random Forest (RF)	statistical algorithm: MACS peak caller	Machine Learning: Random Forest (RF)
Sensitivity	-	92% of 'RT-affecting' mods.	-	90-95 %*	-	96 %$^{*(i)}$
Specificity	-	-	-	-	-	97 % (1-FPR)$^{*(i)}$
Positive Predictive Value (PPV)	87.5-95 %*	-	∼70-80 %*	-	-	97 % (3 % FDR)$^{*(i)}$
Negative Predictive Value (NPV)	-	-	-	75-84 %*	-	96 %$^{*(i)}$
False Discovery Rate (FDR)	5-12.5 %*	∼15 %*	-	-	-	worst: ≤8%$^{*(iv)}$ (m^1A vs. other mods.)
Area Under Curve (AUC from ROC)	∼90-95 %*	-	92-93/94 %*	-	-	≥94 %$^{*(ii)}$

3.1.11.1 Parameters that Shape m^1A's RT-signature

One important message implied in the presented data is the very substantial effect of misincorporation by the reverse transcriptase, which results in a non-negligible read-through efficiency. Occasional 'correct' complementation with dTTP can even partially disguise actual occupancy. According to literature, read-through by RT-enzymes *in vitro* may depend on a variety of parameters, such as Mg^{2+} ions and dNTP concentrations [143] e.g. in case of 2'-OMe modifications [144]. As hinted in our definition of the term *RT signature* (section 3.1), the nature of the RT itself is important in the encounter with an RNA modification [145], which is currently investigated in Prof. Dr. Mark Helm's research group. Furthermore, pH, ion strength and divalent cations can be expected to play a crucial role as well. Whereas these *in vitro* parameters could serve as additional dimensions of potentially differential effects on the individual modification types, they were kept constant in the present study. In doing so, we focused fully on the identification of factors residing in the RNA template itself, such as the immediate sequence context.

Our investigations into the influence of the neighboring nucleotides -1, +1 and +2 unraveled the otherwise unsystematic mismatch compositions by a clear influence of the +1 residue, 3'-adjacent to the m^1A site. While total arrest and mismatch rate seemed unaffected, this fundamental insight has significant implications for all endeavors of transcriptome-wide m^1A *de novo* prediction that rely on mismatch composition. But the findings of this project have bearings beyond: the negative effect of higher order RNA structure on primer extension efficiency has long been known. This issue is frequently encountered in structural probing of rRNA [95], where a strong noise from structure-driven RT-arrests makes signal interpretation towards DMS-generated m^1A residues difficult. Our data, however, suggest that in certain cases RNA structure may even facilitate read-through, exemplified by comparison of m^1A_{2142} in yeast' 25S rRNA (3.1.3, Fig. 3A) and m^1A_9 in the revolver oligonucleotides (3.1.6, Fig. 8A). The latter can reasonably be assumed to be weakly structured [78]. While all revolver m^1A sites showed ~80% arrest and ~45% mismatch rate, the ribosomal sites caused strongly diverging arrest, albeit being equally modified to ~70%. Considering the quite similar immediate sequence neighborhood of both rRNA sites, the discrepancy is indicative of other factors beyond occupancy, such as higher level structures. These influences are yet to be envisaged in ensuing projects.

3.1.11.2 Determinants of Resolution Capacity

As became apparent throughout the discussion of strategy and results of m^1A signature characterization, various determinants set limits to the resolution of the approach and its present application. Even despite the structural influences discussed in the context of ribosomal m^1A sites (section 3.1.3, Fig. 3A), signature intensities are still merely semi-quantitative estimates also of fractional m^1A occupancies in weakly structured RNAs, which deteriorates the overall signature resolution. However, individual m^1A sites show robust correlation of sequencing profiles with corresponding LC-MS/MS data under different modification levels (see section 3.1.4, Fig. 5).

Although system inherent factors such as site-specific read-through efficiency, varying modification levels and sequence dependence infer variability of m^1A's RT signature, they are not considered actual blur, but constitute the natural range of signals ideally recognized in profile inspection. In contrast, the choice of positive and negative training examples based on the available database annotations, certainly introduces biased sequence contexts. That said, best prediction quality by maximum availability of training instances (leave-one-out validation, section 3.1.8) without any debiasing or strategic adaption to a target RNA pool may come at a cost in impartiality, when deploying the signature model to another RNA landscape. Also reliability of the database annotations themselves is decisive for the quality of training input.

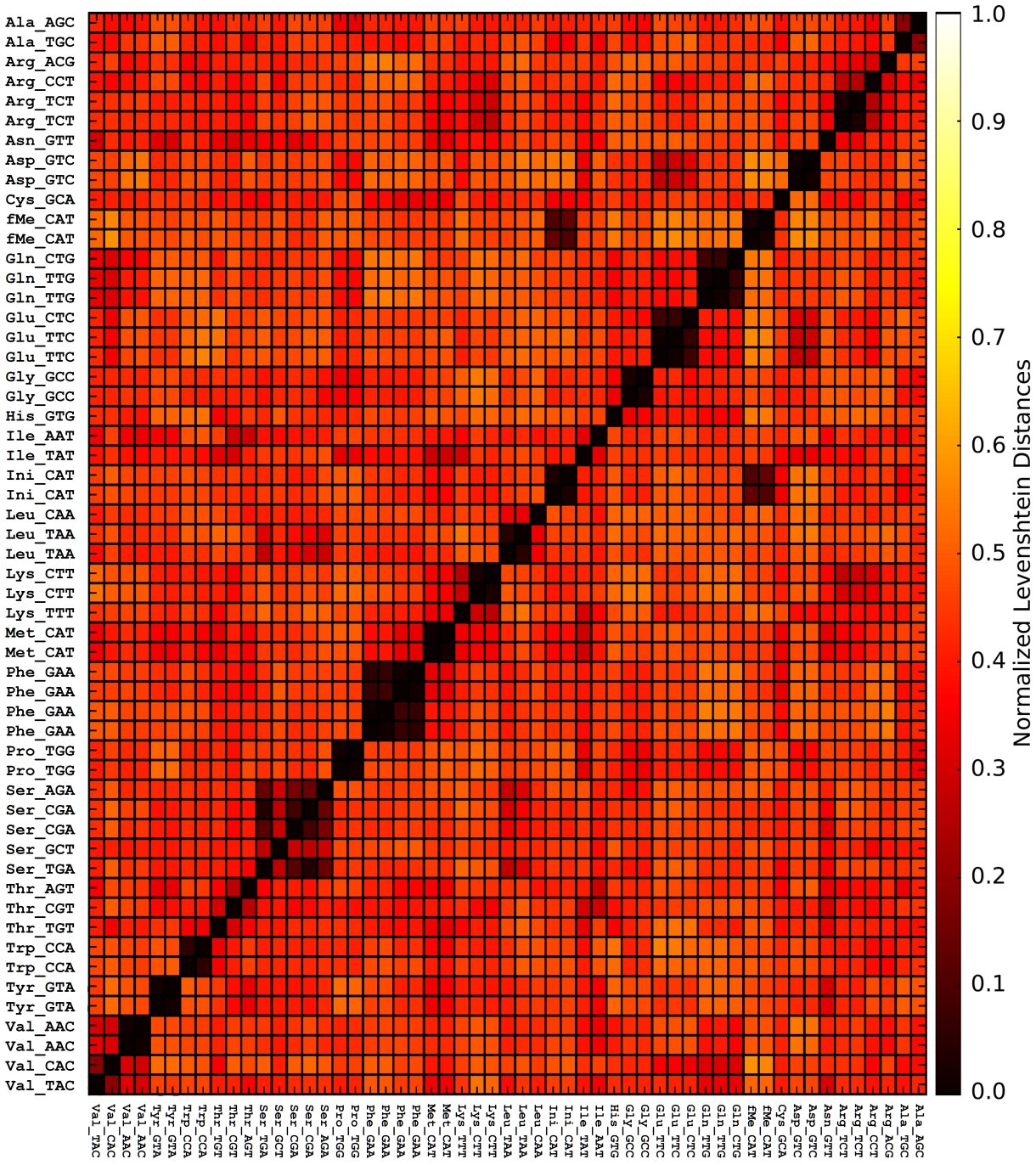


Figure 17: Pairwise Levenshtein edit distances of cytosolic tRNA sequences of *S. cerevisiae*. Normalization was done by division of each distance value by the length of the longer of two compared sequences.

In contrast to these principal limitations, also biases from sequence processing were of concern. Herein, proper handling of ambiguous target sites for reads mapped to multiple similar tRNA sequences plays a central role, since mismatch and arrest information in the reads should represent modification levels of tRNA species they really originate from. By using the k=1 (termed 'k1 regime') setting for Bowtie2, we decided to report one arbitrary alignment only, whenever the mapper found multiple valid target sites for a read. While in its standard setting Bowtie2 instead reports the 'best' alignment, the k1 regime allows to report 'secondary' (with respect to alignment score) mapping sites. Not surprisingly, isoacceptors, i.e. tRNAs charged with the same amino acid [117], show much higher sequence similarity than different acceptors do (Fig. 17), such that most concerns about mismapping must be directed to isoacceptors. Based

on this result, we compared the cross-mapping tendencies (reads mapped to tRNAs of different isoacceptor groups) of three report settings and showed that their impact on m^1A's signature parameters is negligible (Appendix section 6.4, Fig. 28). As therein discussed (page 71), we could keep the initially used k1 regime throughout the project for consistence reasons based on its minor impact, but point out that standard settings ('best') should be preferred for future mapping tasks. In contrast to the mapping strategy, we expect variation and impact of certain experimental parameters more significant, such as salt conditions, temperature and the type of the RT enzyme. These parameters are known to affect polymerization characteristics e.g. in PCR reactions [146], and will be addressed in near future.

3.1.11.3 Potential Applications and Scope

In the rapidly developing RNA modification field, there is an urgent need for new approaches, which can be applied to solve a variety of biological questions. These imply not only transcriptome-wide detection of modified nucleotides but also address quantification of the modification occupancy in response to stress conditions. An example of the latter are elevated temperatures, which were recently shown to ablate a thiol-modification in yeast [147, 148]. Analogously, we have compared m^1A signatures in tRNAs from yeast under normal growing conditions versus 39°C. Although quantification accuracy is limited (compare section 3.1.4, Fig. 5), total ablation of m^1A could be ruled out. In contrast, transcriptome-wide studies found m^1A levels in mRNA dynamically regulated by different stress stimuli, e.g. glucose starvation, and observed variation between tissues [85, 71]. Other applications reside in studies on detection of antibiotic resistance. In analogy to the m^1A_{964} we confirmed in *S. pactum*, where the methylation mediates resistance to pactamycin [84], also modifications of similar RT signature, such as $m^{6,6}A$ could be envisaged, provided a larger set of *bona fide* positive training instances and moderate parameter adaption. This modification could for instance be monitored at position 1519 of bacterial rRNA, where its absence causes resistance to kasugamycin [149, 150].

Although transcriptome-wide application of our Random Forest machine learning model was beyond the scope of this work, the crucial determinants for success of such an endeavor became evident from the supervised validation scenarios. On the one hand, the amount of both, positive and negative training instances is crucial for prediction accuracy. On the other hand, negative training data should include examples of sufficiently pronounced RT signatures, especially if the model is supposed to be deployed on another RNA landscape of distinct modification texture. Also the quality of real m^1A training signatures is of major concern. Besides a biological bias in sequence context, e.g. due to evolution of m^1A methyltransferases [88, 87], also the way training pools are compiled based on database annotations has a strong influence (technical bias). Herein, two interests collide, aiming at maximum, preferably impartial coverage of possible sequence contexts, while calibrating the model to an m^1A signature distribution as natural as possible.

Besides the discovery of new m^1A sites, such as the homology based predictions in trypanosomal tRNAs (Appendix section 6.5, Fig. 29), this project serves as a proof of principle for modification detection based on RT signatures without chemical labeling. Now that the method is established, it can easily be configured for detection of other eligible modifications. The *eligibility* assay introduced in section 3.1.10 revealed $m^{2,2}G$ as a highly promising candidate to be analyzed. The Random Forest prediction performance for this double-methylation illustrated in Fig. 18 is comparable to that for m^1A, thanks to extraordinarily distinct arrest and mismatch rates compared to other guanosine derivatives, and a representative amount of training instances. Interestingly, a second methyl-group may affect RT behavior more than twice as much than a single methyl-group does, exemplified by comparison of $m^{2,2}G$'s and m^1G's signatures.

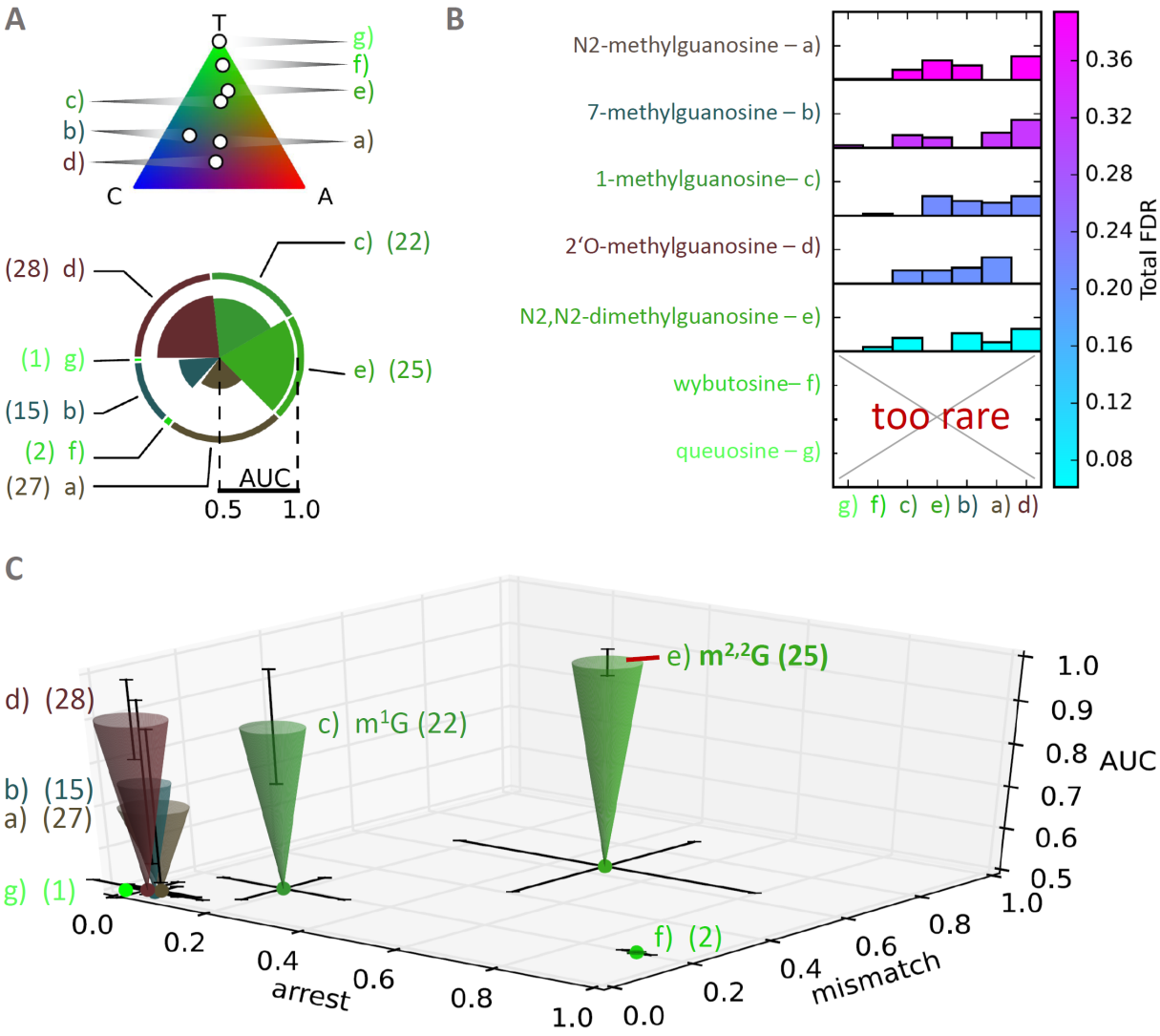


Figure 18: Eligibility chart - Random Forest performance by RT signatures: guanosines. All annotated (MODOMICS) modification sites in yeast cyt. tRNA & rRNA and human mitoch. tRNA sequences were grouped by modification type. Results were determined in 10 repetitions of a 5-fold stratified cross-validation using equal amounts of a specific modification (minimum required frequency = 5) vs. a random composition of 'other'-labeled modifications. **(A)** Colors of pie chart code for mismatch composition displayed in ternary plot. Pie radii reflect Area Under Curves (AUC) from Receiver Operating Characteristic (ROC) curves of a Random Forest model tested for discrimination performance of the modification types. Circular fractions of pies represent the relative frequencies (abs. frequencies displayed in round brackets) of 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 medication frequencies). **(C)** Random Forest performance (AUC) represented as cone height vs. arrest rates, mismatch rates and mismatch compositions (colors). Radii were squared (=normalized) for presentation reasons. Black whiskers indicate standard deviations.

Once the repertoire of natively detectable RT signatures is exhausted, the approach can be configured to find modifications by comparison of usual profiles to those obtained after RT-affecting modification-specific chemical treatments. Similarly, differential profiles for variegated types of RT enzymes as an additional dimension can reveal characteristic responses to non-canonical residues. Visual inspection and automated screening of both, native and differentially treated profiles are the focus of the next chapter of this work, which presents a novel analysis tool for NGS data, *CoverageAnalyzer*.

3.2 CoverageAnalyzer

Besides a holistic experimental concept, characterization of m^{1}A's RT signature substantially relied on specialized software engineered during this project according to the outlined objectives. The need for suitable NGS data processing tools, candidate screening along with site specific extraction and tailored visualization of signature features resulted in the all-in-one platform *CoverageAnalyzer*. This centerpiece of our pipeline generates a direct visual and numerical response to *in vivo*, *in vitro* and *in silico* conditions shaping RT signatures.

Outline. A typical session, conceived to identify, highlight and collect sequence positions that show native or conditional RT signatures is depicted in Fig. 19. Users specify an arbitrary set of NGS mapping profiles (a), which are converted to *CoverageAnalyzer*'s internal format. Therein generated statistics on coverage, mismatch sites, RT arrests and other properties allow subsequent selection (b) of interesting sequences for closer inspection. Detailed plots of the selected profiles (c), allow indication of characteristic sequence positions according to variable feature thresholds. Depending on the sort of data, visualizations of signature features can be used to reveal modification sites either by independent inspection (c_i) or by comparative plotting (c_{i-iii}) of data sets from differentially treated samples. Both options allow the export of numerical feature information, too. For scenarios, in which features shall be spotted in large sequence pools, a screening environment (d_i) allows candidate casting by highly detailed queries. *Via* batch plotting of search results (d_{ii}), the analyst can inspect hundreds of candidates in a highly time-saving manner, skipping forward images outside of *CoverageAnalyzer*, or perform arbitrary ensuing statistical analysis and plotting of exported signature data points.

Format. *CoverageAnalyzer* was implemented as a JAVA/Python hybrid, and released in editions for 64bit Windows, Linux and MacOSX systems. The software runs on modern desktop computers and notebooks (i5+ processor and 4GB+ memory recommended). The program takes advantage of JAVA's platform in-

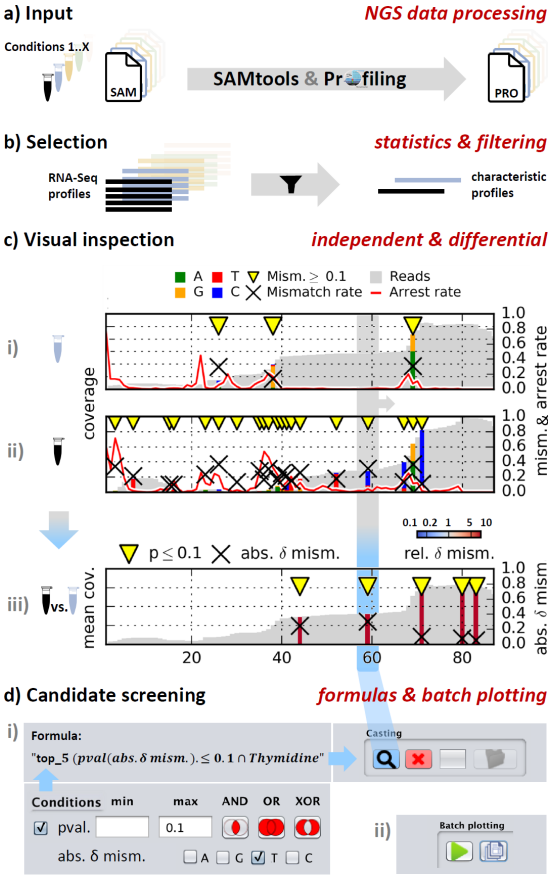


Figure 19: CoverageAnalyzer - Outline. **a**) Input SAM files are processed to a positional profile. **b**) Sorting and filtering of data by various statistical criteria. From the depicted result table, users select sequences for visualization. **c**) Visualization tab. Independent plots and differential 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. Submitted to *Bioinformatics*.

dependence, object orientation, speed and intuitive control elements of its Swing GUI (graphical user interface), while using the highly flexible matplotlib [151] Python library for customized visual data representations. Setup archives including test data sets are hosted on SourceForge under https://sourceforge.net/projects/coverageanalyzer along with descriptions, a detailed manual and an audiovisual screencast that guides novice users through installation and typical analysis steps.

While technical aspects are detailed in Materials and Methods section 5.10, the following sections present the functional range of *CoverageAnalyzer* in the order shown in the scheme in Fig. 19.

3.2.1 Input and Selection

Upon launching *CoverageAnalyzer*, users are prompted to specify NGS profiles in SAM [152] format along with the FASTA reference they have used for mapping. In a live scheme on the Input tab (Fig. 20), the analyst can follow the current progress of data conversion, which generates sorted and indexed BAM and Pileup files *via* included SAMtools [152] before storing positional information in the *Profile* format (see section 5.5). Meanwhile, detailed statistics are created, facilitating preselection of mapping profiles in the next tab of the interface. The format conversion steps take place within the greater context of our workflow presented in section 3.3. Materials and Methods section 5.10 describes special provisions for acceleration, and exemplifies the formulas used for calculation of statistical indices. Ensuing sessions on the same data, do not require repeated data processing and automatically restore statistics from a backup file in the analysis folder. However, the user may manipulate intermediate formats such as Pileup (e.g. for mapping based overhang trimming) and start over processing from the corresponding stage upon prior deletion of the consecutive files (e.g. *Profile*).

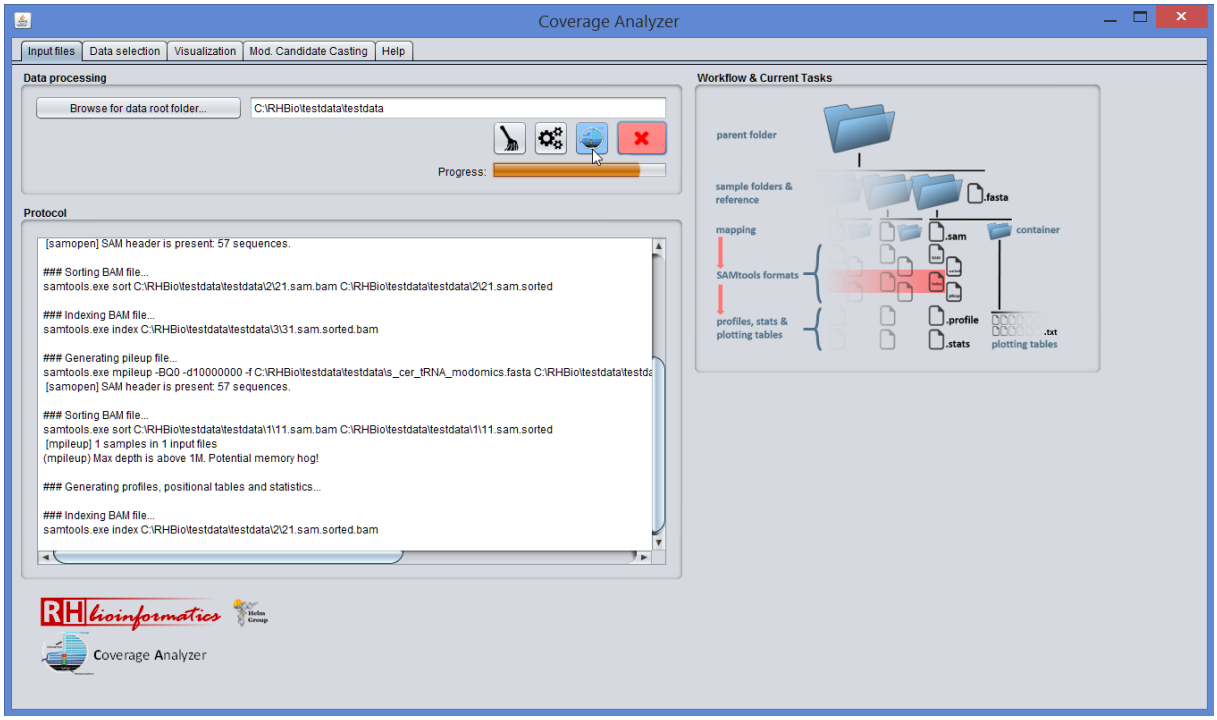


Figure 20: CoverageAnalyzer - Input tab. NGS data sets in SAM format are specified *via* file browser or text field, processed to internal formats and analyzed statistically in preparation for targeted inspections. Progress is shown in a live scheme highlighting the current steps in red, before the results are presented in the Selection tab (Fig. 21).

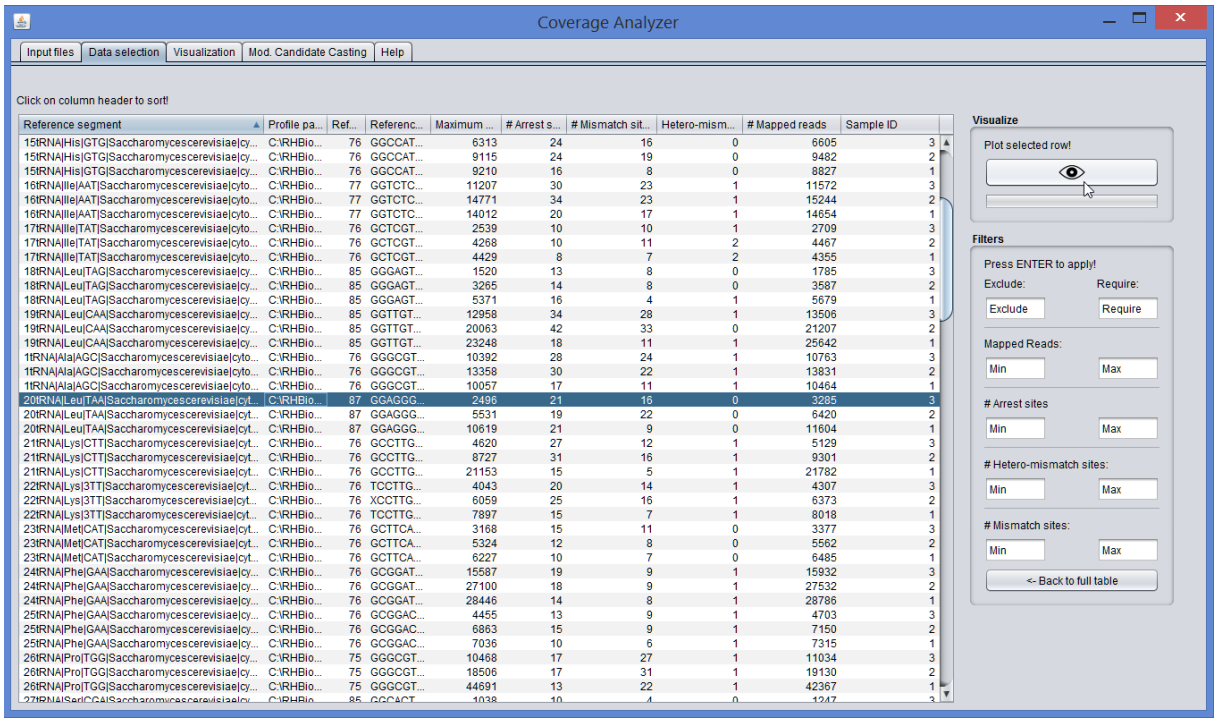


Figure 21: CoverageAnalyzer - Selection tab. NGS profiles of reference segments can be filtered by minimum and maximum cutoffs and sorted by various mapping characteristics including RT signature features. Subsequently, selected profiles can be sent to the Visualization tab (Fig. 22).

The Selection tab (Fig. 21) allows flexible filtering and sorting of reference segments by various criteria. Besides requirement or exclusion of name parts or subsequences of reference segments, also reference length and importantly indicators of mapping characteristics can be used to drastically reduce even large transcriptomal sequence pools to a manageable subset that is worth closer inspection. Useful hints herein are the number of 'significant' arrest sites and mismatch sites, especially those with two or more mismatch components, hence classified as heterogeneous. Calculation formulas of the latter are found in Materials and Methods 5.10. Once the user has identified his profile of interest, also taking the number of mapped reads and peak coverage into account, the table entry can be selected and sent to closer inspection in the Visualization tab (Fig. 22).

3.2.2 Visualization

In the plotting environment of *CoverageAnalyzer*, selected mapping profiles can be evaluated by detailed visual inspection of RT signature features. By exact specification of positional intervals, or usage of a range slider and zoom buttons, the user can navigate to the sequence region of interest. Display of sequence information and activation of a low-detail mode are automatically toggled depending on zoom level and plot dimensions. The latter can be controlled by dragging the window frame or the right-side split pane. In combination with various options for details such as grid lines, legends and sequence letters, these features allow generation of print-ready, high-quality custom plots, which are automatically saved on the hard drive and named by time codes. Two visualization modes can be chosen, which are available under both, independent and differential analysis as described in the next paragraphs.

3.2.2.1 Independent Inspection

In a standard scenario, a user may load several plots representing the same sample, but different reference sequences. Depending on the types of modifications and potential treatments he investigates, he would then inspect the profiles either in 'Mismatch Patterns (MP)' mode, 'Context

Sensitive Arrest (CSA)' mode or both consecutively, typically specifying a minimum threshold for mismatch rate or arrest rate respectively, in order to control the number of highlighted positions. MP mode, which was used for all profiles shown throughout this thesis, reveals full details of positional mismatch compositions for all above-threshold sites, and optionally indicates the course of arrest rate by a red line. The second mode, which is exemplified in Fig. 22, highlights context sensitive arrest rates (CSA) introduced in section 3.1.3 as a complementation of the standard arrest rate curve. Both modes allow focus restrictions on reference base types of preferential interest. Explicit restriction to the central sequence position is applicable, too.

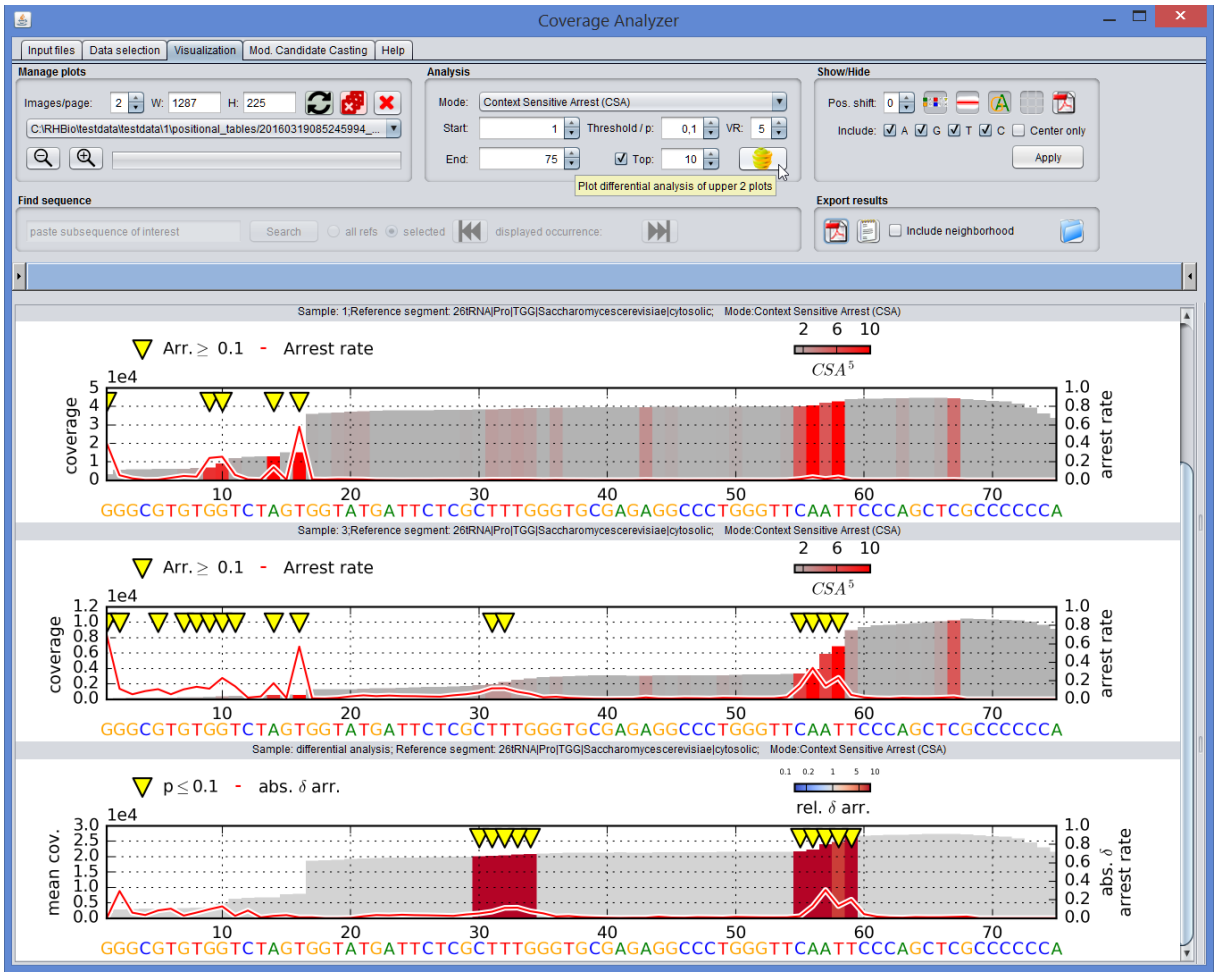


Figure 22: CoverageAnalyzer - Visualization tab. Controls for plot **management** (selection, refresh and delete buttons, images per page spinner, zooming and progress bar) are placed on the upper left, while plot scope can also be specified *via* start and end in the **analysis** panel in the upper center as well as by the light-blue range slider right above the plots. Further applicable analysis parameters are the plotting modes ('Mismatch Patterns (MP)' vs. 'Context Sensitive Arrest (CSA)' with selectable visual range (VR)) alongside respective minimum thresholds, a mark-top-x-results-only option and a differential plotting function. The upper right panel toggles **show**/**hide** options for plot details and automatic PDF generation. It also applies numerical specifications from other panels. Visual and numerical **export** can be managed on the subjacent panel. The example shows mapping profiles of yeast's $tRNA^{Pro\_TGG}$ from a wild type (top) and a chemically treated (middle) sample, which are differentially plotted (bottom) showing the top ten significant positions according to p-value from Fisher's Exact Test. For the differential visualization, the coverages are plotted as average, while the arrest rate (or in MP mode the mismatch rate respectively) is shown as absolute difference by the red line and as relative difference (middle vs. upper plot) by a color gradient. Independent and differential analysis can also be done in automatic high-throughput mode (Fig. 23).

3.2.2.2 Differential Analysis

In contrast to individual inspection, differential plots highlight positions according to maximum thresholds for p-values calculated *via* Fisher's Exact Test, ranking differential positions by significance. A feature also available under individual inspection, becomes particularly useful in differential analysis: where minimum (or maximum in case of p-values) thresholds can be tedious to regulate, in order to obtain the desired discriminatory power for highlighting of interesting events, the mark-top-x-results-only option is a helpful alternative. The result (bottom of Fig. 22) is displayed based on an averaged coverage profile, derived from the single plots of the same RNA sequence present in two samples. *CoverageAnalyzer* indicates both, absolute and relative differences at one glance. Typical applications of differential analysis are wildtype vs. knockout (e.g. of a methyltransferase) scenarios, or comparisons of samples exposed to varied concentrations of chemical reagents that modify specific nucleotides (or modifications) for induction or alteration of interference with RT enzymes as exemplified in Fig. 22. Variation of *in vivo*, *in vitro* and *in silico* conditions entails manifold additional occasions that make differential analysis necessary. Whether assessing effects of stress, characteristics of polymerase species, or optimality of mapping parameters, the user of *CoverageAnalyzer* is provided with direct feedback, not only visually, but also by a thorough composition of all substantial RT signature parameters in numerical format.

3.2.2.3 Export

Positional details from both, independent and differential plots can be exported graphically as single (PNG, PDF) or line-by-line (PDF) images. Deviations of reference positions from canonical numbering due to variable loops or because of purposeful manipulation of the mapping template can be be accounted for by specifying a correctional shift. Numerical export of position specific signature parameters allows collection and arbitrary plotting and statistical analysis by external software (SPSS, QtiPlot, Origin, Excel...) or preparation as input for machine learning models.

3.2.3 High-Throughput Candidate Screening

As experience has shown, collection of modification site specific data points by manual inspection may quickly reach the limit of feasibility, when moving from single sequences to comprehensive sequence pools such as tRNA sets, whole rRNAs or even transcriptomes or genomes. This problem was solved by *CoverageAnalyzer*'s environment for high-throughput screening (Fig. 23) of mapping profiles for candidate positions according to user-defined rules for recognition of complex RT signatures. In contrast to the basic means of common variant callers and SNP identification tools, *CoverageAnalyzer* allows the definition of a highly detailed query, based on combinations of up to more than a dozen different criteria in arbitrary complexity. Independent screening is carried out one-by-one or for the entirety of available samples (data sets). Correspondingly, differential screening is run for a single sample pair or for all possible pairs.

3.2.3.1 Formula Editor

Either for unexperienced users or in simpler application scenarios, the 'required', 'sufficient' or 'mutual exclusive' filter modes can be a recommendable choice. These, generate linear connections of checkbox-activated criteria for significance (coverage, 3'-adjacent coverage) and signature (arrest rate, mismatch, mismatch per arrest, CSA and *diversity* score recognizing heterogeneous mismatch sites as detailed in Appendix 6.6). In the standard case, connectors are all of the same type of Boolean operators, i.e. all AND, all OR, or all XOR respectively. More individualized formulas can be created in custom mode, in which conditions are applicable *via* Boolean buttons in arbitrary order. Beyond the rule work of conventional binding priority (AND > XOR > OR), *CoverageAnalyzer* assists in correct placement of user-desired parenthesis by its built-in automatic recognition and quick-selection algorithm for Boolean expressions. Differential screening

restricts the set of available features to mismatch and arrest, but differentiates between absolute and relative differences, which can be ranked by p-values in the sense of 3.2.2.1. Finally, filter formulas can be stored as presets (e.g. for certain modification species) and restored in later sessions. Besides the formula, search parameters for exclusion of positions and base types are available. Upon a casting run, independent or differential candidates can be statistically analyzed and visualized by external software as mentioned for manually exported signatures in 3.2.2.3.

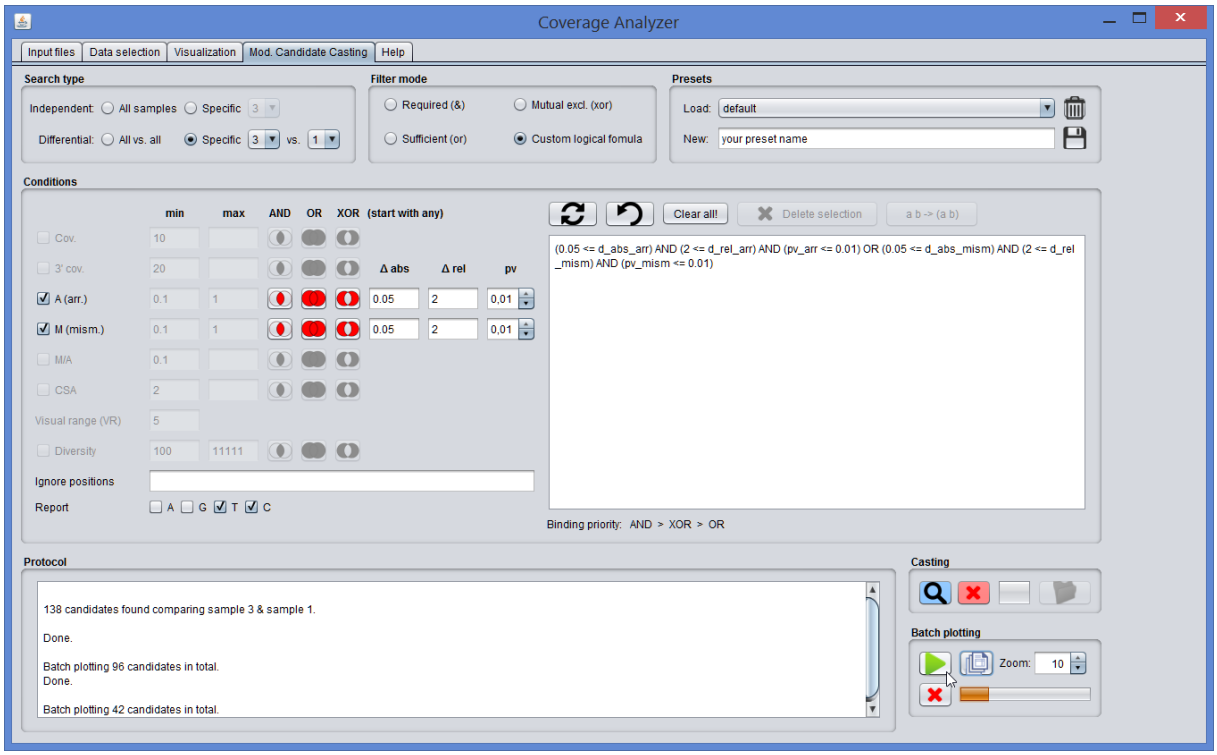


Figure 23: CoverageAnalyzer - Candidate Casting tab. According to the specified **search type**, single or multiple data sets can be screened independently or differentially based on customizable **conditions**, which can be combined using Boolean operators in four **filter modes**. Custom logical formulas can be generated using an intuitive editor that allows composition of filter criteria in arbitrary order and hierarchy featuring parenthesis assistance by automatic expression recognition. Formulas can be saved as reusable **presets** for future sessions. Once completed a filter rule, **casting** can be launched, saving results in *Candidates* format on the hard drive alongside the formula they were found with. By means of the **batch plotting** function, the user generates snapshot image files showing each candidate in profile context for convenient and efficient visual assessment.

3.2.3.2 Batch Plotting

An indispensable feature for efficient case-by-case assessment is the batch plotting functionality, which enables the user to generate snapshot images on the hard drive for each of the candidate sites. Such visual feedback can reveal important details, such as sequence neighborhood or signature features that have not been taken into account by the screening formula. Ideally, this highly efficient and convenient way of qualitative review can drastically narrow down the field of candidates to be further pursued, and thus be a valuable preparation step for a directed design of verification experiments.

3.2.4 Further Comments

Accumulation of alpha errors is a common problem in multiple testing, such as candidate screening on large sequence pools. While it is always up to the user how p-values are used to gauge the significance of findings, we recommend to use techniques like the Bonferroni correction [153]

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 [110]. Both strategies are planned to be included in future releases of *CoverageAnalyzer*. Further improvements include a sequence search already indicated in Fig. 22 as well as a target range of positions to be screened in candidate casting.

3.3 Bioinformatic Workflow

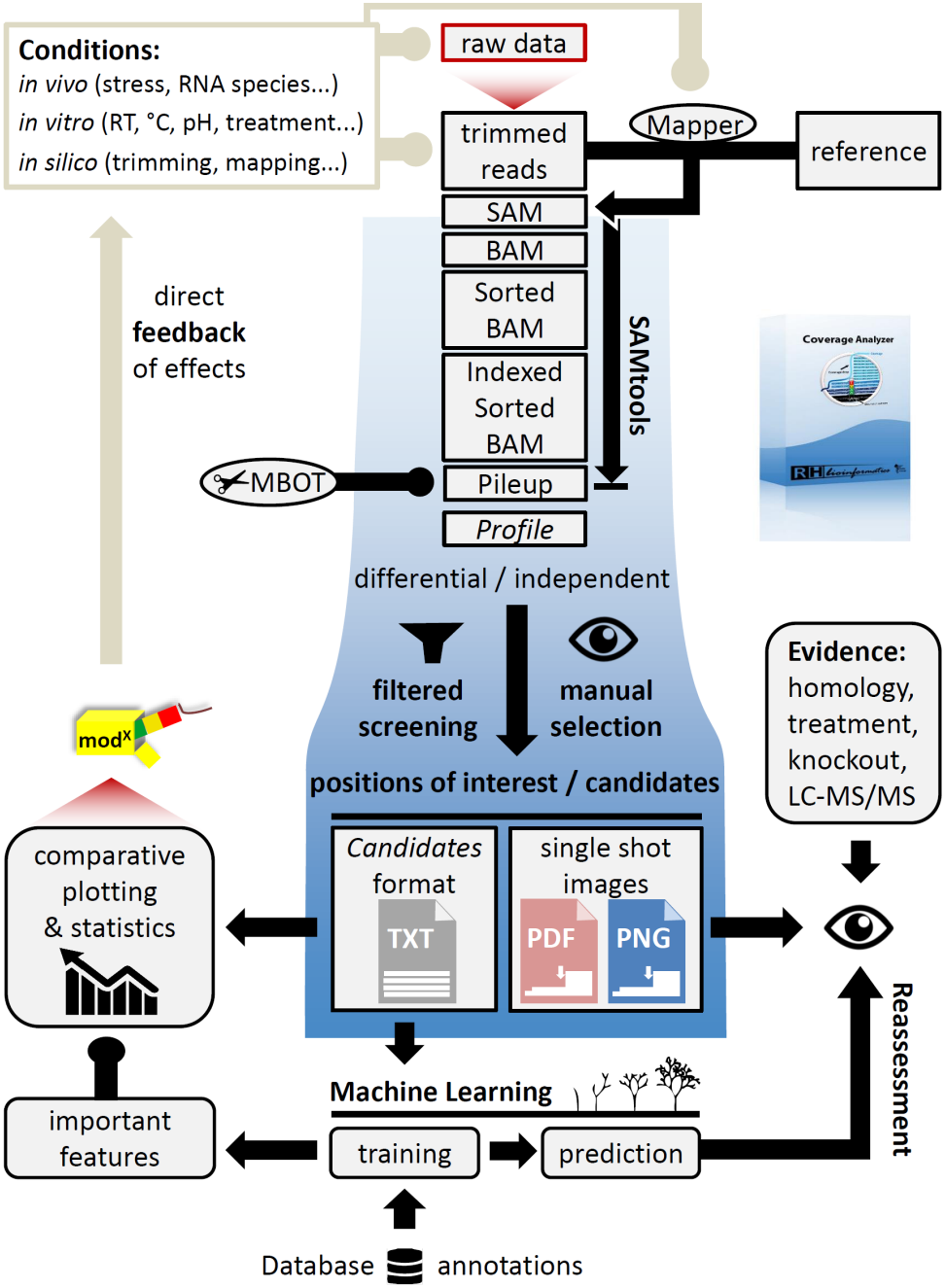


Figure 24: Pipeline for DeepSeq based RNA modification analysis. Input: Raw sequence libraries in FASTQ format. Outcome: i) modification specific RT signatures, ii) trained machine learning models iii) homologous identifications, iv) candidates from visual inspection, threshold based screening or model based prediction, v) visual and numerical feedback for experimental conditions. MBOT = Mapping Based Overhang Trimming (see Methods section 5.3 and 5.5). Processing steps of (binary) Sequence Alignment/Map data formats (SAM, BAM...) are described in section 5.5.

Summary. One goal of this work was the establishment of a comprehensive bioinformatic concept for analysis of NGS profiles with respect to characteristic RT signatures that, provided sufficient native or induced discriminatory potential, can ideally be used for identification of RNA modifications. As becomes clear from the introduced past efforts in detection and description of RT effects, standardized technical approaches may pave the way to a certain degree, but actual breakthroughs in the field require intense focus on single members of the family of modified RNA residues. Therefore, instead of all-encompassing wideband application to multiple modification types in parallel, demonstration of the pipeline's suitability was enacted in form of a holistic characterization of m^1A's RT signature, a highly topical target covering and demanding the entirety of analytical stages in our workflow. This is illustrated by the generalized scheme in Fig. 24, presenting the methodological result of this work.
Besides its role as user interface, *CoverageAnalyzer* covers a substantial portion of data processing steps, and serves as central junction to all orbiting objectives of the analysis pipeline. While *in vivo* conditions, such as stress (section 3.1.11.3) or RNA species were intentionally variegated to investigate influences on m^1A levels, *in vitro* and *in silico* parameters were kept constant at settings empirically chosen to obtain a pronounced and sharp fingerprint of the modification. Nevertheless, the conception of the workflow allows a direct numerical and visual reflection of effects of changes in the experimental setup. Such can include employment of alternative types of RT enzymes or modifications of the library preparation protocol e.g. in terms of adapter ligations and priming strategy, affecting raw read sequences and thus demanding appropriate trimming steps. More typical examples of differential feedback reside in the application of chemical treatments or specific antibodies in order to highlight certain kinds of modified nucleotides in sequence context, even though this study characterized m^1A based on its native signature, only.
Whether acquired through *CoverageAnalyzer*'s differential analysis functionality, or by independent inspection of plots and statistics, extracted information on native or induced RT signatures can be utilized for evidence based identification of unannotated modification sites. Therein, sequence homology, knockout comparisons, chemical treatments and LC-MS/MS data may serve with helpful evidence. In turn, existing database annotations provide the basis for machine learning methods that use the measured properties of modification instances as training data.
Apart from application of the resulting model for detection of novel modification sites, the supervised training procedure itself can already reveal useful information, including feature importance and discriminatory power of signatures, assessed under different conditions in cross-validation scenarios. In this way, one can refine feature engineering on the one hand. On the other hand, under consideration of findings from statistical analysis of signatures, one can derive feedback and strategies for experimental design and generation of sequence libraries. An example is the identified sequence dependence in composition of m^1A-induced mismatches, which was addressed by the *revolver* concept.

4 Conclusion & Outlook

This study was conceived to develop a computational framework for characterization and identification of RNA modifications based on complex RT signatures in Deep-Seq profiles, choosing m^1A as demonstration object in a proof-of-principle. Documented biological relevance, a reasonable number of annotated examinable instances, and existing knowledge of m^1A's native RT arrest and misincorporation tendencies were the major motives. From that starting point, the endeavor aimed to uncover a high-resolution RT fingerprint of m^1A by combination of a tailored biomolecular approach with integral bioinformatic analysis of a detailed digital readout.

4.1 Achievements, Scope and Impact

According to the defined project emphases, a modular analysis workflow was established, including *CoverageAnalyzer* as a stand-alone GUI software for processing, inspection and screening of NGS profiles. Application of these computational solutions directly promoted the successful characterization of m^1A's RT signature and its utilization for identification of unreported sites.

4.1.1 m^1A's RT Signature

Proof of principle. By a holistic experimental and analytical concept, we demonstrated the suitability and significant advantage of a library preparation method that captures full-length and abortive cDNA products. Although m^1A has proven to be a particularly favorable target compared to other modifications, the approach allowed its recognition still solely by a native RT signature, i.e. without any specific chemical treatment. A key factor therein was to identify important characteristics and teach them to a machine learning model. The results show a clear benefit of a combined access to arrest and mismatch information, which should be kept in mind for any modification mapping scenario, in which both aspects can be made available, either directly or by chemical treatment.

Sequence dependent, high-resolution signature. Demonstration of technical suitability of capture and utilization of such a two-fold readout was only part of the study's objective. The major strength of this work lies in the substantial improvement of m^1A's signature resolution. On the one hand, our main set of features, arrest rate, mismatch rate and mismatch components, already allowed a precise statistical description of the modification's fingerprint. On the other hand, we found that arrest and mismatch can vary remarkably even between sites of equal modification levels, which is plausible for strongly structured rRNAs with contingent local influences on RT error patterns. The possibility that structure can even increase read-through, has important implications for RNA probing studies, which directly correlate fractional RT block with fractional m^1A occupancy. By introducing a Context Sensitive Arrest rate (CSA), we provided a helpful parameter to gauge the credibility of candidate sites with respect to observed arrest rates in their immediate environment. The fact that our Random Forest machine learning model preferentially used CSA, reflects the importance of this resolution improvement. A plus of antibody-based methods is the option to estimate m^1A stoichiometry by microarrays, subtracting a gene's proportional expression value by non-precipitated transcripts from the one by all transcripts (100%), leaving the relative number of m^1A-bearing RNAs. Nevertheless, from the comparison of LC-MS/MS data and RT patterns one can conclude that our signatures provide lower boundaries for modification levels and therefore may be used at least as semi-quantitative readout. Moreover, good linear single-site correlation of signature intensity with m^1A occupancy was shown, not only speaking for integrity of experimental and computational workup, but also preparing the ground for future applications like signature-based stress response studies.
Arguably the most interesting finding of this work, a dependence of the mismatch composition on the base configuration at the +1 position 3'-adjacent to m^1A, has drastically enhanced the signature resolution. Based on the observed distinct +1-dependent clustering of mismatch patterns,

assessment of modification candidates with respect to sequential neighborhood could be used to notably increase prediction performance. Adoption of the *revolver* idea would be a promising step in analysis also of other modification types with potential dependence of RT signatures on sequence context, especially under low availability of natural instances.
While large-scale *de novo* prediction was beyond the scope of this project, recently published transcriptome-wide m^1A sites called *via* peak assessment after antibody pull-down, enzymatic demethylation and Dimroth conversion provide an interesting target pool for future investigations. For the moment, a review of the predicted m^1A instances under consultation of the +1 base configuration, may be undertaken to estimate confidence of sites that provide sufficient coverage. By means of undersampling, one can simulate the statistical deviation of measured mismatch compositions from a virtual underlying m^1A with given +1 neighbor in dependence on the number of supporting reads. This allows calculation of p-values for hypothesis-based judgment on acceptance or rejection of new candidates according to plausibility of their m^1A-similarity in terms of mismatch composition given their read counts. Caution should be exercised, since the RT type used in the large-scale prediction runs differs from ours, and so might the pattern of sequence dependence. Provided high-throughput mappings achieve more sequencing depth in future attempts, assessment of mismatch patterns with respect to +1 nucleotides might be adopted as additional criterion in site calling algorithms, calibrated according to preferences of underlying RT enzymes. One should bear in mind, that the statistical distribution of base configuration in m^1A's sequence context in mRNA is most likely different from the one observed in our tRNA/rRNA pool, not least with regard to some enriched m^1A-associated sequence motifs found by *Li et al. 2016* [71].

Limitations. Constraints of resolution, generalizability and applicability of the characterization and identification of m^1A's RT signature can be found on both, the technical and the biological level. Sources of variance and bias in the final results reside in flexible modification occupancy as well as in absolute and relative molecular representation of RNA species. Moreover, RT reaction conditions, non-uniform PCR amplification, sequencing errors, untrimmed library preparation artifacts and mapping ambiguity can influence the outcome. While most of these factors introduce noise to the sequencing profiles, the discussed findings regarding mapping strategies demonstrate that not all variables necessarily affect actual RT signatures at modification sites. Statistical variation of RT patterns between biological and technical replicates was minor. Modeling of a collective signature furthermore neutralizes at least unsystematic fluctuation of single instances, wherefore concerns of impaired classification power by potential qualitative issues can be directed toward unseen input data rather than against inappropriate learning content.
A limitation intrinsic to RNA-Seq is caused by the enormous differences in gene expression levels. It takes ∼10 billion 50 bp reads to quantify 60% of the known human transcripts with no more than 20% error [154], which is not only a matter of cost but also of computational feasibility. Thus, capturing complex signatures with a coverage allowing to read mismatch components in single digit percentage resolution is restricted to the subset of high-abundant RNAs, unless workarounds like CaptureSeq are used to bundle sequencing power in the RNA deep-field [155]. Even better for analysis of modification signatures is immunoprecipitation of RNAs bearing the target, concentrating coverage to narrow regions of relevant RNA fragments only, and not wasting depth in case of fractional occupancy, since all reads represent modified molecules. In the light of dropping sequencing costs and increasing computational power, our RNA-Seq based approach has a good long-term perspective, though. Another advantage is its comparability to the technical workup of a wide range of existing data sets, allowing direct application of our m^1A signature model or adoption of the analytical concept.
The results from simulations of model behavior in supervised scenarios of varied difficulty attest remarkable classification performance. Nevertheless, premature conclusions drawn in view of deployment of the signature model for m^1A prediction on transcriptome-wide scale easily ignore some pitfalls. Supposing that input data for *de novo* prediction is either generated under sufficient methodological conformity with our training background or ideally allows reconstruc-

tion of the signature model based on annotated instances resequenced together with the target transcriptome, another challenge comes up: Whereas a deployed model's average *a priori* distinction power for a certain positive or negative query instance is constant, e.g. at the obtained 96% sensitivity and 97% specificity, the reliability of made predictions highly depends on the relative frequencies of both classes in an unknown query sequence pool. The claimed size of a non-redundant transcriptome varies significantly, e.g. depending on the tissue type: Between 1.5 and 5.3% of ∼3.2 billion genomic base pairs are transcribed in human [156]. Since such a sequence pool is sparsely decorated with m^1A, as deducible from the recent publications of m^1A predictions in mRNA, rejections typically become highly reliable in lack of chance to even encounter positive instances, yielding a good Negative Predictive Value (NPV). Conversely and much more detrimentally, a rising False Discovery Rate (FDR) displaces the Positive Predictive Value (PPV) almost completely due to α-error accumulation by multiple hypothesis testing, such that few m^1A candidates are real positives.

Immediate and implied value of novel sites. The question whether an integrally characterized picture of m^1A's RT signature allows for sensitive and specific identification of the modification was positively answered by the outcome of diverse supervised machine learning scenarios. Instead of immediately launching out automated transcriptome-wide prediction, only to yield large candidate sets to be intensely reviewed and experimentally verified, we preferred to demonstrate the practical value of our signature description by comparison of patterns in homologous RNAs. The results reinforce the view of m^1A as highly conserved modification in different RNA types across species and even kingdoms. On the one hand, the presented novel sites in murine rRNA, human mtRNA and trypanosomal tRNA exemplify the direct application potential of findings of this work for annotation of unreported modification sites within an ample published repertoire of sequenced model organisms or more exotic species. On the other hand, novel sites, ideally those verified by isolation and LC-MS/MS analysis, can be integrated into the m^1A signature model, in order to further improve recognition power in *de novo* prediction on non-homologous RNAs. Homology-based co-occurrence of m^1A often goes along with strong sequence similarity like in the example of human, murine and yeast rRNA. However, here identified or previously known m^1A instances in tRNAs of limited sequence identity demonstrate that actual underlying triggers for conserved methylation are rather structural and thus functional commonalities, of note for the selection of new targets for homology-based identification.

4.1.2 Bioinformatic Solutions

Section 3.3 summarized the rational behind the analytical concept developed for characterization of m^1A's RT signature and identification of modifications in sequencing profiles. The established workflow demonstrates the important interplay between standardized data processing and automated screening on the one hand, and detailed visual inspection and in-depth customized analysis on the other.

Analysis interface. The demand for a tailored tool allowing to retrieve compact numerical and graphical representations of RT signatures from NGS data in order to describe or identify modifications in sequence pools that exceed capacity of unaided review, was supplied by the engineered all-in-one platform *CoverageAnalyzer*. Compatible with SAM, a universal format of mapping results, the software comprises the processing steps up to *Profile*, a human-readable tabular representation of signature-relevant features for each position of reference sequences. Versatile filtering and sorting options facilitate the selection of RNAs of interest for closer inspection. The latter takes place in a multifunctional environment for navigation, plotting and export of signature features, which supports independent and differential analysis of NGS profiles. Manual collection of data points and images is complemented by a powerful automated screening facility, parsing detailed conditional queries of arbitrary complexity phrased by the user

to either gather information from multiple known sites of interest or perform *de novo* casting of modification candidates. One can project that *CAn*'s intuitive graphical interface will add to usage intensity, reaching a clientele without background in bioinformatics. As intended, development of *CoverageAnalyzer* and characterization of m^1A in parallel have mutually promoted their progress. While the software supplied all numerical data and graphical representations of sequencing profiles presented in this work, its functional repertoire was guided directly by the needs of the study. This synergy is one bottom line of the project. A second is that the seamless layout of *CAn* indeed builds a bridge between case-by-case scrutiny and broad-scale survey, as demonstrated for m^1A and continued with the look-out for further eligible targets. Thus, the program closes a technical and conceptual gap in curation and extension of the knowledge base around RNA modifications.

Machine learning. From the positive outcome of various tested scenarios of m^1A instances composed with data points from modified and canonical nucleotides, one can conclude that Random Forests are suitable models for the identification of complex RT signatures. Yet, we learn that this complexity, namely the diverse mismatch composition, is not necessarily utilized in the same way or priority order by computational models as it would be used under human assessment. Instead, it essentially depends on texture and quality of negative instances and is driven by information content. Therefore, also regarding further discussed strengths of RF models, potential improvements are hardly to be found in the way *how* (by which model) features are evaluated, but reside in *what* additional features provide unemployed discriminatory information. Such could be modification specific differential behavior of distinct types of RT enzymes. Then, to cope with the course of dimensionality, the demand for training material increases. To calibrate models towards difficult query data, m^1A-similar negative instances are invaluable and make the difference in training quality and thus in robust classification performance. In turn, for recognition of low-diversity signatures like the one of m^3C (page 76), consisting of homogenous (T) mismatches combined with a sizeable arrest rate, the number of training instances is probably less critical. In fact, one might want to replace an RF model by simple threshold based screening in this case, in favor of transparency and control.

Methodological innovations. In the course of this study a variety of algorithmic and statistical approaches was developed that provided valuable insights into m^1A's RT signature on the one hand, but also represent methodological building blocks reusable in future projects. Among these, innovative value can be attributed to definition and technical acquisition of parameters describing RT arrest. Previously published measures to assess RT stops at potentially modified sites are mostly conceived for differential approaches. Some compare positional counts of mapped 5' read ends by a fold-change between NGS profiles from native and treated samples under normalization with respective total read counts on a reference sequence [95]. Similarly, others compute the positional gain of such 5' counts upon treatment and normalize it with the average positional sum of stops from native and treated samples in a surrounding sliding window [45]. The challenge in direct interpretation of values of the latter kind lies in the semidefinite scale and in the arbitrariness of window size. Instead, HRF-Seq for RNA probing uses the more intuitive 'termination-coverage ratio', an equivalent to what we defined as RT arrest rate $\in [0, 1]$, yet on a single-position basis, but again integrated into a formula for differential comparison [157]. To our knowledge, we were the first to describe technical retrieval of a non-differential positional arrest rate from the popular and universally used Pileup format. Besides using it for characterization of native RT signatures, we implemented the steps in a distributable tool, *CoverageAnalyzer*. The finding that RT signatures of various modifications exhibit certain preference areas for the ratio of mismatch rate m and arrest rate a, encourages also future utilization of another novel parameter developed in this work, m/a. Finally, CSA takes up the sliding window principle from differential studies, here accounting for potential regional tendencies of a, e.g. due to RNA structure, but was designed for non-differential analysis of RT arrest.

Innovations regarding statistical analysis include a novel distance measure for mismatch compositions and a permutation-based matching assessment of 4:4 assignments between data points and clusters, applied for comparison of *revolver* mismatch patterns and natural data points. Conventional indicators of clustering quality were furthermore complemented by a special application of inferential statistics, namely the prediction of 3'-neighboring bases from the mismatch composition of m^1A data points. Practical value of this approach resides in hypothesis testing of potential sequence dependence of RT signatures of other modification types in the future.
Another novel method, presented in the context of mapping strategy (Appendix, page 71), is a normalization technique for Levenshtein distances between sequences, for the purpose of comparability of distance values obtained for sequence pairs of different lengths. Finally, a mapping based overhang trimming algorithm, MBOT, represents a completely new approach to sequence trimming, namely removal of auxiliary sequence elements stemming from library preparation, here ligation assistance overhangs, by comparison of mapped reads and references.

4.2 Prospects

Having characterized m^1A's RT signature holistically, established an analytical concept and developed specialized software for identification of RNA modifications based on NGS profiles, a broad application spectrum awaits future activities. While some findings of this work also raised several interesting ensuing questions, certain potential was recognized for optimization and expansion of methods and tools themselves.

4.2.1 Refinement and Scaling

Polishing and future measurement of signatures. For mere fine-correction of present signatures, or in view of engagement in transcriptomal detection, the experimental and computational pipeline is under steady review. Currently, the library preparation protocol is optimized, e.g. by introducing a circularization step that deprecates overhang-based ligation and certain purification requirements. Goals of the new strategy are improved quality and increased yield of mappable target sequences, while trimming is simplified compared to the current version. Other setscrews to be further optimized but also scrutinized regarding potential modification-specific impact on signatures are experimental parameters like ion strengths, dNTP concentrations, temperature and the RT type itself. On the computational side, future processing of raw reads will involve removal of PCR duplicates for bias reduction. Mapping policies are planned to be revised regarding the discussed advantage of reporting 'single-best' alignments. A challenge would then be a potential switch to the much more comprehensive genomic references e.g. for remeasurement of m^1A signatures in order to extend the model by tRNA instances from sequences missing in the current set of modification-annotated references. Optimal handling of ambiguous mappings is a long-term task, and RT signatures of tRNA instances of modifications can't ever be free of error. A promising alternative to tRNA training data could be the acquisition of recently published m^1A sites in mRNA, provided these are reproducible with sufficient coverage.

Exploration of sequence context. For future investigations on sequence dependence of RT signatures, it is advisable to cover the sequence space of modification instances as complete and uniform as possible, allowing for expedient statistics. The presented sequence contexts of m^1A sites analyzed in this work reflect the non-random distribution of neighboring bases of modifications, a phenomenon especially intrinsic to nucleotides modified in dependence on certain sequence motifs. In particular for the latter scenario, design of synthetic oligoribonucleotides will be a promising alternative, if statistics and machine learning suffer from short natural supply of certain base combinations. Examination of the '+1 information' and appropriate implementation into RF models are promising prospects in terms of m^1A as well as of other modification species. A remaining challenge is the investigation of physicochemical circumstances that cause

+1 dependence as observed for m^1A's RT signature, or even potential cooperative influences of neighboring nucleotides. Valuable insights therefor could be derived from crystal structures.

Further development of CoverageAnalyzer. Whereas the formula based screening facility of *CAn* readily processes long genomic references efficiently, detailed visual inspection was so far conceived rather for the size range of transcripts. In the current version, low-detail mode is activated in favor of navigation performance, if the displayed sequence exceeds a length of 1000 kb. Future porting of the plotting component from Python to JAVA will significantly improve responsiveness of zooming and resizing of plots, allowing for dynamic user experience in analysis even of chromosomal mapping profiles. For the latter, future activation of the at present disabled seek-and-visit sequence search function, will be instrumental. As an alternative to file based export of statistics, real-time display of positional signature information upon mouse hover is planned to be integrated, too. Another valuable functional expansion would be an interface to online resources from NCBI and Gene Ontology, allowing on-demand retrieval of functional details for transcripts that bear modification candidates.

4.2.2 Applications and Transfer

De novo prediction. Obvious potential for application of the signature model and the analytical workflow resides in large-scale prediction of modification sites on transcriptomal level, following a two-fold goal: Reproducibility of published sites can shed light on reliability of such predictions. On the other hand, extension of the map of *bona fide* modification sites by high-confidence hits may provide a better idea of distributional patterns, and can yield targets for isolation and LC-MS/MS-based confirmation if candidates occur in a context of special biological interest. Discussed limitations and caveats in terms of size and texture of the target sequence pools should be tackled by sequencing depth and calibration of prediction models.

Biological questions. Recent studies reporting numerous occurrence of m^1A in mRNA revealed a dynamic response of the modification status to stress stimuli like heat or starvation, a positive correlation with protein production and distinct tissue-specific modification levels [85]. However, concrete association with certain diseases was not intensified and awaits future analysis. Cases like m^1A_{964} in rRNA from *S. pactum*, mediating pactamycin resistance, or $m^{6,6}A_{1519}$ in rRNA of *E. coli* preserving kasugamycin sensitivity [149], call out on application of our approach for detection of antibiotic resistance by focused monitoring of relevant sites also in other organisms. The methods and software developed in this work could also be used for identification of modifications that represent markers for medical conditions.

Further eligible modifications. Charts comparing RT signatures of modifications of all four standard nucleotides revealed $m^{2,2}G$ as a highly promising target for future transfer of our analytical concept, being separable from other guanosines by distinct arrest and mismatch rates, the latter featuring a certain heterogeneity. With its number of annotated tRNA instances qualifies $m^{2,2}G$ for machine learning. Care should be taken to avoid confusion with the signature of the cognate m^1G, exhibiting a similar mismatch composition. In fact, the observed m^1G signatures have weaker intensities than those of $m^{2,2}G$, but may simply stem from lower occupancies. Another eligible target is m^3C, displaying almost complete T-transition in mapping profiles in $tRNA^{Ser}$ and $tRNA^{Thr}$, accompanied by strong arrest rates. Transcriptome-wide search for candidates would ideally be undertaken as differential screening using data from a control sample featuring *Trm*140 knockout, the human ortholog of yeast's m^3C methyltransferase [158]. This way, one might be able to enqueue another modification type into the growing map of the epitranscriptome. In the latter context, the developed workflow and software are instrumental for testing of new chemical treatments that induce specific RT signatures for further unexplored representatives of the rich pool of RNA modifications.

4.3 Quintessence

The combination of reverse transcription and deep sequencing is a powerful instrument for identification of nucleotide modifications in RNA. Our holistic characterization of m^1A beyond the hitherto point of view demonstrated that modified residues can have complex native RT signatures providing significant distinction potential. Full access to the latter puts high requirements on specialized analytical methods and tools, which were developed in this work. By establishment of an integral concept for processing, inspection and screening of NGS data, complemented by machine learning, statistics and a tailored experimental scheme, we created the framework for studies of various eligible modifications with native or inducible RT effects. *CoverageAnalyzer*, engineered as universal standalone bioinformatic platform for specialized analysis of RT signatures in NGS profiles, was the key software on the way to the findings revealed in this work:
As demonstrated for m^1A, misincorporation patterns in RT signatures of modifications can be influenced by the nature of the 3'-neighboring nucleotide. Differentially structured macromolecular contexts should be considered as determinants of signature intensity, too. Whereas estimation of occupancy is restricted to semi-quantitative assessment, m^1A's specific RT effects allow its qualitative confirmation or identification based on sequence homology. Distinction power of RT signatures modeled by machine learning methods essentially depends on balanced and complete availability of arrest and misincorporation information. Moreover, success in identification of modification types is bound to quality and texture of both, training and target data. The findings of this work have significant bearings on the understanding and appropriate utilization of RT signatures. Together with custom-made methodology and a novel multifunctional tool, they represent a valuable contribution to future progress in exploration of the epitranscriptome.

5 Materials & Methods

5.1 RNA Sources

Yeast rRNA was prepared as described in [81], yeast tRNA as described in [114]. The synthetic oligonucleotides were purchased from IBA (Göttingen, Germany). Sequence information is provided in Appendix section 6.3, table 6. An *S. pactum* strain, DSM40530 (DSMZ, Braunschweig, Germany) was cultivated as recommended for liquid media growth [159] with slight modifications. The total bacterial RNA was extracted with TRIzol® reagent according to the manufacturers protocol.

5.2 Library Preparation & Sequencing

The preparation of all RNA samples for sequencing was conducted by Lyudmil Tserovski. He used a specific protocol to generate sequence libraries designed for detection of RT signatures by capturing full-length as well as abortive RT products. The method is based on a previously published version [114, 115], but was optimized in many aspects as detailed in *Hauenschild et al. 2015* [113].

The steps can be summarized as follows for total / ribosomal RNA. rRNA was fragmented ($ZnCl_2$, heat) and size-selected by excision of 50-150 nt bands after polyacrylamide gel electrophoresis (PAGE). A dephosphorylation of both extremities by fast alkaline phosphatase (FastAP), was followed by 3' adapter ligation (RAdapter, activated *via* chemical preadenylation using imidazolide of 5'-AMP's free acid) by T4 RNA Ligase 2, as described in [114] (see also Fig. 25, step 1). This adapter included a single C at its 5' end followed by a random 9 nt sequence (N9), used as individual barcode for every cDNA molecule produced during RT. Excess adapters were treated by 5'-Deadenylase, heat-denatured, and digested by Lambda exonuclease. Next, Reverse Transcription (RT) was performed using RT Primer (Fig. 25, steps 2 & 3) complementary to RAdapter. Superscript III reverse transcriptase synthesized cDNA at 50°C in a 1h reaction. Primers were digested using Lambda exonuclease, followed by treatment with single strand specific Exonuclease I and dephosphorlylation of dNTPs by FastAP. The cDNA 3' ends were tailed using terminal deoxynucleotide transferase (TdT) and ribocytidine triphosphate (rCTP) under conditions shown to result in tails of three C's in > 90%

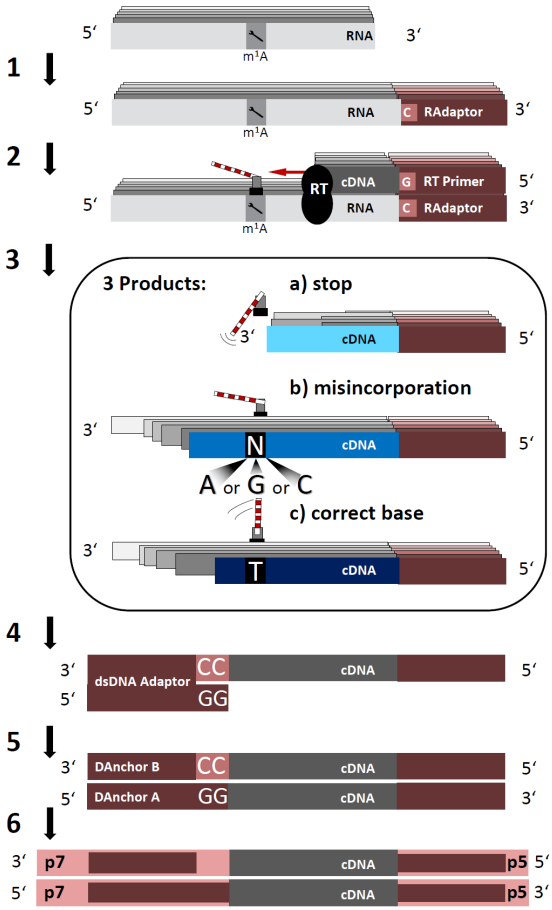


Figure 25: Library preparation protocol. The procedure captures cDNA from both, abortive and full-length reverse transcription events. Primer information is given in Table 6. **(1)** Ligation of preadenylated ssRNA adapter (RAdaptor) at the template's 3' end. **(2)** RT start from hybridized RT Primer. **(3)** Formation of three product types: abortive (a) and read-through with misincorporation (b) or correct base (c) at m^1A position. **(4)** C-Tailing and ligation of double stranded adaptor to the 3' end of the cDNA. **(5)** Complementation to ds cDNA. **(6)** PCR step yields template to be amplified and sequenced using barcoded p5 and p7 Illumina primers. Adopted from *Hauenschild et al. 2015* [113].

of the molecules [114]. A double-stranded DNA adapter was prepared from single stranded DNA oligonucleotides, DAnchor A & DAnchor B, and ligated to the tailed 3' end of the cDNA using T4 DNA ligase (Figure 25, step 4). Taq-Polymerase was used to obtain double stranded products (Fig. 25, step 5) and to amplify them in 7-12 PCR cycles using p5 and p7 Illumina primers (Fig. 25, step 6) containing 8 nt barcodes for multiplexed sequencing. The resulting amplicons contained p5 and p7 sequences required for flow cell clustering, barcodes, sequences of $read_1$'s and $read_2$'s sequencing primers and the target sequence corresponding to the RNA template used for RT. *Via* PAGE size-separation, the molecule size of interest (150-300 bp, which exceeds length of potential adapter dimers) was excised and submitted to *Illumina* sequencing on a *MiSeq* platform in Prof. Dr. Yuri Motorin's laboratory (Nancy, France. See Tab. 1 for details on read lengths). In paired-end mode, $read_1$ corresponds to the 3'-end of RNA template in antisense direction (= 5'-end of cDNA in sense) and $read_2$ to the RNA 5'-end in sense direction (= 3'-end of cDNA in antisense).

5.3 Trimming

The raw reads from the sequence libraries specified in Tab. 1 were demultiplexed in Prof. Dr. Yuri Motorin's laboratory (Nancy, France) using the Casava pipeline and processed further in Prof. Dr. Mark Helm's lab by Ralf Hauenschild using a custom Python/Java based pipeline for FASTQ files. Corresponding to the preparation settings in section 5.2, this accommodated removal of auxiliary sequence elements (primers, adapters, barcodes) not detected in Casava trimming, as well as ligation-assistance overhangs. Leftovers of the auxiliary sequence elements were trimmed by searching for their suffixes/prefixes ($\geq$5) at read starts/ends under a one-mismatch-tolerance accounting for sequencing errors. Overhangs are reported to have a length of three in $\geq 90\%$ of the cases under optimized conditions [114]. Thus, after removal of other artifacts, trimming of up to three 3'-terminal Cs from $read_1$ and up to three 5'-leading Gs from $read_2$, mapping could be carried out widely unimpeded by mismatches due to overhangs. This implied toleration of a small fraction of reads that had actual overhang lengths ≤ 2 and hence was trimmed unjustifiedly, if its cDNA templates ended/began with non-artificial C(s)/G(s). However, therein implied harsh reduction of overhang-related mismatches payed off in a better tolerance of modification-induced mismatch positions, when mapping the reads. A specialized method for posterior trimming of overhangs of lengths ≥ 4 based on the mapping result is presented in section 5.5.

5.4 Reference Sequences & Mapping

Except for sample IDs 9, S23 and S24, reference sequences for all samples listed in Tab. 1 were obtained from MODOMICS [4]. In case of the eligibility charts, the pool of tRNAs (yeast cytosolic, human mitochondrial) was nonredundantly complemented by sequences from the Sprinzl tRNAdb database [139]. Reads from synthetic oligonucleotides were mapped to custom shorter reference templates derived from the $tRNA^{Lys}$ sequence from MODOMICS. *S. pactum* data was mapped to rRNA reference gi|636560031|ref|NR_116091.1| from NCBI RefSeq [160] database, and *T. brucei* data to tRNA references from triTrypDB [161]. The raw FASTQ sequence libraries from Tab. 1 are available upon request from Prof. Dr. Mark Helm. Mapping of the trimmed reads was carried out using Bowtie2 [106] in global alignment ('end-to-end') mode without soft-clipping. Splicing junctions were not taken into account with respect to our focus on mature RNAs. Also paired-end read information was not relevant to the mapping strategy, which was chosen for tRNA and rRNA sequences. In tRNA scenarios, the respective reference sequences were provided to the mapper at once in a single FASTA file. One mismatch ('-N 1') was tolerated in the seed of six nucleotides ('-L 6') with setting '-k 1' reporting only one (the first) alignment declared as valid by Bowtie2 (see discussion in section 3.1.11.2).

5.5 Postprocessing

The mapping results in SAM (Sequence Alignment/Map) format were converted to BAM (Binary Sequence Alignment/Map) files using SAMtools [152]. With the latter tool, the data was also sorted and indexed, and finally converted to Pileup format *via* the *mpileup* function ('-BQ0', 'd10000000' to avoid dumping of read information in regions of high coverage). At the Pileup

Algorithm: Mapping Based Overhang Trimmer (MBOT)

m := reference sequence
l := length of m
B := $\{A, G, T, C\}$ set of base types
$S_m(i)$:= returns set of reads mapped to position i of m, spanning positions $[i..]$
$X_m(B_x, i)$:= returns set of reads $r \in S_m(i)$ starting with B_x
$O_m(B_x, i)$:= returns no. of reads covering pos. [.. i-1] with G and i with B_x
$C_m(B_x)$:= $\{i \in [1..l-1]$ where $m[i] \neq \mathrm{G} \wedge m[i+1] = B_x\}$ set of positions i eligible for frequency check of tailing of base type B_x within m's sequencing profile.
$T_m(B_x)$:= $\underset{i \in C_m(B_x)}{\text{median}} \left\{ \frac{O_m(B_x, i+1)}{|X_m(B_x, i+1)| + O_m(B_x, i+1)} \right\}$ approx. tailing frequency of type B_x, w.r.t. presumable real-sequence starts

```
1:  procedure MBOT(m, l)
2:      for i ← {1 to l − 1} do
3:          E_A, E_G, E_T, E_C ← 0                          ▷ Counters for statistical trimming events
4:          Q_{r[i+1]} ← T_m(r[i+1])·|X_m(r[i+1],i+1)| / (1−T_m(r[i+1]))   ▷ Expected no. of overhang bases at i
5:          for r ∈ X_m(G, i) do                            ▷ Trimming candidates
6:              if m[i] ≠ G then                            ▷ Read begins with G-mismatch?*
7:                  r ← r[2 ..]                             ▷ Remove first base from read!**
8:              else                                        ▷ Follow G-stretch (if any) till...
9:                  k ← 1
10:                 while r[k] = G ∧ m[i + k] = G do
11:                     k ← k + 1
12:                 end while                               ▷ ...non-G in m or r
13:                 if r[k]=G then                          ▷ Evidence found?
14:                     r ← r[k + 1..]                      ▷ Remove first k bases (Gs) from read!
15:                 else                                    ▷ No evidence found
16:                     if E_{r[i+1]} < Q_{r[i+1]} then     ▷ Quota left?
17:                         r ← r[2 ..]                     ▷ Statistical trimming!
18:                         E_{r[i+1]} ← E_{r[i+1]} + 1
19:                     end if
20:                 end if
21:             end if
22:         end for
23:     end for
24: end procedure
```

*Since the Pileup format represents both, C- and G-overhangs in read_1 / read_2 sequences as Gs (actually lower or upper case), i.e. from the perspective of the reference sequence, the algorithm does not have to differentiate between Cs and Gs from raw reads. Positional indices in reads mapped refer to their orientation along the reference after alignment, irrespective whether aligned in sense or reverse complement.
** Whenever reads are trimmed, all variables and functions report the respective updated information of the modified mapping profile, when being called.

stage, the mapping-based trimming of presumable overhang relics mentioned in section 5.3 took place. A Python procedure named MBOT was written to perform the necessary operations according to the above pseudocode of the algorithm. While in three of four cases overhangs can be trimmed thanks to the mismatch between read Gs and reference As, Ts and Cs, the fourth case, a G in the reference, prohibits the differentiation whether a read's tailing G is part of the overhang or true sequence. This case was accommodated by empirical estimation of 'tailing efficiencies' for the four base types on the respective reference. To this end, for each reference sequence m, the profile was screened from 5' to 3' in order to generate four sets $C_m(B_x)$ of eligible positions i of non-G base configuration and 3'-neighbored by the corresponding base type B_x, i.e. A, G, T or C. For the positions in each $C_m(B_x)$, the algorithm determined the relative frequency of reads covering the reference position from their start to i with G-mismatches under all reads having their first base match at $i+1$, i.e. presumably having their real-sequence starting with B_x at $i+1$. The medians $T_m(B_x)$ of these four collections were saved. Albeit termed 'tailing frequencies', these statistics described only relics of overhangs that are left over from the pre-mapping trimming step. Using the resulting ratios as approximation of the likelihood P_{four} by which overhangs are elongated by (at least) a 4^{th} C, MBOT estimated the probability of encountering overhang relics of length n as ${P_{four}}^n$.

First, the main procedure iterated over all positions i of reference m and applied a three-options trimming strategy for the reads r starting at i. Those r starting with a clear G-mismatch were immediately trimmed (lines 6-7). In contrast, those r starting with an arbitrary number of G-matches were traced to the first non-G in r or m at a position $i+k$ (lines 8-12). If r had a G at $i+k$, this was considered evidence of an overhang spanning a G-stretch from i to $i+k-1$ in the reference, and the stretch was trimmed in r (lines 13-14). Conversely, if the G-span was ended by a reference non-G, i.e. no evidence could be found, the read was trimmed off its first base (G), with respect to a contingency E_X of trimmable *bona fide* tailed reads covering i (lines 16-18). E_X is determined based on the empirical median tailing quota of base type X of r at $i+1$ using the number of reads starting with X at $i+1$ as untailed relative complement $1-T_m(r[i+1])$.

Using *median* instead of mean tailing frequencies reduced potential biases from effects like modification-caused high G-mismatch sites or SNPs. The tailing mechanism underlying to this trimming concept is interpreted and modeled as a base-dependent likelihood of repeated elongation events, estimated by $T_m(B_x)$. Index k used for evidence based trimming (lines 13-14) was limited to four, such that G-spans in the reference were followed up to a maximum length of four bases (corresponding to an overhang probability of $T_m(B_x)^4$), as a compromise of computation effort, reliability of the probability based tailing hypothesis and corresponding expectable reduction of bias in mismatch signals.

5.6 Signature Extraction

The Pileup files were further processed to custom tab-separated text files termed *Profile* which contained the important features quickly accessible for each reference position. *Profile* served as source format for Random Forest input generation as well as for visualization and screening as described in section 3.2. While the actual format distinguished between sense and antisense mapping by counting lower and upper case read bases and also counts Ns (unknown read bases), the example in Fig. 4 is shown in a simplified version, uniting sense and antisense mapped read bases as capital letters.

Table 4: Profile format. Adopted from *Hauenschild et al. 2015* [113].

reference	position (i)	ref_base (R)	coverage (c)	matches	A	G	T	C	arrest (a)	mismatch (m)
tdbD00003245|Sacch..cer..|4932|Val|AAC	59	A	7793	3822	3822	858	3057	56	0.612	0.515
...										

Mismatch rate m and arrest rate a, as introduced in section 3.1.3, are determined from the Pileup format, counting commas (,) and periods (.) in the row (pile) p_i as sense and antisense matches and $\wedge$ characters as read starts, such that

$$m_i := 1 - \frac{\sum_{x \in p_i} 1,\ if\ x \in \{comma,\ period\},\ 0\ else}{c_i} \quad \text{and} \quad a_i := \frac{\sum_{x \in p_{i+1}} 1,\ if\ x = \wedge,\ 0\ else}{c_{i+1}}.$$

When *Profile* is screened for modification candidates via *CoverageAnalyzer* (section 3.2), Context Sensitive Arrest rate (CSA, defined in section subsubsec:mismatchrate), 3'-neighboring reference base and respective relative mismatch components $mism_1$, $mism_2$, $mism_3$ are determined and stored together with the other features relevant for signature definition and machine learning (3'coverage 3'c, mismatches per arrest $\frac{m}{a}$, and the *diversity* score mentioned in section 3.1.8 and discussed in Appendix, page 74). The storage format was termed *Candidates* and is exemplified in Tab. 5. For statistical characterization of m^1A's effect on RT behavior, the signatures were extracted from *Candidates* format at the database-annotated positions.

Table 5: Candidates format.

reference	position	c	3'c	a	m	$\frac{m}{a}$	CSA	mism_G	mism_T	mism_C	3'base	*diversity*
22tRNA|Lys|3TT|Sacch..cer..|cyt.	58	1962	2296	0.151	0.891	5.895	0.151	0.604	0.373	0.0240	G	11001.0 ...

5.7 Descriptive Statistics

Synthetic vs. natural m^1As. **(Fig. 9F)** Accordance of m^1A mismatch compositions from revolver oligoribonucleotides with mismatch data from natural instances was evaluated by an outperformance assay. Based on the known +1 base identities of the four revolver data points, the minimum edit distance of each revolver data point to the center (mean) of its corresponding natural cluster was calculated using the same distance formula as for Silhouette coefficients, cohesion and separation (page 25). The *M*ean of these four *D*istances to *C*luster centers, MDC, was now determined for all possible permutations of the true revolver↔natural 1:1 assignment to $4!-1 = 23$ alternative 1:1 assignments. Accordance was evaluated based on two measures: (i) the relative number of permutations outperformed (in reciprocal value of MDC, i.e. ${MDC_{true}}^{-1}$) by the true assignment in terms of MDC, and (ii) the performance (${MDC_{true}}^{-1}$) normalized with (divided by) that of the best assignment: $\frac{{MDC_{true}}^{-1}}{{MDC_{best}}^{-1}}$, which equals $\frac{MDC_{best}}{MDC_{true}}$. The permutation test was repeated using MDCs with weighted contributions of clusters based on respective numbers of data points (cluster size, corrects for differing *a priori* cluster likelihoods and expectable standard error of centers) and mean deviations from their means (intra-cluster variation), which led to comparable results.

Cohesion, separation and Silhouette Coefficients **(Fig. 9E).** Cohesion describes the inverse of average intra-cluster deviation from the respective center, while separation denotes the average inter-cluster center distance. Edit distances were calculated for clusters with more than one data point only. For cohesion, the performance scale was normalized constituting the weakest performance as the maximum possible average deviation of data points from cluster centers, which amounts to 2/3 in the hypothetic worst case of uniform distribution of a cluster's data points to the three corners of a ternary plot. As described for MDCs in the previous paragraph, calculation of means from the four cohesion values was done using relative weights corresponding to the number of data points in each cluster. The maximum theoretical average inter-cluster center distance was used for normalization of the separation parameter. It was set to 100 percentage points, corresponding to three perfectly condensed clusters of contrary misincorporation characteristics.

5.8 Machine Learning

Using the R package 'Random Forest' [121] with standard settings (500 trees, *m-try* = $\sqrt{m}$, where m = total no. of features), machine learning was carried out for scenarios (i-iv) as described in the Results sections 3.1.8, 3.1.9 and 3.1.10. For each scenario, 10 repetitions of a 5-fold stratified cross-validation were performed using a Python script, which managed R based Random Forest training and testing on shuffled data sets. From exported confusion matrices of each fold, sensitivities and specificities as well as positive and negative predictive values were calculated and averaged first within single runs, then in total. This way, standard deviations could be determined as presented in Tab. 7.

In our investigation on sequence context dependent misincorporation (3.1.8), a Random Forest was trained to infer the base type of a corresponding neighbor position from the mismatch composition. Empirical optimization suggested 4-fold rather than 5-fold stratified cross validation for maximum prediction performance. As a consequence, 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) and (63.6, 86.5, 70.6) were obtained for the RF model on average for positions (-1, +1, +2) respectively.

5.9 LC-MS/MS Analysis

Quantification of m^1A in yeast 28S rRNA, trypanosomal tRNA and synthetic oligonucleotides was conducted by Katharina Schmid & Kathrin Thüring using an HPLC-DAD-MS/MS approach (details in *Hauenschild et al. 2015* [113]). Single tRNA species were isolated from *Trypanosoma brucei*'s total RNA (details in *Rubio et al. 2013* [124]) by hybridization with complementary biotinylated DNA-oligonucleotides and subsequent immobilization on streptavidin-coated magnetic beads (Dynabeads). After denaturation, the target $tRNA^{Arg(UCG)}$ was captured using the sequence biotin-CGGCAGGACTCGAACCTGCAACCCTCA. Purification steps included washing (SSC buffer), denaturing polyacrylamide gel electrophoresis (PAGE) and ethanol precipitation. Next, the samples were prepared for LC-MS/MS analysis by digestion into nucleosides. As stable internal standard (SIL-IS as described in *Kellner et al. 2014* [37]) ^{13}C-labeled total RNA from *S. cerevisiae* was added. Calibration solutions and digested RNA were analyzed on an Agilent 1260 HPLC series equipped with a diode array detector (DAD) and a triple quadrupole mass spectrometer (Agilent 6460). Upon a photometrical measurement of column effluents by DAD, MS was operated in positive ion mode using a time-segmented multiple-reaction monitoring (MRM mode) allowing separation of m^1A from other methylated adenosine derivatives by exclusive elution times. Using SIL-IS, a response factor could be determined for ^{12}C-m^1A and ^{13}C-m^1A peak areas and allowed m^1A quantification in the RNA samples. Quantification of A (adenosine) was performed by peak extraction from UV chromatograms and the m^1A amounts were normalized to the A contents of the analyzed RNA molecules.

5.10 CoverageAnalyzer - Software Engineering

Distribution and Dependencies. *CoverageAnalyer* was published under GNU GPL licence version 3, which is displayed during setup. For the Windows release, an SFX setup archive was built using 7-Zip SFX Maker. The Linux and MacOSX editions were packed by conventional zip compression. The software does not interfere with existing Python installations on the user machine, but comes with its own tailored Python environment. Using the package manager Miniconda, the setup routine downloads and installs all required libraries and dependencies, including numpy, scipy (for Fisher's exact test) and matplotlib. For Windows, SAMtools was included as an executable, whereas the Shell-Script setup routine on MacOSX installs SAMtools *via* Homebrew and Linux does *via* apt-get install. JAVA Runtime Environment 1.7+ is required. On Windows, launching CoverageAnalyzer.EXE (created *via* launch4j) redirects the user to Oracle's

download website, if necessary. Under MacOSX, JRE is usually preinstalled, and under Linux, the installation is automatically ensured by the installation script *via* apt-get install. Included JAVA dependencies are the libraries *itextpdf* (PDF generation), *prefuse* (sequence range slider), *commons-exec* (threading) and *swing-layout*.

Tuning Strategies. In order to enable usability of *CoverageAnalyzer* for longer mapping templates, such as rRNA, mRNA and even chromosomal reference sequences, a tiling method was implemented, which partitions the *Profile* data into chunks of 1000 lines saved as positional files. These files are indexed by according $x\_y.txt$ name tags, where x represents the reference number and y the y^{th} block. This allows fast access to a query region of interest without reading or memorizing the leading positions 1 to $y-1$ of the reference. Further acceleration of the program was accomplished by a low-detail mode automatically toggled in response to the zoom level. Moreover, implementation of multithreading drastically reduced runtime of both, input processing and candidate screening. Parallelization enables simultaneous conversion of mapping data from several samples to *Profiles* and blockwise positional tables, according to the number of available CPUs. Casting of candidates in multiple samples is parallelized in the same manner, but additionally takes advantage of a highly performant hashing based decomposition and interpretation mechanism for Boolean expressions and hierarchical conditions based on the user's screening formula. The interplay of the involved functions and objects is illustrated in a simplified scheme in Appendix section 6.1, Fig. 26.

Mapping statistics. In the Selection tab of *CoverageAnalyzer*, sorting criteria comprise reference ID, file path, length, sequence (first 100 nt), coverage peak, number of 'high-arrest' sites (S_A), of 'high-mismatch' sites (S_M), of 'heterogeneous mismatch' sites (S_H) and of mapped reads. Let c be the coverage at position i of reference f of length n. Let R be the reference base at i. Let $F_b(f,i) := \frac{obs.(b,i)}{c(f_i)}$, where $b \in \{A,G,T,C\}$ be the observed frequency of base type b covering i in f. Thus, $mF(f,i) := \{F_b(f,i)$, with $b \neq R\}$ is the set of mismatching $F_b(f,i)$. All i with $c(f_i) \geq 20$ contribute to S_{H_f}, if two or more mismatch components exhibit a minimum relative coverage contribution of 0.1:

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

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

6 Appendix

6.1 CoverageAnalyzer - Example of Software Architecture

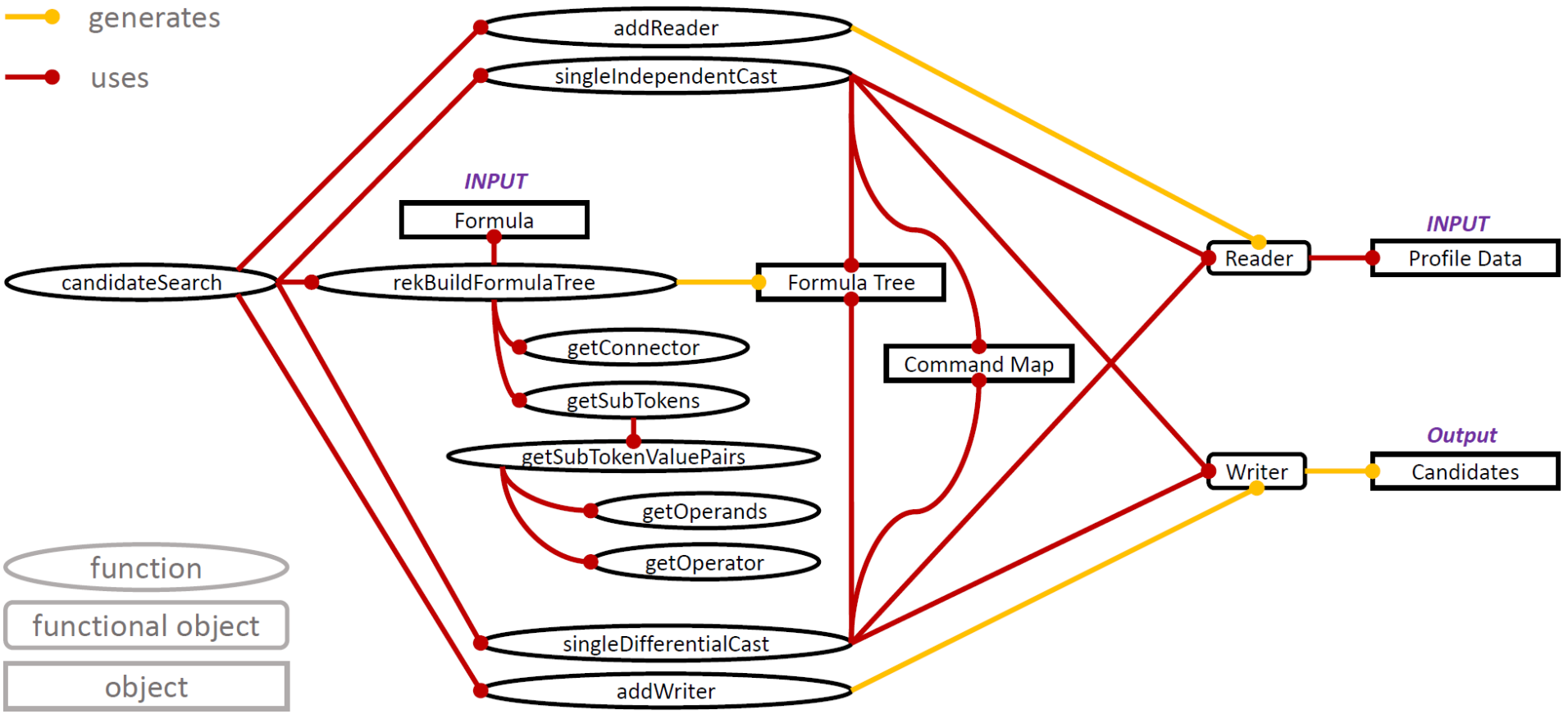


Figure 26: CoverageAnalyzer - Architecture of candidate screening system. Simplified scheme of the interplay of functions and subfunctions, interpreting the tree of conditions represented by the user-generated Boolean formula, and screening *Profiles* for agreeing candidate sites. A recursive function decomposes the input formula into its tokens (conditions), which consist of connectors (Booleans) and subtokens. In this way, a tree structure is built from the condition hierarchy. The leaf level is occupied by statements of the types $y <= x$ or $x <= z$, which can occur as pairs such as e.g. $(0.05 <= mism.rate, mism.rate <= 0.9)$. Therein, x, y and z are recognized as values or parameters and $>=$ as operator. Then throughout screening, a hash based 'Command Map' channels candidate sites along the formula tree evaluating their features in a depth-first approach until a termination condition accepting or rejecting the candidate.

6.2 Wildtype and Knockout Profiles

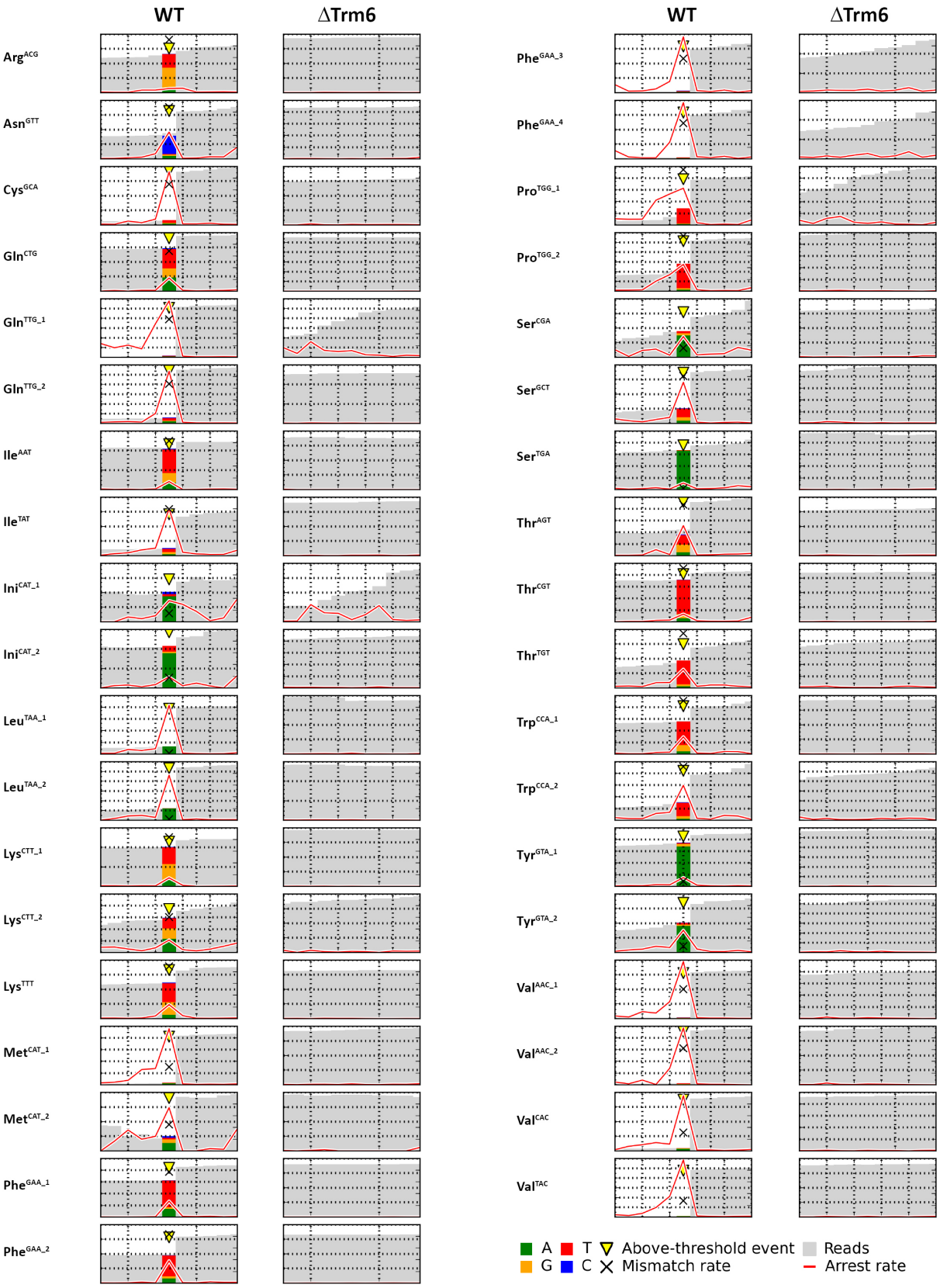


Figure 27: m^1A_{58} signature compilation from 37 cytosolic tRNAs of wildtype and m^1A-negative (ΔTrm6) *S. cerevisiae*. Plot scope: 5 bp upstream and 5 bp downstream of m^1A. Arrest and mismatch rates range on a $[0, 1]$ scale, normalized to height of each plot as in results subsection 3.1.3, figure 3. Adopted from *Hauenschild et al. 2015* [113].

6.3 Library Preparation

Table 6: Library preparation: sequence elements

Element	Sequence
RAdapter	5'-P-CNNNNNNNNNAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-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'-AATGATACGGCGACCACCGAGATCTACAC55555555ACACTCTTTCCCTACACGACGCTCTTCCGATCT-3'

Barcodes used with P7	Sequence	Barcodes used with P5	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^1A-G	12	5'-CACUGUAAm^1AGCUAACUUAGC-3'
revolver m^1A-C	13	5'-CACUGUAAm^1ACCUAACUUAGC-3'
revolver m^1A-U	14	5'-CACUGUAAm^1AUCUAACUUAGC-3'
revolver m^1A-A	15	5'-CACUGUAAm^1AACUAACUUAGC-3'
hybridization oligo for $tRNA^{Arg\_UCG}$	S24	biotin-CGGCAGGACTCGAACCTGCAACCCTCA

6.4 Multiple Mapping on tRNAs

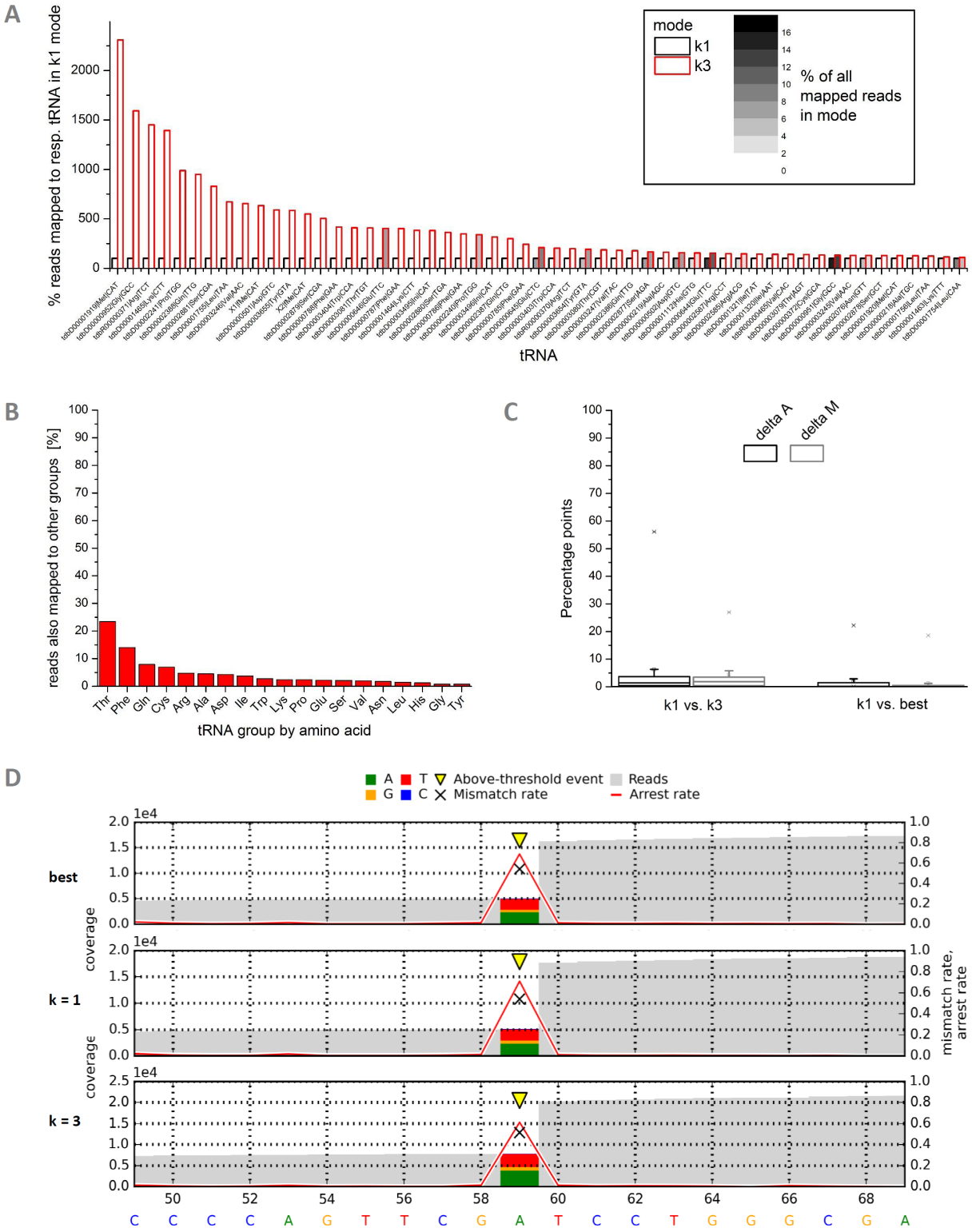


Figure 28: Impact and handling of multiple mapping problem. (**A**) Relative read count comparison of k = 1 ('report best only', set as 100% for each tRNA) and k = 3 ('report best three') modes for valid alignments by Bowtie2. (**B**) Isotype confusion behavior for k = 3. The red bars indicate the relative 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 $tRNA^{Val}\_^{AAC}$. Adopted from *Hauenschild et al. 2015* [113].

Since under k1, Bowtie2 can stop looking for better alternatives once it has found a valid alignment, instead of testing all other possible mapping sites, k1 can be interesting in runtime-critical applications. Because runtime is not an issue w.r.t. our limited reference sequence pool, the only scenario, in which k1 is preferable to 'best', can occur if a read has mismatches due to modifications. In such a case, 'secondary' alignments, inferior in mapping quality compared to 'best' alignments, may be ignored under standard settings, which could redirect important mismatch information to similar but unmodified sequences, resulting in biased signatures. Report settings with k$\geq$ 2 report valid alignments per read up to a certain limit, e.g. up to three results for k=3 ('k3'). In case of several isoforms of one tRNA acceptor type, such a setting tends to distribute reads to more than one target site (multiplication) and thus equalizes the coverage profiles among isoforms. Of course this enforces erroneous mappings, since every read can have only one real molecular origin. Certainly, none of the strategies can solve the mapping problem perfectly. However, with further progress of the project, it turned out that standard ('best') setting is the overall most recommendable, in order to retrieve correct molecular origins when dealing with mutually similar tRNA isotype sequences.

Because earlier results had been generated under k1, the setting was kept for the subsequent data sets presented in this work for consistence reasons, but only after excluding a significant impact on the quality of m^1A's RT signature. In the first step of an in-depth analysis, the reference pool used for mapping of yeast cytosolic tRNAs was analyzed for pairwise sequence similarities based on the Levenshtein [162] edit distance. By normalizing each distance with the length of the longer of both sequences, absolute distances were penalized harder for shorter sequences than for longer ones, resulting in a penalty in relation to sequence length, thus comparable between all pairs of tRNAs. As can be recognized in section 3.1.11, Fig. 17, sequence similarity is much higher among isotypes than between different acceptor groups ($tRNA^{Ini}$ and $tRNA^{fMe}$ are both initiator tRNAs). Therefore, under any of k1, k3 or 'best' settings, most concerns of mismapping should be directed to the similar isoacceptors, while cross-group similarities are minor in comparison. Among isoacceptors, mismapping can be considered less harmful, assuming that also the modification levels responsible for mismatches in the reads should be much more similar among isotypes than across groups.

When the amounts of mapped reads was relatively compared between k1 and k3 for each single reference (Fig. 28A), the disadvantages of both strategies became apparent. For several references, k3 can in fact outnumber the reported mappings of k1 by more than an order of magnitude. This happens preferentially in case of shorter reads from subsequences shared in similar variants by multiple tRNAs. Under k3, these reads can be mapped also to those tRNAs listed after the tRNA sequence, which yields the first valid alignment in the reference file. k1 in contrast, never reaches these alternative alignments due to the sequential processing order of references. Thus, reporting the first valid alignment only, introduces a strong group-internal representation bias among isotypes on the one hand, but avoids over-representation of tRNA groups with many isotypes, which was observed under k3. Fig. 28B shows that even under k3, most multiple mappings happen within the same isotype group, which is a first requirement when characterizing m^1A's RT signature under a k1 setting. Actual legitimation of k1 for our purpose was shown by comparison of the signature features under settings k1, k3 and 'best'. As illustrated in Fig. 28C and exemplified by the $tRNA^{Val\_AAC}$ profile, the tRNA profiles had reference sequences unique enough to be almost unaffected by the reporting mode in terms of m^1A signature.

6.5 Trypanosomal m^{1}As

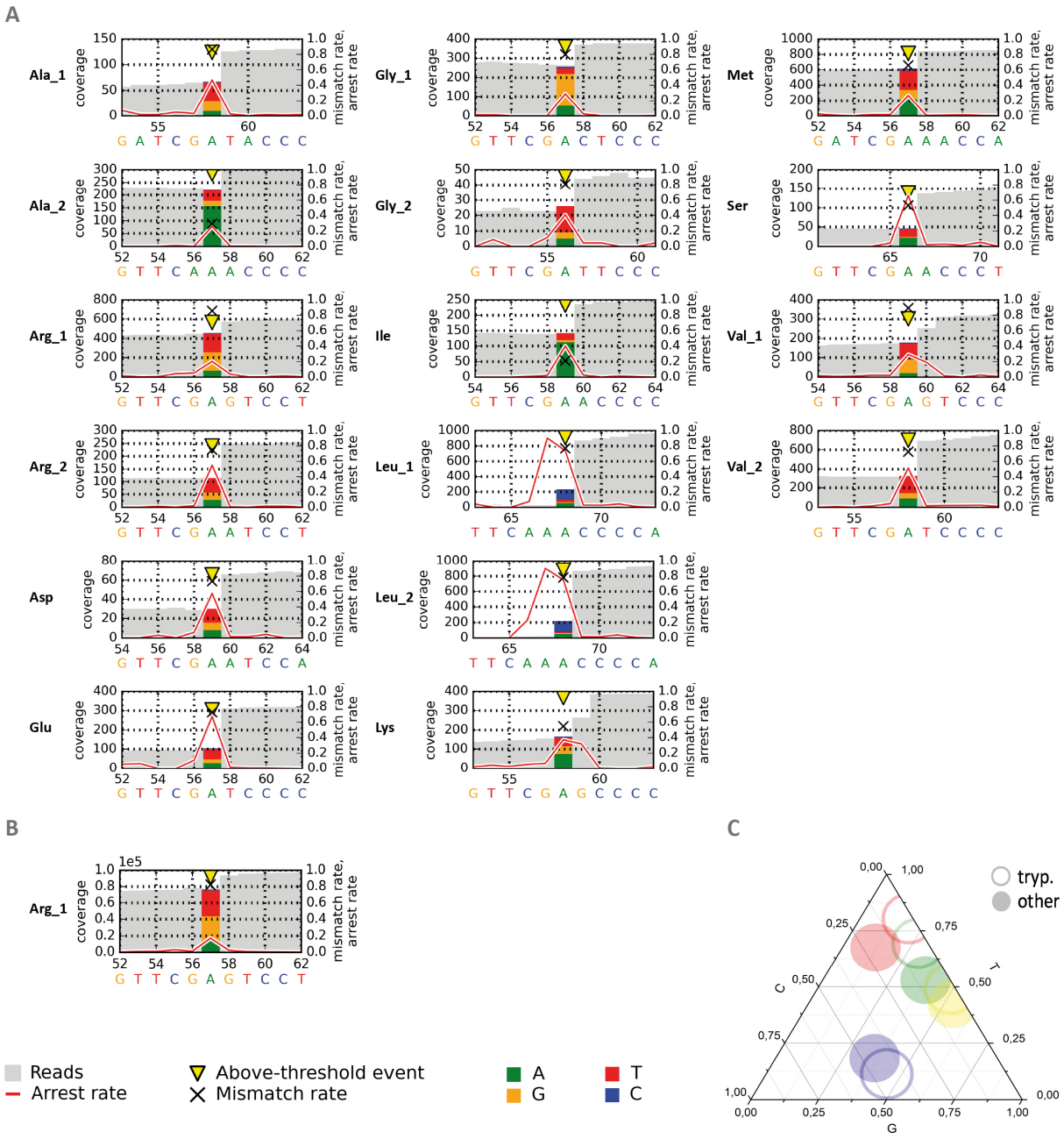


Figure 29: Application to unannotated trypanosomal m^{1}A. (**A**) 16 single tRNA profiles from total tRNA preparation from *Trypanosoma brucei*. (**B**) Sequencing profile of a purified sample of tRNA^{Arg_UCG}. (**C**) Mismatch composition by base configuration at position +1. The data points taken from (**A**) were treated in the same way as described in section 3.1.6 and averaged, then visualized as open circles. For comparison, remaining m^{1}A data averaged from Figure 7C are plotted in full circles. Adopted from *Hauenschild et al. 2015* [113].

6.6 Prediction Dynamics

Table 7: 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 fold-wise outcomes for the corresponding performance measure. Settings (i) and (ii) are defined in section 3.1.8. Adopted from *Hauenschild et al. 2015* [113].

performance for class m^1A (avg. of 10 runs)	low resemblance (i)	high resemblance (ii)
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 8: Random Forest performance - 10 rep. leave-one-out cross-validation. Settings (i) and (ii) are defined in section 3.1.8. Adopted from *Hauenschild et al. 2015* [113].

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

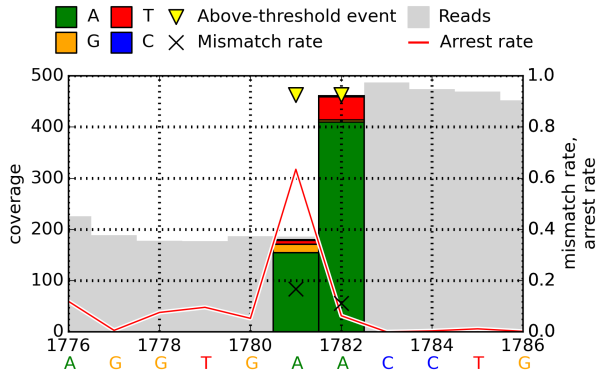
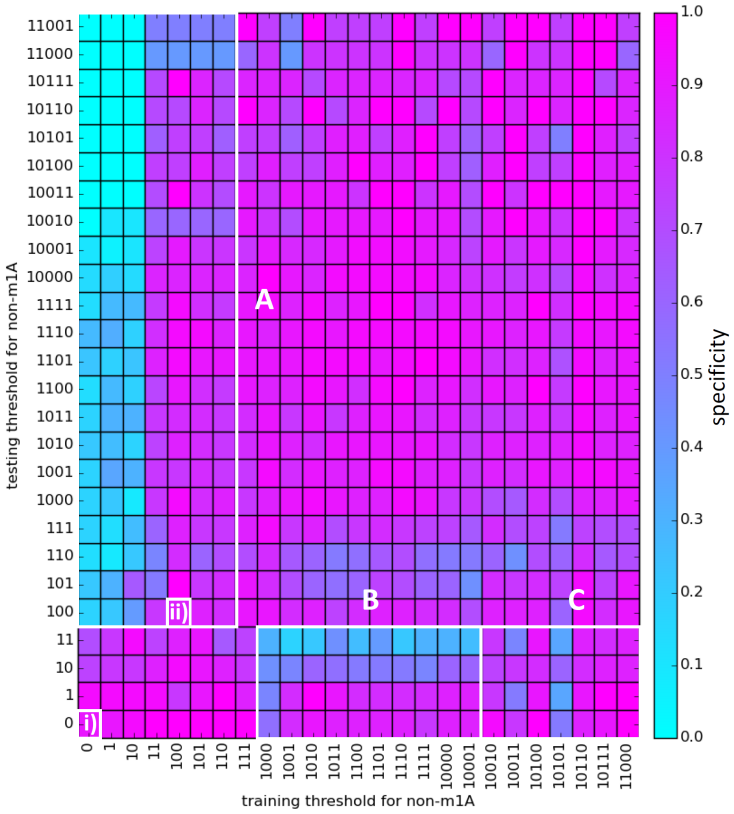


Figure 30: RT signatures of $m^{6,6}As$ 1781 and 1782 in yeast 18S rRNA. For a position p, the arrest rate reflects the relative amount of mapped reads ending at $p+1$, i.e. not covering p. Adopted from *Hauenschild et al. 2015* [113].

Figure 31: Excursion. Performance by texture of training and testing data: specificity. For each tile of the heatmap, non-m^{1}A signatures for training and testing are selected according to a minimum threshold for their *diversity* score. *Diversity* d of a single non-m^{1}A signature is represented as a 5-bit binary code, e.g. 10101 with the i^{th} bit set to 1, if the i^{th} of properties $\{(m \geq 0.1 \ \wedge \ m - max(m^G, m^T, m^C) \geq 0.1, m \geq 0.2, a \geq 0.2, CSA \geq 2, 10 \geq m/a \geq 0.1\}$ is fulfilled by the instance. Correspondingly, $d := \sum_{i=0}^{4} f \cdot 2^i$, where $f = 1$ if condition $5 - i$ is fulfilled and $d \in [0, 31]$ is the decimal representation. In fact, most actual m^{1}A sites have a diversity $\geq$ 11100. Thus, a higher non-m^{1}A *diversity* score mimics higher similarity to m^{1}A. The bit code serves two purposes, providing a rough impression of pronounced global signature parameters at first glance, while uprating particularly such non-m^{1}As meeting the most characteristic features of m^{1}A by the implied exponential weighting. Random Forest (RF) performance (here specificity with m^{1}A as *positive* classification outcome) was then determined in a 10-rep. 5-fold stratified cross-validation with data points selected like in settings i) and ii) in section 3.1.8. The brindle appearance of the heatmap can be ascribed to both, the stepwise sudden activation of *diversity*'s features i judging eligible non-m^{1}As by one single arbitrary cutoff, and to subsequent combinatorial activation of subordinate thresholds $\in [i+1 \ , 5]$. In these terms, *diversity* has a highly discrete character and is, due to its strict exponential weighting of arbitrarily ordered parameters, only a non-monotonous estimation function of actual m^{1}A similarity. A smoother result would be obtained for a weight-neutral continuous similarity measure S based on the 5-dimensional feature space, but the qualitative indication, i.e. which features are mainly responsible for this resemblance, would be lost in that case and feature values could still fluctuate intensely among elected non-m^{1}As of similar S. The very limited amount of *diverse* non-m^{1}As progressively promoted disordered model performance on the right-side and upper border areas of the heat map. The lack of instances made clipping of the assay necessary, where meaningful model testing in the validation scheme was not feasible anymore, close to $d = 11111$. For better orientation, we marked settings i) and ii) from section 3.1.8 in the map. Region (**A**) demonstrates how the rejection of non-m^{1}As with increasingly diverse RT signature is corrupted (turquoise) under soft training conditions too rich in plain adenosines. (**A**) is flanked by a huge field of combinations with widely acceptable outcome. (**B**) illustrates how the high-weighted mismatch feature $m \geq 0.2$ ($d \geq 1000$) decimates the training population to a learning background based on which the RF failed in rejection of most non-m^{1}As that have a $CSA \geq 2$ or fulfill $10 \geq m/a \geq 0.1$ in the tests. Only the again intensely decimated non-m^{1}A testing instances $a \geq 0.2$ (atop of (**B**)), which often naturally exhibit mismatch properties, could be sensitively rejected (purple). (**C**) basically resumes (**B**), but had intermittent good specificities, where a certain combination of population-shrinking thresholds coincidentally favored the model's rejection quota for non-m^{1}A in the tests. Of interest, the maximum number of decimation steps by an order of magnitude a starting population of non-m^{1}As can cope with before the assay needs to be clipped, is an upper limit of therein possible improvements of specificity by an order of magnitude. With respect to the above discussion of *diversity*'s scope, i.e. weighted quick-selection and qualitative information on m^{1}A candidates, we emphasize the limited use of this parameter for other purposes. Here, *diversity* was used in an excursion demonstrating the core message, the impact of training quality on model performance, which should be taken into account in any potential application.



6.7 Discrimination of Modification Types

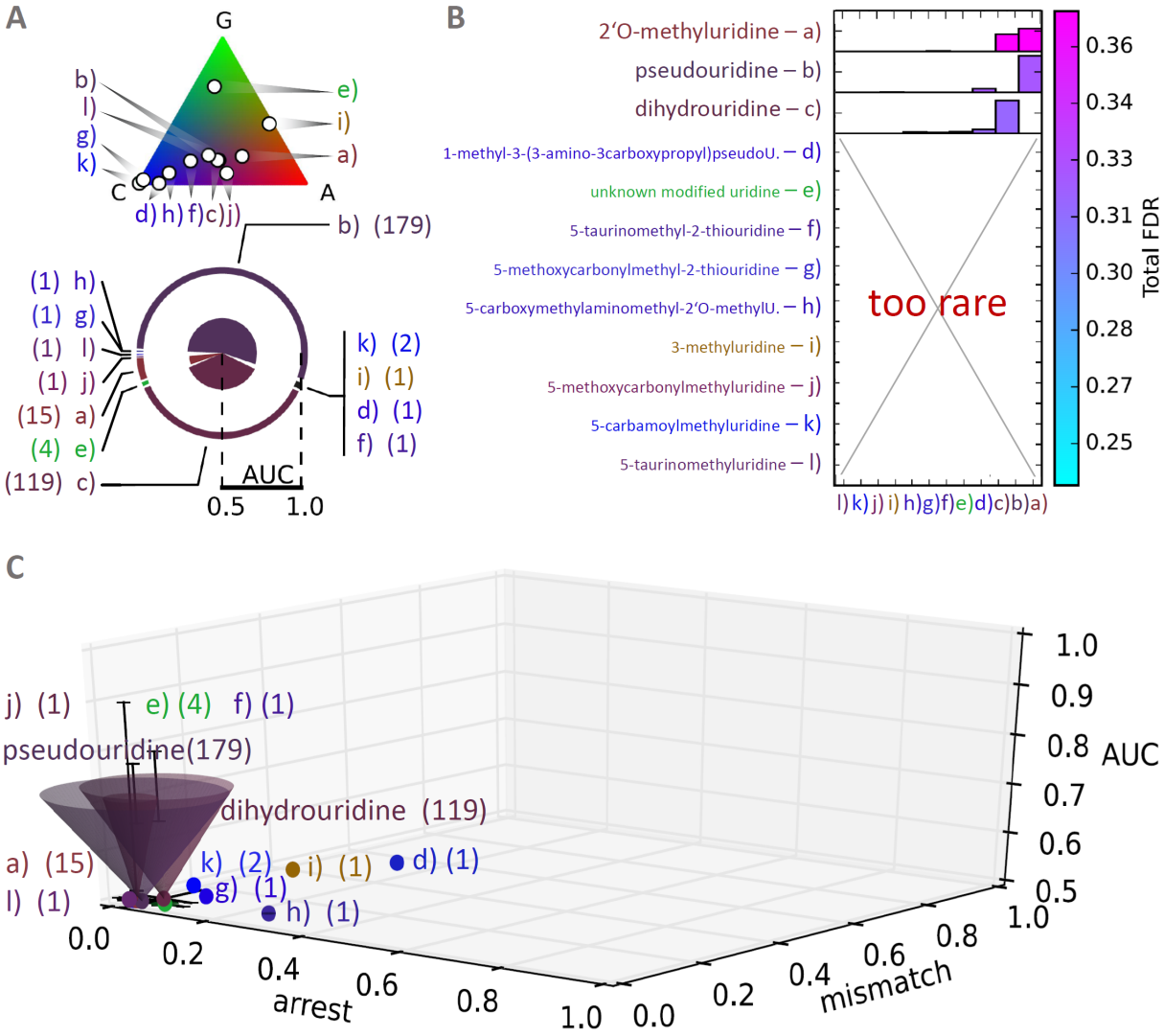


Figure 32: Eligibility chart - Random Forest performance by RT signatures: uridines. All annotated (MODOMICS) modification sites in yeast cyt. tRNA & rRNA and human mitoch. tRNA sequences were grouped by modification type. Results were determined in 10 repetitions of a 5-fold stratified cross-validation using equal amounts of a specific modification (minimum required frequency = 5) vs. a random composition of 'other'-labeled modifications. **(A)** Colors of pie chart code for mismatch composition displayed in ternary plot. Pie radii reflect Area Under Curves (AUC) from Receiver Operating Characteristic (ROC) curves of a Random Forest model tested for discrimination performance of the modification types. Circular fractions of pies represent the relative frequencies (abs. frequencies displayed in round brackets) of 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 medication frequencies). **(C)** Random Forest performance (AUC) represented as cone height vs. arrest rates, mismatch rates and mismatch compositions (colors). Radii were squared (=normalized) for presentation reasons. Black whiskers indicate standard deviations.

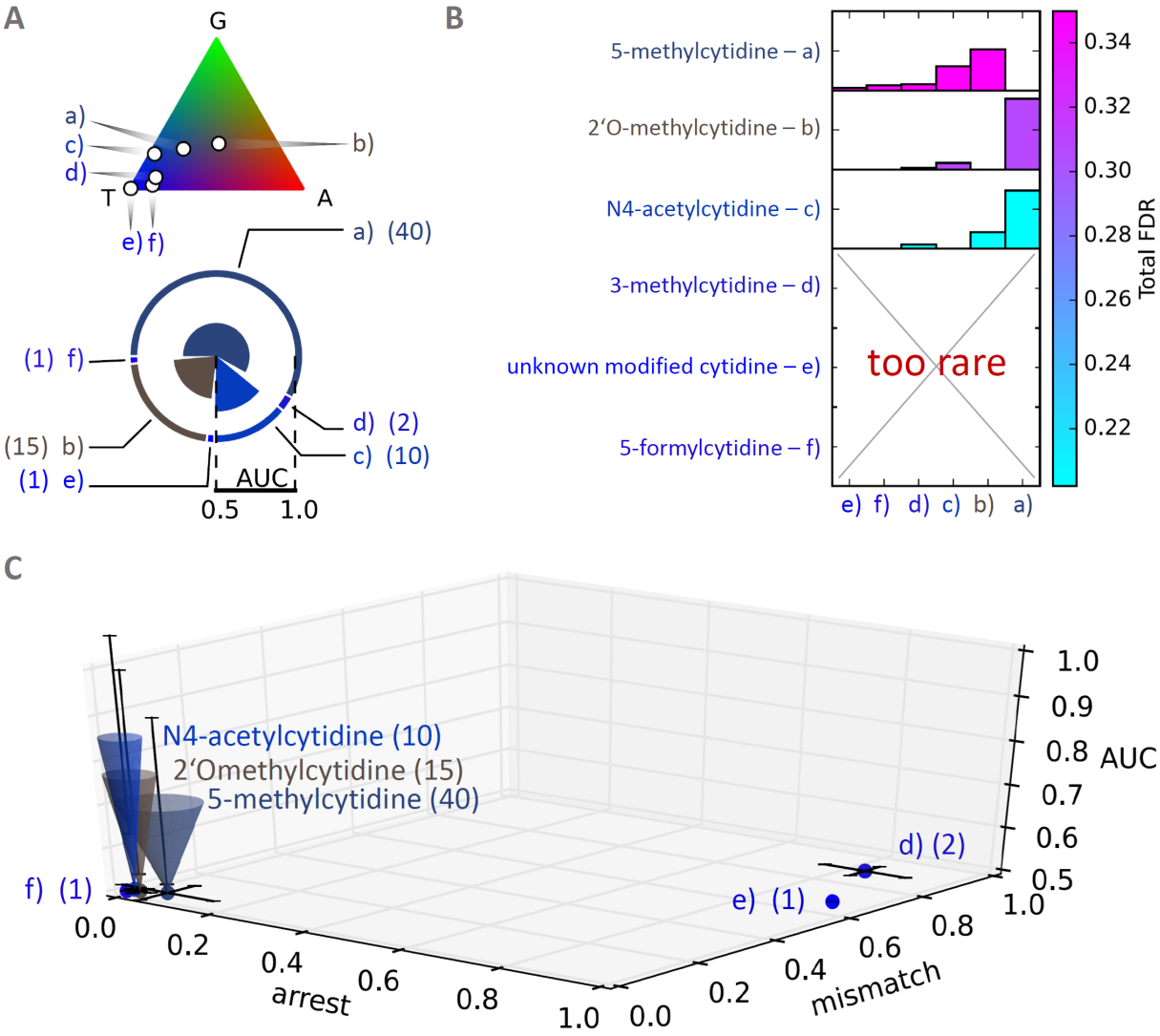


Figure 33: Eligibility chart - Random Forest performance by RT signatures: cytidines. All annotated (MODOMICS) modification sites in yeast cyt. tRNA & rRNA and human mitoch. tRNA sequences were grouped by modification type. Results were determined in 10 repetitions of a 5-fold stratified cross-validation using equal amounts of a specific modification (minimum required frequency = 5) vs. a random composition of 'other'-labeled modifications. **(A)** Colors of pie chart code for mismatch composition displayed in ternary plot. Pie radii reflect Area Under Curves (AUC) from Receiver Operating Characteristic (ROC) curves of a Random Forest model tested for discrimination performance of the modification types. Circular fractions of pies represent the relative frequencies (abs. frequencies displayed in round brackets) of 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 medication frequencies). **(C)** Random Forest performance (AUC) represented as cone height vs. arrest rates, mismatch rates and mismatch compositions (colors). Radii were squared (=normalized) for presentation reasons. Black whiskers indicate standard deviations.

6.8 LC-MC/MS

Table 9: QQQ parameters of dynamic MRM method. Adopted from *Hauenschild et al. 2015* [113]. Measurements done by Katharina Schmid and Kathrin Thüring.

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 10: LC-MS/MS quantification of m^{1}A. Adopted from *Hauenschild et al. 2015* [113]. Measurements done by Katharina Schmid and Kathrin Thüring.

ID	Sample	Material	Interest	m^{1}A/molecule	% m^{1}A/A
3	*S. cer.* Δtrm6	total tRNA	m^{1}A$_{58}$ knockout		0.04
4	*S. cer.* wt	total tRNA	positive control		3.89
5	*S. cer.* 25S wt	rRNA	wildtype	1.50	
6	*S. cer.* 25S Δrrp8	rRNA	single knockout m^{1}A$_{645}$	0.73	
7	*S. cer.* 25S Δbmt2	rRNA	single knockout m^{1}A$_{2142}$	0.77	
8	*S. cer.* 25S Δrrp8 + Δbmt2	rRNA	double knockout m^{1}A$_{645}$ and m^{1}A$_{2142}$	0.01	
9	*S. pactum*	total RNA	m^{1}A on SSU of rRNA		0.15
10	*H. sapiens*	rRNA	Homologous identification		0.15
11	*M. musculus*	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	in vitro transcr.	negative control	0.00	
S24	Signature vs. occupancy	tRNA^{Arg_UCG}	Novel site	1.01	

Bibliography

[1] B. R. Graveley, "Alternative splicing: increasing diversity in the proteomic world," *TRENDS in Genetics*, vol. 17, no. 2, pp. 100–107, 2001.

[2] D. C. Di Giammartino, K. Nishida, and J. L. Manley, "Mechanisms and consequences of alternative polyadenylation," *Molecular cell*, vol. 43, no. 6, pp. 853–866, 2011.

[3] A. J. Hamilton and D. C. Baulcombe, "A species of small antisense rna in posttranscriptional gene silencing in plants," *Science*, vol. 286, no. 5441, pp. 950–952, 1999.

[4] 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 Res*, vol. 41, no. Database issue, pp. D262–7, 2013.

[5] M. Roovers, J. Wouters, J. M. Bujnicki, C. Tricot, V. Stalon, H. Grosjean, and L. Droogmans, "A primordial rna modification enzyme: the case of trna (m1a) methyltransferase," *Nucleic acids research*, vol. 32, no. 2, pp. 465–476, 2004.

[6] F. F. Davis and F. W. Allen, "Ribonucleic acids from yeast which contain a fifth nucleotide," *J Biol Chem*, vol. 227, no. 2, pp. 907–915, 1957.

[7] Y. Saletore, K. Meyer, J. Korlach, I. D. Vilfan, S. Jaffrey, and C. E. Mason, "The birth of the epitranscriptome: deciphering the function of rna modifications," *Genome Biol*, vol. 13, no. 10, p. 175, 2012.

[8] T. Suzuki, "A complete landscape of post-transcriptional modifications in mammalian mitochondrial trnas," *Nucleic Acids Res*, vol. 42, no. 11, pp. 7346–7357, 2014.

[9] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, D. Baltimore, and J. Darnell, "Processing of rrna and trna," 2000.

[10] H. Grosjean, *Fine-tuning of RNA functions by modification and editing*. Springer, 2005.

[11] E. M. Phizicky and J. D. Alfonzo, "Do all modifications benefit all trnas?," *FEBS letters*, vol. 584, no. 2, pp. 265–271, 2010.

[12] T. Suzuki, A. Nagao, and T. Suzuki, "Human mitochondrial trnas: biogenesis, function, structural aspects, and diseases," *Annual review of genetics*, vol. 45, pp. 299–329, 2011.

[13] J. E. Jackman and J. D. Alfonzo, "Transfer rna modifications: nature's combinatorial chemistry playground," *Wiley Interdisciplinary Reviews: RNA*, vol. 4, no. 1, pp. 35–48, 2013.

[14] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg, "Compilation of trna sequences and sequences of trna genes," *Nucleic acids research*, vol. 26, no. 1, pp. 148–153, 1998.

[15] Y. S. Polikanov, S. V. Melnikov, D. Söll, and T. A. Steitz, "Structural insights into the role of rrna modifications in protein synthesis and ribosome assembly," *Nature structural & molecular biology*, vol. 22, no. 4, pp. 342–344, 2015.

[16] Y. Motorin and M. Helm, "trna stabilization by modified nucleotides," *Biochemistry*, vol. 49, no. 24, pp. 4934–4944, 2010.

[17] K. R. Noon, R. Guymon, P. F. Crain, J. A. McCloskey, M. Thomm, J. Lim, and R. Cavicchioli, "Influence of temperature on trna modification in archaea: Methanococcoides burtonii (optimum growth temperature [topt], 23 c) and stetteria hydrogenophila (topt, 95 c)," *Journal of bacteriology*, vol. 185, no. 18, pp. 5483–5490, 2003.

[18] R. K. Kumar and D. R. Davis, "Synthesis and studies on the effect of 2-thiouridine and 4-thiouridine on sugar conformation and rna duplex stability," *Nucleic acids research*, vol. 25, no. 6, pp. 1272–1280, 1997.

[19] P. Romby, P. Carbon, E. Westhof, C. Ehresmann, J.-P. Ebel, B. Ehresmann, and R. Giegé, "Importance of conserved residues for the conformation of the t-loop in trnas," *Journal of Biomolecular Structure and Dynamics*, vol. 5, no. 3, pp. 669–687, 1987.

[20] D. E. Bergstrom and N. J. Leonard, "Photoreaction of 4-thiouracil with cytosine. relation to photoreactions in escherichia coli transfer ribonucleic acids," *Biochemistry*, vol. 11, no. 1, pp. 1–9, 1972.

[21] A. Favre, A. Michelson, and M. Yaniv, "Photochemistry of 4-thiouridine in escherichia coli transfer rna 1 val," *Journal of molecular biology*, vol. 58, no. 1, pp. 367–379, 1971.

[22] S. Gehrig, M.-E. Eberle, F. Botschen, K. Rimbach, F. Eberle, T. Eigenbrod, S. Kaiser, W. M. Holmes, V. A. Erdmann, M. Sprinzl, *et al.*, "Identification of modifications in microbial, native trna that suppress immunostimulatory activity," *The Journal of experimental medicine*, vol. 209, no. 2, pp. 225–233, 2012.

[23] S. Jöckel, G. Nees, R. Sommer, Y. Zhao, D. Cherkasov, H. Hori, G. Ehm, M. Schnare, M. Nain, A. Kaufmann, *et al.*, "The 2'-o-methylation status of a single guanosine controls transfer rna–mediated toll-like receptor 7 activation or inhibition," *The Journal of experimental medicine*, vol. 209, no. 2, pp. 235–241, 2012.

[24] B. El Yacoubi, M. Bailly, and V. de Crécy-Lagard, "Biosynthesis and function of posttranscriptional modifications of transfer rnas," *Annual review of genetics*, vol. 46, pp. 69–95, 2012.

[25] H. Grosjean, D. Söll, and D. Crothers, "Studies of the complex between transfer rnas with complementary anticodons: I. origins of enhanced affinity between complementary triplets," *Journal of molecular biology*, vol. 103, no. 3, pp. 499–519, 1976.

[26] G. R. Björk, K. Jacobsson, K. Nilsson, M. J. Johansson, A. S. Byström, and O. P. Persson, "A primordial trna modification required for the evolution of life?," *The EMBO journal*, vol. 20, no. 1-2, pp. 231–239, 2001.

[27] A. G. Torres, E. Batlle, and L. R. de Pouplana, "Role of trna modifications in human diseases," *Trends in molecular medicine*, vol. 20, no. 6, pp. 306–314, 2014.

[28] E. M. Phizicky and A. K. Hopper, "trna biology charges to the front," *Genes & development*, vol. 24, no. 17, pp. 1832–1860, 2010.

[29] Y.-T. Yu, M.-D. Shu, and J. A. Steitz, "Modifications of u2 snrna are required for snrnp assembly and pre-mrna splicing," *The EMBO Journal*, vol. 17, no. 19, pp. 5783–5795, 1998.

[30] V. Evdokimova, P. Ruzanov, H. Imataka, B. Raught, Y. Svitkin, L. P. Ovchinnikov, and N. Sonenberg, "The major mrna-associated protein yb-1 is a potent 5' cap-dependent mrna stabilizer," *The EMBO journal*, vol. 20, no. 19, pp. 5491–5502, 2001.

[31] G. Jia, Y. Fu, X. Zhao, Q. Dai, G. Zheng, Y. Yang, C. Yi, T. Lindahl, T. Pan, Y.-G. Yang, *et al.*, "N6-methyladenosine in nuclear rna is a major substrate of the obesity-associated fto," *Nature chemical biology*, vol. 7, no. 12, pp. 885–887, 2011.

[32] C. Benedict, J. A. Jacobsson, E. Rönnemaa, M. Sällman-Almén, S. Brooks, B. Schultes, R. Fredriksson, L. Lannfelt, L. Kilander, and H. B. Schiöth, "The fat mass and obesity gene is linked to reduced verbal fluency in overweight and obese elderly men," *Neurobiology of aging*, vol. 32, no. 6, pp. 1159–e1, 2011.

[33] L. Keller, W. Xu, H.-X. Wang, B. Winblad, L. Fratiglioni, and C. Graff, "The obesity related gene, fto, interacts with apoe, and is associated with alzheimer's disease risk: a prospective cohort study," *Journal of Alzheimer's Disease*, vol. 23, no. 3, pp. 461–469, 2011.

[34] K. D. Meyer and S. R. Jaffrey, "The dynamic epitranscriptome: N6-methyladenosine and gene expression control," *Nature Reviews Molecular Cell Biology*, vol. 15, no. 5, pp. 313–326, 2014.

[35] T. P. Hoernes, N. Clementi, K. Faserl, H. Glasner, K. Breuker, H. Lindner, A. Hüttenhofer, and M. D. Erlacher, "Nucleotide modifications within bacterial messenger rnas regulate their translation and are able to rewire the genetic code," *Nucleic acids research*, p. gkv1182, 2015.

[36] L. Warren, P. D. Manos, T. Ahfeldt, Y.-H. Loh, H. Li, F. Lau, W. Ebina, P. K. Mandal, Z. D. Smith, A. Meissner, *et al.*, "Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mrna," *Cell stem cell*, vol. 7, no. 5, pp. 618–630, 2010.

[37] S. Kellner, A. Ochel, K. Thuring, 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 Res*, vol. 42, no. 18, p. e142, 2014.

[38] C. Brandmayr, M. Wagner, T. Brückl, D. Globisch, D. Pearson, A. C. Kneuttinger, V. Reiter, A. Hienzsch, S. Koch, I. Thoma, *et al.*, "Isotope-based analysis of modified trna nucleosides correlates modification density with translational efficiency," *Angewandte Chemie International Edition*, vol. 51, no. 44, pp. 11162–11165, 2012.

[39] S. L. Hiley, J. Jackman, T. Babak, M. Trochesset, Q. D. Morris, E. Phizicky, and T. R. Hughes, "Detection and discovery of rna modifications using microarrays," *Nucleic acids research*, vol. 33, no. 1, pp. e2–e2, 2005.

[40] 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 Res*, vol. 13, no. 23, pp. 8339–8357, 1985.

[41] E. Y. Levanon, E. Eisenberg, R. Yelin, S. Nemzer, M. Hallegger, R. Shemesh, Z. Y. Fligelman, A. Shoshan, S. R. Pollock, D. Sztybel, *et al.*, "Systematic identification of abundant a-to-i editing sites in the human transcriptome," *Nature biotechnology*, vol. 22, no. 8, pp. 1001–1005, 2004.

[42] Z. Wang, M. Gerstein, and M. Snyder, "Rna-seq: a revolutionary tool for transcriptomics," *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57–63, 2009.

[43] 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.

[44] G. Zheng, Y. Qin, W. C. Clark, Q. Dai, C. Yi, C. He, A. M. Lambowitz, and T. Pan, "Efficient and quantitative high-throughput trna sequencing," *Nature methods*, vol. 12, no. 9, pp. 835–837, 2015.

[45] 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. 143–146, 2014.

[46] 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.

[47] S. Schwartz, D. A. Bernstein, M. R. Mumbach, M. Jovanovic, R. H. Herbst, B. X. Leon-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, no. 1, pp. 148–162, 2014.

[48] 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, no. 8, pp. 592–597, 2015.

[49] C. Hoang and A. R. Ferré-D'Amaré, "Cocrystal structure of a trna ψ55 pseudouridine synthase: nucleotide flipping by an rna-modifying enzyme," *Cell*, vol. 107, no. 7, pp. 929–939, 2001.

[50] J. Ge and Y.-T. Yu, "Rna pseudouridylation: new insights into an old modification," *Trends in biochemical sciences*, vol. 38, no. 4, pp. 210–218, 2013.

[51] K. Karikó, H. Muramatsu, J. M. Keller, and D. Weissman, "Increased erythropoiesis in mice injected with submicrogram quantities of pseudouridine-containing mrna encoding erythropoietin," *Molecular Therapy*, 2012.

[52] J. Karijolich and Y.-T. Yu, "Converting nonsense codons into sense codons by targeted pseudouridylation," *Nature*, vol. 474, no. 7351, pp. 395–398, 2011.

[53] S. R. Jaffrey, "An expanding universe of mrna modifications," *Nature structural & molecular biology*, vol. 21, no. 11, pp. 945–946, 2014.

[54] 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, no. 10, pp. 733–740, 2010.

[55] 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*, vol. 127, no. 2, pp. 461–465, 2015.

[56] I. Behm-Ansmant, M. Helm, and Y. Motorin, "Use of specific chemical reagents for detection of modified nucleotides in rna," *J Nucleic Acids*, vol. 2011, p. 408053, 2011.

[57] M. Schaefer, T. Pollex, K. Hanna, and F. Lyko, "Rna cytosine methylation analysis by bisulfite sequencing," *Nucleic Acids Res*, vol. 37, no. 2, p. e12, 2009.

[58] 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 m 5 c within archaeal mrnas," *PLoS Genet*, vol. 9, no. 6, p. e1003602, 2013.

[59] S. Hussain, J. Aleksic, S. Blanco, S. Dietmann, and M. Frye, "Characterizing 5-methylcytosine in the mammalian epitranscriptome," *Genome Biol*, vol. 14, no. 11, p. 215, 2013.

[60] J. E. Squires, H. R. Patel, M. Nousch, 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*, p. gks144, 2012.

[61] 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.

[62] 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, no. 7397, pp. 201–206, 2012.

[63] D. Dominissini, S. Moshitch-Moshkovitz, M. Salmon-Divon, N. Amariglio, and G. Rechavi, "Transcriptome-wide mapping of n6-methyladenosine by m6a-seq based on immunocapturing and massively parallel sequencing," *Nature protocols*, vol. 8, no. 1, pp. 176–189, 2013.

[64] 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, no. 7, pp. 1635–1646, 2012.

[65] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A.-C. Jungkamp, M. Munschauer, *et al.*, "Transcriptome-wide identification of rna-binding protein and microrna target sites by par-clip," *Cell*, vol. 141, no. 1, pp. 129–141, 2010.

[66] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan, "N6-methyladenosine-dependent rna structural switches regulate rna-protein interactions," *Nature*, vol. 518, no. 7540, pp. 560–564, 2015.

[67] K. Chen, Z. Lu, X. Wang, Y. Fu, G.-Z. Luo, N. Liu, D. Han, D. Dominissini, Q. Dai, T. Pan, *et al.*, "High-resolution n6-methyladenosine (m6a) map using photo-crosslinking-assisted m6a sequencing," *Angewandte Chemie International Edition*, vol. 54, no. 5, pp. 1587–1590, 2015.

[68] 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, no. 8, pp. 767–772, 2015.

[69] T. Pan, "N6-methyl-adenosine modification in messenger and long non-coding rna," *Trends in biochemical sciences*, vol. 38, no. 4, pp. 204–209, 2013.

[70] B. Delatte, F. Wang, L. V. Ngoc, E. Collignon, E. Bonvin, R. Deplus, E. Calonne, B. Hassabi, P. Putmans, S. Awe, *et al.*, "Transcriptome-wide distribution and function of rna hydroxymethylcytosine," *Science*, vol. 351, no. 6270, pp. 282–285, 2016.

[71] X. Li, X. Xiong, K. Wang, L. Wang, X. Shu, S. Ma, and C. Yi, "Transcriptome-wide mapping reveals reversible and dynamic n1-methyladenosine methylome," *Nat Chem Biol*, vol. advance online publication, Feb 2016. Article Report.

[72] N. Shigi, T. Suzuki, M. Tamakoshi, T. Oshima, and K. Watanabe, "Conserved bases in the tψc loop of trna are determinants for thermophile-specific 2-thiouridylation at position 54," *Journal of Biological Chemistry*, vol. 277, no. 42, pp. 39128–39135, 2002.

[73] L. Droogmans, M. Roovers, J. M. Bujnicki, C. Tricot, T. Hartsch, V. Stalon, and H. Grosjean, "Cloning and characterization of trna (m1a58) methyltransferase (trmi) from thermus thermophilus hb27, a protein required for cell growth at extreme temperatures," *Nucleic acids research*, vol. 31, no. 8, pp. 2148–2156, 2003.

[74] J. Anderson, L. Phan, and A. G. Hinnebusch, "The gcd10p/gcd14p complex is the essential two-subunit trna(1-methyladenosine) methyltransferase of saccharomyces cerevisiae," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 10, pp. 5173–5178, 2000.

[75] S. Kadaba, A. Krueger, T. Trice, A. M. Krecic, A. G. Hinnebusch, and J. Anderson, "Nuclear surveillance and degradation of hypomodified initiator trnamet in s. cerevisiae," *Genes Dev*, vol. 18, no. 11, pp. 1227–1240, 2004.

[76] M. Kempenaers, M. Roovers, Y. Oudjama, K. L. Tkaczuk, J. M. Bujnicki, and L. Droogmans, "New archaeal methyltransferases forming 1-methyladenosine or 1-methyladenosine and 1-methylguanosine at position 9 of trna," *Nucleic acids research*, vol. 38, no. 19, pp. 6533–6543, 2010.

[77] M. Sakurai, T. Ohtsuki, and K. Watanabe, "Modification at position 9 with 1-methyladenosine is crucial for structure and function of nematode mitochondrial trnas lacking the entire t-arm," *Nucleic acids research*, vol. 33, no. 5, pp. 1653–1661, 2005.

[78] M. Helm, H. Brule, F. Degoul, C. Cepanec, J. P. Leroux, R. Giege, and C. Florentz, "The presence of modified nucleotides is required for cloverleaf folding of a human mitochondrial trna," *Nucleic Acids Res*, vol. 26, no. 7, pp. 1636–1643, 1998.

[79] F. Voigts-Hoffmann, M. Hengesbach, A. Y. Kobitski, A. Van Aerschot, P. Herdewijn, G. U. Nienhaus, and M. Helm, "A methyl group controls conformational equilibrium in human mitochondrial trnalys," *Journal of the American Chemical Society*, vol. 129, no. 44, pp. 13382–13383, 2007.

[80] C. Peifer, S. Sharma, P. Watzinger, S. Lamberth, P. Kötter, and K.-D. Entian, "Yeast rrp8p, a novel methyltransferase responsible for m1a 645 base modification of 25s rrna," *Nucleic acids research*, vol. 41, no. 2, pp. 1151–1163, 2013.

[81] S. Sharma, P. Watzinger, P. Kötter, and K.-D. Entian, "Identification of a novel methyltransferase, bmt2, responsible for the n-1-methyl-adenosine base modification of 25s rrna in saccharomyces cerevisiae," *Nucleic acids research*, vol. 41, no. 10, pp. 5428–5443, 2013.

[82] W. Schmidt, H. H. Arnold, and H. Kersten, "Biosynthetic pathway of ribothymidine in b. subtilis and m. lysodeikticus involving different coenzymes for transfer rna and ribosomal rna," *Nucleic Acids Res*, vol. 2, no. 7, pp. 1043–1051, 1975.

[83] R. Srivastava and K. P. Gopinathan, "Ribosomal rna methylation in mycobacterium smegmatis sn2," *Biochem Int*, vol. 15, no. 6, pp. 1179–1188, 1987.

[84] 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," *J Bacteriol*, vol. 173, no. 22, pp. 7213–7218, 1991.

[85] 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 n1-methyladenosine methylome in eukaryotic messenger rna," *Nature*, vol. advance online publication, Feb 2016. Article.

[86] J. D. Finkelstein and J. J. Martin, "Homocysteine," *The international journal of biochemistry & cell biology*, vol. 32, no. 4, pp. 385–389, 2000.

[87] H. Takuma, N. Ushio, M. Minoji, A. Kazayama, N. Shigi, A. Hirata, C. Tomikawa, A. Ochi, and H. Hori, "Substrate trna recognition mechanism of eubacterial trna (m1a58) methyltransferase (trmi)," *J Biol Chem*, vol. 290, no. 9, pp. 5912–5925, 2015.

[88] D. Hamdane, A. Guelorget, V. Guérineau, and B. Golinelli-Pimpaneau, "Dynamics of rna modification by a multi-site-specific trna methyltransferase," *Nucleic acids research*, vol. 42, no. 18, pp. 11697–11706, 2014.

[89] E. Vilardo, C. Nachbagauer, A. Buzet, A. Taschner, J. Holzmann, and W. Rossmanith, "A subcomplex of human mitochondrial rnase p is a bifunctional methyltransferase-extensive moonlighting in mitochondrial trna biogenesis," *Nucleic acids research*, p. gks910, 2012.

[90] Y. Motorin, S. Muller, I. Behm-Ansmant, and C. Branlant, "Identification of modified residues in rnas by reverse transcription-based methods," *Methods Enzymol*, vol. 425, pp. 21–53, 2007.

[91] J. G. Underwood, A. V. Uzilov, S. Katzman, C. S. Onodera, J. E. Mainzer, D. H. Mathews, T. M. Lowe, S. R. Salama, and D. Haussler, "Fragseq: transcriptome-wide rna structure probing using high-throughput sequencing," *Nature methods*, vol. 7, no. 12, pp. 995–1001, 2010.

[92] M. Kertesz, Y. Wan, E. Mazor, J. L. Rinn, R. C. Nutter, H. Y. Chang, and E. Segal, "Genome-wide measurement of rna secondary structure in yeast," *Nature*, vol. 467, no. 7311, pp. 103–107, 2010.

[93] J. B. Lucks, S. A. Mortimer, C. Trapnell, S. Luo, S. Aviran, G. P. Schroth, L. Pachter, J. A. Doudna, and A. P. Arkin, "Multiplexed rna structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (shape-seq)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 27, pp. 11063–11068, 2011.

[94] S. A. Mortimer, C. Trapnell, S. Aviran, L. Pachter, and J. B. Lucks, "Shape-seq: High-throughput rna structure analysis," *Curr Protoc Chem Biol*, vol. 4, no. 4, pp. 275–297, 2012.

[95] J. Talkish, G. May, Y. Lin, J. L. Woolford, and C. J. McManus, "Mod-seq: high-throughput sequencing for chemical probing of rna structure," *RNA (New York, N.Y.)*, vol. 20, no. 5, pp. 713–720, 2014.

[96] S. Auxilien, G. Keith, S. F. Le Grice, and J. L. Darlix, "Role of post-transcriptional modifications of primer trnalys,3 in the fidelity and efficacy of plus strand dna transfer during hiv-1 reverse transcription," *J Biol Chem*, vol. 274, no. 7, pp. 4412–4420, 1999.

[97] M. J. Renda, J. D. Rosenblatt, E. Klimatcheva, L. M. Demeter, R. A. Bambara, and V. Planelles, "Mutation of the methylated trna(lys)(3) residue a58 disrupts reverse transcription and inhibits replication of human immunodeficiency virus type 1," *J Virol*, vol. 75, no. 20, pp. 9671–9678, 2001.

[98] Y. Zhang, F. Yuan, X. Wu, O. Rechkoblit, J.-S. Taylor, N. E. Geacintov, and Z. Wang, "Error-prone lesion bypass by human dna polymerase η," *Nucleic Acids Research*, vol. 28, no. 23, pp. 4717–4724, 2000.

[99] H. Yang, Y. Zhan, D. Fenn, L. M. Chi, and S. L. Lam, "Effect of 1-methyladenine on double-helical dna structures," *FEBS letters*, vol. 582, no. 11, pp. 1629–1633, 2008.

[100] H. Yang and S. L. Lam, "Effect of 1-methyladenine on thermodynamic stabilities of double-helical dna structures," *FEBS letters*, vol. 583, no. 9, pp. 1548–1553, 2009.

[101] S. Findeiss, D. Langenberger, P. F. Stadler, and S. Hoffmann, "Traces of post-transcriptional rna modifications in deep sequencing data," *Biol Chem*, vol. 392, no. 4, pp. 305–313, 2011.

[102] 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.

[103] S. Blanco, S. Dietmann, J. V. Flores, S. Hussain, C. Kutter, P. Humphreys, M. Lukk, P. Lombard, L. Treps, M. Popis, *et al.*, "Aberrant methylation of trnas links cellular stress to neuro-developmental disorders," *The EMBO journal*, vol. 33, no. 18, pp. 2020–2039, 2014.

[104] A. Hodgkinson, Y. Idaghdour, E. Gbeha, J.-C. Grenier, E. Hip-Ki, V. Bruat, J.-P. Goulet, T. de Malliard, and P. Awadalla, "High-resolution genomic analysis of human mitochondrial rna sequence variation," *Science*, vol. 344, no. 6182, pp. 413–415, 2014.

[105] 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, no. 12, pp. 1684–1692, 2013.

[106] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," *Nat Methods*, vol. 9, no. 4, pp. 357–359, 2012.

[107] D. Sims, I. Sudbery, N. E. Ilott, A. Heger, and C. P. Ponting, "Sequencing depth and coverage: key considerations in genomic analyses," *Nature Reviews Genetics*, vol. 15, no. 2, pp. 121–132, 2014.

[108] M.-s. Kim, B. Hur, and S. Kim, "Rddpred: a condition-specific rna-editing prediction model from rna-seq data," *BMC Genomics*, vol. 17, no. 1, p. 85, 2016.

[109] B. Panwar and G. P. Raghava, "Prediction of uridine modifications in trna sequences," *BMC bioinformatics*, vol. 15, no. 1, p. 326, 2014.

[110] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300, 1995.

[111] M. Zaringhalam and F. N. Papavasiliou, "Pseudouridylation meets next-generation sequencing," *Methods*, 2016.

[112] W.-J. Sun, J.-H. Li, S. Liu, J. Wu, H. Zhou, L.-H. Qu, and J.-H. Yang, "Rmbase: a resource for decoding the landscape of rna modifications from high-throughput sequencing data," *Nucleic acids research*, p. gkv1036, 2015.

[113] 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, no. 20, pp. 9950–9964, 2015.

[114] H. Cahová, M. L. Winz, K. Höfer, G. Nübel, and A. Jäschke, "Nad captureseq indicates nad as a bacterial cap for a subset of regulatory rnas," *Nature*, vol. 519, no. 7543, pp. 374–377, 2015.

[115] M. L. Winz, "Biological, chemical and computational investigations on rna function and modification, ph.d. thesis, heidelberg university.," 2014.

[116] T. M. Lowe and S. R. Eddy, "trnascan-se: a program for improved detection of transfer rna genes in genomic sequence," *Nucleic acids research*, vol. 25, no. 5, pp. 955–964, 1997.

[117] R. Giege, M. Sissler, and C. Florentz, "Universal rules and idiosyncratic features in trna identity," *Nucleic Acids Res*, vol. 26, no. 22, pp. 5017–5035, 1998.

[118] J. D. Engel, "Mechanism of the dimroth rearrangement in adenosine," *Biochemical and biophysical research communications*, vol. 64, no. 2, pp. 581–586, 1975.

[119] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[120] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[121] Liaw, A., & Wiener, M, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[122] N. B. Leontis and E. Westhof, "Conserved geometrical base-pairing patterns in rna," *Quarterly reviews of biophysics*, vol. 31, no. 04, pp. 399–455, 1998.

[123] B. E. Maden, "The numerous modified nucleotides in eukaryotic ribosomal rna," *Prog Nucleic Acid Res Mol Biol*, vol. 39, pp. 241–303, 1990.

[124] M. A. Rubio, Z. Paris, K. W. Gaston, I. M. Fleming, P. Sample, C. R. Trotta, and J. D. Alfonzo, "Unusual noncanonical intron editing is important for trna splicing in trypanosoma brucei," *Mol Cell*, vol. 52, no. 2, pp. 184–192, 2013.

[125] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC bioinformatics*, vol. 9, no. 1, p. 1, 2008.

[126] R. L. Somorjai, B. Dolenko, and R. Baumgartner, "Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions," *Bioinformatics*, vol. 19, no. 12, pp. 1484–1491, 2003.

[127] A. Bureau, J. Dupuis, K. Falls, K. L. Lunetta, B. Hayward, T. P. Keith, and P. Van Eerdewegh, "Identifying snps predictive of phenotype using random forests," *Genetic epidemiology*, vol. 28, no. 2, pp. 171–182, 2005.

[128] D. L. Sampson, T. J. Parker, Z. Upton, and C. P. Hurst, "A comparison of methods for classifying clinical samples based on proteomics data: a case study for statistical and machine learning approaches," *PloS one*, vol. 6, no. 9, p. e24973, 2011.

[129] W. G. Touw, J. R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, and S. A. van Hijum, "Data mining in the life sciences with random forest: a walk in the park or lost in the jungle?," *Briefings in bioinformatics*, p. bbs034, 2012.

[130] R. Bellman, "Dynamic programming," *Rand Corporation*, p. 342, 1957.

[131] T. K. Ho, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 20, no. 8, pp. 832–844, 1998.

[132] M. W. Mitchell, "Bias of the random forest out-of-bag (oob) error for certain input parameters," *Open Journal of Statistics*, vol. 1, no. 03, p. 205, 2011.

[133] E. J. Humphrey, J. P. Bello, and Y. LeCun, "Moving beyond feature design: Deep architectures and automatic feature learning in music informatics.," in *ISMIR*, pp. 403–408, Citeseer, 2012.

[134] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Rocr: visualizing classifier performance in r," *Bioinformatics*, vol. 21, no. 20, pp. 3940–3941, 2005.

[135] C. T. Leondes, *Image processing and pattern recognition*, vol. 5. Elsevier, 1998.

[136] C. Gini, "Variability and mutability, contribution to the study of statistical distribution and relaitons," *Studi Economico-Giuricici della R*, 1912.

[137] C. Strobl and A. Zeileis, "Danger: High power!–exploring the statistical properties of a test for random forest variable importance," in *Proceedings of the 18th International Conference on Computational Statistics, Porto, Portugal*, Citeseer, 2008.

[138] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC bioinformatics*, vol. 8, no. 1, p. 1, 2007.

[139] F. Juhling, M. Morl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Putz, "trnadb 2009: compilation of trna sequences and trna genes," *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D159–62, 2009.

[140] M. Wildauer, G. Zemora, A. Liebeg, V. Heisig, and C. Waldsich, "Chemical probing of rna in living cells," *Methods Mol Biol*, vol. 1086, pp. 159–176, 2014.

[141] Y. Ding, Y. Tang, C. K. Kwok, Y. Zhang, P. C. Bevilacqua, and S. M. Assmann, "In vivo genome-wide profiling of rna secondary structure reveals novel regulatory features," *Nature*, vol. 505, no. 7485, pp. 696–700, 2014.

[142] G. St Laurent, M. R. Tackett, S. Nechkin, D. Shtokalo, D. Antonets, Y. A. Savva, R. Maloney, P. Kapranov, C. E. Lawrence, and R. A. Reenan, "Genome-wide analysis of a-to-i rna editing by single-molecule sequencing in drosophila," *Nature structural & molecular biology*, vol. 20, no. 11, pp. 1333–1339, 2013.

[143] V. Goldschmidt, J. Didierjean, B. Ehresmann, C. Ehresmann, C. Isel, and R. Marquet, "Mg2+ dependency of hiv-1 reverse transcription, inhibition by nucleoside analogues and resistance," *Nucleic Acids Res*, vol. 34, no. 1, pp. 42–52, 2006.

[144] B. E. Maden, "Mapping 2'-o-methyl groups in ribosomal rna," *Methods*, vol. 25, no. 3, pp. 374–382, 2001.

[145] 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, p. 8, 2013.

[146] K. L. Tee and T. S. Wong, "Polishing the craft of genetic diversity creation in directed evolution," *Biotechnol Adv*, vol. 31, no. 8, pp. 1707–1721, 2013.

[147] F. Alings, L. P. Sarin, C. Fufezan, H. C. Drexler, and S. A. Leidel, "An evolutionary approach uncovers a diverse response of trna 2-thiolation to elevated temperatures in yeast," *RNA*, vol. 21, no. 2, pp. 202–212, 2015.

[148] L. Han, Y. Kon, and E. M. Phizicky, "Functional importance of psi38 and psi39 in distinct trnas, amplified for trnagln(uug) by unexpected temperature sensitivity of the s2u modification in yeast," *RNA*, vol. 21, no. 2, pp. 188–201, 2015.

[149] K. Ochi, J.-Y. Kim, Y. Tanaka, G. Wang, K. Masuda, H. Nanamiya, S. Okamoto, S. Tokuyama, Y. Adachi, and F. Kawamura, "Inactivation of ksga, a 16s rrna methyltransferase, causes vigorous emergence of mutants with high-level kasugamycin resistance," *Antimicrobial agents and chemotherapy*, vol. 53, no. 1, pp. 193–201, 2009.

[150] T. L. Helser, J. E. Davies, and J. E. Dahlberg, "Change in methylation of 16s ribosomal rna associated with mutation to kasugamycin resistance in escherichia coli," *Nat New Biol*, vol. 233, no. 35, pp. 12–14, 1971.

[151] J. D. Hunter *et al.*, "Matplotlib: A 2d graphics environment," *Computing in science and engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[152] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and Genome Project Data Processing, Subgroup, "The sequence alignment/map format and samtools," *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, 2009.

[153] C. Bonferroni, "Sulle medie multiple di potenze," *Bollettino dell'Unione Matematica Italiana*, vol. 5, no. 3-4, pp. 267–270, 1950.

[154] P. P. Łabaj, G. G. Leparc, B. E. Linggi, L. M. Markillie, H. S. Wiley, and D. P. Kreil, "Characterization and improvement of rna-seq precision in quantitative transcript expression profiling," *Bioinformatics*, vol. 27, no. 13, pp. i383–i391, 2011.

[155] A. Roberts, L. Pachter, *et al.*, "Rna-seq and find: entering the rna deep field," *Genome Med*, vol. 3, no. 11, pp. 74–74, 2011.

[156] M. Pertea, "The human transcriptome: an unfinished story," *Genes*, vol. 3, no. 3, pp. 344–360, 2012.

[157] L. J. Kielpinski and J. Vinther, "Massive parallel-sequencing-based hydroxyl radical probing of rna accessibility," *Nucleic acids research*, p. gku167, 2014.

[158] A. Noma, S. Yi, T. Katoh, Y. Takai, T. Suzuki, and T. Suzuki, "Actin-binding protein abp140 is a methyltransferase for 3-methylcytidine at position 32 of trnas in saccharomyces cerevisiae," *RNA*, vol. 17, no. 6, pp. 1111–1119, 2011.

[159] M. D. Shepherd, M. K. Kharel, M. A. Bosserman, and J. Rohr, "Laboratory maintenance of streptomyces species," *Curr Protoc Microbiol*, vol. Chapter 10, p. Unit 10E 1, 2010.

[160] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "Ncbi reference sequences (refseq): a curated nonredundant sequence database of genomes, transcripts and proteins," *Nucleic acids research*, vol. 35, no. suppl 1, pp. D61–D65, 2007.

[161] M. Aslett, C. Aurrecoechea, M. Berriman, J. Brestelli, B. P. Brunk, M. Carrington, D. P. Depledge, S. Fischer, B. Gajria, X. Gao, *et al.*, "Tritrypdb: a functional genomic resource for the trypanosomatidae," *Nucleic acids research*, vol. 38, no. suppl 1, pp. D457–D462, 2010.

[162] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 845–848, 1965.

Acknowledgments

Funding. This work was supported by the DFG SPP1784, MH3397/12-1 and MH3397/8-1 to Prof. Dr. XXXX XXXX and a fellowship to Ralf Hauenschild for the *International PhD Programme* (IPP) at the Institute of Molecular Biology (IMB) Mainz, funded by the Boehringer Ingelheim Foundation.

Personal. It is with immense gratitude that I acknowledge the support by my supervisor Prof. Dr. XXXX XXXX. Working in the highly interdisciplinary environment of his group was an inspiring and enriching experience. Special thanks go to Dr. XXXX XXXXXXXXXX, who became a close friend and certainly knows how much I owe him. Thank you, XXXXX XXXXX, for the patience with your 24/7 scientist boyfriend. I thank all current and former members of the XXXX Group for the nice times we had together! As undeniable contribution to the so far best years in my life, each of the following colleagues and friends made me smile at least once:

XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX,
XXXXXXX XXXXXX,
XXXXXXX XXXXXX, XXXXXXX XXXXXX, Mr. XXXXXXX, Mr. XXXXXXX, XXXXXXX XXXXXX,
XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX,
XXXXXXX XXXXXX,
XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXXd, XXXXXXX XXXXXX, XXXXXXX XXXXXX,
XXXXXXX XXXXXX,
XXXXXXX XXXXXX, XXXXXXX XXXXXX & XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX,
XXXXXXX XXXXXX,
XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX,
XXXXXXX XXXXXX,
XXXXXXX XXXXXX, XXXXXXX XXXXXX, XXXXXXX XXXXXX

As direct contributors to my project, I acknowledge, XXXXXXXXX XXXXXX and XXXXXXX XXXXXXX for the LC-MS/MS analyses, XXXXXXX XXXXXX for help with *S. pactum* cultivation, XXXXXXXX XXXXXXX for RNA extraction, Prof. Dr. XXXX XXXXXXX for sequencing, Prof. Dr. XXXXXXX XXXXXXXXXXX and Prof. Dr. XXXXXX XXXXXXX for their advice and the TAC meetings, Junior Prof. Dr. XXXXX XXXX for readiness as examiner, XXXXX XXXXXX and the IMB for their support and XXXXXXX XXXXXXXXX for his enormous effort in a pivotal role within the Modificators Task Force. Finally, I want to thank XXXXXXX XXXXXX for unterrified endurance and standby in challenging surgeries in the depths of *CoverageAnalyzer*'s programming code.

Statement of Authorship

I declare on oath that this document and the accompanying software have been composed exclusively by myself and describe my own work, unless otherwise acknowledged in the text. The thesis has not been accepted in any previous application for a degree. All verbatim extracts have been distinguished by quotation marks, and all sources of information have been specifically acknowledged.

Mainz, 11 May 2017 

Location, Date First name Surname

Curriculum Vitae

Ralf Johannes Hauenschild

Address: Kanalstraße 8
D-63785 Obernburg
Germany

Nationality: German

Date of Birth: 07 February, 1987

Place of Birth: Erlenbach am Main

Education:

1997-2006	Julius Echter Gymnasium - Center of Excellence, Elsenfeld, Bavaria Allgemeine Hochschulreife (Abitur)
2007-2010	Johann Wolfgang Goethe University, Frankfurt am Main Bachelor of Science in Bioinformatics (B.Sc.)
2010-2012	Johann Wolfgang Goethe University, Frankfurt am Main Master of Science in Bioinformatics (M.Sc.)
2012-2016	Johannes Gutenberg University, Mainz Fellow in the *International PhD Programme* (IPP) of the Institute of Molecular Biology (IMB) as candidate for Dr. rer. nat. in the research group of Prof. Dr. Mark Helm at the Institute of Pharmacy and Biochemistry (IPB).

Publications of contents of this work:

2015	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, no. 20, pp. 9950-9964, 2015.
2016	L. Tserovski, V. Marchand, R. Hauenschild, F. Blanloeil-Oillo, M. Helm and Y. Motorin, „High-throughput sequencing for 1-methyladenosine (m1A) mapping in RNA“, *Elsevier Methods*, vol. 107, pp. 110-121, 2016.

Comment: Due to time constraints, submission of *CoverageAnalyzer* manuscript was retracted from the journal specified in the publication list of the print version of this thesis. The work was instead published as:

2016	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, pp. 42, 2016.